ARTICLES

No Electron Left Behind: A Rule-Based Expert System To Predict Chemical Reactions and Reaction Mechanisms

Jonathan H. Chen[†] and Pierre Baldi[*,‡,†]

Institute for Genomics and Bioinformatics and Department of Computer Science, School of Information and Computer Sciences and Department of Biological Chemistry, University of California, Irvine, Irvine, California 92697-3435

Received May 3, 2009

Predicting the course and major products of arbitrary reactions is a fundamental problem in chemistry, one that chemists must address in a variety of tasks ranging from synthesis design to reaction discovery. Described here is an expert system to predict organic chemical reactions based on a knowledge base of over 1500 manually composed reaction transformation rules. Novel rule extensions are introduced to enable robust predictions and describe detailed reaction mechanisms at the level of electron flows in elementary reaction steps, ensuring that all reactions are properly balanced and atom-mapped. The core reaction prediction functionalities of this expert system are illustrated with applications including: (1) prediction of detailed reaction mechanisms; (2) computer-based learning in organic chemistry; (3) retrosynthetic analysis; and (4) combinatorial library design. Select applications are available via http://cdb.ics.uci.edu.

1. INTRODUCTION

Among the most fundamental problems in organic chemistry is predicting the course and major products of arbitrary reactions. In addition to being a fundamental scientific problem, reaction prediction is also important for several practical applications including the planning of new chemical experiments and syntheses. Seminal work in computer-aided reaction prediction was achieved with the CAMEO[1] and EROS[2] systems, and several other projects have made their own advances (e.g., Beppe,[3] ROBIA,[4] SOPHIA,[5] ToyChem[6]); however, most computer reaction prediction systems have fallen out of support over time. Thus developing an expert system capable of reliable reaction predictions remains one of the most important and unsolved problems in chemoinformatics.[7,8]

The relative lack of emphasis and support for reaction prediction is surprising given its fundamental importance for organic chemistry, especially considering the amount of attention given to the complementary problem of retrosynthesis. Although these two problems are closely intertwined, historically more attention has been given to computer-aided retrosynthetic analysis,[9] where one wishes to identify a synthetic pathway to yield a desired target product. A likely reason for this imbalance is the more obvious relevance of retrosynthesis toward obtaining important small molecules, including the majority of pharmaceutical drugs and natural products. Even within the scope of retrosynthetic analysis, however, reaction prediction is of direct relevance to solving one of the two key components of the analysis problem. The first component of the problem is the generation of retrosynthetic suggestions, while the second component is the validation of these suggestions as viable synthetic reactions. Without consideration for reactivity issues in the second component, generating retrosynthetic suggestions is relatively straightforward. A common approach involves searching a database of reactions or transformation rules for reaction centers that match the target compound of interest and proposing analogous transformations. Figure 1 illustrates how such suggestions, based on analogous examples, often do

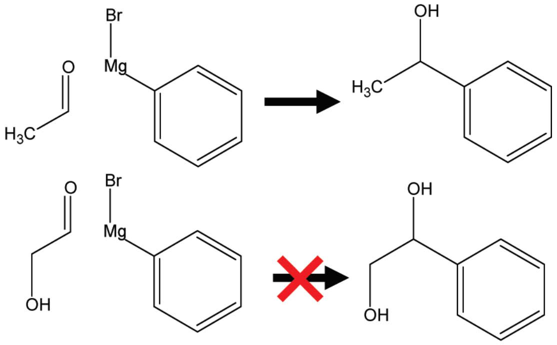


Figure 1. Retrosynthetic suggestions, illustrating the need for reaction validation capabilities. The first example illustrates a simple benzyl alcohol target compound and a proposed pair of precursor molecules to synthesize the target by a Grignard reaction. The second example illustrates a nearly identical target compound and the precursors that would be proposed by naively applying the analogous retrosynthetic transformation. This second suggestion is invalid because it does not consider the acid−base side reaction between the alcohol and organometallic reagent that will ruin the intended result.

[*] Corresponding author e-mail: pfbaldi@ics.uci.edu.

[†] Institute for Genomics and Bioinformatics and Department of Computer Science, School of Information and Computer Sciences.

[‡] Department of Biological Chemistry.

10.1021/ci900157k CCC: $40.75
Published on Web 08/31/2009

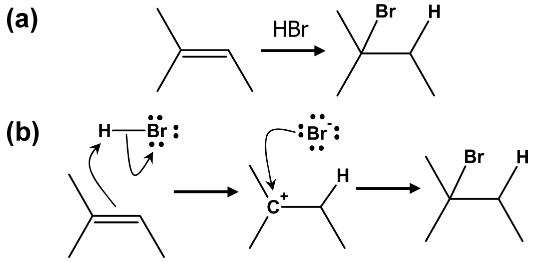


Figure 2. a - Representation for the overall "macroscopic" reaction of an alkene with hydrobromic acid, indicating the starting material, reagent, and final product. In the context of the system, the alkene starting material reactant and the selection of "HBr" as a reagent represents the expected input, while the alkyl bromide product represents the primary output. b - Detailed reaction mechanism for an alkene hydrobromination reaction, illustrating the underlying "microscopic" elementary processes that the overall reaction is based upon. This represents the detailed expected output when applying a reagent model for hydrobromic acid to the alkene reactant.

not consider functional group compatibilities and other unexpected reactivity issues that will invalidate the proposed reaction.

Existing computer-aided synthesis design systems have each addressed this problem of interfering chemical functionality to different degrees. The classic solution is to add "exclusion rules" to the suggested transformations. For the example in Figure 1, an exclusion rule could be added stating that this organometallic addition should only be suggested if none of the participating molecules contains an OH group. However, the problem is more complex because there are many other exclusion rules that would also be necessary in this example, such as the absence of SH, NH, other carbonyl, or nitrile groups. A more versatile option that has the potential to completely solve this problem is to develop a robust reaction predictor that can foresee these unexpected side reactions. To address the reaction validation component of retrosynthetic analysis, a reaction predictor could simply execute a virtual reaction on any proposed precursors to verify that the intended target is actually produced.

Beyond the scope of supporting retrosynthetic analysis, a robust reaction predictor would have many other immediate applications. For example, a reaction predictor could (1) systematically generate many reactions to power combinatorial library design and development;[10] (2) dynamically generate and validate content to support chemical education;[11] (3) propose mechanisms to explain the course of a reaction;[12,13] and (4) reveal previously undiscovered and useful reactivity.

2. METHODS

2.1. System Overview. We have developed a reaction expert system to predict the major products of a reaction, given a combination of starting materials and reagents. This functionality is implemented through two primary modules, a knowledge base of transformation rules and an inference engine to process those rules (Figure 3).

A key design decision for the system is determining what the knowledge base of transformation rules represents and, in particular, at what level of detail does the system model the predicted reactions. Most past systems have used a knowledge base of transformation rules that reflect the overall reactions from starting materials to final products (Figure 2a). However, using a single rule to reflect an overall "macroscopic" reaction obscures the "microscopic" elementary steps that underlie multistep reaction mechanisms (Figure 2b). To capture this mechanistic detail, the individual rules in our system are instead designed to mirror elementary reaction steps, from which the "macroscopic" reactions can be derived.

