A Graph-Based Toy Model of Chemistry

Gil Benkö[†], Christoph Flamm[†,*], and Peter F. Stadler[†,¶,‡]
[†]Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien
[¶]Bioinformatik, Institut für Informatik, Universität Leipzig
[‡]Santa Fe Institute, New Mexico
[*]Corresponding author: Tel: ++43 1 4277 52739, Fax: ++43 1 4277 52793, Email: `xtof@tbi.univie.ac.at`

August 19, 2002

Large scale chemical reaction networks are a ubiquitous phenomenon, from the metabolism of living cells to processes in planetary atmospheres and chemical technology. At least some of these networks exhibit distinctive global features such as the "small world" behavior. The systematic study of such properties, however, suffers from substantial sampling biases in the few networks that are known in detail. A computational model for generating them is therefore required.

Here we present a Toy Model that provides a consistent framework in which generic properties of extensive chemical reaction networks can be explored in detail and that at the same time preserves the "look-and-feel" of chemistry: Molecules are represented as labeled graphs, i.e., by their structural formulae; their basic properties are derived by a caricature version of the Extended Hückel MO theory that operates directly on the graphs; chemical reaction mechanisms are implemented as graph rewriting rules acting on the structural formulae; reactivities and selectivities are modeled by a variant of the Frontier Molecular Orbital Theory based on the Extended Hückel scheme. The approach is illustrated for two types of reaction networks: Diels-Alder reactions and the formose reaction implicated in prebiotic sugar synthesis.

1 INTRODUCTION

Extensive chemical reaction networks arise in very different situations, from the metabolic networks of living cells [1] to the chemistry of planetary atmospheres [2] and combinatorial chemistry, see e.g. [3]. It is therefore of immediate interest to determine which features are generic properties of large-scale reaction networks and which properties are the consequence of a particular chemistry. For instance, do all large reaction networks exhibit the power-law degree distribution that is indicative of small world networks [4, 5], as suggested by data reported in [6]? If this hypothesis should prove to be true it immediately raises the question whether there are other significant differences that imply a natural classification of naturally occurring reaction networks. Unfortunately, the available data in most cases suffer from severe sampling biases because reactions are typically considered only if they link a relatively small number of chemical species of interest. This limitation calls for a computational model that allows an unbiased construction of realistic reaction networks.

In chemistry the changes of molecules upon interaction are not limited to quantitative properties of physical state, such as free energy or density, because molecular interactions do not only produce more of what is already there. Rather, novel molecules can be generated. This is the principal difficulty for any theoretical treatment of the situation. Chemical combinatorics makes it impossible to think of molecules as atomic names whose reactive relationships are tabulated. A computational approach to large scale reaction networks thus requires an underlying model of an *artificial chemistry* to capture the unlimited potential of chemical combinatorics. The investigation of generic properties of chemistries requires the possibility to vary the chemistry itself; hence a self-consistent albeit simplified combinatorial model seems to be more useful than a knowledge-based implementation of the real chemistry which inevitably is subject to sampling biases. The level of realism required for our purposes furthermore does not justify the significant financial burden of accessing chemical reaction databases at a larger scale.

Several approaches to designing such an artificial chemistry have been explored in recent years. The spectrum ranges from chemically accurate quantum mechanical simulations to abstract computational models. Walter Fontana's `AlChemy` [7, 8], for example, represents molecules as λ-calculus expressions and reactions are defined by the operations of "application" of one λ-term to its reaction partner. The result is a new λ-term. Related models are based on a wide variety of different computational paradigms from strings and matrices to Turing machines and graphs [9, 10, 11, 12, 13, 14]. It is worth noting in this context that chemical reactions can in turn be regarded as a model of computation, a possibility that is realized e.g. in the *Chemical Abstract Machine* [15]. The abstract computational models are very useful for understanding algebraic properties of reaction systems; the notion of a self-maintaining set may serve an example. On the other hand, these models lack a natural definition of an energy function and in most cases there is no natural analogue of conservation of mass and atom types. For a recent review of artificial chemistries we refer to [16].

We argue that an energy function is indispensable for any model that is realistic enough to allow us to consider, say, the differences between the metabolic network of *E. coli* and the reactions of hydrocarbons in Jupiter's atmosphere. The reason is that energetic considerations impose constraints that severely limit which ones of the logically possible molecules actually exist, and how they can react with each other. The energy function also determines the directionality of the reactions. Unfortunately, detailed quantum chemical computations are by far too expensive in terms of computer resources. We therefore propose a *Toy Model* of chemistry that is computationally inexpensive and still retains the "look and feel" of the real thing.

Our approach is based on the way how chemical reactions are explained in introductory Organic Chemistry classes: in terms of structural formulae (graphs) and reactions mechanisms (rules for modifying graphs). In fact, graphs are probably *the* natural and the most familiar representation of molecules. Indeed, the description of molecular structures is one of the roots of graph theory [17, 18]. By construction, the graph representation abstracts spatial information to mere adjacency. Thereby we avoid the most time-consuming computation step: embedding the atoms in 3D by means of finding the minima on a potential energy surface [19]. On the other hand, the restriction to graphs implies that several features of real molecules cannot even be defined within the model: (1) There is no distinction between different conformers and, in particular, between *cis* and *trans* isomers at a $\mathsf{C} = \mathsf{C}$ double bond. (2) there is no notion of asymmetric atoms and chirality. In section 2 we show that a caricature version of quantum chemistry can be used on vertex (atom) and edge (bond) labeled graphs. A recent model of interstellar hydrocarbon interconversions [20] follows a similar philosophy.

Once we represent the molecules as (labeled) graphs it becomes natural to view reactions as graph transformations. In other words, a reaction is an instruction or a rule defining how the educt graph must be reshaped by means of insertion, deletion, and relabeling of edges and vertices in order to obtain the product graph. A graph rewriting rule is specified in terms of a graphical pre-condition and a post-condition. A graph rewrite system [21] interprets the graph rewrite rule and performs the graph rewriting step if the graphical pre-condition is matched in a host graph. This is equivalent to finding a subgraph isomorphic to the rule's pre-condition. The subgraph isomorphism problem is in general NP-complete [22]. Following the strategies described by Dörr [23] it is nevertheless possible to solve the subgraph isomorphism in linear time for certain classes of vertex and edge labeled graphs. In section 3 we describe the rewriting part of the Toy Model in some more detail. Reactivities for a particular rewrite can be computed from the Klopman-Salem formula [24, 25]. In particular, the regioselectivity of reaction mechanisms (i.e., which subgraph isomorphism is used if there more than a single one) can therefore be determined within the framework of the model. A chemical application of graph rewriting in a different context, namely to enzymatic DNA processing, is described in [14].

In section 4 we use two well-known examples of chemical reaction networks, the formose reaction and a repetitive Diels-Alder network to demonstrate that the Toy Model is indeed a chemically sensible construction. A number of possibilities for future extensions and refinements of the toy model are briefly considered at the end of this presentation.

