

# Atom Mapping with Constraint Programming

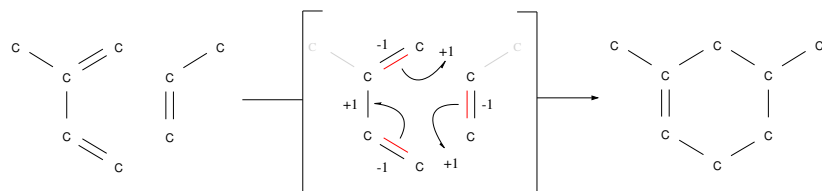
Martin Mann<sup>1</sup>, Feras Nahar<sup>1</sup>, Heinz Ekker<sup>5</sup>, Rolf Backofen<sup>1-4</sup>, Peter F. Stadler<sup>5-9</sup>, and Christoph Flamm<sup>5</sup>

<sup>1</sup>Bioinformatics, Department for Computer Science, University of Freiburg, George-Köhler-Allee 106, 79106 Freiburg, Germany, <sup>2</sup>Centre for Biological Signalling Studies (BIOSS), University of Freiburg, Germany, <sup>3</sup>Centre for Biological Systems Analysis (ZBSA), University of Freiburg, Germany, <sup>4</sup>Center for non-coding RNA in Technology and Health, University of Copenhagen, Denmark, <sup>5</sup>Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, 1090 Vienna, Austria, <sup>6</sup>Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany, <sup>7</sup>Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany, <sup>8</sup>Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, 04103 Leipzig, Germany, and <sup>9</sup>Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA  
{mmann,studla,xtof}@tbi.univie.ac.at

**Abstract.** Chemical reactions consist of a rearrangement of bonds so that each atom in an educt molecule appears again in a specific position of a reaction product. In general this bijection between educt and product atoms is not reported by chemical reaction databases, leaving the Atom Mapping Problem as an important computational task for many practical applications in computational chemistry and systems biology. Elementary chemical reactions feature a cyclic imaginary transition state (ITS) that imposes additional restrictions on the bijection between educt and product atoms that are not taken into account by previous approaches. We demonstrate that Constraint Programming is well-suited to solving the Atom Mapping Problem in this setting. The performance of our approach is evaluated for a subset of chemical reactions from the KEGG database featuring various ITS cycle layouts and reaction mechanisms.

## 1 Introduction

A chemical reaction describes the transformation of a set of educt molecules into a set of products. In this process, chemical bonds are re-arranged, while the atom types remain unchanged. Thus, there is a one-to-one correspondence, the so-called *atom map* (or atom-atom mapping), between atoms in educts and products. Atom maps convey the complete information necessary to disentangle the mechanism, i.e. the bond re-arrangement, of a chemical reaction via the identification of bonds that differ in educt and product molecules. The changing parts of the molecules are described by a so called intermediate transition state (ITS) [17, 24] that allows, for instance, a classification of chemical reactions [31, 33, 45]. Atom maps are a necessary requisite for computational studies of an



**Fig. 1.** Example of a Diels-Alder reaction. The imaginary transition state (ITS) is an alternating cycle defined by the bonds that are broken (in red) and the bonds that are newly formed.

organisms metabolism. For instance, they allow for consistency checks within metabolic pathway analyses [3] and play a role in the global analysis of metabolic networks [7, 26]. Practical applications include, for example, the tracing or design of the metabolic breakdown of a candidate drug, which constitutes an important issue in drug design studies [39].

For chemical reactions only the product and educt molecules are directly observable. The atom map therefore often remains unknown and has to be inferred from partial knowledge. Experimental evidence may be available from isotope labeling experiments. Here, special isotopes, i.e. atoms with special variations, are introduced into educt molecules that can then be identified in product molecules by means of spectroscopy techniques [44]. Such data, however, is not available for most reactions. The complete experimental determination of an atom map is in general a complex and tedious endeavor. Reaction databases, such as KEGG, therefore do not generally supply atom maps. The computational construction of atom maps is therefore an important practical problem in cheminformatics.

Several computational approaches for this problem have been developed over the last three decades (for a recent review see [8]). The educts and products are described as two not necessarily connected labeled graphs  $I$  and  $O$ , respectively. Vertex labels define atom types, while edge labels indicate bond types. The atom map is then determined as the solution of a combinatorial optimization problem resulting in a bijective mapping of all vertices of the educt molecule graph to corresponding vertices in the product molecule graphs. An illustration is given in Fig. 1.

The most common formulations are variants of the maximum common subgraph (isomorphism) problem [15]. Already the earliest approaches analyzed the adjacency information within educts and products [14, 34]. The Principle of Minimal Chemical Distance, which is equivalent to minimizing an edge edit distance, was invoked in [28], using a branch and bound approach to solve the corresponding combinatorial optimization problem. Maximum Common Edge Subgraph (MCES) algorithms search for isomorphic subgraphs of the educt/product graphs with maximum number of edges [13, 22, 23, 33, 40], an NP-hard problem. Furthermore, the use of specialized energetic [2, 30] or weighting [32] criteria allows for the identification of the static parts of the reaction and, subsequently, of

the atom mapping. A detailed investigation of the MCES from an Integer Linear Programming (ILP) perspective can be found in [6].

Akutsu [1] showed that the MCES approach fails for certain reactions. As an alternative, the Maximum Common Induced Subgraph (MCIS) problem was proposed as a remedy. This problem is also NP complete. Approximation results can be found in [27]. Algorithms for the MCIS iteratively decompose the molecules until only isomorphic sub-graphs remain [1, 7, 11, 12]. Recently, an ILP approach incorporating stereochemistry was presented [16].

Neither the solutions of the MCES nor the MCIS necessarily describe the true atom map. Indeed, both optimality criteria are artificial and can not be derived from basic principles of chemical reactions. In fact, it is not hard to construct counter-examples, i.e., chemical reactions whose true atom maps are neither identified by MCES nor by MCIS. The re-organization of chemical bonds in a chemical reaction is far from arbitrary but follows strict rules that are codified e.g. in the theory of imaginary transition states (ITS) [17, 24]. The ITS encodes the redistribution of bond electrons that occurs along a chemical reaction. Bond electrons define the atom-connecting chemical bonds and their according bond orders. Their redistribution is expressed in terms of the deletion or formation of bonds as well as changes of in the oxidation state of atoms, the latter resulting from non-bound electrons that are freed from or integrated into bonds. The ITS can be used to cluster, classify, and annotate chemical reactions [17, 24, 25]. These studies revealed, that only a limited number of ITS “layouts” are found among single step reactions and that these layouts represent a cyclic electron redistribution pattern usually involving less than 10 atoms [25]. In a most basic case, an elementary reaction, the broken and newly formed bonds form an alternating cycle (see Fig. 1) covering a limited even number of atoms [18], usually less than 8 [24]. In the case of homovalent reactions, i.e., those in which the number of non-bound electron pairs of all atoms (defining their oxidation state) remains unchanged, this cycle is elementary. That is, the transition state is a single, connected even cycle, along which bond orders change by  $\pm 1$  [25]. This property imposes an additional, strong condition of the atom maps that is not captured by the optimization approaches outlined in the previous paragraphs. Here, we explicitly include it into the specification of the combinatorial problem.

