Conjecturing Hidden Entities by Means of Simplicity and Conservation Laws: Machine Discovery in Chemistry

Raúl E. Valdés-Pérez

School of Computer Science and
Center for Light Microscope Imaging and Biotechnology
Carnegie Mellon University; Pittsburgh, PA 15213
internet address: (valdes@cs.cmu.edu)

Abbreviated Title: **Conjecturing Hidden Entities in Chemistry**

Abstract

We show that combinatorial search, constrained by experimental evidence, domain knowledge, and simplicity, is sufficient to discover credible explanatory hypotheses in a scientific task of current importance.

The MECHEM system automates the scientific task of determining the pathway of steps underlying a chemical reaction. MECHEM carries out a systematic search for pathway hypotheses that satisfy constraints and explain the available evidence. During its search, the system conjectures unseen entities as needed to formulate hypotheses, but only after determining by experimentation or other means that all simpler hypotheses are inadequate. The generality of MECHEM's methods is established by pointing out similarities to other scientific tasks and programs.

1. Introduction

This paper will show that combinatorial search, constrained by experimental evidence, domain knowledge, and simplicity, is sufficient to discover credible explanatory hypotheses in a scientific task of current importance. This result suggests that other scientific tasks are susceptible to similar approaches, defining a class of scientific problems that are well regarded as a combinatorial search constrained by simplicity.

The MECHEM system automates the problem of determining the pathway of steps underlying a chemical reaction. MECHEM carries out a systematic search for pathway hypotheses that satisfy constraints and explain the available evidence. During its search, MECHEM conjectures unseen entities as needed to formulate hypotheses, but only after determining by experimentation or other means that all simpler hypotheses are inadequate.

Our intent has been to study deeply a specific, creative, scientific task, with the goal of successfully automating the task or creating a truly useful assistant. A successful automation serves to guarantee the efficacy of whatever methods are invented or adapted for the task. The generality of the methods is to be established partly by further experiments, and partly by analysis of whether the task had peculiar characteristics unlikely to be found elsewhere.

We proceed first by describing the task addressed by MECHEM and reviewing the needed chemistry knowledge. Then the design of MECHEM is presented in detail, with a focus on the hypothesis-generation procedure. Then the system is applied to two chemical reactions to show how MECHEM is able to discover credible hypotheses. Finally, MECHEM is compared to other AI programs with respect to the methods used and the domain addressed.

2. The Chemistry Problem

This section introduces the task, describes the elementary chemical knowledge needed to understand it, and places the task in its scientific context. Finally, prior work in AI on automating the task is discussed.

2.1. Elucidation of reaction pathways

The reaction-pathway elucidation problem in chemistry is as follows. Given empirical data on starting materials and observed products (not necessarily final products) of a specific chemical reaction,

1. infer simple, plausible pathways that explain the empirical data and are consistent with background theory, and
2. seek and apply further evidence to discriminate among the plausible pathways.

Figure 1: Reaction Products of Ornithine, Ammonia, and Carbon Dioxide

.

For example, we consider a reaction that consumes the starting materials ornithine, ammonia, and carbon dioxide and that is observed to form the products urea, arginine, and water. A depiction of the reaction in term of molecular structures is in Figure 1.

These reaction data can be explained by either of these two pathways, or sets of reaction steps:

1. $$\begin{aligned} ornithine + ammonia + carbon\ dioxide &\rightarrow water + C_6H_{13}N_3O_3 \\ ammonia + C_6H_{13}N_3O_3 &\rightarrow arginine + water \\ arginine + water &\rightarrow urea + ornithine \end{aligned}$$

2. $$\begin{aligned} ornithine + carbon\ dioxide &\rightarrow water + C_6H_{10}N_2O_3 \\ 2\,(ammonia) + C_6H_{10}N_2O_3 &\rightarrow arginine + water \\ carbon\ dioxide + arginine &\rightarrow urea + C_6H_{10}N_2O_3 \end{aligned}$$

in which the unseen chemical species $C_6H_{13}N_3O_3$ and $C_6H_{10}N_2O_3$ have been conjectured, i.e., were not observed, hence were not part of the initial data. (ornithine is $C_5H_{12}N_2O_2$, ammonia is NH_3, carbon dioxide is CO_2, water is H_2O, arginine is $C_6H_{14}N_4O_2$, and urea is CH_4N_2O. The reader can verify that all steps are balanced.) The first pathway shown was a discovery made by Hans Krebs [13] which was studied historically by Holmes [8], and which was the subject of a cognitive model by Kulkarni and Simon [14].

The *idealized* problem of pathway elucidation is to find the exact set of elementary steps that occur in a given chemical reaction. In current practice it is infeasible to determine the pathway with certainty, for several reasons. The most recalcitrant reason is that, given usual experimental data and a pathway to explain those data, one can always create a more elaborate pathway that explains the data equally well. Hence, the *practical* problem of pathway elucidation is to find a set of steps that is consistent with all or most of the experimental data, is consistent with current knowledge of what types of reaction tend to occur (background theory), and is simple, i.e., no more complex than needed.

2.2. Pertinent chemistry knowledge

This section describes the minimal chemistry knowledge needed to understand the task domain of MECHEM. To supplement this description, we recommend [12] as an elementary introduction to the basic concepts.

The notation $A + B \rightarrow X + Y$ expresses a process, or reaction, by which the chemical substances A and B are transformed into new substances X and Y. The starting materials are the chemicals that are placed in a medium to start a reaction or sequence of reactions. The notation '$\leadsto$' is introduced to indicate a qualitative input/output relation among substances, e.g., $A, B \leadsto X, Y, Z$ means that A, B are starting materials and X, Y, Z are formed during the reaction. The materials need not balance across the '$\leadsto$,' since a complete accounting of the mass is not implied.

A molecular formula (or just formula) is a list of the type and number of atoms making up a chemical substance, e.g., CH_2O for formaldehyde. In this paper, a chemical species will denote a unique formula, and vice versa.

An elementary reaction is a single process in which one or more molecules are re-arranged into new products, but no constituent reactions underlie the elementary reaction. During an elementary reaction among multiple reactant molecules, the reactants must collide in order for the reaction to occur. Reaction steps as depicted here are *not necessarily* elementary unless stated otherwise. A pathway is defined as a set of reactions that are not necessarily elementary. Each individual reaction of a pathway is a step.

Ordinary chemical reactions conserve mass; molecules are transformed into product molecules only in ways that conserve total numbers and types of atoms. The concentration of a reaction species refers to the number of species molecules per unit volume; concentrations can vary over time as species react or are formed.

2.3. Role of pathway elucidation in chemistry

Chemists often undertake to elucidate a reaction pathway with the goal of adding to scientific knowledge. A second reason arises in the context of enhancing the yield of a synthesis, because even partial knowledge of the pathway can suggest changes to the reaction conditions that may improve the yield. For example, another reagent might be introduced in order to increase the rate of a desirable step, or to alter its equilibrium.

Studies of the pathways underlying chemical reactions appear frequently in chemistry and chemical engineering journals. Even a paper whose main purpose is to describe a new chemical synthesis typically includes at least some discussion of the pathway underlying the synthesis (e.g., see [37]).

The current state of automated pathway elucidation is evaluated in these recent, skeptical comments [4] (the term 'mechanism' is used instead of 'pathway,' but the two are closely related):

> It is not possible to prescribe a systematic procedure for the definition of a reaction mechanism. A novice might think that a set of standard moves might be made in predetermined order to enable the chemist to move steadily toward his goal; these might include certain standard preliminary experiments, a stage of initial decisions whereby certain broad categories of possibilities were eliminated, further standard experiments appropriate to the remaining categories, another stage of decision-making, and so forth until the work was finished. Although there are certain general questions that the investigator is advised to answer by experiment (...), such a straightforward approach seldom leads directly to an adequate definition of the reaction mechanism. The diversity of mechanistic possibilities is so great and still so incompletely known that a neat outline of how to determine a reaction mechanism cannot at present be constructed.

One goal of our research is to provide a systematic and reliable procedure for the discovery and experimental test of pathway hypotheses. A fuller discussion of pertinent chemistry knowledge, and of prior attempts to automate reasoning about chemical reactions, is contained in [29].

2.4. Prior work

The only prior work in AI on discovering reaction pathways is that of Soo et al. [24] and Kulkarni and Simon [14]. Soo et al. describe a prototype program that uses published rules to interpret experimental concentrations data represented in a special form (Lineweaver-Burk plots). Given a catalogue of reaction pathways, the program tested each pathway against plausibility constraints inferred using the rules. Hypothesis generation was not otherwise addressed, nor reasoning about other types of experimental evidence. Also, knowledge of all reaction species and their roles in the reaction (i.e., reactant or product) was assumed.

Kulkarni and Simon developed a cognitive model of the complete line of investigation leading to the discovery of a biochemical reaction pathway by Hans Krebs, but they focused on the heuristic strategies that Krebs used to design and interpret experiments. They did not attempt to automate in a systematic manner the general problem of reaction-pathway elucidation.

3. Automated Discovery of Chemical Reaction Pathways

Figure 2 is a block diagram of a complete system in which the implemented parts of MECHEM would be embedded; the only aspect of the diagram not currently addressed by our work is deciding whether the current hypotheses and evidence constitute a scientific contribution (deciding when to suspend work on a problem is discussed briefly in [29]).