2 MOLECULES

In the Born-Oppenheimer approximation the properties of a molecule can at least in principle be derived from the wave function Ψ of its electrons, which in turn can be obtained from the atomic coordinates. Consequently, much of theoretical chemistry is concerned with solving the time-independent Schrödinger equation

$$\widehat{H}\Psi_\alpha = E_\alpha \Psi_\alpha\,, \tag{1}$$

where the electronic Schrödinger operator $\widehat{H}$ contains the coordinates of the atomic nuclei as parameters. The energy calculation used in our Toy Model can be viewed as an extreme simplification of this approach based on the Extended Hückel Theory (EHT) [26]. In this spirit we start with a set of atomic orbitals $\{\chi_i\}$ as a basis and expand the molecular orbital (MO) in the form

$$\Psi_\alpha = \sum_i c_{\alpha,i}\chi_i \tag{2}$$

In the Hückel MO theory [27] only one p-orbital per atom is considered, hence there is a one-to-one correspondence with the spectral theory of the underlying molecular graph, see e.g. [28]. In EHT one typically considers all AOs of the valence shell.

The Hamilton matrix $\mathbf{H}$ and the overlap matrix $\mathbf{S}$ are defined in the usual way by means of the matrix elements

$$H_{ij} = \int \chi_i \widehat{H} \chi_j d\tau \tag{3}$$

$$S_{ij} = \int \chi_i \chi_j d\tau. \tag{4}$$

The Schrödinger equation (1) then takes the form

$$\mathbf{H}\vec{c}_\alpha = E_\alpha \mathbf{S}\vec{c}_\alpha, \tag{5}$$

where $\vec{c}_\alpha$ denotes the vector of coefficients $c_{\alpha,i}$ belonging to the molecular orbital Ψ_α with orbital energy E_α. Let n_α be the number of electrons in orbital Ψ_α. Then the total electronic energy of the molecule is

$$E = \sum_\alpha n_\alpha E_\alpha\,. \tag{6}$$

The electronic population in the atom orbital i is given by

$$q_i = \sum_\alpha n_\alpha c_{\alpha,i}^2 \tag{7}$$

where we assume that the vectors $\vec{c}_k$ are normalized. With the notation $i@a$ for the atom orbital i at atom a we obtain the charge density at atom a in the form

$$q(a) = z_a - \sum_{i@a} q_i \tag{8}$$

where z_a is the number of valence electrons of atom a. The charge density q is the natural starting point for modeling chemical or physical properties of the molecule.

The EHT uses the Wolfsberg-Helmholtz approximation [29]

$$H_{ij} = \kappa(H_{ii} + H_{jj})S_{ij}/2\,. \tag{9}$$

to parametrize the Hamilton matrix in terms of the overlap integrals S_{ij} between any two orbitals and the *atomic valence state ionization potentials* I_i which are the negative diagonal elements of the Hamiltonian matrix, $H_{ii} = -I_i$.

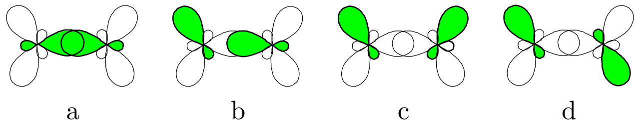

Figure 1: Overlap along a bond (a), "semi-direct" overlap, where only one of the orbitals is directed along the bond (b), and the two possibilities of "indirect" overlaps of two sp^2 orbitals at adjacent atoms (c,d). In the graph-theoretical model (c) and (d) are equivalent because the orientation in the plane is not a property of the molecular graph. In the current implementation "indirect" overlaps are neglected.

In our implementation we use the $1s$ orbital for hydrogen and the usual Slater-type hybrid AOs (sp^3, sp^2, and sp) for carbon, nitrogen, and oxygen. Hybrid orbitals are used because they allow us to simplify the model further by assuming that (1) only orbitals that are localized at neighboring atoms have non-zero overlap and (2) the overlap integrals S_{ij} depend only on the type and orientation of the involved orbitals, see Fig. 1.

Additional rules are added to account for resonance structures that occur when more than one Lewis structure can be drawn for a molecule. For example, lone pairs on one atom can interact with π-systems on the adjacent atoms. Furthermore we treat the bonds in strained (three- and four-membered) rings separately. More details on the parametrization and tables of the parameter values are given in the appendix. In its current implementation, the energy calculation is limited to neutral molecules. It seems straight forward, however, to extend the model to account for charged species and radicals within the same framework.

A molecule is therefore completely determined by a vertex labeled graph Γ, Fig. 2, which was introduced by O. Polanski [30]. The vertices of Γ are the atom orbitals (labeled by atom type and hybridization); edges denote overlaps of adjacent orbitals. This *orbital graph* Γ is obtained in an unambiguous way from the chemical structure formula by means of the VSEPR rules [31]. It follows that, in the framework of the Toy Model, the structure formula already encapsulates the complete information about the molecule.

Obviously this is a rather crude approximation that, in particular, disregards the influence of three-dimensional space by reducing the molecular structure to connectivity information. Nevertheless, we obtain a qualitatively reasonable behavior of the electronic energies as shown by the comparisons between computed and experimental energies, Fig. 3.

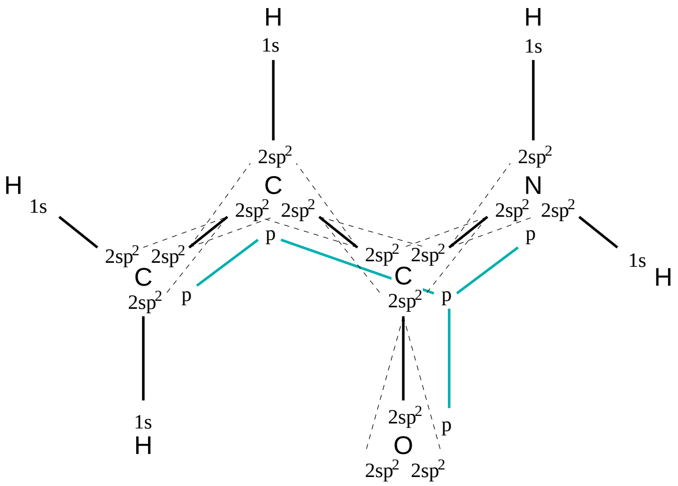

Figure 2: Orbital graph of propenamide $H_2C=CH-CONH_2$. Direct, semi-direct σ-overlaps, and π-overlaps are represented by solid black, dashed, and solid grey lines.

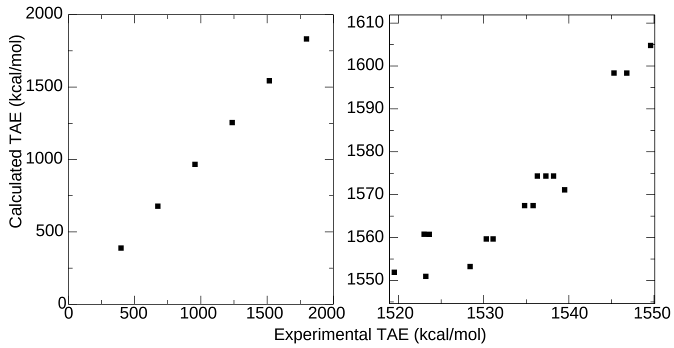

Figure 3: Plots of calculated versus experimental Total Atomization Energies (TAE). Left: Homologous series of n-alkanes from methane to hexane. Right: C_4H_{10} isomers, in order of increasing experimental TAE those are 1-hexyne, 2- and 3-hexyne, 3,3-dimethyl-1-butyne, 1,5-hexadiene, Z- and E-1,4-hexadiene, Z- and E-1,3-hexadiene, Z,Z- and E,Z- and E,E-2,4-hexadiene, bicyclo[3.1.0]hexane, 4- and 3-methylcyclopentene, 1-methylcyclopentene. Experimental TAE values are taken from [32].

