The *ToyChem* Package:

A Computational Toolkit Implementing a Realistic Artificial Chemistry Model[0]

Gil Benkö[a,b], Christoph Flamm[c], and Peter F. Stadler[a,d,e]

[a]Bioinformatics Group, Department of Computer Science, University of Leipzig
[b]Graduiertenkolleg Wissensrepräsentation, Department of Computer Science, University of Leipzig
[c]Institute of Theoretical Chemistry and Structural Biology, University of Vienna, Austria
[d]Interdisciplinary Center for Bioinformatics, University of Leipzig
[e]Santa Fe Institute, Santa Fe, NM, USA

Most models of artificial chemistries are far away from the "look-and-feel" of a real-world chemistry. Usually, abstract algebraic entities are used that do not lend themselves to a natural definition of reaction enthalpies or to the incorporation of the crucial conservation properties of mass and atom types. In this short contribution we describe an improved version of an artificial chemistry model that stays close enough to a quantum-chemical description to be recognizable as an approximation to organic chemistry while at the same time allowing for a computationally efficient implementation that makes large-scale simulation feasible. Molecules are represented by their molecular graphs whose energy is defined via a simplified Extended Hückel Theory approach based on the orbital graphs. The model is implemented as an `Ansi C++` library, as a stand alone executable, and as a simple webserver. The software is distributed free of charge under the GNU Public License.

1. Introduction

A distinguishing feature of chemistry is the fact that, upon interaction, molecules not only change quantitative physical properties such as free energy or density. Indeed, the possibility to generate novel molecules as a crucial defining feature of Chemistry. It can thus be thought of as an algebraic system defined by the set of possible chemical reactions. The set

[0]Dedicated to Edward C. Kirby in celebration of his 70th birthday

of conceivable molecules is infinite in principle, and very large in practice with currently more than 25 million [1] described compounds. Thus the mere tabulation of all possible molecules and all their reactions is infeasible. Instead, an explicit representation of molecules and their interactions, i.e., a *algebraic* computational model of chemistry, is required 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 which inevitably is subject to sampling biases. The fact that most chemical databases are not freely accessible adds to the attractivity of artificial chemistry models. For a recent review of artificial chemistries we refer to Ref. [2].

Several approaches towards designing artificial chemistries have been explored in recent years. The spectrum ranges from chemically accurate quantum mechanical simulations to abstract computational models. The required level of realism depends of course on the purpose of the model. Walter Fontana's `AlChemy` [3, 4], 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. Models in a similar spirit have been implemented using a variety of computational paradigms from strings, to matrices, to Turing machines and graphs [5, 6, 7, 8, 9, 10]. Abstract computational models are very useful for understanding the purely algebraic properties of reaction systems; indeed, important concepts in chemical self-organization, such as the notion of self-maintaining sets and chemical organizations [4, 11, 8, 12] have arisen from the study of such abstract models. On the other hand, these models lack a natural definition of an energy function and, in most cases, there is no natural analogue to the conservation of mass and atom types.

In a previous study we have introduced a *Toy Chemistry* [13] that retains the look-and-feel of chemistry by representing molecules as vertex and edge-labeled graphs, whose vertex labels correspond to atom types and whose edge-labels can be interpreted as bond types. A minimal version of quantum mechanics is employed to define an energy function that determines thermodynamical equilibria in this universe. In this short contribution we describe a software library and associated tools that make an expanded implementation of the *Toy Chemistry* freely accessible to the scientific community.