A *chemically correct* atom map is a bijective map between the vertices of the educt and product graphs such that:

1. The map preserves atom types
2. The total bond orders (including lone electron pairs) are preserved. Each broken bond thus must be compensated by a newly formed bond or a change in the oxidation number of an atom.
3. The broken and newly formed bonds constitute a chemically reasonable intermediate transition state (ITS) following [25]. In the case of elementary chemical reactions, the transition state is an alternating cycle.

A formal definition of the combinatorial problem will be given in the following section. While cyclic transition states are very common, more “complex transition states” appear in non-elementary reactions, i.e., compositions of elemen-

tary reactions. Furthermore, even in elementary reaction, it is not true that the shortest ITS cycle is necessarily chemically correct. Empirically, transition states are most frequently six-membered cycles, while cycles of length 4 or 8 are less abundant [17–19, 24]. As a consequence, we will consider several variants of the chemical reaction mapping problem:

1. **Decision problem:** Is there an atom map with cyclic ITS? Of course one may restrict the question to ITS cycles of length  $k$ .
2. **Optimization problem:** Find the minimal length  $k$  of an ITS cycle that enables an atom map.
3. **Enumeration problem:** Find all atom maps with cyclic ITS (of length  $k$ ).

Given a straightforward encoding of molecular graphs in terms of vertex indices, atom labels, and adjacency information, the atom mapping problem is naturally open to be treated as a constraint satisfaction problem with finite integer domains. This approach is particularly appealing when additional information on the ITS, e.g. its size or atoms involved in the ITS, are known.

## 2 Constraint Programming Formulation of the Atom Mapping Problem

We focus on the identification of the cyclic ITS. Once the ITS has been identified the overall atom mapping is easily derived. We formulate separate constraint satisfaction problems for different ITS layouts and cycle lengths. A fast graph matching approach is used subsequently to extend each ITS to a global atom mapping. In this section we follow closely [36]. We first formally define the problem, which is followed by a description of our constraint programming approach for identifying the cyclic ITS. Finally we discuss how to extend an ITS candidate to a complete atom mapping for the chemical reaction.

### 2.1 Problem Definition

Both educts and products of a chemical reaction are each represented by a single, not necessarily connected, undirected graph defined by a set of vertices  $V$  and a set of edges  $E = \{ \{v, v'\} \mid v, v' \in V \}$ . The educt (input) graph is denoted by  $I = (V_I, E_I)$  and the product (output) graph by  $O = (V_O, E_O)$ . Here, each molecule corresponds to a connected component. Vertices represent atoms and are labeled with the respective atom type accessible via the function  $l(v \in V_I \cup V_O)$ . The principle of mass conservation implies  $|V_I| = |V_O|$ , i.e. no atom can dissolve or appear during a reaction. Edges encode covalent chemical bonds between atoms. For the CSP formulation we label each edge  $\{x, y\} \in E_I \cup E_O$  with the number of shared electron pairs, i.e. its bond order: single, double or triple bonds are represented by a single edge with labels 1, 2, or 3, respectively. Non-bonding electron pairs of an atom, which define its oxidation state, are represented by loops labeled with the according number of unbound pairs.

We use an adjacency matrix  $\mathcal{I}$  to encode the edge labels of the educt graph (and a corresponding matrix  $\mathcal{O}$  for the products). The matrix elements  $\mathcal{I}_{v,v'}$  denote the number of shared bond electron pairs for the edge between the atoms  $v$  and  $v'$  in the educt graph  $I$ . In practice  $\mathcal{I}_{v,v'} \in \{0, 1, 2, 3\}$ . Non-bonding electron pairs (loops) are represented by the diagonal entries  $\mathcal{I}_{v,v}$  and  $\mathcal{O}_{v,v}$ .

Consider a bijective function  $m : V_I \rightarrow V_O$  mapping the vertices of  $I$  onto the vertices of  $O$  and a matrix  $\mathcal{Q}$  with rows and columns indexed by  $V_I$ . Then  $\mathcal{Q} \circ m$  is the matrix with entries  $\mathcal{Q}_{m(x),m(y)}$ , i.e. with rows and columns indexed by  $V_O$ . Thus the *reaction matrix*  $\mathcal{R}^m = \mathcal{O} - (\mathcal{I} \circ m)$  is well defined and encodes the bond electron differences between educt and product.

**Definition.** An *atom mapping* is a bijective mapping  $m : V_I \rightarrow V_O$  such that

1.  $\forall_{x \in V_I} : l(x) = l(m(x))$  (preservation of atom types)
2.  $\mathcal{R}^m \vec{1} = 0$  (preservation of bond electrons)

The reaction matrix  $\mathcal{R}^m$  encodes the imaginary transition state (ITS) [17, 24]. This definition of  $m$  is a slightly more formal version of the Dugundji-Ugi theory [14]. Our notation emphasizes the central role of the (not necessarily unique) bijection  $m$ . Since we consider  $I$  and  $O$  as given fixed input, the atom mapping  $m$  uniquely determines  $\mathcal{R}^m$ . The triple  $(m, I, O)$ , furthermore, completely defines the chemical reaction. It therefore makes sense to associate properties of the chemical reaction directly with the atom map  $m$ .

Equivalently, the ITS can be represented as a graph  $R = (V_R, E_R)$  so that  $E_R$  consists of the ‘‘changing’’ edges that lose or gain bond electrons during the reaction, i.e.  $\mathcal{I}_{v,v'} \neq \mathcal{O}_{m(v),m(v')} \rightarrow \mathcal{R}_{v,v'}^m \neq 0$ . The set of atom vertices  $V_R \subseteq V_O$  covers all vertices with at least one adjacent edge in  $E_R$ . Each edge  $\{v, v'\} \in E_R$  is labeled by the electron change  $\mathcal{R}_{v,v'}^m \neq 0$ , i.e. its change in bond order. See Fig. 2 for an example.