3 REACTIONS

Graph rewrite systems, also called graph grammars, operate on edge and vertex labeled graphs [21]. Intramolecular reactions such as rearrangements and substitution reactions are naturally implemented as rewrite rules that act on the molecular graphs. A rewrite rule consists of three parts, a left graph, a right graph and the context. The context of a rule is the part of the graph which remains unchanged during a rewriting step. A rewrite rule is applicable to a molecular graph if it contains a subgraph that is isomorphic to the rule's left-hand side (which is the union of left graph and context). In Fig. 4 the rewrite rule for intermolecular Diels-Alder rearrangement [33] is shown.

The formalism of graph rewriting is a more general and more versatile framework for specifying chemical reactions than e.g. the Dugundji-Ugi theory [34, 35]. This generality comes at a cost: not every graph rewriting rule is meaningful as a chemical reaction mechanism. Most importantly, chemical reactions do not create, annihilate,

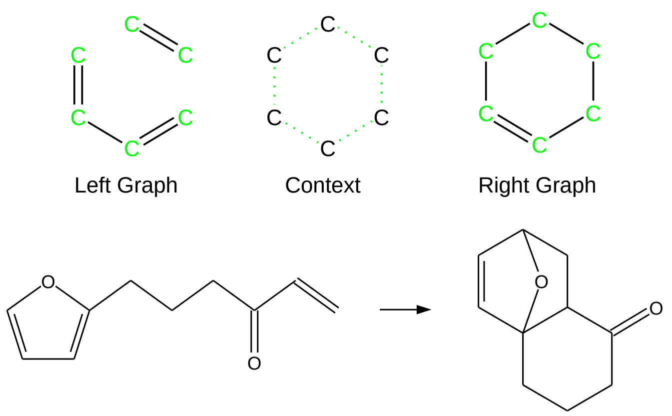

Figure 4: Intramolecular Diels Alder rearrangement (iDAR). Top: rewrite rule; since all bounds change their type during the rewrite, the context consists of the six C-atoms only. Bottom: Application of iDAR to the synthesis of a bridged ring system.

or change atoms. Thus chemical rewrite rules must satisfy the principle of *conservation of vertex labels*. Furthermore, the total number of valence electrons must be conserved. Currently we consider only single, double, and triple bonds. Hence we require *conservation of total bond order* for any chemical transformation. Since the rewrite rules are graphs themselves, it is of course easy to verify these two conservation laws by simply comparing the list of labels and the total bond order of the left and the right graph of the rule.

The graph rewrite engine is implemented in `Haskell`, a lazy functional programming language [36]. Since it is not easy to glue together pieces of code written in functional and imperative programming languages (e.g. `C`), the engine is designed as a client/server application. The client sends a graph to the server, which performs the rewrite step and sends the transformed graph back to the client. The rewrite behavior of the server only depends on the set of rewriting rules which are read from a file at server startup. This program architecture allows us to easily fit the rewrite engine to the needs of a particular task by simply changing the client. The server can be run in two rewriting modes: random rewrite and priority rewrite. In the former mode a rewrite rule is picked at random from the set of potentially applicable rules, while in the latter mode the rule with the highest "priority value" is chosen.

The graph rewrite framework can be applied to modelling bimolecular reactions. As an example consider the Aldol condensation [37], Fig. 5, which e.g. forms the core step of the formose reaction [38]. The idea is to split the bimolecular reaction mechanism into two half reaction rules, one for each educt molecule, and a joining rule, that describes how the two educt molecules are joined together. The half reaction rules describe the local changes within each of the reacting molecules, whereas the joining rule captures the inter-molecular bond formation. In reactions such as Cannizzaro's disproportionation [39] or Olefin metathesis [40] the joining rule is of course empty.

Let us now consider the rewriting step for a bimolecular reaction in detail. First the client sends the two educt graphs to the server. The server then constructs all sub-

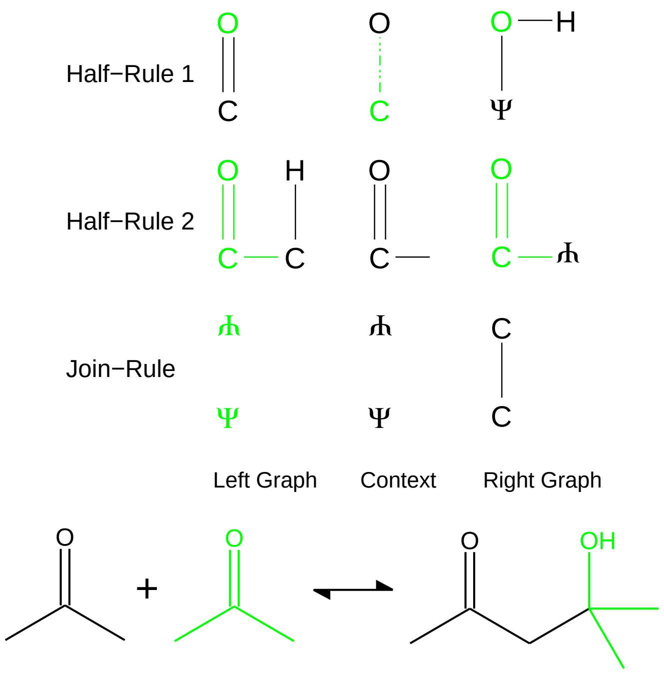

Figure 5: Aldol Reaction. Top: rewrite rule; the two half-rules describe the locale changes in the reacting molecules during Aldol condensation, while the half-rule-join describes only intermolecular changes; notice the special label Ψ, acting as anchor for the intrermolecular bond to be formed. Bottom: Application to the synthesis of β-hydroxy-carbonyls.

graph isomorphisms for the left hand side of both half-rules for both graphs. If the list of subgraph isomorphisms for one of the two half-rules is empty for both graphs, the rule is not applicable and the server sends the two graphs unaltered back to the client. This case corresponds to an "elastic collision".

Otherwise the server picks a half rule at random for the first graph and then a corresponding half rule for the second graph. This corresponds to choosing a reaction channel if there is more than one subgraph isomorphism. Then the rewriting is performed for both half-rules, then the join half-rule is applied, and finally the resulting graph is sent back to the client.

