

## UvA-DARE (Digital Academic Repository)

### Graph-theoretical approach to algorithmic construction of complex reaction networks

Orlova, Y.

**Publication date**

2020

**Document Version**

Other version

**License**

Other

[Link to publication](#)

**Citation for published version (APA):**

Orlova, Y. (2020). *Graph-theoretical approach to algorithmic construction of complex reaction networks*.

**General rights**

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

# 1

## Introduction

Observing real life phenomena as mirrored by computer generated representations becomes more and more indispensable next to and even in the context of experimental investigations in modern chemical research. Software packages and numerical simulators allow systematic and precise ways of studying the behaviour of chemical systems. Time and length scales of the underlying models range from atomic to macroscopic scales, from quantum states of atoms to the complex topology of polymer networks. These tools and methods require representations of chemical entities - atoms, molecules, particles, *etc.* - that serve as a language into which these entities have to be translated in order to become palatable for a computer. Subsequently, these objects can be subjected to the laws of mathematics to describe chemical interactions and predict the outcomes of virtual experiments. Such algorithmic solutions are intended to improve innovative power to resolve new issues that could not be resolved before by human activity.

One instance of human limitation, where a computer algorithm can make the difference, is the extensive combinatorics involved in the number of all choices to make when inventing a new chemical synthesis route<sup>1-3</sup>. Of course, making such choices has always been at the heart of chemistry and has produced many successful syntheses. However, to identify the underlying structure of the decision process towards a new synthesis route has only recently become a topic of interest. One of the important notions here is that, although not always being aware of it, chemists designing a new route work according to a plan. In order for a computer to take over elements of human chemical

planning and generate a synthesis route, one needs to formalise such a planning process. The corresponding chemical discipline designing such a computer aided activity is called 'synthesis planning'<sup>4,5</sup>.

Just as all computer representations of chemical entities, as remarked above, a chemist's plan has to be expressed in a certain language, which would consist of 'chemical words' that are connected to express meaning by using 'formal grammar' rules. Conveniently, this abstracting from specific domain knowledge to generic planning principles like rules proceeds in similar ways to many other domains than chemistry. The comparison can be made with chess playing, where the role of rules is evident. The number of rules in chess playing is in the order of ten, while for chemical synthesis it is in the order of hundred to thousand<sup>4,5</sup> – hence the challenge in the chemical area is considerably larger.

This thesis will not address chemical synthesis planning, but instead will be focused on predicting material properties. The synthesis planning example has been chosen as it represents a state-of-the-art example of the concepts that we will employ in a different domain. We develop automated procedures for the prediction of nanoscale topology of an infinite network formed by a natural polymer. We expect that our approach has promising features for other complex materials and processes as well.

## 1.1. Application area: natural polymers

The application area we have chosen for this thesis is the drying or polymerization of linseed oil, the unraveling of the principles of formation and degradation of this bio-based polymer<sup>6</sup>. Our findings are valuable for both improvement of the durability of newly produced material and design of preservation strategies for historic objects made with the use of bio-based polymers. The latter practice is highly relevant for the field of restoration and conservation of art objects, especially old oil paintings<sup>7,8</sup>, which are known to be chemically unstable objects<sup>9</sup>.

Little is known about the reaction networks behind synthesis and degradation of bio-based polymer materials, such as wood, cellulose, and many natural resins that became widely used centuries ago<sup>10,11</sup>. The term 'polymers' is applied to a variety of synthetic and natural materials, which are composed of an infinite number of interconnected molecules. Polymerization is a process where previously disconnected molecules, monomer units, start connecting to each other under the influence of external conditions, such as light, temperature, presence of other molecules, etc<sup>12,13</sup>. Atomic structure of the monomer units determines polymer's functions, which range from sustaining growth and development in the living organisms, as in DNA molecules<sup>14,15</sup>, to preventing metal

constructions from corrosion, as in paints and varnishes<sup>16,17</sup>. In synthetic polymers, the monomer units are simple and small molecules that during polymerization do not undergo other alterations than losing one or more double bonds<sup>18,19</sup>. Hence, polymer models have adopted the idea of 'repeating units' starting from a work by Flory<sup>20</sup>. Polymers form a special class of molecules, as, in some instances, they are infinite in size networks (in terms of the number of monomers that form a polymer)<sup>18</sup>, while, other polymers have finite size and resemble chains with or without branches<sup>21</sup>. In either case, the concept of repeating units is still valid. Currently available modeling studies of polymerization regard well-defined reaction mechanisms, in which both the repeated units and the complete set of reactions are known in advance. Efficient tools have been developed for such systems describing macromolecular properties in terms of sizes of chains, branches, functional groups, etc.<sup>22–26</sup>.

The classical polymer approach, however, is not valid in every respect for bio-based polymers: extracted from natural sources, they polymerize under certain conditions resulting in materials, which are characterized by rather complex chemistry<sup>11,27–29</sup>. Properties of such polymers have been discovered empirically, via trial and error, and the information regarding those properties has been partially preserved in old books and recipes<sup>10,30,31</sup>. 'Monomers' constituting natural polymer networks may undergo many side reactions, as in the case of 'drying oils', either increasing their weight via reactions with oxygen<sup>32</sup>) or breaking into smaller fragments by  $\beta$ -scission reactions<sup>33</sup>. Thus, the monomer units lose their identity and the idea of 'repeating units' no longer applies, as the original monomers are typically transformed into many thousands of different species. Studying such complex polymeric materials requires advanced modeling and analytical techniques, as a classical polymer reaction engineering approach is not applicable for such systems.

