Stony Brook University

Physical polymerization mechanisms in the chemistry-to-biology transition

A Dissertation presented

by

Elizaveta Guseva

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Physics and Astronomy

Stony Brook University

December 2016

Stony Brook University

The Graduate School

Elizaveta Guseva

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation

Ken A. Dill – Dissertation Adviser
Professor, Department of Physics and Astronomy

Marivi Fernandez Serra – Chairperson of Defense
Professor, Department of Physics and Astronomy

Thomas Weinacht – Committee Member
Professor, Department of Physics and Astronomy

Thomas MacCarthy – Outside Committee Member
Professor, Department of Applied Mathematics and Statistics

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Physical polymerization mechanisms in the chemistry-to-biology transition

by

Elizaveta Guseva

Doctor of Philosophy

in

Physics and Astronomy

Stony Brook University

2016

Studying complex systems and emergent phenomena is very popular today. The reason is that we desperately need more knowledge about many complex systems such as cells, organisms, society and emergent phenomena on the internet. Applying physical and quantitative methods to such systems resulted in many discoveries, yet a lot of knowledge is missing. In particular, we dont fully understand living systems including their emergence. What are the minimal requirements for life? How to make a chemical system capable of inheritance and open ended evolution? If a system is capable of Darwinian evolution, is it necessarily a living system? Modern life relies in its functioning (including inheritance and capability to evolve) on long polymeric molecules: proteins and nucleic acids. Because of their indispensable role in cells it is very important to understand the origins of these biological polymers as well as their role in the emergence of inheritance, evolution and metabolism. Are long biological polymers enough to jump-start life? We propose physical mechanisms of emergence of long bio-polymers in the prebiotic world. We use HP lattice model to model polymerization, interaction and folding of short chains of hydrophobic (H) and polar (P) monomers. We show that such chains fold into relatively compact structures exposing hydrophobic patches. These hydrophobic patches act as primitive versions of modern pro-

teins catalytic site and assist in polymerization of other HP-sequences. These HP-sequences form autocatalytic, self-sustaining dynamical systems capable of multimodality: ability to settle at multiple distinct quasi-stable states characterized by different groups of dominating polymers. We study properties of these systems to see their role in the chemistry-to-biology transition. We also propose a stochastic simulation algorithm for modeling agent-based complex systems which is particularly well suited for polymeric systems with several types of monomers. This algorithm is efficient for sparse systems: systems where the number of the species which could possible be generated is much higher than the number of species actually generated. It allows for simulation of systems with unlimited number of molecular species.

Contents

Acknowledgments

This work uses materials from the following papers written in collaboration with Prof. Ronald Zuckermann, Prof. Ken Dill (who supervised this work) and Anton Bernatskiy:

1. Elizaveta A Guseva, Ronald N. Zuckermann, and Ken A. Dill. How did prebiotic polymers become informational foldamers? (https://arxiv.org/abs/1604.08890).

2. Anton V. Bernatskiy and Elizaveta A. Guseva. Exact rule-based stochastic simulations for the system with unlimited number of molecular species. (https://arxiv.org/abs/1609.06403)

All citations are provided with the permission of their corresponding authors

Chapter 1: Knowledge landscape

> I, however, believe that there is at least one philosophical problem in which all thinking people are interested. It is the problem of cosmology: the problem of understanding the world – including ourselves, and our knowledge, as part of the world. All science is cosmology, I believe, and for me the interest of philosophy, no less than of science, lies solely in the contributions which it has made to it. For me, at any rate, both philosophy and science would lose all attraction if they were to give up that pursuit.
>
> Popper [1, 1959, preface]

1.1 Defining life

As human beings we can easily separate living objects from non-living ones or at least we think we can. Some people even argue that life is something that "you know when you see it". It seems as if it was almost clear that life is a special class of physical phenomena. Yet, as scientists, as physicists, can we say it is?

It appears natural to start answering this question with a definition of life. If there is no proper definition how can we argue on the matter? Attempts to define life however encounter significant difficulties. Hundreds of different definitions of life have been suggested over the years. In 2002 alone 40 definitions of life were proposed [2]. And this is precisely the problem: why would we need so many definitions if we know what we are talking about? Indeed, if we look at the multitude of these definitions we can notice that there appears to be a certain inherent problem with the definition of life, as Andy Pross noticed [3]. For almost every definition one can find a population of living organisms that doesn't fit the definition or a non-living object that does. Many the definitions are incompatible with each other. One can find a great deal of the definitions and good discussion in [2, 3].

One of the problems with defining life is that process of defining is to a degree similar to a process of defining a chair. An intuitive definition of a chair would be "a char is what we sit on". Yet somebody can say "I use

chairs to stand on, so according to your definition my chair isn't a chair" or "I sit on my floor. Is my floor also a chair?" The critique which is addressed to the definitions of life is very similar.

As an example we can consider the definition of life by NASA "Life is a self-sustained chemical system capable of undergoing Darwinian evolution"[4]. This is a great definition. But what about a population of males stuck at a Mars colony soon-to-be out of food. They cannot procreate and cannot evolve. At the same time they are clearly alive. One can argue that a population should be considered within a broader context of the whole humanity. Yet if it was a self-sustaining colony with both genders present, we wouldn't need to add extra factors into the definition.

These kind of problems with definitions might stem from enormous complexity of living entities and enormous complexity of the outside world into which they are immersed. Attempts to capture this complexity in one formal definition resulted in the multitude of the definitions each of which is concentrated on certain aspects of living systems. Because the question of origin of life is an extremely captivating one, it attracted people from various fields of study: biology, chemistry, geology, physics, computer science and engineering. Approaches taken by the researchers from different fields reflect the paradigm existing in the field at the moment, which is reflected in the definitions of life they use in their works.

With the rise of the theory of information and spread of its adepts into the field of biology the view of life as information processing became increasingly popular and is now essentially a dominating view among many researchers from different fields[5]. This approach seems to be very attractive. The life in this paradigm is a self-sustained system capable of information transfer and evolution and the origin of life is a process of creating new information out of randomness. While the approach is very attractive, it has its drawbacks. For example computer viruses are capable of informational self replication, but they are not considered alive. As well as not every dissipative system that appears ordered can self sustain can be called living. These two time scales (quick metabolic reactions of self-maintenance and slow process of inheritance and evolution) are not the only ones which a characteristic of life. Living organisms are also behave and develop: the latter one is a very distinctive feature of life as the opposite to non-living dissipative systems and computer or biological viruses. Omitting these stages from definition of life and disregarding them while searching for the origin of life may be the problem which stagnates the origin of life research[5].

The approach taken in this work is strongly influenced by the information theory. We are going to consider physical principles which can be a foundation of spontaneous information creation as well as possible limitations of those physical principles.

1.2 What is so special about life?

First of all, because there is no proper definition of life even within a certain approach, we have to exhaustively list all the properties which known life exhibits and then rank them according to importance to the property of being "alive". Because the author of this work surely enough thinks within a certain paradigm, not all possible properties will be listed, for a property reflects an angle from which one can look on a certain object and how one logically groups parts and features of this object.

Enormous complexity. For a person who haven't been exposed to biology such as a physicist or an engineer the first encounter with cell biology is a bewildering experience. Organization and complexity even of the simplest living cells is magnificent. Even the simplest[1] organism Candidatus Carsonella ruddii, a bacteria which cannot survive on its own and requires a host to provide for essential nutrients, has a DNA of 213 genes or about 160 00 nucleotides, almost as many types of proteins(182 genes code for proteins), each responsible for a certain function, as well as RNA and pool of smaller molecules[6, 7].

DNA, RNA and proteins are complex entities on their own. DNA is a very long polymeric molecule made from four types of small molecules called nucleotides. The sequence of the nucleotides is called genetic code and it stores information about proteins – the actual workhorse of the cell. Proteins are made out of 20 types of small molecules called amino acids. Some of the proteins are responsible for "reading" DNA in order to make proteins, some for "copying" it in order to make progeny of the cell. In addition there are signaling proteins, proteins which let nutrients inside, proteins which help other proteins to fold properly in order to perform functions, proteins which break nutrients into accessible food, proteins which make cell move and so on and on and on[8]. Besides DNA and proteins, cells have a third type of long polymers – RNA, which is similar to DNA, but is used by cells for

[1] the organism having the shortest known DNA

both functioning and information storage[2]. When proteins "read" DNA, they produce RNA, which is then being used by ribosome (a molecular machine which is a mixture of RNAs and proteins) to produce proteins (see figure 1). Even this simplistic view isn't simple at all, there are endless details which were omitted here.

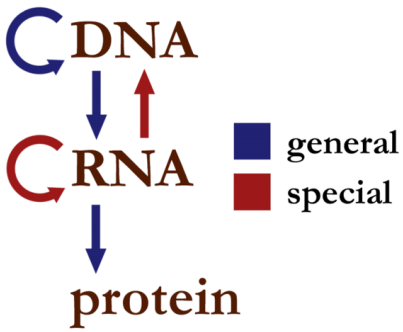

Figure 1: When proteins "read" DNA, they produce RNA, which is then being used by ribosome (a molecular machine which is a mixture of RNAs and proteins) to produce proteins (this picture is from Wikimedia Commons)

When one thinks about this complexity with regards to the origin of life a series of questions naturally arise. Does it have to be so complicated to function? What is the simplest living system possible? How could such a complexity originate in the chaos of prebiotic Earth? These are not idle questions. In fact there's an argument that the complexity of the living systems produces the huge space of possibilities which enable life to be what it is[9, 10]. Just 182 genes are about 160 000 nucleotides. At the every genome copying event any nucleotide can be substituted with 3 other. This produces about 4^{160000} possible genomes.

Autonomy and purposeful character of life. In biology it is very hard to exclude the notion of purpose from scientific discussion. This notion arises naturally: when talking about evolution, we can clearly distinguish between adaptation and maladaptation, thus distinguishing between so to say "right" and "wrong" actions of a biological entity which benefit or hurt a given entity. Nor in physics neither in chemistry we cannot talk about any kind of purpose

[2] This fact when discovered had tremendous influence on the origin of life community and changed the way how we think about it for many years.

or notion of "self", while one of the important questions of evolutionary theory is the question of what is the entity which benefits from the "right" actions. "The central theoretical problem of teleonomy[3] will be that of the nature of the entity for whose benefit adaptations may be said to exist."[11] In the scientific argument on what constitutes a single organism, for example, one of the definitions is: "It's one set of genetically identical cells that are in communication with one another that have a common *purpose* or at least can *coordinate themselves* to do something."[12] The notion of self and notion of purpose are emergent phenomena in biology.

Inheritance. The property of inheritance in perhaps the most obvious one. Every living cell passes its DNA to its daughters with high precision. All descendants look similar to their ancestors. The information how to develop, look and very often how to behave is preserved between the generations with great precision. While ability to replicate itself isn't unique to life – computer viruses do it very well. Heritability is a minimum requirement for life. It has to make more of *itself* in order to pass traits to further generation and participate in natural selection.

Exponential growth. If one looks at the living organisms it appears that they, given enough food and good conditions, would multiply exponentially. This observation led to a popular requirement for self-replicating systems to undergo exponential growth [13, 14, 15] Such systems, given enough food, can grow indefinitely while being constantly diluted, in particular moved to a new vessel containing new food supplies.

1.3 Evolvability

> I have hitherto sometimes spoken as if the variations ... had been due to chance. This, of course, is a wholly incorrect expression, but it serves to acknowledge plainly our ignorance of the cause of each particular variation.
>
> Darwin [16, Chapter 5]

[3] "Teleonomy is the quality of apparent purposefulness and of goal-directedness of structures and functions in living organisms brought about by natural laws." (Wikipedia definition)

Another very characteristic property of life is evolvability. There are very few definitions of life that do not include this peculiar ability. Yet defining evolution is also tricky. Generally speaking evolution is a fascinating ability of living organisms to thrive not even despite imperfections of inheritance, but precisely due to the imperfections. That is notoriously different from the properties of the complex machines created by humans. If one makes a copy of a blueprint with a mistake and then makes a copy of this copy also with a mistake and so on, they will likely end up with a machine which doesn't work, even if they try to do it with hundreds of thousands machines and check if at least one will benefit from a random change[4]. The ability of the living organisms to evolve is indeed an amazing one. Maybe that is why in the origin of life as well in artificial life research it is ubiquitous to concentrate on the evolutionary properties of life as the defining feature[4, 17, 18].

At this point in order to continue I have to define genotype and phenotype, because those concepts play the key role in discussions of evolution.

Genotype or genome can be thought about as an information carrier, for like a tape, that keeps information about organism and which is being copied in order to replicate the organism.

Phenotype, on the other hand, is the "composite of an organism's observable characteristics or traits, such as its morphology, development, biochemical or physiological properties, phenology, behavior, and products of behavior (such as a bird's nest)"[19]. There is a non-trivial relationship between phenotype and genotype, which itself is a topic of discussions [20, 21, 22]

It is very common to distinguish two types of evolution – adaptive and innovative[23]. The first one is the classic Darwinian evolution. The principles of Darwinian evolution are:
" *1. Different individuals in a population have different morphologies, physiologies, and behaviors (phenotypic variation).*
2. Different phenotypes have different rates of survival and reproduction in different environments (differential fitness).
3. There is a correlation between parents and offspring in the contribution of each to future generations (fitness is heritable)[24]."
Thus, given that a certain genotype appears and it corresponds to a phenotype with higher fitness, it will get fixed in the population. Here, evolution

[4]A biologist would say that the experiment was performed on a population scale. In fact this is the only proper way to talk about evolution, that is why I had to make a comment about hundreds of thousands of machines

is an ability to preserve "good" changes. The second type of evolution is the ability to invent new functions, to grow more complex so to say.

The reason why one needs to introduce innovative evolution in addition to adaptive is that natural selection doesn't have a creative power. It allows the best innovations to spread, but cannot invent them. As Hugo de Vries said : "natural selection may explain the survival of the fittest, but it cannot explain the arrival of the fittest" [25]. Thus in order to understand the origins of life and its diversity and resilience we need to understand origins of innovability. It was shown[22] that modern biological systems which are capable of innovative evolution share one property in common. This property is a relation between genotype and phenotype. In particular, living systems have vast genotype networks. Consider a genotype to be a node in a graph and a mutation to be an edge which connects to genotypes which differ only in this mutation. Because there are many possible mutations, each genotype has many neighboring genotypes. What is interesting is that a single mutation (a hop in the graph of genotypes) will likely not result in the change of a phenotype. Moreover, after making one more mutation there is a high chance again that phenotype will stay the same. Having very many consecutive mutations can change genotype drastically, yet the phenotype may stay the same[22, 23]. The other property is that every genotype in the graph mentioned above has great phenotypic diversity of its neighbors. It means that every genotype has many neighbors with viable phenotypes, each of which is different from the current and different from each other. This allows living systems, so to say, explore their space of genotypes through the large network and discover new phenotypes (inventions)[22, 23].

There is yet another type of evolution researchers care about – this is open ended evolution. There is no unique definition or cohesive theory of open-ended evolution[26]. Roughly speaking it can be though of as one that continually produces novel forms[27]. The problem with open-ended evolution is that it is relatively easy to make a system which would adapt to certain external conditions, but then will stay unchanged in its local maximum of fitness[28, 26]. Life on the other hand is known to continue to evolve even without change in the external conditions. In the famous experiment on evolution when Richard Lenski has been tracking genetic changes in 12 initially identical populations of E. *coli*, the bacteria continued to evolve after 60 000 generations or 27 years[29]! The understanding of open-ended evolution is important for the origin of life community as well as for the machine learning one. In order to claim that one has discovered how the life could

have originated on Earth or can originate in principle one has to replicate at least the most important properties of life; and open-endnesness of evolution is one of them. To the best of my knowledge this problem hasn't been solved yet[28]. For the machine learning community this problem is important because search of global optima of the objective function is often stuck in local minima and there is an argument[30] that if an open-ended dynamics to be created in the system the difficulty mentioned above can be overcome.

In machine learning research, however, one need not to search for the origins of open-ended evolution, but just for a set-up which will be able to provide it. While there is a certain controversy if it has been achieved yet[31, 32, 26], I believe that it was.

One can consider novelty search[30] as an example. In novelty search, one considers a population of agents, each of which has a defined genotype and a certain phenotype. Populations go through a process of reproduction and selection. The objective of the selection is however *novelty*: how different is a specimen under consideration compared to *all specimens that ever existed*. The authors of the novelty algorithm claim that it achieves open-endnesness[30].

In this approach as well as in another one, which I believe achieved open-ended evolution – MAP-elites [33] – individuals usually have well-defined large genotypes with phenotypes corresponding to them. In biological terms it would mean that the individuals already have DNA which codes for proteins. In addition to that researchers can always add extra requirements, such as a complete history of individuals in novelty search[30] or artificially defined ecological niches like in MAP-elites[33] In the origin of life research on the other hand the knowledge of how all of the above can emerge from the "prebiotic soup" of small molecules is the question. Emergence of an open-ended evolution contrary to construction makes the origin of life research so hard. The next chapter will give a brief overview of more than a hundred of years of research. It is not intended to be complete but rather to explain main challenges and reasons behind our approach to the question.

Chapter 2. From non-life to life. What was there before life and why there is a problem

> But even more important was that Miller's experiments moved life's origin from philosophical speculation to the realm of hard, experimental science.
>
> A.Wagner [10, Chapter 2]

This chapter is concerned with models for the origin of life on the planet Earth. More precisely it attempts to give a review of existing hypotheses, logical links that connect them as well as logical gaps and inconsistencies.

To say that the origination of life on Earth is a mystery is to say nothing. After decades of research in biology, geology, chemistry and physics we are still clueless about several stages of the origin of life. The problem is perhaps that unlike with origins of different species one cannot blame the origin of life onto evolutionary forces, because one has to explain the origin of evolution itself. Specifically one has to explain an open-ended evolution and this is not an easy task. During the last couple of decades computer scientists and engineers joined the quest of looking for ways an open-ended evolution could emerge on its own. Yet all to no avail. I claim, the belief that evolution will solve all our problems with creating recognizably living entities has to an extent hampered the process. Researchers try to see systems experience an evolutionary dynamics as early as possible. To show that a system, no matter how simple it is, is "evolvable" (and thus as a prerequisite for evolvability as capable of heritability) is a very important part of to proof for a researcher to avoid critique[34, 35, 36, 37]. Yet this might have been not the case on the early Earth and we need to consider this possibility. It is also possible that there are certain requirements on the complexity of a system capable of open-ended or innovative evolution[9].

Heritability itself is not an easy task to achieve. While computer scientists can easily invent *in silico* systems which will replicate with a desired degree of faithfulness[38, 26, 33], chemists and biologists have managed to implement reasonable replication *in vitro* only for very simplest and thoroughly designed systems[13, 39]. Yet heritability is a necessary component of evolvability. The very definition of evolution relies on it[5].

[5]See Chapter 1

Another reason for the enigmatic nature of the origins of life is the timing and speed of the origin. We all know how fragile living organisms are: it is very easy to turn one into inanimate matter, but impossible to revert things back. Nevertheless early Earth, where and when life originated and persisted, was an incredibly harsh environment. After Earth was formed about $4.5 \cdot 10^9$ years ago[40] it was in the molten state because of high activity of volcanoes and frequent collision with other bodies[41]. Gradually surface of the Earth cooled down and it accumulated the atmosphere. Then around $4.1 - 3.8 \cdot 10^9$ years ago it went through a stage called Late Heavy Bombardment, which was marked (as it's clear from its name) by disproportionately large number of collision with asteroids[42] The Earth was significantly hotter and tectonic activity was higher, being responsible for constant circulation of the material. The atmosphere was uninviting – high pressure hot hydrogen and carbon dioxide[43]. Yet earliest evidences of life correspond to a very close time – biogenic graphite was discovered in 3.7 billion-year-old meta-sedimentary rocks in Western Greenland[44, 45, 46].

Invention of such complex structures out of nothing by basically means of random search in the given time frame is an impressive achievement. This is not the most life is capable of, however. The Murchison meteorite[6] contained common amino acids and a complex mixture of alkanes[7][47, 48].

Not all stages of life are equally hard to explain. The early stage of abiogenesis are uniformly agreed upon. The hypothesis that appeared the earliest happens to be the most accepted now. It is a "primordial soup" theory stated by Alexander Oparin and John Haldane in the 1920s. It was summarized by Robert Shapiro in the following form[49, p.100].

> "Early Earth's reducing atmosphere, being exposed to energy sources of various forms, produced simple organic compounds, such as for example amino acids. These compounds formed a "soup" which may have had various concentrations at different locations. Further transformation brought up more complex organic polymers and eventually life."

This hypothesis was confirmed in 1952 in the famous Miller-Urey experiment in which Stanley Miller and Harold Urey imitated the early Earth's

[6] a more than 100kg meteorite that fell near Murchison, Victoria, in Australia, in 1969.
[7] similar to that found in the MillerUrey experiment

environment and produced 5 different amino acids out of the mixture of water (H_2O), methane (CH_4), ammonia (NH_3), and hydrogen (H_2)[50]. "Later experiments produced many other of lifes construction materials, including sugars and parts of DNA" [51]

Exciting as it is, primordial soup theory has one pretty obvious issue. "Further transformation brought up more complex organic polymers and eventually life"[49] isn't precisely the recipe for life formation and requires a consistent theory which would be able to shine some light on what kind of "transformations" are capable of bringing about life. I am going to describe the most popular view on the order and nature of the transformations that had to take place.

To determine the way prebiotic chemistry followed we need to know when it is time to stop and let evolution to take on. The simplest form of life we know is bacterial. Probably because of that it is traditional to imagine the final destination of prebiotic chemistry to be some sort of proto-bacteria.