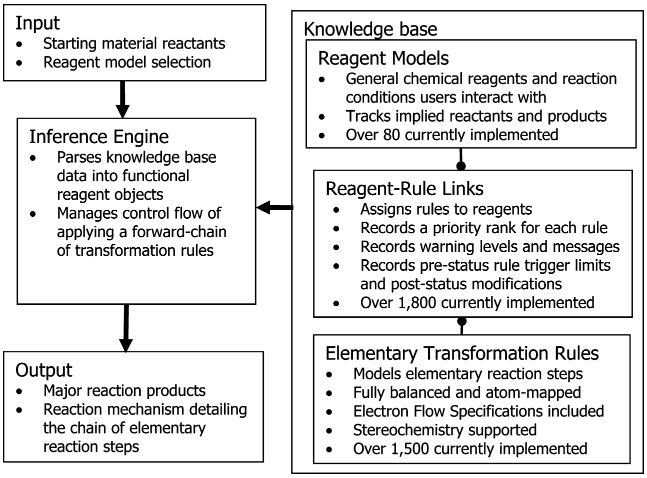

Input
Starting material reactants
Reagent model selection
Inference Engine
Parses knowledge base data into functional reagent objects
Manages control flow of applying a forward-chain of transformation rules
Output
Major reaction products
Reaction mechanism detailing the chain of elementary reaction steps
Knowledge base
Reagent Models
General chemical reagents and reaction conditions users interact with
Tracks implied reactants and products
Over 80 currently implemented
Reagent-Rule Links
Assigns rules to reagents
Records a priority rank for each rule
Records warning levels and messages
Records pre-status rule trigger limits and post-status modifications
Over 1,800 currently implemented
Elementary Transformation Rules
Models elementary reaction steps
Fully balanced and atom-mapped
Electron Flow Specifications included
Stereochemistry supported
Over 1,500 currently implemented

Figure 3. Overall architecture of the system. The knowledge base is implemented in a database, and the right column provides a simplified view of the database schema. There exists a one-to-many relationship between reagents and reagent-rule links and likewise between transformation rules and reagent-rule links. The combination of the previous relationships creates a many-to-many relationship between reagents and rules.

While the system's transformation rules model reactions at the level of elementary processes, users are typically not interested in directly observing this level of detail. Instead, users typically prefer interacting at the level of overall reactions or even more broadly at the level of general reagents and reaction conditions. To accommodate this high level interaction, the detailed transformation rules are aggregated into reagent models that represent general chemical reagents and reaction conditions (e.g., hydrobromic acid), which can then predict the overall course of specific reactions (e.g., alkene hydrobromination). Furthermore, to develop richer and more robust predictions, the elementary transformation rules are extended with additional information and control logic such as mechanistic electron flow specifications and priority values.

2.2. Elementary Transformation Rules. The core elementary rules in the system describe chemical structure transformations using the SMIRKS language, a simple extension of the SMILES (molecule) and SMARTS (chemical pattern matching) languages,[14] which is processed using the OEChem toolkit[15] from OpenEye Scientific Software. Though the SMIRKS specification does not require it, all the reaction equations represented by the transformation rules in the system are fully balanced with reactant atoms precisely mapped to corresponding product atoms. Ensuring that all reaction equations are fully balanced and atom-mapped is a detail often neglected by chemical data systems and even human chemists, but it is critical to ensure that transformation rules model elementary reaction steps rigorously. Table 1 lists examples of SMIRKS transformation rules that correspond to the elementary steps of the reaction mechanism

Table 1. SMIRKS Transformation Rules Corresponding to a Simple Alkene Hydrobromination Reaction Model[a]

SMIRKS	description
[C:1]=[C:2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition
[C+:1].[-:2]≫ [C+0:1][+0:2]	carbocation, anion addition

[a] Each item in brackets corresponds to an atom in the reaction equation. The “≫” symbol delimits reactants from products. The numbers following colons are atom-map indexes used to specify which reactant atoms correspond to which product atoms. Further specification of the SMIRKS language can be found in the references.[14]

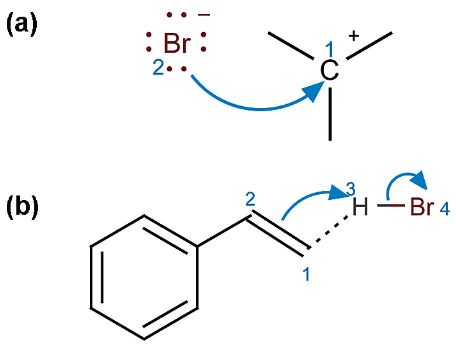


Figure 4. a - Arrow pushing mechanism diagram generated when applying the SMIRKS reaction transformation rule [C+:1].[-:2]≫[C+0:1][+0:2] and electron flow specification 2=1 to a carbocation electrophile and a bromide anion nucleophile. This represents the movement of two electrons from atom 2 to atom 1. b - Arrow pushing mechanism diagram generated when applying the SMIRKS reaction transformation rule [C:1]=[C:2][c:10].[H:3][Cl,Br,I:4]≫[H:3][C:1][C+:2][c:10].[Cl,Br,I;0:4] and electron flow specification 2,1=1,3;3,4=4 to styrene in hydrobromic acid, representing protonation of the nucleophilic π orbital electrons. The dashed line between carbon 1 and hydrogen 3 indicates the site of a forming bond in the mechanism step.

depicted in Figure 2b. Currently over 1500 distinct transformation rules have been manually composed in our system.

2.2.1. Electron Flow Specifications. The reaction transformation rules developed for this expert system are designed to mirror elementary reaction steps, which makes it relatively straightforward to extend their function to generating curved arrow mechanism diagrams.[12,13] This is achieved by attaching to each elementary transformation an additional string indicating where the flow of electrons should begin and end within the reaction intermediates. Figure 4a,b illustrates this method by applying a SMIRKS transformation rule to predict the product of an elementary step in combination with an electron flow specification.

The electron flow specification language, described below and illustrated in Figure 4a,b, was created for this reaction expert system as a SMIRKS language extension to support mechanistic detail in reaction transformation rules. The typical form of one of these specifications is “n_1,n_2=n_3,n_4” where n_1, n_2 are the indexes associated with the source atoms flanking the bond of origin for the electron flow arrow, while n_3, n_4 are the indexes associated with the target atoms flanking the new bond that will be formed by the elementary reaction step. A similar string like “n_1,n_2-n_3,n_4” represents the movement of a single electron (i.e., a free radical reaction) instead of the more typical movement of a pair of electrons. The complete set of symbols used in this language is listed in Table 2.

Table 2. Definition of the Symbols That Can Be Used in the Electron Flow Specification Language

symbol	description
;	delimits specifications for diagrams with multiple electron flow arrows
=	represents a double-headed arrow for the movement of 2 electrons, delimiting source from sink atoms
-	represents a single-headed “fishhook” arrow for the movement of 1 electron (i.e., free radical reactions), delimiting source from sink atoms
n_i	numerical indexes which identify atoms representing sources and sinks for the arrows
,	atom delimiter for when the source or target of the arrow consists of multiple atoms (i.e., 2 atoms specified to indicate bond electrons)—The order in which the atoms are listed here does not affect the resulting diagram.