Fast formation of the insoluble fraction during the polymerization process leads to inability of experimental as well as modeling techniques to characterize both soluble and insoluble fractions of the material on the same level of detail. Experimental techniques such as nuclear magnetic resonance (NMR) and Fourier-transform infrared spectroscopy (FTIR) measure the amounts of functional groups present in the whole material. However, these measurements do not provide any spatial information about the distribution of these functional groups. Mass spectrometry (MS) and size exclusion chromatography (SEC) characterize the material in terms of weights and sizes of its fractions, which can further be attributed to particular molecular structures<sup>6,34,35</sup>. Although analytical techniques generate vast amounts of experimental data, interpretation of these data is not always a straightforward task. As in MS measurements, soluble fractions

in aged bio-based polymers are characterized by a wide range of measured masses that are challenging to assign to molecular structures manually<sup>34</sup>. Overlapping FTIR signals of esters, aldehydes and carboxylic acid make it challenging to isolate a signal corresponding to solely carboxylic acid, which in the presence of metal ions contributes to the formation of metal soaps - potentially dangerous degradation effects in oil paintings<sup>36</sup>. Because overcoming experimental challenges might be expensive and time consuming, we will take an alternative direction and improve computational modeling techniques to handle complex polymerization processes.

Varying in complexity and precision level, existing models cover different aspects of the formation of natural polymers. Quantum chemical calculations explore reactions at high resolution: which transformations happen to a structure of a molecule, and which reaction pathway is the most energetically feasible. Such modeling has been shown to elucidate free radical reactions, which happen too fast to be tracked experimentally<sup>33,37-39</sup>. Molecular dynamics simulations are used to describe atoms' positions and trajectories for such tasks as assessing molecular flexibility, exploring conformational changes, or gaining information about molecular interactions. As for modeling polymeric structures, such studies are mainly performed on exploring conformations of known structures of proteins<sup>40</sup> or formation of synthetic polymers with small repeating monomer units<sup>41</sup>. Due to the size limitations and complex reaction schemes, modeling of the formation of natural polymers and accessing their properties with molecular dynamics is restricted. Therefore, a more coarse-grained level of modeling as compared to molecular dynamics, is employed: kinetic modeling. A kinetic model is a system of nonlinear, ordinary differential equations (ODEs) describing balances of the amounts of all species present in the system over time. Solving the kinetic model for a time interval provides the concentration profiles of all species over time<sup>42</sup>.

Having introduced modeling and experimental challenges regarding the formation of natural polymers, we proceed with the introduction to concepts and tools, which are going to be used in this thesis for modeling and interpretation of experimental measurements for complex polymers.

## 1.2. Grammar and chemistry

The resemblance of chemistry to language – chemical words to real words – has been extensively discussed in a series of papers by the group of Grzybowski<sup>43-45</sup>. As it has been observed by Jean-Marie Lehn, atoms resemble letters while molecules are like words made of letters<sup>46</sup>. Later, we will denote this principle of organization 'morphology' - the set of rules that describes how to organize letters into words, with its chemical

equivalent: how to organize atoms into molecules and molecular substructures. Cadeddu *et al.* statistically analyze molecules from large databases as well as words from books in the English language<sup>43</sup>. They formulate ‘maximum common substructures’ (MCS) and analyze MCSs in both molecular structures and natural language. The MCSs are comparable to the parts of the molecules that determine the molecules’ function or reactivity, the functional groups. The frequency of MCS in language and in molecular databases follow the same distribution implying that the analysis in natural language corpora is applicable to chemical databases<sup>45</sup>. In this work, we depart from the same hypothesis based on the similarity of language and chemistry and proceed with exploring structure of language, the concept of grammar and how it is transferred to the rules of chemistry.

We will use the term grammar as the overarching term describing sets of rules on various hierarchical levels, according to which we construct words and meaningful sentences. The concept of a hierarchical structure is borrowed from the linguistic theory by Chomsky<sup>47</sup>. Grammar comprises morphology and syntax, where morphology rules govern the structure of the words composed of letters, and syntax determines how words are changed of shape and position in sentences. Chemical grammar employs the resemblance between language and chemistry. Thus, morphology corresponds to the rules of chemical bonding according to which atoms are connected in a particular pattern to form functional groups. Chemical syntax consists of the rules according to which functional groups change by undergoing reactions and form molecules. When considering polymers, one might formulate an extra hierarchical level where molecules may arrange themselves into polymers according to the rules of connecting monomers. To complete the analogy with language one might associate the highest level of building up a polymer using a given set of monomer structures with writing a grammatically correct text using a given lexis.

Projecting this reasoning to the problem of the formation of complex polymer, we likewise split our challenging modeling task into smaller steps applied to the different hierarchical levels of a polymer. The first task is to formulate reaction rules in terms of functional groups of the molecules, or, in other words, to formulate the chemical syntax. Like a chemical reaction from a textbook, a reaction rule is a relation specifying the change of a reactant into a product functional group, which comes on the place of the reactive site. The reaction rule applies when two molecules with the required reactive sites are found in a chemical system. Enumeration of such reaction rules can be done manually, as we did in our application, for smaller systems as well as using machine learning methods for extracting relations from chemical databases for bigger systems<sup>3</sup>.

The task we have set ourselves is to find the connectivity patterns of the many different monomer species that resemble the nodes of an infinite polymer network with a resolution down to the atomic level. Hence, for our modeling approach we suggest the representation of the monomer units imposing a particular manner in which they are connected to other monomers. The connections that form the 'boundaries' of the monomer units are represented by 'half edges'<sup>48</sup>. These half-edges denote the connections to neighboring monomers without being explicitly connected. Thus, monomers are characterized by the type and number of connections, or crosslinks, they can gain while traveling along polymerization reaction pathways. Using this concept, the 'monomer approach', we are able to reconstruct the great variety of possible monomer units according to the given reaction rules, or chemical grammar. The monomer approach allows to fully define the polymer, thus avoiding the need to explicitly describe excessively large sizes of the molecules and the 'explosion' of monomer combinations.

The monomer approach unavoidably gives rise to the question, from the perspective of a monomer species captured within its boundaries, how the world looks like beyond the boundaries. As reconstruction and analysis of all possible polymers is computationally infeasible, we need to employ a probabilistic method, which will provide us average properties of an ensemble of polymer networks characteristic to a given set of their monomer constituents. This is where the innovation in the form of random graphs enters allowing the inference of the topology of the (infinite) polymer network on the basis of probability distributions of the number and type of links present on monomer units<sup>49-51</sup>. This modeling approach has been successfully applied to a chemically less complex system of multi-functional acrylate polymerization<sup>18</sup>.

