JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

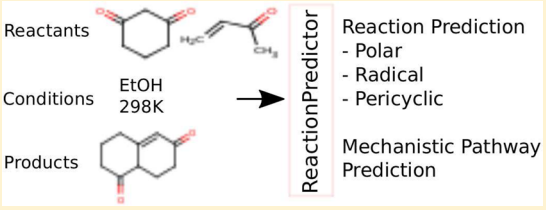
pubs.acs.org/jcim

ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning

Matthew A. Kayala and Pierre Baldi*

Institute for Genomics and Bioinformatics and Department of Computer Science, University of California, Irvine, CA, United States

ABSTRACT: Proposing reasonable mechanisms and predicting the course of chemical reactions is important to the practice of organic chemistry. Approaches to reaction prediction have historically used obfuscating representations and manually encoded patterns or rules. Here we present ReactionPredictor, a machine learning approach to reaction prediction that models elementary, mechanistic reactions as interactions between approximate molecular orbitals (MOs). A training data set of *productive* reactions known to occur at reasonable rates and yields and verified by inclusion in the literature or textbooks is derived from an existing rule-based system and expanded upon with manual curation from graduate level textbooks. Using this training data set of complex polar, hypervalent, radical, and pericyclic reactions, a two-stage machine learning prediction framework is trained and validated. In the first stage, filtering models trained at the level of individual MOs are used to reduce the space of possible reactions to consider. In the second stage, ranking models over the filtered space of possible reactions are used to order the reactions such that the productive reactions are the top ranked. The resulting model, ReactionPredictor, perfectly ranks polar reactions 78.1% of the time and recovers all productive reactions 95.7% of the time when allowing for small numbers of errors. Pericyclic and radical reactions are perfectly ranked 85.8% and 77.0% of the time, respectively, rising to >93% recovery for both reaction types with a small number of allowed errors. Decisions about which of the polar, pericyclic, or radical reaction type ranking models to use can be made with >99% accuracy. Finally, for multistep reaction pathways, we implement the first mechanistic pathway predictor using constrained tree-search to discover a set of reasonable mechanistic steps from given reactants to given products. Webserver implementations of both the single step and pathway versions of ReactionPredictor are available via the chemoinformatics portal http://cdb.ics.uci.edu/.



INTRODUCTION

Predicting the mechanistic outcome of chemical reactions given input reactants and reaction conditions is a fundamental scientific problem. Teaching how to make quick and accurate inferences about likely mechanisms or final products is a key goal of any undergraduate or graduate level organic chemistry curriculum. Furthermore, this reaction prediction ability is used in common tasks such as synthesis planning, product determination, or generation of plausible chemical explanations. A high-throughput reaction prediction system could be used for virtual space exploration,[1] assessing synthetic accessibility, or as a validation component of a retro-synthesis system. In previous work (Kayala et al.[2]), we presented a prototype machine learning approach to reaction prediction that performs well on a small data set of polar reactions and restricted atom types. Here, we present ReactionPredictor, a complete machine learning reaction prediction system that predicts experimentally observed polar, radical, and pericyclic reactions, handles expanded atom types, and incorporates the first ever published mechanistic pathway search for multistep reaction prediction.

The earliest attempts to make reactivity inferences using computational techniques revolved around the problem of retro-synthesis, beginning with the seminal work of Corey and Wipke.[3] Historical efforts in this field are reviewed in the work of Todd,[4] but there has also been a recent revival of interest in the problem with a commercial product[5] for medicinal chemists and in the context of assessing synthetic accessibility in virtual screening.[6−9] Reaction prediction is the inverse of retro-synthesis; rather than make a prediction about what reactions can construct a given product, reaction prediction systems aim to predict *productive* reactions, i.e. reactions known to occur at reasonable yields and rates given reactants and reaction conditions often verified through inclusion in the literature. The two problems are intimately related though, since a successful reaction prediction system can be used to validate a retro-synthetic proposal.

The state of the art in reaction prediction is reviewed in the work of Todd[4] and Kayala et al.;[2] however, we briefly outline the three major poles around which efforts have revolved. The first major pole of reaction prediction approaches is physical simulation of the chemical reaction transition energies, for example finding saddle points in an energy landscape.[10] As they are extremely low-level and often based on principled quantum mechanical approximations, these approaches can be very accurate. However, physical simulations require careful manual setup and extensive computation, and thus are not currently usable for high-throughput reaction prediction tasks. Furthermore, human chemists certainly do not perform in-depth quantum mechanical calculations when making judgments about reactivity. While it is not necessarily, the goal of a reaction prediction system to mimic the human decision

Received: June 29, 2012
Published: September 16, 2012

   dx.doi.org/10.1021/ci3003039 | *J. Chem. Inf. Model.* 2012, 52, 2526−2540

process, the fact that humans are capable of identifying chemical reactions without principled quantum mechanical calculations signifies that other methods for making reactivity decisions surely must exist.

The second major, and most extensively investigated, pole of reaction prediction approaches is rule-based expert systems. Efforts in this pole, from the pioneering CAMEO[11] and EROS[12] systems through the state-of-the-art,[13,14] have been reviewed extensively.[2] In general, rule-based expert systems have been found to suffer from several drawbacks: (1) They require manual encoding of chemical knowledge. (2) They do not scale well. After a certain complexity threshold has been reached, expansion of the underlying rule library can require updating a significant portion of existing rules to handle exceptions or resolve conflicts.[14] (3) They are not generalizable. On one hand, if a particular type of chemistry has not been encoded, then it will not be predicted by a rule-based system. On the other hand, patterns and rules written to be very abstract and general lead to large numbers of false positive misclassifications. These drawbacks are in addition to representing reactions at the level of overall transformations, obfuscating the underlying sequence of elementary concerted mechanisms. The state-of-the-art rule-based expert system Reaction Explorer[13,14] is the notable exception, with rules written to handle single mechanisms.

The third major pole of reaction prediction approaches is inductive machine learning. Initial work in this area began in the 1990s,[15] though there has been recent work on this for retro-synthesis.[5,6] Other recent work in machine learning involves either narrowly scoped regioselectivity classification problems[16] or classifying the "formability" of bonds to guide a retro-synthetic search.[7] Our previous work on the prototype machine learning framework[2] represented the first statistical approach to the general reaction prediction problem.

Our early prototype[2] revolves around two key ideas. The first key idea is that reactions should be represented at an elementary, mechanistic level. Overall transformation representations historically used by most rule-based systems obfuscate the underlying physicochemical properties of the reactions. Considering reactions at the level of discrete concerted electron movements, i.e., at the level of mechanisms, should lend itself to statistically learning generalizable underlying physicochemical patterns and trends that would be impossible to detect at the level of overall transformations. Thus, in previous and current work, we use the term *reaction* to refer exclusively to mechanistic, elementary reactions.

The second key idea presented in the early prototype is the simple model of elementary reactions as the interaction of a pair of idealized molecular orbitals (MOs), i.e., a reaction is the movement of electrons from a source MO to a sink MO. Using this formulation, it is possible to enumerate all possible mechanistic reactions over arbitrary molecules by simply enumerating all pairs of MOs. We defined the notion of productive reactions as those known to occur at reasonable rates and yields and verified by inclusion in the literature or textbooks. Using a data set of labeled productive reactions restricted to nonpericyclic, polar, two-electron reactions over a a small set of allowed atom types, we described a two-stage machine learning framework to build an accurate and generalizable reaction prediction system given input reactants and reaction conditions. In the first-stage, filtering models trained to predict source and sink MOs are used to reduce the space of reactions to consider. Then in the second stage, ranking models are used to provide an ordering of possible filtered reactions.

Here we describe ReactionPredictor, a complete machine learning reaction prediction approach that incorporates the key ideas of the early prototype with new reaction type models, a large new training data set, and significant improvements in machine learning techniques. Some of the key differences between the early prototype and ReactionPredictor are highlighted in Figure 1. The coverage of ReactionPredictor is

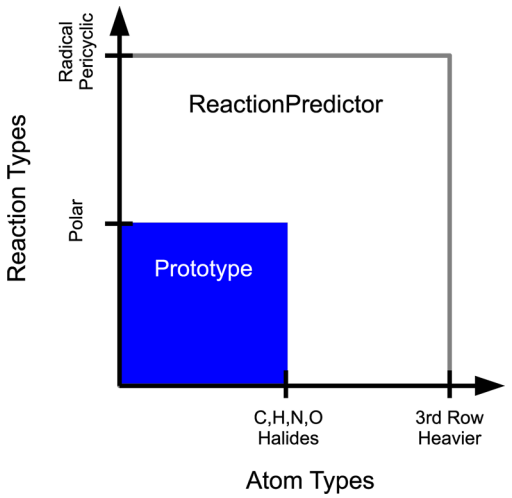


Figure 1. Illustration of how ReactionPredictor compares to the previous early prototype. The prototype made predictions over polar reactions and a restricted set of atom types. ReactionPredictor covers everything in the early prototype with the addition of radical mechanisms, pericyclic mechanisms, and expanded atom types.

a superset of the prototype, including many more atom types along with radical and pericyclic mechanisms. A high-level outline of the prediction workflow for ReactionPredictor is shown in Figure 2.