Instead of picking a subgraph isomorphism at random from the list, it is possible to consider all reaction channels and to compute a reactivity index for each of them. The client can then pick the reaction channel (pair of subgraph isomorphisms for the two half-rules). Consider two "systems" of atoms $\mathcal{A}$ and $\mathcal{B}$. In the case of bimolecular reactions of course $\mathcal{A}$ and $\mathcal{B}$ are the two molecules. Within the context of the Toy Model it seems natural to start with the Klopman-Salem formula [24, 25, 41], eq.(10) below, that predicts the energy increment incurred by combining systems $\mathcal{A}$ and $\mathcal{B}$ in the following form:

$$\begin{aligned}\Delta E = &\sum_{a\in\mathcal{A},b\in\mathcal{B}} G_{ab} + \sum_{a\in\mathcal{A},b\in\mathcal{B}} \frac{q(a)q(b)}{\epsilon r_{ab}} \\ &- \left(\sum_{\alpha\in\mathcal{A}}^{\text{occ}} \sum_{\zeta\in\mathcal{B}}^{\text{unocc}} F^{\alpha,\zeta} + \sum_{\alpha\in\mathcal{B}}^{\text{occ}} \sum_{\zeta\in\mathcal{A}}^{\text{unocc}} F^{\alpha,\zeta} \right)\end{aligned} \tag{10}$$

Here r_{ab} is the bond length (which of course is a tabulated parameter in our setting) and ϵ is the dielectric constant of the reaction medium. The interaction terms have the following explicit form:

$$G_{ab} = -\sum_{i@a}\sum_{j@b}(q_i+q_j)H_{ij}S_{ij}\ ,$$
$$F^{\alpha,\zeta} = \frac{2}{E_\zeta - E_\alpha}\left(\sum_{a\in\mathcal{A}}\sum_{i@a}\sum_{b\in\mathcal{B}}\sum_{j@b} c_{\alpha,i}c_{\zeta,j}H_{ij}\right)^2\ , \tag{11}$$

where $\alpha \in \mathcal{A}$ and $\zeta \in \mathcal{B}$ is an occupied and an unoccupied MO, respectively. With the abbreviation

$$W_{ab}^{\alpha\zeta} = \sum_{i@a}\sum_{j@b} c_{\alpha,i}c_{\zeta,j}H_{ij} \tag{12}$$

we obtain a four-point term

$$F_{ab;a'b'} = 2\sum_{\alpha\in\mathcal{A}}^{\text{occ}}\sum_{\zeta\in\mathcal{B}}^{\text{unocc}} \frac{W_{ab}^{\alpha\zeta}W_{a'b'}^{\alpha\zeta}}{E_\zeta - E_\alpha} + 2\sum_{\alpha\in\mathcal{B}}^{\text{occ}}\sum_{\zeta\in\mathcal{A}}^{\text{unocc}} \frac{W_{ba}^{\alpha\zeta}W_{b'a'}^{\alpha\zeta}}{E_\zeta - E_\alpha} \tag{13}$$

that allows us to write ΔE as an expansion of atom pairs and quadruples. Within the approximation of the Toy Model all contributions (with the exception of the Coulomb term) that do not belong to new bonds (or bonds with increasing bond order) vanish because their overlap integrals are zero. Thus

$$\Delta E = \sum_{(a,b)}\left(G_{ab} + \frac{q(a)q(b)}{\epsilon r_{ab}} - F_{ab;ab}\right) - \sum_{(a,b)\neq(a',b')} F_{ab;a'b'} \tag{14}$$

where the sums run only over newly formed bonds (a,b). The situation can be simplified further by considering only the frontier orbitals [42], i.e. the HOMO of one system and the LUMO of the other one. In this case the sums in equ.(13) reduce to a single term. Often this is approximated by $\Delta E = \xi/(E_\zeta - E_\alpha)$ with an empirical constant ξ that depends only on the reaction mechanism [41]. We have used this simplification for generating the two examples in section 4, Figs. 7 and 8. The same formalism can be applied to intra-molecular reactions by setting $\mathcal{A} = \mathcal{B}$; in eq.(13) we then retain only one of the two double sums (which become identical in this case).

The reactivity ΔE allows us to model regio-selectivity. If more than one subgraph isomorphism, i.e., more than one possible reaction channel, has been found one simply has to evaluate ΔE for all of them. Then the rewrite with the smallest ΔE value is chosen. Of course, the reaction scheme could then be modified to select a reaction channel with a probability proportional to its Boltzmann weight $\exp(-\Delta E/RT)$, i.e., according to Arrhenius' law. This would be the natural starting point for the stochastic simulation of a reaction network e.g. using Gillespie's approach [43].

4 NETWORKS

A chemical reaction

$$a\mathsf{A} + b\mathsf{B} + \cdots \to v\mathsf{V} + w\mathsf{W} + \ldots \tag{15}$$

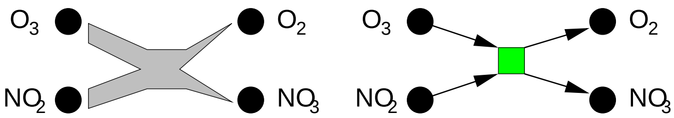

Figure 6: Representation of a chemical reaction $NO_2 + O_3 \to NO_3 + O_2$ as a directed hypergraph $\mathcal{H}(V,E)$. The chemical species are the vertices $\mathsf{X} \in V$. Each reaction is represented by a single directed hyperedge connecting educts with products. Directed hypergraphs are conveniently displayed as bipartite directed graphs. Here the reactions are represented as a second type of vertices. Directed edges connect educts with the reaction vertex and the reaction vertex with products of the reaction.

can be described as a directed hypergraph $\mathcal{H}(V,E)$ in which chemical species are the vertices [44]. Each reaction forms a hyperedge $\boldsymbol{\rho} \in E$ that connects educts with products. Alternatively, the reactions are represented as a second class of vertices. Directed edges then connect the educts with the reaction vertex and the reaction vertex with the products, Fig. 6.

The algebraic representation of $\mathcal{H}$ is the *stoichiometric matrix* $\mathbf{S}$. Its entries are the *stoichiometric coefficients* $s_{\mathsf{X}\boldsymbol{\rho}}$, i.e., the numbers of molecules of species X that are produced ($s_{\mathsf{X},\boldsymbol{\rho}} > 0$) or consumed ($s_{\mathsf{X},\boldsymbol{\rho}} < 0$) in reaction $\boldsymbol{\rho}$. Reversible reactions are considered as two separate reactions. We remark that $\mathbf{S}$ is the starting point for quantitative approaches to analyzing large networks such as flux analysis [45, 1] and control analysis [46].

Suppose we are given a list of reaction mechanisms and an initial list $\mathfrak{L}_0$. The reaction network can be built up systematically by means of "orderly generation" [47, 48]. Performing all unimolecular reactions on each molecule $\mathsf{M} \in \mathfrak{L}_0$ and all bimolecular reactions with each pair of molecules $(\mathsf{M}_1, \mathsf{M}_2) \in \mathfrak{L}_0 \times \mathfrak{L}_0$ we obtain a new list $\mathfrak{L}_1'$ and a list of new molecules $\mathfrak{L}_1 = \mathfrak{L}_1' \setminus \mathfrak{L}_0$. The recursion then proceeds in the obvious way:

$$\mathfrak{L}_{k+1}' = \left(\bigcup_{j=0}^{k-1}\mathfrak{L}_j\right) \times \mathfrak{L}_k \cup (\mathfrak{L}_k \times \mathfrak{L}_k) \tag{16}$$

and $\mathfrak{L}_{k+1} = \mathfrak{L}_{k+1}' \setminus \bigcup \mathfrak{L}_k$.

In order to check whether a newly generated molecule m is already contained in a previous list a comparison of the structural formulae must be performed. This amounts to a test of graph isomorphism, for which neither an efficient algorithm nor proof of NP-completeness is known in general [49]. The chemically relevant problem of testing graph isomorphism with bounded vertex degree (i.e., bounded valency of the atoms) can be solved in polynomial time [50]. We transform the molecular graphs into their *canonical SMILES* representation [51]. The isomorphism test then reduces to simple string comparison.