$\mathcal{I}$	$v_1$	$v_2$	$v_3$	$v_4$	$v_5$	$v_6$	$v_7$	$v_8$		$\mathcal{O}$	$v'_1$	$v'_2$	$v'_3$	$v'_4$	$v'_5$	$v'_6$	$v'_7$	$v'_8$		$\mathcal{R}^m$	$v'_1$	$v'_2$	$v'_3$	$v'_4$	$v'_5$	$v'_6$	$v'_7$	$v'_8$			
$v_1$	0	1	0	0	0	0	0	0		$v'_1$	0	1	0	0	0	0	0	0		$v'_1$	0	0	0	0	0	0	0	0	0		
$v_2$	1	0	1	2	0	0	0	0		$v'_2$	1	0	2	1	0	0	0	0		$v'_2$	0	0	+1	-1	0	0	0	0	0		
$v_3$	0	1	0	0	2	0	0	0		$v'_3$	0	2	0	0	1	0	0	0		$v'_3$	0	+1	0	0	-1	0	0	0	0	0	
$v_4$	0	2	0	0	0	0	0	0		$v'_4$	0	1	0	0	0	1	0	0		$v'_4$	0	-1	0	0	0	+1	0	0	0	0	
$v_5$	0	0	2	0	0	0	0	0		$v'_5$	0	0	1	0	0	0	1	0		$v'_5$	0	0	-1	0	0	0	+1	0	0	0	
$v_6$	0	0	0	0	0	0	2	1		$v'_6$	0	0	0	1	0	0	1	1		$v'_6$	0	0	0	+1	0	0	-1	0	0	0	
$v_7$	0	0	0	0	0	2	0	0		$v'_7$	0	0	0	0	1	1	0	0		$v'_7$	0	0	0	0	+1	-1	0	0	0	0	
$v_8$	0	0	0	0	0	1	0	0		$v'_8$	0	0	0	0	0	1	0	0		$v'_8$	0	0	0	0	0	0	0	0	0	0	0

**Fig. 2.** Adjacency matrices  $\mathcal{I}$  for the reaction given in Fig. 1. The vertices  $v_i \in V_I$  and  $v'_j \in V_O$  are numbered in top-down-left-right order of their appearance in Fig. 1. The atom mapping  $m(v_i) = v'_i$  defines  $\mathcal{R}^m$  and thus the ITS graph  $R$  covers only vertices  $v'_2$  to  $v'_7$  since  $v'_1$  and  $v'_8$  do not show any bond electron changes.

It is important to note that the existence of an atom mapping  $m$  as defined above does not necessarily imply that  $\mathcal{R}^m$  is a chemically plausible ITS.

We say that two edges  $\{v, v'\}, \{v', v''\} \in E_R$  in  $R$  are *alternating* if  $\mathcal{R}_{v,v'}^m \neq 0$  and  $\mathcal{R}_{v',v''}^m + \mathcal{R}_{v',v''}^m = 0$ . A *simple cycle* in  $R$  of size  $k > 2$  is given by the vertex sequence  $(v_1, v_2, \dots, v_k, v_1)$  with  $v_i \in V_R$ ,  $\{v_i, v_{i+1}\} \in E_R$ ,  $\{v_k, v_1\} \in E_R$ , and

$\forall i < j \leq k : v_i \neq v_j$ . Such a simple cycle is called alternating if all successive edges as well as the cycle closure  $\{v_2, v_1\}, \{v_1, v_k\}$  are alternating.

**Definition.** An atom map  $m$  is *homovalent* if  $\mathcal{R}_{v,v}^m = 0$  for all  $v \in V_R$ . A homovalent reaction is *elementary* if its ITS  $R$  is a simple alternating cycle. Thus  $\mathcal{R}_{v,v}^m \in \{-1, 0, +1\}$  holds for all elementary homovalent reactions.

In the following we outline a novel algorithm for finding atom maps for a given ITS graph  $R$  that is guaranteed to retrieve all possible mappings given the educt and product graphs  $\mathcal{I}$  and  $\mathcal{O}$ , respectively. To simplify the presentation, first only elementary homovalent reactions are considered. Generalizations are discussed in Sec. 3.

## 2.2 Constraint Programming Approach

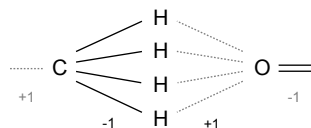
The central problem to find an elementary homovalent atom mapping is to identify the alternating cycle defining the ITS  $R$  given the adjacency information of the educts  $\mathcal{I}$  and products  $\mathcal{O}$ . This can be done via solving the Constraint Satisfaction Problem (CSP) as presented below. Note, due to the alternating edge condition within the ITS, we have to consider cycles with an even number of atoms only. In practice, the ITS of elementary homovalent reactions involves  $|V_R| = 4, 6, \text{ or } 8$  atoms [18].

**Basic CSP Formulation:** In the following, we will present a first basic CSP for an ITS of size  $k = |V_R|$  that we already introduced in [36]. It is given by the triple  $(X, D, C)$  defining the set of variables  $X$ , according domains  $D_i$ , and the set of constraints  $C$  to be fulfilled by any solution.

We construct an explicit encoding of the ITS atom mapping using  $k$  variables representing the cycle in  $I$  and another set for the mapped vertices in  $O$ , i.e.,  $X = \{X_1^I, \dots, X_k^I\} \cup \{X_1^O, \dots, X_k^O\}$  with domains  $D_i^I = V_I$  and  $D_i^O = V_O$ . Note, we do *not* directly encode the overall atom mapping problem but the identification of the two ITS subgraphs in the educts and products. Given this information, the overall atom mapping is easily identified as explained later.

To find a bijective mapping we have to ensure  $\forall i \neq j : X_i^I \neq X_j^I$  and  $\forall i \neq j : X_i^O \neq X_j^O$ , i.e., a distinct assignment of all variables. To enforce atom label preservation we need arc consistency for  $l(X_i^I) = l(X_i^O)$ , i.e. we have to enforce  $\forall e \in D_i^I : \exists p \in D_i^O : l(e) = l(p)$  as well as  $\forall p \in D_i^O : \exists e \in D_i^I : l(p) = l(e)$ . Analogously, homovalence is represented by  $(\mathcal{I}_{X_i^I, X_i^I} - \mathcal{O}_{X_i^O, X_i^O}) = 0$ . Due to the alternating bond condition, each atom can lose or gain at most one edge during a reaction. Thus, we can further constrain the variables with  $|\text{degree}(X_i^I) - \text{degree}(X_i^O)| \leq 1$ ; where  $\text{degree}(v)$  gives the out-degree of vertex  $v$ .

Finally, we have to encode the alternating cycle structure of the ITS in the mapping, i.e., for the sequence of bonds with indices 1-2-...- $k$ -1. For all index pairs within the cycle  $(i, j)$  we therefore require pairs with even index  $i$  to correspond the formation of a bond, i.e., we enforce  $(\mathcal{O}_{X_i^O, X_j^O} - \mathcal{I}_{X_i^I, X_j^I}) = 1$ , while all odd indices  $i$  are bond breaking  $(\mathcal{O}_{X_i^O, X_j^O} - \mathcal{I}_{X_i^I, X_j^I}) = -1$  accordingly.