While using this specification language, certain nuances in electron arrow pushing diagrams must be highlighted. One potential issue is that the specification may seem to imply that arrows can originate from the nuclei of atoms when in reality they are meant to represent the movement of the electrons. Obviously, the intended meaning in these scenarios is that the arrows represent the movement of the electrons (lone pair or free radical) associated with the atom and not of the actual atom nucleus. Thus the specification language assumes that the user is capable of identifying lone pair and free radical electrons. Unfortunately, ChemAxon’s MarvinView module,[16] used for the system’s visualization of mechanism diagrams, does not presently include proper support for explicit lone pair or free radical entities. Instead, the MarvinView arrows in these cases must currently be drawn as originating from an atom, despite the atom’s electrons being the intended origin of these arrows.

Another nuance for these curved arrow mechanism diagrams is revealed when an electron source is a bond. In such cases, the target atom list *must* contain one of the source atoms to yield unambiguous results. In Figure 4b, this is represented by the dashed bond between atom 1 and 3, indicating a forming bond. Without this information (the dashed line or any equivalent), the reader is uncertain as to whether atom 1 or 2 should be bonded to atom 3 after the electrons have moved.

2.2.2. Stereochemistry. Many reaction examples in this manuscript have their stereochemistry simplified for clarity, but the actual system enforces that all molecules processed have complete stereochemistry specified. This ensures that any reactions that actually have stereospecific outcomes are modeled appropriately based on fully specified inputs. To achieve this, any molecule processed by the system that contains unspecified stereocenters has all of its stereocombinations enumerated to represent the corresponding racemic mixture (Figure 5). For unspecified E vs Z stereochemistry of double bonds, rather than enumerate both possibilities, the system preferentially selects the isomer that keeps the largest substituents *trans* with respect to each other.

For reactions that actually have stereospecific or stereoselective outcomes, the SMIRKS language already has the expressive power to represent these transformations as illustrated in Figure 6. These representations are based on the use of the “@” symbol to specify atom chirality and the “/” and “\” symbols to specify bond chirality.[14]

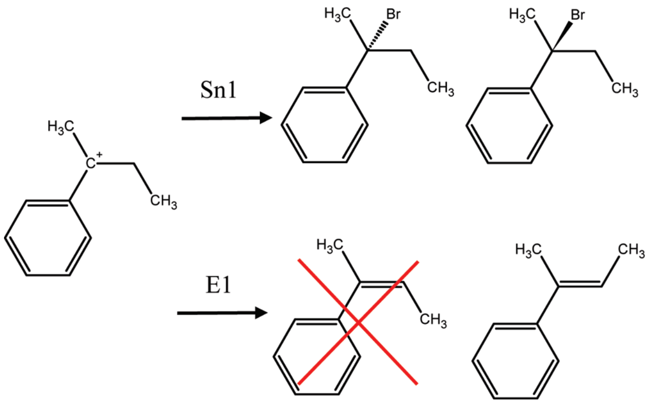


Figure 5. Products generated by Sn1 and E1 reactions applied to an achiral benzyl carbocation. No stereo selectivity exists for this Sn1 substitution reaction, so the system will enumerate both enantiomers as possible products of the reaction to represent a racemic mixture. The E1 elimination reaction could theoretically produce two diastereomeric products, but the system will preferentially select the one which keeps the largest substituents *trans* with respect to each other, to reflect the typical pattern of stereo selectivity in these reactions.

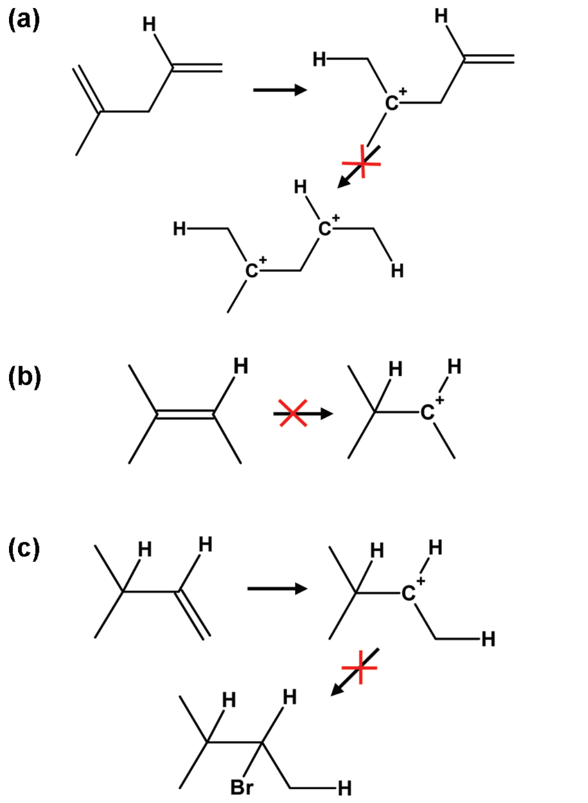
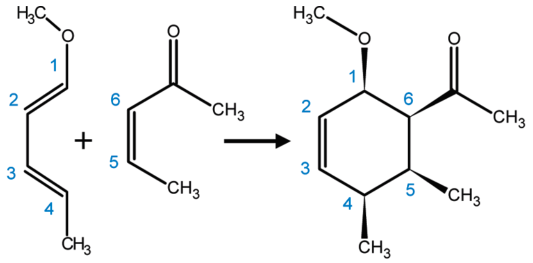


Figure 7. a - Potential prediction mistake: Adding another proton to a carbocation intermediate, yielding an unlikely species containing multiple atoms with a positive charge. b - Potential prediction mistake: Ignoring regioselective preference for carbocations on more substituted carbons. Protonation that yields a secondary carbocation is highly unlikely when the protonation could occur on the other end of the double bond, yielding a tertiary carbocation of greater stability. c - Potential prediction mistake: Ignoring the possibility of unintended reactivity such as carbocation rearrangements. The anion (Br^-) does not add directly to the secondary carbocation. Instead, because the carbocation is adjacent to a tertiary center, a 1,2 hydride shift will alter the site of the carbocation.

2.2.3. Potential Prediction Mistakes. Transformation rules as described thus far are still insufficient to provide robust reaction predictions. If alkene hydrobromination reactions were based solely on the two SMIRKS rules from Table 1, many mistakes would be made, such as those illustrated in Figure 7a−c.

To develop more robust predictions that address the above issues, we must add more specific transformation rules and prioritize the list of rules by an appropriate precedence order.

2.3. Reagent-Rule Links. Reagents are basically ranked collections of elementary transformation rules. In the database implementation of the knowledge base, reagent-rule link tables assign transformation rules to reagent models, with an additional priority rank value to indicate which rules should be attempted first before descending down the precedence order. To a first order approximation, the rules are ranked in terms of the relative "reactivity" of the step being modeled. All rule links have a priority rank specified to enforce complete ordering of the rules within a reagent model, though ties are allowed to represent elementary reaction steps that are equally likely to occur (e.g., Sn1 vs E1 termination of a carbocation intermediate). Table 3 includes a subset of the linked rules from the complete HBr



```
[*:10]/[CX3;$(*[O,N,S]):1]=[C:2]\-[C:3]=[CX3:4]/[*:11].
[*:12]\[CX3:5]=[CX3:6]/[$(*=[O,N,S]),$(C#N):13]>>
[*:10][CX4;@:1]1-/[C:2]=[C:3]\-[CX4;@@:4]([*:11])[CX4;@:5]([*:12])[CX4;@@:6]1[*:13]
```

Figure 6. Product prediction for a Diels−Alder reaction using the accompanying SMIRKS transformation rule. The example illustrates the expressive power of the rule to enforce the regioselectivity, stereospecificity, and stereoselectivity of the reaction. Regioselectivity: carbon 1 preferentially assumes an *ortho* position with respect to carbon 6, based on the pattern of their substituents. Stereospecificity: The (*E,E*) diene results in substituents at carbon 1 and 4 that are *syn* with respect to each other and likewise for the *Z* dienophile resulting in *syn* substituents at carbon 5 and 6. Stereoselectivity: Assuming kinetic preference for the *endo* product over the *exo* product, the orientation of stereocenters at carbon 1 and 4 is defined with respect to those at carbon 5 and 6. This example illustrates one product generated by the stereospecific Diels−Alder reaction, but the actual reaction will yield a racemic mixture including the enantiomeric product. In such cases, a second copy of the SMIRKS rule is included. The second rule is modified with inverted stereospecification symbols to yield the respective enantiomeric product.