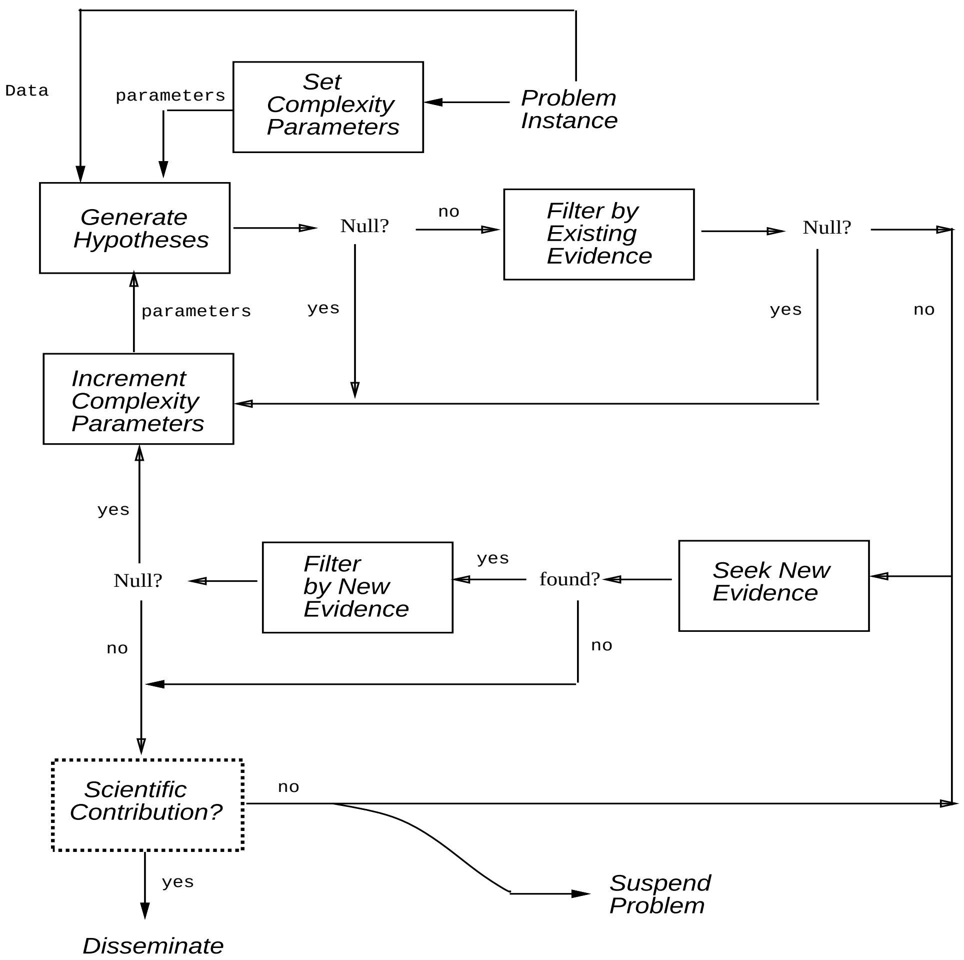

Figure 2: A Model for Discovery of Reaction Pathways

As depicted, the theory-formation cycle starts with the initial data from a problem instance. These data suggest initial values of the hypothesis complexity parameters, which are the number of steps and species to be contained in a pathway hypothesis. Then the hypothesis generator cycles until it finds a non-empty set of pathways, at which point these are tested against the available evidence not already incorporated in the generator. If all pathways are discredited to the point of rejection, then hypothesis generation resumes with incremented complexity parameters.

If the current pathways explain the existing evidence, then an experimentation cycle is entered. Initially, evidence is sought to discriminate among them. Any evidence found is applied to filter the pathways; if all of the latter are rejected, then hypothesis formation resumes after incrementing the complexity. If some pathways survive, then one might decide that a scientific result has been achieved, e.g., a small set of highly credible pathways.

If, on the other hand, no evidence can be found with an effort deemed appropriate, then still one might decide that a disseminable result has been reached, or perhaps decide to suspend the problem, or decide to incur a greater cost in order to find more evidence. Strategic decisions such as whether to suspend work on a problem, publish one's results, or incur greater cost to obtain evidence are not yet addressed by our work. Their inclusion in the diagram is intended to convey a picture of the theory-formation cycle that complements MECHEM, which to date primarily addresses detailed tactical reasoning about hypotheses and evidence.

This paper will emphasize MECHEM's hypothesis generator, which carries out a systematic search guided by simplicity and domain constraints such as laws of conservation. Pruning of hypotheses by means of evidence not usable during generation will be illustrated in the reaction examples later in the paper, in order to show how hypothesis generation recurs when simple hypotheses are inadequate. Our methods for testing hypotheses are of less interest here, but have constituted theoretical contributions to chemistry in their own right, and are cited where appropriate.

3.1. Overview of MECHEM

The hypothesis generator within MECHEM is charged with forming candidate hypotheses while incorporating much of the available evidence during generation. The generator forms hypotheses using *minimally* the following empirical data about a reaction:

- The molecular formulas of all starting materials.
- The molecular formulas of observed products.

A complete list of reaction products is not needed (nor is usually available in practice), but knowledge of each additional product can be expected to increase the credibility of the pathway hypotheses found, as well as the efficiency of the search.

MECHEM infers a set of pathways, each of which explains the formation of the specified reaction products from the starting materials. Since the space of possible pathways is unbounded, the program uses simplicity as a heuristic to guide its attention. The program's simplicity criteria are the number of pathway species and of steps; fewer of each is simpler.

We offer a small, contrived example to illustrate MECHEM's behavior. The data input to MECHEM are $O_2 \leadsto O_{16}$. In response to the **find_pathways** command, the program searches the pathway space, printing the subspaces as they are examined. Finally, MECHEM searches a subspace (4 species and 3 steps) that contains solutions, and outputs all of them (only one in this case).

```
<cl> (setq pathways (find_pathways '((2)) '(16) '(0)))
Searching with 2 species ...
          With 1 step
Searching with 3 species ...
          With 1 step
Searching with 4 species ...
          With 2 steps
```

```
        With 3 steps

(((((2) (2)) ((0) (4))) (((4) (4)) ((0) (8))) (((8) (8)) ((0) (16)))))

<cl> (mapc #'depict pathways)
202 => 04
204 => 08
208 => 016
```

The DEPICT program prints pathways legibly and conventionally. The number of pathway species and steps are 4 and 3, respectively (S=4,R=3). The pathway found indicates that the species O_{16} is formed by successive addition reactions that start with O_2. There are no pathways simpler than this in the spaces searched by MECHEM. The pathway is an *explanation* of the input data in terms of intelligible chemical processes.

3.2. The a-priori pathway space

A pathway can be represented as a set of reactions (not necessarily elementary) of the type $reactants \rightarrow products$, where the reactants and products are species. A key observation is that elementary chemical reactions are of small molecularity, meaning that reactions are limited to one, two, or three reactants. Moreover, a three-reactant step such as $A{+}B{+}C \rightarrow products$ rarely occurs as an *elementary* step. Most elementary steps are of the form $A + B \rightarrow products$, $2A \rightarrow products$, or $A \rightarrow products$.

In any case, a three-reactant step $A{+}B{+}C \rightarrow products$ can be modelled as $A + B \rightarrow AB$, $AB + C \rightarrow products$ at the cost of introducing an intermediate AB into the model. A three-product step $reactants \rightarrow X + Y + Z$ can be modelled analogously. Therefore, to narrow the a-priori search space, MECHEM only generates pathway steps having at most two reactants and two products, without losing the ability to model adequately three-reactant (or -product) steps.

Next we examine the prima-facie size of this space, focusing on a fixed hypothesis complexity S,R. Given a template of R steps, one assigns the S species to the 2R reactant and 2R product slots in all possible ways, allowing assignment of any species to multiple slots. All the complete slot assignments that fulfill the constraints are collected. Since steps of schematic form $A \rightarrow Y{+}Z$ and its reverse must be allowed, the S species are augmented with the null species, whose formula is all zeros; hence S+1 species are actually assigned instead of S.

The size of this pathway space is $(S{+}1)^{4R}$. A simple, typical case might involve S=8,R=3, meaning that 9^{12} pathways are possible, a gigantic number. To make this abstract problem tractable, an algorithm must be able to, first, use heuristics to prune partially-formed pathways early during generation, and second, reduce the number of redundant generation paths leading to identical pathways.

A straightforward way to search an S,R subspace would be to define a procedure $search(S,R)$ which has no memory of search in simpler spaces. Instead, it is desirable to exploit prior search in the hypothesis-formation algorithm. This has been accomplished in MECHEM, as is made clear in the description of the algorithm in section 3.6 below.

3.3. Constraints on pathways

Section 3.1 mentioned simplicity as a key constraint on the order of pathway generation. This section lists most of the constraints on candidate pathways that are applied to every problem; MECHEM only generates pathways consisting of schematic steps $reactants \rightarrow products$ that satisfy the list below. Some other constraints in current use are optional; an example is discussed in section 3.11.

Constraint on individual species.

- A legal formula contains only non-negative integer coefficients.

It is nonsense to speak of minus-one hydrogen, or half a carbon, for example.

Constraints on pathway steps.

- For each chemical element, the number of its reactant atoms in any step equals the number of its product atoms in the same step. That is, steps must be balanced; number and type of atoms are conserved.
- No species appears both among the reactants and among the products of a single step.
- Every reaction step has either one or two reactant species.
- Every reaction step has either one or two product species.