More precisely it is vastly agreed upon that the first substances which one can call *living* should be *autonomous cell-like objects* (vesicles), which grow and split into daughter cells[4, 3]. These vesicles are seen as capable of undergoing a process of adaptive evolution and thus rather faithful *information transfer* to further generations. The process of transfer being error-prone drives the *evolution*. The vesicles are also often seen to experience a set of *metabolic reactions*. These entities further can be taken on by the creative forces of open ended evolution and develop into bacterial cells with present day complex molecular machinery.

The reasoning seems very intuitive:

- An entity must be *encapsulated* in order to keep concentrations of relevant chemicals high. Being encapsulated also helps to maintain notion of "self" by providing feedback, say between presence of genes and how they affect cell performance.

- It has to be reasonably *good in replication* in order to be evolvable. If new "better" traits are not remembered by descendants, then it would be impossible to preserve them and thus evolve.

- Constant growth and replication requires cell to consume, transform and excrete molecules. In order to do this it must have a primitive *metabolic system*.

And the view is surely enough very popular. It has its opponents though. Bacteria as we know them might be a rather late stage of life development. Reasonable criticism comes from the fact that early life didn't have to look like modern; it could have been easily totally different, less efficient and eventually phased out during evolution[3]. There's also a possibility that encapsulation isn't necessary for evolution and steady growth[52, 5]. Seeing life as an information transfer and production machine also can be a very limiting approach[5].

Nonetheless, mainstream view establishes the stages of prebiotic chemistry that are thought to have happened. Traditionally we can speak about three approaches to the order of events of that phase of primordial soup theory when long biopolymers, evolution and life in general appear. These would be:

- **Encapsulation-first.** One needs small vesicles which would play a role of small test tubes, where relevant molecules will have high concentration and wouldn't diffuse away. Vesicles also put informational and metabolic molecules together, thus generating a feedback link between them. Thus appearance of vesicles is a necessary requirement for information preservation, metabolism, autonomy and evolution.

- **Metabolism-first.** A minimum requirement for life is metabolism: organisms must harvest energy and produce molecular building blocks from it. Thus first of all they need a network of chemical reactions to exist.

- **Information-first.** Without ability to replicate itself and pass information to future generations there is no Darwinian evolution. So minimal requirement for life is heritability.

Because it is virtually impossible that all three – encapsulation, informational polymers and metabolism – appeared simultaneously there's an argument about which came first. The necessity of encapsulation seems very reasonable and is widely accepted in the origin of life community [53, 54, 55, 56]. It means that the main disagreement is about whether metabolism or informational polymers must be the necessary first step for life to emerge on Earth. If we look at the problem naively it seems to be a chicken and egg type issue: in order to produce reasonably long biopolymers which can faithfully replicate one needs metabolic apparatus; yet without inheritance how can

cell "remember" its metabolic machines. Despite this complication, there is a certain similarity in the approaches metabolism-first and information-first camps take. Both of them concentrate on producing (*in vitro* or *in silico*) autonomous self-reproducing entities capable of an open ended evolution(see for example [57, 54, 5, 28]). In the following two sections I describe the main theories and critique of the camps.

2.1 Information First: RNA world hypothesis

Perhaps it was the discovery of DNA structure[58, 59] in 1953 and proposal of the central dogma of molecular biology in 1956[60] that pushed scientific community towards informational approach towards the origins of life. The structure of the DNA as well as RNA offers a great way to implement chemical inheritance and evolution. The information is stored digitally on a linear structures (genes) which are read only in one direction. Information can be restored and copied. In addition the process of copying is error-prone with a very low rate and there is a repair mechanism (in modern cells) enhancing faithfulness of replication[61, 62, 63].

DNA has four types of nucleotides it is constructed from. They are called guanine, thymine , adenine, and cytosine and abbreviated into G, T, A and C correspondingly. Nucleotides can form relatively weak bonds with other non-adjacent monomers called hydrogen bonds. These bonds are called base-pairs because any one of the monomers can bond (form a pair) only with a certain other monomer: A with T and C with G[64]. (see figure 2 (a)).

The nature of the bonds allows DNA to be a perfect molecule for preserving information between generations. Each DNA strand in the living cells forms a double helix with another, complement DNA strand. Because hydrogen bonds are not covalent, they can be broken relatively easily. Therefore the two strands of DNA can be pulled apart. Because strands are complementary to each other information stored on one is duplicated in the other. During the DNA replication in the cell this allows for the easy information transfer between generations. RNA is very similar to DNA, but due to a bit different structure the base pairing makes it fold on itself like proteins do[8]. Thus nucleic acids seem a very natural candidate for the basis of life.

Early on it was thought that life strictly separates information in the form of DNA/RNA and function in the form of proteins. As a result, despite the attractiveness of nucleic acids as first substances of life, it wasn't clear how metabolic reactions including formation of the DNA or RNA could be

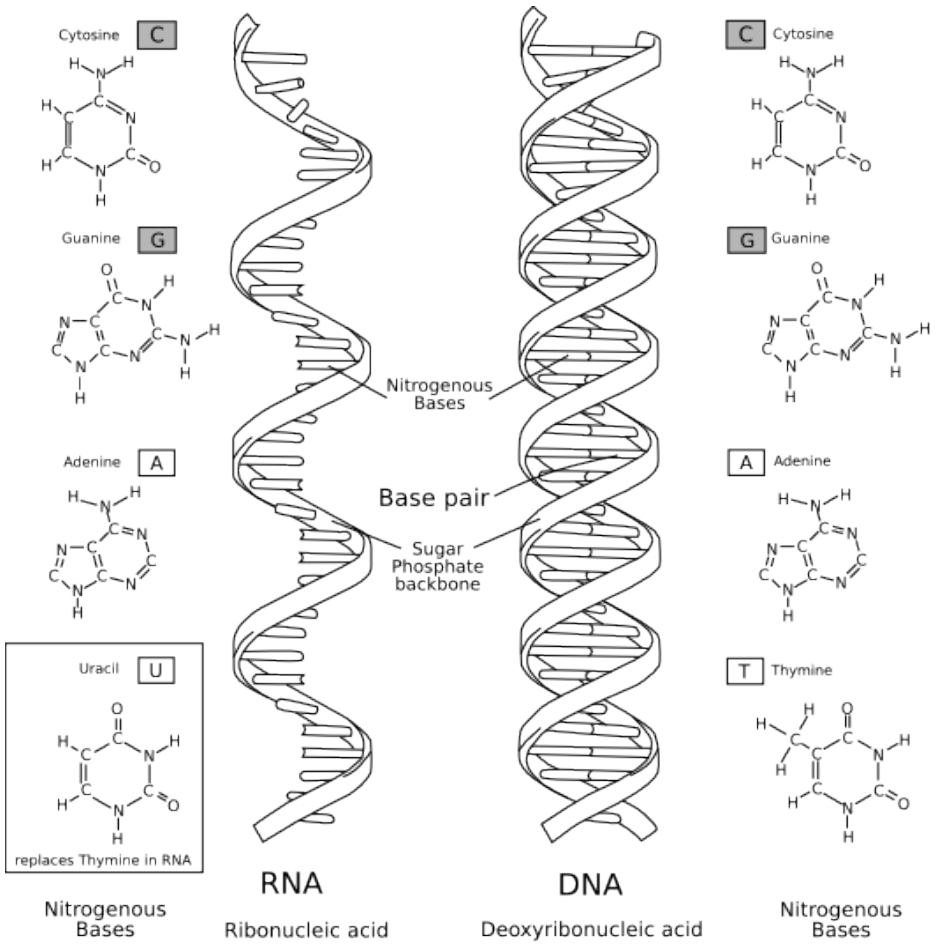

Figure 2: Nucleotides can form relatively weak bonds with other non-adjacent monomers called hydrogen bonds. These bonds are called base-pairs because any one of the monomers can bond (form a pair) only with a certain other monomer: A with T and C with G (this picture is from Wikimedia Commons)

possible without proteins, which in their modern form are impossible without nucleic acids. Consequently the scientific community still was puzzled with the chicken and egg problem. After the discovery of functional properties of RNA[65, 66, 67] (see3) though the spirit of great optimism towards the informational approach flooded the mainstream. Ability of RNA to perform catalytic functions like proteins do and its prominent role in the protein building molecular machine ribosome[68] fascinated minds of the scientists.

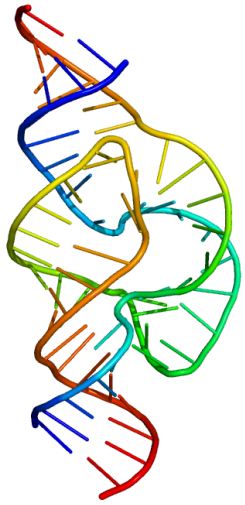

Figure 3: 3D structure of hammerhead RNA. Hammerhead RNA is a small RNA enzyme (ribozyme) which is capable of self-cleavage. Self-cleavage properties were first discovered in 1986[69, 70]. It is one of the first discovered ribozymes and is thought to play major role in origination of life because of its ubiquity[71, 72]. (this picture is from Wikimedia Commons)

The hypothesis that self-replicating RNA (ribozyme) is a precursor to all the life on Earth is called the RNA-world hypothesis. Besides the ability of RNA to serve both as functional and informational polymers, another argument in favor of RNA-world is that all modern cells have the same mechanism for peptide synthesis and this is RNA which catalyzes the peptide bond formation[73, 74]. A recent study[75] argues that that ribosomal RNA was that self-replicating entity which is a precursor to all modern life. They have found that rRNA of E. *coli* K12 encodes the entire set of tRNA's of this microbe as well as some proteins. The authors examined only one specie, thus the results can be an anomaly, however if replicated in other species they indeed indicate that the ribosome is very likely have evolved prior to cellular life and could be a precursor to all modern life.

The RNA world hypothesis is indeed a beautiful one. A molecule which can be used for self replication and for catalyzing reactions can be both a metabolizing agent and a unit of information storage. It is very easy to imagine how such a molecule could undergo adaptive evolution. It can also possibly be an intermediate step between single self-replicating entity and

complex metabolic apparatus of a cell[75].

No wonder so much effort has been put into explaining origin of RNA from the primitive chemical building blocks and also producing effective and prebiotically plausible self-replicating RNA molecules[76, 77, 78]. However it is not that easy. Most of the today's catalysts are extremely long. They couldn't possibly be the first molecules. Significant effort was put into creating a prebiotically plausible replicase. The best results are an 189-letter-long single piece self-copying RNA[79, 80] and a pair of cross-replicating ribozymes which catalyze each other formation from 4 oligonucleotides[14], each ribozyme being around 60-70 nucleotides long.

One of the arguments against the RNA-world hypothesis is that the replicases mentioned above were designed by people with very good knowledge of biochemistry[81]. It is hard to imagine the first replicase could be very efficient or very precise. Yet for decent inheritance and thus evolution one needs a very accurate replicase. This problem is called "error catastrophe". It was discovered my Manfred Eigen[82]. He calculated accuracy required to avoid error catastrophe. The longer the sequence the more accurately it has to replicate itself. He found that a replicase with fifty nucleotides can make less than one mistake; the replicase with 100 nucleotides also cannot make more than one mistake[82, 83]. Current 189 letters candidate for the original replicase makes several times as many mistakes[84]. It is obviously not good enough, even despite the fact that it was carefully designed.

Another critique of the RNA-world is a supply and demand issue. Even if there was a faithful replicase, it would result in the exponential growth of the population. This is surely good, this is what life does, but on the other hand exponential growth requires enormous amount of raw material. As Andreas Wagner analyzed in [23] if RNA replicase copies one letter per second after six hours RNA replicase system will consume 1 ton of nucleotides. This amount of very special food is hard to produce in big enough concentrations without biological source. In addition to that the "food" itself isn't the easiest thing to make[85]. These two problems are often given by the proponents of metabolism first hypothesis[23, 81, 36] as an argument for either "citric acid cycle world" or "protein world"[8]. The problem with nucleic acids is that unlike proteins, which are composed out of relatively easy to synthesize amino acids, they are composed out of much more complex nucleotides (each of nucleotides itself is made out of three sub-units[64])[85]. Nucleotides

[8]see 2.2 for more details

aren't made easily prebiotically and don't readily couple together even when activated[85].

It is possible that RNA itself is a result of biological evolution[86, 87]. A solution which is still information-first was proposed by Nick Hud[85] who argued that RNA is a product of evolution of proto-RNA, a more prebiotically plausible RNA-like oligomer. They also propose a possible pathways of the evolution. This solution indeed can help with with the problem of nucleotides, but doesn't help with the supply issue. It is also important to notice that protein world also faces the supply and demand problem, though not in such a harsh form because amino acids are an easier material to make.

2.2 Metabolism-first: proteins and citric-acid cycle

The problem with food supply has a plausible solution in the metabolism-first set of hypotheses. It is also a geologically plausible solution. To ensure exponential growth of RNA, one needs a lot of nucleotides, which also are not simple molecules and have to be produced from simpler materials. The source of a tremendous amount of this simpler material was found in hydrothermal vents [88]. Hydrothermal vents can provide energy for running metabolic cycle known as citric acid cycle which is capable of producing nucleotides. The citric acid cycle is a metabolic pathway which is used by aerobic organisms to make energy[8] (see figure 4). It is ubiquitous in nature and thus seems a plausible precursor to life.

On the other hand, citric acid cycle has its problems. First it has many parasitic reactions which make sustaining the main body of the cycle implausible[36]. Second, studies of a metabolism first model GARD[89, 35] indicate that it is very improbable for metabolic network to undergo adaptive evolution[34, 28]. Hence, while very attractive and powerful, the idea of citric acid cycle as a precursor of all life remains a controversial one.

Another candidate is protein-like molecules autocatalytic cycles In modern organisms metabolic networks are enormously complex and many pathways heavily rely on proteins. That is why they are attractive candidates for metabolism-first models. The amino acids – monomeric units of proteins – are also very easily made in prebiotic reactions. Protein-based metabolic cycles were proposed and popularized by Stuart Kauffman[90]. The main idea is simple. Polypeptides can catalyze two types of reactions: condensation and cleavage of other polypeptides, thus it is a possibility that they can form reflexively autocatalytic sets. It is argued[90] that the emergence of such sets

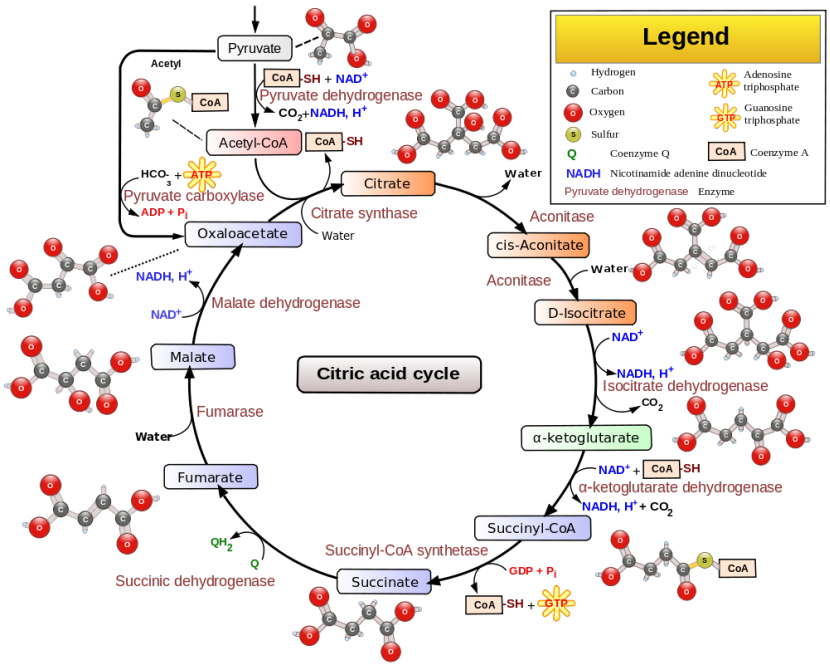

Figure 4: Citric acid cycle. Citric acid cycle is autocatalytic: it makes two molecules of citric acid out of one through a chain of ten chemical reactions. Portions of this metabolic cycle appear in the most ancient forms of life. It can run in two directions. These properties make it a very attractive candidate for the "grandfather of all metabolic activities"[23]. However there is no experimental support yet for emergence of the cycle from prebiotic soup. (picture is from Wikimedia Commons)

is an inevitable property of sufficiently complex sets of polypeptides.

This research sparked interest to autocatalytic sets and many scientist attempted to produce and studying them[91, 92, 93, 94, 95].

Autocatalytic sets are subject to the same criticism as citric acid cycle and partially RNA: it requires a high monomer supply, very well may be not evolvable and may suffer from parasitic reactions. Yet it as well as information-first theories deserves more investigation or a powerful alternative theory.

2.3 On the way to discover origins of life: aims of this thesis

Both information-first and metabolism-first camps share their target: they are looking for self-sustained, autocatalytic, self-replicating capable of adaptive evolution, which would emerge out of prebiotic soup. Very often theoretical models and in vitro experiments are criticized for lack of some features of living organisms. It is not necessarily true, however, that life evolved right out of prebiotic soup. Self-sustainability, evolvability, heritability and resemblance of current life didn't have to occur at the same stage of life origin. It is possible that every particular step may only provide a good ground for the next one. In this dissertation I don't claim we found a perfect model of life origins, but rather a self-sustaining system with complex and selective dynamics, which we claim is an important step towards further stages of evolution.

First, I believe, it is important to discover a prebiotically plausible system, which doesn't die out. It doesn't have to have any form of inheritance. It just need to produce biomass and have a complex dynamic behavior – an Oparin's "self-reproducing garbage bag". If such a system is not stuck in one dynamical attractor, but on the contrary, is capable of investigation of the phase space, capable of innovability, it can discover new interactions and molecules. Systems which discover molecules which can be used for implementing "memory" would survive because it will be able to preserve good innovations. The separation into memory and function present in current organisms can be achieved by continuing exploration. Thus our research starts with search of self-sustaining system capable of discovering new molecules and interaction and physical mechanisms which are responsible for emergence of such a system from prebiotic soup. The next important step is to find principles which can explain discovery of inheritance or memory.

Bio-polymeric systems are good candidates for a role of such systems. First, formation of a complex autocatalytic network of polymers cannot be excluded probabilistically[90, 96]. Second, it was shown that if autocatalytic networks include rare uncatalyzed reactions then new autocatalytic sets can arise and system can avoid a curse of single attractor[17, 28]. Third, if polymers in such systems can grow long enough they will automatically discover new polymers, which will have new functions. However there is a problem with the polymeric solution. It is what we call a Flory problem or problem of short lengths: spontaneously polymerizing molecules have an exponen-

tial length distribution and don't grow long enough to perform functions of modern proteins and RNAs.

The Flory length problem arises due to the fact that even if there's enough raw material RNA and protein chains tend to be just few monomers long. This is a significant challenge. For proteins and RNA to function and to store meaningful amount of information the length should be significant. It is estimated that the shortest length for proteins and RNA's is around 60-100[76]. Nor RNA neither proteins are capable of self polymerization. Many attempts to polymerize them in the lab using various prebiotically plausible catalysts resulted in rather short chains[9]. I explore possible solution to the Flory length problem in Chapter 3

In Chapter 5 I investigate dynamical and evolutionary properties of this solution.

We have also developed computational methods for studying chemical systems with emergent behavior. These methods allow to simulate systems with unlimited number of molecular species, allowing for exploration of complex systems with non-trivial dynamics[97]. They are strongly rooted in physics and can account for small fluctuations, which likely played important role in the emergence of first biological systems. Thus I believe the proposed method[97] can be an important tool in the investigations of complex chemical systems with emergent behavior. The method is explained in section 4.

[9]See chapter 3 for details

Chapter 3. Solution to the problem of short length

This chapter is a copy of the chapters 3-6, 8 of the paper [81]

3.1 The "Flory Length Problem": polymerization processes produce mostly short chains

Prebiotic polymerization experiments rarely produce long chains. It is commonly assumed that the chain lengths of proteins or nucleic acids that could have initiated the transition to biology must be at least 30-60 monomers long [98]. Both amino acids or nucleotides can polymerize under prebiotic conditions without enzymes, but they produce mostly short chains [99, 100, 101, 102, 103]. Leman et al. showed that carbonyl sulfide (COS), a simple volcanic gas, brings about the formation of oligo-peptides from amino acids under mild conditions in aqueous solution in minutes to hours. But the products are mainly dimers and trimers [102]. Longer chains can sometimes result through adsorption to clays [104, 105] or minerals [106, 76], from evaporation from tidal pools [107], from concentration in ice through eutectic melts [108], or from freezing [109] or temperature cycling. Even so, the chain-length extensions are modest.

For example, mixtures of Gly and Gly_2 grow to about 6-mers after 14 days [110, 111] on mineral catalysts such as calcium montmorillonite, hectorite, silica or alumina. Or, in the experiments of Kanavarioti, polymers of oligouridylates are found up to lengths of 11 bases long, with an average length of 4 [108] after samples of phosphoimidazolide-activated uridine we frozen in the presence of metal ions in dilute solutions. Similar results are found in other polymers: a prebiotically plausible mechanism produces oligomers having a combination of ester and amide bonds up to length 14 [112].

It is puzzling how prebiotic processes might have overcome what we call the "Flory Length Problem" – i.e. the tendency of any polymerization process to produce a distribution in which there are more short chains and fewer long chains. Standard polymerization mechanisms lead to the the Flory or Flory-Schulz distribution of populations $f(l)$, whereby short chains are exponentially more populated than longer chains [113],

$$f(l) = a^2 l(1-a)^{l-1}, \tag{1}$$

where l is the chain length and a is the probability that any monomer addition is a chain termination. The average chain length is given by $\langle l \rangle = a(2-a)$; see Figure 5(a).

Prebiotic monomer concentrations are thought to have been in the range of micromolar to millimolar [114, 115, 116, 108, 117]. Given micromolar concentrations of monomers, and given $\langle l \rangle = 2$, the concentration of 40-mers would be $\approx 10^{-19}$ mol/L. Figure 5(b) shows that where the chain-length distributions are known for prebiotic syntheses, they are well fit by the Flory distribution (or exponential law $f(l) \propto \text{constant}^l$) [118, 119]).

