E. E. Litsa, M. I. Peña, M. Moll, G. Giannakopoulos, G. N. Bennett, and L. E. Kavraki, "Machine Learning Guided Atom Mapping of Metabolic Reactions," Journal of Chemical Information and Modeling, vol. 59, no. 3, pp. 1121– 1135, 2019.

http://dx.doi.org/10.1021/acs.jcim.8b00434

Machine Learning Guided Atom Mapping of Metabolic Reactions

Eleni E. Litsa,[†] Matthew I. Peña,[‡] Mark Moll,[†] George Giannakopoulos,[¶] George N. Bennett,[‡] and Lydia E. Kavraki[*,†]

†*Department of Computer Science, Rice University, 6100 Main St., Houston, TX 77005 USA*

‡*Department of BioSciences, Rice University, 6100 Main St., Houston, TX 77005 USA*

¶*SKEL Lab, Institute of Informatics and Telecommunications, NCSR Demokritos, Agia Paraskevi 15310, Greece*

E-mail: kavraki@rice.edu

Abstract

The atom mapping of a chemical reaction is a bijection between the atoms in the reactant molecules and the atoms in the product molecules. It encodes the underlying reaction mechanism and, as such, constitutes essential information in computational studies in metabolic engineering. Various techniques have been investigated for the automatic computation of the atom mapping of a chemical reaction approaching the problem as a graph matching problem. The graph abstraction of the chemical problem though, eliminates crucial chemical information. There have been efforts for enhancing the graph representation by introducing the bond stabilities as edge weights, as they are estimated based on experimental evidence. Here, we present a fully automated optimization-based approach, named AMLGAM (Automated Machine Learning Guided Atom Mapping), that uses machine learning techniques for the estimation of the bond stabilities based on the chemical environment of each bond. The optimization method finds the reaction mechanism which favors the breakage/formation of the less stable bonds. We evaluate our method on a manually curated dataset of 382 chemical reactions and we run our method on a much larger and diverse dataset of 7,400 chemical reactions. We show that the proposed method improves the accuracy over existing techniques, based on results published by earlier studies on a common dataset, and is capable of handling unbalanced reactions.

1 Introduction

Within an organism, thousands of chemical reactions are catalyzed as part of the metabolic processes that take place to sustain life. A metabolic process consists of a series of chemical reactions that either break down molecules, supplying the cells with energy and building blocks, or synthesize new molecules necessary for the function of the cell. A chemical molecule consists of a set of atoms bound together with a specific arrangement. At each reaction step of a metabolic process, a set of molecules, called reactants, is transformed into a new set of molecules, called products, by rearranging the atoms of the reactant molecules. The transformation consists of bond breakages and bond formations that change the arrangement and the distribution of the atoms among the reactant molecules. The correspondence between the atoms in the reactant molecules and the atoms in the product molecules is given by the atom mapping (AM) of the chemical reaction.

The AM describes the underlying reaction mechanism since it encodes the changes that take place, i.e., which bonds break or form, during the reaction. Therefore, it complements the definition of a chemical reaction along with the sets of the reactant and the product molecules. As such, the AM is necessary for the development of computational tools that involve processing or simulation of chemical reactions especially for drug design studies. In metabolic engineering for example, the AM is used for assessing the feasibility of computationally derived metabolic pathways towards the production of therapeutic compounds.[1–3] In computer-aided synthesis, AMs are required in order to extract reaction rules from known reactions and utilize them to predict unknown reactions. Similarly, in the assessment of drug efficacy and safety, atom mappings are required for the prediction of drug metabolism.[4] Moreover, AM information is necessary in simulations of tracer experiments which are essential in various studies such as in metabolic flux analysis.[5,6] AMs have also been used for classifying chemical reactions based on the electron redistribution pattern of the reaction[7] and for retrieving reactions from chemical databases.[8]

Despite its importance, AM data are not available for many known chemical reactions. Most chemical databases do not provide AM data and those that do, may not provide AMs for their entire content. Determining the AM for a chemical reaction requires expert knowledge and expert annotation of chemical reactions in databases, which consist of thousands of entries, is not a feasible task. There is an increasing effort to fill this gap by developing computational tools for automatically determining the AM of a chemical reaction.

The AM problem, i.e., the automatic derivation of the AM of a chemical reaction using computational methods, has been formulated borrowing concepts from graph theory.[9] A chemical molecule can be represented as a graph in which the vertices correspond to atoms and the edges correspond to the bonds formed between the atoms. Within this context, the AM of the reaction is a graph matching between the graphs that correspond to the reactant and product molecules. The AM problem is equivalent to finding the bonds that create isomorphic subgraphs between the reactants and the products or otherwise the broken and formed bonds of the reaction.[10] The computational approaches to the AM problem can be divided into two major categories: the approaches that search for subgraph isomorphisms between the reactant and the product graph and the optimization-based approaches that minimize the number of reacting bonds.

The graph abstraction of the AM problem provides a formal formulation of the problem but at the same time imposes limitations. The graph representation captures the connectivity between the atoms but it does not encode the chemical properties of the molecules. Subgraph isomorphism-based techniques identify the common substructures between the reactant and the product molecules but in some cases this information is not enough to determine the AM.[1] Optimization-based techniques, on the other side, rely on the assumption that a reaction proceeds with the minimum number of reacting bonds, an assumption that does not hold for all chemical reactions. There have been attempts to enhance the graph representation of the molecules by introducing edge weights that reflect the bond stabilities.[11,12] In these approaches, the bond stabilities have been determined by experts, based on experimental evidence, and have been found to be inadequate to capture the variability imposed by the chemical environment of the bond.

An additional challenge in the AM problem is the case of unbalanced reactions. Unbalanced reactions are reactions in which a bijective mapping between the atoms of the two sides of the reaction, with respect to the atom species, is not possible due to missing atoms. While no reaction can be unbalanced in nature, such cases correspond to incomplete entries in chemical databases. Imbalances in the number of atoms between the reactant and the product molecules may be the result of wrong stoichiometry in the reaction equation, non-recorded molecules or even molecules with a wrong chemical formula. Such cases are quite prevalent in reaction databases with the two most popular databases, KEGG[13] and Metacyc,[14] containing around 10% and 5% unbalanced reactions,[15] respectively. According to another source though, reactions with a small number of missing atoms can constitute up to 40-50% of the content of reaction databases.[7] Therefore, an AM algorithm ideally should be capable of dealing with such cases and especially with reactions with a small number of missing atoms.

In this paper, we present an optimization-based AM algorithm that takes into account the stability of each bond. The contributions of this work are the following:

1. We provide a machine learning (ML) framework for estimating the bond stabilities based on local topological and atomic features regarding the bond itself as well as the connected atoms. To our knowledge, this is the first method that uses machine learning for incorporating chemical knowledge in the graph formulation of the AM problem.

2. The presented algorithm supports the computation of AMs for unbalanced chemical reactions and indicates equivalent mappings due to indistinguishable atoms.

3. We evaluate our method on a manually curated dataset of 382 balanced chemical reactions and compare it against existing AM tools based on results published by earlier studies. Furthermore, we run our method on a much larger and diverse dataset of 7,400 chemical reactions including unbalanced reactions.

More specifically, we adopt an optimization-based approach for the AM problem that minimizes the weighted graph edit distance between the reactants and the products. The edge weights correspond to bond stabilities and are estimated using ML techniques based on local features regarding the bond itself and the surrounding atoms. The obtained mapping corresponds to the reaction mechanism which favors the breakage of the less stable bonds. The optimization problem is formulated as a mixed integer linear programming (MILP) problem.[12,16] We handle unbalanced reactions by relaxing the constraints of the MILP problem. Finally, we indicate equivalent mappings due to indistinguishable mappings and provide alternative mappings if multiple optimal solutions exist.

2 Related Work

The current approaches to the AM problem can be divided into two major categories: 1) common substructure-based approaches and, 2) optimization-based approaches. A thorough review of the two approaches has been presented by Chen et al. in 2013.[17]

In the first category, the AM is determined by identifying the structures that are preserved through the reaction. The preserved structures are identified by detecting isomorphic subgraphs between the reactant and the product molecules. Most approaches rely on an iterative application of a maximum common subgraph (MCS) algorithm between the reactants and the products. MCS algorithms, although are characterized by high complexity, have been studied extensively for the comparison of chemical compounds, and multiple variations have been developed.[18] The first MCS based approach to the AM problem was presented by Arita in 2003.[1] Arita's study is of particular interest because he reported the cases where his method failed to identify the correct mapping and some of those errors are inherent limitations of the MCS approach.[1] The most recent MCS based approaches are the reaction decoder tool (RDT) by Rahman et al. (2016)[19] and the canonical labeling for clique approximation (CLCA) algorithm by Kumar and Maranas (2014).[20] The RDT tool is an ensemble method based on 4 different MCS based algorithms which can also handle unbalanced chemical reactions. The CLCA algorithm is a more efficient MCS approach based on canonical naming and local search algorithms which can also handle unbalanced reactions. As a side note, the CLCA algorithm was compared against optimization-based tools in a very thorough study, demonstrating the weaknesses of each approach.[20] A common substructure based approach that does not rely on the use of an MCS algorithm was published by Akutsu in 2004[10] in an effort to overcome the inherent limitations of the MCS based methods reported by Arita. Although Akutsu's method was specifically designed for a certain reaction category, named exchange reactions, his work offered important insights in the complexity of the AM problem.

Optimization-based approaches rely on the principle of the minimum chemical distance according to which a chemical reaction proceeds with the minimum structural change, which corresponds to a transformation with the minimum number of broken and formed bonds.[21] The mapping that corresponds to the minimum number of bond changes is determined by minimizing the edit distance between the reactant and the product graph. The edit distance corresponds to the bonds that change during the reaction. A minimum edit distance (MED) approach to the AM problem was presented by First et al. in 2011,[16] where the optimization problem was formulated as a mixed integer linear programming (MILP) problem. The MILP formulation allowed a more thorough representation of the chemical problem through the constraints and the objective function comparing to the common substructure based approaches. In particular, bond order changes, changes in hydrogen atoms and stereochemistry are also taken into account along with the bond formations and breakages. However, the capabilities of this approach are limited by the assumption that an optimal solution minimizes the number of changed bonds which does not hold for all reactions. In an effort to alleviate that assumption, Latendresse et al., in 2012, presented a modified version of First's MILP approach which takes into account the bond stability in the optimization function.[12] In that framework, the mapping is determined by maximizing the stability of the preserved bonds which implicitly minimizes the weighted edit distance (MWED). In that work, the values of the bond stabilities have been determined manually by chemists based on experimental evidence and depend on the species of the connected atoms and the bond order. However, the authors identified cases where the proposed stability values do not lead to the correct mapping and for those cases they adjusted the values. The authors compared the proposed MWED approach against First's MED approach and according to the reported results, the introduction of the bond stabilities gives an advantage over the simple MED method. The importance of introducing the bond stabilities in the optimization problem had been indicated earlier, in 2008, by Körner and Apostolakis.[11,22] They not only took into account the bond stabilities for determining the atom mappings but also elaborated the theoretical underpinnings of this approach. In particular, the weighted edit distance is regarded as an approximation of the transition state energy of the reaction, and therefore, its minimization leads to the reaction mechanism with the minimum activation energy.[11] An important distinction is that although in this approach the AM problem is formulated as an optimization problem, it is approached using common substructure techniques.