The last three constraints are design decisions to reduce the scope of the program in favor of tractability. However, little scope is lost by the last two constraints, as explained previously.

Constraints on pathways.

- The pathway eventually forms from the starting materials all the specified reaction products.
- All reaction steps are distinct.
- All pathway species must be formable from the starting materials. This rules out, for example, the pathway $A+B \rightarrow T+X$, $C \rightarrow B+Y$ in which A,B are the starting materials. Since C is not present at the start and cannot be formed by the pathway, the step containing it (and thence the pathway) is spurious.[1]

- For any pathway containing variables (i.e., unknown species), there must exist an assignment of legal formulas to the variables such that the steps balance.[2]

All of these are commonsensical.[3]

Constraint on sets of pathways.

- No two pathways are identical nor become identical under any of the following transformations: 1. replacing a step $R_1 + R_2 \rightarrow products$ by $R_2 + R_1 \rightarrow products$, 2. replacing a step $reactants \rightarrow P_1 + P_2$ by $reactants \rightarrow P_2 + P_1$, and 3. permuting two steps. This constraint is enforced by the canonic pathway notation developed in section 3.9 below.

The justification for these constraints is also commonsensical.

3.4. Search heuristics

The crucial heuristic to focus search of the pathway space is simplicity, defined in terms of number of steps and species. At any moment during execution, MECHEM is examining pathways containing S species and R steps. Given fixed S,R, the system searches exhaustively in the subspace demarcated by these values of the complexity parameters. When search in one subspace proves fruitless, MECHEM searches an incrementally more complex subspace, but exploits its prior search when possible.

MECHEM always prefers increments in the number of steps R to increments in S, that is, it adds steps one at a time. An exception occurs when it becomes evident that further increments in R are fruitless, i.e., all current pathways are at dead ends. In such cases, MECHEM increments S, resets R to its minimum value, and resumes the search. A sound criterion to detect dead ends is discussed in [29].

3.5. Conjecturing unseen species

When forming pathway hypotheses, the seen species consist of the starting materials and any other reaction products whose identities are known. Generally, the pathway hypotheses discovered by MECHEM contain unseen species, that is, species that are not *known* at the start. By what mechanism are unseen species introduced into pathways?

When beginning to search the class of S species, where S $\geq$ the number of seen species, MECHEM creates a variable for each unknown species, e.g., X,Y,Z in the case of three unknowns. The total number of unknowns created is $S - \#(seen)$. Any acceptable, complete pathway is required to make use of these unknowns, since simpler pathways with fewer unknowns would already have been considered earlier in the search.

When MECHEM adds a reaction step while generating a candidate pathway, the possible products for the new step are chosen from any of the species, seen or unseen. As more steps using the unknowns are added to the pathway, more algebraic constraint is placed on the unknowns, since every step gives rise to a vector-algebraic equation by replacing the '$\rightarrow$' by '='. At some point unique formulas can be inferred for the unknowns.

We illustrate with an example. Suppose that pathways are required on the following data for the class S=6,R=3:

$$C_8H_{18}O_3 + H_2O \ \rightsquigarrow\ C_4H_8O_2 + C_2H_6O$$

MECHEM considers this initial one-step fragment:

$$C_8H_{18}O_3 + H_2O \rightarrow X + Y$$

A subsequent extension to the fragment

$$Y \rightarrow H_2O + C_4H_8O_2$$

adds enough constraint to infer the values of X, Y. After one further extension, an acceptable pathway is reached:

$$\begin{array}{rcl} C_8H_{18}O_3 + H_2O & \rightarrow & \langle C_4H_{10}O\rangle + \langle C_4H_{10}O_3\rangle \\ \langle C_4H_{10}O_3\rangle & \rightarrow & H_2O + C_4H_8O_2 \\ H_2O + \langle C_4H_{10}O\rangle & \rightarrow & 2\,(C_2H_6O) \end{array}$$

The bracketed, unseen species are $C_4H_{10}O$ and $C_4H_{10}O_3$. This is how the integration of the *seen* species with the *unseen* (and initially unknown) species is accomplished. If an instantiation leads to an illegal molecular formula, then the last fragment extension is discarded, since it caused the instantiation. The general point is that a conservation law - mass balance - is used to infer the identity of conjectured species.

3.6. Hypothesis generation in MECHEM

The basic algorithm underlying MECHEM is represented by the lisp-like pseudo-code below. (Recall that a fragment is a partial pathway, i.e., a set of steps, not necessarily complete in the sense of forming all the required products).

```
find_pathways(materials, products)
  S = #(materials) + #(products)
  loop
```

```
    *fragments* = '(nil)   ;;; global variable
    min_steps = ceiling((S - #(materials)) / 2)
    extendnum = min_steps
    until (null *fragments*) do
        extend_fragments_by(extendnum)
        pathways = test_for_solutions(*fragments*)
        if pathways then return(pathways)
        extendnum = 1
    S = S + 1   ;;; postulate a new variable
    push(Xs,products)  ;;; add it to products

extend_fragments_by(remaining_steps)
  if remaining_steps = 0 then return
  else
  for fragment in *fragments* do
     for reactants in get_reactants_set(fragment) do
        for products in get_products_set(fragment) do
           new_fragment = append(fragment,make_step(reactants,products))
           infer_and_instantiate_values_of_unknowns(new_fragment)
           if legal?(new_fragment) then push(new_fragment,new_fragments)
  *fragments* = new_fragments
  extend_fragments_by(remaining_steps - 1)
```

The inputs to **find_pathways** are the formulas of starting **materials** and **products**. The minimum number of species in any acceptable pathway must include all of these, accounting for the first line in **find_pathways**. The parameter **extendnum** receives the value of **min_steps**, the minimum number of steps possessed by an acceptable pathway; the formula is justified by the fact that each step can contribute at most two new products. Thereafter in the loop, ***fragments*** having **extendnum** steps are created by the recursive **extend_fragments_by**, then are tested for the presence of acceptable pathway solutions. The initial fragment submitted for extension is the empty (zero-step) fragment **nil**.

If satisfying pathways are found, the search ends by returning them. Otherwise, **extendnum** is set to 1, so that all of the ***fragments*** are extended by one step, and the presence of acceptable pathways is once again checked. The inner iteration in **find_pathways** continues until the set of fragments becomes empty, which occurs eventually because of the **go-forward** heuristic discussed in section 3.10 below.[4] When the fragments become empty, S is incremented and the inner loop resumes as before.

The recursive function **extend_fragments_by** extends each fragment in ***fragments*** by the number of **remaining_steps**. Each normal invocation of the function extends the fragments by one step, decrements **remaining_steps**, and recurs. However, when **remaining_steps** reaches a value of zero, it means that the extension is completed so no further recursion is carried out.

For each fragment, a list of possible reactants and a list of possible products of a new step are created. All the possible new steps are generated by combining each pair of possible reactants with each pair of possible products. After checking that the step is not unbalanced, nor violates any other constraint on steps, the new step is added to the fragment, the values of any species unknowns are inferred - if possible - and instantiated in the **new_fragment**. If **new_fragment** is a legal fragment, it is added to the list of new_fragments built by the outer iteration. Finally, **extend_fragments_by** recurs, until the extensions of specified length are reached.

3.7. An execution trace

We illustrate the basic algorithm with traces from MECHEM's behavior on reaction example 6 from Table 4 in section 4.3 below. The reaction data are:

$$C_7H_9N,\ CH_2O \ \leadsto\ C_{17}H_{18}N_2,\ H_2O$$

First, MECHEM generates all possible initial reactants:

$$\begin{array}{rcl@{\qquad}rcl@{\qquad}rcl} CH_2O & \rightarrow & \ldots & C_7H_9N & \rightarrow & \ldots & CH_2O + C_7H_9N & \rightarrow & \ldots \\ 2(CH_2O) & \rightarrow & \ldots & 2(C_7H_9N) & \rightarrow & \ldots & & & \end{array}$$

followed by generating complete steps, some of which are immediately rejected. For example, the step $C_7H_9N \rightarrow H_2O + X$ is rejected because X receives the illegal value $C_7H_7NO_{-1}$. The first steps that are not immediately rejected are these eleven, where the conjectured species are left as variables for clarity.

$$\begin{array}{rcl@{\qquad}rcl} CH_2O & \rightarrow & X+Y & C_7H_9N & \rightarrow & X+Y \\ CH_2O & \rightarrow & H_2O+X & 2(C_7H_9N) & \rightarrow & X \\ & & & 2(C_7H_9N) & \rightarrow & X+Y \\ 2(CH_2O) & \rightarrow & X & CH_2O+C_7H_9N & \rightarrow & X \\ 2(CH_2O) & \rightarrow & H_2O+X & CH_2O+C_7H_9N & \rightarrow & H_2O+X \\ 2(CH_2O) & \rightarrow & X+Y & CH_2O+C_7H_9N & \rightarrow & X+Y \end{array}$$