2. The Model

Molecules are represented by means of their chemical graphs. Their properties within the Toy Universe are completely described by the *orbital graph* [14]: the vertices are the atom orbitals (labeled by atom type and hybridization, Fig. 1) and the edges denote overlaps of interacting orbitals. The orbital graph, Fig. 2, is obtained in an unambiguous way from the chemical

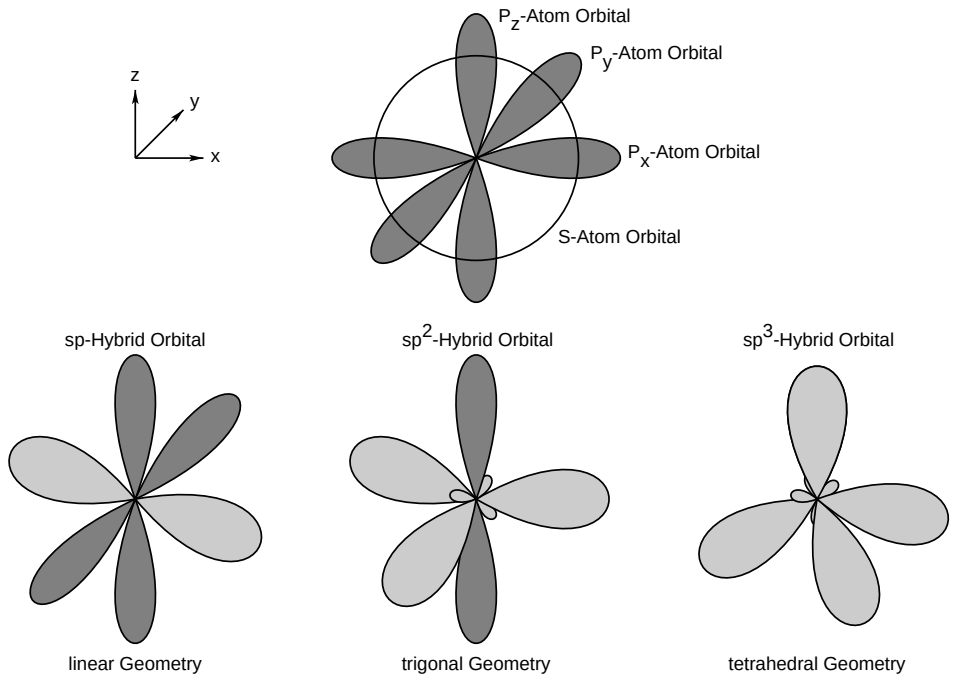

FIGURE 1. Hybrid Orbitals. The four atom orbitals $2p_x$, $2p_y$, $2p_z$ (dark gray dumbbells) and $2s$ (unfilled circle) can form three distinct hybrid orbitals sp, sp^2 and sp^3 with completely different geometry. Triple bonds are formed by sp-, double bonds by sp^2- and single bonds by sp^3 hybridized atoms.

structure formula by means of the VSEPR rules [15]. The energy calculation used in our Toy Model is an extreme simplification of the Extended Hückel Theory (EHT) [16]. Starting from a basis set of atomic orbitals (AOs) $\{\chi_i\}$ we expand the molecular orbital (MO) in the form

$$\Psi_\alpha = \sum_i c_{\alpha,i}\chi i \tag{1}$$

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 as $H_{ij} = \int \chi_i \widehat{H} \chi_j d\tau$ and $S_{ij} = \int \chi_i \chi_j d\tau$. In this setting the Schrödinger equation can be written in terms of the coefficients $c_{\alpha,i}$ as

$$\mathbf{H}\vec{c}_\alpha = E_\alpha \mathbf{S}\vec{c}_\alpha, \tag{2}$$

where $\vec{c}_\alpha$ denotes the vector of coefficients $c_{\alpha,i}$ belonging to the molecular orbital Ψ_α with orbital energy E_α. We assume that the vectors $\vec{c}_k$ are normalized.