Viewing the problem of polymerization from the perspective of chemical grammar, we employ the hierarchical approach to model infinite size polymers as a finite set of their monomer constituents. Random graphs provide the set of rules to form a polymer network from its monomer constituents carrying half edges. Moreover, it offers a manner to infer average properties of the polymer network material. In the following subsection we proceed with the overview of tools that can be used to implement this modeling concept as a computer algorithm.

### 1.3. Graph theory and chemistry

The resemblance between grammar and chemistry allows us to split the problem of modeling material properties of a complex polymer in two main steps: 1) apply chemical syntax to functional groups of monomer units; 2) infer polymer properties from the probability distributions of these monomer units. The reconstruction of all structures

of monomers with more than one functional group is a labour intensive task. This task is similar to reconstructing all the sentences given a specific set of words. However, knowing the set of rules, one may reconstruct the desired set of monomer structures in an automated manner by programming the chemical rules in a computer. This program would enumerate all possible combinations of functional groups appearing in the molecules under construction while following the reaction rules, thus solving the huge combinatorial problem.

The core of the algorithm we choose employs the resemblance between a molecular structure and a graph as a mathematical object. Molecules have previously been described as labeled graphs with nodes corresponding to atoms and edges between nodes corresponding to bonds between atoms<sup>52,53</sup>. Such a representation allows to translate a molecule into a computer-readable object and yields access to using the latest developments in the field of graph theory on molecular data. The molecular graph is defined by two key elements: 1) an adjacency matrix, which captures the connectivity and bond order between atoms; 2) a list of node labels corresponding to chemical elements of each atom. This representation enables utilizing graph isomorphism tests, which identifies isomeric species or particular substructures responsible for the reactivity of a given molecule<sup>54</sup>.

Thus, the molecular graph representation and graph theory constitutes the machinery to algorithmically reformulate chemistry as a set of rules acting on the reactive parts of molecules and transforming them into unstable intermediate and stable product species. By repeatedly applying chemical rules to molecular graphs, we obtain, in a systematic manner, a complete reaction scheme involving manually intractable numbers of molecular species (in the order of thousand and higher) and all reactions between them. This procedure is called 'automated reaction network generation'.

The automated reaction network generation algorithm starts with an initial set of unchanged molecules represented as molecular graphs and a predefined list of reaction rules, the chemical syntax. A reaction rule, also called 'reaction template', is a rule according to which a reactant functional group of a molecule is transformed into a product functional group<sup>55</sup>. The first step of the algorithm is to recognize a reactant functional group on the initial molecule using a subgraph isomorphism test<sup>54</sup> and then apply a transformation dictated by one or more chemical reactions known to occur at the recognized reactant site. According to the reaction rule, a different molecular substructure, that corresponds to the product functional group, then replaces the original substructure of the reactant. Such a transformation creates a new molecule, which may undergo further transformations. The algorithm stops when no new structures

are produced according to the predefined reaction rules. Once the recovery of new molecular species is completed, the algorithm constructs a reaction network in the form of a directed bipartite graph that captures all molecular interactions<sup>56</sup>.

Numerous software packages, which are based on automated reaction network generation, have been developed since the pioneering works by Yoneda<sup>57</sup> and Clymans *et al.*<sup>58</sup>. The automated discovery of reaction networks became widely used by other researchers who adapted the idea to model various chemical processes<sup>59–64</sup>. An extensive overview and comparison of the automated reaction network generation techniques is provided in the works by Vernuccio and Broadbelt<sup>65</sup> and Simm<sup>66</sup>.

Automatic reaction network generation enables modeling of complex chemical processes, such as: metabolic processes<sup>67,68</sup>, prebiotic scenarios<sup>69</sup>, pyrolysis<sup>70</sup>, protein interactions<sup>71–74</sup> and regulation of cell functions<sup>75</sup>. These processes have in common that they are dealing with a large number of molecules undergoing complex reaction pathways. In our application, we are confronted with an additional issue, polymerization, which has never been addressed before in the context of automatic reaction network generation. The molecular graph representation is computationally intensive as it requires large random-access memory (RAM), so the automated reaction network generation becomes prohibitively slow when molecular graphs grow in size. By considering polymers as hierarchical structures, as discussed in the section above, we limit the size of the molecules to the size of monomers while preserving their connectivity patterns. This monomer approach is compatible with the concept of automated reaction network generation methodology, and hence we could implement this monomer approach in the molecular graph representation. Instead of joining molecular graphs and forming an even larger graph, we introduce an extra index in atom labels, which is responsible for indicating the presence of a crosslink on a given atom. If two atoms take part in a connection that links two monomers forming a dimer, then these atoms carry an extra label, which indicates the type of the crosslink. Hence, the monomers do not change in size significantly, and they remain not explicitly connected. Implementing this concept in the automated reaction network generation solves the problem of ever growing polymers and allows enumerating the structures of all unique monomer units that constitute the polymer network. Returning to the analogy with language, we have thus introduced an algorithm that constructs all possible sentences on the basis of predefined lexis and syntax. The resulting reaction network is further transformed into a kinetic model, which describes the concentration profiles, or probability distributions, of all the reconstructed monomer constituents of the polymer.

Inferring polymer properties can then be obtained using random graphs<sup>49</sup>. This the-

ory offers a potential to extract properties of infinitely large irregular polymer structures from the probability distribution describing the concentrations of crosslinked monomers. This distribution is also known as a degree distribution, where a degree is a number of connections, or crosslinks, per monomer. Using kinetic modeling followed by the application of random graph modeling allows obtaining the characteristics of complex natural polymers like glass transition temperature, gel point, polymer size distributions (radius of gyration) of soluble fractions, etc. <sup>18,41</sup>.