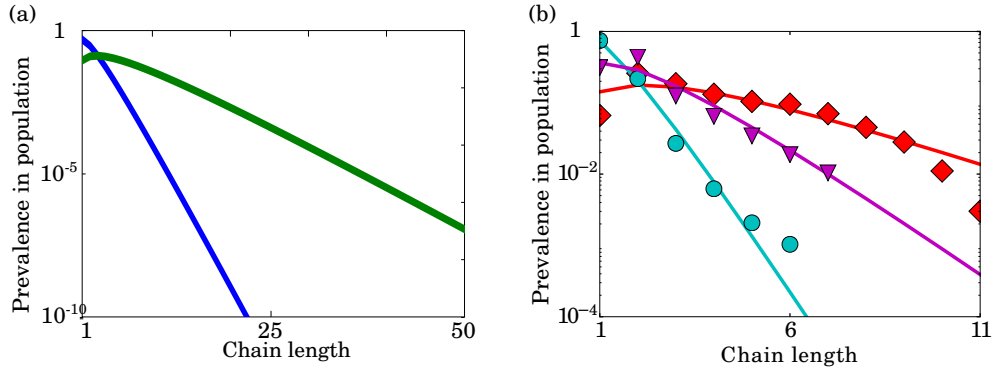

Figure 5: **Polymerization processes lead to mostly short chains.** (a) Spontaneous polymerization processes typically lead to a Flory distribution of chain lengths. Green line gives $\langle l \rangle = 6$, blue corresponds to $\langle l \rangle = 2$ (b) Fitted distributions from experiments on prebiotic polymerization: red – Kanavarioti [108], cyan – Ding [120], magenta – Ferris [121]

3.2 The foldamer-autocat mechanism: Short HP chains fold and catalyze the elongation of other HP chains

We propose that the key to the Chemistry-to-Biology transition may have been *foldable polymers* ("foldamers"). Today's biological foldamers are predominantly proteins (although RNA molecules and synthetic polymers can also fold [122, 123, 124]). Many foldamers adopt specific native conformations, mainly through a binary solvation code of particular sequence patterns of the H (hydrophobic) and P (polar) monomers [125]. We call these HP copolymers.

Since today's bio-catalysts are proteins, it is not hard to imagine that early primitive proteins could have been primitive catalysts. Precision and

complexity are not required for peptides to perform biological functions. For example, proteins generated from random libraries can sustain the growth of living cells [126]. And, specific binding actions between random peptides and small molecules are not rare [127]. Below, we describe results of computer simulations that lead to the conclusion that short random HP chains carry within them the capacity to autocatalytically become longer and more protein-like.

3.2.1 Here are the premises of the model

1. Some random HP sequences can fold into compact structures.
2. Some of those foldamers will have exposed hydrophobic "landing pad" surfaces.
3. Foldamers with landing pads can catalyze the elongation of other HP chains.
4. These foldamer-catalysts form an autocatalytic set.

Here is evidence for these premises.

1. Non-designed random HP sequences are known to fold. HP polymers have been studied extensively as a model for the folding and evolution of proteins [128, 125, 129, 130, 131]. Those studies show that folded structures can be encoded simply in the binary patterning of polar and hydrophobic residues, with finer tuning by specific interresidue contacts [132, 133]. This is confirmed by experiments [134, 135, 136, 137]. Moreover it was shown [138] that sequences capable of collapsing into compact structures can be prebiotically selected just under the forces of hydrolysis and aggregation.
2. Exposed hydrophobic clusters and patches are common on today's proteins. A study of 112 soluble monomeric proteins [139] found patches ranging from 200 to $1,200\mathring{A}^2$, averaging around $400\mathring{A}^2$; they are often binding sites for ligands or other proteins. Modern proteins have many sites of interaction with other proteins, typically nearly a dozen partners. Almost 3/4 of protein surfaces have geometrical properties that are amenable to interactions and those sites are enriched in hydrophobes [140].

3. Surface hydrophobic patches on proteins are often sites of catalysis [141, 139, 142, 143]. For example, hydrophobic clusters on the surface of lipases serve as initiation sites where the hydrophobic tail of a surfactant interacts with the patch first [142]. A hydrophobic cluster on Cytochrome-c Oxidase is known to increase k_{cat}[143].

4. Primitive proteins might have catalyzed peptide-chain elongation. Of course, today's cells synthesize proteins using ribosomes, wherein the catalysis is carried out by RNA molecules. Yet, there are reasons to believe that peptide chain elongation might alternatively be catalyzable by proteins. First, peptide chain elongation entails a condensation step and the removal of a water molecule [64, chapter 3, p. 82]. Dehydration reactions can occur in water if carried out in nonpolar environments [144, 145], such as protein surfaces. Second, a major route of protein synthesis in simple organisms such as bacteria and fungi utilizes nonribosomal peptide synthetases, and which don't involve mRNAs [146, 147].

3.3 Modeling the dynamics of HP chain growth and selection

The dynamics of the model. We assume that chain polymerization takes place within a surrounding solution that contains a sufficient supply of activated H and P monomers. Since living systems – past or present – must be out-of-equilibrium, this assumption is not very restrictive. In our model, activated H and P monomers are supplied by an external source at rate a. A given chain elongates by adding a monomer at rate β. Just to keep the bookkeeping simple, we consider a steady state process in which molecules are removed from the system by degradation or dilution at the same rate they are synthesized. We assume chains can undergo spontaneous hydrolysis due to interaction with water; any bond can be broken at a rate h. Without loss of generality we define the unit rate by setting $\beta = 1$. All other rates are taken relative to this chain-growth rate.

Chain folding in the model. In addition, our model also allows for how the collapse properties of the different HP sequences affect the populations that polymerization produces. A standard way to study the properties of HP sequence spaces is using the 2D HP lattice model [128, 125]. In this model, each monomer of the chain is represented as a bead. Each bead is either H

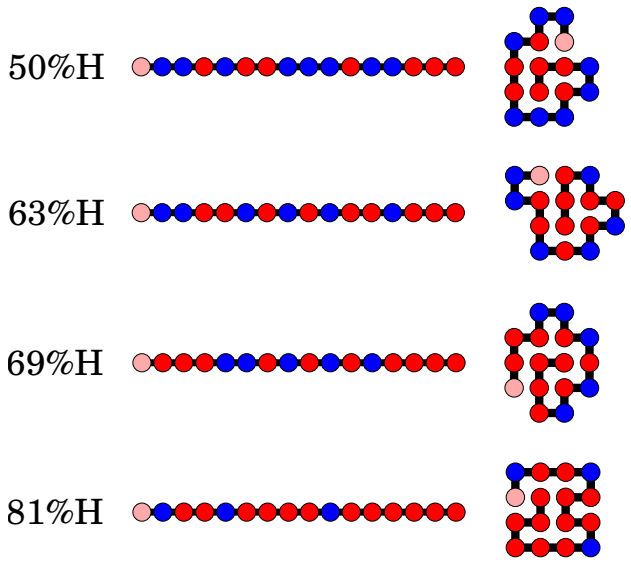

Figure 6: **Examples of HP sequences that fold to unique native structures in the HP lattice model. Red (or pink if in the beginning of the sequence) corresponds H monomers and, blue to P.**

or P. Chains have different conformations, represented on a 2-dimensional square lattice. The free energy of a given chain in a given conformation equals (the number of HH noncovalent contacts) $\times$ (the energy e_H of one HH interaction). Some HP sequences have a single lowest-free-energy structure, which we call *native*, having native energy E_{nat}:

$$E_{nat} = n_{h\phi} e_H. \tag{2}$$

where $n_{h\phi}$ is the number of HH contacts in the native structure of that particular sequence.

A virtue of the HP lattice model is that for chains shorter than about 25 monomers long, every possible conformation of every possible sequence can be studied by exhaustive computer enumeration. Thus folding and collapse properties of whole sequence spaces can be studied without bias or parameters. Prior work shows that the HP lattice model reproduces many of the key observations of protein sequences, folding equilibria, and folding kinetics of proteins[148]. A main conclusion from previous studies is that a non-negligible fraction of all possible HP sequences can collapse into compact and structured and partially folded structures resembling native proteins [128];

see fig. 6. The reason that the 2-dimensionality adequately reflects properties of 3-dimensional proteins is because the determinative physics is in the surface-to-volume ratios (because the driving force is burial of H residues). And, it is helpful that the 10-30-mers that can be studied in 2D have the same surface-to-volume ratios as typical 3D proteins, which are 100-200-mers [149].

We assume that folded and unfolded states behave differently, as they do in modern proteins. We suppose that a folded chain is prevented from further growth, and also are protected from hydrolysis. This simply reflects that open chains are much more accessible to degradation from the solvent or adsorption onto surfaces than are folded chains. Even so, folding in our model is a reversible, as it is for natural proteins, so some small fraction of the time even folded chains are unfolded, and in that proportion, our model allows further growth or degradation. For this purpose, we estimate the folding and unfolding rate coefficients for any HP sequence as [150]:

$$\ln\left(\frac{k_f}{k_u}\right) = -\Delta G/kT = E_{nat}/kT - N\ln z, \tag{3}$$

where z is the number of rotational degrees of freedom per peptide bond.

Catalysis in the model. Some HP sequences will fold to have exposed hydrophobic surfaces. These surfaces could act as primitive catalysts, as modern proteins do more optimally today. Fig. 7 illustrates a common mechanism of catalysts; namely translational localization of the reacting components. A protein A (the catalyst molecule) has a hydrophobic "landing pad" to which a growing reactant chain B and a reactant monomer C will bind, localizing them long enough to form a bond that grows the chain. How much rate acceleration could such a localization give? Here is a rough estimate.

For chain elongation, the catalytic rate will increase if the polymerization energy barrier is reduced by hydrophobic localization, by a factor $\beta_{\rm cat}/\beta_{\rm no\,cat} \propto \exp(E_H \cdot n_c/kT)$, where n_c is the number of H monomers in the landing pad (see figure 10). The free energy of a typical hydrophobic interaction is 1-2 kT. We take the minimum size of a landing pad to be 3. For a landing pad size of 3-4 hydrophobic monomers, this binding and localization would reduce the kinetic barrier by 3-8 kT, increasing the polymerization rate by 10 to 3000 times. Of course, this rate enhancement is much smaller than the 10^7-fold of modern ribosomes [151], but even small rate accelerations might have been relevant for prebiotic processes.

In order to simulate this dynamics, we run stochastic simulations. We

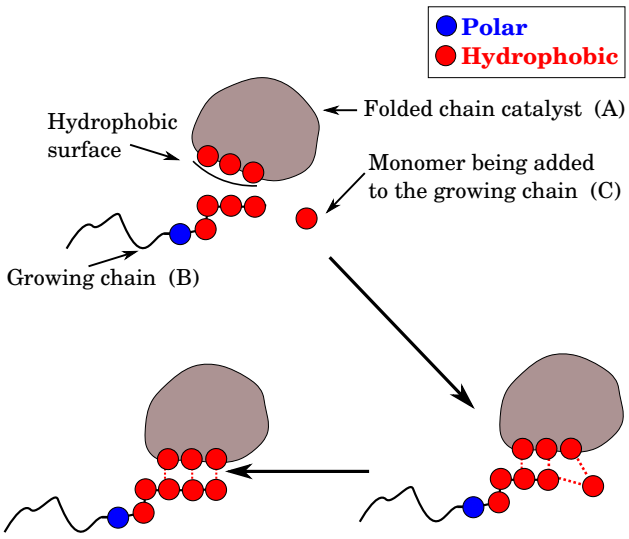

Figure 7: **Some HP foldamers have hydrophobic patches, which serve as "landing pads" that can catalyze the elongation of other HP chains.** Chain A folds and exposes a hydrophobic sticky spot, or landing pad, where another HP molecule B, as well as an H monomer C, can bind. This localization reduces the barrier for adding monomer C to growing chain B.

used Expandable Partial Propensity Method (EPDM) [97, not published][10].

3.4 Results

3.4.1 Folding alone does not solve the Flory Length Problem. But folding plus catalysis does.

We compare three cases: Case 1 is a reference test in which sequences grow and undergo hydrolysis but no other factors contribute, Case 2 allows for chain folding, but not for catalysis and Case 3 allows for both chain folding and catalysis. Case 1 simply recovers the Flory distribution, as expected, with exponentially decaying populations with chain length (see figure 8 gray lines). In the Case 2 when chains can fold, they can bury some monomers in their folded cores. Thus, chains that are compact or folded degrade more slowly than chains that don't fold. Figure 9 (Case 2) indeed shows that folded polymers have higher populations than unfolded ones. This result is in the

[10]Description and the corresponding C++ library, can be found at: https://github.com/abernatskiy/epdm.

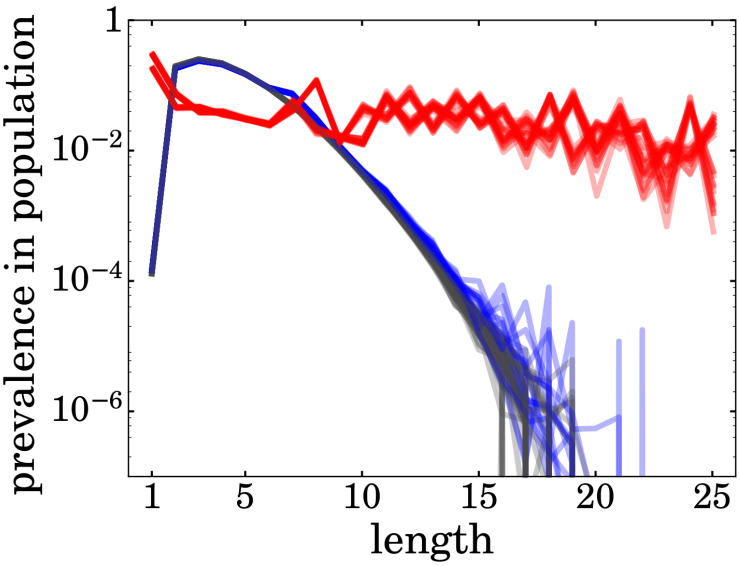

Figure 8: **Chains become elongated by foldamer-catalyst HP sequences. Case 1 (gray):** A soup of chains has a Flory-like length distribution in the absence of folding and catalysis. **Case 2 (blue):** A soup of chains still has a Flory-like length distribution in the absence of catalysis (but allowing now for folding). **Case 3 (red):** A soup of chains contains considerable populations of longer chains when the soup contains HP chains that can fold and catalyze. We run 30 simulations for every case. To produce each line we took a time average over 10^6 time points in the steady state interval, then counted molecules for each length and divided it it by the total molecular count.

agreement with the work of Shakhnovich et al. [138]; they showed that compact structures are favored under conditions of aggregation and hydrolysis. However folding alone does not solve the Flory length problem (see Figure 9 case 2 and Figure 8 blue lines). However, shows that this situation does not solve the Flory length problem either. Folding does increase the populations of some foldamer sequences relative to others, but the effects are too small affect the shape of the overall distribution (see figure 8 blue lines). Case 3 gives considerably larger populations of longer chains than cases 1 or 2 give (red lines on figure 8). When chains can both fold by themselves and also catalyze the elongation of others, such polymerization processes will "bend" the Flory distribution. This effect is robust over an order of magnitude of the hydrolysis and dilution parameters. The result is that some HP chains can fold, expose some hydrophobic surface, and reduce the kinetic barrier for

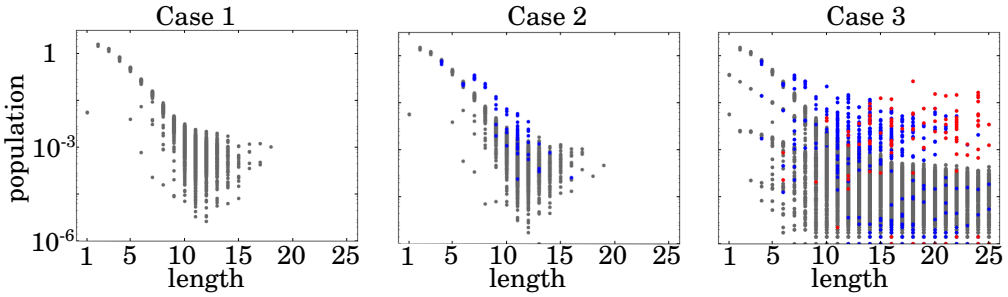

Figure 9: **The distributions over individual sequences are highly heterogeneous.** We show the populations (molecule counts of individual sequences) for the three cases: in case 1 we don't allow folding or catalysis, in case 2 we allow folding but not catalysis, and in case 3 both folding and catalysis are allowed. For all the cases gray dots represent populations of the sequences that cannot fold, blue – sequences that fold, but cannot catalyze and red – sequences which act as catalysts and for which at least one elongation reaction has been catalyzed. For cases 1 and 2, populations of the sequences of the given length are distributed exponentially. Thus we can take mean or median population for the given length as a faithful representation of the behavior of average sequence of that length. The case 3 is drastically different: the populations of the sequences of the given lengths are distributed polynomially. While most of the sequences have very low population for the longer chains, several sequences (mostly autocatalytic ones) have very high ones and constitute most of the biomass. For the case 3 neither mean or median are good representations of the behavior of the chains, as we can see from the figure, all the chains basically separate into two groups with different distributions, this information cannot be shown in the mean or median. Every point on the panels is a time average over 10^6 time points in the steady state interval. Lower limit of 10^{-6} is due to computational precision.

elongating other chains. These enhanced populations of longer chains occur even though the degree of barrier reduction is relatively small.

Case 3 is qualitatively different than cases 1 and 2. Even though cases 1 and 2 have substantial variances, they have well-defined mean values that diminish exponentially with chain length. Case 3 has much bigger variances, and a polynomial distribution of chain lengths, so neither the mean nor median are good representations of the behavior of the chains; see figure 9(Case 3).

3.4.2 The foldamer-catalyst sequences form an autocatalytic set.

The present model makes specific predictions about what molecules constitutes the autocatalytic set – which HP sequences and native structures are in it, and which ones are not. Figure 10 shows a few of the HP sequences that fold to single native structures. Figure 10 (a) shows those foldamers that are catalysts while Figure 10 (b) shows those foldamers that are not catalysts.

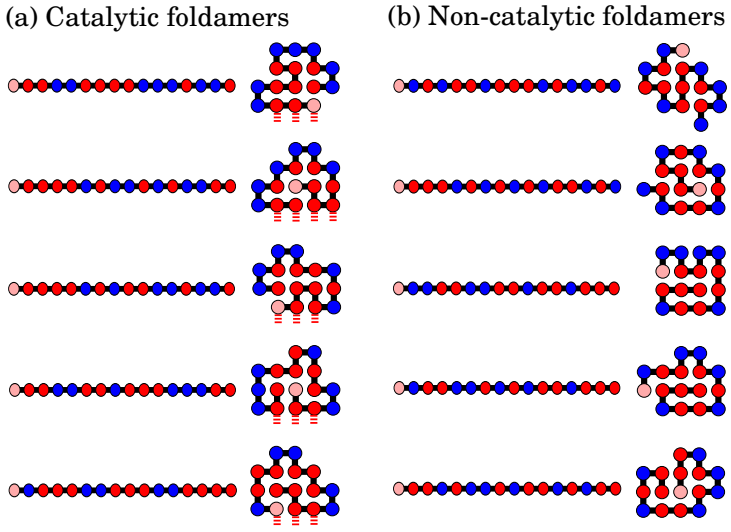

Figure 10: **(a) HP lattice chains that fold and are autocatalytic.** They fold into unique structures and have landing pads that can catalyze the elongation of each other. **(b) HP chains that fold, but are not catalytic.** Most chains are not catalysts, but the size of the autocatalytic set is non-negligible; see Fig. 12.

In short, all HP sequences that are foldamer-catalysts are members of the autocatalytic set: any two HP foldamer-catalyst sequences are autocatalytic for each other. Figure 11 shows two examples of autocatalytic paired chain elongations. The top row of Figure 11 shows *crosscatalysis*: a polymer A elongates a polymer B while B is also able to elongate A. The bottom row of Figure 11 shows *autocatalysis*: one molecule C elongates a another C molecule in solution.

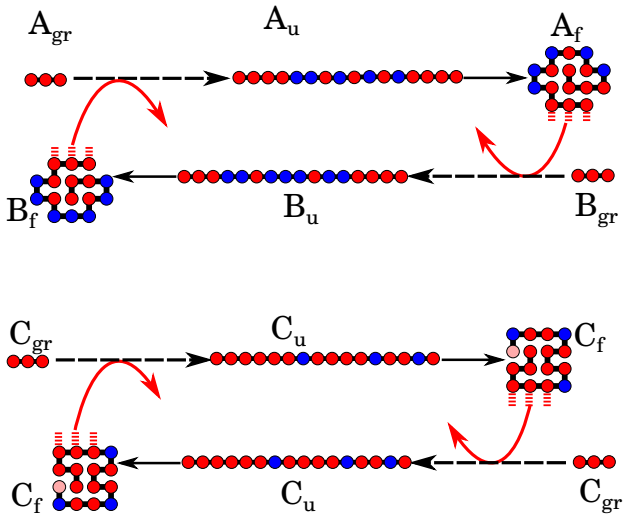

Figure 11: **Top: Cross-catalysis of 2 different sequences. Bottom: Autocatalysis of 2 copies of an identical sequence.** Dashed arrows (**- - -**) represent multiple reactions of chain growth. Among them there are both $\cdots HH + H \rightarrow \cdots HHH$ catalyzed reactions and spontaneous chain elongations. Catalysis is represented by red solid arrows (—). Solid black lines (—) are folding reactions. Chains, which we call "autocatalytic" experience catalysis during one (or more often several) of the steps of elongation. Then, when they reach the length at which they can fold (A_u, B_u, C_u), they fold and serve as catalysts them selves (A_f, B_f, C_f). Mutual catalysis can happen between different sequences (here A and B) and between different instances of the same sequence (here C).