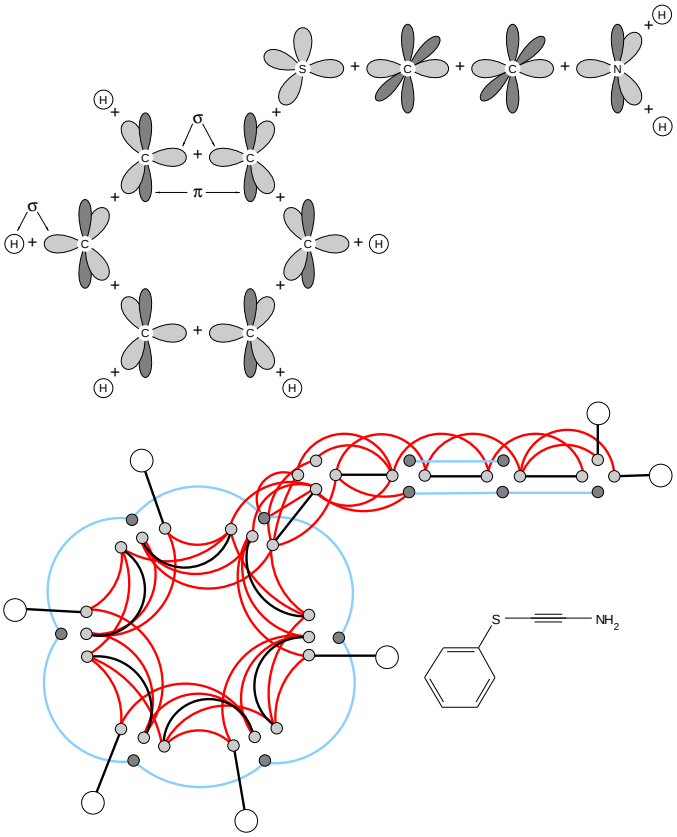

FIGURE 2. Orbital Graph. The upper part of the figure shows the decomposition of the structural formula (insert lower right) according to the VSEPR rules into hybrid orbitals. The lower part shows the orbital graph of the same molecule which is used for the energy calculation within the `ToyChem` Model. The color code is as follows: (i) Nodes indicate orbitals (atom labels have been omitted): p (dark gray), sp^n (light gray) and s (white) (ii) Arcs indicate overlaps between orbitals: σ-bonds (black), semi-direct, hyperconjugation, and fictitious interactions [13](red) and π-bonds (light blue).

EHT uses the Wolfsberg-Helmholtz approximation [17]

$$H_{ij} = \kappa(H_{ii} + H_{jj})S_{ij}/2 \tag{3}$$

to parametrize the Hamilton matrix in terms of the *atomic valence state ionization potentials* $H_{ii} = -I_i$ and the overlap integrals S_{ij} between any two orbitals. Within our model we tabulate S_{ij} and I_i as function of the atom and orbital types. Our implementation uses the $1s$ orbital for hydrogen and the usual Slater-type hybrid AOs (sp^3, sp^2, and sp) for carbon, nitrogen,

oxygen, phosphorus, and sulfur. 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. The following bond types are taken into account: σ-bonds, π-bonds, backbonding and hyperconjugation through indirect sp^n/sp^n and sp^n/p interactions, "banana" overlaps in short rings, as well as stronger backbonding for phosphorus and sulfur atoms. The weaker overlap in hydrogen bonds or three-center bonds is also taken into account too. For details we refer to Ref. [13].

The physical properties of a molecule are determined by the eigenvalues E_α and their associated eigenvectors $\vec{c}_\alpha$ of $\mathbf{H}$ and the numbers n_α of electrons in the MO Ψ_α and the numbers z_a of valence electrons of atom a. For example, the total electronic energy of the molecule is $E = \sum_\alpha n_\alpha E_\alpha$, and the electronic population in the atom orbital i is given by $q_i = \sum_\alpha n_\alpha c_{\alpha,i}^2$. The *charge density* at atom a can then be obtained as $q(a) = z_a - \sum_{i@a} q_i$ where the sum runs over all orbitals i located at atom a.

3. Features and Limitations

The library `ToyChemLib` provides a complete implementation of the model described in the previous section.

- Neutral and charged molecules as well as radicals can be evaluated.
- Parameter values are currently implemented for H, C, N, O, P and S. Additional atoms can easily be added by extending the parameter file without changes in the software itself.
- The spectrum of molecular orbitals, charges and total energies are computed.
- A spectral embedding procedure [18] is used to compute dipole moments.
- A simple model for computing solvation energies can be used to simulate environments with multiple phases.