The remainder of the paper is organized as follows: First, we construct a training data set of productive reactions using an existing rule-based system and by manually curating complex reactions from graduate level organic chemistry texts. Second, we describe idealized MO reaction models to include hypervalent mechanisms, single electron mechanisms, and cyclic electron mechanisms, thus capturing higher molecular weight atoms, radical chemistry, and pericyclic chemistry respectively. Next, we describe the feature representations, feature selection methods, and details of the statistical learning techniques used for the filtering and ranking models. We show results of the trained single-step predictors. Then, we describe a reaction type classification model that can predict the most applicable polar, radical, or pericyclic reaction type models with high accuracy using individual ranking results and simple molecular features. Examples of multistep pathway prediction using the single step predictors and the reaction type model in a tree-search algorithm are shown. We conclude with a discussion of the successes, limitations, and future directions of the work.

MATERIALS

As a data-driven method, the machine learning approach to reaction prediction depends on training data of labeled reactions known to occur at reasonable rates and yields, which we call productive reactions. In an early prototype,[2] a small restricted chemistry data set, which we denote as EP,

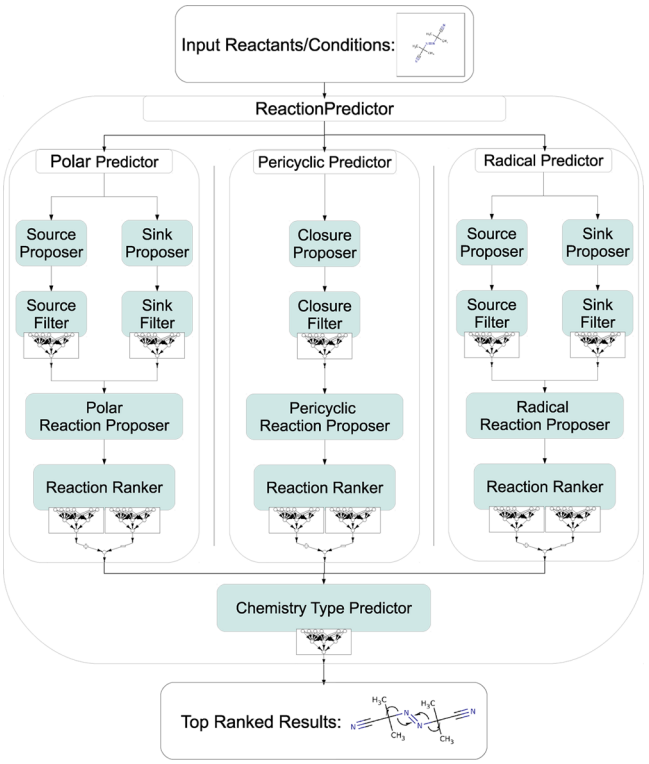


Figure 2. High-level ReactionPredictor workflow. Input reactants and conditions are pushed through three separate reaction type ranking models: polar, pericyclic, and radical. Within each reaction type ranking model, first single artificial neural networks are trained and used for initial filtering at the level of atoms or MOs. Then pairs of coupled artificial neural networks are trained and used to rank pairs of interacting MOs within each reaction type. The outputs from the individual ranking models are then combined with a final artificial neural network trained and used to make a decision between the three reaction types, resulting in a final ranked list of reactions from the chosen reaction type .

derived from the Reaction Explorer rule-based system[14] was used for training and validation. The reactions generated from Reaction Explorer come from a test suite of thousands of reactions manually curated from several undergraduate-level organic chemistry textbooks to ensure the correctness of the rule-based system. The Reaction Explorer-generated reactions are thus valid mechanistic reactions. This data set was limited to nonpericyclic, two-electron polar reactions over C, H, N, O, Li, and the halides, i.e., a large proportion of a standard undergraduate curriculum. In this work, the training data set is expanded in two ways: (1) by deriving reactions from Reaction Explorer with more complex reaction and atom types and (2) by manual curation of complex reactions from graduate level organic mechanisms textbooks.[17,18] The resulting data set encompasses a large proportion of a standard graduate level organic chemistry curriculum.

The EP data set includes 2989 labeled polar reactions. However, the Reaction Explorer system[14] contains transformation rules and a test set of labeled reactions, derived from the literature or textbooks, for radical and pericyclic reactions as well as more atom types, such as sulfur, phosphorus, and magnesium, than covered in our previous work. Including these expanded reaction and atom types yields 2194 new polar reactions, 46 radical reactions, and 247 pericyclic reactions.

As a rule-based system though, Reaction Explorer is inherently limited in the types and amount of chemistry that can be covered. Furthermore, the rule-based system does not scale well; adding new transformation rules to cover more chemistry often requires updating all existing rules to handle exceptions.[14] In contrast to a rule-based system, the machine learning approach to reaction prediction only needs examples of reactions rather than encoded expert knowledge. To take advantage of this fact and to circumvent the issues of scale in adding new chemistry to Reaction Explorer, more example reactions are added to the training set via manual curation from graduate level textbooks. Two common texts on mechanisms are covered: Grossman's *The Art of Writing Reasonable Organic Mechanisms*[17] and Carey and Sundberg's *Advanced Organic Chemistry A: Structure and Mechanisms.*[18] From these texts, we curate 368 polar, 51 radical, and 47 pericyclic reactions not previously found in the Reaction Explorer sets. The sizes of the different chemistry subsets of this combined data set, denoted RP, are shown in Table 1.

Table 1. Size of Data Sets over Sources and Reaction Types

source	polar	radical	pericyclic
EP[2]	2989	0	0
new Reaction Explorer[14]	2194	46	247
graduate texts[17,18]	368	51	47
RP	5551	97	294

The curated reactions represent novel chemistry not covered by the expert system. The vast majority of the curated reactions cannot be generated by the Reaction Explorer rules. Each of the 80 possible Reaction Explorer reagent models was used to attempt to predict the outcome of sequences of the manually curated reactions (as they are presented in their texts). From this experiment, only 10 polar, 4 radical, and 7 pericyclic curated reactions can be predicted by any Reaction Explorer reagent model. Thus, the expert knowledge to make predictions over the remaining reactions is not encoded in the rule-based system. This was validated qualitatively as well with manual inspection of the underlying rules and curated reactions. Furthermore, only 25 of the curated polar reactions and none of the curated radical and pericyclic reactions are predicted by the early prototype.

METHODS

There are two main components of the machine learning reaction prediction approach. The first component is a proposal model that can take arbitrary molecules and propose all possible reactions for a particular mechanism of action. It is not too important that this proposal model be *specific*, by proposing only reasonable reactions. However, it is important that the proposal model is *sensitive*, by having proposals encompass all reasonable reactions. Furthermore, it is important that the proposal model is *decomposable*, by having a the reaction representation that can be broken into smaller components to use for a coarse-grained filtering. The polar reaction proposal model described in our previous work is exactly such a system. Polar reactions are composed of a source and a sink MO. The proposal model is a simple algorithm to generate all possible source and sink MOs. All possible source and sink MO pairings encompass all reasonable polar reaction mechanisms and many nonspecific, unreasonable mechanisms. Filtering at the level of

the source or sink MOs independently allows coarse-grained filtering of the total reactions proposed.

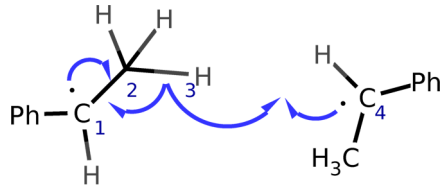
The specificity of the reaction prediction system is handled by the second major component: machine learning representations, filtering models, and ranking models. Filtering models are trained on the decomposable pieces of the reaction model, e.g., on source and sink MOs separately, to filter the number of reactions that are considered. Then ranking models are trained on the remaining reactions to build a final reaction prediction system that is both specific and sensitive.

Proposal Models. The polar reaction proposal model is described thoroughly in ref 2; however, the key details are briefly described below for completeness. A polar reaction can be modeled as movement of two electrons from a source MO to a sink MO. Lone pairs or bonds are potential source MOs, while bonds and empty orbitals are potential sink MOs. Lone pairs or π-bonds adjacent to π-bond source MOs can be chained to allow longer range resonance rearrangement, e.g., the lone pair of an enolate traveling through the C−C π-bond. Empty orbitals or bonds adjacent to general bond sink MOs can be chained in a similar manner, e.g. in E2 reactions. Source and sink MO combinations are not allowed to overlap, i.e., to make cycles. The decomposability of this model can be used in a two-stage prediction system. First, we train two different filtering models, one for source MOs and one for sink MOs. The source filter predicts the likelihood of each atom being the main atom of a source MO, and the sink filter predicts the similar likelihood of each atom being the main atom of a sink MO. All reactions with source MOs with atoms that score poorly in the source filter are disregarded. Similarly, reactions with sink MOs with atoms that score poorly in the sink filter are disregarded. Then ranking models are trained with the remaining reactions to make the final reaction prediction.

In this work, the polar reaction proposal model is extended to handle third-row and heavier atoms, such as sulfur or phosphorus, by incorporating a simple hyper-valency model. In this model, heavier atoms can use *d*-orbitals as empty orbitals or to participate in π-bonding. All other aspects of the polar proposal model remain the same as previously described.