Table 3. Example of 10 Prioritized Transformation Rules, Relating to Alkene Hydrobromination Reactions, out of the 92 Rules Used in the Complete Robust HBr Reagent Model[a]

SMIRKS	description	electron flow	priority
[H:10][CH1:1][C+;!H0:2]≫ [C+:1][C+0:2][H:10]	carbocation, hydride shift from tertiary	1,10=10,2	10
[C:10][CH0:1][C+;!H0:2]≫ [C+:1][C+0:2][C:10]	carbocation, methyl shift from quaternary	1,10=10,2	9
[H:10][CH2:1][CH2+:2]≫ [C+:1][C+0:2][H:10]	carbocation, hydride shift from secondary	1,10=10,2	8
[C+:1].[-:2]≫ [C+0:1][+0:2]	carbocation, anion addition	2=1	7
[C:1]=[C;$(*O):2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition, alkoxy	2,1=1,3;3,4=4	6
[C:1]=[C;$(*a):2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition, benzyl	2,1=1,3;3,4=4	5
[C:1]=[C;$(**=*):2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition, allyl	2,1=1,3;3,4=4	4
[C:1]=[CH0:2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition, tertiary	2,1=1,3;3,4=4	3
[C:1]=[CH1:2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition, secondary	2,1=1,3;3,4=4	2
[C:1]=[C:2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition	2,1=1,3;3,4=4	1

[a] Included for each transformation rule is not only the SMIRKS pattern and description but also a relative priority rank to indicate the order in which the rules should be attempted. The existence of several variants for similar rules and the customized priority ordering enables robust reaction predictions that address the issues noted in Figure 7. An electron flow specification accompanies each rule to support curved arrow mechanism diagrams.

reagent model. Patterns worth noting that address the issues mentioned in Figure 7 include the following:

1. "Carbocation, anion addition" is ranked higher than any "alkene, protic acid addition" to prevent production of a species with multiple positive charges.

2. Several "alkene, protic acid addition" rules exist, differentiated by what kind of carbocation each would yield. These are ranked to ensure the more stable carbocations will be formed before any others are attempted.

3. Carbocation rearrangement rules are added with high priority, and again note that several exist to account for the different possibilities, ranked respectively.

In addition to a priority ranking value, the rule link records include other supporting information and control logic. For example, warning levels and messages are attached to the carbocation rearrangement rule links in Table 3 such that, when these rules are triggered, the system can alert the user to these reaction steps that are likely to be unintended or undesirable. The system can also control the timing of certain rule triggers based on the assignment of a status number to (intermediate) reactant molecules. Rule links include a poststatus number such that application of the rule transforms not only the reactant molecule structure but also the status number associated with it. The convention established in the system is that starting material reactants begin with a status number of 0 and typically result in a status number of 100 to reflect conversion into a final product. This is particularly useful for managing reagent models representing multistage reactions, such as first treating a substrate with $LiAlH_4$ and then following with aqueous workup. To achieve this two-stage effect, the rules associated with the $LiAlH_4$ step modify the input reactant's status number from 0 to 100, while the rules associated with the aqueous workup have prestatus conditions that will not allow them to trigger until the input reactants have achieved a status number of at least 100.

2.4. Reagent Models. When chemists depict reagent usage on paper, such as the hydrobromination example in Figure 2a, they typically just write "HBr" over a reaction arrow. To model what is represented there, the reagent is assigned a collection of elementary transformation rules with priority rank values, such as those in Table 3. To complete the reagent model, additional information on implied reactants and products must be tracked. This information is often neglected by chemical data systems and human chemists alike, but it is necessary to satisfy the fully balanced and atom-mapped reaction equations. For example, alkene hydrobromination reactions (Figure 2a) should be aware of implied "HBr" reactants to clearly specify the source of those product atoms. Similarly, condensation reactions (Figure 8) should be aware of implied "H_2O" products to clearly specify the fate of those reactant atoms, instead of implying that they were annihilated.

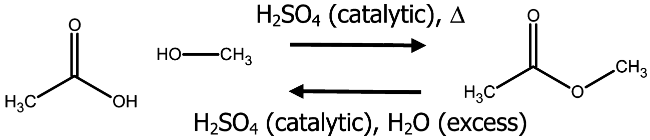


Figure 8. Reversible Fischer esterification reaction whose preferred outcome can be altered by selecting a reagent model representing a particular set of reaction conditions. In this example, both the forward and reverse reaction can be driven using an aqueous acid reagent, but selection of a concentrated acid, under heat, will tend to distill any H_2O, driving the reaction toward the ester. Selection of a dilute acid with excess H_2O solvent will drive the reaction toward the hydrolyzed carboxylic acid and alcohol. The reagent models for these two cases are nearly identical in their reaction rule content, but differences are achieved by respectively adjusting the priority rank of the hydrolysis vs esterification rules in each reagent model.

2.4.1. Recurring Reactivity Patterns and Variable Reaction Conditions. In developing the reagent models, many recurring mechanisms and reactivity patterns are found. As a result, general reaction rules can be reused in many reagent models. For example, to develop a reagent model for aqueous sulfuric acid that can add H_2O to alkenes, we can reuse almost every single rule in Table 3 without having to duplicate any effort. The only change necessary is to replace the "carbocation, anion addition" rule with a similar rule for "carbocation, hydroxyl addition". This kind of rule reuse is facilitated by the reagent-rule link records which allow for a many-to-many relationship between reagents and rules. Furthermore, a simple reagent inheritance framework is supported where all of the general rules associated with a common pattern (e.g., carbocation chemistry) are linked to an abstract reagent model and then the actual, concrete reagent models that users interact with (e.g., the HBr reagent)

Table 4. List of Reaction Topics Currently Covered in the System[a]

section	description
5	alkenes
9.04	substitution (nucleophilic) of alkyl halides
9.05	elimination reactions of alkyl halides
10	alcohols and epoxides
11.04	epoxides and organometallic compounds
11.05	oxidation of alcohols and alkenes
14	alkynes
15	dienes, conjugation, Diels−Alder
16	electrophilic aromatic substitution
17	allylic and benzylic reactivity
17.02	alkanes, radical reactions
18	transition metal (Pd) catalysis
18.04	SnAr and benzyne reactions
19	aldehydes and ketones
20.1	redox of alcohols and carbonyls
21	carboxylic acid derivatives
22	enolate chemistry
22.04	aldol chemistry and Michael addition
22.05	Claisen condensations
22.08	organometallic addition, conjugate addition
23	amines
23.1	arenediazonium reactions
24	naphthalene and heteroaromatic EAS reactions
24.05	pyridine derivatives
25	pericyclic reactions
26.04	amino acid synthesis
26.07	peptide synthesis
27	carbohydrates

[a] Section numbers correspond to the Loudon textbook, though the system is not tied to any particular content source since it is designed to model the fundamental underlying chemistry. Gaps in the section numbers correspond to textbook chapters that do not include any relevant reactions to model, such as chapters on stereochemistry or spectroscopy.

simply “inherit” any rules from the abstract reagent model without having to reconnect each individual rule to the concrete reagent model.