Note, that the graph (or network) representation appears more than once in this grammar-based methodology. One should distinguish three instances of graphs in our modeling approach:

- **Molecular graph** represents molecular structure, where nodes are atoms connected with edges, which are bonds. On account of the monomer approach, molecular graphs represent monomer constituents of the polymer network with their adjacent connections. In the analogy with language, these molecular graphs represent the sentences that form a text.
- **Reaction network** is a directed bipartite graph with two types of nodes: species and reactions. It is a network of states assumed by monomers while following along reaction pathways. In this network, an edge points from a reactant species to a reaction in which the species takes part, and from a reaction to a product species, which is formed due to this reaction. Here species are monomer constituents of the polymer. The reaction network captures the whole reaction scheme as interactions between nodes corresponding to monomers via nodes corresponding to reactions.
- **Polymer network** is represented as a random graph in terms of the finite ensemble of the monomers with particular connectivity patterns. This representation corresponds to all possible networks that can be formed with given degree distribution, or, in terms of the language, all possible texts that can be written using sentences built with given lexis.

To summarize, based on the similarity between language and chemistry, we employ a hierarchical approach to view complex infinite size polymers and propose a modeling methodology for obtaining polymer properties. The tools required to perform this modeling task are: 1) automated reaction network generation modified with the monomer approach used to uncover all monomer constituents of natural polymers; 2) kinetic modeling of the monomers used to obtain their degree distribution governed by the chemistry formulated as chemical grammar; 3) random graph modeling, which uses the degree distribution to infer polymer properties, such as gel point, gel fraction,

sizes of soluble fractions, polymer size distributions (radius of gyration), glass transition temperature, etc.

## 1.4. Case study

Automated procedures for reaction network generation have previously addressed metabolic processes, prebiotic scenarios, protein interactions, etc. In this dissertation we focus on drying and degradation of oil paint in the context of restoration and preservation of old artworks. Oil paint has been widely used by artists since the 15th century<sup>76</sup> and became the key polymer in modern paint and varnish industry<sup>77</sup>. Oil paint consists of binding medium, pigments and additives. When exposed to air and light, binding medium undergoes a polymerization process, which causes the phase transition of paint from its liquid state to a finger-dry film<sup>28,34</sup> (see Figure 1.1). Linseed oil became one of the most common binding media for oil paint due to its optimal rate of drying, which is not too fast and is suitable for an artist to work on paintings for more than one day.

Polymerization of linseed oil is a rather complex process, as it is governed by fast free-radical oxidation, in which numerous reactions happen simultaneously. Oil paintings are not chemically stable objects, as the polymerization phase is followed by the degradation due to slow decomposition of the polymer network causing the migration of low molecular weight molecules in the film<sup>6</sup>. These phenomena may result in various degradation effects, such as formation of metal soaps, loss of pigment, cracks on the surface of paint due to uneven accumulation of stress and strain<sup>78–81</sup>. Understanding the dynamics of polymerization of linseed oil is highly important for the analysis of the state of a painting and the improvement of preservation methods<sup>82,83</sup>.

Understanding chemical processes in linseed oil is also challenging from an experimental point of view due to the overlap of numerous chemical processes happening in a paint film. To isolate particular reaction pathways, scientists have studied model systems, which are simpler in structure than the full linseed oil molecule<sup>34,35,77,84</sup>. In our modeling methodology, we adopt a similar strategy and start with common model systems: triolein and ethyl linoleate (EL). Triolein is a triglyceride, which consists of glycerol connected to three oleic acids via ester bonds. EL is one of three fatty acids, oleic, linoleic and linolenic, in the structure of linseed oil. EL is characterized by unsaturations, which make its chemistry more complex than the oleic fatty ester but less complex than the linolenic fatty ester.

**Triolein** Triolein consists of a glycerol connected to three oleic acids via ester bonds. Studying this model system, we adapt our modeling methodology to work with triglyc-

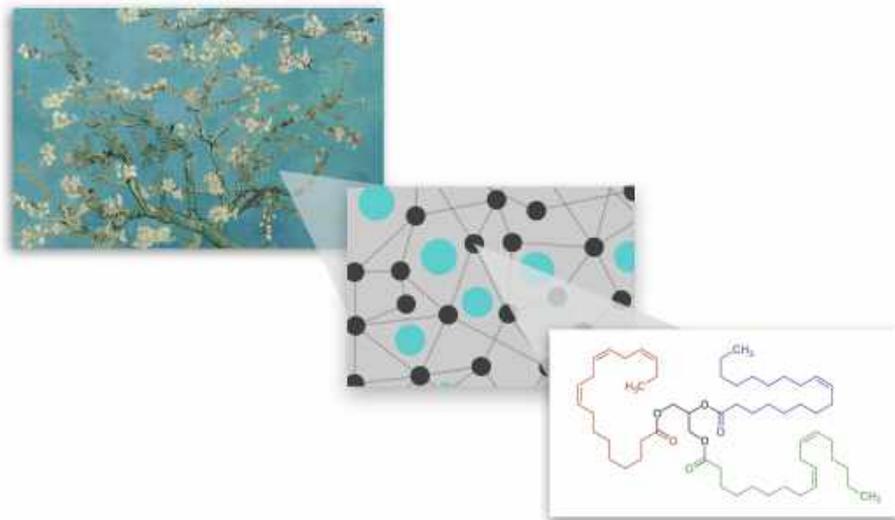


Figure 1.1: Schematic representation of linseed oil polymer network in the context of oil painting.

eride structure, such as in linseed oil. This molecule is a suitable model to study hydrolysis of ester bonds and oxidative scission of oleic esters.

**Ethyl linoleate** EL is a common model system for linseed oil. EL is a fatty ester with two unsaturations. EL has one bis-allylic carbon that is susceptible to fast hydrogen abstraction having a lower energy barrier than the allylic hydrogen present in oleic fatty ester. The amount of linoleic and linolenic fatty esters in triglycerides therefore influences the oil's ability to dry. EL may form up to three crosslinks per monomer of three types: alkyl, ether and peroxy, forming highly crosslinked polymer networks via a free radical polymerization process. EL has been studied in experimental as well as in computational modeling works<sup>85,86</sup>. While triolein allows isolating the chemistry around glycerol and  $\beta$ -scission of fatty esters, EL is a suitable model system for the complex autoxidation chemistry occurring at the linoleic fatty esters of linseed oil molecules.