To extend reaction prediction to more complex radical and pericyclic chemistry, two new separate decomposable reaction proposal models are described, one for radical reactions and one for pericyclic reactions. The radical reaction proposal model is similar to the polar model by being decomposable into source and sink MOs, disallowing cycles. However, all radical reactions involve the movement of only a single electron, and radical species can behave as single electron source or sink MOs. The machine learning component for this reaction type is similar to the polar; filtering models are trained on source and sink atoms for filtering, and ranking models are used for the final reaction prediction. The reaction proposal model is a little more specific than in the polar case. A few limiting rules are used, for example disallowing C−C σ bond homolysis unless in the presence of radicals, in a strained ring system, or under heat or photoexcitation conditions. A radical reaction example with corresponding source and sink MO notation is shown in Figure 3.

Modeling pericyclic reactions which involve cycles of electron movement is more complicated. One could write out any pericyclic reaction as a pair of source and sink MOs using the string orbital representation described in previous work. The end of a chained source will meet the end of a chained sink, making a cycle of electron movement. However, devising a



Source: $(C_4, n, \emptyset, \emptyset)$
Sink: $(H_3, \sigma^*, C_2, (C_1, n, \emptyset, \emptyset))$

Figure 3. Example of radical reaction and source-sink formulation. The source and sink MOs are written using the recursive notation of (main atom, orbital type, neighbor atom, chain orbital) used previously.[2]

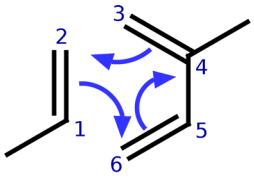
pericyclic proposal model based on pairs of sources and sinks presents several problems. The most difficult problem being that any pericyclic reaction can be described by several equivalent source and sink MO representations, each depending on the choice of source and direction of electron movement.

To bypass this issue of rotation and directional invariance, pericyclic reactions are proposed via a model based on 20 highly nonspecific, but sensitive, sets of SMARTS[19] patterns that canonicalize the orbital labeling scheme. The SMARTS patterns are constructed by writing the most general rule possible for each type of pericyclic reaction in our data set, e.g., a general SMARTS for [4 + 2]-cycloadditions, another general SMARTS for retro-ene reactions, etc. This nonspecific proposal model is vastly different from a rule-based expert system such as Reaction Explorer. The same underlying machinery of molecular graph matching is used. However, the pericyclic reagent models of Reaction Explorer include several hundreds of complex and specific transformation patterns to create a very specific reaction prediction system, while the pericyclic proposal model uses an order of magnitude less patterns for a nonspecific but sensitive proposer. Furthermore, additional proposal rules can easily be added to the pericyclic proposer without breaking any existing patterns. An example pericyclic reaction with all the possible source, sink formulations is shown in Figure 4.

For decomposability, we define the concept of a *closure atom*. A closure atom is an atom at which a σ-bond is made or broken during a pericyclic mechanism, i.e., a site of a molecular skeleton change during the reaction. Each of the pericyclic SMARTS patterns has the closure atoms noted by simply determining where the skeletal changes will occur in the pattern. The two-stage machine learning framework uses this concept to limit the number of reactions to consider: first filtering models are trained using closure atoms, and only reactions that have matching SMARTS patterns on the filtered closure atoms are considered in the final ranking models.

The machine learning details of the filtering and ranking models are very similar for all three reaction types . First, different filtering models are trained to independently predict atom-level labels (polar source, polar sink, radical source, radical sink, and pericyclic closure). Then, using the first-stage filtering predictions and the proposal model, different ranking models are trained, one each for the three reaction types. A final problem of deciding which of the polar, radical, or pericyclic ranking models will yield the most probable reaction prediction can be approached using to top ranked predictions from the separate ranking type models. Details for each of the machine

dx.doi.org/10.1021/ci3003039 | J. Chem. Inf. Model. 2012, 52, 2526−2540



Source	Sink
$(C_1,\pi,C_2,\emptyset)$	$(C_6,\pi^*,C_5,(C_4,\pi^*,C_3,(C_2,p,\emptyset,\emptyset)))$
$(C_5,\pi,C_6,\emptyset)$	$(C_4,\pi^*,C_3,(C_2,\pi^*,C_1,(C_6,p,\emptyset,\emptyset)))$
$(C_3,\pi,C_4,\emptyset)$	$(C_2,\pi^*,C_1,(C_6,\pi^*,C_5,(C_4,p,\emptyset,\emptyset)))$
$(C_2,\pi,C_1,\emptyset)$	$(C_3,\pi^*,C_4,(C_5,\pi^*,C_6,(C_1,p,\emptyset,\emptyset)))$
$(C_4,\pi,C_3,\emptyset)$	$(C_5,\pi^*,C_6,(C_1,\pi^*,C_2,(C_3,p,\emptyset,\emptyset)))$
$(C_6,\pi,C_5,\emptyset)$	$(C_1,\pi^*,C_2,(C_3,\pi^*,C_4,(C_5,p,\emptyset,\emptyset)))$

SMARTS Proposal Pattern

```
[C,N,O:1]=[C,N:2][C,N:3]=[C:4].[C,O,N:10]#,=[C,N,S:11]
```

Figure 4. Example of a pericyclic reaction, all possible source-sink formulation of the reaction, and SMARTS proposal pattern for [4 + 2]-cycloadditions proposing this reaction. Representing this Diels–Alder reaction as a pair of source and sink MOs is ambiguous. There are six different ways to represent the same reaction depending on choice of source and direction of electron movement. The SMARTS proposal pattern is sensitive enough to propose all reasonable [4 + 2]-cycloadditions but is nonspecific and proposes too many reactions. Carbons 1, 6, 2, and 3 are the closure atoms of the reaction.

learning framework components are described in the following sections.

Machine Learning Methods. For both the atom-level filtering and reaction-level ranking, we must choose an input feature representation, i.e., decide which physicochemical facts about the atoms, reactants, or reaction conditions to provide as input to our learning algorithms. Then, we must perform feature selection to reduce the size of input features to make learning tractable and avoid overfitting. Finally, we must train and evaluate our learning techniques.

Before describing the machine learning components, we first describe the available data: The overall reaction prediction system takes as input a set of reactant molecules and a description of reaction conditions, which we call a (r, c) tuple, and produces a final ranking of the most productive reactions. The intermediate filtering models make predictions for each set of topologically distinct atom and reaction conditions, which we call an (a, c) tuple. By topologically distinct atoms, we mean that only one atom in each atom symmetry class is considered. Reaction conditions are described with a vector of attributes to represent temperature, solvent, and photoexcitation conditions described in detail in the Feature Representations section below. However any mechanistic interaction with solvent is explicitly captured as an elementary reaction in the data set.

We label an (a, c) tuple as *polar source reactive* if it is the site of a source MO in the polar reaction subset of the RP data set and label it as *polar source nonreactive* otherwise. Labels for polar sink, radical source, radical sink, and pericyclic closure atoms are similarly constructed. Table 2 shows the number of labeled atom, reaction condition (a, c) tuples available from the different chemistry data sets and the number of input reactants, reaction conditions (r, c) tuples for the different data sets. Table 3 shows the number of productive reactions and nonproductive reactions, along with the number of nonproductive reactions to consider given perfect atom level predictions. Note that, even with perfect predictions, atom-level filtering is coarse-grained.

Feature Representation. In an early prototype, atoms and conditions were represented by a vector of physicochemical and

Table 2. Sizes of Input Data Sets[a]

	polar		radical		pericyclic
type	source	sink	source	sink	closure
reactive (a, c)	2302	3270	70	76	523
nonreactive (a, c)	57040	56072	866	860	1658
total (a, c)	59342		936		2181
total (r, c)	5176		90		169

[a]The data is broken down by reaction type and label. (a, c) represents atom and reaction condition tuples. (r, c) represents reactants and reaction conditions tuples.

Table 3. Sizes of Labeled Reaction Data Sets[a]

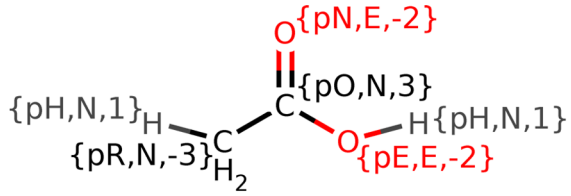
reactions	polar	radical	pericyclic
productive	5551	97	294
nonproductive	27851806	7750	3475
total	27857357	7847	3769
total perfect atom filtering	66909	274	541

[a]The productive reactions are those curated from Reaction Explorer or graduate texts. The "Total Perfect Atom Filtering" row says how many total reactions there are to consider in each reaction type if the atom-level predictions make perfect predictions. Atom-level filtering is coarse-grained.

topological features. The physicochemical properties included 14 real-valued features such as the molecular weight of the entire molecule, formal and partial[20] charges at and around the atom, a steric score that is the exponentially decaying sum of atom sizes in the atom's neighborhood, a span feature denoting distance from center of the 2D graph,[21] and the smallest ring of which the atom is a member. The topological features are similar to molecular fingerprints, commonly constructed by enumerating different paths and trees for the entire molecular graph, known to be effective in QSAR applications.[22,23] Instead of paths and trees over the entire molecule though, the atom-level topological feature vectors are constructed by enumerating paths and trees for particular atoms and their local neighborhoods. In addition to atom fingerprints over the standard molecular graph, previous work also included atom fingerprints over a pharmacophore graph[24] which can be seen as a graph with the same connectivity, but with a reduced alphabet (number of atom and bond types), intuitively allowing chemical motifs with similar reactivity but different atom types to share features. Combining all of these atom features on the EP data set gives 17 228 possible feature types. For the prototype, we performed feature selection through ad-hoc frequency and fingerprint depth limits, accepting only features occurring more that 25 times and a max path-depth of 3, a max tree-depth of 2, and a max path- or tree-depth through π-systems of 6. These frequency and depth cutoffs gave a final atom feature vector size of 1516 features.