This reuse of rules, representing recurring reactivity patterns, makes it convenient to develop reagent models representing similar but variable reaction conditions. For example, there are several reagent models in the system representing treatment with aqueous (sulfuric) acid, but these are qualified with variations in their reaction conditions such as “catalytic,” “cold, dilute”, “hot, dilute”, “cool, concentrated/fuming”, and “hot, concentrated”. These reagent models have identical effects in many cases, as they are based on a common set of reaction rules, but they can include additional rules or override rule priority rankings to achieve variable results. Figure 8 illustrates a reversible reaction which can be completed in either direction by selecting the appropriate reaction condition variant of an aqueous acid reagent model. These reagent models are nearly identical in reaction rule content, but the model representing a dilute acid with excess H_2O has a rule for substrate hydrolysis, which is priority ranked higher than in the complementary reagent model for concentrated acid (minimal or distilled H_2O). Finally, a few “reagent” models in the system do not even represent an actual chemical reagent but rather fairly generic solvent conditions in cases where the reactions are driven primarily by the reactants(Table 5).

Table 5. Listing of 80 Reagent Models Currently Implemented in the System That Users Can Combine with Reactant Molecules To Predict the Course and Major Products of Reactions[a]

reagent model descriptions	
Pd(0) (catalyst)	Clemmensen reduction (acid)
pericyclic reactions (thermal)	Wolff−Kishner reduction (base)
mix reactants, aprotic	oxidation (base, permanganate)
mix reactants, protic	oxidation (acid, chromate)
hydrogen fluoride (Friedel−Crafts catalyst)	oxidation (MnO2, benzylic, partial)
Lewis acid (Friedel−Crafts catalyst)	oxidation (PCC)
sulfuric acid (catalytic)	oxidation (nitric acid)
sulfuric acid (cold, dilute)	SOCl2
sulfuric acid (hot, dilute)	PBr3
sulfuric acid (cool, concentrated/fuming)	tosylation
phosphoric acid (hot, concentrated)	triflate preparation
bromination, Lewis acid	POCl3
nitric acid	P2O5
NaOH	acetic anhydride
NaOEt	DCC
NaH	Mg (Grignard)
NaNH2	lithium
LDA	organocuprate preparation
hydroboration-oxidation	organostannane preparation
hydrobromination	bromination, radical
hydrobromination, peroxide	NBS, peroxide
bromination	PPh3, BuLi (phosphonium ylide prep)
bromohydrin	TMSCl, Et3N
Br2, H3O+	Fmoc Amine Protection
Br2, NaOH	NH4+ F-
hydrogenation	Fmoc deprotection (piperidine)
hydrogenation, partial	TFA (para-oxy benzyl deprotection)
hydrogenation, Pd/BaSO4	NH4Cl, NaCN
Na, NH3	cyanohydrin
O3, CH3SCH3	benzylic halide (para-oxy)
O3, H2O	arenediazonium prep (HCl)
OsO4, NaHSO3 (syn dihydroxylation)	arenediazonium prep (H2SO4)
periodic acid (HIO4)	arenediazonium prep (HBr)
peroxyacid (mCPBA)	arenediazonium (F)
Sharpless epoxidation (+)-DET	arenediazonium (Cl)
Sharpless epoxidation (-)-DET	arenediazonium (Br)
LiAlH4	arenediazonium (I)
DIBALH	arenediazonium (C#N)
NaBH4	arenediazonium (H) hypophosphorus acid
NaBH3CN	Hofmann elimination

[a] A few models do not reflect an actual chemical reagent but instead represent reactions driven primarily by the reactants in a generic solvent. In particular, there are “reagents” for simply mixing the reactants in different solvent types (“mix reactants, protic” and “mix reactants, aprotic”) and one for mixing the reactants under heat to model thermal pericyclic reactions (“pericyclic reactions (thermal)”).

3. RESULTS

3.1. Core Predictive Capabilities. The core reaction prediction capabilities supported by the system are illustrated in Figures 9 and 10. The products and curved arrow mechanism diagrams for these reactions are not precoded results but rather are dynamically predicted within a second based on the input reactant and reagent combinations.

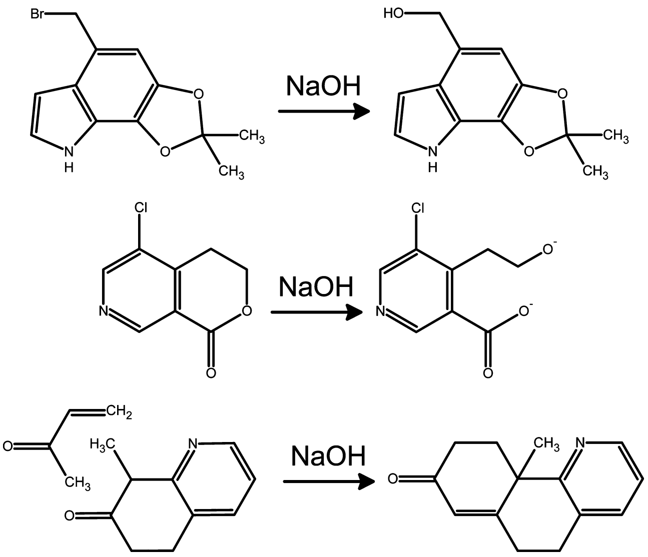
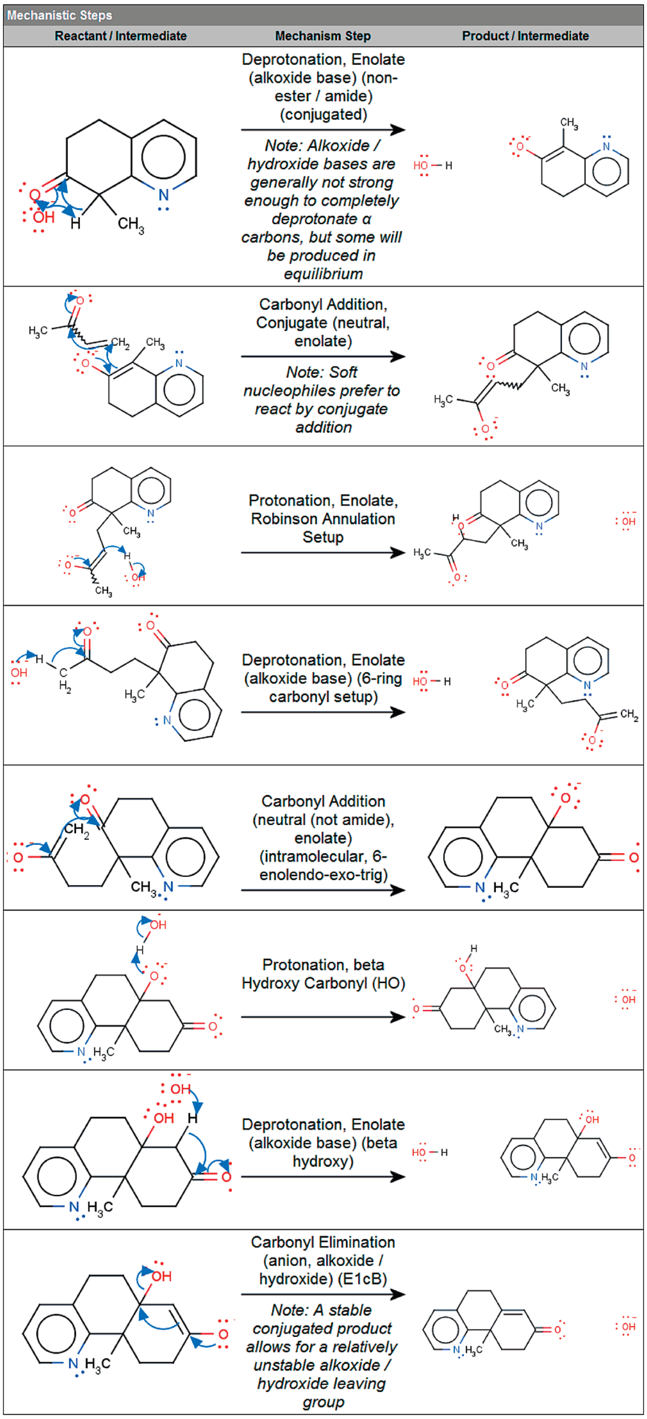