Methodologies that combine characteristics of both approaches, optimization and common substructure, have also been investigated. In these methods, the AM is determined in two stages: first the preserved structures are identified using an MCS algorithm and next the atoms in the unmatched structures are mapped using either heuristics or following an optimization based strategy that minimizes the reacting bonds. The Automapper tool from ChemAxon, the ICMAP tool from InfoChem (2013)[7] as well as Fooshe's ReactionMap (2013),[23] all fall under this category.

At this point, it should be noted that the lack of standard benchmark datasets has hampered the evaluation of the existing approaches. Many approaches do not assess the accuracy of the computed mappings and among those that do, the chosen datasets differ a lot in terms of reaction complexity as well as dataset size, hindering a comparative evaluation. On top of that, the validity of the reference mappings is always in question, especially in the case of mappings that have been derived computationally. An effort to create a manually curated dataset of chemical reactions for the comparative evaluation of the existing AM approaches was recently made by a research group in the Luxembourg Centre for Systems Biomedicine.[24] In this work, the authors created a dataset of 512 manually curated chemical reactions which was used to compare 6 existing approaches.

Regarding unbalanced reactions, many computational methods are not designed to handle such cases, including the MILP based approaches presented by First and Latendresse.[12,16] Among the methods that can be applied on unbalanced reactions, they deal with such cases by either re-balancing the reaction prior to the mapping computation or by allowing unmapped atoms.[7,11,20] An unbalanced reaction is re-balanced by adding molecules in the reaction equation that balance out the number of atoms between the two sides of the reaction for each species. In the simple case of missing oxygen atoms, water molecules can be added. If larger parts of the reaction are missing though, the problem of re-balancing the reaction equation becomes more complicated. In addition to that, optimization methods that re-balance the reaction equation rely on the additional assumption that the added structures do not undergo any structural change or otherwise they do not contribute in the cost function.[11]

In this paper, we present an optimization-based approach to the AM problem which takes into account the stability of each bond. We define the bond stability probabilistically and rely on ML techniques for its estimation taking into account features describing the bond locally within the molecule. We follow the MILP formulation of the optimization problem[12,16] which we modify in order to handle unbalanced chemical reactions without re-balancing the reaction equation.

The use of ML in the area of cheminformatics has been established as a tool to discover quantitative structure-activity relationships in chemical molecules, known as QSAR analysis. ML algorithms are used in order to model the relationship between molecular structures and certain properties such as toxicity and solubility.[25] In these methods, the chemical molecules are described as either binary vectors, called molecular fingerprints, that indicate the presence of certain substructures, or as a set of physiochemical descriptors that quantify certain properties (topological, geometrical, thermodynamic, electronic, constitutional) of the molecule.[26] In most cases, the descriptors regard the molecule in its entirety; however, more localized descriptors at a bond or atom level have also been used, such as, in the prediction of drug metabolism through the identification of possible sites of metabolism[4] and in the automatic assignment of EC numbers in metabolic reactions.[27] In the same framework, here we present a method that aims to correlate bond descriptors with the stability of the bond using ML. In our study, the descriptors contain topological and atomic information regarding the bond and the surrounding atoms. Although to our knowledge this is the first approach that uses ML to aid the computation of AMs in chemical reactions, it is worth mentioning that Muller et al., presented an ML-based method for the automatic identification of erroneous mappings that are computationally derived.[28]

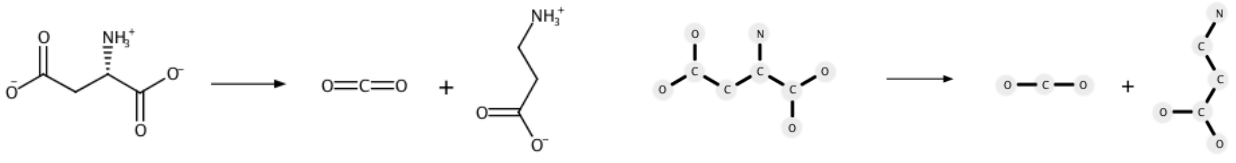
3 Problem Definition

A chemical reaction is denoted as:

$$r_1 + \ldots + r_m \longrightarrow p_1 + \ldots + p_n$$

where r_i, $i = 1, \ldots, m$ are the reactant molecules and p_i, $i = 1, \ldots, n$ are the product molecules.

Each reactant molecule r_i is represented as a graph with A_{r_i} being the vertex set corresponding to the set of atoms of the molecule r_i and B_{r_i} being the edge set corresponding to the set of bonds that are formed between the atoms in A_{r_i}. Accordingly, each product molecule p_i is represented as a graph with A_{p_i}, B_{p_i} being the vertex and the edge set of the graph. The graph representation can be extended to a chemical reaction as follows: A chemical reaction is represented by a pair of graphs R, P, with $R = (A_r, B_r)$ being the union of the graphs that correspond to the reactant molecules r_i, and $P = (A_p, B_p)$ being the union of the graphs of the product molecules p_i. Sets A_r and A_p contain all atoms appearing in the reaction equation except hydrogen atoms. Hydrogen atoms are highly reactive atoms and their position can change very rapidly. As such, the atom mappings of hydrogen atoms do not give important insights on the reaction mechanism and are therefore not computed. Figure 1 shows an example of a graph representation of a chemical reaction, with the reactant graph consisting of a single component and the product graph consisting of two components.



(a) Chemical reaction

(b) Graph representation

Figure 1: A chemical reaction and the corresponding graph representation

The atom mapping of a chemical reaction $r_1 + \ldots + r_m \longleftrightarrow p_1 + \ldots + p_n$, with reactant graph $R = (A_R, B_R)$ and product graph $P = (A_P, B_P)$, is a mapping $f : A_R \longrightarrow A_P$, which satisfies the following properties:

(i) The mapping preserves the atom species: if $f(i) = j$ then $s(i) = s(j)$ where $s(i)$ denotes the species of atom i.

(ii) The mapping f is bijective: for each atom $i \in A_r$ there is an atom $j \in A_p$ such that $f(i) = j$ and for each atom $j \in A_p$ there is an atom $i \in A_r$ such that $f^{-1}(j) = i$. Although the atom mapping of a chemical reaction is bijective in nature, from a computational perspective there are two cases where the bijective property is violated: unbalanced reactions and reactions that involve molecules with equivalent atoms.

- In unbalanced reactions, some atoms do not appear in both sides of the reaction and therefore a bijection from A_r to A_p is not possible. Still, a mapping exists which is defined in a subset of the total atoms: $f : A_R^B \subseteq A_R \longrightarrow A_P^B \subseteq A_P$ where $A_R^B = A_P^B$ is the set of atoms that appear in both sides of the reaction.
- Equivalent atoms correspond to different atom entities that are chemically indistinguishable. If two atoms are equivalent then they should be mapped to the same atom (or atoms) and therefore the mapping function is no longer bijective. More specifically, if $i \in A_R$ and $j \in A_P$ and f is an AM function, then:

$$\text{if } i \text{ and } i^* \text{ are equivalent and } f(i) = j \text{ then } f(i^*) = j \tag{1}$$

$$\text{if } j \text{ and } j^* \text{ are equivalent and } f^{-1}(j) = i \text{ then } f^{-1}(j^*) = i \tag{2}$$

Equivalences between atoms occur in the following cases:

(a) Multiple copies of the same molecule in the same side of the reaction. In that case, the corresponding atoms (atoms with the same position within each molecule) are equivalent. If for example, there are two water molecules in the reactants side then the two oxygen atoms are equivalent.

(b) Molecules with symmetries. In this case, the equivalent atoms belong to the same molecule and the molecule has planes of symmetry. Computationally, we determine such equivalences by comparing the BFS (Breadth First Search) traversal of the chemical graph starting from each atom, taking into account visited atoms and bonds, with the bond defined by its bond order and stereochemistry. The two oxygen atoms in carbon dioxide as well as the highlighted carbon atoms in acetone of Figure 2 are equivalent atoms that fall under this case.

(c) Equivalences due to resonance. In this case, the equivalent atoms belong in the same molecule, they have the same species, they are connected to the same atom, which can be either C, P, N or S, and they are not connected with any other atom. It should be noted that, in this case, the two bonds that attach the two atoms to the same atom do not have the same order (if the bond order is the same then the atoms are still equivalent according to criterion (b)) however, the shared electrons for each bond are de-localized due to resonance and therefore the two atoms cannot be distinguished. The two highlighted oxygen atoms in acetoacetate of Figure 2 are equivalent according to this criterion.

(iii) The mapping represents the reaction mechanism i.e., which bonds change during the reaction. In this approach, we assume that a chemical reaction proceeds with the minimum activation energy and therefore the computed mapping should induce the breakage and formation of the bonds with the lowest stability.

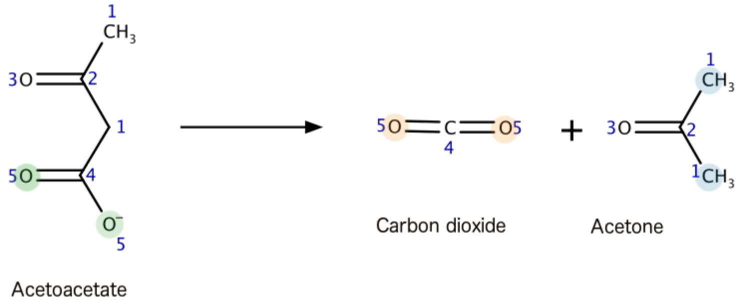


Figure 2: Equivalent atoms due to the resonance phenomenon in acetoacetate and due to symmetries in carbon dioxide and acetone

4 Methodology

We present an automated, machine learning guided, atom mapping method (AMLGAM). The AMLGAM method determines the AM by minimizing the weighted edit distance (MWED) between the reactant and the product graph, with the edge weights representing the bond stabilities. The optimization problem is defined as a mixed integer linear programming problem.[12,16] The stability of a bond is an indication of the difficulty for breaking that bond during the reaction and we use probabilistic techniques for estimating it. The computed AM corresponds to the reaction mechanism that favors the breakage or formation of the less stable bonds. This way we seek to identify the reaction mechanism that requires the minimum amount of energy for the reaction to proceed, or otherwise the minimum activation energy based on a crude approximation of the activation energy that relies on bond stabilities.[11]

4.1 Assumptions

The presented framework relies on the assumption that each chemical reaction proceeds with the minimum activation energy.[11] More specifically, the assumptions of the AMLGAM method are:

1. A chemical reaction proceeds with the minimum activation energy.
2. The energy that a bond requires to break or form is proportional to the stability of the bond.
3. The total amount of energy that a reaction requires to proceed is approximated as the sum of the energies of the bonds that completely break or form.
4. The chemical reaction is a single step transformation.
5. The complete breakage of a double or triple bond consists of two concurrent events: a bond order change (reduction) and a bond breakage. Here, for simplicity we assume that the two events are independent.

Based on these assumptions, we minimize the sum of the stabilities of the reacting bonds, as they are approximated by the ML model, in an effort to find the AM that corresponds to the minimum activation energy. The assumptions 1 – 4 are in agreement with the theoretical background that Körner and Apostolakis' approach relies on,[11] while the last assumption is introduced for facilitating the calculation of the stabilities with probabilistic methods as further discussed in section 4.3. In the case of unbalanced reactions, the presented method relies on the additional assumption that atoms are missing only from one side of the reaction equation. This additional assumption does not restrain the application of the method in cases where there are missing atoms in both sides of the reaction but lowers the confidence on the computed mapping.