For the prototype, feature vectors to represent full reactions are constructed by concatenating the atom features for the source and sink MOs and then including features to further describe the orbitals and net molecular changes induced by the reaction. The orbital level features are constructed for the sink and sink MOs separately and include information such as: the type of orbital (σ^*, π^*, n, p, π, σ), the type of chained orbital, and counts and charges over atoms in the orbitals. The net molecular change features are constructed by computing simple molecular fingerprints for the reactants and products including features such as counts over bond types, rings, aromatic atoms,

dx.doi.org/10.1021/ci3003039 | J. Chem. Inf. Model. 2012, 52, 2526–2540



Standard	Pharmacophore	Electron	Oxidation
H,σ,O	pH,σ,pE	N,σ,E	$(1),\sigma,(-2)$
H,σ,O,σ,C	pH,σ,pE,σ,pO	N,σ,E,σ,N	$(1),\sigma,(-2),\sigma,(3)$
$H,\sigma,O,\sigma,C,\pi,O$	$pH,\sigma,pE,\sigma,pO,\pi,pN$	$N,\sigma,E,\sigma,N,\pi,E$	$(1),\sigma,(-2),\sigma,(3),\pi,(-2)$
$H,\sigma,O,\sigma,C,\sigma,C$	$pH,\sigma,pE,\sigma,pO,\sigma,pR$	$N,\sigma,E,\sigma,N,\sigma,N$	$(1),\sigma,(-2),\sigma,(3),\sigma,(-3)$

Figure 5. Carboxylic acid with different labeling schemes for atom fingerprints. The standard molecular graph is shown. The label used by the pharmacophore, electron, and oxidation schemes, respectively, is shown to the side of each atom. Below the graph are examples of paths up to depth two emanating from the acidic proton for each of the fingerprint types.

rotatable bonds, and multiple separated formal charges. Then, we calculate the net change in the molecule induced by the reaction by simply subtracting the reactants fingerprint vector from the products fingerprint. This captures molecular motifs "created" or "destroyed" during the course of the reaction. In the prototype, using path-depth 3, tree-depth 4, and π-depth 4 cutoffs for the source and sink atom features, the above reaction feature construction procedure yields a total of 1677 reaction features.

Initial experiments on the RP data set using the feature representations and selection methods developed for the prototype performed poorly. The RP data set contains more diverse and complicated types of chemistry, making it more difficult to learn effective models. Furthermore with the expanded RP atom types, the number of fingerprint features become very large and very sparse with many features only appearing a handful of times in the entire data set. To overcome these issues, we create new feature classes and use a principled feature selection method.

Inspired by the pharmacophore graph representation, we induce further feature sharing between motifs with similar reactivity through atom fingerprints over several other reduced alphabet graphs. Specifically, we use an electron-labeled graph and an oxidation state-labeled graph. The electron-labeled graph labels each atom as either empty (empty orbitals), electron-filled (nonbonded electrons), or neutral (no empty orbitals or nonbonded electrons), and labels bonds as σ (single aliphatic bond) or π (multiple or aromatic bond). The oxidation state-labeled graph labels each atom with its oxidation state and bonds with the same σ/π scheme. The oxidation state is calculated using common rules, where the oxidation state for each atom is simply the standard valence of the atom minus the number of electrons assigned to the atom if all the bonds in the molecule were ionic and bond electrons are donated to the atom with the lower atomic electronegativity. An example molecular graph with the different labeling schemes and paths is shown in Figure 5.

To represent reaction conditions, we use the same features previously described with the addition of a binary photoexcitation feature. The features used previously are temperature (Kelvin), anion solvation potential, and cation solvation potential, where the solvation potentials are unitless numbers between 0 and 1 which can be used to capture common solvent trends such as polar/nonpolar or protic/aprotic. For example, water, a polar protic solvent, is represented by anion and cation solvation potentials of 1.0. Tetrahydrofuran, a mildly polar, aprotic solvent, is represented with an anion solvation potential of 0.4 and a cation solvation potential of 0.8. Hexane, a nonpolar solvent, is represented with cation and anion solvation potentials of 0.0. The reaction conditions are set for each reaction in our training data either by a mapping from the corresponding Reaction Explorer reagent model or from mapping the solvent type mentioned in the relevant textbook.

Reaction level feature representations are calculated in the same way as in the prototype, with the exception of pericyclic reaction features. Pericyclic reactions do not decompose into a single source and single sink MO representation and have varying numbers of closure atoms. Because of this, only the net molecular change features are used to represent pericyclic reactions.

Feature Selection. As the features depend on the occurrence of particular topological paths or trees in the input molecules, the number of possible features is different for each particular reaction type subset of the RP data set. There are 120 882 possible polar atom features, 5928 possible radical atom features, and 7342 possible pericyclic atom features. These atom level features vastly outnumber the labeled examples in our training data. Attempting to train on such high-dimensional data will result in severely overfit models.

To choose a set of informative and minimally redundant features, we use the Max-Relevance-Min-Redundancy (mRMR) feature selection technique.[25] This technique iteratively builds up a set of features that simultaneously maximizes a relevance measure $D_{\text{Rel}}(f, c)$, denoting similarity of feature f to the class feature c, and minimizing a redundancy measure $D_{\text{Red}}(f, S)$, denoting similarity of feature f to the set of already selected features S. A new feature f is iteratively added to the set of selected features S by choosing f from the nonselected features that optimizes:

$$\arg\max_{f}[D_{\text{Rel}}(f, c) - D_{\text{Red}}(f, S)]$$

The original mRMR implementation uses the mutual information between two features $\mathbb{I}(f_1, f_2)$ for both the relevance and similarity measures, i.e.,

$$D_{\text{Rel}}(f, c) = \mathbb{I}(f, c)$$

and

$$D_{\text{Red}}(f, S) = \frac{1}{|S|}\sum_{f_s \in S}\mathbb{I}(f, f_s)$$

In our application, the class distribution is highly skewed, with an order of magnitude less reactive atoms than nonreactive

dx.doi.org/10.1021/ci3003039 | J. Chem. Inf. Model. 2012, 52, 2526−2540

atoms in each labeling problem and a similarly skewed class distribution when considering productive and nonproductive reactions. Cieslak and Chawla[26] show that information based feature selection criteria, such as mutual information or equivalently a Kullback−Leibler divergence measure, are very sensitive to skew. In the same work, they suggest a different *f*-measure, the Hellinger distance of class-conditionals with respect to a feature, as a skew insensitive alternative. Let c be a binary class label and f be a binary feature, with c_+ as the positive class, e.g., the set of reactive filled atoms, c_- as the negative class, e.g., the set on nonreactive filled atoms, $c_{+,f=1}$ the set of data points in the positive class and with feature f present, and with $c_{-,f=1}$, $c_{+,f=0}$, and $c_{-,f=0}$ similarly defined. The Hellinger distance between c_+ and c_- with respect to f is

$$H_f(c_+, c_-) = \left[\left(\sqrt{\frac{|c_{+,f=1}|}{|c_+|}} - \sqrt{\frac{|c_{-,f=1}|}{|c_-|}}\right)^2 + \left(\sqrt{\frac{|c_{+,f=0}|}{|c_+|}} - \sqrt{\frac{|c_{-,f=0}|}{|c_-|}}\right)^2\right]^{1/2}$$

giving a number between 0 and $\sqrt{2}$. To ensure a similar scale with redundancy measures, we use a scaled Hellinger distance for the relevance measure,

$$D_{\mathrm{Rel}}(f, c) = \frac{1}{\sqrt{2}} H_f(c_+, c_-)$$

In exploratory experiments for the redundancy criteria, the Jaccard−Tanimoto similarity,

$$D_{\mathrm{Red}}(f, S) = \frac{1}{|S|} \sum_{f_s \in S} \frac{|f \cap f_s|}{|f \cup f_s|}$$

works well compared to other measures including mutual information and cosine similarity and, thus, is used in the rest of this work.

For the atom-level problems, we start by removing duplicate features, leaving 100 553 polar atom features, 5257 radical atom features, and 6765 pericyclic atom features. Twenty-six real-valued features in each set are simply included without going through the feature selection algorithm. Then in each atom level problem, we select a different number of discrete features using mRMR to make a final total of 1500 polar features, 100 radical features, and 200 pericyclic features. To avoid overfitting, the feature set sizes for each individual model are selected such that there are an order of magnitude less features than labeled data points. The mRMR feature selection is run treating each each feature as binary for the relevancy and redundancy criteria.

For the reaction level feature selection, we begin with the same duplicate removal. Because of the low-dimension of the pericyclic reaction feature space, all possible 51 nonduplicate pericyclic reaction features are used without any feature selection. For the polar and radical reaction data, we include all real-valued features and high-frequency features occurring in more than half the data without selection. This gives 79 polar reaction features and 55 radical reaction features. We then use mRMR feature selection treating discrete features as binary for each problem to construct a total of 1500 polar reaction and 100 radical reaction features.