The `ToyChemLib` package addresses mostly researchers in Artificial Chemistry and is designed to make it easy for users to implement their own simulations based upon the Toy Chemistry universe. To this end we provide `Perl` bindings to the functions implemented in `ToyChemLib`. This is convenient e.g. for writing `cgi`-scripts such as the web server that is included as a programming example with the `ToyChemLib` distribution.

We also provide an interactive program `ToyChemEnergy` that can be used to compute the spectrum $\{E_\alpha\}$, the charge densities $q(a)$, the total atomization energy E, as well as a dipole moment [18]. In addition, the orbital graph can be output in `GML` format [19]. The program takes a structure formula in the widely used `SMILES` format [20]. This tool is intended as simple way to become familiar with the model without the need to write programs.

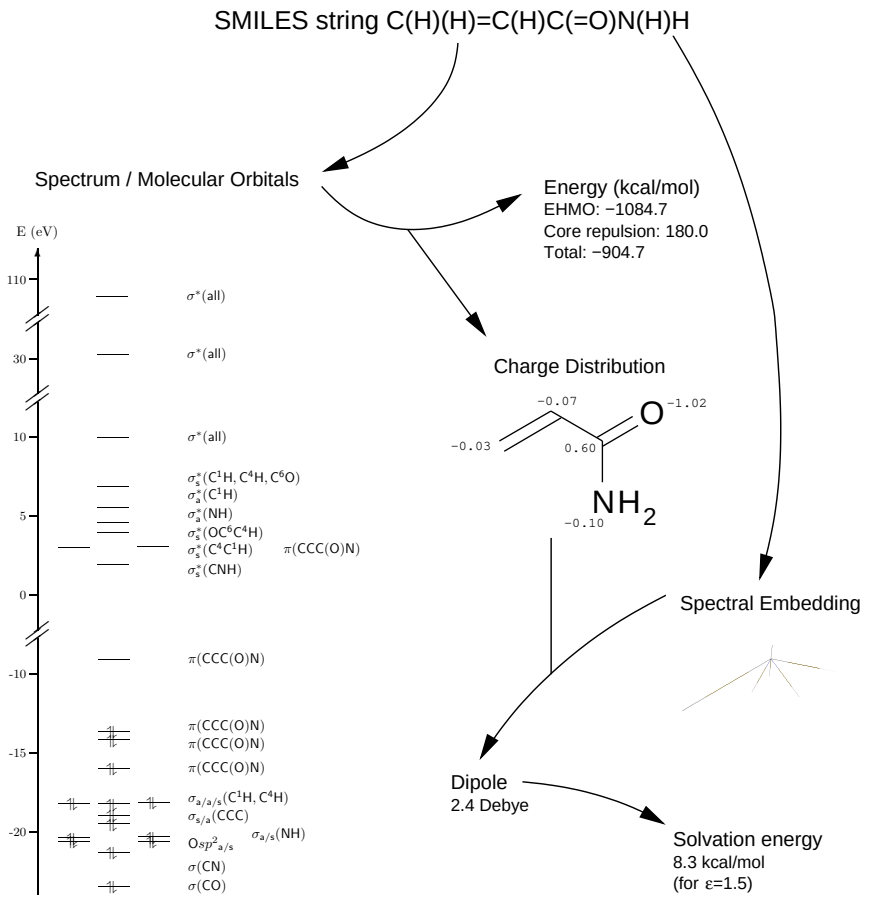

FIGURE 3. Example application of `ToyChemEnergy`. The `SMILES` string of propenamide is used as input. It can be used for an embedding using spectral graph drawing, or can be expanded to an orbital graph, which serves to calculate the spectrum/molecular orbitals. These in turn provide the energy, charge distribution, and, in combination with the spectral embedding, yield the dipole moment and a solvation energy.