3.4.3 The size of the autocatalytic set grows with the size of the sequence space.

An important question is how the size of an autocatalytic set grows with the size of the sequence space. Imagine first the situation in which the chemistry-to-biology transition required one or two "special" proteins as autocatalysts. This situation is untenable because sequence spaces grow exponentially with chain length. So, those few particular special sequences would wash out as biology moves into an increasingly larger sequence space sea. In contrast, Figure 12 shows that the present mechanism resolves this problem. On the one hand, the fraction of HP sequences that are foldamers is always fairly small (about 2.3% of the model sequence space), and the fraction of HP sequences that are also catalysts is even smaller (about 0.6% of sequence space). On the other hand, Figure 12 shows that the populations of both foldamers and foldamer-cats grow in proportion to the size of sequence space.

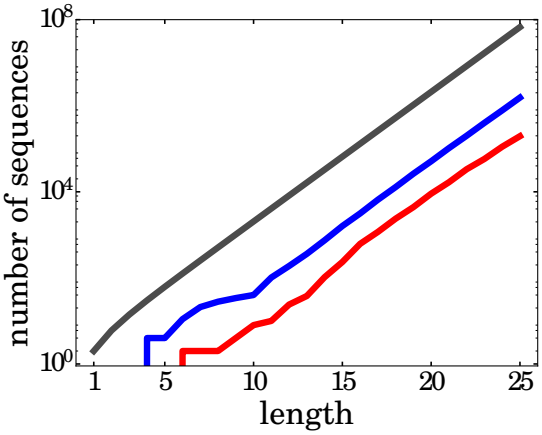

Figure 12: **Different sequence spaces grow exponentially with chain length.** (Gray) The number of all HP sequences. (Blue) The number of foldamers. (Red) The number of foldamer catalysts.

The implication is that the space of autocats in the CTB might have been huge.

Figure 13 makes a closely related point. It shows that for longer chains, the fraction of biomass that is produced by autocatalysts completely takes over and dominates the polymerization process, relative to just the basic polymerization dynamics itself, even though the catalytic enhancements are quite modest. This is due to two factors: (1) the number of autocatalysts grow longer sequences (see fig.12 and (2) folding alone is not sufficient to populate longer chains. We find that the average hydrophobicity of the dominant sequences in these runs is 68%.

At this point, we note what our model is, and what it is not. Our model is not intended as an accurate atomistic depiction of a real catalytic mechanism. It is a coarse-grained toy model, of which there will be variants. The mechanism we explore here is the translational localization of the two reactants, polymer B and monomer C, in the chain extension reaction. And, while this model is 2-dimensional, extensive previous studies have shown that it captures many important principles of folding and sequence-to-structure relationships. At the present time, this type of model is the only unbiased, complete and practical way to explore plausibilities of physical hypotheses such as the present one.

We note that the present model is not necessarily exclusive to proteins.

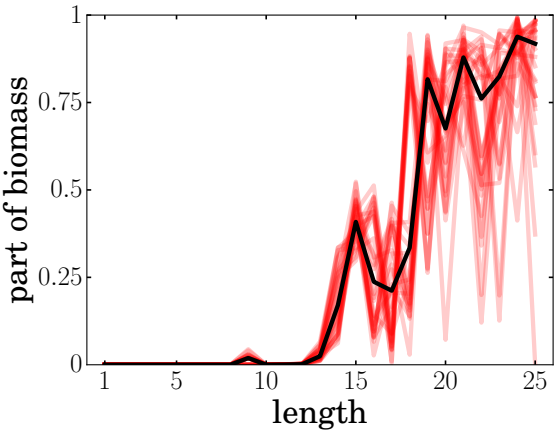

Figure 13: **The longer chains the chains, the bigger the contribution of the autocatalysts.** Each red line shows how the contribution of autocatalytic chains to the biomass of the given length grows with chain length. Different red lines correspond to different simulation runs. The black line shows the median over 30 simulations.

Nucleic acid molecules are also able to fold in water, indicating differential solvation. While our present model focuses on hydrophobic interactions, it is simply intended as a concrete model of solvation, that could more broadly include hydrogen bonding or other interactions. So, while our analysis here is only applicable to foldamers, that does not mean it is limited to proteins. The unique power that foldable molecules have for catalyzing reactions – in contrast to other nonfoldable polymeric structures – is that foldamers lead to precisely fixing atomic interrelationships in relative stable ways over the folding time of the molecule. It resembles a microscale solid, with the capability that substrates and transition states can recognize, bind, and react to those stable surfaces. For example, serine proteases utilize a catalytic triad of 3 amino acids. So, foldability in some type of prebiotic polymer, could conceivably have had a special role in allowing for primitive catalysis. Here, we use a toy model to capture that simple idea, namely that a folded polymer can position a small number of residues in a way that can catalyze a reaction.

3.5 Models and methods

In this section I will go over the physics and computer simulations in more details. This section is an expansion of the Supplementary Information to the [81].

We model our system with a set of likely prebiotic chemical reactions which could happen in cell-like vesicles. To model the sequence-structure relationship for polymers, we use 2D HP lattice model. To determine if a given sequence can fold or act as a catalyst we have to determine if it has a unique free energy conformational minimum. We need to do so because when sequence has the only conformational minimum it tends to stay longer in it and thus is being more stable compared to sequences with several conformational minima. The problem of determining if the sequence has a unique minimum is NP-complete[152] and thus limits our analysis to relatively short chains. Because of that we had to limit length of possible polymers to 25mers. Binary polymers up to length 25 give us about 10^8 possible molecules. Depending on how strongly coupled the system is there will be $\propto 10^8$ to $\propto 10^{16}$ possible reactions. Such tremendous number of reactions is hard to process, yet every particular system cannot have so many species. Moreover, because we essentially try to model small cell-like vesicles, the system would have very small number of molecules (much less than N_A). The best method to determine the behavior of such systems is stochastic simulations. They take into account non-deterministic nature of chemical reactions and allow fluctuations to play important role in the dynamics of a chemical system, which is likely to be important in the emergent phenomena such as the origin of life. The method we use is called the Expandable Partial-Propensity Direct Method (EPDM)[97]. This method allows to work with unlimited number of possible species and reactions if only few of them appear in the system in the same time.

In every simulation we were looking for several properties of the system. We were mostly interested in steady state statistics. In order to determined a steady state we ran several simulations under different conditions and looked at the time, when total number of species stops growing and when length distribution stops changing. Across the set of conditions used in [81] and in this work we found that steady state was reached by 40s of the simulation time. Then we ran simulation for 100 more seconds. Because for every run recordings were made every 0.0001s, this guaranteed 10^6 time points during the steady state to calculate various statistics.

The main statistic we were interested in is a steady state length distribution. To calculate it we took the average population of every sequence over time during the million time steps. Then we summed all the populations of a given length, obtained total populations for all n-mers, $n \in [1, 25]$, and then computed every population as:

$$p_n = \frac{\sum \text{all n-mers}}{\sum \text{total population}} \tag{4}$$

giving probability of finding an n-mer of a randomly chosen molecule in the system. We also looked at statistics of the length distributions over an ensemble of simulations with identical parameters starting with the same initial conditions as well as dependency of the distributions on the parameters.

3.5.1 Experiment 1: Does our bare polymerization reproduces the Flory distribution?

In our first experiment we modeled unassisted prebiotic polymerization. We assumed that polymerization took place in small cell-like vesicles, which serve to preserve high concentration of the reacting molecules. Monomers (H and P) can diffuse into the vesicle. Because polymerization reaction is thermodynamically unfavorable in water[105] these monomers must carry extra chemical energy which can be released during the polymerization and thus drive polymerization reaction. We call such monomers activated ($H^\star$ and $P^\star$). Activation of the monomers can happen with either help of light or CO/H_2S[115] or ATP[153] fore example. Activated monomers can interact with polymers to produce longer polymers. We set the rate of this reaction to $\beta = 1$. This doesn't affect generality of our calculations and sets up a reference rate. All other rates are relative to the polymerization rate.

$$\text{activated 1-mer} + n\text{-mer} \xrightarrow{\beta} (n+1)\text{mer} \tag{5}$$

We assume unaided polymerization is a rather slow process and thus rate of import of activated monomers a has to be much faster. In our simulations we vary this rate in the wide range of $a \in [100, 5000]$.

$$\varnothing \xrightarrow{a} H^\star \text{ or } P^\star \tag{6}$$

Another process which prebiotic polymers would naturally undergo is hydrolysis: water present in the vesicle will break the bonds between monomers.

This process has a constant rate h per bond.

$$n\text{-mer} \xrightarrow{h} l\text{-mer} + (n-l)\text{-mer} \tag{7}$$

Hydrolysis rates strongly depend on the temperature. Hydrolysis rate of proteins under neutral conditions and room temperature are known. Typical values for the half-time for the hydrolysis of a bond under neutral conditions and room temperature are on the order of hundreds of years[11]. Here, we explored a range of hydrolysis rates that are about $0.01 - 1$ of the polymerization rate. This allows us to deduce spontaneous polymerization rates, which in our case are on the order of days.

This elementary model describes basic chemical events which a mixture of prebiotic polymers in a vesicle would undergo. Another important process, which has to be added is vesicle growth. Vesicles are known to grow and split spontaneously as the mass inside them increases[156]. Thus molecules either dilute away or leave the system as vesicles divide. In this model we consider bulk reactions and describe this process by the deletion reaction with the rate d:

$$\text{anything} \xrightarrow{d} \varnothing \tag{8}$$

In our simulations this rate is in the range $d \in [0.01, 0.1]$. The results of the simulations are described in the section 3.4 The source file of the model and parameters of the simulation are located at `https://github.com/gelisa/hp_world_data/tree/master/001`

3.5.2 Experiment 2. What is the effect on the distribution of just HP folding?

In our next experiment we were checking if folding affects the dynamics of the HP polymerization. To do so we allowed HP-polymers with unique conformational minima to undergo folding and unfolding reactions.

Folding and unfolding reactions happen much faster than the polymerization processes, with corresponding rate coefficients of $k_f \gg k_u \gg \beta$:

$$\begin{aligned} \text{folded chain} &\xrightarrow{k_u} \text{unfolded chain} \\ \text{unfolded chain} &\xrightarrow{k_f} \text{folded chain} \end{aligned} \tag{9}$$

[11]The hydrolysis rate constants of oligopeptides in neutral conditions are of the order of $10^{-11} - 10^{-10}$: $1.310^{-10} M^{-1}s^{-1}$ for benzoylglycylphenylalanine ($t_{1/2} = 128y$)[154], $6.310^{-11} M^{-1}s^{-1}(t_{1/2} = 350y)$ for glycylglycine and $9.310^{-11} M^{-1}s^{-1}$ for glycylvaline [155].

We used the most realistic values we could obtain for these rates and for the folding free energies for proteins. We took E_{nat} from the HP model, known folding free energies from experimental data [157, 158], and we used the relationship [150]:

$$\ln\left(\frac{k_f}{k_u}\right) = -\Delta G/kT = E_{nat}/kT - N\ln z, \tag{10}$$

where z is the number of rotational degrees of freedom per peptide bond. To account for the difference between the 2D model and real 3D proteins, we calibrated the parameters taken from the literature to yield unfolding/folding rates that are meaningful in the context of the other rates in our model: folding is much faster than growth and for any of the sequence in our pool $k_f/k_u \in (10^2, 10^4)$ [157, 158] for 3D proteins. Because the literature models are only mean-field, averaged over sequences, and in order to retain sequence dependence here, we set the unfolding rate of all sequences to the average for their lengths, and assigned all the sequence dependence to k_f. So, we used: significantly.

$$\begin{aligned} k_u &= \exp[12 - 0.1\sqrt{N} - E_H(0.5N + 1.34)], \\ k_f &= k_u \exp(\Delta G) \end{aligned} \tag{11}$$

The model is not sensitive to varying these parameters over a wide range. We use $E_h \approx (1-2)$kT, so $k_{unf} \approx 10^2$, which leads to a range of unfolding rates from one unfolding per hour to one unfolding per day. Folding rates vary from a reaction per hour to a reaction per fraction of a second.

In our simulations we started with the same initial population as in Experiment 1. To calculate the result in length distribution, we computed the average population of every sequence for each trajectory over time over all the recordings after 40s, resulting in a million time steps. The results of the experiment are in the section 3.4. The source file of the model and parameters of the simulation are located at `https://github.com/gelisa/hp_world_data/tree/master/002`

3.5.3 Experiment 3. What is the effect on the distribution of both folding and catalysis?

In this experiment we studied our main hypothesis: that hydrophobic interaction responsible for the folding of HP-sequences also is capable of facilitating autocatalytic growth of the polymers by catalyzing H to H bond formation.

The catalytic step is:

$$\text{Catalyst} + H + \underbrace{\cdots HH}_{l-1} \xrightarrow{\beta_{cat}} \text{Catalyst} + \underbrace{\cdots HHH}. \tag{12}$$

The rate enhancement is $\beta_{cat} = \beta \cdot \exp(E_h \cdot n_c/kT)$, where hydrophobic sticking energy is e_H, the number of contacting hydrophobes is n_c, which varies in the range $3-6$. With the hydrophobic energies of $e_H = 1 - 2\text{kT}$, this gives catalysis rates around hours to days per reaction. Because the EPDM supports only binary reactions, we divided the reaction above into to steps: interaction of catalyst with a monomer with rate β and the reaction of this complex with a polymer has the rate β_{cat}.

In addition to folding in this *in-silico* experiment, we also accounted for the pairwise contact interactions between two proteins, with the parameters as indicated above. We explored ranges of parameters. We observed significant stability of the length distribution towards change of h and d in the range: $0.05 \lessapprox d \approx h \lessapprox 0.5$. The distributions we observe are quite sensitive to the choice of hydrophobic energy, as expected for chemical reactions, since this enters into the exponent of the rate expression. In the generally physical range of $e_h = 1 - 3kT$, we observe a bending of the Flory distribution, as noted in the text. The results of the experiment are in the section 3.4. The source file of the model and parameters of the simulation are located at `https://github.com/gelisa/hp_world_data/tree/master/003`

3.6 Conclusion

This section is based on the conclusion section of [81]. Life requires some form of autocatalysis [90, 159, 160]. Molecular mechanism which can explain it is only know for very small and simple molecules[13] or for exquisitely designed RNA sequences[161]. Yet the mechanism which can be responsible for emergence of Kauffman-style set of polymers is unknown. We have found that autocatalysis is inherent in the process of polymerization of HP-polymers. Due to the hydrophobic interaction small fraction of randomly synthesized HP polymers can fold into stable compact states. A fraction of those folded structures have a set of hydrophobic monomers exposed to the surface. They provide "landing pads": hydrophobic surfaces that can help to catalyze the elongation of other HP oligomers (see Figure 11).

The fraction of all the HP sequences that can fold to unique structures (2.3% for lengths up to 25-mers) is not negligible. 12.7% of them have cat-

alytic surfaces which can facilitate growth of other HP sequences. This constitutes 0.3% of the whole sequence space. These ratios are big enough for the sequences to be found by random exploration of the sequence space and remain do not change significantly at least up to 25-mers; see figure 12. It also has been shown that biologically active proteins can be designed based on the HP folding rule [162]. This and the reasonably high frequency of active polymers predicted by HP model suggests that discovery of autocatalytic biologically active sequences during the random prebiotic polymerization is plausible.

Chapter 4: Exact rule-based stochastic simulations for systems with unlimited number of molecular species

This section is an exact copy of [97]

Frequently used abbreviations and symbols

SSA – Stochastic Simulation Algorithm
DM – Direct Method[163]
PDM – Partial propensity Direct Method[164]
SPDM – Sorting Partial propensity Direct Method[164]
EPDM – Expandable Partial propensity Direct Method
N – number of molecular or agent species
M – number of all possible reactions or agent interactions
α – maximum number of possible reactions or agent interactions between any pair of molecular or agent species

4.1 Introduction

Mathematical modeling of chemical reactions is an important part of systems biology research.

The traditional approach to modeling chemical systems is based on the law of mass of action equations which assumes reactions to be macroscopic, continuous and deterministic. It doesn't account for the discreteness of the reactions or temporal and spatial fluctuations, describing only the average properties instead. This makes it ill-suited for modeling nonlinear systems or systems with a small number of participating molecules, such as living cells.

In contrast, *stochastic simulation algorithms* (SSAs) do account for discreteness and inevitable randomness of the process. As a result, these methods are becoming increasingly popular in modern theoretical cell biology[165, 166, 167]. Additionally, they have more physical rigor compared to the empirical law of mass action ordinary differential equations[168].

It is worth mentioning that replacing molecules with any other interacting agents does not invalidate the approach, as long as certain conditions are met. This makes it suitable for modeling many non-chemical systems in areas such as population ecology [169, 170], evolution theory [171, 172], immunology[173], epidemiology [174, 175, 176], sociology [177, 178], game

theory [179], economics [180, 181], robotics [182, 183] and information technology [184].

The majority of contemporary SSAs are based on the algorithm by Gillespie[185, 163, 168] known as *direct method* (DM). It describes a system with a finite state space undergoing a continuous-time Markov process. The time spent in every state is distributed exponentially; future behavior of the system depends only on its current state. Gillespie has shown[168] that for chemical systems these assumptions hold if the system is well-stirred and molecular velocities follow Maxwell-Boltzmann distribution. Because of DM's good physical basis its solutions are as accurate or more accurate than the law of mass equations, which isn't necessarily a good predictor of mean values of molecular populations[163].

Time complexity of DM is linear in the number of reactions. For highly coupled systems that means that the run-time grows as a square of the number of species. This poses a problem for many models in systems biology which describe systems with a multitude of molecular species connected by complex interaction networks. Storage complexity of the DM is linear in the number of possible reactions.

Many alternative SSAs which improve the time complexity of DM (see Table 1 for an incomplete list) were developed. Approximate SSAs make such improvements at the cost of introducing additional approximately satisfied assumptions, while exact SSAs achieve better run-time by performing a faster computation that is equivalent to the one performed by DM. Exact methods have been developed with time per reaction that is linear[164] or, for sparsely connected networks, even constant[186] in the number of molecular species.

One deficiency shared by most SSAs is that they operate with a static list of all possible reactions and species, while for some systems of interest maintaining such a list is not possible. An example of such system is heterogeneous polymers undergoing random polymerization, an object of interest for the researchers of prebiotic polymerization and early evolutionary processes. The number of heteropolymers that can be produced by such a process is infinite. Even if we only consider species of length up to L, we have to deal with $\mathcal{O}(p^L)$ species, where p is the number of monomer species. This causes both time and storage complexity of stochastic models to grow exponentially with L, which limits the studies to very short polymers.

However, if an algorithm can maintain a dynamic list of molecular species and interactions, the complexity of modeling such systems can be substantially reduced. In our example, the set of possible species is so large that

even for moderate cutoff L the vast majority of all possible species will have a population of zero. If at any time step a specie has a population of zero, no reaction can happen in which this specie is a reagent. Therefore, the reactions involving such a specie need not be tracked until some reaction occurs in which the specie is a product.

In addition, many studies are concerned with emergent phenomena, which often involve the system exhibiting some dynamical pattern, e.g. sitting in a stable or chaotic attractor. Such phenomena tend to only involve a subset of possible states and transitions and not explore the space of all possibilities uniformly. This may limit the species and reactions involved in the phenomenon to a small subset of all possible species and reactions. For example, in the study by Guseva et al.[81] effective populations of all sequences are $\propto 10^3$, while the number of all the possible species is $\propto 10^7$. The model is tightly coupled, thus time costs are 10^4 times higher and memory costs are 10^8 times higher than they could be if only the final subset of reactions and species was considered and the SSA retained the time and storage complexity of the best SSAs for static lists (e.g. PDM[164]).

However, not only this subset is unknown prior to the experiment, but the transient to the behavior of interest may involve many more species and reactions than the behavior itself. Thus, the simulation cannot be constrained to use a smaller list of species and reactions if such list is static.

One class of systems in which reactions cannot be listed occurs in solid state chemistry. To simulate those, Henkelman and Jonsson[187] developed an algorithm in which the list of possible reactions is generated on the fly and used in the standard Gillespie's algorithm. This approach was formulated as a part of simulation framework specific to solid state chemistry. It remained obscure outside of the solid state chemistry community and was rediscovered independently by authors of the present work early in the course of its preparation.

Here we present *extendable partial-propensity direct method* (EPDM): a general purpose, exact SSA in which species and reactions can be added and removed on the fly. Unlike Henkelman and Jonsson's approach, our algorithm is based on the *partial-propensity direct method* (PDM) by Ramaswamy et al.[164] and retains its linear time complexity in the number of species. Storage complexity is linear requirements in the number of reactions.

Only the species with nonzero populations and the reactions involving them are recorded and factored into the complexity. This reduces time complexity of executing one reaction by a factor of N_{tot}/N where N_{tot} is the

number of all possible species and N is the number of species with nonzero population at the time of the reaction. Storage complexity is reduced by the square of that factor for densely connected reaction networks.

We test the performance of a C++ implementation of our algorithm[12] for two chemical systems, investigate scaling and compare the performance with two other SSAs. The results confirm our predictions regarding the algorithm's complexity.

Exact	• Direct method (DM) [185, 163] • First reaction method (FRM) [185] • Gibson-Bruick's next-reaction method (NRM)[188] • Optimized direct method (ODM) [189] • Sorting direct method (SDM) [190] • Partial propensity direct method (PDM) [164]
Approximate	• τ-leaping [191, 192, 193, 194] • k_α-leaping [191] • Implicit τ-leaping [195] • The slow-scale method [196] • R-leaping [197] • L-leap [198] • K-leap [199]

Table 1: List of some exact and approximate stochastic simulation algorithms

4.2 Direct stochastic algorithms now

4.2.1 Gillespie Algorithm

Being an exact SSA, our algorithm performs an optimized version of the same basic computation as the Gillespie's DM[200, 163]. DM is based on the following observation:

> Probability that any particular interaction of agents occurs within a small period of time is determined by and proportional to a

[12]https://github.com/abernatskiy/epdm

product of a rate constant specific to this interaction and a number of distinct combinations of agents required for the interaction.