Learning Techniques. Artificial neural network models are trained to build predictors for each of the five atom level classification problems, source and sink for radical and polar chemistry and closure for pericyclic chemistry. The networks use sigmoidal activation functions, with one hidden layer, and a single sigmoidal output node. Standard back-propagation with a cross-entropy cost function and L2 regularization is used for learning. Training is performed using stochastic gradient descent and an adaptive per-weight learning rate scheme.[27] Tenfold cross-validation is used for validation. The architecture parameters, number of hidden nodes, weight-decay parameter, and epochs for convergence are chosen for each classification problem by grid search and one round of internal 5-fold cross-validation on a single training set of the overall 10-fold cross-validation. The atom level classifiers must be highly sensitive. The decision threshold for each fold of cross-validation is fit as the average of the maximum decision thresholds giving a false negative rate (FNR) of 0 from internal 5-fold cross-validation.

Given filtered reaction sets, predicting the most productive reaction given reactants and conditions is a ranking problem. We use a pairwise approach with shared-weight artificial neural networks described in detail previously.[2] The overall network is given ordered pairs of reactions. The two networks are tied together with a single sigmoidal comparator output node with fixed ±1 weights.

An ensemble of 10 ranking machines are trained with a random sample of 10 pairs from each tuple of reactants and conditions. Training and architecture fitting of the individual pairwise networks are similar to the atom level networks. Back-propagation with cross-entropy error and L2 regularization is used for learning. Tenfold cross-validation is used for validation. The architecture parameters for each reaction problem are fit with internal 5-fold classification on a single training set of the overall cross-validation. Training is performed with stochastic gradient descent and the same adaptive per-weight learning rate scheme. Final predictions can be made either with a majority vote within the ensemble or by ranking the average internal network predictions. The second provides a small performance increase and is reported.

RESULTS

Atom Level Filtering. First, classifiers are trained on each of the atom level labels and decision thresholds are fit with careful internal cross-validation. These predictors are used to filter the reactions to consider in the ranking. The results of these predictors are measured by the true negative rate (TNR), the percent of negative class labels correctly discarded, and the false negative rate (FNR), the percent of positive class labels incorrectly discarded. We similarly look at the TNR and FNR

Table 4. Atom Filtering Results

type	problem	CV TNR % (SD)	CV FNR % (SD)	best TNR %
polar	source	84.9 (3.4)	1.9 (1.1)	90.1
	sink	68.8 (4.2)	1.3 (0.9)	79.6
	reaction	85.8 (5.3)	3.1 (0.8)	91.7
radical	source	95.7 (2.5)	1.4 (2.8)	97.2
	sink	71.7 (6.5)	3.1 (4.5)	76.8
	reaction	83.5 (6.6)	2.2 (3.7)	86.9
pericyclic	closure	79.3 (2.9)	1.1 (1.5)	79.3
	reaction	83.1 (1.5)	1.8 (2.6)	84.1

Table 5. Size of Filtered Reaction Data Sets[a]

reactions	polar	radical	pericyclic
productive	5551	97	294
nonproductive total	27851806	7750	3475
nonproductive perfect atom filtering	61358	177	247
nonproductive train atom filtering	2296883	1015	551

[a]To assess the reaction ranking methods, we use filtered reaction data sets obtained by applying atom level predictors trained on all available data. We show the number of total nonproductive reactions, the number of nonproductive reactions if we had perfect atom level filtering, and the number of productive reactions for comparison.

Table 6. Reaction Prediction Results[a]

type	*i*	NDCG@*i* (SD)	*n*	%*w*-*n* (SD)
polar	1	0.811 (0.022)	0	80.5 (2.1)
	2	0.879 (0.015)	1	92.1 (1.3)
	3	0.899 (0.012)	2	95.9 (0.9)
	4	0.907 (0.011)	3	97.5 (0.9)
	5	0.911 (0.011)	4	98.5 (0.6)
radical	1	0.787 (0.076)	0	78.7 (7.6)
	2	0.833 (0.068)	1	85.4 (6.3)
	3	0.856 (0.054)	2	91.0 (5.2)
	4	0.869 (0.061)	3	93.3 (6.5)
	5	0.882 (0.052)	4	96.6 (5.2)
pericyclic	1	0.881 (0.057)	0	88.1 (5.7)
	2	0.949 (0.021)	1	99.6 (1.2)
	3	0.957 (0.021)	2	100 (0)

[a]Two metrics for assessing the predictions are shown: normalized discounted cumulative gain at different list sizes *i* (NDCG@*i*) and percent within-*n* (%*w*-*n*). NDCG@*i* is a common information retrieval metric, and %*w*-*n* is the percent of all reactant/conditions sets where all productive reactions are in the top ranked list with at most *n* nonproductive reactions.

for the reaction filtering using the trained predictors. Complete atom-level prediction results for the polar, radical, and pericyclic problems are shown in Table 4. The final column shows the TNR rates for atom level predictors trained on all available data and a decision threshold chosen such that the FNR = 0.

In polar cross-validation experiments, an order of magnitude reduction in the number of nonproductive reactions to consider can be achieved with very few errors. One can see from the TNR rates for the subproblems that predicting sink labels is more difficult than predicting source labels. A potential reason is that hydrogen is often a reasonable sink site, i.e,. in acid−base reactions, but is rarely a source site, i.e., only in hydride transfer or reductions with hydride reagents. This makes sense, acid−base reactions are common in organic chemistry and our training set, while hydride reagents are much less prevalent. Even when only considering topologically distinct atoms, hydrogen represents a significant portion of the atoms in organic molecules and the polar sink predictor is forced to be much less specific. However, even with the low TNR rates for the sink predictor, combining both source and sink predictors provides excellent reaction filtering results.

The radical and pericyclic atom level predictions show a similar level of filtering ability. Like the polar atom level problems, the radical sink labels are more difficult to learn. However, the final reaction level filtering for all problems gives about an order of magnitude reduction in the number of reactions to consider with very few mistakes. After having shown excellent cross-validation results, we use the predictors trained on all the available data to construct filtered data sets to assess the reaction ranking methods. The sizes of these reduced data sets are given in Table 5. A combined overall validation experiment is described after independently assessing the ranking procedure.

Reaction Ranking. Even after filtering the reaction data sets using the trained atom level predictors, there is still a large imbalance between numbers of true positives and true negatives in the labeled data sets for each reaction type. In spite of this, the pairwise ranking techniques are able to learn very accurate ranking models in cross-validation experiments. The results are shown in Table 6.

Two different metrics are used to assess the ranking results: normalized discounted cumulative gain at various list sizes *i* (NDCG@*i*) and percent within-*n* (%*w*-*n*). The NDCG@*i* is a standard information retrieval metric[28] that quantifies the ranking in the top *i* results by summing a measure of usefulness (or gain) for a particular result that decays exponentially with position. The metric is normalized such that the best possible ranking for list size *i* has NDCG@*i* = 1. To give some intuition of what this means, the NDCG@1 row gives the fraction of reactant, conditions (*r*, *c*) tuples that rank a productive reaction in the top position. The metric is less intuitive at larger list sizes. A more intuitive metric is the %*w*-*n* measure. This is simply the fraction of (*r*, *c*) tuples that have at most *n* nonproductive reactions in the smallest ranked list containing

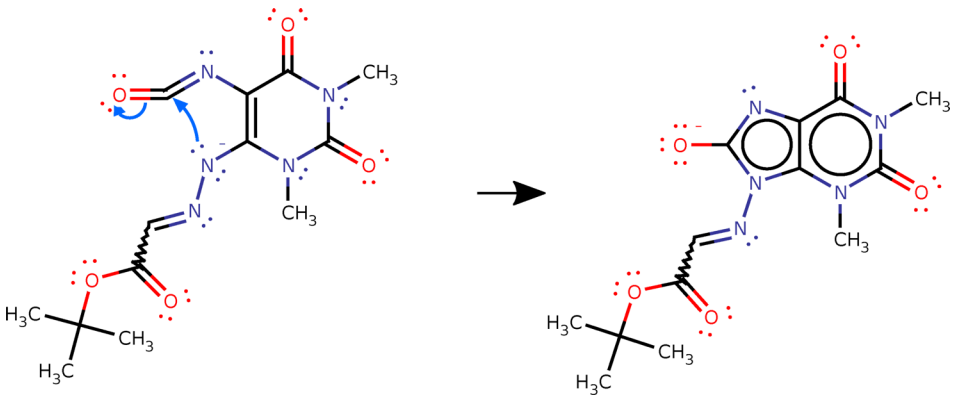


Figure 6. Perfectly predicted polar mechanistic step. The ring-closing addition mechanism to form the substituted purine is correctly ranked as the most productive polar reaction.

dx.doi.org/10.1021/ci3003039 | J. Chem. Inf. Model. 2012, 52, 2526−2540

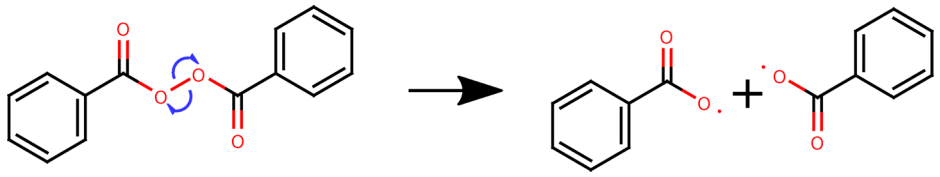


Figure 7. Perfectly predicted radical mechanistic step. The homolytic dissociation of the O−O σ bond in the benzoyl peroxide is correctly ranked as the most productive radical reaction.