## 1.5. Outline of the thesis

The structure of the thesis follows the steps we took to develop the methodology for modeling complex polymerization processes. We will first present the modeling methodology and then introduce the chemistry of linseed oil. This forms the basis, on which we further build the chemical syntax required for the application of the methodology to

model triolein and EL, as model systems of linseed oil.

Chapter 2 offers a thorough description of the automated reaction network generation (ARNG) methodology and its modification to include polymer structures. Polymer growth is addressed by implementing the monomer approach in the algorithm of automated reaction network generation. The grammar-based methodology is applied to model co-polymerization of inimer and isobutylene, which is a rather simple system chosen to demonstrate the ARNG. The chapter describes the monomer approach, the reconstruction of the reaction network, and its transformation into a dynamical system describing the concentration profiles of functional groups and crosslinks in the polymer over time. The result of such a kinetic model is further used as an input to the random graph model to predict the molecular weight distribution at the late stage of polymerization. These results are compared with the outcomes from kinetic Monte Carlo simulations.

Developing models employing the automated reaction network generation type of algorithms requires knowledge of chemical processes happening in the system. This knowledge is presented as a database of reactions that act on molecular graphs to formulate reaction networks. Therefore, Chapter 3 offers a comprehensive survey of linseed oil chemistry based on several older reviews, augmented by important experimental and modeling works of more recent date. This chapter serves as a reference for a chemical database, which describes reactions, or chemical syntax, involved in the polymerization of oils as well as our case studies, triolein and ethyl linoleate. Additionally, we discuss the experimental data that are used to validate computational models. Finally, we suggest reaction processes, which need further kinetic investigation.

In Chapter 4 we apply the automated reaction network generation to uncover the autoxidation reaction scheme of triolein. The methodology provides molecular graphs of triolein species, which are formed while following predefined autoxidation process. Knowing the structures of the molecules, we calculate their masses and compare them with the masses measured with electrospray ionization mass spectrometry (ESI-MS). MS spectra of aged triolein demonstrate a wide range of peaks. The diversity of measured masses challenges manual identification of the peaks. As an aid to identification, we relate masses of computationally generated molecular species to the measured ones. With this approach, we are able to identify the majority of high intensity peaks and gain better understanding of the chemical processes happening in the autoxidation of triolein.

In Chapter 5 we proceed with the case study of EL, and, using this example, we draw a parallel between chemistry and language. The analogy is based on Chomsky's idea of

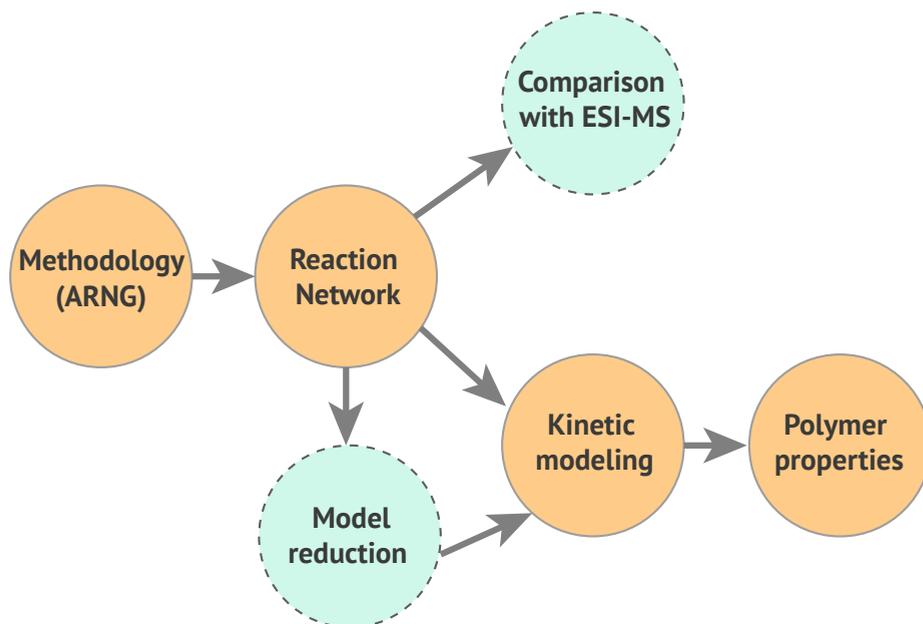


Figure 1.2: The components of our modeling methodology. Orange represents the sequence of the main components that lead to the polymer properties of the material, while light green represents auxiliary components that were discovered in the process of the development of the methodology.

formal grammar<sup>47</sup>. The chapter starts with the analogy between chemistry and language and proceeds with the case study of EL, which is modeled using the automated reaction network generation methodology. Firstly, we reconstruct a detailed reaction network in terms of unique monomer constituents of the polymer network of EL. As in Chapter 4, the molecular graphs of monomers and automatically reconstructed dimers are used to identify peaks in ESI MS spectra of EL. The reaction network is further translated into a kinetic model describing the evolution of the concentrations of all important functional groups that are formed during the drying of EL. Furthermore, the Random Graph model is used to obtain the average size distribution of soluble fractions during and after drying, formation of the gel fraction (the insoluble fraction) and the component size distribution on every step of the drying process. The results of the kinetic model and the random graph model are validated with findings from the experimental studies.

The scheme in Figure 1.2 shows our major findings and their inter-dependence. The core of the thesis is a novel modeling methodology, which generates a reaction network of infinitely large chemical systems. The reaction network, together with molecular graphs of all species, is then used to assist the peak identification problem in ESI-MS

measurements. The reaction network is then automatically transformed into the kinetic model of the system, which describes the evolution of the concentrations of molecular species. Then from the solution of the kinetic model we extract the connectivity profiles of individual monomers, which serve as an input for random graph theory to obtain polymer properties of the material.