**Fig. 3.** Symmetries resulting from interchangeable hydrogens. The figure presents three successive atom assignments within an ITS mapping. Bonds present in  $I$  are given in black, bonds to be formed to derive  $O$  are dotted and gray. The ITS describes the loss of an hydrogen for the carbon (bond order decrease) and the bond formation between the decoupled hydrogen with the oxygen next in the ITS. It becomes clear that all 4 hydrogens are not distinguishable, which results in 4 possible symmetric ITS mappings.

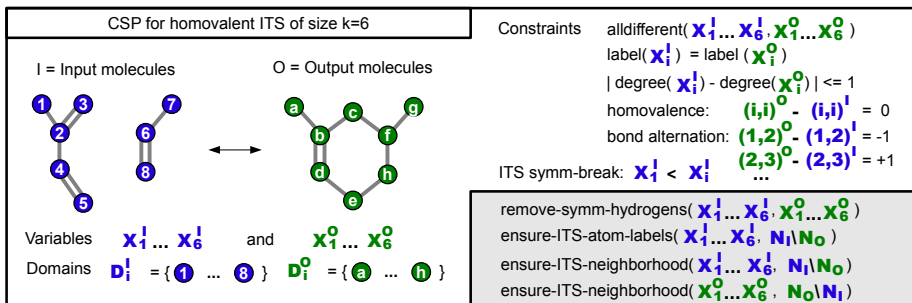
The homovalent ITS layout is rotation symmetric in itself (see Fig. 6). To partially counter this, we introduce order constraints on the input variables:  $(\forall i > 1 : X_1^I < X_i^I)$ ; where  $X_i < X_j$  denotes  $\exists(x, y) \in D_i \times D_j : x < y$  using e.g. an index order on the vertices. This ties the smallest cycle vertex to the first variable  $X_1^I$  and prevents the rotation-symmetric assignments of the input variables. Note, since we constrain the bond  $(1, 2)$  to be a bond breaking ( $\mathcal{O}_{X_1^O, X_2^O} - \mathcal{I}_{X_1^I, X_2^I} = -1$ ), the direction of the cycle is fixed and all direction symmetries are excluded as well.

As we will show in the evaluation (Sec. 3), the basic CSP will produce many ITS candidates that do not enable an atom mapping over the whole educt and product graphs. Therefore, we introduce an extended version of this CSP that incorporates further constraints derived from the input.

**Extended CSP Formulation:** Investigating the given educt and product graph, we can exclude a large set of symmetric solutions that arise due to an exchange of hydrogens. The latter can form at most one single bond to other atoms. Thus, if a hydrogen participates in the ITS, its adjacent atom will do as well (since the bond is to be broken in the ITS). Most adjacent atoms are non-hydrogens, like carbon atoms, that can have multiple adjacent hydrogens. Since there is exactly one bond breaking and formation for each ITS atom, only one such adjacent hydrogen will be part of the ITS. This results in a combinatorial explosion due to the symmetries of adjacent hydrogen atoms. An example is given in Fig. 3.

To break this type of symmetry, we select for each non-hydrogen one adjacent “master” hydrogen and remove all other sibling hydrogens from the domains, both for educt and product variables  $X^I$  and  $X^O$ , respectively.

Furthermore, we can extend and tune the CSP formulation by comparing the graph structure of educts and products. To this end, we generate the sets  $N_I$  and  $N_O$  of local neighborhoods of all atoms (vertices) for the educt and product



**Fig. 4.** Overview of the extended CSP for a homovalent ITS of size  $k = 6$  where the extensions of the basic CSP are given in the gray box in the lower right.

graph, resp., given by

$$N_I = \{ N(v) \mid v \in V_I \} \text{ with} \quad (1)$$

$$N(v) = ( l(v), \{ \mathcal{I}_{v,v'} \oplus l(v') \mid \text{where } v \neq v' \in V_I \wedge \mathcal{I}_{v,v'} > 0 \} ) \quad (2)$$

where  $N(v)$  is a tuple of the label of atom vertex  $v$  and an encoding of the set of all adjacent edges for this vertex. Note,  $\oplus$  denotes string concatenation.  $N_O$  is derived accordingly. For example, the neighborhood sets for the reaction from Fig. 1 are

$$N_I = \{ 2 \times (\mathbf{C}, \{1\mathbf{C}\}), 3 \times (\mathbf{C}, \{2\mathbf{C}\}), 2 \times (\mathbf{C}, \{1\mathbf{C}, 2\mathbf{C}\}), (\mathbf{C}, \{1\mathbf{C}, 1\mathbf{C}, 2\mathbf{C}\}) \}$$

$$N_O = \{ 2 \times (\mathbf{C}, \{1\mathbf{C}\}), 3 \times (\mathbf{C}, \{1\mathbf{C}, 1\mathbf{C}\}), (\mathbf{C}, \{1\mathbf{C}, 2\mathbf{C}\}), (\mathbf{C}, \{1\mathbf{C}, 1\mathbf{C}, 1\mathbf{C}\}),$$

$$(\mathbf{C}, \{1\mathbf{C}, 1\mathbf{C}, 2\mathbf{C}\}) \}$$

The subtraction  $N_I \setminus N_O$  gives the local neighborhoods that are unique within the educts and thus are part of the ITS, i.e. have to be changed during the reaction. Therefore, we can derive a lower bound on the number of atoms of a certain type that are participating in the ITS. In the example this results in  $N_I \setminus N_O = \{ 3 \times (\mathbf{C}, \{2\mathbf{C}\}), (\mathbf{C}, \{1\mathbf{C}, 2\mathbf{C}\}) \}$  revealing that at least 4 C-atoms of two types are ITS members.

Given this information, we formulate an extended version of the basic CSP. An arc-consistent global constraint on  $X^I$  is added, which enforces the occurrence of the identified ITS atom labels. This is automatically propagated on  $X^O$  via the atom label preservation constraints. In addition, we enforce that a valid assignment of the variables  $X^I$  and  $X^O$  preserves the ITS neighborhoods  $N_I \setminus N_O$  and  $N_O \setminus N_I$ , respectively. To minimize propagation cost, this is ensured by a simple n-ary constraint propagating, which is propagated only after all variables have been confined to a single value. The full CSP is depicted in Fig. 4.

Although the CSPs from above are defined for domains of vertices  $v \in V_I \cup V_O$ , they can be easily reformulated using integer encodings of the atom indices



allowing for the application of standard constraint solvers such as **Gecode** [42]. This enables the use of efficient propagators for most of the required constraints, such as the algorithm of Regin [41] for globally unique assignments. Only a few binary constraints, e.g. to ensure atom label preservation or the cyclic bond pattern, require a dedicated implementation as discussed in Sec 4.

All solutions for these CSPs are chemically valid ITS candidates. In order to check whether or not a true ITS is found we have to ensure that the remaining atoms, i.e., those that do not participate in the ITS, can be mapped without further bond formation or breaking. This is achieved using a standard graph matching approach as discussed in the following.