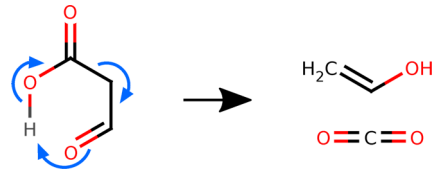


Figure 8. Perfectly predicted pericyclic mechanistic step. The retro-ene reaction that takes 3-oxopropanoic acid to ethenol and carbon dioxide is correctly ranked as the most productive pericyclic reaction.

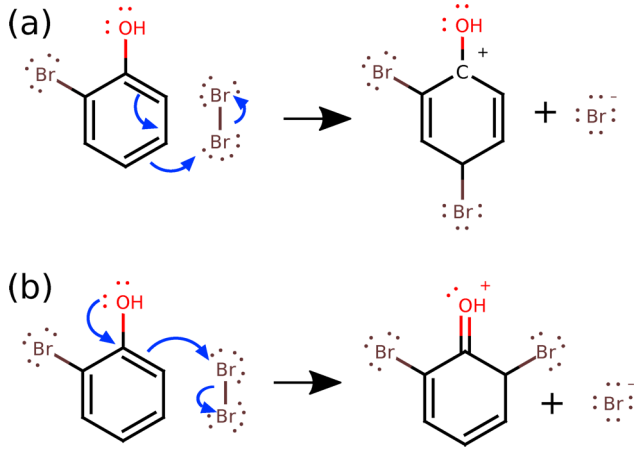


Figure 9. Polar reaction mechanism off-by-one. The substituted benzene can reasonably participate in electrophilic substitution reactions leading to new substituents at positions either ortho or para to the hydroxyl group. Mechanism a, labeled productive by the Reaction Explorer data, shows the para substitution reaction. The ranking model predicts the ortho substitution reaction slightly higher. However, this is not an unreasonable chemical prediction.

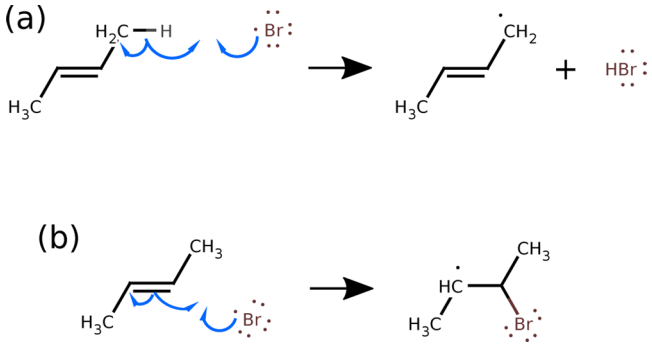


Figure 10. Radical reaction mechanism off-by-one. The proton abstraction mechanism shown in part a is the correct productive mechanism for this set of reactants. The ranking model ranks the addition reaction involving the π-bond as slightly more productive. Both reactions are reasonable.

all productive reactions. In other words, if n errors are accepted, %w-n denotes how often all productive reactions are recovered.

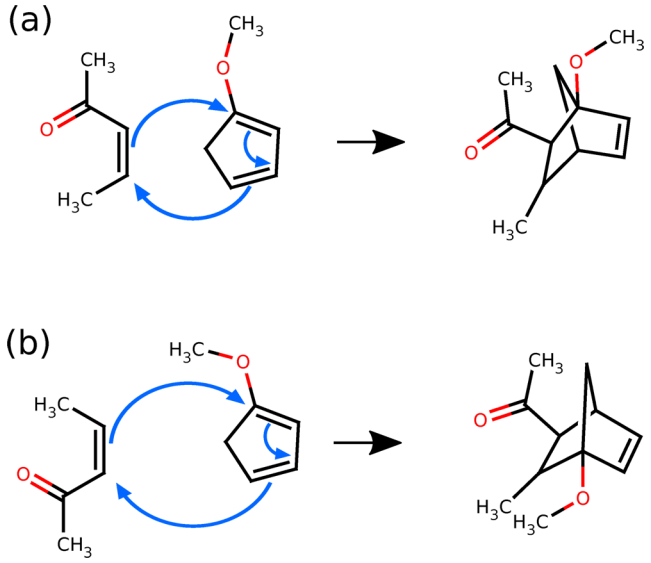


Figure 11. Pericyclic reaction mechanism off-by-one. The reaction predictor misranks the two regioisomeric reactions shown. The correct mechanism shown in part a is favored because of the correct addition of electrons from the diene at the electrophilic Michael receptor site on the dienophile. The mechanism in part b giving the opposite regioisomer is incorrectly ranked slightly more productive, likely due to lack of training data to properly learn this regioselectivity.

Table 7. Reaction Prediction Results from Combined Filtering and Ranking Validation[a]

type	i	NDCG@i (SD)	n	%w-n (SD)
polar	1	0.788 (0.021)	0	78.1 (2.2)
	2	0.853 (0.016)	1	89.3 (1.6)
	3	0.873 (0.014)	2	93.1 (1.1)
	4	0.880 (0.012)	3	94.7 (0.8)
	5	0.885 (0.013)	4	95.7 (0.9)
radical	1	0.77 (0.084)	0	77.0 (8.4)
	2	0.822 (0.079)	1	84.6 (7.8)
	3	0.839 (0.065)	2	89.0 (7.0)
	4	0.851 (0.067)	3	91.1 (7.1)
	5	0.859 (0.064)	4	93.3 (6.2)
pericyclic	1	0.858 (0.049)	0	85.8 (4.8)
	2	0.929 (0.030)	1	97.9 (3.1)
	3	0.936 (0.029)	2	98.1 (3.0)

[a]The NDCG@i and %w-n are shown. As expected, there is not much degradation from the results presented in Table 6.

The %w-0 measure is the fraction of (r, c) queries with all productive reactions ranked higher than any nonproductive. The %w-4 measure is the fraction of (r, c) queries with all productive reactions recovered with at most four errors.

The ranking model results for all three reaction types show similar behavior. As one can see in Table 6, the ranking is perfect a large proportion of the time (80.5% for polar, 78.7%