MECHEM then extends these one-step fragments and discards all resulting two-step fragments that are recognizable dead ends ([29] has details on how this is tested). The remaining, promising fragments are the following four.[5]

$$\begin{array}{lrcl@{\qquad}lrcl} 1. & CH_2O+C_7H_9N & \rightarrow & H_2O+X & 3. & CH_2O+C_7H_9N & \rightarrow & H_2O+X \\ & CH_2O+X & \rightarrow & H_2O+Y & & 2(X) & \rightarrow & C_7H_9N+Y \\ 2. & CH_2O+C_7H_9N & \rightarrow & H_2O+X & 4. & CH_2O+C_7H_9N & \rightarrow & H_2O+X \\ & CH_2O+X & \rightarrow & Y & & 2(X) & \rightarrow & Y \end{array}$$

Finally, these five extensions of the two-step fragments are found to be legitimate solutions:

1. $$\begin{aligned} CH_2O + C_7H_9N &\rightarrow H_2O + X \\ CH_2O + X &\rightarrow H_2O + Y \\ X + Y &\rightarrow C_{17}H_{18}N_2 \end{aligned}$$

2. $$\begin{aligned} CH_2O + C_7H_9N &\rightarrow H_2O + X \\ CH_2O + X &\rightarrow C_9H_{11}NO \\ X + C_9H_{11}NO &\rightarrow H_2O + C_{17}H_{18}N_2 \end{aligned}$$

3. $$\begin{aligned} CH_2O + C_7H_9N &\rightarrow H_2O + X \\ 2(X) &\rightarrow C_7H_9N + Y \\ X + Y &\rightarrow C_{17}H_{18}N_2 \end{aligned}$$

4. $$\begin{aligned} CH_2O + C_7H_9N &\rightarrow H_2O + X \\ 2(X) &\rightarrow Y \\ X + Y &\rightarrow C_7H_9N + C_{17}H_{18}N_2 \end{aligned}$$

5. $$\begin{aligned} CH_2O + C_7H_9N &\rightarrow H_2O + X \\ 2(X) &\rightarrow Y \\ CH_2O + Y &\rightarrow H_2O + C_{17}H_{18}N_2 \end{aligned}$$

We have left the variables uninstantiated for legibility; MECHEM always instantiates variables as soon as their values can be inferred. To infer values of unknowns, Gaussian elimination (over formulas, i.e., vectors, not scalars) is used on a representation of pathways as sets of equations expressed as matrices.

It should now be clear how search in a space S,R exploits prior search in the space S,R-1: initial pathway fragments of a length R_0 are created, after which increments to R involve simply extending the current fragments. However, when S is incremented, prior records of search are discarded. Changing this seems quite complicated, and has not been attempted.

3.8. Another view of MECHEM's search

Section 3.6 described MECHEM's search in terms of pseudo-code. An alternative description focuses on the program's ability to introduce new terms representing hidden entities into its search, which are then instantiated by balancing the pathway steps.

Abstractly, the system carries out a *succession* of *heuristic*, *breadth-first* searches. Each *successive* search involves postulating a new hidden species in the form of a variable, thus changing the available vocabulary with which to form hypotheses. Each level in the *breadth-first* search adds a new step, corresponding to a more complex hypothesis. Finally, the search is *heuristic* because domain constraints plus experimental evidence are applied both to guide the addition of new steps and to prune unpromising nodes in the tree. Subprograms called by the hypothesis generator, which test pathways against specific evidence, can be seen as pruning nodes from the tree.

It is not logically necessary for the search to be successive, one could imagine a procedure that uses simplicity in the same way, but does not need to build a new tree from scratch for each vocabulary change. However, as already mentioned, devising such a procedure seems quite complex and has not been studied.

3.9. Removing redundant generation paths

A straightforward implementation of the basic algorithm can be very inefficient because a same fragment is reached by many different generation routes. It is important for the hypothesis generator to be as nonredundant as possible. This section shows how to remove redundance by enforcing a canonic representation of pathways [31]. This aspect of MECHEM follows in the spirit of prior work in the DENDRAL project [18] on establishing a canonic representation of molecules, although the details are quite different.

I - Species must be available

A first observation is that the steps of any legal pathway can be ordered from top to bottom such that each reactant of any step is a starting material or is formed by another step above it in the pathway. By generating only fragments that have this property, we remove some redundance while retaining completeness. The schematic pathway on the left below (with starting materials A, B) does not have the property, whereas the pathway on the right does (note that the two become identical after permuting the 2nd and 3rd steps).

$$\begin{array}{rcl} A+B & \rightarrow & X+Y \\ Y+Z & \rightarrow & T \\ B+X & \rightarrow & 2Z \end{array} \qquad \begin{array}{rcl} A+B & \rightarrow & X+Y \\ B+X & \rightarrow & 2Z \\ Y+Z & \rightarrow & T \end{array}$$

The line from the basic algorithm

```
extend_fragments_by(fragments,remaining_steps)
  ...
    for reactants in get_reactants_set(fragment) do
  ...
```

is made to have this property by restricting an extension to draw reactants only from the starting materials and products of the fragment being extended.

II - Steps should be added "soon"

The following five pathways are identical except for permutations between steps, yet they obey the previous restriction. Further conditions are needed to eliminate this redundance.

1.
$$\begin{aligned} A+B &\rightarrow X \\ 2B &\rightarrow Y+Z \\ Y+Z &\rightarrow T \\ X+Z &\rightarrow T+A \end{aligned}$$

2.
$$\begin{aligned} A+B &\rightarrow X \\ 2B &\rightarrow Y+Z \\ X+Z &\rightarrow T+A \\ Y+Z &\rightarrow T \end{aligned}$$

3.
$$\begin{aligned} 2B &\rightarrow Y+Z \\ A+B &\rightarrow X \\ Y+Z &\rightarrow T \\ X+Z &\rightarrow T+A \end{aligned}$$

4.
$$\begin{aligned} 2B &\rightarrow Y+Z \\ A+B &\rightarrow X \\ X+Z &\rightarrow T+A \\ Y+Z &\rightarrow T \end{aligned}$$

5.
$$\begin{aligned} 2B &\rightarrow Y+Z \\ Y+Z &\rightarrow T \\ A+B &\rightarrow X \\ X+Z &\rightarrow T+A \end{aligned}$$

We require that a new step be added to a fragment "soon" after it becomes enabled. This statement is formalized as follows.

Each species is assigned an <u>index</u> according to when the species is first formed. The starting materials and null species receive index 0; a product of the ith step in the fragment receives index i unless it possesses a smaller index. Each step's reactants give rise to a list of two indices, sorted in *descending order*. We require that a legal arrangement of fragment indices appear in *ascending* order, treating the indices as multi-place numbers.

The previous five pathway orders give rise to these indices:

1.	2.	3.	4.	5.
[0 0]	[0 0]	[0 0]	[0 0]	[0 0]
[0 0]	[0 0]	[0 0]	[0 0]	[1 1]
[2 2]	[2 1]	[1 1]	[2 1]	[0 0]
[2 1]	[2 2]	[2 1]	[1 1]	[3 1]

Only the second and third cases meet the new criterion. Can another criterion be devised to select only one of these two remaining cases? The next paragraphs develop a criterion that uniquely orders two steps such as $A+B \rightarrow X$ and $2B \rightarrow Y+Z$, both of which have indices [0 0] if A,B are starting materials.

III - Breaking ties among indices

We need to define a unique ordering on steps that distinguishes between two steps having identical indices. This is achieved by first placing an arbitrary, fixed order on all species, including the seen and unseen species. Then, any two steps can be ordered by comparing their four corresponding species (two reactants and two products). The arbitrary order on species can be just a list L; for example, the simple problem in section 3.5 could give rise to a list

$$[C_8H_{18}O_3,\ H_2O,\ C_4H_8O_2,\ C_2H_6O,\ C_0H_0O_0,\ X,\ Y]$$

in which the fifth entry is the null species. It is important to require that the initial unknowns (X, Y in this example) conserve their location in the order even after they are instantiated.

With the above three restrictions, we are close to determining a unique order for fragment generation. [29] proves that a pathway generator using the above three constraints is *partially* nonredundant in the following sense: no two *uninstantiated* fragments, derivable one from the other by permutation of steps, are both saved by the generator for further extension. Hence, such duplicate fragments cannot contribute to downstream redundant branching. Some redundance persists, but it shows up only when instantiation of pathway unknowns is considered.

Without a canonic representation such as above, MECHEM's search task would be completely impractical for reasons of combinatorics, which are quite severe in the present case.

3.10. Early pruning of pathway fragments

MECHEM uses several sound heuristics to prune fragments that cannot yield an acceptable pathway even after further extensions. These tests, some of which are entailed by the constraints in section 3.3, are carried out within the basic algorithm at this line:

```
extend_fragments_by(fragments,remaining_steps)
      ...
        if legal?(new_fragment) then push(new_fragment,new_fragments)
      ...
```

We will describe only one of these heuristics, the **go-forward** heuristic, which by itself nearly halves the search time on many examples tested.

During search of an S,R subspace, MECHEM usually[6] ignores any fragment extension forming only products that are either starting materials or are already formed within the fragment. The justification is that any acceptable pathway incorporating such a step could not be a simplest pathway, since the pathway omitting the step would also be acceptable and would be simpler.

Elsewhere [36] we discuss pruning of pathway hypotheses in a systematic way - for any pathway problem, not just chemistry - based on the level of aggregation at which domain knowledge is applied.

3.11. Plausible molecular formulas

Some molecular formulas are implausible because they imply too many nonstandard chemical valences for the molecule's atoms, e.g., CH_6 implies that carbon has a valence of six. We make use of this plausibility judgment as a heuristic, which is described fully in [29] and has been tested successfully on a database of about 800 compounds developed by Peter Karp of SRI. In MECHEM any pathway fragment conjecturing an implausible species is rejected.