Figure 9. Progressively more complex reactions predicted by the system (Sn2 substitution, nucleophilic acylation/saponification, Robinson annulation), all based on a single common reagent model (NaOH). This reagent model contains relevant reactivity rules to represent a set of general reaction conditions as opposed to one rule for each specific reaction.



Mechanistic Steps

Reactant / Intermediate	Mechanism Step	Product / Intermediate
	Deprotonation, Enolate (alkoxide base) (non-ester / amide) (conjugated) *Note: Alkoxide / hydroxide bases are generally not strong enough to completely deprotonate α carbons, but some will be produced in equilibrium*	
	Carbonyl Addition, Conjugate (neutral, enolate) *Note: Soft nucleophiles prefer to react by conjugate addition*	
	Protonation, Enolate, Robinson Annulation Setup	
	Deprotonation, Enolate (alkoxide base) (6-ring carbonyl setup)	
	Carbonyl Addition (neutral (not amide), enolate) (intramolecular, 6-enolendo-exo-trig)	
	Protonation, beta Hydroxy Carbonyl (HO)	
	Deprotonation, Enolate (alkoxide base) (beta hydroxy)	
	Carbonyl Elimination (anion, alkoxide / hydroxide) (E1cB) *Note: A stable conjugated product allows for a relatively unstable alkoxide / hydroxide leaving group*	

Figure 10. Reaction mechanism details page generated by the system to illustrate the chain of elementary reaction steps used to predict the outcome of the Robinson annulation reaction at the end of Figure 9. Each step includes a system-generated curved arrow mechanism diagram and an accompanying verbal description. Some steps include additional informative or cautionary notes to assist the user.

Furthermore, the reagent models implemented in the system do not correspond to specific "name reactions[17]" such as a "Michael addition" or "Claisen condensation". Instead, reagent models represent general reaction conditions such as treatment of the reactants with NaOH. The NaOH reagent model, for example, contains rules giving it the ability to predict many different reactivity patterns, including Sn2/E2 chemistry, nucleophilic acylations, aldol chemistry, and more. This general predictive power is possible without "memorizing" every possible "macroscopic" reaction pattern because the reagent rule sets are designed to match the "microscopic" elementary reaction steps at the mechanistic level of detailed electron flows. This allows the system to automatically derive specific overall, balanced and atom mapped, reaction patterns based on their underlying mechanisms.

Currently over 80 reagent models (listed in Table 5) have been developed for the system based on over 1500 prioritized SMIRKS transformation rules. The examples described above were chosen for simplicity of presentation, but significantly more complex reactivity has been modeled with these rules. Reaction topics currently implemented are listed in Table 4, based on sections of content adapted from the Bruice,[18] Loudon,[19] and Smith[20] organic chemistry texts.

3.2. System Validation Process. The examples in Figure 9 illustrate a few specific reactions the system is known to reproduce accurately and consistently, but to ensure prediction validity across a range of possible inputs, we have manually composed over 4500 specific reaction test cases for the system. These test cases systematically cover a range of relevant functional group combinations, including negative test cases where the most reasonable prediction is that "no reaction" will occur (e.g., treatment of a saturated hydrocarbon with acid or base). As part of a rigorous unit testing process, new test cases are added whenever the system's rule set is expanded or modified. Before any changes to the rule set are accepted, all new and prior test cases are verified to ensure prediction validity remains intact. Furthermore, as a limited form of crowd-sourcing, users can submit a "challenge" if they believe any reaction predicted by the system is incorrect. After a few years of system service and over 3000 users, a few dozen challenges have been submitted, but less than a handful actually identified legitimate prediction errors in the system. These few legitimate challenges alerted us to make changes in the rule set to correctly handle novel inputs, but in most of the remaining cases, the challenges came from student users who did not fully understand the chemistry involved.

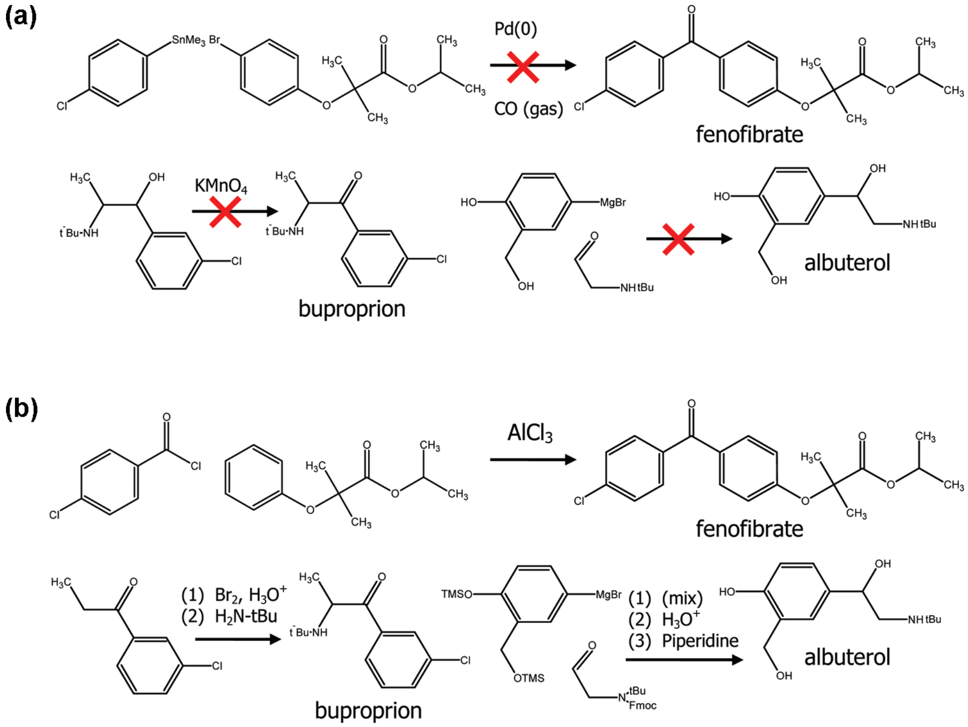


Figure 11. a - Examples of major pharmaceutical drugs and synthetic reactions proposed by naively applying a typical retrosynthetic pattern matching approach. The expert system's robust reagent models can provide the critical "expertise" of synthesis plan validation by identifying all of these proposed plans as ineffective due to unintended side reactions. Fenofibrate: The proposed organostannane precursor is difficult to prepare in the presence of another aryl halide. Buproprion: overoxidation of the benzylic alcohol is likely. Albuterol: organometallic Grignard reagent cannot be produced in the presence of acidic OH groups. b - Examples of possible synthetic reactions for several pharmaceutical drugs that the expert system's reagent models can reproduce, validating the intended reaction products.