Since we developed EPDM with chemical applications in minds, we will hereafter refer to agents of all kinds as molecules. However, as long as the observation above holds, the method is valid for any other kind of agent (see Introduction).

The algorithm starts with initialization (**Step 1**) of the types and numbers of all the molecules initially present in the system, reaction rates and the random number generator. Then propensities of all reactions are computed (**Step 2**). Propensity a_μ of a reaction R_μ ($\mu \in \{1, 2, ..., M\}$) is proportional to its reaction rate c_μ:

$$a_\mu = h_\mu c_\mu. \tag{13}$$

Here, h_μ is the number of distinct molecular reactant combinations for reaction R_μ at the current time step, a combinatorial function which depends on the reaction type and the numbers of molecules of all reactant types [185]. Total propensity is the sum of propensities of all reactions:

$$a = \sum_{\mu=1}^{M} a_\mu. \tag{14}$$

The next step (**Step 3**), called Monte Carlo step or sampling step, is the source of stochasticity. Two real-valued, uniformly distributed random numbers from $[0, 1)$ are generated. The first (r_1) is used to compute time to next reaction τ:

$$\tau = \frac{1}{a} \ln\left(\frac{1}{r_1}\right). \tag{15}$$

The second one (r_2) determines which reaction occurs during the next time step τ. The j-th reaction occurs if

$$\sum_{\mu=1}^{j-1} a_j \leqslant a r_2 < \sum_{\mu=1}^{j} a_j. \tag{16}$$

The next step is update (**Step 4**): simulation time is increased by τ generated at Step 3, molecules counts are updated using the stoichiometric numbers of the sampled reaction and propensities are updated in accordance with the new molecular counts.

The last step is iteration: go back to Step 3 unless some termination condition is met. Termination should occur if no further reactions are possible (i.e. when the total propensity $a = 0$). Optional termination conditions may include reaching a certain simulation time, performing a given number of reactions, reaching some steady state etc.

At every step the algorithm looks through the list of all M possible reactions. Therefore, the time it takes to process one reaction (i.e. perform Steps 3 and 4) is proportional to M:

$$t_{reac} = \mathcal{O}(M). \tag{17}$$

There must be a record for every reaction, so the space complexity is also $\mathcal{O}(M)$.

4.2.2 Partial propensity methods

For many systems it is valid to neglect reactions which involve more than two molecules or agents. This premise allows for a class of partial-propensity direct methods. The method described in the present work is among those.

If α is the maximum number of reactions which may happen between any pair of the reagent species then $M = \mathcal{O}(\alpha N^2)$. Then the expression for the time complexity of DM (17) becomes quadratic in N:

$$t_{reac} = \mathcal{O}(\alpha N^2). \tag{18}$$

Partial-propensity direct method (PDM) and sorting partial-propensity direct method (SPDM) [164] improve this bound to $\mathcal{O}(\alpha N)$ by *associating each reaction with one of the involved reagents* and *sampling the reactions in two stages*. In the first stage the first reactant specie of the reaction to occur is determined; this takes $\mathcal{O}(N)$ operations. The second stage determines the second specie and a particular reaction to occur; this takes $\mathcal{O}(\alpha N)$ operations, and the total complexity of the sampling step adds up to $\mathcal{O}(\alpha N)$.

Since any specie can be involved in at most $\mathcal{O}(\alpha N)$ uni-molecular or bi-molecular reactions, only $\mathcal{O}(\alpha N)$ values have to be updated when the molecular counts change. This enables PDM and SPDM to perform the update step without worsening the time complexity of the sampling step.

The final time complexity of these algorithms

$$t_{reac} = \mathcal{O}(\alpha N) \tag{19}$$

holds irrespective of the degree of coupling of the reaction network. The number of records required for sampling is still $\mathcal{O}(M) = \mathcal{O}(\alpha N^2)$.

Note that for sparsely coupled reaction networks, time complexity can be improved further to $\mathcal{O}(1)$[186]; here we won't concern ourselves with such reaction networks.

4.3 Algorithm description

4.3.1 Reaction grouping

Similar to PDM and SPDM[164], our method relies on splitting the reactions into groups associated with one of the reagents. To simplify this procedure, we reformulate all reactions with up to two reagents as bimolecular by introducing the virtual *void specie* $\varnothing$ which always has a molecular count of 1. It can interact with other, real species and itself, however its stoichiometry is always zero. With these new assumptions, unimolecular reactions are reformulated as follows:

$$\begin{aligned} &S_i \to \text{products} \Rightarrow \\ &1\, S_i + 0\, \varnothing \to \text{products}. \end{aligned} \tag{20}$$

Source reactions are also reformulated:

$$\begin{aligned} &\varnothing \to \text{products} \Rightarrow \\ &0\, \varnothing + 0\, \varnothing \to \text{products}. \end{aligned} \tag{21}$$

Our algorithm keeps a list of species existing in the system to which entries can be added. When a new specie S_i is added to the list, the algorithm considers every specie S_j in the updated list. For each (unordered) pairing $\{S_i, S_j\}$ a list of possible reactions is generated. If the list is not empty, it is associated with S_j, the specie that has been added to the list earlier unless it coincides with S_i. If $S_i = S_j$, the list is associated with S_i.

For example, the first step of the initialization stage involves adding the first element to the list of species which is always the void specie $\varnothing$. At this point there are no species in the list aside from $\varnothing$, so the algorithm checks which kinds of reactions may happen between $\varnothing$ and itself. Due to the reformulation (21) this will involve all source reactions. Their list will be generated and associated with the newly added specie $\varnothing$.

Another example. Suppose the list of known species is $[\varnothing, S_1]$ and we're adding a specie S_2 which reacts with S_1, itself and also participates in some

unimolecular reactions. The list becomes $[\varnothing, S_1, S_2]$ and the algorithm proceeds to pair up S_2 with every specie in it and generate reaction lists.

Due to the reformulation (20) it will find all the unimolecular reactions for the pair $\{S_2, \varnothing\}$. The list of these reaction will be associated with the previously known specie of the pair, $\varnothing$.

When considering the pair $\{S_2, S_1\}$ it will find all the reactions between S_1 and S_2 and associate them with S_1.

Finally, it will consider a pair $\{S_2, S_2\}$ and find its reactions with itself. Since there are no previously known species in the pair, the reactions will be associated with S_2.

4.4 Propensities generation

As the reactions are generated and associated with species we also compute and store some propensities.

For all reactions R_μ with up to two participating molecules full propensities a_μ are

$$\begin{aligned}
&a_\mu = n_i n_j c_\mu \text{ for a bimolecular reaction} \\
&\quad \text{of distinct species } i \neq j,\ S_i + S_j \to \text{products}, \\
&a_\mu = \frac{1}{2} n_i (n_i - 1) c_\mu \text{ for a bimolecular reaction} \\
&\quad \text{of identical species } 2S_i \to \text{products}, \\
&a_\mu = n_i c_\mu \text{ for a unimolecular reaction} \\
&\qquad\qquad\qquad\qquad S_i \to \text{products}, \\
&a_\mu = c_\mu \text{ for a source reaction } \varnothing \to \text{products}.
\end{aligned} \tag{22}$$

Since the void specie $\varnothing$ has a fixed population of 1 we can use the formula for the unimolecular reactions for source reactions as well: $a_\mu = 1 \cdot c\mu$ for $\varnothing \to$ products. Then for all reactions we can define *partial propensity w.r.t. the reactant specie* S_i as

$$\pi_\mu^{(i)} \equiv a_\mu / n_i. \tag{23}$$

For the reactions with up to two reagents, partial propensities are[164]

$$\begin{aligned}
&\pi_\mu^{(i)} = n_j c_\mu \text{ for a bimolecular reaction} \\
&\quad \text{of distinct species } i \neq j,\ S_i + S_j \to \text{products}, \\
&\pi_\mu^{(i)} = \frac{1}{2}(n_i - 1)c_\mu \text{ for a bimolecular reaction} \\
&\quad \text{of identical species } 2S_i \to \text{products}, \\
&\pi_\mu^{(i)} = c_\mu \text{ for a unimolecular reaction} \\
&\qquad\qquad S_i \to \text{products}, \\
&\pi_\mu^{(\varnothing)} = c_\mu \text{ for a source reaction } \varnothing \to \text{products}.
\end{aligned} \tag{24}$$

Suppose some specie S_j is added to the list of known species as described in section 4.3.1. For every specie S_i in the updated list we generate a list of possible reactions between S_i and S_j and associate it with S_i. For every reaction R_{ijk} in the list, we compute its rate constant c_{ijk} and its partial propensity $\pi_{ijk}^{(i)}$ w.r.t. S_i using formulas (24).

We also keep some sums of propensities to facilitate the sampling. Given a list of reactions between S_i and S_j associated with S_i, we define $\Psi_{ij}^{(i)}$ as

$$\Psi_{ij}^{(i)} = \sum_{k=1}^{\alpha_{ij}} \pi_{ijk}^{(i)}, \tag{25}$$

where α_{ij} is the number of reactions possible between S_i and S_j.

$\Lambda_i^{(i)}$ is the sum of partial propensities of all reactions associated with S_i:

$$\Lambda_i^{(i)} = \sum_{j=1}^{m_i} \Psi_{ij}^{(i)}. \tag{26}$$

Here, m_i is the number of lists of reactions with other species associated with S_i.

Σ_i is the full propensity of all reactions associated with S_i:

$$\Sigma_i = n_i \Lambda_i^{(i)}. \tag{27}$$

and a is defined as a total full propensity of all reactions in the system:

$$a = \sum_{i=1}^{N} \Sigma_i. \tag{28}$$

Data type	Necessary members				Significance
	Containers	Scalars	References	Functions	
`TotalPopulation`	Linked list of `Population`s	Total propensity a (28), current time t, random generator state	–	–	Represents the whole system being modeled
`Population`	Linked list of `Relation`s, linked list of `RelationAddress`es	`Specie` S, molecular count n, propensity sums Σ (27) and Λ (26)	–	–	Represents a population of molecules of a specie S
`Relation`	Linked list of `Reaction`s	Partial propensity sum Ψ (25)	To `RelationAddress` pointing to this `Relation`	–	Data on all reactions possible between a pair of species, to be stored withing the list of an *owner* specie
`RelationAddress`	–	–	To a `Relation`, its owner's `Population`, list containing this `RelationAddress` and itself	–	Reference to a `Relation` to be kept by a non-owner specie, useful in propensity updates and population deletions
`Reaction`	Array of pairs (ID, σ) of IDs and stoichiometries of participating species	Rate constant c, partial propensity π w.r.t. the owner specie (24)	–	–	Represents a reaction
`Specie`	–	Unique, compact specie identifier ID	–	`Specie::reactions(Specie)`	An extended specie representation capable of keeping extra information to generate lists of possible reactions with `reactions()` method quickly

Table 2: List of data structure types used by the algorithm

4.4.1 Data model

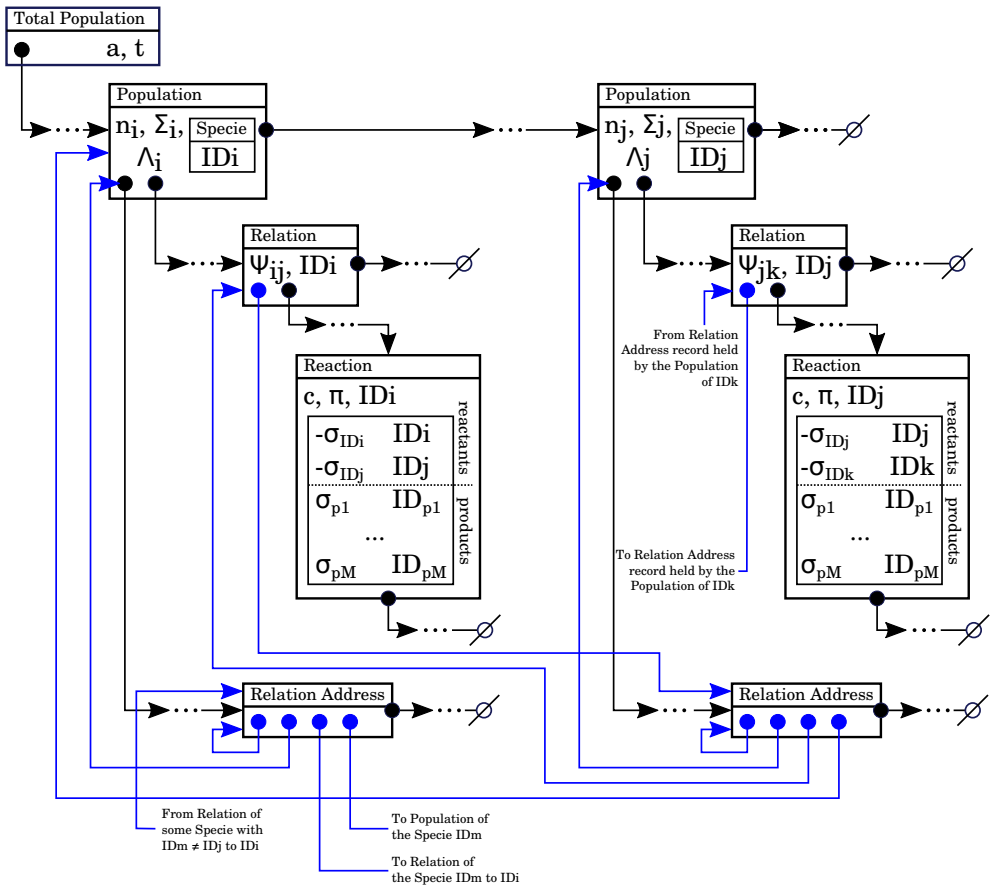

Figure 14: EPDM data model. Circles with arrows denote references. References shown in black form the hierarchical linked list; in implementation, many of those are hidden within the `std::list` container. Auxiliary references added for dynamic updates are shown in blue. Letters and boxes within the boxes indicate non-reference member variables. Crossed empty circle represents a null reference.

Similar to Steps 3 and 4 in Gillespie's DM (see section 4.2.1), execution of each reaction in our algorithm involves two steps: *sampling* and *updating*. We'll refer to the data used in the sampling process as *primary* and call all the data which is used only for updating *auxiliary*.

To minimize the overhead associated with adding and removing the data

we use a hierarchical linked list (multi-level linked list of linked lists). We define several data structure types (see table 2), many of which include lists of instances of other types. Complete data model is shown in Figure 14.

Top level list is stored within a structure of type `TotalPopulation`. Aside from the list, the structure holds the data describing the system as a whole: total propensity a (see (28)), current simulation time t and the random generator state.

Elements of the top level list are structures of type `Population`, each of which holds the information related to the molecular population of a particular specie S_i. The information about the specie itself, to which we will in the context of its `Population` refer as the *owner specie*, is represented by a member object of user-defined class `Specie` (see note 2 at the end of this section), containing a unique string specie identifier ID_i as a member variable. `Population` also contains the number of molecules in the population n_i, total propensity of the population Σ_i and total partial propensity $\Lambda_i^{(i)}$ w.r.t. S_i of all reactions associated with S_i (see eqs. (27) and (26)).

The `Population` structures are appended to the top level list sequentially over the course of the algorithm execution (see section 4.3.1). Each of them holds, in addition to the data mentioned above, two linked lists: of structures of type `Relation` and of structures of type `RelationAddress`. The former stores the list of all reactions possible between the owner specie and some other specie that has been added later in the course of the algorithm's operation. `RelationAddress` structures hold the references used to access `Relation`s in which the owner specie participates, but not as an owner. Those include all the species added before the owner.

It can be observed that the `Population` which has been added first will necessarily has an empty list of `RelationAddress`es and a potentially large list of `Relation`s, with as many elements as there are species with which the first specie has any reactions. On the other hand, the `Population` added last will not be an owner of any `Relation`s and its list of those will be empty. In this case, all information about the specie's reactions is owned by other species and only available within the `Population` through the list of `RelationAddress`es.

A `Relation` structure owned by S_i holds a linked list of all possible `Reaction`s between S_i and S_j and their total partial propensity $\Psi_{ij}^{(i)}$ (see eq. (25)). `Reaction`s store the reaction rate c_{ijk}, partial propensity $\pi_{ijk}^{(i)}$, a table of stoichiometric coefficients and IDs of all participating species, in-

cluding products. Additionally, each of them stores an auxiliary reference to the `RelationAddress` structure pointing to this `Relation`.

`RelationAddress` is an auxiliary structure holding references to a `Relation` in which the owner specie participates without owning it, and to the `Population` of the specie which does own the `Relation`. It also contains the references to itself and to the list holding it.

Notes:

1. We will say that a `TotalPopulation` is *valid* iff all variables mentioned in eqs. (24–28) satisfy these equations and all references mentioned in the present section are valid and point where intended.

2. Unique specie string ID format and `Specie` objects are user-defined and must be freely convertible between each other. The utility of having a separate `Specie` class lies in it having a user-defined method `reactions()` : $(self, \texttt{Specie}) \to ListOfReactions$ which produces a list of all reactions possible between the specie described by a caller object and a specie described by the the argument object. For some systems, this computation can be made much faster if some auxiliary data can be kept in the structure describing a specie. String IDs on the other hand are intended as memory efficient representations of specie data which are used when this computation may not be needed, e.g. to represent reaction product not yet present in the system or for specie comparison.

   Our method does not rely on this separation other than as a means of optimization.

3. Additionally, in our implementation we provide a global structure to hold the parameters of the system which may influence the behavior of the `reactions()` method of class `Specie`.

4.4.2 Adding a `Population`

Algorithm 1 Adding a `Population`

1: **function** AddPopulation(tp, ID, n)
2: Append a new `Population` p_N to tp's list
3: Convert ID to a `Specie` object S
4: $S_N \leftarrow S$, $n_N \leftarrow n$
5: $\Lambda_N^{(N)} \leftarrow 0$, $\Sigma_N \leftarrow 0$
6: **for all** `Populations` p_i in tp's list **do**
7: $ListOfReactions \leftarrow S_N$.`reactions`($S_i$)
8: **if** $ListOfReactions$ is empty **then**
9: **continue**
10: **end if**
11: Append a new `Relation` ρ_{iN} to p_i's list
12: Append all `Reactions` in $ListOfReactions$ to ρ_{iN}'s list
13: $\Psi_{iN}^{(i)} \leftarrow 0$
14: **for all** `Reactions` R_{iNk} in ρ_{iN}'s list **do**
15: Compute $\pi_{iNk}^{(i)}$ using (24)
16: $\Psi_{iN}^{(i)} \leftarrow \Psi_{iN}^{(i)} + \pi_{iNk}^{(i)}$
17: **end for**
18: $\Lambda_i^{(i)} \leftarrow \Lambda_i^{(i)} + \Psi_{iN}^{(i)}$
19: Append a new `RelationAddress` structure RA_{Ni} to p_N's list
20: Store references to ρ_{iN}, p_i, RA_{Ni} and p_N's list of `RelationAddresses` at RA_{Ni}
21: Store reference to RA_{Ni} at ρ_{iN}
22: **end for**
23: Recompute a using (28)
24: **end function**

Suppose that we have a valid `TotalPopulation` with $N-1$ `Populations`. To add a `Population` based on a specie ID and a molecular count n, we follow the method described in section 4.3.1 (see also algorithm 1). We begin by appending a `Population` p_N to `TotalPopulation`'s list. p_N's `Specie` structure S_N is converted from ID, its molecular count $n_N \leftarrow n$ and propensities $\Lambda_N^{(N)} \leftarrow 0$, $\Sigma_N \leftarrow 0$.

Next, for each `Population` p_i in `TotalPopulation`'s updated list (including the newly added p_N) we compute a list of all possible reactions between S_N and p_i's `Specie` S_i. If the list is not empty, it is then converted into `Relation` ρ_{iN} and stored at p_i's list of those. To do the conversion, we compute partial propensities $\pi^{(i)}_{iNk}$ and $\Psi^{(i)}_{iN}$ using eqs. (24) and (25). We also store copies of S_i's ID at every `Reaction` in the list and at the `Relation` ρ_{iN} itself, to keep track of the specie w.r.t. which the current partial propensities are computed. ρ_{iN}'s reference to the `RelationAddress` structure is invalid at this point.

After appending the `Relation` we update p_i's propensities: $\Lambda^{(i)}_i \leftarrow \Lambda^{(i)}_i + \Psi^{(i)}_{iN}$, $\Sigma_i \leftarrow n_i \Lambda^{(i)}_i$.

We then proceed to create a `RelationAddress` structure RA_{Ni} pointing to ρ_{iN}. A blank structure is appended to p_N's list. We take references to ρ_{iN}, p_i, RA_{Ni} and p_N's list of `RelationAddress`es and save them at RA_{Ni}.

Then, we take the reference to RA_{Ni} and store it at ρ_{iN}. At this point, all references in the whole structure are valid and correct.

Processing each p_i in the `TotalPopulation`'s list takes $\mathcal{O}(\alpha)$, so the whole procedure up to this point takes $\mathcal{O}(\alpha N)$. At this point we need to update the total propensity of the system a by recomputing it, which takes $\mathcal{O}(N)$[13].

The total number of operations it takes to add one `Population` is $\mathcal{O}(\alpha N)$. The operation preserves the validity of the data structure.

4.4.3 Initialization stage

To initialize the data structure, we make a `TotalPopulation` with an empty list of `Population`s, initialize $a \leftarrow 0$, $t \leftarrow 0$ and the random generator with a seed. The only variable that needs to take a particular value for the `TotalPopulation` to be valid is a and it has the correct value of 0; therefore, it is a valid `TotalPopulation`. We build the data structure by adding the initial populations to this structure as described in section 4.4.2. The resulting structure is valid because we only used validity-preserving operations.