### 2.3 Overall Atom Mapping Computation

Given the CSP formulation from above, we can enumerate all valid ITS candidates. For a CSP solution we denote with  $a_i^I$  and  $a_i^O$  the assigned values of the variables  $X_i^I$  and  $X_i^O$ , respectively. Once the ITS candidate is fixed, we can reduce the problem to a general graph isomorphism problem with a simple re-labeling of the ITS edges. Thus, we derive two new adjacency matrices  $\mathcal{I}'$  and  $\mathcal{O}'$  from the original matrices  $\mathcal{I}$  and  $\mathcal{O}$ , resp., as follows: For all atom pairs  $(i, j)$  within the cyclic index sequence 1-2-...- $k$ -1, we change the corresponding adjacency information to a unique label using  $\mathcal{I}'_{a_i^I, a_j^I} = \mathcal{O}'_{a_i^O, a_j^O} \in \{f, b\}$  encoding if a bond between the mapped ITS vertices is formed ( $f$ ) or broken ( $b$ ). All other adjacency entries are kept the same as in  $\mathcal{I}$  and  $\mathcal{O}$ , respectively.

Given these updated, "ITS encoding" adjacency matrices  $\mathcal{I}'$  and  $\mathcal{O}'$ , the identification of the overall atom mapping  $m$  reduces to the graph isomorphism problem based on  $\mathcal{I}'$  and  $\mathcal{O}'$ . Thus, all exact mappings of  $\mathcal{I}'$  onto  $\mathcal{O}'$  are valid atom mappings  $m$  of an elementary homovalent reaction, since the encoded ITS respects all constraints due to the CSP formulation.

### 2.4 Implementation Details

Our C++ implementation of the approach currently takes a chemical reaction in SMILES format [43], identifies chemically correct atom mappings, and returns these in annotated SMILES format. The latter provides a numbering of mapped atoms in the educts and products.

Molecule parsing, writing, and graph representation uses the chemistry module of the Graph Grammar Library (GGL) [35]. Note, we do an explicit hydrogen representation within the CSP formulation as in [16], since most homovalent elementary reactions involve the replacement of at least one hydrogen. Unfortunately, the compact string encoding of molecules in SMILES format does not explicitly represent hydrogens. Thus, we use the hydrogen correction procedures of the GGL to complete educt and product molecule input. The CSP formulation and solving is done within the Gecode framework on finite integer domains [42]. The final graph matching is done using the state-of-the-art VF2-algorithm [10], which is among the fastest available [9].

The CSP uses standard binary order constraints and the n-ary distinct and counting constraints provided by the Gecode library. Dedicated binary constraints propagating on unassigned domains have been implemented for preservation of atom label, degree, and homovalence. The alternating cycle is implemented by a sequence of  $k$  constraints propagating the edge valence change of  $\pm 1$ . The ITS local neighborhood preservation to be enforced in the extended CSP is implemented by a dedicated n-ary constraint over all variables propagating on assignments only.

We are using a Depth-First-Search where the branching strategy chooses first variables with minimal domain size and first assigns non-hydrogen indices before hydrogen vertices are considered. The latter increases the performance to find the first solution since most reaction mechanism are constructed of at least 50% non-hydrogen atoms. Once a non-hydrogen is selected, propagation will ensure that adjacent hydrogens are considered for the neighbored variables within the ITS cycle encoding if appropriate.

For each ITS mapping identified, a full reaction atom mapping is derived via VF2-based graph matching. Therein, the discussed problem of hydrogen interchangeability (see extended CSP formulation) is faced again and would result in symmetric overall atom mappings. This is countered by first producing intermediate "collapsed" educt/product graphs, where all adjacent non-ITS hydrogens are merged into the atom labels of their adjacent non-hydrogens. This preserves the adjacency information and enables a unique mapping via VF2 excluding the hydrogen-symmetries. Furthermore, this compression speeds up the graph isomorphism identification since the graph size is approximately halved.

While not described here, the CSPs can be easily extended to find candidates for the entire atom mapping by introducing additional matching variables for all atoms participating in the reaction, all constrained to preserve atom label, vertex degree, and bond valence information. But first tests (not shown) revealed that the increase in CSP size and accordingly search and propagation effort needed does not repay due to the efficiency of the VF2 graph isomorphism approach used. Therefore, we omitted this approach from this work.

### 3 Application and Evaluation

In order to investigate the impact of our extended CSP formulation over the basic version, we selected a subset of homovalent elementary reactions from the KEGG LIGAND database [29]. The reactions have been chosen to provide various ITS and reaction sizes for evaluation. The average size of the selected reactions, i.e. the average number of atoms, is about 30 (Tab. 2 column 2) while the whole KEGG database shows an average of 50 atoms per reaction. The example reactions cover homovalent ITS sizes of  $k = 4, 6, \text{ and } 8$  as introduced. Since there is no atom mapping information provided within the KEGG database, the example reactions had to be identified manually based on chemical knowledge. This again highlights the need for an automated identification of chemically feasible

Reaction	Educts	Products
R00013	<chem>C(=O)=O, C(C(=O)O)(C=O)O</chem>	$2 \times$ <chem>C(=O)(C=O)O</chem>
R00018	<chem>N, N(CCCCN)CCCN</chem>	$2 \times$ <chem>C(CCN)CN</chem>
R00048	<chem>CC(O)CC(=O)OC(C)CC(O)=O, O</chem>	$2 \times$ <chem>CC(O)CC(O)=O</chem>
R00059	<chem>N(C(=O)CCCCN)CCCCC(=O)O, O</chem>	$2 \times$ <chem>C(CC(=O)O)CCCN</chem>
R00207	<chem>P(=O)(O)(O)O, O=O, CC(=O)C(=O)O</chem>	<chem>P(=O)(OC(=O)C)(O)O, OO, C(=O)=O</chem>

**Table 1.** Elementary homovalent reactions from the KEGG LIGAND database [29] used for the evaluation of the approach. The educt and product molecules are given in SMILES notation [43].

Reaction	Atoms	CSP Type	$k$	Time 1st Sol.	Sol. Sol.	Sol. CSP	Time all Sol. CSP	VF2
R00013	14	Basic Ext. {2C}	6	0.03 <b>0.02</b>	1	346 76	0.8 <b>0.05</b>	0.03 0.02
R00018	36	Basic Ext. {2N}	4	10.4 <b>0.28</b>	1	73,924 36	2.62 <b>0.44</b>	19.9 0.01
R00048	30	Basic Ext. {2O}	4	0.1 <b>0.02</b>	2	26,178 24	1.44 <b>0.42</b>	6.1 0.03
R00059	44	Basic Ext. {H,C,N,O}	4	0.34 <b>0.03</b>	1	194,210 4	9.45 <b>2.08</b>	63.15 0.01
R00207	20	Basic Ext. {C,4O}	8	0.02 <b>0.01</b>	1	20,640 24	1.11 <b>0.56</b>	4.05 0.02

**Table 2.** Evaluation of the reactions from Tab. 1. Timings are given in seconds. For extended CSPs, the minimal set of ITS participating atoms is listed in column 3. Column “Sol. CSP” gives the number of CSP solutions (ITS candidates) tested via VF2 for final atom mappings.

atom mappings as provided by our approach. The selected homovalent reactions are given in Tab. 1 with their respective KEGG ID, educts and products.