## References

- [1] T. Li, L. Liu, N. Wei, J.-Y. Yang, D. G. Chapla, K. W. Moremen, G.-J. Boons, *et al.*, *An automated platform for the enzyme-mediated assembly of complex oligosaccharides*, *Nature chemistry* **11**, 229 (2019).
- [2] P. S. Gromski, A. B. Henson, J. M. Granda, and L. Cronin, *How to explore chemical space using algorithms and automation*, *Nature Reviews Chemistry* , 1 (2019).
- [3] W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin, and B. A. Grzybowski, *Automatic mapping of atoms across both simple and complex chemical reactions*, *Nature Communications* **10**, 1434 (2019).
- [4] S. Hanessian, J. Franco, and B. Larouche, *The psychobiological basis of heuristic synthesis planning-man, machine and the chiron approach*, *Pure & App. Chem.* **62**, 1887 (1990).
- [5] O. Ravitz, *Data-driven computer aided synthesis design*, *Drug Discovery Today: Technologies* **10**, e443 (2013).
- [6] M. Lazzari and O. Chiantore, *Drying and oxidative degradation of linseed oil*, *Polymer Degradation and Stability* **65**, 303 (1999).
- [7] T. L. Dawson, *Examination, conservation and restoration of painted art*, *Coloration Technology* **123**, 281 (2007), <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1478-4408.2007.00096.x> .
- [8] I. D. van der Werf, K. J. van den Berg, S. Schmitt, and J. J. Boon, *Molecular characterization of copaiba balsam as used in painting techniques and restoration procedures*, *Studies in Conservation* **45**, 1 (2000), <https://doi.org/10.1179/sic.2000.45.1.1> .
- [9] J. Mallegol, J.-L. Gardette, and J. Lemaire, *Long-term behavior of oil-based varnishes and paints i. spectroscopic analysis of curing drying oils*, *JAOCS* **76**, 967 (1999).

- [10] A. Gacek, *On the Making of Local Paper. A Thirteenth Century Yemeni Recipe*, 99-100 (Publications de l'Université de Provence, 2002).
- [11] N. I. Nikitin, *The chemistry of cellulose and wood*, (1966).
- [12] J. Mallécol, J.-L. Gardette, and J. Lemaire, *Long-term behavior of oil-based varnishes and paints. photo- and thermooxidation of cured linseed oil*, *JAOCS* **77**, 257 (2000).
- [13] B. A. van Driel, T. A. Wezendonk, K. J. van den Berg, P. J. Kooyman, J. Gascon, and J. Dik, *Determination of early warning signs for photocatalytic degradation of titanium white oil paints by means of surface analysis*, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **172**, 100 (2017).
- [14] J. St-Pierre, M.-F. Hivert, P. Perron, P. Poirier, S.-P. Guay, D. Brisson, and L. Bouchard, *Igf2 dna methylation is a modulator of newborn's fetal growth and development*, *Epigenetics* **7**, 1125 (2012), pMID: 22907587, <https://doi.org/10.4161/epi.21855> .
- [15] E. Becker, N. C. Herrera, F. Q. Gunderson, A. I. Derman, A. L. Dance, J. Sims, R. A. Larsen, and J. Pogliano, *Dna segregation by the bacterial actin alfa during bacillus subtilis growth and development*, *The EMBO Journal* **25**, 5919 (2006), <https://www.embopress.org/doi/pdf/10.1038/sj.emboj.7601443> .
- [16] E. Armelin, R. Pla, F. Liesa, X. Ramis, J. I. Iribarren, and C. Alemán, *Corrosion protection with polyaniline and polypyrrole as anticorrosive additives for epoxy paint*, *Corrosion Science* **50**, 721 (2008).
- [17] P. Spellane, *A DC electrochemical method for studying the inhibition of metal corrosion by chromate-containing paint*, *Progress in Organic Coatings* **35**, 277 (1999).
- [18] V. Schamboeck, I. Kryven, and P. D. Iedema, *Acrylate network formation by free-radical polymerization modeled using random graphs*, *Macromolecular Theory and Simulations* , 1700047 (2017).
- [19] Y. R. Zhao, K. B. McAuley, P. D. Iedema, and J. E. Puskas, *Advanced Monte Carlo modeling using weight-based selection of arborescent polyisobutylene molecules in a batch reactor*, *Macromolecular Theory and Simulations* **25**, 134 (2016).
- [20] P. J. Flory, *Molecular Size Distribution in Three Dimensional Polymers. I. Gelation1*, *Journal of the American Chemical Society* **63**, 3083 (1941).

- [21] P. D. Iedema and H. C. J. Hoefsloot, *Synthesis of branched polymer architectures from molecular weight and branching distributions for radical polymerisation with long-chain branching, accounting for topology-controlled random scission*, *Macromolecular Theory and Simulations* **10**, 855 (2001).
- [22] H. Tobita and A. E. Hamielec, *Polymerization processes, 2. modeling of processes and reactors*, in *Ullmann's Encyclopedia of Industrial Chemistry* (Wiley-VCH Verlag GmbH & Co. KGaA, 2000).
- [23] P. Iedema and N. Kolhapure, *Mathematical methods*, in *Handbook of Polymer Reaction Engineering* (Wiley-VCH Verlag GmbH, 2008) pp. 431–532.
- [24] P. A. Mueller, J. R. Richards, and J. P. Congalidis, *Polymerization reactor modeling in industry*, *Macromolecular Reaction Engineering* **5**, 261 (2011).
- [25] D. R. D'hooge, P. H. Steenberge, M.-F. Reyniers, and G. B. Marin, *The strength of multi-scale modeling to unveil the complexity of radical polymerization*, *Progress in Polymer Science* **58**, 59 (2016).
- [26] E. Mastan, X. Li, and S. Zhu, *Modeling and theoretical development in controlled radical polymerization*, *Progress in Polymer Science* **45**, 71 (2015).
- [27] H. Wexler, *Polymerization of drying oils*, *Chemical Reviews* **64**, 591 (1964).
- [28] R. van Gorkum and E. Bouwman, *The oxidative drying of alkyd paint catalysed by metal complexes*, *Coordination Chemistry Reviews* **249**, 1709– (2005).
- [29] Juita, B. Z. Dlugogorski, E. M. Kennedy, and J. C. Mackie, *Low temperature oxidation of linseed oil: a review*, *Fire Science Reviews* **1** (2012).
- [30] L. de Viguerie, P. Payard, E. Portero, P. Walter, and M. Cotte, *The drying of linseed oil investigated by fourier transform infrared spectroscopy: Historical recipes and influence of lead compounds*, *Progress in Organic Coatings* **93**, 46 (2016).
- [31] J. D. van den Berg, N. D. Vermist, L. Carlyle, M. Holcapek, and J. J. Boon, *Effects of traditional processing methods of linseed oil on the composition of its triacylglycerols*, *Journal of Separation Science* **27**, 181 (2004), <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jssc.200301610> .
- [32] N. Fox and G. Stachowiak, *Vegetable oil-based lubricants—a review of oxidation*, *Tribology International* **40**, 1035 (2007).