Diels-Alder Reaction

The Diels-Alder reaction has been studied extensively both because of its importance in natural products synthesis and because it can be understood in detail by means of simple semi-empirical methods. It involves the reaction

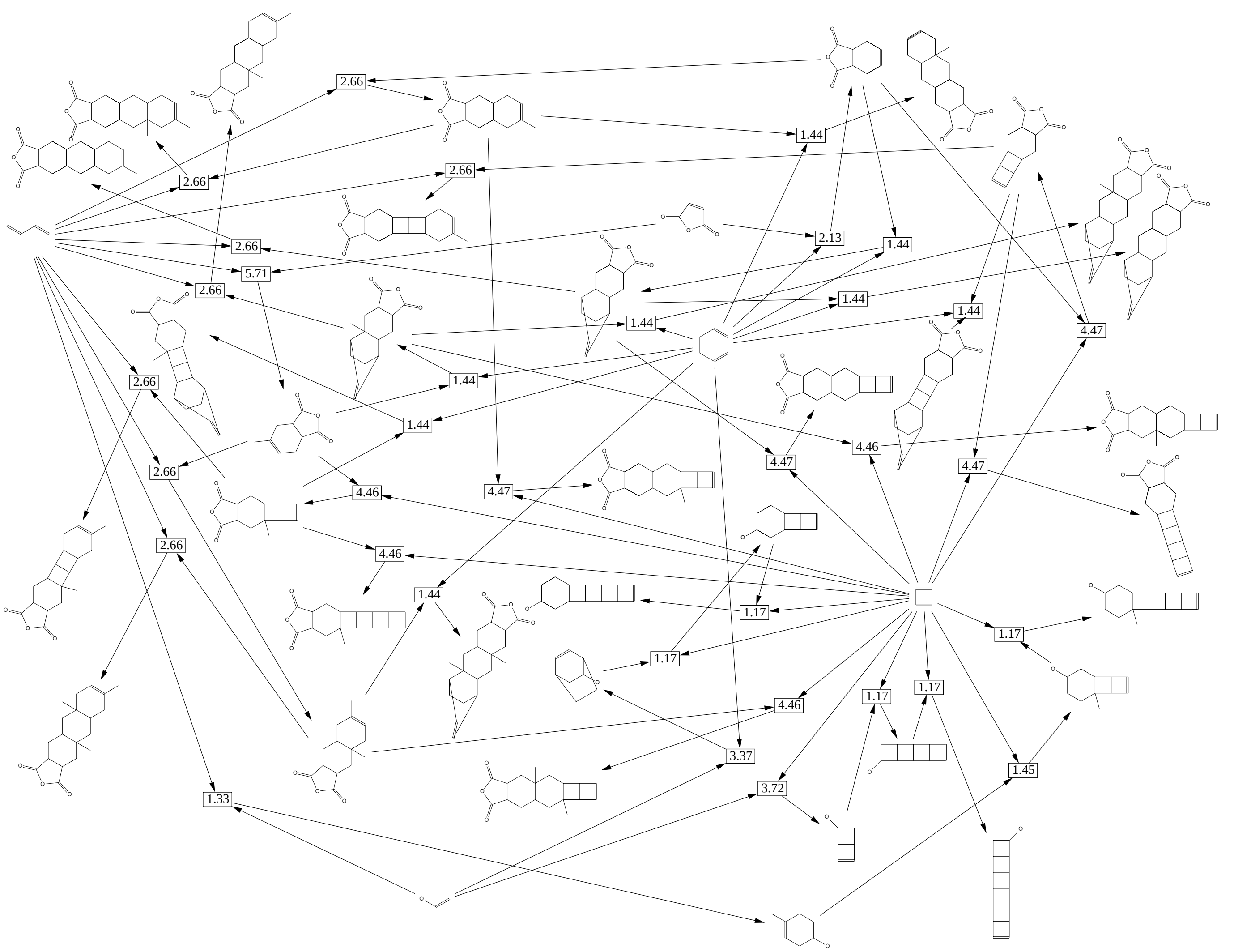

Figure 7: Network of Diels-Alder reaction constructed with 3 iterations of the orderly generation algorithm. The initial mixture consists of cyclobutadiene, ethenol, phtalic anhydride, methylbutadiene, and cyclohexa-1,3-diene. Each rectangle represents one reaction, its label indicates the reaction rate using the proportionality constant ξ from [52].

between two linear π-systems of length 2 and 4, respectively [52]. The product is again a π-system and thus may react again in a Diels-Alder reaction. Recently, it has been used to synthetize particular classes of polymers [53].

Fig. 7 displays the reaction network obtained by repetitive Diels-Alder reactions of a simple initial mixture.

Formose Reaction

The synthesis of sugars from formaldehyde under alkaline conditions ("formose reaction") was discovered more than a century ago [54]. It is one of the earliest examples of a reaction network that is collectively autocatalytic in the sense that the reaction products catalyze their own formation. The condensation of formaldehyde proceeds by means of repeated aldol condensations and subsequent dismutations [55, 38].

The formose reaction has been studied in much detail because of its importance as a potential prebiotic pathway [56]. More than 40 different sugars have been identified in the reaction mixture [57]. The network produced by the Toy Model is shown in Fig. 8.

5 DISCUSSION

We have described here a Toy Model that is at least close to a minimal implementation of an artificial chemistry exhibiting what we consider the defining features of "real" chemistry. We represent molecules explicitly as arrangements of atoms (labeled graphs) and define an energy function along the lines of quantum chemistry. This energy model forms the basis of full-fledged chemical thermodynamics and kinetics. Chemical reactions are implemented as graph rewriting rules that have to obey the principle of conservation of matter. These features distinguish our Toy Model from artificial chemistries that are defined on abstract algebraic structures such as lambda calculus, Turing machines, or term rewriting.

A number of extensions of the present Toy Model are desirable. For instance, the current implementation of the model considers only neutral molecules composed of C, H, O, and N. An extension to an expanded set of chemical elements, most importantly S, P, Si, and the halogenes is straightforward. The inclusion of charged particles and radicals also does not seem to pose problems in the current framework. Additional types of chemical bonds, in partic-