For each reaction, we applied our approach using both the basic and extended CSP formulation to evaluate the impact of the latter for various reaction and ITS cycle sizes. In Table 2 we report runtime, search, and solution details for the smallest ITS size  $k$  that yields a solution. For smaller values of  $k$ , the infeasibility tests were done within fractions of seconds and are therefore omitted.

Our atom mapping approach finds a first atom mapping for homovalent elementary reactions within milliseconds. It is clear that the additional constraints within the extended CSP formulation significantly increase the performance of the approach. This becomes even more striking when considering the timings for full solution enumeration. The extended CSP produces several orders of magnitude less ITS candidates (column “Sol. CSP”). Since the time consumption of the VF2 algorithm is about linear in the number of ITS candidates to test, this results in according speedups of the overall approach. Still there is room for optimization since the symmetry breaking within the CSP solution enumeration is not complete (see next section).

The strength of the extended CSP comes from the precomputed list of local neighborhoods to be part of the ITS candidate and the “hydrogen symmetry” breaking. For the reactions from Tab. 2, this list comprises on average about half the ITS resulting in the impressive impact of the constraint. For reaction R00059, the list covers the whole ITS with an according immense reduction in ITS candidates.

As already expected based on the results from other approaches [16], only a single or very few reaction mechanisms, i.e. non-symmetric atom mappings, are identifiable, see Tab. 1 column “Sol”.

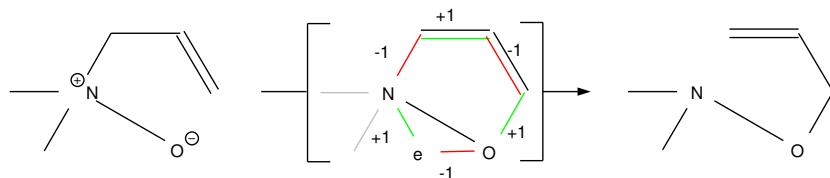
## 4 Development and Future Work

The basic approach was implemented by a user not familiar with constraint programming within 1 month work time given the well documented and easily extendable Gecode library [42] and the chemoinformatics implementations provided by the GGL [35]. Extending the approach and adding the basic functionalities for symmetry exclusion required another week of implementation, such that we got a first prototype within 1.5 months. Given the current framework and available constraint implementations, we expect another month of implementation time to get the final atom mapping program that will cover most of the following features.

*Branching strategies:* The current CSP allows for further performance optimizations when solving the satisfaction problem. We are currently evaluating the impact of different branching strategies on the runtime of the approach. As a first result, a hierarchical value selection that first tries to assign vertices to the variables that are compatible to the neighborhoods participating in the ITS (see extended CSP formulation) and which selects hydrogen representing vertices last seems to allow for a good performance.

*Symmetry breaking:* As it can be seen from Tab. 2, the current CSP formulation still produces symmetric ITS solutions when enumerating all possible atom mappings. We are currently working on strategies to apply further symmetry breaking techniques during the solution enumeration of CSPs, i.e. symmetry breaking during search (SBDS) [4, 21, 5] (or the similar lightweighted dynamics symmetry breaking (LDSB) approach [37]), as well as symmetry exclusion in the final mapping phase. Both requires more sophisticated input analyses as e.g. done in [16].

*CSPs for other ITS layouts:* Of course, not all chemical transformations are based on a homovalent elementary ITS. This will in general be the case for multi-step reactions and for the so-called ambivalent reactions, in which the number of non-bonding electron pairs (and thus the oxidation number of some atoms) changes in the course of a reaction [25]. Figure 5, for example, shows a reaction for which it is not possible to find a simple homovalent circular ITS



**Fig. 5.** The Meisenheimer rearrangement [38] transforms nitroxides to hydroxylamines. It does not admit a simple alternating cycle as ITS when molecules are represented as graphs whose vertices are atoms. An extended representation, in which the additional electron at the oxygen is treated a “pseudo-atom” can fix this issue. See Figure 6 for further details of such an ITS layout.

Reaction		Atoms	$k$	Fig.	Sol.	Sol. CSP	Time all Sol.
$O=[S^{--}] = O . C=CC=C$	$O=S1(=O)CC=CC1$	13	5	6 bottom	1	20	0.01
$C1[C^{--}]C1 . C=C$	$C1C1(C1)CC1$	9	3	6 bottom	1	12	0
$[O^-] [NH2^+] CC=C$	$NOCC=C$	12	5	5, 6 top	1	22	0

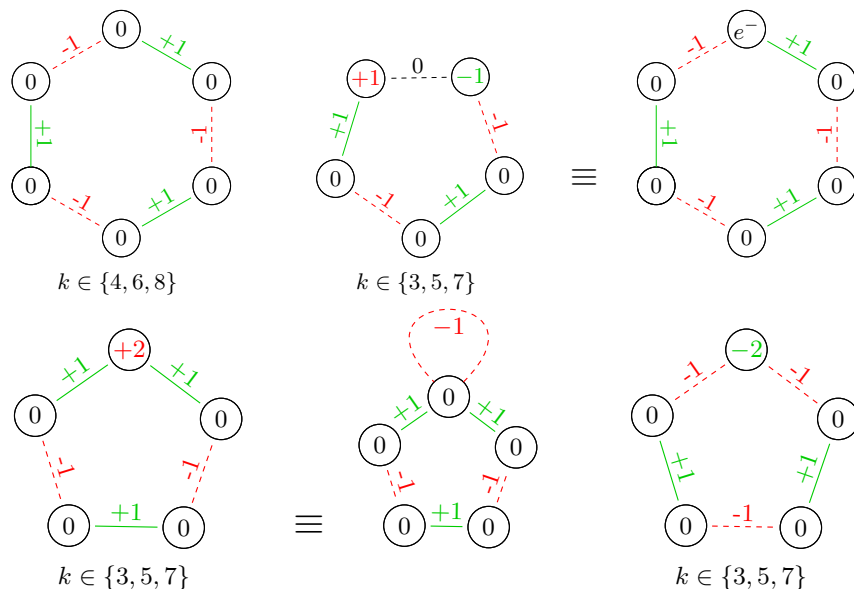
**Table 3.** Evaluation of CSPs for ambivalent reactions with an odd cycle ITS layout given in Fig. 6.

using the presented ITS encoding. Still the reaction shows a cyclic ITS with alternating bond electron changes for all but one bond [17].