4.2 Formulation of the optimization problem

We formulate the optimization problem of minimizing the activation energy as a MILP problem.[12,16] A MILP problem is described as:

$$\min \quad C^T X \qquad s.t. \quad AX \leq B \tag{3}$$

where X is a set of variables, $AX \leq B$ is a set of linear constraints and $C^T X$ is the objective function. A solution of a MILP problem is an assignment to the variables X which optimizes the objective function and satisfies the constraints. In the MILP formulation of the AM problem, the variables correspond to possible mappings between the atoms and the bonds of the two sides of the reaction, the objective function minimizes the weighted edit distance while the constraints impose the bijective property to the mapping and ensure consistency between the atom and the bond mappings.

More specifically, two types of variables are defined: the atom mapping variables and the bond mapping variables. An atom mapping variable a_{ij} is defined for every pair of atoms i, j with $i \in A_r$ and $j \in A_p$ which have the same species $s(i) = s(j)$. A bond mapping variable β_{ijkl} is defined for every pair of bonds (i, j), (k,l), with $(i, j) \in B_r$, $(k,l) \in B_p$ and $s(i) = s(k)$ and $s(j) = s(l)$. It should be noted here that the bonds are considered directed and therefore the mapping β_{ijkl} is not the same as the mapping β_{ijlk}. In the case of symmetric bonds, i.e., bonds that connect atoms of the same species, both mappings β_{ijkl} and β_{ijlk} are possible while in the case of bonds that connect atoms of different species, only one mapping is defined. The atom mapping variables are defined as binary and indicate possible mappings between the atoms in reactants and the atoms in products. Accordingly, the bond mapping variables are also binary variables indicating possible mappings between the bonds in the two sides of the reaction. Here, we recall that the hydrogen atoms are not mapped and therefore we do not define atom mapping variables between hydrogen atoms.

The constraints on the atom mapping variables are defined in order to ensure the bijective nature of the mapping: each atom in the reactants should be mapped to one atom in the products and each atom in the products should be mapped to one atom in the reactants. Since the atom mapping variables are declared as binary variables, taking values 0 or 1, these constraints can be expressed as:

$$\forall i \in A_r \sum_{j \in A_p, s(i)=s(j)} a_{ij} = 1 \tag{4a}$$

$$\forall j \in A_p \sum_{i \in A_r, s(i)=s(j)} a_{ij} = 1 \tag{5a}$$

In the case of unbalanced reactions the bijective property of the mapping is violated. In particular, a one-to-one mapping is not possible for all atom species but may be possible for certain species. We apply constraints 4a and 5a for the atom species that are balanced, i.e., there is an equal number of occurrences of that species in both sides of the reaction. We relax those constraints for the atom species that are not balanced in order to allow some atoms to remain unmapped:

$$\forall i \in A_r^* \qquad 0 \leq \sum_{j \in A_p, s(i)=s(j)} a_{ij} \leq 1 \tag{4b}$$

$$\forall j \in A_p^* \qquad 0 \leq \sum_{i \in A_r, s(i)=s(j)} a_{ij} \leq 1 \tag{5b}$$

where A_r^* and A_p^* are the sets of atoms of the non-balanced species in the reactants and products side, respectively. If, for example, only oxygen atoms are not balanced then the constraints 4b and 5b are applied only for the oxygens while the constraints 4a and 5a are applied for all other atoms.

In order to ensure that a maximal mapping is found and constraints 4b and 5b do not let additional atoms be unmapped, the following constraint is also introduced in the case of unbalanced reactions:

$$\sum_{i \in A_r, j \in A_p} a_{ij} = min(|A_r|, |A_p|) \tag{6}$$

This constraint requires the number of mapped atoms to be equal to the minimum number of atoms between the two sides of the reaction assuming that atoms are missing only in one side of the reaction.

The constraints on the bond mapping variables are defined in order to ensure consistency between the bond mappings and the atom mappings. If a bond $(i, j) \in B_r$ is mapped to a bond $(k,l) \in B_p$ then the atom i should be mapped to k and the atom j should be mapped to l (we recall here that the bonds are considered directed). This constraint is formulated as:

$$\forall \beta_{ijkl} \quad (\beta_{ijkl} \leq a_{ik} \wedge \beta_{ijkl} \leq a_{jl}) \tag{7}$$

The objective function minimizes the weighted edit distance between the reactant and the product graph. In particular, it penalizes the broken and formed bonds with a cost equal to the stability of each bond. The objective function favors the breakage/formation of the most unstable bonds:

$$\min \quad \sum_{(i,j) \in B_r} S_{ij}(1 - \sum_{(k,l) \in B_p} \beta_{ijkl}) + \sum_{(k,l) \in B_p} S_{kl}(1 - \sum_{(i,j) \in B_r} \beta_{ijkl}) \tag{8}$$

The first term of the objective function penalizes the bonds that completely break, while the second term penalizes the bonds that are formed. The cost of a broken or a formed bond (i, j) is equal to the stability S_{ij} of the bond. The estimation of the bond stability S_{ij} for a bond (i, j) is described in the following section. Here we recall that the variables β_{ijkl} are binary variables that indicate a possible mapping between the bond $(i, j) \in B_r$ and the bond $(k,l) \in B_p$. If a bond $(i, j) \in B_r$ breaks, this means that it is not mapped to any bond in B_p and therefore $\beta_{ijkl} = 0$, $\forall (k,l) \in B_p$ and consequently $1 - \sum_{(k,l) \in B_p} \beta_{ijkl} = 1$. Similarly, for a formed bond $(k,l) \in B_p$, there is no mapped bond in reactants and therefore $\beta_{ijkl} = 0$, $\forall (i, j) \in B_r$.

It should be noted, that the objective function in (8) includes only the bonds that completely break or form while bond order changes are not explicitly taken into account. However, a change in the bond order normally induces a bond breakage or formation and therefore bond order changes are implicitly taken into account through penalizing the induced bond formations or breakages. Changes in the bond stereochemistry are also not included in the objective function due to the uncertainty of the event. A change in the bond stereochemistry may mean that either the bond simply changed stereochemistry or the bond broke and reformed with a different stereochemistry.

4.3 Estimation of bond stabilities using machine learning

We estimate the stability of a bond following a probabilistic approach. More specifically, we define the stability S_{ij} of a bond (i, j) as the probability of the bond being preserved after the reaction or otherwise the probability of not breaking the bond (i, j).

$$S_{ij} = 1 - Pr_{ij}(break) \tag{9}$$

After a chemical reaction, a bond formed between two atoms in reactants, will either remain intact, preserving its bond order and stereochemistry, or it will react. A reacting bond either changes bond order, or changes stereochemistry or completely breaks. Here, we assume that the complete breakage of a double or triple bond implies two concurrent events: a reduction in the bond order and a breakage of a first order bond. We calculate the probability that a first order bond will break as $Pr(break) = Pr(react) \cdot Pr(break|react)$. For a second order bond, that probability is obtained as $Pr(break) = Pr(react) \cdot Pr(BO|react) \cdot Pr(break|react)$ where BO represents the event *bond order change*. Generalizing the equations above, the stability of a bond (i, j) with bond order o_{ij} is obtained as:

$$S_{ij} = 1 - Pr(react) \cdot Pr(BO|react)^{o_{ij}-1} \cdot Pr(break|react) \tag{10}$$

assuming that the events *bond order change* and *bond breakage* are independent.

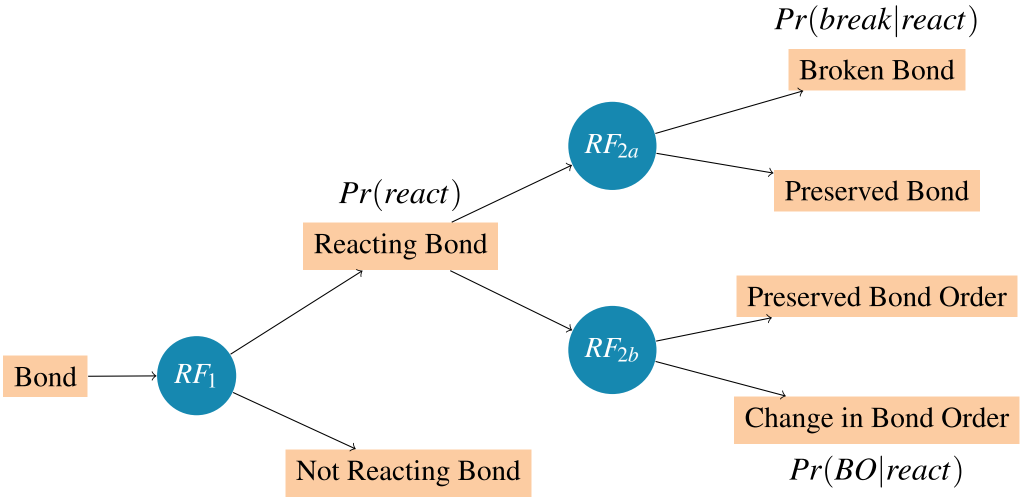


Figure 3: Hierarchical classification model for predicting the fate of a bond after a reaction occurs

The calculation of the probabilities that appear in equation (10), for determining the stability of a bond, is not trivial. The propensity of a bond to break depends on various factors on multiple levels that need to be considered: at a bond level (e.g., the species of the connected atoms, the topology of the bond and the neighboring atoms) at a molecular level (e.g., the presence of functional groups at more distant sites within the molecule) and at a reaction level (e.g., the reaction conditions and the enzyme that catalyzes the reaction). Here, we follow an ML approach for estimating those probabilities based on bond-level features and information regarding the surrounding atoms. More specifically, in order to capture the semantics discussed above, a hierarchical classification model is built as shown in Figure 3. In this representation, the circles correspond to binary classifiers while the rectangles correspond to the possible outcomes. The classifiers have been implemented as random forests (RF).[29] The first classifier (RF_1) is used in order to determine the reactivity of the bond, i.e., the probability that the bond reacts. At a second level and given that the bond reacts, two additional classifiers (RF_{2a} and RF_{2b}) are used in order to assess the likelihood of completely breaking the bond and the likelihood of changing bond order, respectively.

The RFs are trained on a set of labeled chemical bonds. We have constructed the set of labeled bonds, based on a set of atom mapped chemical reactions, as follows: each mapped reaction is decomposed into all bonds $(i, j) \in B_r \cup B_p$. Each bond is labeled as: (1) unchanged, (2) altered, i.e., there is a change in the bond order, (3) completely broken/formed. The RF_1 is trained to distinguish the reacting from the non-reacting bonds. All bonds that remain intact through the reaction are labeled as negative examples (label 1), while the bonds that undergo any change (labels 2 and 3) are labeled as positive. In the next level, two more RFs are trained for predicting 1) the probability for a single bond to break (RF_{2a}), and 2) the probability for a bond to change bond order (RF_{2b}). The RF_{2a} is trained only on first order bonds because here we model the complete breakage of a double or triple bond as a reduction in the bond order and a simultaneous breakage of the first order bond. The RF_{2b} is trained on all reacting bonds. The double and triple bonds that completely break/form (label 3) are considered positive examples for both RFs. It should be pointed out that since stereochemistry changes are not included in the objective function, there is no RF intended for predicting the probability that a bond changes stereochemistry.

For the construction of the training set, each bond of each chemical reaction creates a new entry. This means that it is possible to have multiple entries with identical feature vectors in the training set even under different labels. In theory, training a model with contradicting entries should be avoided. However, here we are interested in capturing the tendency of a bond to react rather than predicting the actual label. Therefore, the whole dataset is potentially more informative than removing either repeating or contradicting entries.