3.3. Chemical Education. A specific application of the expert system that illustrates many of its predictive capabilities is a chemical education system to support the learning of organic chemistry reactions, syntheses, and mechanisms.[11] This educational application challenges students to solve organic synthesis and mechanism elucidation problems, but, unlike typical online learning applications, the underlying expert system enables teaching support for instructors and a richer learning experience for students. For instructors, this includes automated problem generation and grading. For students, this includes also automatic problem generation as well as the fostering of inquiry-based learning[21] where students can conduct and observe virtual experiments by selecting their own novel reactant and reagent combinations. This chemical education application has been tested in several courses at the University of California, Irvine where correlative evidence indicates that students who use the system score on average ~10% better on examinations than those who do not.[11]

3.4. Validation of Synthesis Plans. A natural extension of the reaction prediction functionality is to apply it toward solving retrosynthetic design problems.[22] To a large degree, all that is necessary is to take the transformation rules which normally convert reactants into predicted products and invert them to instead convert target products into proposed precursors. Making these kinds of retrosynthetic suggestions is relatively straightforward and commonplace among computer-aided synthesis design tools,[9,23,24] but the additional value gained here is that any proposed precursors can be passed back through the forward reaction prediction reagent model to validate that the intended target will actually be produced by the proposed precursors. This forward validation step can further contribute a reliability score to any proposed reaction by predicting whether a mixture of different products or stereoisomers could cripple the yield of the intended product.

3.5. Combinatorial Library Design. Another application of the reaction prediction technology is in combinatorial library design. This applies for both general library design,[10] where virtual molecules are systematically enumerated from an initial pool of building blocks, and targeted library design,[22] where virtual molecules that are structurally similar to a target compound are constructed from building blocks that are similar to substructures of the target.

The additional value of the reaction prediction technology is that it provides a natural solution to the library design problem of generating virtual compounds of reasonable synthetic feasibility.[25] Design of a combinatorial library by enumerating all possible structures up to some constraint of atom number[26] or reaction types[27] can generate many possible structures but leaves open the challenge of filtering down to those that could be readily synthesized. Rather than developing heuristic rules or scoring functions to estimate synthetic feasibility, systematically applying the expert system's reagent models to an initial pool of available starting materials will generate a large virtual library of compounds while simultaneously proposing a reasonable synthetic reaction to produce each compound. Furthermore, the robust reagent models allow the generation process to easily filter out any proposed reactions that would yield undesirable side reactions or mixtures as illustrated in Figure 12.

3.6. Reaction Discovery. Given the preprogrammed nature of a rule-based system, it seems unlikely that this system could discover any new types of reactions that were not

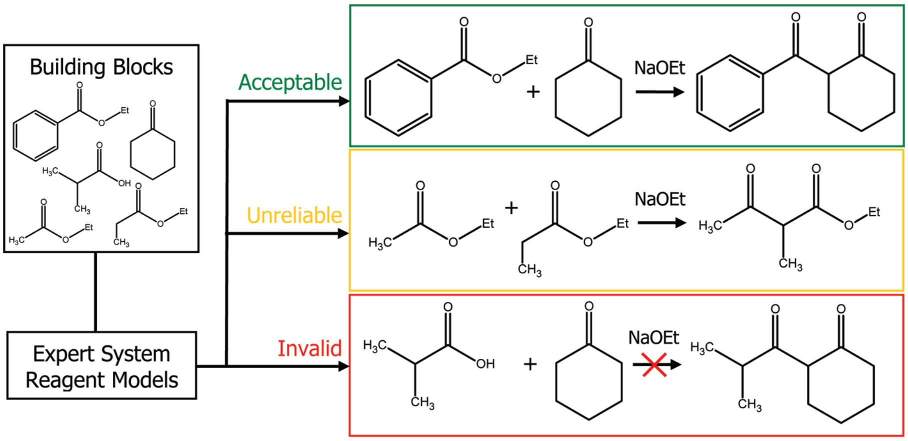


Figure 12. Flowchart for designing a combinatorial library using the expert system's reagent models. A collection of available building blocks is passed through the reagent models in all (pairwise) combinations to predict reasonable virtual products along with a specific synthetic reaction proposal for each. The reagent models will naturally sort the results into relevant subsets based on the proposed synthetic reactions. An example product and respective synthetic reaction is illustrated for each of these subsets. Acceptable: Products generated from proposed reactions that the system validates as reasonable and effective for use in the library. Unreliable: Products generated from proposed reactions that may work but are likely to produce an unreliable mixture of many side products. Invalid: Products for which no proposed synthetic reaction is acceptable, due to side reactions that will disrupt the intended result.

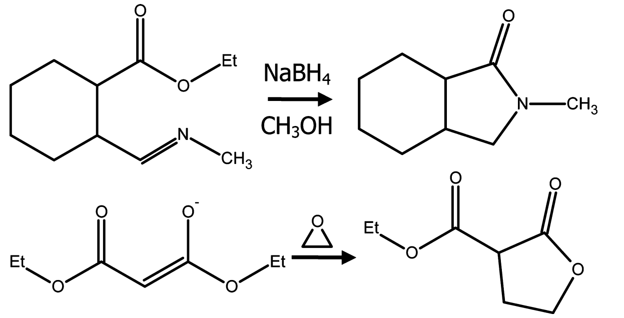


Figure 13. Reactions predicted by the system with results that defied straightforward expectations. The results are based on additional combinations of elementary reaction steps that the system recognizes can be chained together. The first reaction is expected to be a simple hydride reduction of the imine to produce a nitrogen anion that is subsequently neutralized by the protic solvent. Instead, the system recognizes that the nitrogen anion intermediate is a strong nucleophile that can attack the nearby ester to form a lactam before the solvent neutralization. The second reaction is expected to open the epoxide by the enolate nucleophile. The system does predict that effect, but it also recognizes the epoxide opens up to yield an oxygen anion which is a strong nucleophile that can reach back to attack the original ester to form a lactone.

already known by the knowledge engineer who authored the rules. While this may be true in terms of discovering individual elementary reaction processes, the many possible ways elementary steps can be composed into overall reactions may discover novel results. Figure 13 illustrates example reactions predicted by the system where a straightforward result was expected, but the system continued to identify and apply transformation rules for reasonable elementary reaction steps which resulted in different overall reaction patterns.

4. DISCUSSION

A reaction expert system founded upon fundamental reaction prediction capabilities has been developed to provide a platform for addressing problems ranging from retrosynthetic analysis and combinatorial library design[22] to mechanism elucidation and chemical education.[11] The prediction system is based on over 1500 manually composed transformation rules representing fully balanced and atom-mapped elementary reaction steps with over 4500 test cases to validate prediction accuracy and consistency.