We have extended the CSP-based framework outlined above to reactions with arbitrary cyclic ITS layouts, which allows for any defined bond and atom valence changes (i.e. charge changes) within the ITS. Figure 6 exemplifies odd ITS cycle layouts for ambivalent reactions [19]. The main difference to homovalent reaction CSP is the relaxation of the homovalence constraint, which is not enforced for all participating atoms [19]. Furthermore, the preservation of bond electrons for some ITS bonds instead of a change is enforced. The latter holds for instance for the bond connecting  $N^+$  and  $O^-$  in Fig. 5.

Table 3 presents the timing results for our prototypical implementation of the ambivalent ITS layouts given in Fig. 6. The model is based on the extended CSP formulation for elementary reactions. Also for such ambivalent reactions, our CP-based atom mapping approach enumerates all possible atom mappings within milliseconds, as reported for homovalent reactions in Tab. 2. Note, the ambivalent CSPs require a different, ITS-specific symmetry breaking and thus have to enforce different static order constraints compared to the homovalent CSP. The ambivalent layouts given in Fig. 6 show no symmetry in itself such that actually no order constraint is needed here. The driving force of the CSP performance is the propagation of the oxidation state change for the atoms that get charged. This poses a very strong constraint for the ambivalent ITS identification.

We are currently identifying and verifying further ITS layouts, some of the already available layouts are given in Fig. 6. Considering the reaction classifi-



**Fig. 6.** Currently supported ITS layouts: The number within the vertices corresponds to atomic oxidation state changes, red bonds are broken green bonds are formed. (top) Homovalent elementary reactions result in even sized cycles with no oxidation state changes at the atoms (see Fig. 1). Note that odd cycles with two oppositely charged atoms separated by a non-changing pseudo bond (black dashed edge labeled 0 see Fig. 5) are equivalent to the next larger even sized cycle with a virtual vertex for the moving charge (vertex label  $e^-$ ). (bottom) Ambivalent elementary reactions involving non bonding electrons result in odd sized cycles and oxidation state changes of one atom. Note that this situation is equivalent to a non-elementary cycle with alternating bond labeling (bottom middle)

cation work in [25, 24, 17, 18], we expect a very limited number of possible ITS layouts within a few hundreds at most given the physics underlying chemical reactions. The overall approach will select, based on the provided input and the local neighborhood analyses presented for the extended CSP, the suitable ITS layouts and their respective CSPs and search for valid atom mappings.

*Multi-step reactions:* The current framework is designed to identify chemically feasible atom mappings for single-step reactions. Nevertheless, there cases where short-lived intermediate molecule structures are formed that are directly react further into the final products. Unfortunately, these intermediate structures are usually unknown, such that we cannot apply the presented approach.

As discussed by Hendriksen [24], often only two joint reactions with a single unknown intermediate are observed. We therefore plan to create “fused” ITS layouts based on our single-step ITS encodings that will allow for the correct identification of atom mappings for multi-step reactions and reveal the individual steps and intermediate structures. For the combination of ITS layouts, we

are currently investigating the multi-step reaction analyses by Fujita [20] and Herges [25].

*Webserver:* The final atom mapping framework will be available both as stand alone tool as well as via a web front end including a visual depiction of the atom mappings. An according webservice framework ready for the integration is already available.

*Graph Grammars and Atom Flow Network Generation:* Atom mappings are the base to generate and analyze the atom flow in reaction networks [7, 26]. Here, the chemical validity of the atom maps is of particular importance to ensure correct atom flow analyses. We will use our atom mapping approach to generate chemical graph rewrite rules that will be used within our GGL framework [35] to expand according reaction networks where molecular graph rewrite directly provides the atom flow information within the network.

## 5 Discussion

We have presented here the first constraint programming approach to identify chemically feasible atom mappings based on the identification of a cyclic intermediate transition state (ITS). The incorporation of the cyclic ITS structure within the search ensures the chemical correctness of the mapping that is not guaranteed by standard approaches that attempt to solve Maximum Common Edge Subgraph Problems [1]. To our knowledge, this is the first approach explicitly incorporating the cyclic ITS structure into an atom mapping procedure.

The formulation of the CSP using only the atoms involved in the ITS results in a very small CSP that can be solved efficiently. Thus, it is well placed as a filter for ITS candidates for the subsequent, computationally more expensive graph matching approaches. The solutions of such an extended CSP are the desired chemically feasible atom mappings  $m$ . We apply advanced symmetry breaking strategies and thus can enumerate the different chemical mechanisms underlying a reaction for a given ITS cycle size.

The feasibility of the approach was demonstrated here for the special case of elementary, homovalent reactions, i.e., for reactions in which the transition state is an elementary cycle with an even number of atoms. The CSP formulation can be easily extended to arbitrary cyclic ITS layouts. Usually, such reactions are not homovalent, i.e., at least one atom participating in the ITS is gaining or losing non-bonding electrons, which requires some moderate changes in the formulation of the constraints. We are currently identifying all feasible ITS layouts and are developing a generic CSP formulations. This will result in a powerful approach to identify atom mappings with chemically valid ITSs.

Constraint programming was shown to be a very promising approach to solve atom mapping problems since it provides a very flexible framework to incorporate combinatorial constraints determined by the underlying rules of chemical transformations.