Adding every population takes $\mathcal{O}(\alpha N)$ operations and it must be repeated N times. This brings the complexity of the initialization step to $\mathcal{O}(\alpha N^2)$.

[13]It can be done during the iteration in $\mathcal{O}(1)$, but we chose to recompute it for improved numerical accuracy.

4.4.4 Deleting a `Population`

Algorithm 2 Deleting a `Population`

1: **function** DELETEPOPULATION(tp, p_i)
2: **for all** `Relation`s ρ_{ij} in p_i's list **do**
3: Follow ρ_{ij}'s reference to `RelationAddress` pointing at it, RA_{ji}
4: Use the reference at RA_{ji} to itself and to the list holding it to remove it from the list
5: **end for**
6: **for all** `RelationAddress` RA_{ij} in p_i's list **do**
7: Follow RA_{ij}'s reference to the `Relation` ρ_{ji} and to the `Population` p_j owning it
8: Remove ρ_{ji} from p_j's list
9: **end for**
10: Delete all the date in the p_i structure
11: Remove p_i from tp's list
12: **end function**

Our algorithm is designed to keep track only of the species with a nonzero molecular count and reactions involving them. To accomplish that, we use the addition operations described in the previous section and deletion operation described here (see also algorithm 2). Deletion is only ever applied to `Populations` of species with zero molecules. All propensities of reactions involving such species are zeros; this enables us to simplify the procedure.

To remove a `Population` p_i, we begin by removing all `RelationAddress` structures pointing at `Relation`s in p_i's list. Each `Relation` ρ_{ij} in p_i's list contains a reference to the `RelationAddress` structure pointing at it, RA_{ji}, which in turn contains a reference to itself and to a list holding it. We use those to remove each RA_{ji} from its list. Since S_i can be involved in at most N relations, this step takes $\mathcal{O}(N)$ operations.

Next, we remove all `Relation`s in which S_i participates, but which are owned by other species. For each `RelationAddress` RA_{ij} we follow its references to the `Population` p_j of the other specie and to its `Relation` ρ_{ji} with S_i. We use those to remove ρ_{ji} from p_j's list. This step also takes $\mathcal{O}(N)$ operations.

Finally, we delete the whole `Population` structure p_i from `TotalPopulation`'s

list. We recursively remove all the structures in its lists, which takes $\mathcal{O}(N)$ operations for the list of `RelationAddress`es and $\mathcal{O}(\alpha N)$ operations for the list of `Relations`. The resulting `TotalPopulation` is valid since all the propensities of the reactions involving S_i are zeros and the remaining propensity sums has not changed; it also contains no invalid references.

The final complexity of the deletion operation is $\mathcal{O}(\alpha N)$.

4.4.5 Sampling stage

Similarly to DM[185] and PDM[164], our algorithm simulates the system by randomly sampling time to the next reaction and the reaction itself with certain distributions. Here we describe how it happens in our algorithm.

We begin by generating two random numbers r_1 and r_2. The first one is used to compute the time to the next reaction τ exactly as in DM and PDM (see eq. (15)). The second random number r_2 is used to sample the reaction similarly to how its done in PDM[164].

The sampling process has three stages (see algorithm 3). During the first stage (lines 2-6) we determine the first specie participating in the reaction to happen. To this end, we go through the list of `Populations` $p_1 ... p_N$ until the following condition is satisfied:

$$\sum_{i=1}^{J} \Sigma_i \leqslant a r_2 < \sum_{i=1}^{J+1} \Sigma_i. \tag{29}$$

The second stage (lines 7-11) involves finding the second reactant. We look for a `Relation` ρ_{JK} among those attached to the `Population` p_J for which the following condition holds:

$$\sum_{i=1}^{J} \Sigma_i + n_J \sum_{i=1}^{K} \Psi_{Ji}^{(J)} \leqslant a r_2 < \sum_{i=1}^{J} \Sigma_i + n_J \sum_{i=1}^{K+1} \Psi_{Ji}^{(J)} \tag{30}$$

During the third stage (lines 12-16), we determine which of the reactions possible between the two species is going to happen. We go through ρ_{JK}'s list of the `Reaction`s, looking for a reaction R_{JKL} such that

$$\begin{aligned} \sum_{i=1}^{J} \Sigma_i + n_J \sum_{i=1}^{K} \Psi_{Ji}^{(J)} + n_J \sum_{i=1}^{L} \pi_{JKi}^{(J)} \leqslant a r_2 < \\ < \sum_{i=1}^{J} \Sigma_i + n_J \sum_{i=1}^{K+1} \Psi_{Ji}^{(J)} + n_J \sum_{i=1}^{L+1} \pi_{JKi}^{(J)}. \end{aligned} \tag{31}$$

Note how equations (29) and (30–31) are similar, but their implementation in the pseudocode (algorithm 3) is different. This design minimizes sampling errors due to floating point representation, making the first stage exact.

The resulting reaction sampling finds a reaction in exactly the same manner as equation (16) does. However, the first and the second stages of this sampling process take $\mathcal{O}(N)$ steps and the third step takes $\mathcal{O}(\alpha)$ steps, resulting in a total time complexity of $\mathcal{O}(N + \alpha)$. Using (16) directly requires $\mathcal{O}(M)$ steps[185], which is $\mathcal{O}(\alpha N^2)$ for densely connected reaction networks.

Algorithm 3 Reaction sampling

1: **function** SAMPLEREACTION(tp, r_2)
2: $s_1 \leftarrow 0$, $s_2 \leftarrow 0$
3: **while** $ar_2 > s_1$ **do**
4: Get a new `Population` p_i from tp's list
5: $s_2 \leftarrow s_1$, $s_1 \leftarrow s_1 + \Sigma_i$
6: **end while**
7: $g \leftarrow (ar_2 - s_2)/n_i$
8: **while** $g > 0$ **do**
9: Get a new `Relation` ρ_{ij} from p_i's list
10: $g \leftarrow g - \Psi_{ij}^{(i)}$
11: **end while**
12: $g \leftarrow g + \Psi_{ij}^{(i)}$
13: **while** $g > 0$ **do**
14: Get a new `Reaction` R_{ijk} from ρ_{ij}'s list
15: $g \leftarrow g - \pi_{ijk}^{(i)}$
16: **end while**
17: **return** R_{ijk}
18: **end function**

4.4.6 Updating stage

Algorithm 4 Post-sampling data update

1: **function** UPDATEEXISTINGPOPULATIONS(tp, R_{IJK})
2: **for all** species S_i participating in R_{IJK} **do**
3: Search for a `Population` of specie S_i in tp's list using its ID_i
4: **if** S_i has a `Population` p_i in tp's list **then**
5: $n_i \leftarrow n_i + \sigma_i(R_{IJK})$
6: **for all** `RelationAddress`es RA_{ij} in p_i's list **do**
7: Follow the references at RA_{ij} to get the `Relation` ρ_{ji} and its owner `Population` p_j
8: $\Lambda_j^{(j)} \leftarrow \Lambda_j^{(j)} - \Psi_{ji}^{(j)}$
9: $\Psi_{ji}^{(j)} \leftarrow 0$
10: **for all** `Reaction`s R_{jik} in ρ_{ji}'s list **do**
11: Recompute $\pi_{jik}^{(j)}$ using (24)
12: $\Psi_{ji}^{(j)} \leftarrow \Psi_{ji}^{(j)} + \pi_{jik}^{(j)}$
13: **end for**
14: $\Lambda_j^{(j)} \leftarrow \Lambda_j^{(j)} + \Psi_{ji}^{(j)}$
15: $\Sigma_j \leftarrow n_j \cdot \Lambda_j^{(j)}$
16: **end for**
17: **else**
18: ADDPOPULATION(tp, ID_i, $\sigma_i(R_{IJK})$)
19: **end if**
20: **end for**
21: **for all** `Population`s p_i in tp's list **do**
22: **if** $n_i == 0$ **then**
23: DELETEPOPULATION(tp, p_i)
24: **end if**
25: **end for**
26: **end function**

When the reaction to occur R_{JKL} is known, our algorithm proceeds to update the data to reflect the changes in species' populations and propensities (see also algorithm 4). For every specie S_i involved in R_{JKL} we read

its ID_i from the array stored at R_{JKL}. We run a sequential search for this specie's `Population` p_i over the list kept in `TotalPopulation`. If the specie's `Population` is found, its molecular count n_i is updated using the stoichiometric coefficient $\sigma_i(R_{JKL})$:

$$n_i \leftarrow n_i + \sigma_i(R_{JKL}). \tag{32}$$

Stoichiometric coefficients are negative for reagents, so their molecular counts may become zero after this step.

After updating the molecular count, we also update all partial and total propensities which depend on it. From every `RelationAddress` RA_{ij} in p_i's list we obtain references to a `Relation` ρ_{ji} in which S_i participates and to the `Population` p_j of its owner specie. For ρ_{ji} we recompute all $\pi^{(j)}_{jik}$ and $\Psi^{(j)}_{ji}$ from scratch using formulas (24) and (25). For p_j, we update the propensity sums as follows:

$$\begin{aligned} \Lambda^{(j)}_j &\leftarrow \Lambda^{(j)}_j - \Psi^{(j)}_{ji}, \\ \Sigma_j &\leftarrow n_j \Lambda^{(j)}_j. \end{aligned} \tag{33}$$

Since each of the species involved in R_{JKL} may be involved in $\mathcal{O}(\alpha N)$ reaction, updating the structure in this way takes a total of $\mathcal{O}(\alpha N)$ operations.

If the specie S_i is a product, its `Population` may not exist yet. In this case we add a new `Population` of S_i using its ID_i as described in section 4.4.2. The molecular count of the newly added specie is its stoichiometric coefficient in R_{JKL}, $\sigma_i(R_{JKL})$. Additions take $\mathcal{O}(\alpha N)$ operations.

When we're done updating the existing `Population`s and adding the new ones, we iterate through the list of `Population`s again and delete the ones with a molecular count of zero as described in section 4.4.4. The deletion takes $\mathcal{O}(\alpha N)$.

Finally, we recompute the total propensity of the system a using (28).

4.4.7 Summary of EPDM

EPDM is an SSA which only maintains the data about the species with nonzero molecular count (see algorithm 5). This ensures that the number of tracked molecular species N is as low as possible.

Our algorithm uses a data structure described in section 4.4.1. The structure holds one entry for each possible reaction, bringing storage requirements of our algorithm to $\mathcal{O}(\alpha N^2)$.

The data structure is initialized by constructing it to be empty, then adding the specie data for every specie initially present in the system. The process is described in section 4.4.3 and takes $\mathcal{O}(\alpha N^2)$ operations.

Each step of the simulation executes a single reaction. It is composed of a sampling step (see section 4.4.5) and an updating step (section 4.4.6). Each of these takes $\alpha\mathcal{N}$ operations, so the total number of operations needed to simulate one reaction is also $\mathcal{O}(\alpha N)$.

When some reaction produces any number of molecules of a previously unknown specie, specie data is added as described in section 4.4.2. To generate the reactions dynamically, a user-defined function `reactions()` is used which takes two species and produces a list of reactions between them, complete with rates.

When any specie's molecular count reaches zero, its data is pruned from the structure as described in section 4.4.4.

Algorithm 5 EPDM overview

```
1: ▷ Initialization stage, see section 4.4.3
2: Create TotalPopulation tp based on initial conditions
3: while a > 0 and termination conditions not met do
4:    ▷ Sampling stage, see section 4.4.5
5:    Generate two random numbers r_1 and r_2
6:    Compute τ using (15)
7:    R ←SampleReaction(tp, r_2)
8:    ▷ Updating stage, see section 4.4.6
9:    for all species S_i involved in R do
10:      Search tp's list for Population of S_i
11:      if a Population p_i of S_i has been found then
12:         Update S_i's molecular count n_i
13:         Update all partial propensities which
                  depend on n_i
14:      else if no Population of S_i has been found then
15:         Add a Population of S_i to tp's list
16:         ▷ details in section 4.4.2
17:      end if
18:   end for
19:   Find all Populations with n == 0 in tp's list
20:   Remove all found Populations
21:   ▷ details in section 4.4.4
22: end while
```

4.4.8 Implementation

We implemented our algorithm as a C++ framework. The user must define a class `Specie` with a constructor from a string ID which must save the string into the member variable `m_id`. The class must define a method `Specie::reactions(Specie)` returning a list of possible `Reaction`s between the caller `Specie` and the argument.

After implementing the class `Specie`, users can simulate the system. Two stopping criteria are currently available by default: the algorithm can stop either when a certain number of reactions have been executed or a certain simulation time has passed.

A global dictionary with arbitrary parameters loaded from a configuration file is provided for convenience.

The implementation is currently available for Linux and Mac OS X. It is tested with GNU gcc 4.9.3 and GNU make 4.1.

The code is available at `https://github.com/abernatskiy/epdm`.

4.5 Benchmarks

We benchmark performance of EPDM against two direct methods: PDM[164][14] and DM[201][15]

Because our model was designed with complex systems in mind we studied performance only for strongly coupled systems. In both systems every specie can interact with every other specie in a unique way, ensuring that the number of reactions is

$$M = \frac{N(N+1)}{2} \tag{34}$$

and

$$\alpha = 1. \tag{35}$$

Both models are designed in such a way that the total number of species is preserved throughout the simulation time.

Models below are designed to measure the performance of our method. They don't fully illustrate the power of the model because they have fixed number molecules, which is necessary to keep for benchmark. The most striking performance gain is achieved when listing all the species isn't possible in principle or due to computational costs. For example, in case of realistic polymerization and autocatalysis model used to study prebiotic polymerization [81] it was possible to increase the maximum length of simulated polymers from 12 to 25 by employing our algorithm. Note that limit of 25 wasn't due to restrictions of our algorithm, but due to necessity to calculate minimum energy folding configuration of every chain, which is an NP-complete problem.

4.6 CPU time of EPDM is linear for a strongly coupled system

The first model ("colliding particles") is made to test how our algorithm performs on chemical systems where no new molecules are created and no

[14] We took implementation from http://mosaic.mpi-cbg.de/pSSALib/pSSAlib.html

[15] We took implementation from http://sourceforge.net/projects/stochkit/

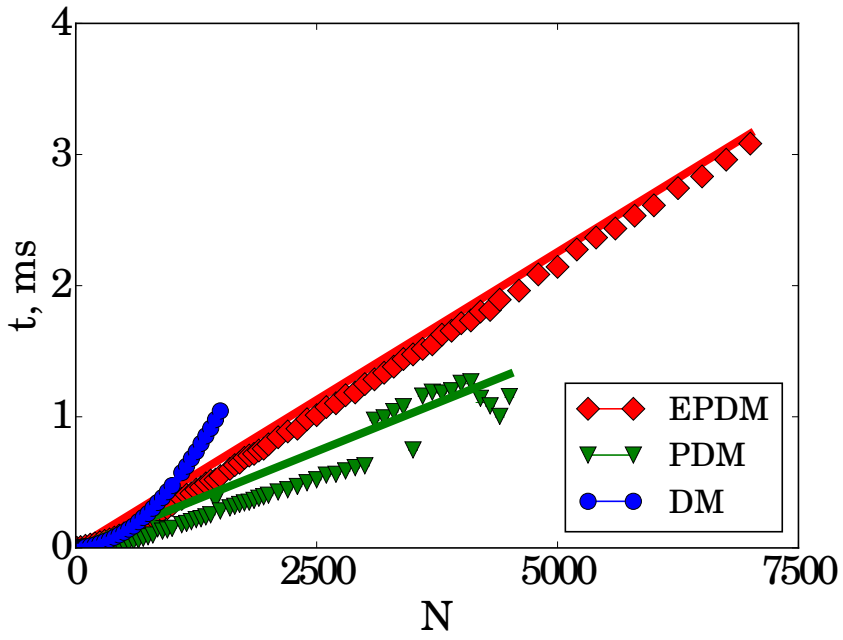

Figure 15: CPU time per reaction as a function of the number of species in the system for "colliding particles" model (section 4.6) for EPDM, PDM and DM.

molecules ever disappear. This is a model of a system consisting of colliding particles of N species. Particles behave like rigid spheres: they collide, and bounce back without internal changes. All of the particle species are known in advance, none are added or removed over the course of the simulation. The system is defined by the following set of equations:

$$P_i + P_j \xrightarrow{k} P_i + P_j, \quad i, j \in [1, N]. \tag{36}$$

In our simulations we vary number of species N from 10 to 7000. Every specie has a population of 50 molecules. Collision rate k is fixed at 0.5 s^{-1}. Every simulation runs until 5000 reactions have occurred. For every value of N, CPU time to simulation completion was measured 10 times and the average time is reported.

Figure 15 shows CPU time it takes per reaction for DM, PDM and EPDM. DM is clearly quadratic. For any given N PDM outperforms the EPDM, but

both are linear in time. It is important to note that DM and PDM were stopped for a relatively low values of N due to excessive RAM consumption (about $\sim$ 120GB) by both of the applications, which we suspect was due to implementation issues (in particular, in XML handling libraries).

4.7 CPU time stays linear when species are actively deleted and created

The second test checks if algorithm keeps linear time when `Population`s are added and removed from the system. The test system ("particles with color") consists of N colliding particles of N types, each of which has an internal property ("color") that is changed in the collision. The following equation defines the system:

$$P_i^\alpha + P_j^\beta \xrightarrow{k} P_i^{\alpha+1} + P_j^{\beta+1}, \quad i,j \in [1,N],\ \alpha,\beta \in [1,\Omega] \tag{37}$$

Indexes i, j enumerate particle types and run from 1 to N. Indexes α, β enumerate colors of particles; particle can have one of Ω colors. During the collision color index of a participating particle goes up $P_i^\alpha \to P_i^{\alpha+1}$, until it reaches maximal index Ω, after which it drops back to 0: $P_i^\Omega \to P_i^0$.

Since every combination of a particle type and a color is considered a separate specie, every reaction causes two species to go extinct and their `Populations` to be deleted. It also adds two new species, which requires adding two new `Populations`. Thus, the total number of species simultaneously present in the system is maintained at exactly the same level. Every specie is represented in the system by a single molecule.

To run such a simulation in DM and PDM frameworks we had to enumerate all $N\Omega$ of the possible species. This slows down the simulation enough to make the comparison impossible beyond a small number of species. The figure 16 shows how CPU time per reaction depends on the number of species in EPDM.

4.8 Conclusion and Discussion

Stochastic simulations are actively used in molecular and systems biology. The bigger and more complex the system, the more important the performance of the simulation algorithm becomes. It is also more burdensome or even impossible to list all the species and reactions for complex systems. We

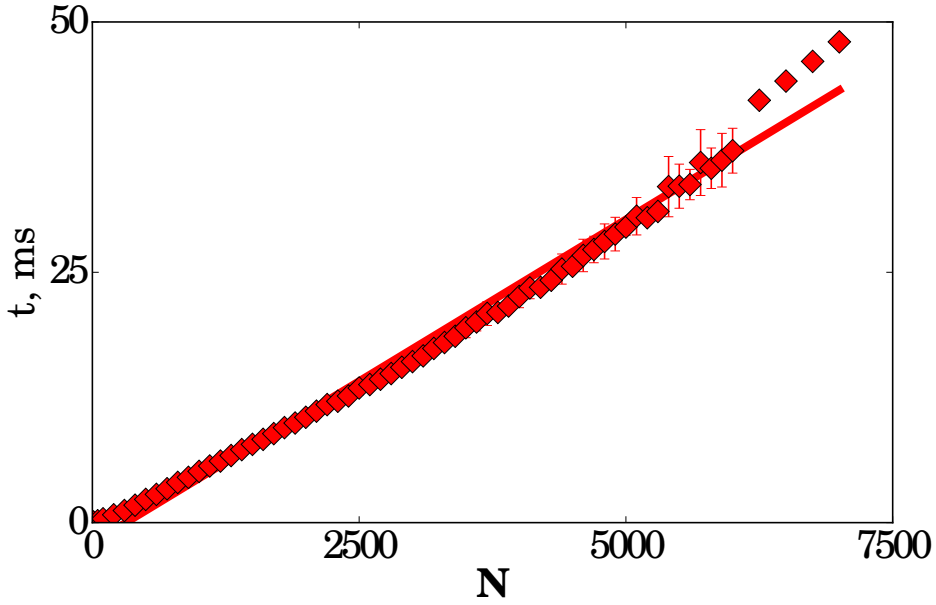

Figure 16: CPU time per reaction for EPDM simulation of "particles with color" (see section 4.7).

introduced a general purpose, exact stochastic simulation algorithm which allows to avoid listing all the possible species and reactions by defining the general rules governing the system instead. Built within the partial propensity framework [164, 186], our algorithm achieves linear time complexity in the number of molecular species.

The algorithm has its limitations. First, it is limited to reactions of maximum of two reactants. In the chemical system that doesn't present much of an issue because reactions with three molecules are significantly rarer then binary reactions and more complex reactions can be represented as sequences of binary reactions. Second, it cannot simulate spatially nonhomogeneous systems. However, as long as all reactions involve no more than two reactants and the observation from section 4.2.1 holds, it is possible to simulate any system for which the set of reactions between any two species is known.

Benchmarks suggest that our algorithm is slower than PDM by a constant factor. Therefore, one should use PDM when it is unlikely that a significant proportion of species will have a molecular count of zero.

Our results suggest that in complex systems (such as polymerization reactions and networks of intra-cellular reactions) and in the case of emer-

gent phenomena studies (e.g. origins of life) EPDM can significantly improve the performance of the stochastic simulations. The software implementation of the algorithm is available as an open source public repository `https://github.com/abernatskiy/epdm`.