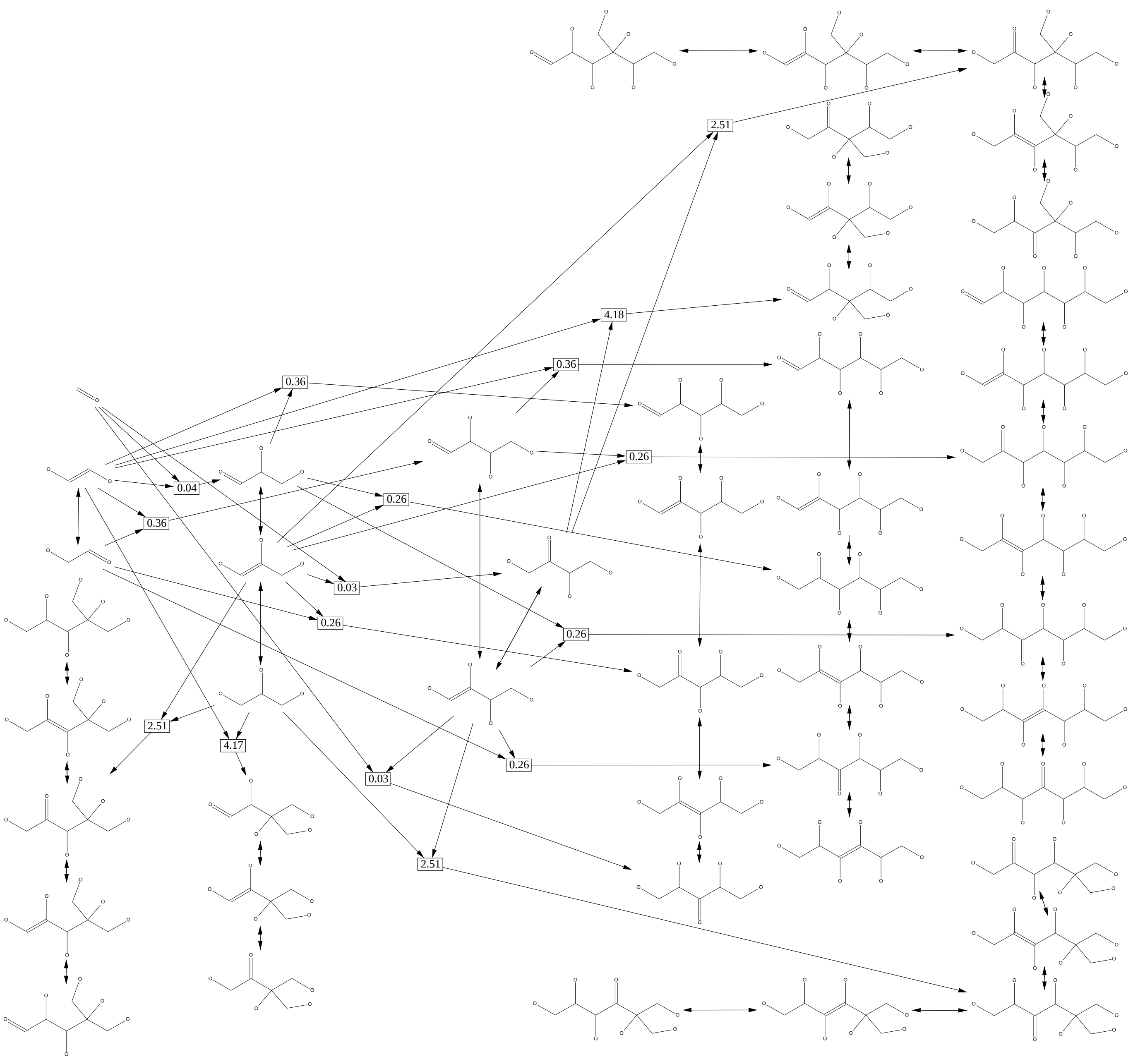

Figure 8: Formaldehyde condensation reaction network. The initial mixture consists of formaldehyde H_2CO and glycol aldehyde $CH_2OH - CHO$ and reacts via aldol condensations and dismutations. The aldol condensation was simulated by the condensation of a keto with an enole group. In order to account for cyclisation, which limits the network, we do not permit carbon chains with more than four members to undergo further aldol condensations. The network generation algorithm thus converges already after two iterations. Reaction rates are computed using the proportionality constant ξ for nucleophilic substitution from [24].

ular hydrogen bonds and the "three center bonds" that are common in boron compounds can be approximated by the orbital graph formalism.

The interaction of a molecule with a more complex environment, in particular a solvent, is easily incorporated into the Toy Model using an implicit solvation model [58] such as Kirkwood's equation

$$\Delta G_{\text{solv}} = -\frac{\epsilon - 1}{2\epsilon + 1}\frac{\boldsymbol{\mu}^2}{a^3} \qquad (17)$$

Here a is the radius of the molecule and $\boldsymbol{\mu}$ is its dipole moment and ϵ is the dielectric constant of the medium. Both a and $\boldsymbol{\mu}$ have to be replaced by appropriate graph descriptors. For example a could be replaced by the Wiener index [59, 60] (with a proper normalization). A topological index for vertex weighted graphs that could serve as a "graph theoretical dipole moment" will be discussed elsewhere.

The reactivities from equ.(10) can be translated into reaction rate contants e.g. using Arrhenius' law. An alternative approach to determining rate constants is QSPR, see e.g. [48]. This class of models is, however, of limited interest for our purposes because it is restricted to reaction mechanism for which a sufficient amount of experimental data is available.

It is straightforward to derive the kinetic differential equations for a given network of reactions using the rules of mass action kinetics. Simulations of this type will provide a very detailed insight into the structure of reaction networks and form the basis for more sophisticated approaches to network analysis [61].

The Toy model implements chemical reactions as explicit rewrite rules. In principle it is possible to simulate the collision of two molecules by assigning a collection of potential new bonds between them. Since the corresponding reactivity ΔE and the over-all reaction energy can be computed, one could in principle simulate reactions at this level. The computational cost would be immense, however. Nevertheless, one could use collision simulations to search for new reaction mechanism. This might be of particular interest when the Toy Model is used to explore "exotic chemistries".

The reaction networks generated by the Toy Model are themselves graphs that can be characterized by a variety of standard measures such as diameter, center, scaling behaviour and small-world classification. The comparison of different networks will reveal generic properties of networks as well as specific features of different classes of reactions. Furthermore, the Toy Model includes parameters (see appendix) that can be varied. The dependency of properties on these parameters could provide a measure for the stability and robustness of the network.

Acknowledgements

Discussions with Othmar Steinhauser and Andreas Svrcek-Seiler, and a travel grant (G.B.) from the University of Vienna are gratefully acknowledged. The SMILES uniquetizer was kindly provided by Delores Grunwald of Computer Science Corp., Duluth, MN.

APPENDIX A: PARAMETERS

The energy calculation in the Toy Model is parametrized in terms of ionization energies I_j and overlap integrals S_{ij} of the usual Slater-type hybrid orbitals. The overlap integrals S_{ij} depend only on the type and orientation of the involved orbitals. The values listed in Tab. 1 apply to σ overlaps of hybridized orbitals that are oriented toward each other along a bond (upper left scheme in Fig. 1) and to π overlaps between p orbitals.

At this stage, the parameters for the "direct" overlaps in Tab. 1 are used to calculate the S_{ij} values for the "semi-direct" and "indirect" by means of the simple scaling factors compiled in Tab. 1. A more sophisticated model for the "semi-direct" and "indirect" overlaps could easily be used in a future implementation of the toy model.

Hyperconjugation [62, 63] denotes the overlap between a p orbital and a sp^3 orbital at an adjacent atom this is not oriented along the bond. The hyperconjugation overlap is included with only one of the three sp^3 orbitals, which is chosen arbitrarily. An alternative way of incorporating the coupling of the σ and the π system is to consider a fictitious overlap of the p orbital with the adjacent sp^3 orbital that is directed along the bond. In the current implementation the fictitious p-sp^3 overlap is set to 0.

The overlaps corresponding to bonds that lie in three- or four-membered rings are scaled by a factor that reflects the fact that the *banana-bonds* in constrained rings are weaker, see Tab. 1.