While this rule-based approach to reaction prediction is already useful in many applications, the approach does have its limitations. Currently just over 80, most common, reagent models are implemented in the system, but it may require hundreds of reagent model variations to achieve comprehensive coverage of the breadth of modern organic chemistry. While additional reagent models can always be added to expand the system's coverage, the size and complexity of the rule set makes progressive addition of rules increasingly more challenging.

An alternative approach to manually composing reaction rules is to automatically generate the rules by perceiving common patterns from reaction databases.[28,29] The automation of these approaches is certainly appealing, but the depth of prediction models they can generate is often limited by the reaction databases available for them to work from. Full access to large reaction databases is often highly restricted, and, even if data mining access is possible, the data are noisy and tend to lack fully balanced and atom-mapped reactions. Furthermore, these data almost always describe overall "macroscopic" reactions with no detail on the underlying mechanisms and elementary reaction steps.

To achieve a greater level of generality and robustness, an alternative reaction predictor design could instead be driven by more fundamental principles of molecular orbital theory[30] and reaction kinetics simulations, though it would probably do so at the cost of longer prediction times. Even for the development of such a principle-driven approach, the rule-based system described here can be useful for generating a virtual database of detailed reaction mechanisms with fully balanced and atom-mapped reaction data to train and validate the principle-driven system. In the meantime, the rule-based expert system already provides a platform for solving a range of important chemistry applications by addressing the fundamental problem of chemical reaction prediction, at a unique mechanistic level of detail, with subsecond prediction times.

ACKNOWLEDGMENT

Work supported by an NIH Biomedical Informatics Training grant (LM-07443-01), NSF grants EIA-0321390 and 0513376, and a Camille and Henry Dreyfus Foundation grant to P.B. We acknowledge OpenEye Scientific Software, Peter Ertl of Novartis (JME Editor), and ChemAxon for academic software licenses. We thank Drs. Suzanne Blum, Zhibin Guan, Elizabeth Jarvo, Susan King, Larry Overman, Scott Rychnovsky, Kenneth Shea, Mare Taagepera, David Van Vranken, Chris Vanderwal, and Gregory Weiss and their students for their feedback and usage of the system in their chemistry courses. We acknowledge Matthew A. Kayala, Peter Phung, and Paul Rigor for contributing to software design and development. We thank Dr. James Nowick for additional feedback and comments.

REFERENCES AND NOTES

(1) Jorgensen, W. L.; Laird, E. R.; Gushurst, A. J.; Fleischer, J. M.; Gothe, S. A.; Helson, H. E.; Paderes, G. D.; Sinclair, S. CAMEO: a program from the logical prediction of the products of organic reactions. *Pure Appl. Chem.* **1990**, *62* (10), 1921–1932.

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

(3) Sello, G. Reaction prediction: the suggestions of the Beppe program. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (6), 713–717.

(4) Socorro, I. M.; Taylor, K.; Goodman, J. M. ROBIA: A Reaction Prediction Program. *Org. Lett.* **2005**, *7* (16), 3541–3544.

(5) Satoh, H.; Itono, S.; Funatsu, K.; Takano, K.; Nakata, T. A Novel Method for Characterization of Three-Dimensional Reaction Fields Based on Electrostatic and Steric Interactions toward the Goal of Quantitative Analysis and Understanding of Organic Reactions. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (4), 671–678.

(6) Benko, G.; Flamm, C.; Stadler, P. F. A Graph-Based Toy Model of Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (4), 1085–1093.

(7) Gasteiger, J.; Engel, T. *Chemoinformatics: A Textbook*; Wiley-VCH: 2003.

(8) Hemmer, M. C. *Expert Systems in Chemistry Research*; CRC Press: Boca Raton, FL, 2008.

(9) Todd, M. H. Computer-Aided Organic Synthesis. *Chem. Soc. Rev.* **2004**, *34* (3), 247–266.

(10) Schnur, D. M. Recent trends in library design: ‘rational design’ revisited. *Curr. Opin. Drug Discovery Dev.* **2008**, *11* (3), 375–380.

(11) Chen, J. H.; Baldi, P. Synthesis Explorer: A Chemical Reaction Tutorial System for Organic Synthesis Design and Mechanism Prediction. *J. Chem. Educ.* **2008**, *2008* (85), 1699.

(12) Miller, A. *Writing Reaction Mechanisms in Organic Chemistry*; Academic Press: San Diego, CA, 1992.

(13) Grossman, R. *The Art of Writing Reasonable Organic Reaction Mechanisms*, 2nd ed.; Springer: New York, NY, 2003.

(14) James, C. A.; Weininger, D.; Delany, J. Daylight Theory Manual. http://www.daylight.com/dayhtml/doc/theory/theory.toc.html (accessed Apr 2009).

(15) OpenEye. *OEChem Toolkit, 1.4.2.* http://www.eyesopen.com (accessed 2009).

(16) ChemAxon. *Marvin Applets, 5.1.2.* http://www.chemaxon.com (accessed 2009).

(17) Li, J. J. *Named Reactions: A Collection of Detailed Reaction Mechanisms*, 3rd ed.; Springer: Berlin, Germany, 2006.

(18) Bruice, P. Y. *Organic Chemistry*, 4th ed.; Prentice-Hall: Upper Saddle River, NJ, 2004.

(19) Loudon, M. *Organic Chemistry*, 4th ed.; Oxford University Press: 2001.

(20) Smith, J. G. *Organic Chemistry*, 2nd ed.; McGraw-Hill: Boston, MA, 2006.

(21) Joolingen, W. R. v.; Jong, T. d.; Dimitrakopoulou, A. Issues in computer supported inquiry learning in science. *J. Comp. Assist. Learn.* **2007**, *23*, 111–119.

(22) Chen, J. H.; Linstead, E.; Swamidass, S. J.; Wang, D.; Baldi, P. ChemDB Update - Full-Text Search and Virtual Chemical Space. *Bioinformatics* **2007**, *23* (17), 2348–2351.

(23) Corey, E. J.; Long, A. K.; Rubenstein, S. D. Computer-Assisted Analysis in Organic Synthesis. *Science* **1985**, *228* (4698), 408–418.

(24) Hanessian, S. Man, machine and visual imagery in strategic synthesis planning: Computer-perceived precursors for drug candidates. *Curr. Opin. Drug Discovery Dev.* **2005**, *8* (6), 798–819.

(25) Allu, T. K.; Oprea, T. I. Rapid Evaluation of Synthetic and Molecular Complexity for in Silico Chemistry. *J. Chem. Inf. Model.* **2005**, *45* (5), 1237–1243.

(26) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47* (2), 342–353.

(27) Cramer, R. D.; Soltanshahi, F.; Jilek, R.; Campbell, B. AllChem: generating and searching 10^{20} synthetically accessible structures. *J. Comput.-Aided Mol. Des.* **2007**, *2007* (21), 341–350.

(28) Chen, L.; Gasteiger, J.; Rose, J. R. Automatic Extraction of Chemical Knowledge from Organic Reaction Data: Addition of Carbon-Hydrogen Bonds to Carbon-Carbon Double Bonds. *J. Org. Chem.* **1995**, *60* (24), 8002–8014.

(29) Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. *J. Chem. Inf. Model.* **2009**, *49* (3), 593–602.

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

CI900157K