As a side note, the hierarchical architecture of the classification model was chosen in order to mitigate the high imbalance in the dataset used to train the RF classifiers which can negatively affect their performance. In particular, we observed that more than 90% of the entries in the bond dataset correspond to bonds that remain intact through the reaction since the big majority of bonds in a reaction do not react. This means that RF_1 will bias toward non-reacting bonds resulting in a classifier with very low sensitivity. The first level of the hierarchical model addresses this issue by filtering out unchanged bonds and creating less unbalanced training sets for the second layer classifiers. A more detailed evaluation of the random forest classifiers is provided in the supplementary material A (section S2). Regarding the classification algorithm for the estimation of the probabilities in equation (10), its basic core is a tree-structure classifier, named decision tree.[30] Decision trees is a method of inductive inference based on a set of training examples. It is a well-suited method for the task of predicting the reactivity of a bond because they create inductive rules that are interpretable by an expert, are invariant to uninformative features and can be applied in data that are represented by both categorical and numerical values without preprocessing. A random forest algorithm is an ensemble method that combines multiple uncorrelated decision tree classifiers, resulting in a more powerful classifier that is less prone to over-fitting.[29] The probability for an event to happen is calculated as the proportion of votes from the ensemble of trees for that event. For example the probability that a bond will react is the proportion of the trees that predict that this bond will react over the total number of trees.

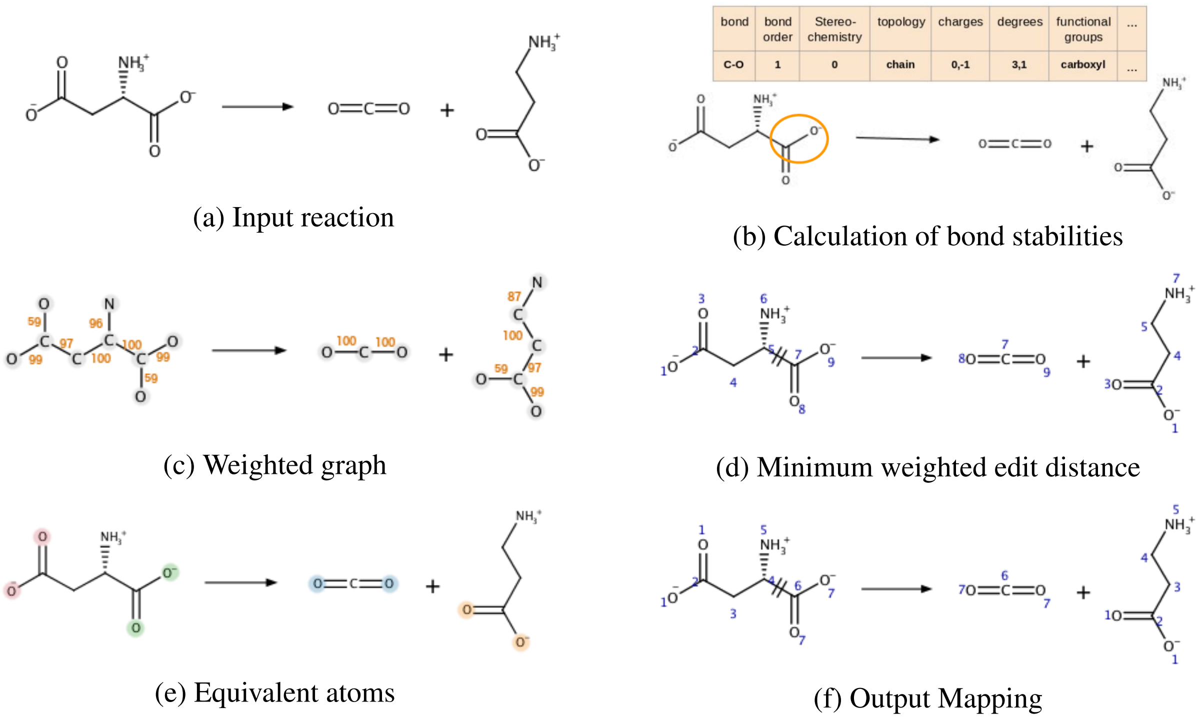
Bond representation For the calculation of the bond stabilities, each bond is represented as a vector which concatenates features that describe the bond locally, regarding the bond itself as well as the neighboring atoms. More specifically, the bond related features are the bond distance (computed based on the coordinates of the atoms as given in the mol or rxn files), bond order and bond stereochemistry. The topology of the bond, that is whether the bond is part of a ring or not, is also recorded. The features that describe each one of the bounded atoms are the atom species, the number of valence electrons, the atomic number, the charge and the number of atoms of each species attached to the bounded atoms. The presence of certain neighboring atoms or the formation of functional groups can also affect the reactivity of a bond. For that reason, the following information is also used in the representation of the bond: i) Whether a bond is part of a functional group. The functional groups that are considered here are: carboxyl group, ketone, aldehyde, ester, amide, phosphate and sulphate. ii) Whether the bond falls under one of the following cases: $x-C=O$, $x-C-O$, $x-CH-O$ where x is either a C, or N or S atom and $x-C$ is the bond under consideration.

The complete list of features that have been used for the representation of the chemical bonds is presented in the supplementary material A (section S1).

It is worth mentioning that since the bond descriptors are local features, the computed probabilities approximate the stability of each bond with respect to the molecule it belongs to, rather than the entire reaction system. With this method, we aim to capture the tendency of a bond to react rather than obtaining a precise prediction by accounting for all possible factors. Although the latter may seem desirable, a more fine representation 1) imposes the risk of overfitting and 2) may require additional input from the user (such as the enzyme number or the reaction conditions) which limits the usability of the method. Indeed, our results showed that even that crude approximation of the actual bond stability can be adequate to guide the search in the optimization problem most likely because the search space is limited by the known structure of the product molecules.

4.4 Alternative Mappings

Although the AM of a chemical reaction is unique, our method outputs multiple mappings in the following two cases: i) the optimization problem results in multiple optimal solutions, and ii) equivalent atoms appear in the reaction. In the first case, the alternative mappings correspond to multiple possible reaction mechanisms. In the case of equivalent atoms, usually the reaction mechanism is unique but there are atoms that cannot be distinguished and therefore multiple mappings are possible. Alternative mappings due to equivalent atoms are computed in a post-processing step and are indicated with the same atom mapping index: Once the optimal mapping (or mappings) has been computed by the optimizer then for each atom we find all equivalent atoms (as described in section 3) and then we update the computed atom mapping function f according to equations (1) and (2).



bond	bond order	Stereo-chemistry	topology	charges	degrees	functional groups	...
C-O	1	0	chain	0,-1	3,1	carboxyl	...

(a) Input reaction

(b) Calculation of bond stabilities

(c) Weighted graph

(d) Minimum weighted edit distance

(e) Equivalent atoms

(f) Output Mapping

Figure 4: Steps involved in the AMLGAM method: (a) The input reaction. (b) Each bond of the reaction is represented as a feature vector for the calculation of the bond stability. (c) The chemical reaction is represented as a pair of weighted graphs with the edge weights being the estimated bond stabilities. (d) The algorithm finds the mapping that corresponds to the minimum weighted edit distance. (e) Equivalences between atoms are determined. (f) The output mapping shows all alternative mappings.

The whole process for computing the AM of a chemical reaction in the AMLGAM method is summarized in Figure 4. The input to the system is the chemical reaction (as an rxn file). Each bond of the given reaction is represented as a vector of local features which is used to determine its stability using the hierarchical classification model. The input chemical reaction is represented as a pair of weighted graphs, the reactant and the product graph, in which the edge weights correspond to the bond stabilities. The AM is determined by minimizing the weighted edit distance between the reactant and the product graph. At a post processing step, equivalences between atoms are determined and the AM is updated accordingly. The method outputs all optimal mappings with the equivalent atoms indicated on each mapping.

5 Method Evaluation

5.1 Experimental Setup

The values of the hyperparameters of the random forests were chosen as follows: The depth of the trees and the number of trees in each forest were tuned based on a validation set from KEGG database. More specifically, we have constructed a dataset of annotated bonds, as described in section 4.3, from a dataset of 6,300 atom mapped reactions obtained by combining RPAIRS from KEGG release 73.1. The selection of the hyperparameters is based on the performance of the RFs on the classification task, i.e., predicting whether a bond reacts, completely breaks or changes bond order. This procedure resulted in a range of optimal values and the final selection of the parameter values was done empirically. In particular, we further increased the number of trees in the forests to ensure convergence of the computed probabilities while we avoided large values for the depth of the trees to prevent over-fitting. As a result, the maximum depth of the trees was tuned to 20 while the number of trees was tuned to 100 for all three RFs. The rest of the parameters were set empirically, as follows: The nodes are split using the 'Gini criterion' and the maximum number of features considered for finding the best split is set equal to the square root of the number of features. The RFs have been implemented using Python's scikit-learn library. For solving the MILP problem we have used the SCIP solver[31] version 4.0. We have set the time limit for the SCIP solver to 1 hour.

5.2 Data

We evaluate our method performing 10-fold cross validation on a manually curated dataset of 382 balanced chemical reactions. This dataset has been derived from a manually curated dataset of 512 metabolic reactions which was constructed for the purpose of comparing six existing AM algorithms in a recent comparative study by Gonzalez et al.[24] This dataset is part of the Recon 3D database[32] and is comprised of 340 reactions from BioPath database[33] and 172 additional reactions that have been added by the curators in an effort to obtain a dataset that represents all 6 EC classes. From the initial set of 512 reactions, we have removed all duplicate entries in order to ensure that the method is evaluated on reactions it has not seen during training. We have also compared our method against the AM tools that have been compared in Gonzalez et al. study[24] based on the results that they have reported.

We additionally run our method on a much larger and diverse dataset from MetaCyc[14] 21.1 database. In particular, this dataset is comprised of 7,380 balanced chemical reactions and 22 unbalanced. We have evaluated our method on the set of balanced chemical reactions by performing 10-fold cross validation. For the unbalanced reactions, we train the RFs using the whole set of 7,380 balanced reactions and we test our method on the set of 22 unbalanced reactions. Although the MetaCyc database provides AMs for more than 13,000 balanced chemical reactions, we have excluded reactions that involve molecules with missing structure (no available mol files), reactions with R-groups, and reactions with more than 200 atoms involved, leaving a dataset of 7,402 reactions. The last case (reactions with more than 200 atoms) corresponds to reactions that include molecules with cofactors and the computational complexity of the AM problem for reactions of that size is prohibitively large and out of the scope of this work. The AMs in MetaCyc are computationally derived based on the MWED approach by Latendresse et al.[12] For the case of unbalanced reactions, MetaCyc does not provide AMs. Also, in contrast to the Recon 3D dataset which has been manually curated, MetaCyc provides multiple alternative mappings for the reactions where the computational method returns multiple optimal solutions. Finally, the AMs in MetaCyc do not indicate equivalent atoms while the manually curated dataset from Recon 3D does. Table 1 summarizes information for the two datasets regarding the total number of reactions, the distribution of reactions in the six EC classes, the number of unbalanced reactions as well as the number of distinct reactant molecules.