## References

1. T. Akutsu. Efficient extraction of mapping rules of atoms from enzymatic reaction data. *J. Comp. Biol.*, 11:449–62, 2004.
2. J. Apostolakis, O. Sacher, R. Körner, and J. Gasteiger. Automatic determination of reaction mappings and reaction center information. 2. Validation on a biochemical reaction database. *J. Chem. Inf. Mod.*, 48:1190–1198, 2008.
3. M. Arita. Scale-freeness and biological networks. *J Biochem*, 138:1–4, 2005.
4. R. Backofen and S. Will. Excluding symmetries in constraint-based search. In *Proc. of 5<sup>th</sup> International Conference on Principle and Practice of Constraint Programming (CP'99)*, volume 1713 of *LNCS*, pages 73–87. Springer, 1999.
5. R. Backofen and S. Will. Excluding symmetries in constraint-based search. *Constraints*, 7(3):333–349, 2002.
6. L. Bahiense, G. Manić, B. Piva, and C.C. de Souza. The maximum common edge subgraph problem: A polyhedral investigation. *Discr. Appl. Math.*, 160:2523–2541, 2012.
7. T. Blum and O. Kohlbacher. Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *Journal of Computational Biology*, 15:565–576, 2008.
8. W.L. Chen, D.Z. Chen, and K.T. Taylor. Automatic reaction mapping and reaction center detection. *WIREs Comput Mol Sci*, 2013. doi:10.1002/wcms.1140.
9. L.P. Cordella, P. Foggia, C. Sansone, and M. Vento. Performance evaluation of the VF graph matching algorithm. In *Proceedings of the 10th International Conference on Image Analysis and Processing, ICIAP '99*, page 1172. IEEE Computer Society, 1999.
10. L.P. Cordella, P. Foggia, C. Sansone, and M. Vento. A (sub)graph isomorphism algorithm for matching large graphs. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(10):1367–72, 2004.
11. J.D. Crabtree and D.P. Mehta. Automated reaction mapping. *J. Exp. Algor.*, 13:1.15–1.29, 2009.
12. J.D. Crabtree, D.P. Mehta, and T.M. Kouri. An open-source Java platform for automated reaction mapping. *J Chem Inf Model*, 50:1751–1756, 2010.
13. M.J.L. de Groot, R.J.P. van Berlo, W.A. van Winden, P.J.T. Verheijen, M.J.T. Reinders, and D. de Ridder. Metabolite and reaction inference based on enzyme specificities. *Bioinformatics*, 25(22):2975–83, 2009.
14. J. Dugundji and I. Ugi. An algebraic model of constitutional chemistry as a basis for chemical computer programs. *Topics Cur. Chem.*, 39:19–64, 1973.
15. H.-C. Ehrlich and M. Rarey. Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review. *WIREs Comput Mol Sci*, 2011. doi:10.1002/wcms.5.
16. E.L. First, C.E. Gounaris, and C.A. Floudas. Stereochemically consistent reaction mapping and identification of multiple reaction mechanisms through integer linear optimization. *J. Chem. Inf. Model.*, 52(1):84–92, 2012.
17. S. Fujita. Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts. *J. Chem. Inf. Comput. Sci.*, 26:205–212, 1986.
18. S. Fujita. Description of organic reactions based on imaginary transition structures. 2. Classification of one-string reactions having an even-membered cyclic reaction graph. *J. Chem. Inf. Comput. Sci.*, 26:212–223, 1986.
19. S. Fujita. Description of organic reactions based on imaginary transition structures. 3. Classification of one-string reactions having an odd-membered cyclic reaction graph. *J. Chem. Inf. Comput. Sci.*, 26:224–230, 1986.



20. S. Fujita. Description of organic reactions based on imaginary transition structures. 5. Recombination of reaction strings in a synthesis space and its application to the description of synthetic pathways. *J. Chem. Inf. Comput. Sci.*, 26:238–242, 1986.
21. I.P. Gent and B.M. Smith. Symmetry breaking in constraint programming. In *Proceedings of ECAI-2000*, pages 599–603. IOS Press, 2000.
22. M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa. Heuristics for chemical compound matching. *Genome Informatics*, 14:144–53, 2003.
23. M. Heinonen, S. Lappalainen, T. Mielikäinen, and J. Rousu. Computing atom mappings for biochemical reactions without subgraph isomorphism. *J. Comp. Biol.*, 18:43–58, 2011.
24. J.B. Hendrickson. Comprehensive system for classification and nomenclature of organic reactions. *J Chem Inf Comput Sci*, 37:852–860, 1997.
25. R. Herges. Organizing principle of complex reactions and theory of coarctate transition states. *Angewandte Chemie Int Ed*, 33:255–276, 1994.
26. T. Hogiri, C. Furusawaa, Y. Shinfukua, N. Onoa, and H. Shimizua. Analysis of metabolic network based on conservation of molecular structure. *Biosystems*, 95(3):175–178, 2009.
27. X. Huang, J. Lai, and S.F. Jennings. Maximum common subgraph: some upper bound and lower bound results. *BMC Bioinformatics*, 7 (S4):S6, 2006.
28. C. Jochum, J. Gasteiger, and I. Ugi. The principle of minimum chemical distance (PMCD). *Angew. Chem. Int. Ed.*, 19:495–505, 1980.
29. M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nuc. Acids Res.*, 40(Database issue):D109–14, 2012.
30. R. Körner and J. Apostolakis. Automatic determination of reaction mappings and reaction center information. 1. The imaginary transition state energy approach. *J. Chem. Inf. Mod.*, 48:1181–1189, 2008.
31. M. Kotera, Y. Okuno, M. Hattori, S. Goto, and M. Kanehisa. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.*, 126:16487–16498, 2004.
32. M. Latendresse, J.P. Malerich, M. Travers, and P.D. Karp. Accurate atom-mapping computation for biochemical reactions. *J Chem Inf Model*, 2013. doi:10.1021/ci3002217.
33. M. Leber, V. Egelhofer, I. Schomburg, and D. Schomburg. Automatic assignment of reaction operators to enzymatic reactions. *Bioinformatics*, 25:3135–3142, 2009.
34. M. Lynch and P. Willett. The automatic detection of chemical reaction sites. *Journal of Chemical Information and Computer Sciences*, 18:154–159, 1978.
35. M. Mann, H. Ekker, and C. Flamm. The graph grammar library - a generic framework for chemical graph rewrite systems. In Keith Duddy and Gerti Kappel, editors, *Theory and Practice of Model Transformations, Proc. of ICMT 2013*, volume 7909 of *LNCS*, pages 52–53. Springer, 2013. Extended abstract at ICMT, long version at arXiv <http://arxiv.org/abs/1304.1356>.
36. M. Mann, H. Ekker, P.F. Stadler, and C. Flamm. Atom mapping with constraint programming. In R. Backofen and S. Will, editors, *Proceedings of the Workshop on Constraint Based Methods for Bioinformatics WCB12*, pages 23–29, Freiburg, 2012. Uni Freiburg. <http://www.bioinf.uni-freiburg.de/Events/WCB12/proceedings.pdf>.
37. C.D. Mears, M.J. Garcia De La Banda, B. Demoen, and M. Wallace. Lightweight dynamic symmetry breaking. In *Proc. of the 8th International Workshop on Symmetry in CSPs (SymCon08)*, 2008. <http://www.aloul.net/symcon/>.

38. J. Meisenheimer. Über eine eigenartige Umlagerung des Methyl-allyl-anilin-N-oxyds. *Chemische Berichte*, 52:1667–1677, 1919.
39. J. Rautio, H. Kumpulainen, T. Heimbach, R. Oliyai, D. Oh, T. Järvinen, and J. Savolainen. Prodrugs: design and clinical applications. *Nature Reviews Drug Discovery*, 7(3):255–270, 2008.
40. J.W. Raymond and P. Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Computer-Aided Mol. Design*, 16:521–33, 2002.
41. J.-C. Regin. A filtering algorithm for constraints of difference. In *Proceedings of the 12th National Conference of the American Association for Artificial Intelligence*, pages 362–367, 1994.
42. Gecode Team. Gecode: Generic constraint development environment, 2013. Available as an open-source library from <http://www.gecode.org>.
43. D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comp. Sci.*, 28(1):31–36, 1988.
44. W. Wiechert.  $C^{13}$  metabolic flux analysis. *Meta Eng*, 3:195–206, 2001.
45. Y. Yamanishi, M. Hattori, M. Kotera, S. Goto, and M. Kanehisa. E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. *Bioinformatics*, 25(12):i179–i186, 2009.