The heuristic depends on calculation of a quantity called degree of unsaturation. This quantity was also used in DENDRAL (see page 12 of [18]). Recall that before DENDRAL could propose plausible structures for a molecule, it had to infer a likely molecular formula. Rejection of unlikely formulas depended on the same calculation, although the exact implausibility criterion differs between DENDRAL and MECHEM.

3.12. Energetic plausibility of reaction steps

The plausibility of a pathway step can be judged by examining a net energy change between reactants and products. The energy concerned is that of free energy, which should be net negative for the more likely reaction steps. For example, the reaction

$$\begin{array}{ccccccc} water & + & formic\ acid & \rightarrow & oxygen & + & methanol \\ -2.1858E8 & + & -3.3623E8 & & 2.6568E7 & + & -1.3827E8 \end{array}$$

implies a net free-energy change of $+4.4E8$ joules/kilomole,[7] obtained by subtracting the reactant energies from the product energies. This net change is an order of magnitude greater than a conservative threshold on plausibility recommended in [9], hence the reaction step is judged implausible. Unfortunately, reliable free-energy values are not always available.

3.13. Other algorithmic aspects

The basic algorithm underlying hypothesis generation in MECHEM has been proved to terminate [29]. The proof constructs a pathway $\mathcal{P}$ that is within the scope of the algorithm and which satisfies the constraints of section 3.3.

We have no proof that MECHEM correctly implements the basic algorithm and the heuristics described in this paper. However, the core of MECHEM was implemented also as a Prolog III program [11], which permitted verifying at least that identical answers were obtained. Five reaction examples from Table 4 below were successfully tested in this way.

Finally, we emphasize that MECHEM's current design makes it insensitive to the order in which evidence is presented to it. The system finds all the simplest hypotheses that can explain a given body of evidence.

4. Reaction Examples

This section will demonstrate that MECHEM is able competently to discover plausible pathways underlying chemical reactions. The system is applied first to a biochemical pathway whose discovery has been studied both historically and from an AI viewpoint. Then MECHEM is applied to a common industrial reaction taken from the scientific literature.

4.1. Urea synthesis

Kulkarni and Simon [14] describe a computer program KEKADA that implemented a historical reconstruction by Holmes [8] of Hans Krebs's celebrated discovery of the cyclical biochemical pathway underlying the synthesis of urea *in vivo* [13]. We have applied MECHEM to the problem of inferring pathways for the urea reaction, using a portion of the evidence discovered by Krebs and KEKADA. Krebs at one point had concluded that three starting materials were involved in forming three reaction products thus:

$$ornithine,\ ammonia,\ carbon\,dioxide \ \rightsquigarrow \ urea,\ water,\ arginine$$

Also, Krebs had inferred from experiment that ornithine catalyzed the reaction, i.e., ornithine was completely re-generated after initially undergoing reaction. Lastly, Krebs knew from the chemical literature that arginine produces ornithine and urea in other contexts (see page 149 of [15] or page 221 of [8]), and considered this fact relevant to the urea pathway under study. We will require that arginine be a precursor of urea, although not necessarily in a single step.

We will use the above data together with a heuristic that merges two pathway steps into one three-reactant step if the following condition is met: a conjectured species appearing in the original two steps is eliminated from the new, merged step. This heuristic is formalized in a small program MERGE-STEPS. The sole purpose of MERGE-STEPS is to allow three-reactant steps like Krebs did. General reasoning about pathway catalysis in MECHEM is formalized in another program YIELD; this program and its supporting theoretical results are reported in [32, 34].

Thus MECHEM, YIELD, and MERGE-STEPS combine to infer the following ten simplest pathways consistent with the above data, of which the pathway proposed by Krebs is shown first:

$$
\begin{array}{lrcl}
1. & ornithine + ammonia + carbon\ dioxide & \rightarrow & water + C_6H_{13}N_3O_3 \\
 & ammonia + C_6H_{13}N_3O_3 & \rightarrow & arginine + water \\
 & arginine + water & \rightarrow & ornithine + urea \\
2. & ammonia + carbon\ dioxide & \rightarrow & water + CHNO \\
 & ornithine + ammonia + CHNO & \rightarrow & arginine + water \\
 & arginine + water & \rightarrow & ornithine + urea \\
3. & ammonia + carbon\ dioxide & \rightarrow & water + CHNO \\
 & ornithine + 2\,(CHNO) & \rightarrow & carbon\ dioxide + arginine \\
 & arginine + water & \rightarrow & ornithine + urea \\
4. & ornithine + carbon\ dioxide & \rightarrow & water + C_6H_{10}N_2O_3 \\
 & 2\,(ammonia) + C_6H_{10}N_2O_3 & \rightarrow & arginine + water \\
 & arginine + water & \rightarrow & ornithine + urea \\
5. & ornithine + ammonia & \rightarrow & water + C_5H_{13}N_3O \\
 & ammonia + carbon\ dioxide + C_5H_{13}N_3O & \rightarrow & arginine + water \\
 & arginine + water & \rightarrow & ornithine + urea \\
6. & ornithine + ammonia & \rightarrow & water + C_5H_{13}N_3O \\
 & carbon\ dioxide + 2\,(C_5H_{13}N_3O) & \rightarrow & ornithine + arginine \\
 & arginine + water & \rightarrow & ornithine + urea \\
7. & ornithine + 2\,(ammonia) & \rightarrow & water + C_5H_{16}N_4O \\
 & carbon\ dioxide + C_5H_{16}N_4O & \rightarrow & arginine + water \\
 & arginine + water & \rightarrow & ornithine + urea \\
8. & ornithine + ammonia & \rightarrow & water + C_5H_{13}N_3O \\
 & carbon\ dioxide + 2\,(C_5H_{13}N_3O) & \rightarrow & ornithine + arginine \\
 & ammonia + arginine & \rightarrow & C_5H_{13}N_3O + urea \\
9. & ornithine + ammonia & \leftrightarrow & water + C_5H_{13}N_3O \\
 & ammonia + carbon\ dioxide + C_5H_{13}N_3O & \rightarrow & arginine + water \\
 & ammonia + arginine & \rightarrow & C_5H_{13}N_3O + urea \\
10. & ornithine + carbon\ dioxide & \leftrightarrow & water + C_6H_{10}N_2O_3 \\
 & 2\,(ammonia) + C_6H_{10}N_2O_3 & \rightarrow & arginine + water \\
 & carbon\ dioxide + arginine & \rightarrow & C_6H_{10}N_2O_3 + urea
\end{array}
$$

Each pathway involves seven species, meaning that MECHEM needed to conjecture two unseen species to find a satisfying pathway; in each case one of the conjectured species was eliminated by MERGE-STEPS. The conjectured species $C_6H_{13}N_3O_3$ is the formula of citrulline, the intermediate proposed by Krebs (ornithine is $C_5H_{12}N_2O_2$, water is H_2O, arginine is $C_6H_{14}N_4O_2$, urea is CH_4N_2O, ammonia is NH_3, and carbon dioxide is CO_2).

Table 1 summarizes the evolution of hypotheses on the urea reaction. Starting only with knowledge of starting materials and of urea, MECHEM finds six 6S2R pathways. Data on the formation of arginine and then of water changes the current pathways found by MECHEM as shown. Then the current 16 pathways, together with the fact that ornithine is a catalyst (the "ornithine effect"), are input to YIELD, which finds that none of the 16 explains the ornithine effect. MECHEM is asked

evidence	#pathways	class
starting materials & urea	0	4S2R
	0	5S2R
	6	6S2R
arginine is formed	0	7S2R
	71	8S3R
water is formed	16	8S3R
ornithine effect	98	8S4R
arginine precursor of urea	45	8S4R
merge step (allow three-reactant steps)	10	7S3R

Table 1: Pruning Hypotheses on Urea Pathway

name	molecule	formula species	concentration at 3 hrs (mole/liter)	concentration at 6 hrs (mole/liter)
acetic acid	CH_3COOH	$C_2H_4O_2$	.093	.13
acetone	$(CH_3)_2CO$	C_3H_6O	.15	.20
ethyl acetate	$CH_3COOC_2H_5$	$C_4H_8O_2$	.004	.04
isopropanol	C_3H_7OH	C_3H_8O	.10	.054
acetaldehyde	CH_3CHO	C_2H_4O	.015	.015
methanol	CH_3OH	CH_4O	.038	.10
carbon dioxide	CO_2	CO_2	.108	.16
formic acid	$HCOOH$	CH_2O_2	-	-

Table 2: Reaction Products of Propane Oxidation

for incrementally more complex pathways, and the new batch in 8S4R is again tested by YIELD, leaving 98 pathways able to explain the ornithine effect. Of these, only 45 pathways show arginine to be a precursor of urea. Then, on each remaining pathway the program MERGE-STEPS is called to merge two steps into one three-reactant step (as was allowed by Krebs) if the merge eliminates a conjectured species; this results in the 10 simplest merged pathways presented above.

This historical reaction example supports the notion that MECHEM is an effectual discoverer in its domain of competence. The system uses data discovered by or available to Krebs to infer systematically a reasonably small set of pathways that includes the one proposed by Krebs.