APPENDIX B: REWRITE RULES

The graph rewrite rules are conveniently specified using the `Graph Meta Language` (GML) [64]. As an example we include here the specifucation of the Diels Alder reaction:

```
# Diels Alder
rule [
 context [
  node [ id 1 label "C" ]
  node [ id 2 label "C" ]
  node [ id 3 label "C" ]
  node [ id 4 label "C" ]
  node [ id 5 label "C" ]
  node [ id 6 label "C" ]
 ]
 left [
  edge [ source 1 target 2 label "=" ]
  edge [ source 2 target 3 label "-" ]
  edge [ source 3 target 4 label "=" ]
  edge [ source 5 target 6 label "=" ]
 ]
 right [
  edge [ source 1 target 2 label "-" ]
  edge [ source 2 target 3 label "=" ]
  edge [ source 3 target 4 label "-" ]
  edge [ source 4 target 5 label "-" ]
  edge [ source 5 target 6 label "-" ]
  edge [ source 6 target 1 label "-" ]
 ]
]
```

REFERENCES

(1) Fell, D., *Understanding the Control of Metabolism*. No. 2 in Frontiers in Metabolism, Portland Press, London, **1997**.

(2) Yung, Y. L.; DeMore, W. B., *Photochemistry of Planetary Atmospheres*. New York: Oxford University Press, **1999**.

(3) Höllering, R.; Gasteiger, J.; Steinhauer, L.; Schulz, K.; Herwig, A., The Simulation of Organic Reactions: From the Degradation of Chemicals to Combinatorial Synthesis. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 482–494.

(4) Watts, D. J., *Small Worlds*. Princeton NJ: Princeton University Press, **1999**.

(5) Albert, R.; Barabási, A.-L., Statistical Mechanics of Complex Networks. *Rev. Mod. Phys.* **2002**, *74*, 47–97.

(6) Gleiss, P. M.; Stadler, P. F.; Wagner, A.; Fell, D. A., Relevant Cycles in Chemical Reaction Network. *Adv. Complex Syst.* **2001**, *4*, 207–226.

Table 1: Parameters for the graph orbital model. The top line gives the Coulomb integrals I for the atom orbitals that are currently implemented. Overlap integrals are listed separately for σ and π bonds. Semi-direct and indirect overlaps and banana-bonds in constrained rings in the sense of Fig. 1 are parametrized as the product of the bonding interaction with a scaling factor.

σ		H	C			N			O	
		s	sp^3	sp^2	sp	sp^3	sp^2	sp	sp^3	sp^2
I		-13.6	-13.9	-14.5	-15.4	-16.6	-17.6	-19.7	-19.2	-20.6
H	s	0.75	0.69	0.65	0.66	0.62	0.63	0.63	0.55	0.57
C	sp^3	0.69	0.65	0.67	0.71	0.60	0.63	0.65	0.54	0.57
C	sp^2	0.65	0.67	0.77	0.80	0.70	0.73	0.77	0.64	0.68
C	sp	0.66	0.71	0.80	0.87	0.77	0.80	0.84	–	–
N	sp^3	0.62	0.60	0.70	0.77	0.58	0.61	0.65	0.63	0.67
N	sp^2	0.63	0.63	0.73	0.80	0.61	0.70	0.73	0.63	0.67
N	sp	0.63	0.65	0.77	0.84	0.65	0.73	0.82	–	–
O	sp^3	0.55	0.54	0.64	–	0.63	0.63	–	–	–
O	sp^2	0.57	0.57	0.68	–	0.67	0.67	–	–	–

π	C	N	O
	p	p	p
I	-11.4	-13.4	-14.8
π	C	N	O
C	0.38	0.31	0.26
N	0.31	0.31	0.26
O	0.26	0.26	0.26

Scaling Factors sp^x-sp^y		
direct	Fig. 1a	1.0
semi-direct	Fig. 1b	0.1
	with H	0.0
indirect	Fig. 1c,d	0.0

σ-π coupling		banana bonds		Wolfsberg-Helmholtz equation	
hyperconjugation	0.8	3-ring	0.7	Scaling factor κ	1.75
fictitious p-sp^3	0.0	4-ring	0.8		

(7) Fontana, W., Algorithmic Chemistry. In: Langton, C. G.; Taylor, C.; Farmer, J. D.; Rasmussen, S., eds., *Artificial Life II*, Santa Fe Institute Studies in the Sciences of Complexity, Redwood City, CA: Addison-Wesley, **1992** 159–210.

(8) Fontana, W.; Buss, L. W., What would be conserved if 'the tape were played twice'? *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 757–761.

(9) Bagley, R. J.; Farmer, J. D., Spontaneous emergence of a metabolism. In: Langton, C. G.; Taylor, C.; Farmer, J. D.; Rasmussen, S., eds., *Artificial Life II*, Santa Fe Institute Studies in the Sciences of Complexity, Redwood City, CA: Addison-Wesley, **1992** 93–141.

(10) Banzhaf, W.; Dittrich, P.; Eller, B., Self-organization in a system of binary strings with spatial interactions. *Physica D* **1999**, *125*, 85–104.

(11) Speroni di Fenizio, P., A less abstract articial chemistry. In: Bedau, M.; McCaskill, J.; Packard, N.; Rasmussen, S., eds., *Artificial Life VII*, Cambridge, MA: MIT Press, **2000** 49–53.

(12) Ugi, I.; Stein, N.; Knauer, M.; Gruber, B.; Bley, K.; Weidinger, R., New Elements in the Representation of the Logical Structure of Chemistry by Qualitative Mathematical Models and Corresponding Data Structures. *Top. Curr. Chem.* **1993**, *166*, 199–233.

(13) Thürk, M., *Ein Modell zur Selbstorganisation von Automatenalgorithmen zum Studium molekularer Evolution*. Ph.D. thesis, Universität Jena, Germany, **1993**.

(14) McCaskill, J. S.; Niemann, U., Graph Replacement Chemistry for DNA Processing. In: Condon, A.; Rozenberg, G., eds., *DNA Computing*, vol. 2054 of *Lecture Notes in Computer Science*, Berlin, D: Springer, **2000** 103–116.

(15) Berry, G.; Boudol, G., The chemical abstract machine. *Theor. Comp. Sci.* **1992**, *96*, 217–248.

(16) Dittrich, P.; Ziegler, J.; Banzhaf, W., Artificial Chemistries — A Review. *Artificial Life* **2001**, *7*, 225–275.

(17) Cayley, A., On the Mathematical Theory of Isomers. *Philos. Mag.* **1874**, *47*, 444–446.

(18) Sylvester, J. J., On an application of the new atomic theory to the graphical representation of the invariants and covariants of binary quantics, with three appendices. *Amer. J. Math.* **1878**, *1*, 64–128.

(19) Heidrich, D.; Kliesch, W.; Quapp, W., *Properties of Chemically Interesting Potential Energy Surfaces*, vol. 56 of *Lecture Notes in Chemistry*. Berlin: Springer-Verlag, **1991**.

(20) Patra, S. M.; Mishra, R. K.; Mishra, B. K., Graph-theoretic study of certain interstellar reactions. *Intern. J. Quantum Chem.* **1997**, *62*, 495–508.

(21) Nagl, M., *Graph-Grammatiken, Theorie, Implementierung, Anwendung*. Braunschweig: Vieweg, **1979**.

(22) Garey, M. R.; Johnson, D. S., *Computers and Intractability*. New York: W. H. Freeman and Co., **1979**.