dx.doi.org/10.1021/ci3003039 | J. Chem. Inf. Model. 2012, 52, 2526−2540

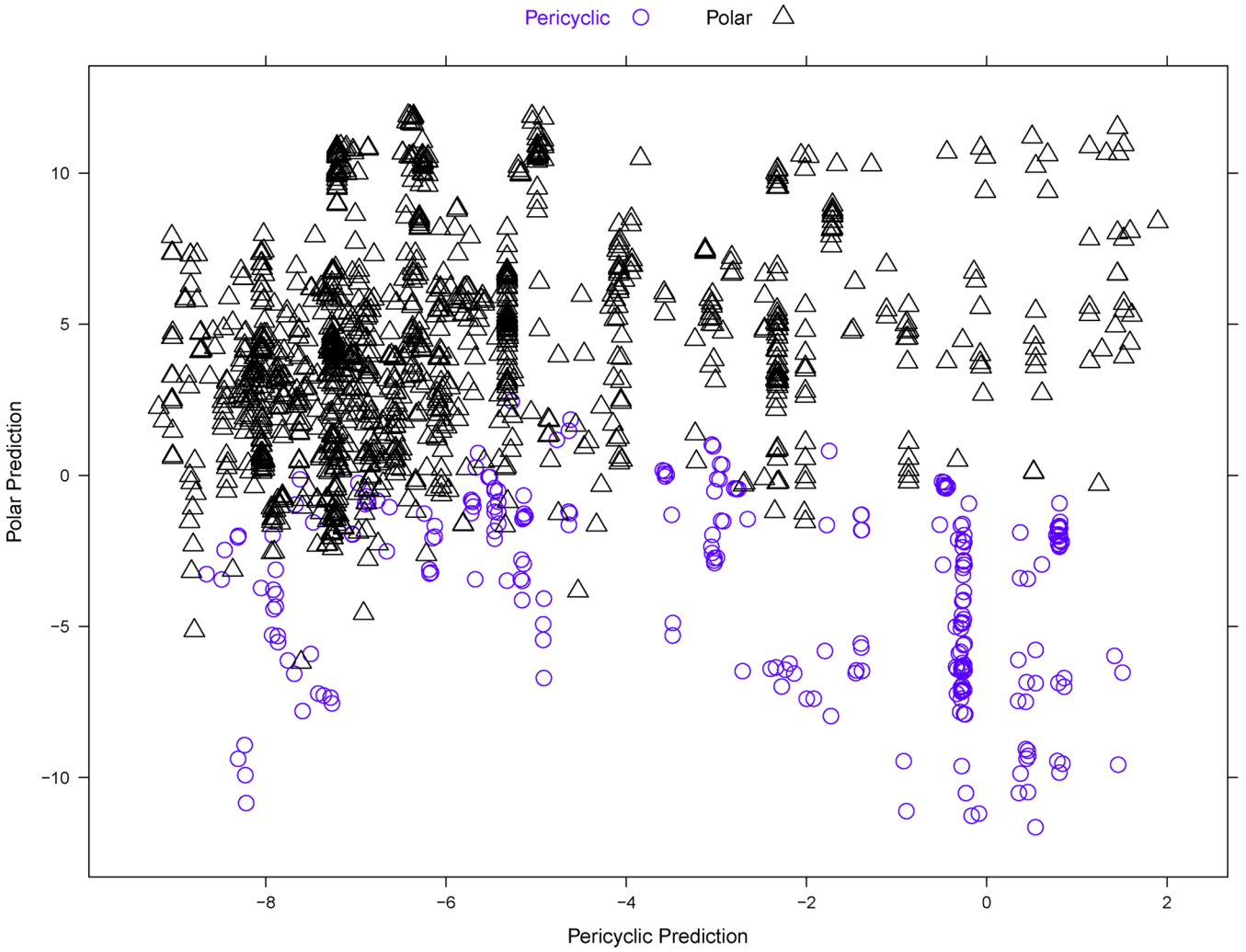


Figure 12. Maximum pericyclic versus maximum polar pseudoenergies. The internal neural network prediction values for the top-ranked pericyclic and polar reactions are plotted for all (r, c) tuples with productive reactions in the RP data set. The (r, c) tuples with a productive pericyclic reaction are represented with purple circles, and the (r, c) tuples with a productive polar reaction are represented with a black triangle. There is no overlap of productive reaction types. One can see that the two reaction class types are somewhat separated, but they are far from being perfectly calibrated.

Table 8. Reaction Type Prediction Results[a]

chemistry type	mol.nn	energy.nn	mol.energy.nn
polar accuracy % (SD)	95.8 (0.6)	98.8 (0.4)	99.7 (0.2)
radical accuracy % (SD)	94.4 (7.9)	94.4 (7.8)	97.8 (3.5)
pericyclic accuracy % (SD)	95.2 (4.8)	95.5 (3.0)	98.5 (2.6)
overall accuracy % (SD)	95.7 (0.8)	98.5 (0.5)	99.6 (0.2)

[a]Mean (SD) accuracies over CV folds are shown.

for radical, and 88.1% for pericyclic). There is a large jump in the percentage of complete recovery if only a single error is allowed (92.0% for polar, 85.4% for radical, and 99.6% for pericyclic). Furthermore, all of these ranking models show excellent recovery of the productive reactions if up to four errors are allowed (98.6% for polar, 96.6% for radical, and 100% for pericyclic.)

Perfectly Predicted Reactions. The vast majority of polar, radical, and pericyclic reactions are perfectly ranked. A few examples of perfectly ranked mechanisms are shown below. Figure 6 shows a perfectly ranked polar reaction, a purine forming ring-closing. Figure 7 shows a perfectly ranking radical reaction prediction, the homolytic dissociation of a peroxide bond. And finally, Figure 8 shows a perfectly ranked retro-ene pericyclic reaction.

Reaction Prediction Errors. Our previous results on a small restricted set of polar reactions showed similar behavior with 89.6%*w*-0 to 99.9%*w*-4. The slight decrease in the recovery measures for the expanded polar data set is not unexpected considering the vast expansion of the size and complexity of chemistry covered. Similar to the previous results, a large proportion of errors are off-by-one errors which are not unreasonable predictions to make. One can easily see from Table 6 that 59% of the polar errors, 28% of the radical errors, and 97% of the pericyclic errors are off-by-one errors. The %*w*-*n* values are not correlated with the number of atoms in the reactants (ρ = 0.04 for polar, ρ = 0.06 for radical, and ρ = 0.18 for pericyclic).

An example of such an off-by-one error is shown in Figure 9. The electrophilic aromatic substitution mechanism leading to a new bromide substituent para to the hydroxyl group shown in part a is the labeled productive reaction. However, the similar reaction leading to the new bromide ortho to the hydroxyl, shown in part b and predicted slightly higher, is not unreasonable given the existing substituents.

dx.doi.org/10.1021/ci3003039 | J. Chem. Inf. Model. 2012, 52, 2526−2540

Table 9. Overall Reaction Prediction Results with Reaction Type Prediction[a]

Type	*i*	NDCG@*i* (SD)	*n*	%*w*-*n* (SD)
Polar	1	0.785 (0.022)	0	77.9 (2.2)
	2	0.851 (0.017)	1	89.1 (1.6)
	3	0.870 (0.014)	2	92.8 (1.2)
	4	0.877 (0.013)	3	94.5 (0.9)
	5	0.882 (0.013)	4	95.5 (0.9)
Radical	1	0.759 (0.083)	0	75.8 (8.5)
	2	0.811 (0.080)	1	83.4 (7.8)
	3	0.827 (0.065)	2	87.9 (7.1)
	4	0.839 (0.068)	3	90.0 (7.0)
	5	0.844 (0.065)	4	92.1 (6.2)
Pericyclic	1	0.847 (0.051)	0	84.6 (4.7)
	2	0.915 (0.030)	1	96.5 (3.2)
	3	0.923 (0.029)	2	96.7 (3.0)
Overall	1	0.786 (0.020)	0	78.2 (2.1)
	2	0.853 (0.018)	1	89.1 (1.5)
	3	0.871 (0.015)	2	92.9 (1.2)
	4	0.878 (0.014)	3	94.6 (0.9)
	5	0.882 (0.013)	4	95.5 (0.9)

[a]The same metrics of NDCG@*i* and %*w*-*n* described in Tables 6 and 7 are used. Results for the entire predictor taking into account the decision about type of reaction model to use are shown. The type of ranking is chosen by the reaction type predictor. The results for individual predictors are upper-bounded by those in Table 7, as with perfect reaction type prediction, they will be equal. Overall the reactions are perfectly ranked no matter what reaction type 78.2% of the time. Impressively, 95.5% of the reactions are recovered when at most 4 errors are allowed.

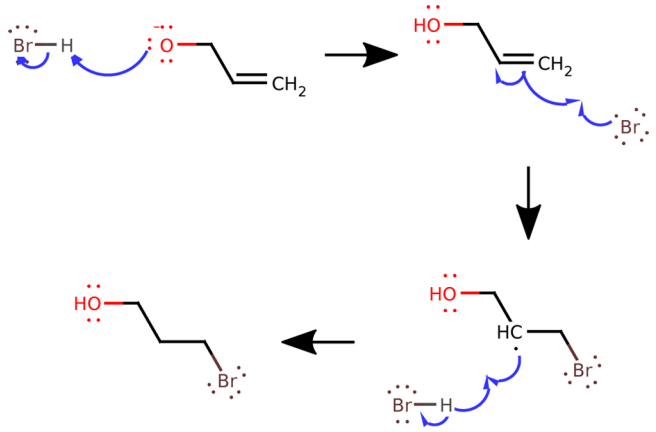


Figure 14. Heterogeneous pathway prediction with perfect ranking. The complete three-step mechanistic pathway, involving a polar step and two successive radical steps, is shown. All three of these predictions are the top ranked at each step. The reaction type predictor correctly decides which reaction model to use at each step. Note that the complete set of molecular species seen at any point in the reaction pathway is carried through and considered at each step, but for clarity, only the reacting molecules are shown.

Similar to the polar ranking model, the off-by-one errors for the radical ranking model are reasonable mechanisms to predict. For example in Figure 10, an addition reaction involving a radical bromine atom is ranked higher than the labeled productive proton abstraction reaction. However, there is no overwhelming chemical reason why the addition reaction should be disfavored, thus the reaction is reasonable to predict and rank highly.

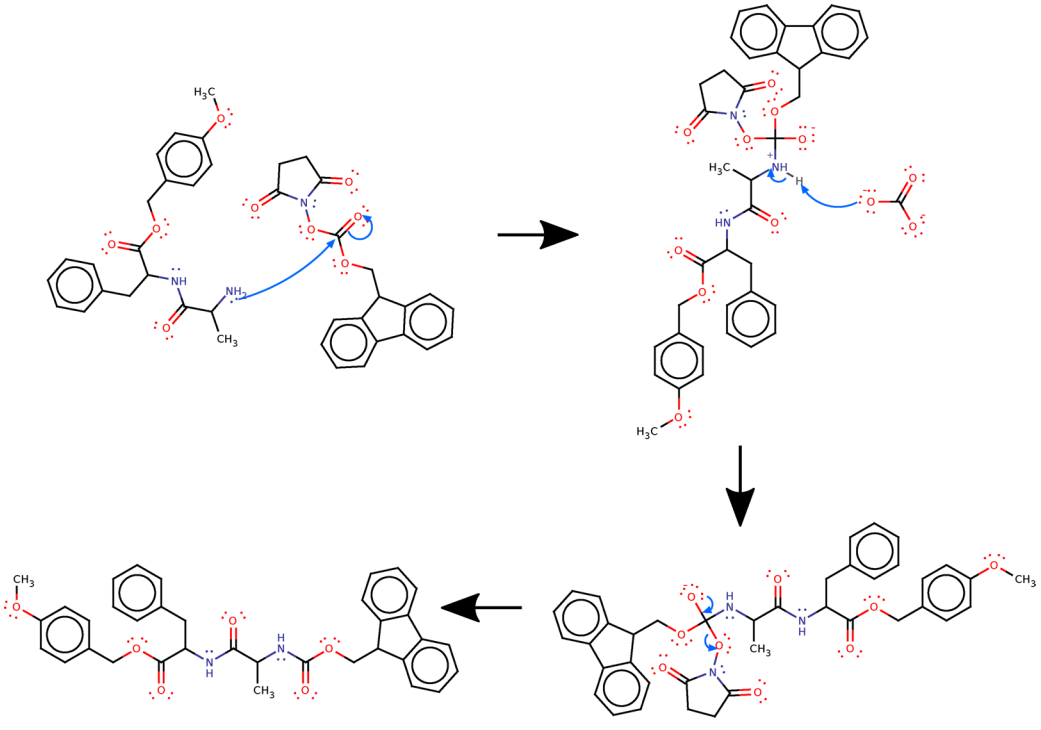


Figure 13. Polar pathway prediction with perfect ranking. The complete three-step mechanistic pathway introducing an FMOC protecting group is shown. All three of these predictions are the top ranked at each step. It is easy for the tree-search to discover this pathway. Note that the complete set of molecular species seen at any point in the reaction pathway is carried through and considered at each step, but for clarity, only the reacting molecules are shown.