4.2. Liquid-phase oxidation of propane

Bulygin et al. [3] published experimental data on the oxidation of propane in the liquid phase, and proposed a pathway to explain the data. These data consisted mostly of the identities of some products of the reaction and a plot of their concentrations over time, which is reproduced as Figure 3. The starting materials of the reaction are propane (C_3H_8) and oxygen (O_2).

The authors list the eight reaction products shown in Table 2. Of these, the concentrations of only

Figure 3: Concentrations Over Time for Products of Propane Oxidation (from Bulygin et al. 72)

seven products are plotted in their paper; formic acid is omitted. In addition, the authors conclude from separate experiments that acetaldehyde is not a precursor of carbon dioxide nor of formic acid.

We proceed to examine the behavior of MECHEM on the available evidence; the results are summarized in Table 3. With the two starting materials and seven plotted products, MECHEM finds no pathways involving only the input species, i.e., in class 9SxR. After incrementing the number of species to 10, MECHEM finds no pathways in 10S4R. Finally, the class 10S5R yields 256 pathways, of which there remain 227 after pruning those containing implausible conjectured species, using the unsaturation heuristic of section 3.11. Of these 227 pathways, the constraint on maximum free-energy change[8] rules out 63 of them because they contain a step that is energetically less plausible.

Up to now, all of the work has been done by MECHEM, and the only experimental evidence used was the identity of products. Next, we use a program called EXTENT which is based on the theory in [27]. EXTENT tests whether a given pathway can adequately account for changes in concentrations measured over time. We have somewhat crudely inferred the concentrations in Table 2 from the plot in Bulygin et al. at reaction times 3 and 6 hours. All the initial concentrations of these species are zero, since they are reaction products. The authors did not plot the concentration of propane, despite its usefulness to us for elucidating the pathway.

We then require that each pathway explain adequately the concentrations transitions between 0, 3,

evidence	#pathways	class
2 starting materials & 7 products	0	9S4R
	0	10S4R
	256	10S5R
degree of unsaturation	227	10S5R
free-energy changes	164	10S5R
concentrations data	24	10S6R
acetaldehyde not precursor of CO_2	0	10S6R
	0	10S7R
	4	11S5R
explain formation of formic acid	1	11S5R

Table 3: Pruning Hypotheses on Propane Reaction

and 6 hours (details are in [29]); the 164 pathways are reduced to 24. If we then adduce the authors' conclusion that acetaldehyde is not a precursor of carbon dioxide, then even these 24 pathways are ruled out, meaning that the entire hypothesis class 10S6R is rejected. Further search involving more steps also fails to uncover pathways consistent with the constraints and evidence.

After incrementing the number of pathway species to 11, the class 11S5R turns up the following 4 pathways that account for all the available theoretical and experimental evidence (conjectured species are depicted as formulas):

1. $$\begin{aligned} propane + oxygen &\rightarrow acetone + H_2O \\ oxygen + acetone &\rightarrow acetaldehyde + CH_2O_2 \\ acetone + CH_2O_2 &\rightarrow isopropanol + carbon\, dioxide \\ oxygen + isopropanol &\rightarrow acetic\, acid + methanol \\ isopropanol + CH_2O_2 &\rightarrow ethyl\, acetate + H_2O \end{aligned}$$
2. $$\begin{aligned} propane + oxygen &\rightarrow acetone + H_2O \\ 2\,(acetone) &\rightarrow isopropanol + C_3H_4O \\ oxygen + isopropanol &\rightarrow acetic\, acid + methanol \\ oxygen + C_3H_4O &\rightarrow acetaldehyde + carbon\, dioxide \\ 2\,(acetaldehyde) &\rightarrow ethyl\, acetate \end{aligned}$$
3. $$\begin{aligned} propane + oxygen &\rightarrow acetone + H_2O \\ 2\,(acetone) &\rightarrow acetaldehyde + C_4H_8O \\ oxygen + C_4H_8O &\rightarrow isopropanol + carbon\, dioxide \\ oxygen + isopropanol &\rightarrow acetic\, acid + methanol \\ isopropanol + C_4H_8O &\rightarrow propane + ethyl\, acetate \end{aligned}$$
4. $$\begin{aligned} propane + oxygen &\rightarrow acetone + H_2O \\ 2\,(acetone) &\rightarrow acetaldehyde + C_4H_8O \\ oxygen + C_4H_8O &\rightarrow isopropanol + carbon\, dioxide \\ oxygen + isopropanol &\rightarrow acetic\, acid + methanol \\ acetic\, acid + C_4H_8O &\rightarrow ethyl\, acetate + acetaldehyde \end{aligned}$$

Consultation of the Merck Index (a chemicals encyclopedia) reveals that the conjectured species (besides water) within the three pathways are formic acid (CH_2O_2), acrolein or an isomer (C_3H_4O), and one of six isomers of C_4H_8O. It is noteworthy that one of the pathways conjectures the unmeasured species formic acid mentioned by the authors.

Bulygin et al. themselves propose a pathway for the reaction, by analogy with another pathway for the liquid-phase oxidation of a distinct hydrocarbon (propane is a member of the hydrocarbons). Their pathway accounts qualitatively for the role of isopropanol as an intermediate, which is clear from the plot; MECHEM's four pathways also account for this role. A drawback of their pathway is its failure to explain the formation of formic acid. (The exact nature of the seventh step below, taken from their diagram, is not clear to this author).

$$\begin{aligned}
propane + oxygen &\rightarrow C_3H_7O_2 + H \\
C_3H_7O_2 + propane &\rightarrow C_3H_8O_2 + C_3H_7 \\
C_3H_8O_2 &\rightarrow acetone + water \\
C_3H_8O_2 &\rightarrow isopropanol + O \\
isopropanol &\rightarrow acetone + H_2 \\
C_3H_7O_2 &\rightarrow acetaldehyde + CH_3O \\
CH_3O + ? &\rightarrow methanol + carbon\ dioxide \\
acetaldehyde + O &\rightarrow acetic\ acid \\
2(acetaldehyde) &\rightarrow ethyl\ acetate
\end{aligned}$$

This pathway differs from the two found by MECHEM mainly in that it involves two precursors $C_3H_7O_2$ and $C_3H_8O_2$ which then lead to the observed products and other intermediates. MECHEM does not find the need to conjecture these precursors, instead proposing that propane and oxygen form acetone and water on the first step, which need not be an elementary step. It is unclear to this author whether the pathway proposed by Bulygin et al. is accepted unreservedly by chemists.

Our conclusion on this reaction example is as follows. The system has used the experimental evidence contained in a chemistry paper to systematically search for the simplest hypotheses that are consistent with its internal chemical theory and which are able to explain the evidence. MECHEM discovered a small number of hypotheses that are simpler than the one proposed by the paper's authors; no simpler pathways can account for the objective evidence. We do not claim that MECHEM's pathways are necessarily plausible to chemists expert in hydrocarbon oxidation, since this area has been well studied, although not completely understood.

4.3. Other examples

MECHEM has been run on a number of other examples for which not enough evidence beyond the reaction products was available; hence the other programs in MECHEM were not exercised on these examples. Statistics on these and on the previous urea and propane reactions appear in Table 4; line 5 is the propane reaction, and line 7 is the urea reaction.

In the table, **#sm** and **#prod** are the number of starting materials and seen products. **#conj** is the number of conjectured (unseen) products. **#pthwys** is the number of simplest pathways found. **complexity** refers to the number of pathway species and steps. **time** is the running time (real time, not cpu time) on an otherwise unloaded DecStation 3100. (Line 11 of the table reports

	#sm, #prod	#conj	#pthwys	complexity	time
1.	2,6	0	1	8S,3R	1 sec
2.	2,2	1	3	5S,2R	1 sec
3.	2,4	1	8	7S,3R	7 sec
4.	2,4	1	28	7S,4R	11 sec
5.	2,7	1	256	10S,5R	22 min
6.	2,2	2	5	6S,3R	18 sec
7.	3,3	2	16	8S,3R	3 min
8.	2,4	2	130	8S,4R	17 min
9.	3,1	3	67	7S,3R	20 min
10.	2,3	3	45	8S,3R	32 min
11.	2,4	3	731	9S,4R	2.8 hrs

Table 4: Statistics on simplest pathways output by MECHEM on reaction examples

remarkably many pathways for such a simple instance. However, we have not used the degree-of-unsaturation nor free-energy heuristics while compiling this table. Just the unsaturation heuristic would reduce the number of pathways from 731 to 57 in this case.)

Current experience with MECHEM suggests several factors that influence execution time. One factor, strongly reflected in the table, is the number of species that need to be conjectured before finding a pathway; in fact the table is sorted with respect to the number of conjectured species. The number of species that will need to be conjectured is not usually evident from the initial data on starting materials and products.

Other factors are suggested by the surprisingly large execution time (22 minutes) on the propane reaction, which involves only one conjectured species. We believe that the relatively large number of balanced steps (10) that can be formed solely with the known species accounts for this time, since the combinatorics becomes worse under this condition. Unlike the previous factor of number of conjectured species, this factor can be calculated before any search is carried out.

5. Discussion

5.1. Related work on chemical reactions

An interesting question is the relation of MECHEM to SYNCHEM [6] and SECS [38], two well-known AI programs that dealt with chemical reactions. There are several key differences between MECHEM and the other two programs, taken together.