(23) Dörr, H., *Efficient Graph Rewriting and Its Implementation*. Berlin Heidelberg: Springer-Verlag, **1995**.

(24) Klopman, G., Chemical reactivity and the concept of charge- and frontier-controlled reactions. *J. Am. Chem. Soc.* **1968**, *90*, 223–243.

(25) Salem, L., Intermolecular Orbital Theory of the Interaction between Conjugated Systems. I. General Theory; II. Thermal and Photochemical Calculations. *J. Am. Chem. Soc.* **1968**, *90*, 543–552 & 553–566.

(26) Hoffmann, R., An Extended Hückel Theory. I. Hydrocarbons. *J. Chem. Phys.* **1963**, *39*, 1397–1412.

(27) Hückel, E., Quantentheoretische Beiträge zum Benzolproblem. I. Die Elektronenkonfiguration des Benzols und verwandter Verbindungen. *Z. Physik* **1931**, *70*, 204–286.

(28) Trinajstić, N.; Mihalić, Z.; Graovac, A., The interplay between graph theory and molecular orbital theory. In: Bonchev, D.; Mekenyan, O., eds., *Graph-Theoretical Approaches to Chemical Reactivity*, Dordrecht, NL: Kluwer, **1994** 37–72.

(29) Wolfsberg, M.; Helmholtz, L., The Spectra and Electronic structure of the Thetrahedral Ions MnO_4^-, $CrO_4^=$, and ClO_4^-. *J. Chem. Phys.* **1952**, *20*, 837–843.

(30) Polansky, O. E., Graphs in quantum chemistry. *MATCH* **1975**, *1*, 183–195.

(31) Gillespie, R. J.; Nyholm, R. S., Inorganic Stereochemistry. *Quart. Rev. Chem. Soc.* **1957**, *11*, 339–380.

(32) Computational Chemistry Comparison and Benchmark DataBase, Release 6a. **2002**, `http://srdata.nist.gov/cccbdb/`.

(33) Diels, O.; Alder, K., Synthesen in der hydroaromatischen Reihe. *Liebigs Ann. Chem.* **1928**, *460*, 98–122.

(34) Dugundji, J.; Ugi, I., Theory of the *be*- and *r*-matrices. *Top. Curr. Chem.* **1973**, *39*, 19–29.

(35) Fontain, E.; Reitsam, K., The generation of reaction networks with RAIN. 1. The reaction generator. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 96–101.

(36) `http://www.haskell.org/`. `Haskell` is a general purpose, purely functional programming language. `Haskell` compilers are freely available for almost any computer.

(37) Wurtz, M. A., Sur un aldéhyde-alcool. *Bull. Soc. Chim. Fr.* **1872**, *17*, 426–442.

(38) Pfeil, E.; Ruckert, H., Die Bildung von Zuckern aus Formaldehyd unter der Einwirkung von Laugen. *Liebigs Ann. Chem.* **1961**, *641*, 121–131.

(39) Cannizzaro, S., Über den der Benzoësäure entsprechenden Alkohol. *Liebigs Ann. Chem.* **1853**, *88*, 129–130.

(40) Ivin, K. J., *Olefin Metathesis*. Academic Press, London, **1983**.

(41) Fleming, I., *Frontier Orbitals and Organic Chemical Reactions*. Wiley: New York, **1976**.

(42) Fukui, K.; Yonezawa, T.; Shingu, H., A molecular orbital theory of reactivity in aromatic hydrocarbons. *J. Chem. Phys.* **1952**, *20*, 722–725.

(43) Gillespie, D. T., Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.* **1977**, *81*, 2340–2361.

(44) Zeigarnik, A. V., On Hypercycles and Hypercircuits in Hypergraphs. In: Hansen, P.; Fowler, P. W.; Zheng, M., eds., *Discrete Mathematical Chemistry*, vol. 51 of *DIMACS series in discrete mathematics and theoretical computer science*, Providence, RI: American Mathematical Society, **2000** 377–383.

(45) Clarke, B. L., Stoichiometric Network Analysis. *Cell Biophys.* **1988**, *12*, 237–253.

(46) Heinrich, R.; Schuster, S., The modelling of metabolic systems. Structure, control and optimality. *Biosystems* **1998**, *47*, 61–77.

(47) Read, R. C., Every one a winner. *Ann. Discr. Math.* **1978**, *2*, 107–120.

(48) Faulon, J.-L.; Sault, A. G., Stochastic Generator of Chemical Structures. 3. Reaction Network Generation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 894–908.

(49) Köbler, J.; Schöning, U.; Torán, J., *The Graph Isomorphism Problem: Its Structural Complexity*. Basel, CH: Birkhäuser, **1993**.

(50) Luks, E., Isomorphism of graphs of bounded valence can be tested in polynomial time. *J. Computer Syst. Sci* **1982**, *25*, 42–65.

(51) Weininger, D.; Weininger, A.; Weininger, J., SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comp. Sci.* **1989**, *29*, 97–101.

(52) Sauer, J., Diels-Alder Reactions Part II: The Reaction Mechanism. *Angew. Chem. Int. Ed.* **1967**, *6*, 16–33.

(53) Morgenroth, F.; Müllen, K., Dendritic and Hyperbranched Polyphenylenes via a simple Diels-Alder Route. *Tetrahedron* **1997**, *53*, 15349–15366.

(54) Butlerow, A., Formation synthétique d'une substance sucrée. *C. R. Acad. Sci.* **1861**, *53*, 145–147.

(55) Breslow, R., On the mechanism of the Formose reaction. *Tetrahedron Let.* **1959**, *21*, 22–26.

(56) Gabel, N. W.; Ponnamperuma, C., Model for origin of monosaccharides. *Nature* **1967**, *216*, 452–455.

(57) Decker, P.; Schweer, H.; Pohlmann, R., Bioids. X. Identification of formose sugars, presumably prebiotic metabolites, using capillary gas chromatography/gas chromatography-mass spectroscopy of *n*-butoxime trifluoroacetates on OV-225J. *J. Chromatography* **1982**, *225*, 281–291.

(58) Cramer, C. J.; Truhlar, D. G., Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* **1999**, *99*, 2161–2200.

(59) Wiener, H., Structure determination of paraffine boiling points. *J. Amer. Chem. Soc.* **1947**, *69*, 17–20.

(60) Gutman, I.; Klavžar, S.; Mohar, B., eds., *Fifty Year of the Wiener Index*, vol. 35 of *MATCH*, **1997**.

(61) Frenklach, M., Modeling of large reaction systems. In: Warnatz, J.; Jäger, W., eds., *Complex chemical reaction systems, mathematical modelling and simulation*, vol. 47 of *Springer Series in Chemical Physics*, Springer-Verlag, Berlin, **1987** 2–16.

(62) Radom, L., Structural Consequences of Hyperconjugation. *Prog. Theor. Org. Chem.* **1982**, *3*, 1–64.

(63) Pophristic, V.; Goodman, L., Hyperconjugation not steric repulsion leads to the staggered structure of ethane. *Nature* **2001**, *411*, 565–568.

(64) The GML Language. The `GML` language allows one to attribute arbitrary information to graphs, their nodes, and their edges. It can therefore be used to emulate almost every other data format. `http://infosun.fmi.uni-passau.de/Graphlet/GML/`.