An example for the functionalities of this interactive program is outlined in Fig. 3.

The library `ToyChemLib`, including the underlying graph library `GraphLib`[21], as well as the executables and `Perl` bindings are distributed free of charge under the GNU Public License.

Our model is limited by its restriction to the connectivity of the molecules, thereby neglecting most of the steric information. Chirality, in particular, is

a property of the three-dimensional embedding of the molecular graph which is not represented in the graph itself. While chirality is of particular interest e.g. in the field of prebiotic chemistry, it is by design excluded from our model universe. An extension of the current implementation in which the edges at each vertex/atom are considered to be cyclically ordered, however, could be used to represent chiral molecules and their reactions. Such an approach would also distinguish E/Z-isomers for each other. Appropriate rules for the behavior of the cyclical ordering of bonds could be used to implement the stereochemistry of reactions such as the Walden inversion.

4. Applications

We have used this method as a part of a chemical reaction network simulation. An implementation of chemical reactions and kinetics was added. Using the Klopman-Salem equation [22, 23], the reactivities of molecules can be derived from their wave functions. This then allowed us to perform predictions on regioselectivity and to automatically generate the system of differential equations for the time evolution of the chemical reaction network in order to study its kinetic properties. Thus we were able to generate chemical reaction networks and study their generic graph-theoretic properties [24].

It has been observed in large networks, like the internet, or regulatory networks, that the vertex degree distribution follows often a power-law ("scale-free"), or that the average shortest path length is relatively short ("small-worlds"). Using our model, we showed that chemical reaction networks do not fall generically into the same class of networks. Some of the networks studied were small-world and scale-free and some failed to show those properties.

Using a simple model for the solvation energy (which is part of the implementation described here), we extended the model of chemical reaction networks by simulating multiple phases [18]. The difference of the phases was determined by their difference in permittivity and thus capability of keeping different molecules in solution. We studied for instance the importance of phase barriers in prebiotic chemistry. Not only the connectivity of the prebiotic reaction network model was changed, but also its kinetic properties were modified as a consequence of a different chemical composition. For example, certain combinations of compounds cannot coexist with noticeable concentration in a single phase because they would immediately react with each other.

Ongoing work involves using our toy model in the study of chemical reactions and in testing scenarios of the origin of metabolism. The model aims to provide a chemically coherent method for researchers in the field of artificial life and the origin of life. Prebiotic scenarios are often modeled using arbitrary parameters and profit from incorporating thermodynamics. They can thus be tested and expanded.

The molecular property calculation can also be useful for the generation of descriptors for QSAR, QSPR, QSMR, (Quantitative Structure-Activity/Product/Metabolism Relationship) or drug data mining. The topological pharmacophore descriptors[25], for example, need only the connectivity of a compound and its electronic properties and are easily calculated within our model. The transparency of the calculation makes it adaptable and has also a certain of didactic value: The setup of the overlap and Hamilton matrix is straightforward. It illustrates the importance of the choice of the basis set, through the almost block-diagonal matrices, and thus explains also the origin of the energy increments in thermochemistry.

Availability. The *ToyChem* Package is entirely open-source and all specifications and source code are freely and publicly available from the `URL`s

`http://www.tbi.univie.ac.at/~xtof/ToyChem/`
`http://www.bioinf.uni-leipzig.de/~gil/ToyChem/`

A `ToyChemEnergy` web service which calculates for an organic molecule provided as `SMILES` string various physico-chemical properties can be found at the `URL`s

`http://www.tbi.univie.ac.at/cgi-bin/ToyChemEnergy.cgi`
`http://www.bioinf.uni-leipzig.de/cgi-bin/gil/getE.cgi`

Acknowledgments. This work was performed under the auspices of the COST Action D27. Partial financial support by the Bioinformatics Initiative of DFG, grant no. BIZ-6/1-2.