First, MECHEM addresses the task of pathway elucidation, whereas SECS/SYNCHEM are concerned with devising chemical syntheses, i.e., finding a set of starting materials and reaction conditions that are likely to form a given target product; in these programs this involves finding as well a set of reaction steps. This difference is highlighted by the data needed by each: SECS/SYNCHEM need the target product, while MECHEM needs the starting materials and at least one product. The greater demand for data on the part of MECHEM does not mean that the program is weaker, since the systems address different problems. The extra data that MECHEM requires allows it to use methods different from SECS/SYNCHEM.

A second difference is that SECS/SYNCHEM make use of extensive databases of starting materials and schematic reactions. The schematic reactions correspond to the operators in the search tree, while the database of starting materials serve as termination conditions in the search: when the leaves of a tree consist of recognized and convenient chemicals, the tree is a good candidate for a successful synthesis. MECHEM does not currently have databases of any kind, with the exception of a list of species free-energies that was obtained from a commercial program called ASPEN.

Another body of work related to MECHEM consists of the programs reported in Langley et al. [17], although the goal of their work was to find simple mechanisms able to reconstruct specific historical discoveries with modest amounts of computation. Three of the four programs they describe (GLAUBER, STAHL, and DALTON) deal with chemical reactions. In terms of the programs' methods, one difference between MECHEM and the former programs is that MECHEM is able to conjecture unseen entities in the course of discovering structural/dynamic models of a reaction. STAHL and DALTON, on the other hand, formulate structural hypotheses about given entities, while GLAUBER is concerned with finding concepts (e.g., *acid, base*) that permit devising simple empirical laws.

A specific comparison with the STAHLp system [21] and its successor REVOLVER [22] is of some interest. STAHLp finds componential models of substances that are implied by a set of given reactions, which is a step in MECHEM's search: given a pathway fragment with species variables, infer its component formula. STAHLp's task also involves finding plausible revisions to the input reactions when an illegal model is inferred. MECHEM, however, simply removes that node from its search. Finally, STAHLp obtains its componential models by a search involving substitution of one variable for another, while MECHEM carries out a Gaussian elimination algorithm to infer the component formulas of species variables.

5.2. Comparison to DENDRAL

The design of MECHEM's hypothesis-generator follows in the spirit of the CONGEN generator of molecular structures that was at the heart of the DENDRAL project [18]. As in the case of DENDRAL, MECHEM's efficiency is crucially dependent on devising a canonic representation of reaction pathways [31], and on using available evidence to prune partial hypotheses as early as possible within the generator.

MECHEM and DENDRAL clearly address different scientific tasks: the determination of reaction pathway and molecular structure, respectively. However, at least one other key difference is salient: simplicity has a strong role in MECHEM, because the practical need to conjecture unseen entities must be controlled by preferring fewer such entities. This accounts for the nested iteration carried out by MECHEM, which successively increments the number of species and steps to be allowed in a hypothesis.

5.3. Impact on chemistry

Three chemical aspects of this work are of interest for AI. First, modern chemists usually refer to changes in molecular structure when reasoning about reaction pathways, although chemical engineers often do not. The version of MECHEM analyzed here reasons about molecular structure only indirectly, and does not even represent molecular structure within the program, although the current version of the system does. This fact is surprising to chemists, but reflects use of the *planning heuristic*: carrying out an efficient, preliminary search in a more abstract space.

Molecular structures are usually represented as graphs with nodes for atoms and edges for bonds. Search at this level implies costly graph-algorithmic manipulations that correspond to bonds fragmenting and forming in a combinatorial way. We envision an automation of pathway elucidation that carries out an initial search in the abstract space of formulas, and descends to the structural space opportunistically to apply further constraints on proposed pathways and steps, and to infer structures for the conjectured species.

A second aspect concerns competing computational approaches to our same task, which are discussed in detail in [29] and will only be sketched here. One tradition within computational chemistry attempts *to predict* the course of a reaction from initial conditions by use of a large knowledge-based system (e.g., [10]) or *ab initio* by numerical simulation of equations. Clearly, if one could *predict* reliably, universally, and tractably, then the problem of experimental *elucidation* would be dissolved. However, perfect predictive ability has not been achieved, and reactions continue to be studied experimentally.

Another tradition representative mainly of chemical engineers addresses the tasks of pathway testing [1] and experimentation [2] based on the laws of chemical kinetics [16]. This statistical approach involves formulating each pathway as a set of ordinary, nonlinear, differential equations, in which the state variables represent species concentrations; recent examples are [7, 19]. This approach has not been concerned with exploiting the rich sources of discriminatory evidence available from chemical theory and experiment, nor with generating pathway hypotheses under realistic assumptions. Instead, its focus has been on constructing a monolithic mathematical apparatus that integrates testing and experiment design, while relying on the chemist to deliver the hypotheses.

The third and final aspect is that we have argued elsewhere [26] that the experimental task of pathway elucidation is a more taxing problem of inductive inference than is commonly imagined by chemists and biochemists. There we adduce experiments with MECHEM to show that when the experimenter must conjecture unseen products (as is usually the case), a surprisingly large number of equally simple and potentially plausible pathways can result. This argument, if accepted, has important methodological consequences for chemical practice.

5.4. Generality

The methodological issue of generality in AI as a whole prompts conflicting viewpoints. For the field of automated scientific discovery, we have laid out elsewhere [23] our view of one good methodology which embraces - unsurprisingly - the work reported here. The rest of this section is limited to a few specific points.

The methods described in this paper are of demonstrated efficacy for the problem of determining reaction pathways. To what extent do these methods depend on idiosyncrasies of the scientific task chosen? For example, a key property of our task is that the maximum number of reactants per step is small, at most three. Is this property not idiosyncratic?

We respond that a considerable number of tasks in science have similar limitations on the space of hypotheses. For example, the problem of determining the hereditary nature of human traits is much simplified by the fact that parents as well as chromosomes come in pairs. If instead the nature of offspring were decided by group sex among an indeterminate number of partners, then the task of genetics would be much harder.

The general approach in this paper of combinatorial search guided by simplicity, in which branches of the search are pruned by means of domain constraints, available data, and further experimentation, seems of some general applicability in science. A group headed by Jan Zytkow has been writing programs to carry out scientific tasks in the domains of particle physics [5] and of classical genetics (personal communication). In these programs, simplicity has a similar role in guiding discovery.

Can MECHEM's method of introducing hidden entities be applied fruitfully to other scientific tasks, or is it peculiar to pathway discovery? A reliable affirmative answer would need to await a MEBIOL, MEGEOL, MEPHYS, or perhaps even a MEOMNI program. In the meantime, we venture the following arguments. Pathway (or mechanism) elucidation is a generic task in science, as keyword searches in library catalogues show. Often in science, one is not able to observe all entities that are needed to devise plausible explanatory models. The need to conjecture unseen entities, and to control their conjecture by simplicity, seem to follow, and these entities can often be determined by domain-specific conservation laws. In fact, Wolfgang Pauli postulated in 1930 the neutrino particle in physics [25] using reasoning quite similar to how MECHEM conjectures unseen species.

5.5. Limitations

Currently, MECHEM is unable, with practicable amounts of computation, to handle cases where the number of conjectured species is greater than four. The execution time of the system increases greatly with this number, and the current practical boundary on a 14 MIPS workstation is four. However, there are many practical cases within this scope.

5.6. Future work

The description in this paper has focused on hypothesis generation within MECHEM, and not on the multiplicity of domain knowledge pertinent to the program's task. One important source of constraint pertains to a representation of species as molecular structures, and to steps as changes in structure. As of this writing, MECHEM does represent species as structures, and incorporates some simple algorithms to detect implausible *elementary* steps. This ability is important for our goal of developing MECHEM to the point where chemists and engineers who need to understand a reaction pathway in detail will use the system. In the near term, we expect that further development will involve incorporating more such domain knowledge, as well as improving search efficiency to extend the scope of application beyond cases of four conjectured species.

In parallel, we are studying other tasks of pathway elucidation in science, e.g., determining endocytic pathways in cell biology [20, 36]. Initial experience suggests that the concepts and procedures developed for MECHEM may find application there. Our interest lies both in automating this new specific problem, and in inducing a general account of pathway discovery from our experiences.

6. Conclusion

This paper describes a system that automates the discovery of the pathway underlying a chemical reaction. This task is traditionally carried out by chemists and engineers and leads regularly to publication of pathway hypotheses in scientific journals. Competence in the task requires the conjecture of unseen entities, since not all the reaction products are typically observed, and also involves reasoning about the selection and interpretation of experiments. The focus of this paper is on the hypothesis generator in MECHEM, and on the pruning of candidate hypotheses by means of separate programs that reason about experimental data.

The principal method of MECHEM involves carrying out a systematic, combinatorial search constrained by simplicity, background theory, and the available experimental data. The crucial ability to conjecture hidden entities is accomplished by postulating a new variable which becomes instantiated by means of conservation laws as the variable is used to build a hypothesis. A complementary view of MECHEM is that it performs a sequence of heuristic breadth-first searches, each new search in the sequence being preceded by a change in the vocabulary available to formulate hypotheses.

The efficacy of MECHEM is demonstrated by its application to two real reactions from the scientific literature; the pathways discovered are credible and of possible scientific value. The generality of MECHEM's methods is argued by reference to analogous scientific tasks, and by similarities between MECHEM and ongoing work elsewhere.