Table 1: Statistics on the two datasets

Database	Reactions	Unbalanced	EC1	EC2	EC3	EC4	EC5	EC6	Reactants
Recon 3D	382	0	124	86	57	51	21	21	314
MetaCyc	7402	22	2515	2232	1157	881	358	237	5118

5.3 Evaluation criteria

The comparison between the computed mapping and the reference mapping is performed by comparing the corresponding sets of broken and formed bonds.[12] If the two bond sets are equal then the two mappings are considered equivalent. A chemical reaction is considered to be correctly mapped, with respect to a reference mapping, if the computed mapping is equivalent with the reference mapping. In the case of multiple computed mappings, we consider the reaction to be correctly mapped if at least one of the computed mappings is equivalent with the reference mapping. The equality between the bond sets is determined by taking into account equivalent atoms: two bonds are considered equivalent if they connect the same pair of atoms or if they connect pairs of equivalent atoms. For the Recon 3D dataset we consider equivalences between atoms as they are indicated by the curators. For the MetaCyc dataset we derive such equivalences computationally, as it is described in section 3, since the annotated mappings do not indicate such equivalences. For the MetaCyc dataset, which provides multiple mappings for certain reactions, we evaluate our method on whether there is an overlap between the computed mappings and the mappings provided by MetaCyc. For the unbalanced reactions, since MetaCyc does not provide AMs, we assess the ability of our method to handle such cases by manually inspecting the computed mappings.

5.4 Results and Discussion

In the following, we present the evaluation of our method on the two datasets from Recon 3D and MetaCyc as well as the comparison of the presented method against the tools that have been compared in Gonzalez's study.[24] Due to the stochastic nature of the RF classifiers, we present the average over 10 runs for all the metrics on the Recon 3D dataset while for the much larger MetaCyc dataset we present the results from a single run. We also briefly discuss our findings from a series of additional experiments performed in order to investigate various classification algorithms for the prediction of bond stabilities as well as additional features for the bond representation.

5.4.1 Recon 3D dataset

According to the results on the 382 distinct reactions from the Recon 3D dataset, the AMLGAM method, presented here, correctly mapped all atoms, on average, in about 359 reactions with an accuracy of 94%. The average time for the MILP solver to find a solution is about 29 seconds. From the set of 382 reactions, 2.6 reactions on average, exceeded the 1 hour time limit. The results on the Recon 3D dataset are summarized in Table 2.

Table 2: Results on the Recon 3D dataset

Number of reactions	382
Correctly mapped reactions	359.3 ± 0.82
Number of timed out reactions	2.6 ± 1.35
Overall accuracy (%)	94.06 ± 0.27
Average time (sec)	28.82 ± 6.54

We additionally analyzed the distribution of the enzymatic reactions over the six EC classes (categorization based on the enzyme that catalyzes the reaction) along with the performance of the AMLGAM method on each class, as shown in Figure 5. The plot shows the number of reactions and the percentage of wrong reactions for each EC class. The highest accuracy occurs for the reactions catalyzed by oxidoreductases and transferases which are the most represented classes in the dataset. The number of training instances from each class can affect the performance of the RF classifiers and cause underfitting if not sufficient training data are available. However, this result may be due to inherent characteristics of each EC class since it appears that the accuracy of other tools on the same dataset follows a similar pattern.[24]

We further assessed our method for correctly mapping carbon atoms, ignoring all other atoms, for two reasons: First, it appears that the accuracy on carbon atoms is of particular interest since many applications that make use of AM data, track only carbon atoms.[2,5] Second, this analysis can give us better insights into the capabilities and weaknesses of the ML-based method. More specifically, since there are more $C-C$ bond instances, given the nature of common biological compounds, we would expect a higher accuracy on carbon mappings for an ML-based algorithm. Indeed, our analysis showed that on average the accuracy on carbon atoms is 97.86% when the overall accuracy considering all atoms is 94.01%. The manual inspection of the reactions for which the computed mappings are not in agreement with the manually annotated mappings revealed that the majority of the wrong mappings concerns oxygen atoms. All reactions for which the computed mapping is not in agreement with the manually annotated mapping are illustrated in the supplementary material A with the differences between the two mappings being highlighted (section S3).

A common case among the reactions in which the computed mapping is not in complete agreement with the manually annotated (10 reactions out of 23 wrong reactions on average) presents the following pattern: (1) there is a reactant molecule that acts as an oxygen donor (usually a water molecule), (2) there is at least one reactant with at least one phosphate structure (or a sulphate) (3) the reaction center is located within the phosphate structure. The disagreement between the computed and the annotated mapping is on whether the phosphate group retains its oxygens or the oxygen from the water molecule is attached to the phosphorus atom after the reaction, as shown in

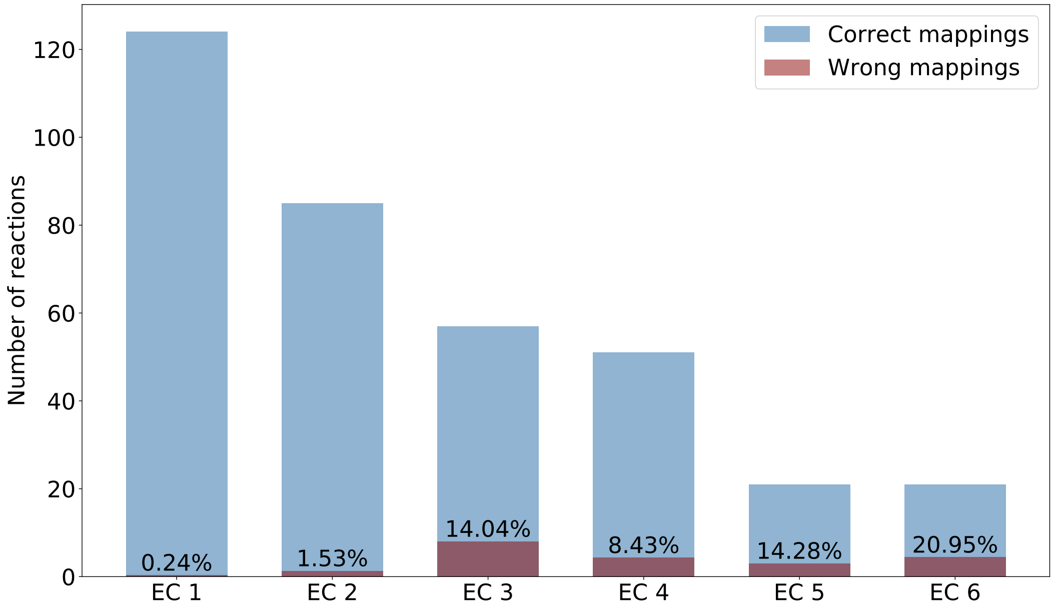
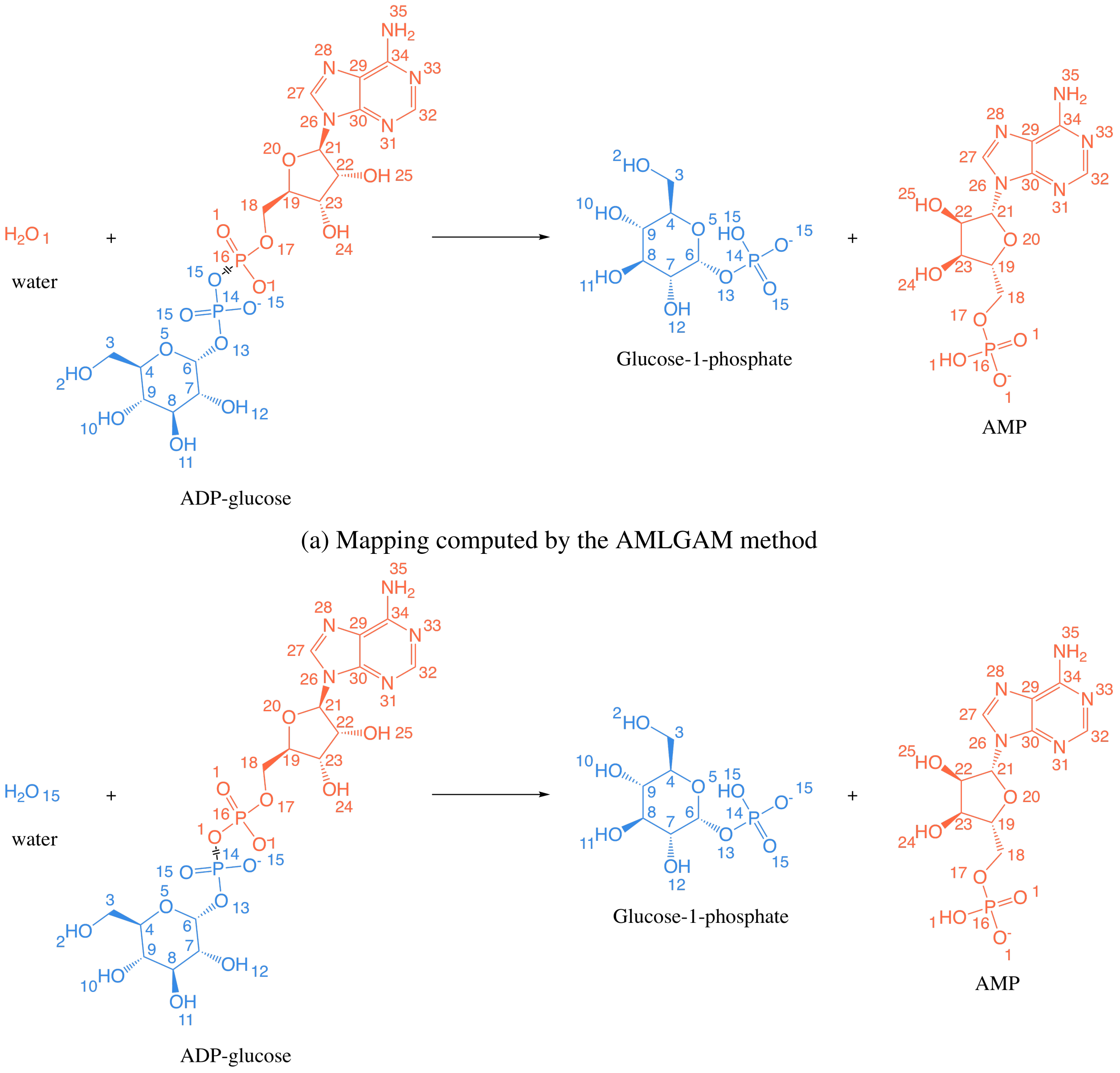


Figure 5: Percentage of reactions that are wrongly mapped by the AMLGAM method for each EC class

Figure 6. In the illustrated example, there is a diphosphate structure and the oxygen atom from the water molecule is attached to a different phosphorus atom regarding the two mappings. Although the curators provide only one mapping, it is not clear whether the annotated mapping is the only possible mapping or the reaction mechanism found by the AMLGAM method is also possible. However, for the evaluation of the accuracy of our method we consider all mappings that are not in complete agreement with the manually annotated mappings (regarding all atoms) as wrong.