Chapter 5: Conclusions and discussion

5.1 Conclusions

Living organisms today use informational polymers such as proteins and nucleic acids for their functioning. It is a common belief that monomeric units of these polymers could have polymerized into short random sequences. However it is not clear how chains long enough for functioning could have been produced and what physical process is responsible for the production of longer non-random chains that could sustain its own production. In this thesis a physical mechanism which explains the emergence of metabolic sets of non-random biopolymers and a plausible mechanism of solving the problem of production of long polymers prebiotically has been proposed.

We have studied physical systems consisting of hydrophobic and polar amino acid-like molecules which are capable of spontaneous random polymerization. We have shown that hydrophobic interaction which drives folding of these polymers can also be a driving force of mutual catalysis. When an oligomer folds and have an exposed hydrophobic patch, this patch can serve as a landing pad for a growing chain and a hydrophobic monomer. This landing pad localizes the growing chain and the monomer and also lowers an activation energy due to hydrophobic interaction. We have showed that such a system can escape Flory problem (the problem of short lengths) and produces self-sustaining sets of non-random polymers.

In order to test our hypothesis, we used Gillespie-like stochastic simulations. We developed an algorithm which allows to simulate systems with potentially infinite number of types of molecules. This algorithm is a general purpose stochastic simulation algorithm that works best for the systems where there are very many (potentially infinite) possible molecular species only few of which are present at every given moment. This algorithm is available through Github[97] for free use.

5.2 Discussion

5.2.1 Evolvability and dynamical behavior of the HP-based autocatalytic ensembles

This section is taken from [81] There are a few problems chemistry-to-biology models in general and autocatalytic models in particular encounter.

One of them is lack of variability and evolvability. Due to the compositional bias or poor dynamical structure of the model such systems converge to one state (attractor or attractor basin) determined by internal dynamics of the system and do not respond to directional selection (see discussions in [119, 28] for example). For a complex system that has many attractors a perturbation can move the system over a threshold to the basin of another attractor. This allows for exploration of the sequence space and thus possible evolvability of the system.

HP ensembles have, we believe, two possible attractors, which allows for the exploration of the sequence space. First, as one can see from the figure 17(a) trajectories split distinctively between two attraction distributions. There are no trajectories that lay in between the two attractors, which shows that there's no switching between the attractors and the separation is not a result of stochasticity. In addition to that each of the distributions has a set of specific sequences which most often dominate the populations. Figure 17(b) shows a few of the structures dominating HP ensembles for the "green" distribution and for the "red" ones. The red and green species differ only in random seeds for the simulations. Each of the two attractors has its own "signature ensemble" of HP sequences that is an emergent property of the dynamics. It is possible that adding more realism to our model (20 monomer types, rather than 2; allowing for longer chains; etc) could lead to larger numbers of attractors. Second, our simulations are limited to 25mers, but in fact the chains can grow longer. This fact allows for the further exploration of the sequence and functionality space beyond what can be seen in our simulations. If we are talking about protein-like molecules, some of the chains will act not only as autocats but also would be capable of binding to other molecules, which could result in a chemical innovation.

5.2.2 Heritability in HP-based autocatalytic systems

One of problems many autocatalytic metabolic systems experience is not only lack of evolvability, but a more basic lack of proper heritability[28]. For a system in order to enjoy a Darwinian evolution it must first of all remember its current state for a mutant to compete with. We studied heritable properties of the HP-systems under random split of the vesicles containing the system. Preliminary results suggests rather poor heritability, but high innovability of the system. The properties of the system aren't however thoroughly studied and is a subject to further research.

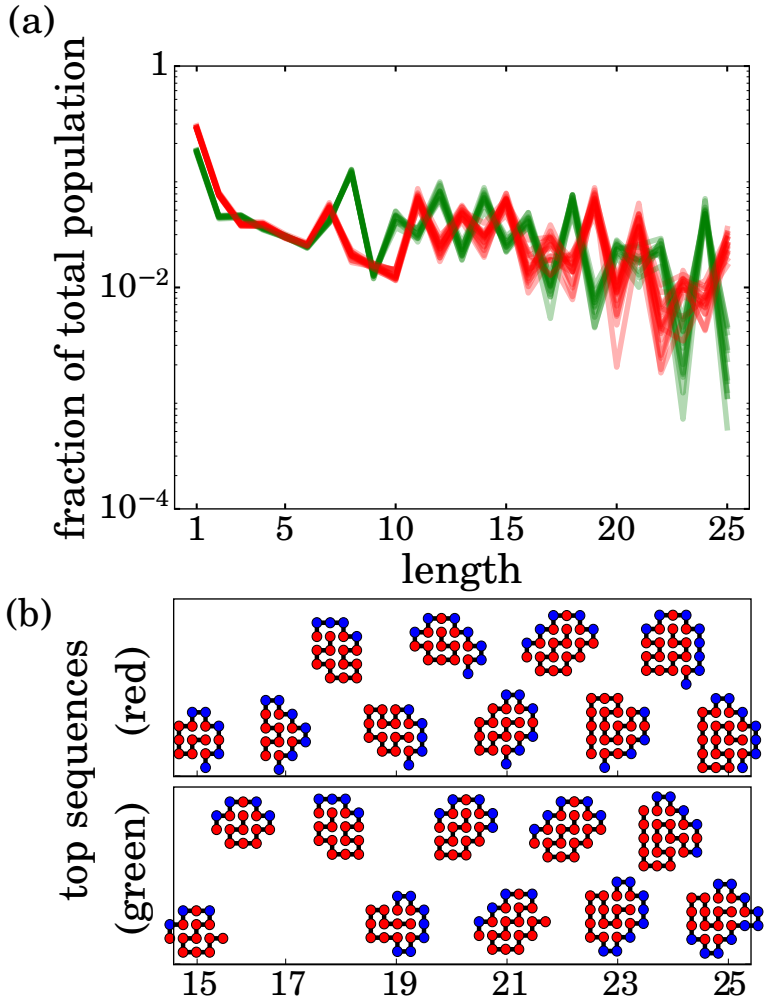

Figure 17: (a) *HP* catalytic system has at least two attractors. The lines are length distributions from case 3. Again, each line represents distribution of length in the steady state for one simulation run. It is clear that there are two kinds of distribution which get realized during the simulations. The system bifurcates either to a state represented by a green line or to one represented by a red one. These are the same lines as on figure 9(a), but separated in two sets by k-means clustering. (b) Structure of the sequences which most often are main contributors into the total population of the polymers of their length. Top panel corresponds to the macrostate shown in red on the panel (a), lower one, to the one shown in green.

References

[1] Karl Popper. *The logic of scientific discovery*, volume 268. 1959.

[2] Radu Popa. *Between necessity and probability: searching for the definition and origin of life*. Springer Science & Business Media, 2004.

[3] Addy Pross. *What is Life?: How chemistry becomes biology*. OUP Oxford, 2012.

[4] Gerald F. Joyce. The RNA World: Life Before DNA and Protein. 1993.

[5] Tom Froese, Nathaniel Virgo, and Takashi Ikegami. Motility at the origin of life: its characterization and a model. *Artificial life*, 20(1):55–76, 1 2014.

[6] Javier Tamames, Rosario Gil, Amparo Latorre, Juli Peretó, Francisco J Silva, and Andrs Moya. The frontier between cell and organelle: genome analysis of Candidatus Carsonella ruddii. *BMC Evolutionary Biology*, 7(1):181, 2007.

[7] Nancy A. Moran and Gordon M. Bennett. The Tiniest Tiny Genomes. *Annual Review of Microbiology*, 68(1):195–215, 9 2014.

[8] B Alberts, A Johnson, J Lewis, M Raff, K Roberts, , and P Walter. *Molecular Biology of the Cell*, volume 54. 2008.

[9] Stuart Kauffman. Beyond Reductionism Twice: No Laws Entail Biosphere Evolution, Formal Cause Laws Beyond Efficient Cause Laws. 3 2013.

[10] Andreas Wagner. *Arrival of the Fittest: Solving Evolution's Greatest Puzzle*. Penguin, 2014.

[11] Richard Dawkins. *The extended phenotype: The gene as the unit of selection*, volume 5. Univesity of Oxford, 1982.

[12] Anne Casselman. Strange but True: The Largest Organism on Earth Is a Fungus. *Scientific American*, 2007.

[13] Gnter von Kiedrowski. A Self-Replicating Hexadeoxynucleotide. *Angewandte Chemie International Edition in English*, 25(10):932–935, 10 1986.

[14] Tracey A. Lincoln and Gerald F. Joyce. Self-Sustained Replication of an RNA Enzyme. *Science*, 323(5918):1229–1232, 2009.

[15] Antonio C. Ferretti and Gerald F. Joyce. Kinetic Properties of an RNA Enzyme That Undergoes Self-Sustained Exponential Amplification. *Biochemistry*, 52(7):1227–1235, 2 2013.

[16] C Darwin and E Mayr. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray, London. *On the Origin of Species by Means of Natural Selection.*, 1859.

[17] Vera Vasas, Chrisantha Fernando, Mauro Santos, Stuart Kauffman, and Ers Szathmáry. Evolution before genes. *Biology direct*, 7(1):1; discussion 1, 1 2012.

[18] Henok Mengistu, Joel Lehman, and Jeff Clune. Evolvability Search: Directly Selecting for Evolvability in order to Study and Produce It. In *Proceedings of the Genetic and Evolutionary Computation Conference*, 2016.

[19] Wikipedia. Phenotype.

[20] Barry Mcmullin, Tim Taylor, Axel Von Kamp, and Maynard Smith. Who Needs Genomes ? pages 1–5, 2001.

[21] Massimo Pigliucci. Evolution of phenotypic plasticity: where are we going now? *Trends in Ecology & Evolution*, 20(9):481–486, 9 2005.

[22] Andreas Wagner. The molecular origins of evolutionary innovations. *Trends in Genetics*, 27(10):397–410, 2011.

[23] Andreas Wagner and William Rosen. Spaces of the possible: universal Darwinism and the wall between technological and biological innovation. *Journal of the Royal Society, Interface / the Royal Society*, 11(97):20131190–, 2014.

[24] R C Lewontin. The Units of Selection. *Annual Review of Ecology and Systematics*, 1(1):1–18, 11 1970.

[25] J Arthur Harris. A New Theory of the Origin of Species. *The Open Court*, (4), 1904.

[26] L. B. Soros and Kenneth O. Stanley. Identifying Necessary Conditions for Open-Ended Evolution through the Artificial Life World of Chromaria. *Proc. of Artificial Life Conference (ALife 14)*, (Alife 14):793–800, 2014.

[27] Russel K. Standish. OPEN-ENDED ARTIFICIAL EVOLUTION. *International Journal of Computational Intelligence and Applications*, 03(02):167–175, 6 2003.

[28] Vera Vasas, Chrisantha Fernando, Andrs Szilágyi, Istvn Zachár, Mauro Santos, and Ers Szathmáry. Primordial evolvability: Impasses and challenges. *Journal of Theoretical Biology*, 381:29–38, 9 2015.

[29] Richard E. Lenski, Michael J. Wiser, Noah Ribeck, Zachary D. Blount, Joshua R. Nahum, J. Jeffrey Morris, Luis Zaman, Caroline B. Turner, Brian D. Wade, Rohan Maddamsetti, Alita R. Burmeister, Elizabeth J. Baird, Jay Bundy, Nkrumah A. Grant, Kyle J. Card, Maia Rowles, Kiyana Weatherspoon, Spiridon E. Papoulis, Rachel Sullivan, Colleen Clark, Joseph S. Mulka, and Neerja Hajela. Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with Escherichia coli. *Proceedings of the Royal Society B: Biological Sciences*, 282(1821):20152292, 12 2015.

[30] Joel Lehman and Kenneth O Stanley. Exploiting Open-Endedness to Solve Problems Through the Search for Novelty. *Artificial Life XI*, (Alife Xi):329–336, 2008.

[31] Alastair Channon and others. Improving and still passing the ALife test: Component-normalised activity statistics classify evolution in Geb as unbounded. *Proceedings of Artificial Life VIII, Sydney, RK Standish, MA Bedau, and HA Abbass,(eds.), MIT Press: Cambridge, MA*, pages 173–181, 2003.

[32] Alastair Channon. Unbounded evolutionary dynamics in a system of agents that actively process and transform their environment. *Genetic Programming and Evolvable Machines*, 7(3):253–281, 9 2006.

[33] Antoine Cully, Jeff Clune, Danesh Tarapore, and J.-B. Mouret. Robots that can adapt like animals. *Nature*, pages 1–26, 2015.

[34] Vera Vasas, Ers Szathmáry, and Mauro Santos. Lack of evolvability in self-sustaining autocatalytic networks constraints metabolism-first scenarios for the origin of life. *Proceedings of the National Academy of Sciences of the United States of America*, 107(4):1470–5, 1 2010.

[35] Omer Markovitch and Doron Lancet. Excess mutual catalysis is required for effective evolvability. *Artificial life*, 18(3):243–66, 1 2012.

[36] Leslie E Orgel. The implausibility of metabolic cycles on the prebiotic Earth. *PLoS biology*, 6(1):e18, 1 2008.

[37] Wim Hordijk. Autocatalytic Sets. *BioScience*, 63(11):877–881, 2013.

[38] A. M. Turing. I.Computing Machinery and Intelligence. *Mind*, LIX(236):433–460, 1950.

[39] Michael P Robertson and Gerald F Joyce. Highly efficient self-replicating RNA enzymes. *Chemistry & biology*, 21(2):238–45, 2 2014.

[40] G Brent Dalrymple. *The age of the Earth*. Stanford University Press, 1994.

[41] N. H. Sleep. The Hadean-Archaean Environment. *Cold Spring Harbor Perspectives in Biology*, 2(6):a002527–a002527, 6 2010.

[42] R. Gomes, H. F. Levison, K. Tsiganis, and A. Morbidelli. Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. *Nature*, 435(7041):466–469, 5 2005.

[43] Minik T. Rosing, Dennis K. Bird, Norman H. Sleep, and Christian J. Bjerrum. No climate paradox under the faint early Sun. *Nature*, 464(7289):744–747, 4 2010.

[44] S. J. Mojzsis, G. Arrhenius, K. D. McKeegan, T. M. Harrison, A. P. Nutman, and C. R. L. Friend. Evidence for life on Earth before 3,800 million years ago. *Nature*, 384(6604):55–59, 11 1996.

[45] Martin J. Whitehouse, Balz S. Kamber, Christopher M. Fedo, and Aivo Lepland. Integrated Pb- and S-isotope investigation of sulphide minerals from the early Archaean of southwest Greenland. *Chemical Geology*, 222(1-2):112–131, 10 2005.

[46] Yoko Ohtomo, Takeshi Kakegawa, Akizumi Ishida, Toshiro Nagase, and Minik T. Rosing. Evidence for biogenic graphite in early Archaean Isua metasedimentary rocks. *Nature Geoscience*, 7(1):25–28, 12 2013.

[47] Keih Kvenvolden, James Lawless, Katherine Pering, Etta Peterson, Jose Flores, Cyril Ponnamperuma, I. R. Kaplan, and Carleton Moore. Evidence for Extraterrestrial Amino-acids and Hydrocarbons in the Murchison Meteorite. *Nature*, 228(5275):923–926, 12 1970.

[48] P. Schmitt-Kopplin, Z. Gabelica, R. D. Gougeon, A. Fekete, B. Kanawati, M. Harir, I. Gebefuegi, G. Eckel, and N. Hertkorn. High molecular diversity of extraterrestrial organic matter in Murchison meteorite revealed 40 years after its fall. *Proceedings of the National Academy of Sciences*, 107(7):2763–2768, 2 2010.

[49] Robert Shapiro. *Origins: A skeptic's guide to the creation of life on earth*. Bantam Dell Pub Group, 1987.

[50] S. L. Miller. A Production of Amino Acids Under Possible Primitive Earth Conditions. *Science*, 117(3046):528–529, 5 1953.

[51] Stanley L Miller. The endogenous synthesis of organic compounds. *The molecular origins of life: assembling pieces of the puzzle*, pages 59–85, 1998.

[52] Doriano Brogioli. Marginally Stable Chemical Systems as Precursors of Life. *Physical Review Letters*, 105(5):058102, 7 2010.

[53] Pier Luigi Luisi and Francisco J. Varela. Self-replicating micelles A chemical version of a minimal autopoietic system. *Origins of Life and Evolution of the Biosphere*, 19(6):633–643, 11 1989.

[54] I. A. Chen. GE PRIZE-WINNING ESSAY: The Emergence of Cells During the Origin of Life. *Science*, 314(5805):1558–1559, 12 2006.

[55] J P Schrum, T F Zhu, and J W Szostak. The origins of cellular life. *Cold Spring Harb Perspect Biol*, 2(9):a002212, 2010.

[56] I. A. Chen and P. Walde. From Self-Assembled Vesicles to Protocells. *Cold Spring Harbor Perspectives in Biology*, 2(7):a002170–a002170, 7 2010.

[57] Martin M Hanczyc, Shelly M Fujikawa, and Jack W Szostak. Experimental models of primitive cellular compartments: encapsulation, growth, and division. *Science*, 302(5645):618–622, 2003.

[58] Rosalind E Franklin and Raymond G Gosling. Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. *Nature*, 172:156–157, 1953.

[59] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, 4 1953.

[60] Francis Harry Compton Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, XII:139–163, 1956.

[61] Hubert P. Yockey. An application of information theory to the central dogma and the sequence hypothesis. *Journal of Theoretical Biology*, 46(2):369–406, 8 1974.

[62] H P Yockey. Origin of life on earth and Shannon's theory of communication. *Computers & chemistry*, 24(1):105–23, 1 2000.

[63] David L Abel and Jack T Trevors. Three subsets of sequence complexity and their relevance to biopolymeric information. *Theoretical Biology and Medical Modelling*, 2(1):29, 2005.

[64] David L. Nelson and Michael M. Cox. *Lehninger Principles of Biochemistry*. W. H. Freeman, 5th edition, 2008.

[65] Kelly Kruger, Paula J. Grabowski, Arthur J. Zaug, Julie Sands, Daniel E. Gottschling, and Thomas R. Cech. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*, 31(1):147–157, 11 1982.

[66] Cecilia Guerrier-Takada, Katheleen Gardiner, Terry Marsh, Norman Pace, Sidney Altman, and others. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3):849–857, 12 1983.

[67] Walter Gilbert. The RNA world, 1986.

[68] P. Nissen. The Structural Basis of Ribosome Activity in Peptide Bond Synthesis. *Science*, 289(5481):920–930, 2000.

[69] G A Prody, J T Bakos, J M Buzayan, I R Schneider, and G Bruening. Autolytic processing of dimeric plant virus satellite RNA. *Science (New York, N.Y.)*, 231(4745):1577–80, 3 1986.

[70] C J Hutchins, P D Rathjen, A C Forster, and R H Symons. Self-cleavage of plus and minus RNA transcripts of avocado sunblotch viroid. *Nucleic acids research*, 14(9):3627–40, 5 1986.

[71] A. Wochner, J. Attwater, A. Coulson, and P. Holliger. Ribozyme-Catalyzed Transcription of an Active Ribozyme. *Science*, 332(6026):209–212, 4 2011.

[72] C. Hammann, A. Luptak, J. Perreault, and M. de la Pena. The ubiquitous hammerhead ribozyme. *RNA*, 18(5):871–885, 5 2012.

[73] George E. Fox. Origin and evolution of the ribosome. *Cold Spring Harbor perspectives in biology*, 2(9):1–18, 2010.

[74] A. S. Petrov, C. R. Bernier, C. Hsiao, A. M. Norris, N. A. Kovacs, C. C. Waterbury, V. G. Stepanov, S. C. Harvey, G. E. Fox, R. M. Wartell, N. V. Hud, and L. D. Williams. Evolution of the ribosome at atomic resolution. *Proceedings of the National Academy of Sciences*, 111(28):10251–10256, 2014.

[75] Meredith Root-Bernstein and Robert Root-Bernstein. The ribosome as a missing link in the evolution of life. *Journal of Theoretical Biology*, 367:130–158, 2015.

[76] J P Ferris, A R Hill, R Liu, and Leslie E Orgel. Synthesis of long prebiotic oligomers on mineral surfaces. *Nature*, 381(6577):59–61, 5 1996.

[77] Matthew W Powner, Batrice Gerland, and John D Sutherland. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature*, 459(7244):239–42, 5 2009.

[78] Brian J. Cafferty and Nicholas V. Hud. Abiotic synthesis of RNA in water: A common goal of prebiotic chemistry and bottom-up synthetic biology. *Current Opinion in Chemical Biology*, 22:146–157, 2014.

[79] H. S. Zaher and P. J. Unrau. Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. *RNA*, 13(7):1017–1026, 7 2007.

[80] L. K. L. Cheng and P. J. Unrau. Closing the Circle: Replicating RNA with RNA. *Cold Spring Harbor Perspectives in Biology*, 2(10):a002204–a002204, 10 2010.

[81] Elizaveta A Guseva, Ronald N. Zuckermann, and Ken A. Dill. How did prebiotic polymers become informational foldamers? 4 2016.

[82] Manfred Eigen. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 58(10):465–523, 10 1971.

[83] Yollete V. Guillen Schlippe, Matthew C. T. Hartman, Kristopher Josephson, and Jack W. Szostak. In Vitro Selection of Highly Modified Cyclic Peptides That Act as Tight Binding Inhibitors. *Journal of the American Chemical Society*, 134(25):10469–10477, 6 2012.

[84] W. K. Johnston, P J Unrau, M S Lawrence, M E Glasner, and D P Bartel. RNA-Catalyzed RNA Polymerization: Accurate and General RNA-Templated Primer Extension. *Science*, 292(5520):1319–1325, 5 2001.

[85] NV Hud, BJ Cafferty, R Krishnamurthy, and LD Williams. The Origin of RNA and My Grandfather's Axe. *Chemistry & biology*, 2013.