- [33] L. H. Oakley, F. Casadio, K. R. Shull, and L. J. Broadbelt, *Examination of mechanisms for formation of volatile aldehydes from oxidation of oil-based systems*, *Industrial & Engineering Chemistry Research* **57**, 139– (2018).
- [34] W. Muizebelt, J. Donkerbroek, M. Nielen, J. Hussem, M. Biemond, R. Kloosen, and K. Zobel, *Oxidative crosslinking of alkyd resins studied with mass spectrometry and NMR using model compounds*, *Journal of Coatings Technology* **70**, 80 (1998).
- [35] Z. Oyman, W. Ming, R. van der Linde, R. van Gorkum, and E. Bouwman, *Effect of [mn(acac)<sub>3</sub>] and its combination with 2,2'-bipyridine on the autoxidation and oligomerisation of ethyl linoleate*, *Polymer* **46**, 1731– (2005).
- [36] L. Baij, L. Chassouant, J. J. Hermans, K. Keune, and P. D. Iedema, *The concentration and origins of carboxylic acid groups in oil paint*, *RSC Adv.* **9**, 35559 (2019).
- [37] L. H. Oakley, F. Casadio, K. R. Shull, and L. J. Broadbelt, *Theoretical study of epoxidation reactions relevant to hydrocarbon oxidation*, *Industrial & Engineering Chemistry Research* **56**, 7454– (2017).
- [38] J. Pfaendtner and L. J. Broadbelt, *Mechanistic modeling of lubricant degradation. 1. structure-reactivity relationships for free-radical oxidation*, *Ind. Eng. Chem. Res.* **47**, 2886 (2008).
- [39] J. Pfaendtner and L. J. Broadbelt, *Mechanistic modeling of lubricant degradation. 2. the autoxidation of decane and octane*, *Ind. Eng. Chem. Res.* **47**, 2897 (2008).
- [40] M. Karplus and J. A. McCammon, *Molecular dynamics simulations of biomolecules*, *Nature structural biology* **9**, 646 (2002).
- [41] A. Torres-Knoop, I. Kryven, V. Schamboeck, and P. D. Iedema, *Modeling the free-radical polymerization of hexanediol diacrylate (HDDA): a molecular dynamics and graph theory approach*, *Soft Matter* **14**, 3404 (2018).
- [42] D. Higham, *Modeling and simulating chemical reactions*, *SIAM Review* **50**, 347 (2008).
- [43] A. Cadeddu, E. K. Wylie, J. Jurczak, M. Wampler-Doty, and B. A. Grzybowski, *Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses*, *Angewandte Chemie International Edition* **53**, 8108 (2014), <https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201403708>

- [44] S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, and B. A. Grzybowski, *Computer-assisted synthetic planning: The end of the beginning*, *Angewandte Chemie International Edition* **55**, 5904 (2016), <https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201506101> .
- [45] M. Wozniak, A. Wołos, U. Modrzyk, R. L. Goński, J. Winkowski, M. Bajczyk, S. Szymkuć, B. A. Grzybowski, and M. Eder, *Linguistic measures of chemical diversity and the “keywords” of molecular collections*, *Scientific Reports* **8** (2018), 10.1038/s41598-018-25440-6.
- [46] J. Lehn, *From molecular to supramolecular chemistry*, in *Supramolecular Chemistry* (John Wiley & Sons, Ltd, 2006) Chap. 1, pp. 1–9, <https://onlinelibrary.wiley.com/doi/pdf/10.1002/3527607439.ch1> .
- [47] N. Chomsky, *Three models for the description of language*, *IRE Transactions on Information Theory* **2**, 113 (1956).
- [48] I. Kryven, *Bond percolation in coloured and multiplex networks*, *Nature Communications* **10** (2019), 10.1038/s41467-018-08009-9.
- [49] I. Kryven, *Analytic results on the polymerisation random graph model*, *Journal of Mathematical Chemistry* (2017), 10.1007/s10910-017-0785-1.
- [50] I. Kryven, J. Duivenvoorden, J. Hermans, and P. D. Iedema, *Random graph approach to multifunctional molecular networks*, *Macromolecular Theory and Simulations* **25**, 449– (2016).
- [51] I. Kryven, *Emergence of the giant weak component in directed random graphs with arbitrary degree distributions*, *Physical Review E* **94**, 012315 (2016).
- [52] L. J. Broadbelt, S. M. Stark, and M. T. Klein, *Computer generated pyrolysis modeling: On-the-fly generation of species, reactions, and rates*, *Industrial & Engineering Chemistry Research* **33**, 790 (1994), <https://doi.org/10.1021/ie00028a003> .
- [53] N. Biggs, E. Lloyd, and R. Wilson, *Graph theory: 1736–1936 clarendon press*, (1976).
- [54] J. Ullmann, *An algorithm for subgraph isomorphism*, *Journal of the Association for Computing Machinery* **23**, 31 (1976).