References

1 Chemical Abstracts Service. CAS registry (2005). URL `http://www.cas.org/cgi-bin/regreport.pl`. The `CAS` Registry provides the largest substance identification system in existence.
2 P. Dittrich, J. Ziegler, and W. Banzhaf. *Artificial Life* **7** (2001) 225–275.
3 W. Fontana. In: C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen (Eds.), *Artificial Life II*. Addison-Wesley, Redwood City, CA, pp. 159–210.
4 W. Fontana and L. W. Buss. *Proc. Natl. Acad. Sci. USA* **91** (1994) 757–761.
5 R. J. Bagley and J. D. Farmer. In: C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen (Eds.), *Artificial Life II*. Addison-Wesley, Redwood City, CA, Santa Fe Institute Studies in the Sciences of Complexity, pp. 93–141.
6 W. Banzhaf, P. Dittrich, and B. Eller. *Physica D* **125** (1999) 85–104.
7 J. S. McCaskill and U. Niemann. In: A. Condon and G. Rozenberg (Eds.), *DNA Computing*, Springer, Berlin, D, vol. 2054 of *Lecture Notes in Computer Science*. pp. 103–116.
8 P. Speroni di Fenizio. In: M. Bedau, J. McCaskill, N. Packard, and S. Rasmussen (Eds.), *Artificial Life VII*. MIT Press, Cambridge, MA, pp. 49–53.
9 M. Thürk. *Ein Modell zur Selbstorganisation von Automatenalgorithmen zum Studium molekularer Evolution*. Ph.D. thesis, Universität Jena, Germany (1993). PhD Thesis.
10 I. Ugi, N. Stein, M. Knauer, B. Gruber, K. Bley, and R. Weidinger. *Top. Curr. Chem.* **166** (1993) 199–233.
11 W. Fontana and L. W. Buss. *Bull. Math. Biol.* **56** (1994) 1–64.
12 P. F. Stadler, W. Fontana, and J. H. Miller. *Physica D* **63** (1993) 378–392.

13 G. Benkö, C. Flamm, and P. F. Stadler. *J. Chem. Inf. Comput. Sci.* **43** (2003) 1085–1093.
14 O. E. Polansky. *MATCH* **1** (1975) 183–195.
15 R. J. Gillespie and R. S. Nyholm. *Quart. Rev. Chem. Soc.* **11** (1957) 339–380.
16 R. Hoffmann. *J. Chem. Phys.* **39** (1963) 1397–1412.
17 M. Wolfsberg and L. Helmholtz. *J. Chem. Phys.* **20** (1952) 837–843.
18 G. Benkö, C. Flamm, and P. F. Stadler. In: H. Schaub, F. Detje, and U. Brüggemann (Eds.), *The Logic of Artificial Life: Abstracting and Synthesizing the Principles of Living Systems*. IOS Press, Akademische Verlagsgesellschaft, Berlin, pp. 16–22. Proceedings of GWAL, Bamberg 14-16 April 2004.
19 The GML language. URL `http://infosun.fmi.uni-passau.de/Graphlet/GML/`. 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.
20 D. Weininger. *J. Chem. Inf. Comput. Sci.* **28** (1988) 31–36.
21 P. M. Gleiss. *Short Cycles*. Ph.D. thesis, University of Vienna (2001).
22 G. Klopman. *J. Am. Chem. Soc.* **90** (1968) 223–243.
23 L. Salem. *J. Am. Chem. Soc.* **90** (1968) 543–552 & 553–566.
24 G. Benkö, C. Flamm, and P. F. Stadler. In: W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim, and J. Ziegler (Eds.), *Advances in Artificial Life*. Springer-Verlag, Heidelberg, Germany, vol. 2801 of *Lecture Notes in Computer Science*, pp. 10–20. Proceedings of the 7th European Conference of Artifical Life, ECAL 2003, Dortmund, Germany, September 14-17, 2003, Proceedings.
25 N. A. Kratochwil, W. Huber, F. Müller, M. Kansy, and P. R. Gerber. *Bioch. Pharm.* **64** (2002) 1355–1374.