[86] G.F. Joyce. Nonenzymatic Template-directed Synthesis of Informational Macromolecules. *Cold Spring Harbor Symposia on Quantitative Biology*, 52:41–51, 1 1987.

[87] A. E. Engelhart and N. V. Hud. Primitive Genetic Polymers. *Cold Spring Harbor Perspectives in Biology*, 2(12):a002196–a002196, 12 2010.

[88] William Martin, John Baross, Deborah Kelley, and Michael J Russell. Hydrothermal vents and the origin of life. *Nature reviews. Microbiology*, 6(11):805–814, 2008.

[89] Daniel Segré, Doron Lancet, Ora Kedem, and Yitzhak Pilpel. Graded autocatalysis replication domain (GARD): kinetic analysis of self-replication in mutually catalytic sets. *Origins of Life and Evolution of the Biosphere*, 28(4-6):501–514, 1998.

[90] Stuart A Kauffman. Autocatalytic sets of proteins. *Journal of Theoretical Biology*, 119(1):1–24, 3 1986.

[91] Dirk Sievers and Gnter von Kiedrowski. Self-Replication of Hexadeoxynucleotide Analogues: Autocatalysis versus Cross-Catalysis. *Chemistry - A European Journal*, 4(4):629–641, 4 1998.

[92] Meng Wu and Paul G. Higgs. Comparison of the Roles of Nucleotide Synthesis, Polymerization, and Recombination in the Origin of Autocatalytic Sets of RNAs. *Astrobiology*, 11(9):895–906, 2011.

[93] Alessandro Filisetti, Alex Graudenzi, Roberto Serra, Marco Villani, Davide De Lucrezia, Rudolf M Füchslin, Stuart a Kauffman, Norman Packard, and Irene Poli. A stochastic model of the emergence of autocatalytic cycles. *Journal of Systems Chemistry*, 2(1):2, 2011.

[94] Wim Hordijk, Mike Steel, and Stuart Kauffman. The Structure of Autocatalytic Sets: Evolvability, Enablement, and Emergence. *Acta Biotheoretica*, 60(4):379–392, 2012.

[95] Wim Hordijk and Mike Steel. A formal model of autocatalytic sets emerging in an RNA replicator system. *Journal of Systems Chemistry*, 4(1):3, 2013.

[96] Wim Hordijk and Mike Steel. Autocatalytic sets extended: Dynamics, inhibition, and a generalization. *Journal of Systems Chemistry*, 3(1):5, 2012.

[97] Anton V. Bernatskiy and Elizaveta A. Guseva. Exact rule-based stochastic simulations for the system with unlimited number of molecular species. 9 2016.

[98] Jack W Szostak and A D Ellington. In Vitro Selection of Functional Nucleic Acids. In *The RNA World*, pages 511–533. Cold Spring Harbor Laboratory Press, 1993.

[99] Everett L Shock. Stability of peptides in high-temperature aqueous solutions, 1992.

[100] R Bruce Martin. Free Energies and Equilibria of Peptide Bond Hydrolysis. *Biopolymers*, 45:351–353, 1998.

[101] M. Paecht-Horowitz, J. Berger, and A. Katchalsky. Prebiotic Synthesis of Polypeptides by Heterogeneous Polycondensation of Amino-acid Adenylates. *Nature*, 228(5272):636–639, 11 1970.

[102] Luke Leman, Leslie E Orgel, and M Reza Ghadiri. Carbonyl sulfide-mediated prebiotic formation of peptides. *Science (New York, N.Y.)*, 306(5694):283–6, 10 2004.

[103] Leslie E Orgel. Prebiotic chemistry and the origin of the RNA world. *Critical reviews in biochemistry and molecular biology*, 39(2):99–123, 2004.

[104] M. Rao, D. G. Odom, and J. Oró. Clays in prebiological chemistry. *Journal of Molecular Evolution*, 15(4):317–331, 12 1980.

[105] Jean-Franois Lambert. Adsorption and polymerization of amino acids on mineral surfaces: a review. *Origins of life and evolution of the biosphere : the journal of the International Society for the Study of the Origin of Life*, 38(3):211–42, 6 2008.

[106] J D Bernal. The Physical Basis of Life. *Proceedings of the Physical Society. Section B*, 62(10):597–618, 10 1949.

[107] Kevin E. Nelson, Michael P. Robertson, Matthew Levy, and Stanley L. Miller. Concentration by Evaporation and the Prebiotic Synthesis of Cytosine. *Origins of Life and Evolution of the Biosphere*, 31(3):221–229, 2001.

[108] Anastassia Kanavarioti, Pierre-Alain Monnard, and David W. Deamer. Eutectic Phases in Ice Facilitate Nonenzymatic Nucleic Acid Synthesis. *Astrobiology*, 1(3):271–281, 9 2001.

[109] Jeffrey L. Bada. How life began on Earth: a status report. *Earth and Planetary Science Letters*, 226(1-2):1–15, 9 2004.

[110] Bernd M Rode, Hoang L Son, Yuttana Suwannachot, and Juraj Bujdak. the Combination of Salt Induced Peptide Formation. (Ii):273–286, 1997.

[111] Bernd Michael Rode. Peptides and the origin of life. *Peptides*, 20(6):773–786, 1999.

[112] Jay G Forsythe, Sheng-Sheng S Yu, Irena Mamajanov, Martha A Grover, Ramanarayanan Krishnamurthy, F M Fernandez, Nicholas V Hud, Facundo M Fern??ndez, and Nicholas V Hud. Ester-Mediated Amide Bond Formation Driven by Wet-Dry Cycles: A Possible Path to Polypeptides on the Prebiotic Earth. *Angew Chem Int Ed Engl*, 10:n/a???n/a, 2015.

[113] Paul J Flory. *Principles of polymer chemistry*. Ithaca, NY : Cornell Univ., 1953.

[114] Roscoe Stribling and Stanley L. Miller. Energy yields for hydrogen cyanide and formaldehyde syntheses: The hcn and amino acid concentrations in the primitive ocean. *Origins of Life and Evolution of the Biosphere*, 17(3-4):261–273, 9 1987.

[115] C Huber and G Wächtershäuser. Peptides by activation of amino acids with CO on (Ni,Fe)S surfaces: implications for the origin of life. *Science (New York, N.Y.)*, 281(5377):670–672, 1998.

[116] A. D. Aubrey, H. J. Cleaves, and Jeffrey L. Bada. The role of submarine hydrothermal systems in the synthesis of amino acids. *Origins of Life and Evolution of Biospheres*, 39(2):91–108, 2009.

[117] Antonio Lazcano and Stanley L Miller. The Origin and Early Evolution of Life: Prebiotic Chemistry, the Pre-RNA World, and Time. *Cell*, 85(6):793–798, 6 1996.

[118] Martin A Nowak and Hisashi Ohtsuki. Prevolutionary dynamics and the origin of evolution. *Proceedings of the National Academy of Sciences*, 105(39):14924–14927, 9 2008.

[119] Julien Derr, Michael L Manapat, Sudha Rajamani, Kevin Leu, Ramon Xulvi-Brunet, Isaac Joseph, Martin A Nowak, and Irene A Chen.

Prebiotically plausible mechanisms increase compositional diversity of nucleic acid sequences. *Nucleic acids research*, 40(10):4711–22, 5 2012.

[120] Ping Z. Ding, Kunio Kawamura, and James P. Ferris. Oligomerization of uridine phosphorimidazolides on montmorillonite: A model for the prebiotic synthesis of rna on minerals. *Origins of Life and Evolution of the Biosphere*, 26(2):151–171, 4 1996.

[121] James P. Ferris. Prebiotic Synthesis on Minerals: Bridging the Prebiotic and RNA Worlds. *Biological Bulletin*, 196(3):311, 6 1999.

[122] Samuel H Gellman. Foldamers: a manifesto. *Accounts of Chemical Research*, 31(4):173–180, 1998.

[123] Byoung-Chul Lee, Ronald N Zuckermann, and Ken A Dill. Folding a Nonbiological Polymer into a Compact Multihelical Structure. *J. Am. Chem. Soc.*, 127:10999–11009, 2005.

[124] E. Capriotti and M. A. Marti-Renom. RNA structure alignment by a unit-vector approach. *Bioinformatics*, 24(16):i112–i118, 8 2008.

[125] Hue Sun Chan and Ken A Dill. Sequence space soup of proteins and copolymers. *The Journal of Chemical Physics*, 95(5):3775, 1991.

[126] Michael a. Fisher, Kara L. McKinley, Luke H. Bradley, Sara R. Viola, and Michael H Hecht. De Novo Designed Proteins from a Library of Artificial Sequences Function in Escherichia Coli and Enable Cell Growth. *PLoS ONE*, 6(1):e15364, 2011.

[127] Izhack Cherny, Maria Korolev, Angela N Koehler, and Michael H Hecht. Proteins from an unevolved library of de novo designed sequences bind a range of small molecules. *ACS synthetic biology*, 1(4):130–8, 4 2012.

[128] Kit Fun Lau and Ken A Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986–3997, 10 1989.

[129] D W Miller and Ken A Dill. A statistical mechanical model for hydrogen exchange in globular proteins. *Protein science : a publication of the Protein Society*, 4(9):1860–73, 10 1995.

[130] K Yue and Ken A Dill. Forces of tertiary structural organization in globular proteins. *Proc Natl Acad Sci U S A*, 92(1):146–150, 1995.

[131] Richa Agarwala, Serafim Batzoglou, Vlado Dancik, Scott E. Decatur, Sridhar Hannenhalli, Martin Farach, S. MUTHUKRISHNAN, and STEVEN SKIENA. Local Rules for Protein Folding on a Triangular Lattice and Generalized Hydrophobicity in the HP Model. *Journal of Computational Biology*, 4(3):275–296, 1 1997.

[132] K Yue and K a Dill. Inverse protein folding problem: designing polymer sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 89(9):4163–4167, 1992.

[133] H Xiong, B L Buckwalter, H M Shieh, and Michael H Hecht. Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proceedings of the National Academy of Sciences of the United States of America*, 92(14):6349–53, 7 1995.

[134] Wendell A. Lim and Robert T. Sauer. The role of internal packing interactions in determining the structure and stability of a protein. *Journal of Molecular Biology*, 219(2):359–376, 5 1991.

[135] S Kamtekar, J M Schiffer, H Xiong, J M Babik, and M H Hecht. Protein design by binary patterning of polar and nonpolar amino acids. *Science (New York, N.Y.)*, 262(5140):1680–1685, 1993.

[136] Y. Wei. Stably folded de novo proteins from a designed combinatorial library. *Protein Science*, 12(1):92–102, 1 2003.

[137] Joseph M. Brisendine and Ronald L. Koder. Fast, cheap and out of control – Insights into thermodynamic and informatic constraints on natural protein sequences from de novo protein design. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 10 2015.

[138] V I Abkevich, A M Gutin, and E I Shakhnovich. How the first biopolymers could have evolved. *Proceedings of the National Academy of Sciences*, 93(2):839–844, 1996.

[139] Philip Lijnzaad, H. J C Berendsen, and Patrick Argos. Hydrophobic patches on the surfaces of protein structures. *Proteins: Structure, Function and Genetics*, 25(3):389–397, 1996.

[140] Sam Tonddast-Navaei and Jeffrey Skolnick. Are protein-protein interfaces special regions on a protein's surface? *Journal of Chemical Physics*, 143(24), 2015.

[141] J. Mitchell Guss and Hans C. Freeman. Structure of oxidized poplar plastocyanin at 1·6 Å resolution. *Journal of Molecular Biology*, 169(2):521–563, 9 1983.

[142] Jan H van Ee and Misset Onno. *Enzymes in Detergency*. CRC Press, 1997.

[143] H. Witt. Tryptophan 121 of Subunit II Is the Electron Entry Site to Cytochrome-c Oxidase in Paracoccus denitrificans. Involvement of a Hydrophobic Patch in the Docking Reaction. *Journal of Biological Chemistry*, 273(9):5132–5136, 2 1998.

[144] K. Manabe, X. M. Sun, and S. Kobayashi. Dehydration reactions in water. Surfactant-type Brønsted acid-catalyzed direct esterification of carboxylic acids with alcohols in an emulsion system [2]. *Journal of the American Chemical Society*, 123(41):10101–10102, 2001.

[145] Kei Manabe, Shinya Iimura, Xiang Min Sun, and Shu Kobayashi. Dehydration reactions in water. Brønsted acid-surfactant-combined catalyst for ester, ether, thioether, and dithioacetal formation in water. *Journal of the American Chemical Society*, 124(40):11971–11978, 2002.

[146] Torsten Stachelhaus, Henning D. Mootz, Veit Bergendahl, and Mohamed A. Marahiel. Peptide Bond Formation in Nonribosomal Peptide Biosynthesis. *Journal of Biological Chemistry*, 273(35):22773–22781, 8 1998.

[147] Mohamed A Marahiel. Working outside the protein-synthesis rules: insights into non-ribosomal peptide synthesis. *Journal of peptide science : an official publication of the European Peptide Society*, 15(12):799–807, 12 2009.

[148] Ken A. Dill. Polymer principles and protein folding. *Protein Science*, 8(6):1166–1180, 1999.

[149] Gilberto Giugliarelli, Cristian Micheletti, Jayanth R. Banavar, and Amos Maritan. Compactness, aggregation, and prionlike behavior of protein: a lattice model study. *Journal of Chemical Physics*, 113(12):5072–5077, 2000.

[150] Kingshuk Ghosh and Ken A Dill. Computing protein stabilities from their chain lengths. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26):10649–54, 6 2009.

[151] A. Sievers, M. Beringer, M. V. Rodnina, and R. Wolfenden. The ribosome as an entropy trap. *Proceedings of the National Academy of Sciences*, 101(21):7897–7901, 5 2004.

[152] Bonnie Berger and Tom Leighton. Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-Complete. *Journal of Computational Biology*, 5(1):27–40, 1 1998.

[153] Ryan R. Julian and J. L. Beauchamp. Abiotic synthesis of ATP from AMP in the gas phase: Implications for the origin of biologically important molecules from small molecular clusters. *International Journal of Mass Spectrometry*, 227(1):147–159, 2003.

[154] Rebecca A R Bryant and David E Hansen. Direct Measurement of the Uncatalyzed Rate of Hydrolysis of a Peptide Bond. *Journal of the American Chemical Society*, 118(23):5498–5499, 1 1996.

[155] Robert M. Smith and David E. Hansen. The pH-Rate Profile for the Hydrolysis of a Peptide Bond. *Journal of the American Chemical Society*, 120(35):8910–8913, 9 1998.

[156] Irene A. Chen and Jack W. Szostak. A Kinetic Study of the Growth of Fatty Acid Vesicles. *Biophysical Journal*, 87(2):988–998, 8 2004.

[157] Kingshuk Ghosh and Ken A Dill. Cellular proteomes have broad distributions of protein stability. *Biophysical Journal*, 99(12):3996–4002, 2010.

[158] Ken A Dill, Kingshuk Ghosh, and Jeremy D Schmit. Physical limits of cells and proteomes. *Proceedings of the National Academy of Sciences of the United States of America*, 108(44):17876–82, 11 2011.

[159] Freeman Dyson. *Origins of Life*. Cambridge: University Press, 1985.

[160] Manfred Eigen and Peter Schuster. The Hypercycle. *Naturwissenschaften*, 65(1):7–41, 1 1978.

[161] Dong-Eun Kim and Gerald F Joyce. Cross-catalytic replication of an RNA ligase ribozyme. *Chemistry & biology*, 11(11):1505–12, 11 2004.

[162] Grant S Murphy, Jack B Greisman, and Michael H Hecht. De Novo Proteins with Life-Sustaining Functions are Structurally Dynamic. *Journal of molecular biology*, 428(2):399–411, 2015.

[163] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 12 1977.

[164] Rajesh Ramaswamy, Nlido González-Segredo, and Ivo F Sbalzarini. A new class of highly efficient exact stochastic simulation algorithms for chemical reaction networks. *The Journal of chemical physics*, 130(24):244104, 6 2009.

[165] Darren James Wilkinson. *Stochastic Modelling for Systems Biology*. CRC Press, 2011.

[166] A Arkin, J Ross, and H H McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics*, 149(4):1633–48, 8 1998.

[167] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, 10 2008.

[168] Daniel T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1-3):404–425, 9 1992.

[169] Tobias Reichenbach, Mauro Mobilia, and Erwin Frey. Mobility promotes and jeopardizes biodiversity in rockpaperscissors games. *Nature*, 448(7157):1046–1049, 8 2007.

[170] A J McKane and T J Newman. Predator-prey cycles from resonant amplification of demographic stochasticity. *Physical review letters*, 94(21):218102, 6 2005.

[171] Ulf Dieckmann and Richard Law. The dynamical theory of coevolution: a derivation from stochastic ecological processes. *Journal of Mathematical Biology*, 34(5-6):579–612, 5 1996.

[172] Jean-Franois Le Galliard, Rgis Ferrière, and Ulf Dieckmann. Adaptive evolution of social traits: origin, trajectories, and correlations of altruism and mobility. *The American naturalist*, 165(2):206–24, 2 2005.

[173] Giulio Caravagna, Alberto dOnofrio, Paolo Milazzo, and Roberto Barbuti. Tumour suppression by immune system through stochastic oscillations. *Journal of Theoretical Biology*, 265(3):336–345, 2010.

[174] J Legrand, R F Grais, P Y Boelle, A J Valleron, and A Flahault. Understanding the dynamics of Ebola epidemics. *Epidemiology and infection*, 135(4):610–21, 5 2007.

[175] Romulus Breban, John M. Drake, David E. Stallknecht, and Pejman Rohani. The Role of Environmental Transmission in Recurrent Avian Influenza Epidemics. *PLoS Computational Biology*, 5(4):e1000346, 4 2009.

[176] Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

[177] Maxim S Shkarayev, Ira B Schwartz, and Leah B Shaw. Recruitment dynamics in adaptive social networks. *Journal of physics. A, Mathematical and theoretical*, 46(24):245003, 2013.

[178] Giuseppe Carbone and Ilaria Giannoccaro. Model of human collective decision-making in complex environments. *The European Physical Journal B*, 88(12):339, 12 2015.

[179] Mauro Mobilia. Stochastic dynamics of the prisoner's dilemma with cooperation facilitators. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 86(1 Pt 1):011134, 7 2012.

[180] Dranreb Earl Juanico. Critical network effect induces business oscillations in multi-level marketing systems. 9 2012.

[181] Miquel Montero. Predator-Prey Model for Stock Market Fluctuations. *SSRN Electronic Journal*, 2008.

[182] S. Berman, A. Halasz, M.A. Hsieh, and V. Kumar. Optimized Stochastic Policies for Task Allocation in Swarms of Robots. *IEEE Transactions on Robotics*, 25(4):927–937, 8 2009.

[183] Spring Berman, Adam Halasz, Vijay Kumar, and Stephen Pratt. Bio-Inspired Group Behaviors for the Deployment of a Swarm of Robots to Multiple Destinations. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 2318–2323. IEEE, 4 2007.

[184] J. Hillston. Fluid flow approximation of PEPA models. In *Second International Conference on the Quantitative Evaluation of Systems (QEST'05)*, pages 33–42. IEEE, 2005.

[185] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 12 1976.

[186] Rajesh Ramaswamy and Ivo F. Sbalzarini. A partial-propensity variant of the composition-rejection stochastic simulation algorithm for chemical reaction networks. *Journal of Chemical Physics*, 132(4):1–6, 2010.

[187] Graeme Henkelman and Hannes Jonsson. Long time scale kinetic Monte Carlo simulations without lattice approximation and predefined event table. *The Journal of Chemical Physics*, 115(21):9657, 2001.

[188] Michael A. Gibson and Jehoshua Bruck. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *The Journal of Physical Chemistry A*, 104(9):1876–1889, 2000.

[189] Yang Cao, Hong Li, and Linda Petzold. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *The Journal of Chemical Physics*, 121(9):4059, 2004.

[190] James M. McCollum, Gregory D. Peterson, Chris D. Cox, Michael L. Simpson, and Nagiza F. Samatova. The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior. *Computational Biology and Chemistry*, 30(1):39–49, 2 2006.

[191] Daniel T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716, 2001.

[192] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. Avoiding negative populations in explicit Poisson tau-leaping. *The Journal of Chemical Physics*, 123(5):054104, 2005.

[193] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. Efficient step size selection for the tau-leaping simulation method. *The Journal of Chemical Physics*, 124(4):044109, 2006.

[194] Xinjun Peng, Wen Zhou, and Yifei Wang. Efficient binomial leap method for simulating chemical kinetics. *The Journal of Chemical Physics*, 126(22):224109, 2007.

[195] Muruhan Rathinam, Linda R. Petzold, Yang Cao, and Daniel T. Gillespie. Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. *The Journal of Chemical Physics*, 119(24):12784, 2003.

[196] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. The slow-scale stochastic simulation algorithm. *The Journal of Chemical Physics*, 122(1):014116, 2005.

[197] Anne Auger, Philippe Chatelain, and Petros Koumoutsakos. R-leaping: Accelerating the stochastic simulation algorithm by reaction leaps. *The Journal of Chemical Physics*, 125(8):084103, 2006.

[198] Xin-jun Peng and Yi-fei Wang. L-leap: accelerating the stochastic simulation of chemically reacting systems. *Applied Mathematics and Mechanics*, 28(10):1361–1371, 10 2007.

[199] Xiaodong Cai and Zhouyi Xu. K -leap method for accelerating stochastic simulation of coupled chemical reactions. *Journal of Chemical Physics*, 126(7):1–10, 2007.

[200] Joseph Leo Doob. *Stochastic Processes*. John Wiley & Sons, Inc.;Chapman & Hall, New York, New York, USA, 1953.

[201] Kevin R Sanft, Sheng Wu, Min Roh, Jin Fu, Rone Kwei Lim, and Linda R Petzold. StochKit2: software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics (Oxford, England)*, 27(17):2457–8, 9 2011.