- [55] P. Röse and J. Gasteiger, *Automated derivation of reaction rules for the EROS 6.0 system for reaction prediction*, *Analytica Chimica Acta* **235**, 163 (1990).
- [56] O. Temkin, A. Zeigarnik, and D. Bonchev, *Chemical Reaction Networks: A Graph-Theoretical Approach* (Taylor & Francis, 1996).
- [57] Y. Yoneda, *A computer program package for the analysis, creation, and estimation of generalized reactions—grace. i. generation of elementary reaction network in radical reactions—grace(i)*, *Bulletin of the Chemical Society of Japan* **52**, 8 (1979), <https://doi.org/10.1246/bcsj.52.8> .
- [58] P. Clymans and G. Froment, *Computer-generation of reaction paths and rate equations in the thermal cracking of normal and branched paraffins*, *Computers & Chemical Engineering* **8**, 137 (1984).
- [59] L. J. Broadbelt, S. M. Stark, and M. T. Klein, *Computer generated pyrolysis modeling: On-the-fly generation of species, reactions, and rates*, *Industrial & Engineering Chemistry Research* **33**, 790 (1994).
- [60] J. Song, *Building robust chemical reaction mechanisms : next generation of automatic model construction software*, Ph.D. thesis, Massachusetts Institute of Technology (2004).
- [61] F. Battin-Leclerc, *Development of kinetic models for the formation and degradation of unsaturated hydrocarbons at high temperature*, *Physical Chemistry Chemical Physics* **4**, 2072 (2002).
- [62] S. Rangarajan, A. Bhan, and P. Daoutidis, *Language-oriented rule-based reaction network generation and analysis: Description of RING*, *Computers and Chemical Engineering* **45**, 114 (2012).
- [63] E. Ranzi, T. Faravelli, P. Gaffuri, and A. Sogaro, *Low-temperature combustion: automatic generation of primary oxidation reactions and lumping procedures*, *Combustion and Flame* **102**, 179 (1995).
- [64] N. M. Vandewiele, K. M. V. Geem, M.-F. Reyniers, and G. B. Marin, *Genesys: Kinetic model construction using chemo-informatics*, *Chemical Engineering Journal* **207–208**, 526 (2012).
- [65] S. Vernuccio and L. J. Broadbelt, *Discerning complex reaction networks using automated generators*, *AIChE Journal* **65**, e16663 (2019).

- [66] G. N. Simm, A. C. Vaucher, and M. Reiher, *Exploration of reaction pathways and chemical transformation networks*, *J. Phys. Chem. A* **123**, 385– (2019).
- [67] H. Jeong, B. Tombor, R. A. Z. N. Oltvai, and A.-L. Barabási, *The large-scale organization of metabolic networks*, *Nature* **407**, 651 (2000).
- [68] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles, *Metabolic network structure determines key aspects of functionality and regulation*, *Nature* **420**, 190 (2002).
- [69] D. Y. Zubarev, D. Rappoport, and A. Aspuru-Guzik, *Uncertainty of prebiotic scenarios: The case of the non-enzymatic reverse tricarboxylic acid cycle*, *Scientific Reports* **5** (2015), 10.1021/acs.jpca.8b10007.
- [70] L. J. Broadbelt and J. Pfandner, *Lexicography of kinetic modeling of complex reaction networks*, *AIChE J* **51**, 2112 (2005).
- [71] M. Nijemeisland, L. K. E. A. Abdelmohsen, W. T. S. Huck, D. A. Wilson, and J. C. M. van Hest, *A compartmentalized out-of-equilibrium enzymatic reaction network for sustained autonomous movement*, *ACS Cent. Sci.* **2**, 843 (2016).
- [72] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, *Lethality and centrality in protein networks*, *Nature* **411**, 41– (2001).
- [73] E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Côté, B. K. Shoichet, and L. Urban, *Large-scale prediction and testing of drug activity on side-effect targets*, *Nature* **486**, 361 (2012).
- [74] G. Shinar and M. Feinberg, *Structural sources of robustness in biochemical reaction networks*, *Science* **5971**, 1389 (2010).
- [75] I. Otero-Muras, P. Yordanov, and J. Stelling, *Chemical reaction network theory elucidates sources of multistability in interferon signaling*, *PLOS Com. Boil.* **13** (2017).
- [76] E. Gombrich, *The Story of Art - 16th Edition*, *Story of Art* (Phaidon Press, 1995).
- [77] W. Muizebelt, J. Hubert, and R. Venderbosch, *Mechanistic study of drying of alkyd resins using ethyl linoleate as a model substance*, *Progress in Organic Coatings* **24**, 263 (1994).

- [78] K. Keune, J. Mass, A. Mehta, J. Church, and F. Meirer, *Analytical imaging studies of the migration of degraded orpiment, realgar, and emerald green pigments in historic paintings and related conservation issues*, *Heritage Science* **4** (2016), 10.1186/s40494-016-0078-1.
- [79] J. J. Hermans, K. Keune, A. van Loon, and P. D. Iedema, *The crystallization of metal soaps and fatty acids in oil paint model systems*, *Phys. Chem. Chem. Phys.* **18**, 10896 (2016).
- [80] L. Baij, J. J. Hermans, K. Keune, and P. Iedema, *Time-dependent ATR-FTIR spectroscopic studies on fatty acid diffusion and the formation of metal soaps in oil paint model systems*, *Angewandte Chemie International Edition* **57**, 7351 (2018).
- [81] G. Eumelen, E. Bosco, A. Suiker, A. van Loon, and P. Iedema, *A computational model for chemo-mechanical degradation of historical oil paintings due to metal soap formation*, *Journal of the Mechanics and Physics of Solids* **132**, 103683 (2019).
- [82] S. Michalski, *A physical model of the cleaning of oil paint*, *Studies in Conservation* **35**, 85 (1990).
- [83] A. Phenix and K. Sutherland, *The cleaning of paintings: effects of organic solvents on oil paint films*, *Studies in Conservation* **46**, 47 (2001).
- [84] H. A. J. Carless and R. J. Batten, *Photosensitised oxidation of model unsaturated lipid systems: (4z,7z)-undeca-4,7-diene and (4z)-undec-4-en-7-yne*, *J. Chem. Soc., Perkin Trans. 1*, 1999 (1987).
- [85] L. H. Oakley, F. Casadio, K. R. Shull, and L. J. Broadbelt, *Microkinetic modeling of the autoxidative curing of an alkyd and oil-based paint model system*, *Applied Physics A: Materials Science and Processing* **121**, 869 (2015).
- [86] L. H. Oakley, F. Casadio, P. K. R. Shull, and P. L. J. Broadbelt, *Modeling the evolution of crosslinked and extractable material in an oil-based paint model system*, *Angewandte Chemie International Edition* **57**, 7413 (2018).