Acknowledgments. This paper is based on the author's Ph.D. thesis in Computer Science at Carnegie Mellon. I am indebted to the members of the thesis committee for their encouragement and suggestions: Herbert Simon (chairman), Bruce Buchanan, Tom Mitchell, and Gary Powers. Pat Langley and Derek Sleeman commented valuably on drafts of this paper. This research was supported by an Air Force Laboratory Graduate Fellowship, administered by the Southeastern Center for Electrical Engineering Education. Support for writing and publication of this paper derives from NSF grant DIR-8920118 to the Center for Light Microscope Imaging and Biotechnology.

Footnotes

1. The current hypothesis-generation algorithm enforces this constraint automatically. However, in previous approaches (e.g., [11]) an explicit check was required.

2. Variables arise when unseen species are conjectured, but there is insufficient constraint to determine their values uniquely; this is discussed below.

3. MECHEM currently uses an optional constraint which depends on distinguishing one target among the products, but this paper omits this rather unimportant distinction for simplicity of exposition. However, the pathways reported on specific reactions were required to fulfill this constraint.

4. This heuristic is sound when all simpler pathways have been rejected without using any reasoning that is "non-monotonic," i.e., that is not invalidated by adding further steps that do not introduce any new species. Most of MECHEM's constraints are "non-monotonic" in this sense.

5. Actually, the line **minsteps = ceiling((S - #(materials)) / 2)**, equals 2 in the current example. The two-step fragments will be built in a depth-first manner by recursion, rather than breadth-first as presented here.

6. Refer to footnote 4 above.

7. Calculated for a temperature of 150 degrees Centigrade and pressure of 50 atmospheres.

8. We have used a threshold of 0 on this example, instead of the more conservative threshold mentioned in section 3.12. We have also used free-energy values that were obtained for 17 species that include some conjectured species.

References

[1] G.E.P. Box and W.J. Hill. Discrimination among mechanistic models. *Technometrics*, 9(1):57–71, 1967. (Reprinted in: The Collected Works of George E.P. Box).

[2] G.E.P. Box and W.G. Hunter. Sequential design of experiments for nonlinear models. In *Proc. of IBM Scientific Computing Symposium on Statistics*, pages 113–137, 1963. (Reprinted in: The Collected Works of George E.P. Box).

[3] M.G. Bulygin, E.A. Blyumberg, and L.A. Al'tshuler. The liquid-phase oxidation of propane. *International Chemical Engineering*, 12(1):50–52, 1972.

[4] J.F. Bunnett. From kinetic data to reaction mechanism. In Claude Bernasconi, editor, *Investigation of Rates and Mechanisms of Reactions*, volume VI of *Techniques of Chemistry*. Wiley Interscience, 4th edition, 1986.

[5] P. Fischer and Jan M. Zytkow. Discovering quarks and hidden structure. In Z.W. Ras, M. Zemankova, and M.L. Emrich, editors, *Proceedings of the 5th International Symposium on Methodologies for Intelligent Systems*, 1990.

[6] H.K. Gelernter, A.F. Sanders, D.L. Larsen, K.K. Agarwal, R.H. Boivie, G.A. Spritzer, and J.E. Searleman. Empirical explorations of SYNCHEM. *Science*, 197:1041–1049, 1977.

[7] Friedrich G. Helfferich. Systematic approach to elucidation of multistep reaction networks. *Journal of Physical Chemistry*, 93:6676–6681, 1989.

[8] Frederic L. Holmes. Hans Krebs and the discovery of the ornithine cycle. In *Proc. of 63rd Annual Meeting of the Federation of American Societies for Experimental Biology*, volume 39, 1980. Symposium on Aspects of the History of Biochemistry.

[9] Olaf A. Hougen, Kenneth M. Watson, and Roland A. Ragatz. *Chemical Process Principles: Part II Thermodynamics*. John Wiley & Sons, 2nd edition, 1966. Pages 1015-16.

[10] William L. Jorgensen, Ellen R. Laird, Alan J. Gushurst, Jan M. Fleischer, Scott A. Gothe, Harold E. Helson, Genevieve D. Paderes, and Shenna Sinclair. CAMEO: A program for the logical prediction of the products of organic reactions. *Pure & Appl. Chem.*, 62(10):1921–1932, 1990.

[11] Jean Jourdan and Raul E. Valdes-Perez. Constraint logic programming applied to hypothetical reasoning in chemistry. In *Logic Programming: Proc. of the 1990 North American Conference*, pages 154–172, Cambridge, Mass., 1990. MIT Press.

[12] Edward L. King. *How chemical reactions occur, an introduction to chemical kinetics and reaction mechanisms*. W.A. Benjamin, 1963.

[13] H.A. Krebs and K. Henseleit. Untersuchungen uber die harnstoffbildung im tierkorper. II. *Klin. Wochenschr.*, 11:1137, 1932.

[14] D. Kulkarni and H.A. Simon. The processes of scientific discovery: The strategy of experimentation. *Cognitive Science*, 12:139–175, 1988.

[15] D.S. Kulkarni. *The Processes of Scientific Research: The Strategy of Experimentation.* PhD thesis, Carnegie Mellon University, 1988. (Computer Science Department).

[16] Keith J. Laidler. *Chemical Kinetics.* Harper & Row, 1987.

[17] P. Langley, H.A. Simon, G.L. Bradshaw, and J.M. Zytkow. *Scientific Discovery: Computational Explorations of the Creative Processes.* MIT Press, Cambridge, Mass., 1987.

[18] R.K. Lindsay, B.G. Buchanan, E.A. Feigenbaum, and J. Lederberg. *Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project.* McGraw Hill, New York, 1980.

[19] P.S. McCroskey. *A Knowledge-Based Approach for the Development of Complex Kinetic Mechanisms.* PhD thesis, Carnegie Mellon University, 1988. (Chemical Engineering Department).

[20] Robert F. Murphy. Processing of endocytosed material. In *Advances in Cell Biology*, volume 2, pages 159–180. JAI Press, 1988.

[21] D. Rose and P. Langley. Chemical discovery as belief revision. *Machine Learning*, 1:423–451, 1987.

[22] Donald Rose and Pat Langley. A hill-climbing approach to machine discovery. *Proceedings of the Machine Learning Conference*, pages 367–373, 1988.

[23] Herbert A. Simon and Raul E. Valdes-Perez. The evaluation and reporting of discovery systems. In *Proceedings of the ML92 Workshop on Machine Discovery*, pages 163–165, 1992.

[24] V. Soo, C.A. Kulikowski, D. Garfinkel, and L. Garfinkel. Theory formation in postulating kinetic mechanisms: Reasoning with constraints. Technical Report CBM-TR-150, Department of Computer Science, Rutgers, 1987.

[25] Christine Sutton. Happy birthday, little neutral one. *New Scientist*, 128(1747):34–36, 1990.

[26] Raul E. Valdes-Perez. Can reaction pathways be elucidated reliably? Unpublished draft.

[27] Raul E. Valdes-Perez. Coarse judgment of reaction model plausibility using linear estimation of reaction extent. Presented at Annual Meeting of American Institute of Chemical Engineers, 1990. Chicago, 11-16 November.

[28] Raul E. Valdes-Perez. Deductive assistance for elucidation of reaction pathways. Presented at Annual Meeting of American Institute of Chemical Engineers, 1990. Chicago, 11-16 November.

[29] Raul E. Valdes-Perez. *Machine Discovery of Chemical Reaction Pathways.* PhD thesis, Carnegie Mellon University, 1990. (School of Computer Science, CMU-CS-90-191).

[30] Raul E. Valdes-Perez. Symbolic computing on reaction pathways. *Tetrahedron Computer Methodology*, 3(5):277–285, 1990.

[31] Raul E. Valdes-Perez. A canonical representation of multistep reactions. *Journal of Chemical Information and Computer Sciences*, 31(4):554–556, 1991.

[32] Raul E. Valdes-Perez. On the concept of stoichiometry of reaction mechanisms. *Journal of Physical Chemistry*, 95(12):4918–4921, 1991.

[33] Raul E. Valdes-Perez. Algorithm to generate reaction pathways for computer-assisted elucidation. *Journal of Computational Chemistry*, 13(9):1079–1088, November 1992.

[34] Raul E. Valdes-Perez. A necessary condition for catalysis in reaction pathways. *Journal of Physical Chemistry*, 96(5):2394–2396, 1992.

[35] Raul E. Valdes-Perez. Theory-driven discovery of reaction pathways in the MECHEM system. In *Proc. of National Conf. on Artificial Intelligence*, 1992.

[36] Raul E. Valdes-Perez, Herbert A. Simon, and Robert F. Murphy. Discovery of pathways in science. In *Proceedings of the ML92 Workshop on Machine Discovery*, pages 51–57, 1992.

[37] Thomas H. Webb and Craig S. Wilcox. Improved synthesis of symmetrical and unsymmetrical 5,11-methanodibenzo[b,f][1,5]diazocines - readily available nanoscale structural units. *Journal of Organic Chemistry*, 55(1):363–365, 1990.

[38] W.T. Wipke, G.I. Ouchi, and S. Krishnan. Simulation and evaluation of chemical synthesis - SECS: An application of artificial intelligence techniques. *Artificial Intelligence*, 11:173–193, 1978. Special issue on applications to the sciences and medicine.

[39] J.M. Zytkow and H.A. Simon. Normative systems of discovery and logic of search. *Synthese*, 74:65–90, 1988.