Among the remaining reactions for which the computed mapping suggests a different reaction mechanism than the manual annotation, we have detected cases with complex reaction mechanisms that are very challenging for an AM tool. Figure 7 illustrates such an example in which even the manually annotated mapping is not accurate. The illustrated reaction is a critical step in glycolysis catalyzed by *phosphoglycerate mutase* (PGM). It converts *3-phosphoglycerate* to *2-phosphoglycerate* through an intermediate compound, *2,3-phosphoglycerate*, with the PGM enzyme contributing one phosphate group. Without knowledge of the intermediate compound, one may assume that this is a simple transfer of the phosphate group from one position to another which corresponds to the minimum edit distance and is what the AMLGAM method outputs. In reality, the *2-phoshoglycerate* retains the phosphate group that comes from the enzyme. Since the AMLGAM method is based on the assumption that each reaction proceeds with the minimum activation energy as a single step transformation, it cannot handle such a case. It should be noted though, that all AM methods that have been tested on that dataset, both optimization-based and common substructure-based, failed in this case as expected due to insufficient information.[24]

The manual inspection of the cases where the computed mapping was not in agreement with

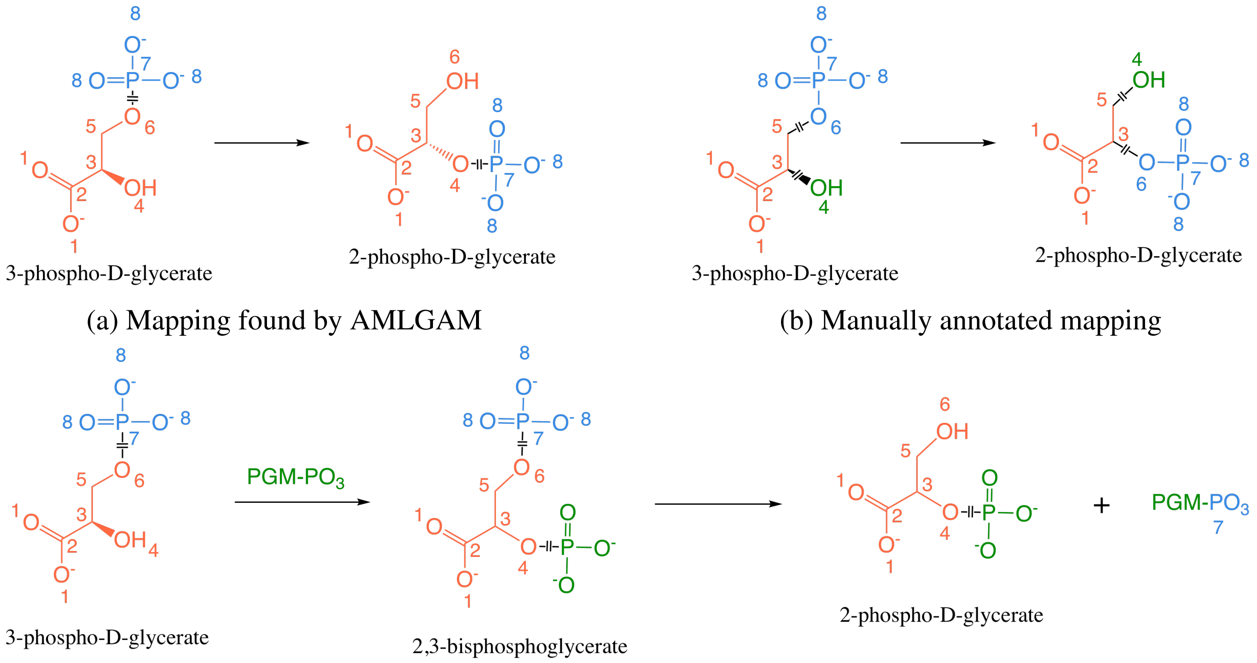


(a) Mapping computed by the AMLGAM method

(b) Manually annotated mapping

Figure 6: The AMLGAM method attaches the oxygen atom of the water molecule to the phosphate group of the *Glucose-1-phosphate* while the manually annotated mapping attaches it to the phosphate group of the *AMP* molecule

the manually annotated, revealed the following two cases: i) minor errors in the annotation of the equivalent atoms (7 cases), and ii) atom equivalences that are not indicated by the curators (6 cases). We have fixed all these cases and we provide the updated dataset in the supplementary file B. It should be mentioned that the errors of case (i) do not affect the accuracy of the compared tools in the earlier comparative study by Gonzalez et al.[24] on which our comparison relies on. For the second case though, it is not clear whether the non annotated equivalences are taken into account for determining the accuracy of the compared AM tools and in general whether any manual



(a) Mapping found by AMLGAM

(b) Manually annotated mapping

(c) Actual reaction mechanism

Figure 7: A chemical reaction with an intermediate step where both, the mapping computed by the AMLGAM method and the manually annotated mapping, do not depict the actual reaction mechanism

inspection was performed for fixing errors. However, the number of problematic cases we have detected is not big enough to create significant changes in the comparison we present here.

5.4.2 Comparison with existing AM tools on the Recon 3D dataset

The comparative evaluation of the AMLGAM method against six existing AM methods on the Recon 3D dataset showed that our ML-based method achieved the highest accuracy. The comparison has been performed based on the results that have been reported in the comparative study by Gonzalez et al.[24] Figure 8 shows the error rate of each method as a percentage of the reactions that disagree with the manually annotated mapping. The two existing MILP-based methods, the MED approach by First et al.,[16] called DREAM, as well as the MWED approach by Latendresse et al.[12] are among the compared algorithms. The comparison with those two algorithms is of particular interest in our study because our method consists an evolution of those two methods. From the remaining four evaluated algorithms, two of them are MCS based techniques (RDT[19] and CLCA[20]) and the other two (AutoMapper and ICMAP[7]) combine techniques from both approaches, optimization and MCS. The best scoring algorithms in terms of accuracy among the 6 existing methods were the RDT tool (91.31%), First's MED method (DREAM) (90.53%) and the CLCA algorithm (91.62%). Tested on the same dataset, the AMLGAM method had an average accuracy of 94.06%. We should mention though, that not all algorithms have been run on the whole dataset of 382 reactions because the authors did not have access in the code of all methods.[24] The number of reactions that have been mapped by each method is shown in Table 3. As an additional note, the accuracies of the existing methods, as presented here, have been adjusted after the removal of the duplicate reactions that existed in the initial dataset (see section 5.2). In the supplementary document C, we provide the results of each method for each reaction in the Recon 3D dataset.

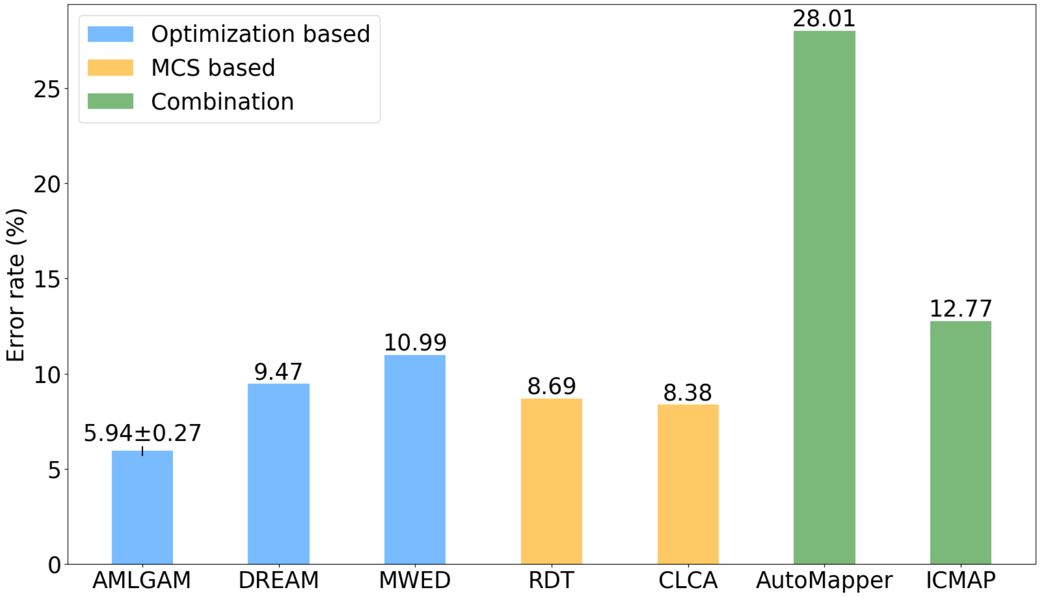


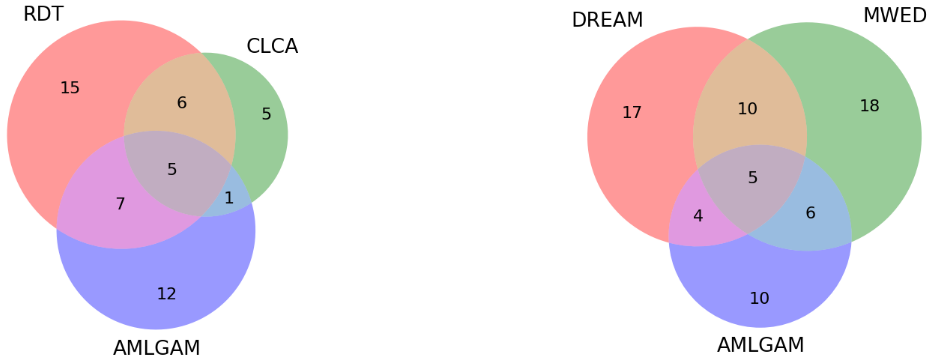
Figure 8: Percentage of wrongly mapped reactions for the compared AM methods on the set of 382 manually curated reactions from Recon 3D database

An interesting finding of the comparative study is that Latendresse's MWED approach did not score better comparing to the simple MED approach of the DREAM tool. This finding contradicts the outcome of an earlier comparison in which the MWED outperformed the DREAM tool.[12] These contradicting results can be an indication of over-fitting of the manually selected weights in the initial dataset which raises concerns regarding the selection of parameters that can handle larger datasets. On the other side, the ML-based MWED method, presented here, provides a more scalable method for determining the bond stabilities which appears to generalize better comparing to the manually chosen values. In the AMLGAM method the stabilities are determined by examining each bond individually and taking into account the chemical context while the manually chosen weights are less flexible regarding the variable chemical environment of each bond.

A more thorough examination of the reactions on which each method fails to find the correct

Table 3: Number of reactions tested by each algorithm

Algorithm	AMLGAM	RDT	CLCA	DREAM	MWED	AutoMapper	ICMAP
Reactions	382	380	203	380	355	382	368



(a) AMLGAM versus MCS techniques

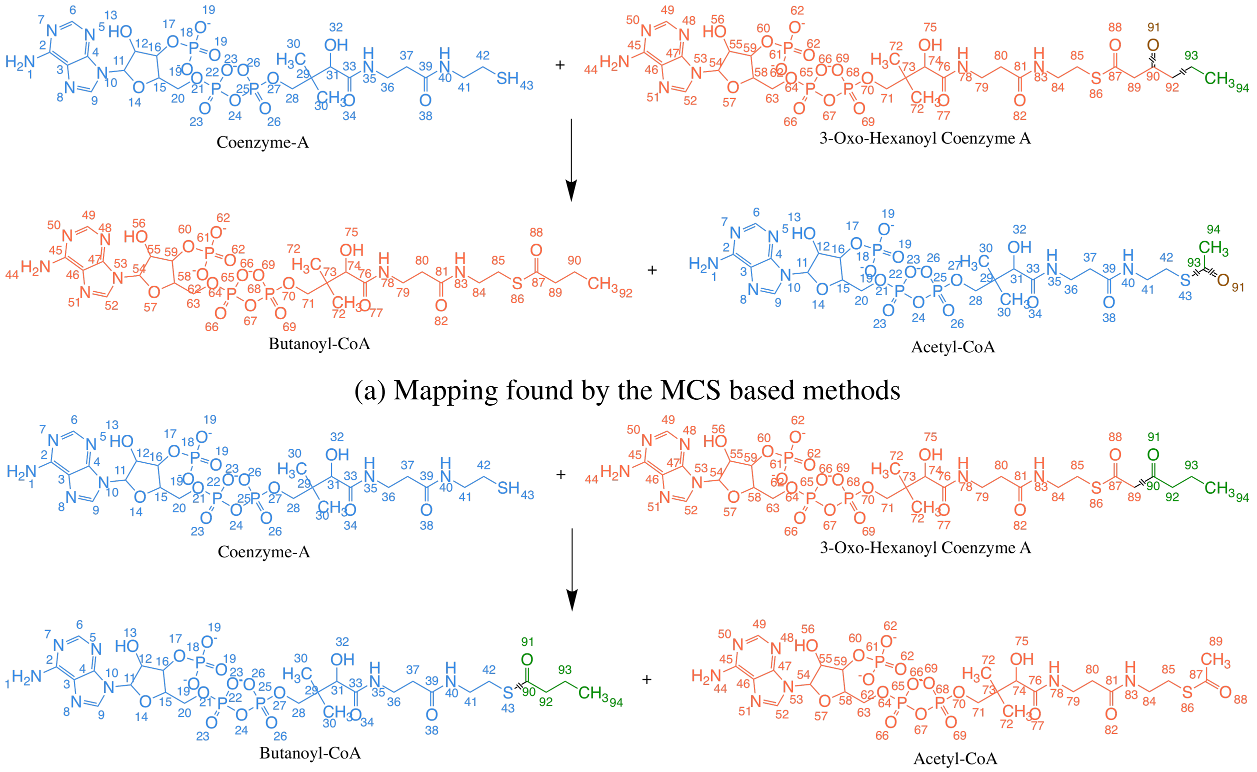
(b) AMLGAM versus optimization techniques

Figure 9: Venn diagrams showing the overlapping between the sets of wrong reactions

mapping revealed that the compared algorithms have different weaknesses. The overlapping between the sets of wrongly mapped reactions is presented in the Venn diagrams of Figure 9. More specifically, the diagrams show the overlapping between the MCS based methods and the AMLGAM method as well as the overlapping between the three optimization (MILP)-based methods including AMLGAM. Indeed, the only noteworthy overlapping occurs between the two MCS based methods. In particular, about 65% of the cases that are wrongly mapped by the CLCA method are also wrongly mapped by the RDT tool. Although this may seem a tentative conclusion, since the CLCA tool has been evaluated on a smaller set of reactions (as shown in Table 3), the manual inspection showed that, indeed, some of the common errors correspond to intrinsic limitations of the MCS approach. More specifically, the common errors between the MCS based methods correspond to reactions that include reactant molecules with common substructures which are preserved through the reaction. One such example is shown in Figure 10 where the two reactant molecules, *Coenzyme-A* and *3-Oxo-Hexanoyl Coenzyme-A* share a common substructure. This reaction scheme is common to many metabolic reactions in fatty acid oxidation. In such a case, an MCS based algorithm may mismatch the preserved structures. This observation confirms Arita's findings who had highlighted the cases on which the MCS based methods fail to identify the correct mapping.[1] On the other side, the optimization based techniques were capable of correctly mapping these cases. Therefore, although the MCS based methods achieved an overall high accuracy it appears that they have intrinsic limitations that the optimization methods can address. However, the capabilities of the optimization-based methods are limited by the assumption of the minimum energy that does not hold for more complex reactions such as reactions with intermediate steps.

5.4.3 MetaCyc dataset

Balanced Reactions Regarding the MetaCyc dataset of 7,380 balanced reactions, we obtained the following results: Our method exceeded the time limit of one hour in 4.6% of the cases (337 reactions). For the remaining cases in which the MILP solver found at least one solution, we compared the mappings computed by the AMLGAM method against the atom mappings provided by the MetaCyc database. The agreement between the two mappings was around 90.16% excluding



(a) Mapping found by the MCS based methods

(b) Mapping found by the AMLGAM method in agreement with the manually annotated mapping

Figure 10: A case where the MCS based methods fail because there are common substructures in the reactants that are preserved through the reaction

timed-out reactions and 86.04% if timed-out reactions are taken into account. The comparison was performed with respect to all atoms except hydrogens which are not mapped by both methods. It should be noted that we have not manually inspected the cases where the mappings computed by the AMLGAM method are not in agreement with the MetaCyc mappings (due to the large size of the dataset) and therefore we believe that the calculated percentage is a lower bound of the accuracy of our method. Furthermore, taken into account that the MetaCyc dataset is much larger, diverse, and noisy comparing to the Recon 3D dataset, the relatively high percentage of agreement shows the potency of the ML-based method. We recall here that the mappings in MetaCyc are computationally derived.

Unbalanced Reactions For the evaluation of the ability of the AMLGAM method to handle unbalanced reactions, we have divided the dataset of unbalanced reactions into three categories: 1) reactions with a small number of missing atoms (1–3 atoms), 2) reactions with wrong stoichiometry, 3) reactions with incomplete information for computing the atom mapping. The first category consists mostly of reactions with missing oxygens, which most likely imply missing water molecules. In this category, the missing atoms appear not to be crucial for the computation of the AM. The second category refers to cases in which the involved molecules are known but the stoichiometry is wrong resulting in unbalanced reactions with usually many missing atoms. The last category contains reactions in which not all involved compounds are known and the structure of the missing molecules is crucial for determining the AM. Table 4 presents the number of reactions and the average number of missing atoms per category as well as the number of reactions for which the AMLGAM method computed a mapping and the number of reactions for which the method timed out without finding a solution.

Table 4: Statistics and results on the dataset of unbalanced reactions

	category 1	category 2	category 3
Number of reactions	9	11	2
Average number of missing atoms	1.3	32.8	19
Number of mapped reactions	9	4	2
Number of timed out reactions	0	7	0

The AMLGAM method computed mappings for all reactions in categories 1 and 3 but timed out for most of the reactions of category 2 which includes the reactions with the highest number of missing atoms on average. Since MetaCyc does not provide AMs for unbalanced reactions we were not able to evaluate the accuracy of the computed mappings. However, manual inspection of the computed mappings gives us high confidence for the AMs of the category 1 reactions and lower confidence for the category 2 reactions, based on the given information. For category 3, the assessment of the computed mapping was not easy since the available information was not adequate to determine the reaction mechanism by inspecting the given reactants and products. In the supplementary material A (section S4), we provide the mappings for all unbalanced reactions for which the AMLGAM method computed a mapping within the one-hour time limit.

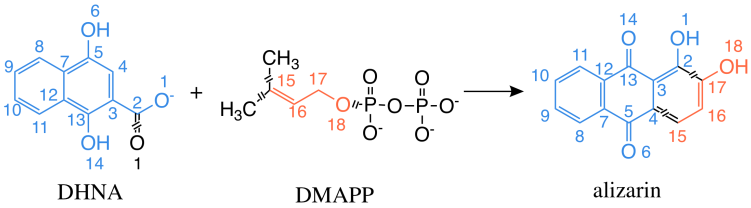


Figure 11: A category 3 reaction where the structure of the missing molecules is crucial for determining the reaction mechanism

In the case of an unbalanced reaction, the AMLGAM method tries to find a maximal mapping allowing unmapped atoms. Figure 11 shows an example of a category 3 reaction in which a diphosphate molecule is missing in the products side and therefore the atoms in the diphosphate group of DMAPP in reactants as well as the two extra carbons are left unmapped. The post-processing step, though, may map such atoms if equivalences between atoms exist. This step is more crucial for category 2 reactions where the wrong stoichiometry may be fixed, to some extend, by identifying equivalent atoms. This effect can be seen in the reaction of Figure 12 where the reactant glutathione molecule is mapped twice in the products side, by identifying the symmetry in glutatione disulfide, implying the existence of two glutathione molecules in reactants. However, most reactions of category 2 are more complicated and either the structure of the missing molecules is not preserved or additional molecules are missing. Indeed, the reaction of Figure 12 is a more complicated case where in addition to the wrong stoichiometry in the reactants side, a water molecule is also missing in the products side. Despite the fact that there are missing atoms in both sides of the reaction, which contradicts the assumption that the imbalance occurs only in one side of the reaction (constraint 6 in section 4.2), the AMLGAM method manages to find a reasonable mapping.

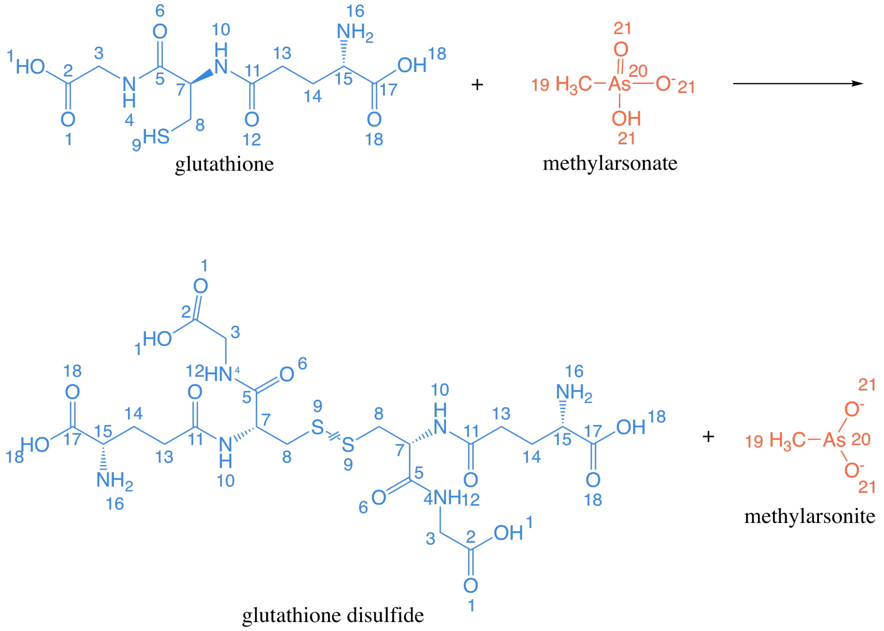


Figure 12: An unbalanced reaction that the AMLGAM method cannot handle because molecules are missing from both sides. The reaction equation can be balanced by adding one glutathione molecule in reactants and one water molecule in products.

5.4.4 Additional experiments

In order to understand the influence of the classification algorithm and the selected features for the bond representation, we performed a series of additional experiments. More specifically, regarding the classification task we investigated the following algorithms: logistic regression, k-nearest neighbors (k-NN), neural networks with logistic loss and random forests. The classification algorithms were evaluated based on the accuracy in the classification task, i.e., predicting the fate of a bond in a chemical reaction. Based on our results, the random forest had the best performance while the k-NN algorithm was found to be the weakest classifier. A comparative analysis of the above models is not presented here as it is out of the scope of this study. It is worth mentioning, though, that our results showed that the performances of the above algorithms did not have significant deviations. On the other side, the choice of the features for representing the bonds appeared to have a significant impact. Although we attempted to perform feature selection prior to the application of the classification algorithm it turned out to worsen the performance. These findings highlight the importance of the selected descriptors in QSAR studies that seek to correlate molecular structures with certain activities.

In our approach, the bond features are manually selected based on our intuition on what affects the reactivity of a bond. Although the present method could facilitate the inclusion of additional features, other than those discussed in section 4.3, we settled in this representation in an effort to 1) avoid overfitting and 2) minimize the input from the user. However, it worths mentioning that we have investigated the inclusion of enzyme information using the EC identification numbers. This addition, though, did not seem to affect the accuracy of the atom mappings on the Recon database. On top of that, many reactions in chemical databases have not been assigned an EC number and therefore requiring such input from the user could be regarded as a limitation. Finally, we should clarify that the bond distance, which is used here as a feature in the bond representation, is calculated based on the atom coordinates of the rxn files. Based on our experiments, the bond distance did not affect the results on the Recon dataset but it had a more significant impact on the MetaCyc dataset. Since this information was available for both datasets, we finally included this feature in the bond representation.

6 Conclusion

We have presented an optimization-based approach for automatically determining the atom mapping (AM) of a chemical reaction, called AMLGAM (Automated Machine Learning Guided Atom Mapping). The computed mapping corresponds to the reaction mechanism which favors the breakage/formation of the less stable bonds. In this work, we define the stability of a bond using a probabilistic framework and we use machine learning techniques for its estimation based on local topological and atomic features that characterize the bond and its surrounding. The optimization problem is solved as a MILP problem which we develop such that the method can handle unbalanced reactions. The AMLGAM method outputs all optimal solutions and indicates equivalent mappings due to indistinguishable atoms. We have evaluated the accuracy of our method on a set of 382 manually curated balanced chemical reactions and we have run our method on a much larger and diverse dataset of 7,400 chemical reactions including unbalanced reactions. We have also compared our method against six AM tools, including common substructure-based and optimization-based methods, based on results from a previous study. The comparison showed that the AMLGAM method achieved the highest accuracy. In particular, we show that it has improved the accuracy of the previous optimization-based techniques while it has correctly handled the intrinsic weaknesses of the MCS-based methods. Tested on a set of unbalanced chemical reactions we showed that our method is capable of dealing with reactions with a small number of missing atoms without the need for re-balancing the reaction equation.

Acknowledgement

The authors thank Dr Ronan Fleming and German Andres Preciat Gonzalez from the Luxembourg Centre for Systems Biomedicine for providing the most updated version of their comparative evaluation of the six atom mapping algorithms.[24] We also thank Dr Mario Latendresse from SRI international for providing clarifications on the implementation of the MWED method.[12] We also thank Sarah Kim for synthesizing and providing the KEGG dataset using the KEGG RPAIRS. Finally, we thank the reviewers for their constructive feedback.

Supporting Information Available

1. Supplementary material A: It includes 1) the features used for the bond representation, 2) evaluation metrics for the random forest classifiers, 3) the disagreements between the AMLGAM mappings and the manually annotated mappings on the Recon 3D dataset, and 4) the AMLGAM mappings for the unbalanced reactions from MetaCyc.

2. Supplementary material B: Updated dataset of the 382 manually curated reactions from Recon 3D database.

3. Supplementary material C: The results of all compared methods, including AMLGAM, on the dataset of 382 reactions. The reactions for which the computed mapping is in agreement with the annotated mapping for a given method are marked as 1 while the disagreements as 0. A NaN entry means that the method has not been tested for that reaction. The results on the existing AM tools (all methods except AMLGAM) correspond to the most updated version of the comparative evaluation conducted by German et al.

Funding

Work on this project by EEL, MIP, MM, GNB and LEK was supported by NSF ABI 1262491 and Rice University funds. Work by GG and, in part, EEL was supported by NCSR Demokritos.

Author's contributions

EEL developed and implemented the method and run the experiments. MIP helped with the analysis of the experiments and supervised the bio-chemical part. GG supervised the computational part. MM, GNB and LEK conceived the project and supervised the overall analysis. EEL wrote the manuscript. All authors approved the manuscript.

References

(1) Arita, M. In Silico Atomic Tracing by Substrate–Product Relationships in Escherichia coli Intermediary Metabolism. *Genome Research* **2003**, *13*, 2455–2466.

(2) Heath, A.; Bennett, G.; Kavraki, L. Finding metabolic pathways using atom tracking. *Bioinformatics* **2010**, *26*, 1548–1555.

(3) Kim, S. M.; Pena, M. I.; Moll, M.; Bennett, G.; Kavraki, L. E. A review of parameters and heuristics for guiding metabolic pathfinding. *Journal of Cheminformatics* **2017**, *9*.

(4) He, S.; Li, M.; Ye, X.; Wang, H.; Yu, W.; He, W.; Wang, Y.; Qiao, Y. Site of metabolism prediction for oxidation reactions mediated by oxidoreductases based on chemical bond. *Bioinformatics* **2017**, *33*, 363–372.

(5) Zupke, C.; Stephanopoulos, G. Modeling of Isotope Distributions and Intracellular Fluxes in Metabolic Networks Using Atom Mapping Matrixes. *Biotechnol Progress* **1994**, *10*, 489–498.

(6) Ravikirthi, P.; Suthers, P. F.; Maranas, C. D. Construction of an E. Coli genome-scale atom mapping model for MFA calculations. *Biotechnology and Bioengineering* **2011**, *108*, 1372–1382.

(7) Kraut, H.; Eiblmaier, J.; Grethe, G.; Low, P.; Matuszczyk, H.; Saller, H. Algorithm for reaction classification. *Journal of Chemical Information and Modeling* **2013**, *53*, 2884–2895.

(8) Moock, T. E.; Nourse, J. G.; Grier, D.; Hounshell, W. D. In *Chemical Structures*; Warr, W., Ed.; Springer-Verlag: Heidelberg, Berlin, 1988; pp 303–313.

(9) Bonchev, D.; Rouvray, D. *Chemical Graph Theory: Introduction and Fundamentals*; Abacus Press: New York; London, 1991.

(10) Akutsu, T. Efficient Extraction of Mapping Rules of Atoms from Enzymatic Reaction Data. *Journal of Computational Biology* **2004**, *11*, 449–462.

(11) Korner, R.; Apostolakis, J. Automatic Determination of Reaction Mappings and Reaction Center Information. 1. The Imaginary Transition State Energy Approach. *Journal of Chemical Information and Modeling* **2008**, *48*, 1181–1189.

(12) Latendresse, M.; Malerich, J.; Travers, M.; Karp, P. Accurate Atom-Mapping Computation for Biochemical Reactions. *Journal of Chemical Information and Modelling* **2012**, *52*, 2970–2982.

(13) Kanehisa M., G. S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **2000**, *28*, 27–30.

(14) Caspi, R.; Altman, T.; Billington, R.; Dreher, K.; Foerster, H.; Fulcher, C.; Holland, T.; Keseler, I.; Kothari, A.; Kubo, A.; Krummenacker, M.; Latendresse, M.; Mueller, L.; Ong, Q.; Paley, S.; Subhraveti, P.; Weaver, D.; Weerasinghe, D.; Zhang, P.; Karp, P. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **2014**, *42*, D459–D471.

(15) Altman, T.; Travers, M.; Kothari, A.; Caspi, R.; Karp, P. D. A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics* **2013**, *14*, 112.

(16) First, E.; Gounaris, C.; C.A., F. Stereochemically Consistent Reaction Mapping and Identification of Multiple Reaction Mechanisms through Integer Linear Optimization. *Journal of Chemical Information and Modelling* **2012**, *52*, 84–92.

(17) Chen, W. L.; Chen, D. Z.; Taylor, K. T. Automatic reaction mapping and reaction center detection. *WIREs Computational Molecular Science* **2013**, *3*, 560–593.

(18) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of Computer-Aided Molecular Design* **2002**, *16*, 2002.

(19) Rahman, S.; Torrance, G.; Baldacci, L.; Cuesta, S.; Fenninger, F.; Gopal, N.; Choudhary, S.; May, J.; Holliday, G.; Steinbeck, C.; Thornton, J. Reaction Decoder Tool (RDT): extracting features from chemical reactions. *Bioinformatics* **2016**, *32*, 2065–2066.

(20) Kumar, A.; Maranas, C. CLCA: maximum common molecular substructure queries within the MetRxn database. *Journal of Chemical Information and Modeling* **2014**, *54*, 3417–3438.

(21) Dugundji, J.; Gasteiger, J.; Ugi, I. The Principle of Minimum Chemical Distance (PMCD). *Angewandte Chemie International Edition* **1980**, *19*, 495–505.

(22) Apostolakis, J.; Sacher, O.; Körner, R.; Gasteiger, J. Automatic Determination of Reaction Mappings and Reaction Center Information. 2. Validation on a Biochemical Reaction Database. *Journal of Chemical Information and Modeling* **2008**, *48*, 1190–1198.

(23) Fooshee, D.; Andronico, A.; Baldi, P. ReactionMap: An Efficient Atom-Mapping Algorithm for Chemical Reactions. *Journal of Chemical Information and Modeling* **2013**, *53*, 1549–9596.

(24) Preciat Gonzalez, G. A.; El Assal, L. R. P.; Noronha, A.; Thiele, I.; Haraldsdóttir, H. S.; Fleming, R. M. T. Comparative evaluation of atom mapping algorithms for balanced metabolic reactions: application to Recon 3D. *Journal of Chemoinformatics* **2017**, *9*.

(25) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in chemoinformatics and drug discovery. *Drug Discovery Today* **2018**, *23*, 1538–1546.

(26) Danishuddin,; Khan, A. U. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discovery Today* **2016**, *21*, 1291–1302.

(27) Latino, D. A. R. S.; Zhang, Q.-Y.; Aires-de Sousa, J. Genome-scale classification of metabolic reactions and assignment of EC numbers with self-organizing maps. *Bioinformatics* **2008**, *24*, 2236–2244.

(28) Mullert, C.; Marcou, G.; Horvath, D.; Aires-de Sousa, J.; Varnek, A. Models for Identification of Erroneous Atom-to-Atom Mapping of Reactions Performed by Automated Algorithms. *Journal of Chemical Information and Modeling* **2012**, *52*, 3116–3122.

(29) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

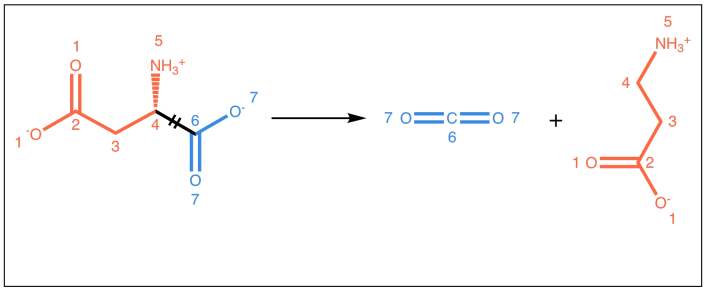
(30) Quinlan, J. Induction of Decision Trees. *Machine Learning* **1986**, *1*, 81–106.

(31) Maher, S. J. et al. *The SCIP Optimization Suite 4.0*; 2017.

(32) Brunk, E.; Sahoo, S.; Daniel, Z.; Altunkaya, A.; Prlić, A.; Mih, N.; Sastry, A.; Preciat, G. G.; Danielsdottir, A.; Noronha, A.; Aurich, M.; Rose, P.; Fleming, R.; Thiele, I.; Palsson, B. Recon 3D: A resource enabling a three-dimensional view of gene variation in human metabolism. *Nature Biotechnology* **2018**, *36*, 272–281.

(33) Forster, M.; Pick, A.; Raitner, M.; Schreiber, F.; Brandenburg, F. The system architecture of the BioPath system. *In Silico Biology* **2002**, *2*, 415–426.

Graphical TOC Entry


1
O
5
NH_3^+
7
1 ^-O
2
3
4
6
O^-
O
7
7 O
C
O 7
6
+
5
NH_3^+
4
3
1 O
2
O^-
1