dx.doi.org/10.1021/ci3003039 | J. Chem. Inf. Model. 2012, 52, 2526−2540

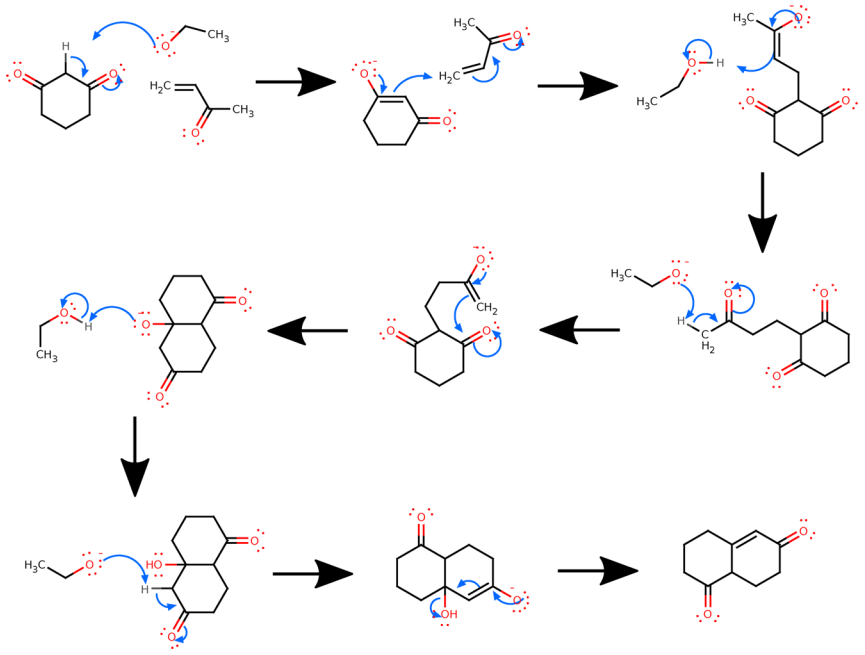


Figure 15. Polar pathway prediction. Th complete eight-step mechanistic pathway of a Robinson Annulation is shown. Not all predictions are the highest ranked at each step. However most are the highest ranked, and the lowest ranked is third. A tree-search with a max depth of eight and branching factor of 3 successfully discovers this pathway. Note that the complete set of molecular species seen at any point in the reaction pathway is carried through and considered at each step, but for clarity, only the reacting molecules are shown.

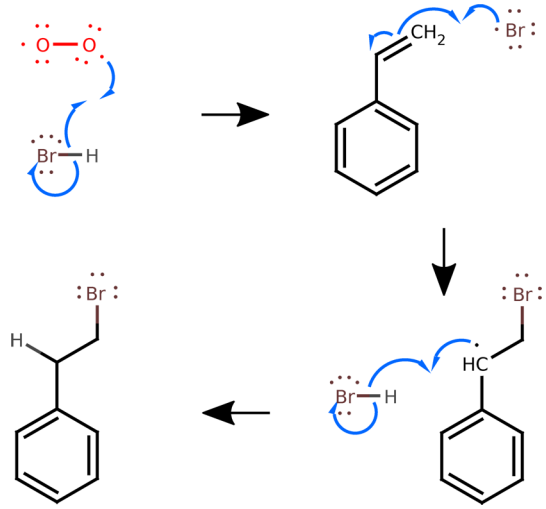


Figure 16. Radical pathway prediction. The radical mechanism leading to anti-Markonikov addition of bromine to an alkene is shown. The input hydrobromide, diatomic oxygen, and substituted benzene are given to the pathway predictor with standard conditions and light. The three correct mechanistic steps are predicted and returned. Again, all molecular species seen at any point in the reaction pathway are considered at each step, but only reacting molecules are drawn for clarity.

Finally, the off-by-one errors for the pericyclic ranking model are also reasonable. An example of a regioselectivity off-by-one error is shown in Figure 11. Here the mechanism shown in part b is incorrectly ranked as slightly more productive than the mechanism in part a. The mechanism in part a is in fact a more favorable reaction due to the electrophilic Michael acceptor site on the dienophile.

The misranking is likely due to the small amount of training data. Regioselectivity issues like the electrophilicity of Michael acceptors are learned by the polar reaction predictor, where there is much more data.

Combined Filtering and Ranking Validation. The filtering and ranking components show excellent performance when trained and assessed independently of each other. The independently trained components are used in the webserver and provide the best estimates of each components performance. The two problems, however, are of course related since the filtering step is meant to make the ranking procedure easier. The question then is how much the error in the filtering stage affects the overall ranking results?

Any false negative predictions at the filtering stage will reduce the ranking performance. However, we expect this reduction to be small for two reasons, First, the filtering errors are small, e.g., the polar filter is expected to make false negative predictions in only 3.1% of the cases. Second, the independent filtering CV experiment is a more stringent assessment than the corresponding stage in a careful overall validation experiment. In the independent filtering CV experiment, sets of individual molecules and conditions tuples (m, c) are used as the unit over which to cross-validate. In an overall cross-validation experiment at the level of reactants and conditions (r, c), we expect an overlap in the set of unique molecules across the training and testing splits, in the case of common solvents or reagents sharing different partners. To assess how much the filtering error is propagated, we use the same CV folds from the independent ranking evaluation in a careful overall validation.

This overall validation experiment works in the following manner. Each CV fold contains a training set of (r, c) tuples and a testing set of (r, c) tuples. For each training/testing setup, we perform the complete pipeline of filter training and then ranking training on only the training (r, c) tuples. Then we evaluate the ranking results on the test (r, c) tuples. The overall performance is shown in Table 7. As expected, the decrease in

dx.doi.org/10.1021/ci3003039 | J. Chem. Inf. Model. 2012, 52, 2526−2540

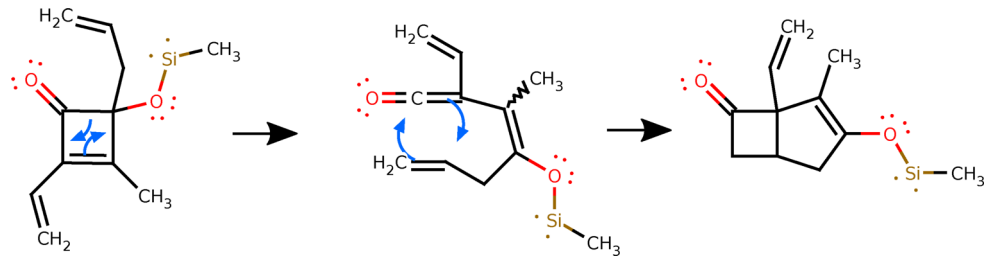


Figure 17. Pericyclic pathway prediction. A two-step pericyclic reaction mechanism is shown. The mechanism begins with an electrocyclic opening of the cyclobutene ring leading to a ketene. Then the ketene participates in a [2 + 2]-cycloaddition to give the final product. The two correct mechanistic steps are predicted and returned.

the %*w*-*n* measures is approximately the CV false negative rate for each reaction type.

Reaction Type Prediction. We have shown that independently the machine learning approach can rank the top polar, radical, and pericyclic reactions with high accuracy (>77%) and with remarkably high recovery rates (>93%) if a small number of errors are accepted. This then poses an interesting question of how these different chemical predictors can be combined. Is it possible to choose which of the three reaction type models to use given an arbitrary input of reactants and reaction conditions (r, c)?

The ranking model internal neural network predictions can be considered "pseudo-energies". Then for a given (r, c) tuple, the internal network value for the best-ranked polar reaction can be considered to be the maximum polar pseudoenergy for the (r, c) tuple, and similarly, the internal network value for the best-ranked radical (or pericyclic) reaction can be considered to be the maximum radical (or pericyclic) pseudoenergy for the (r, c) tuple. If the pseudoenergies are well-calibrated between the reaction types, then plotting the maximum pseudoenergies for different reaction types against each other should yield distinct clusters. For example in a plot of the maximum pericyclic pseudoenergy versus the maximum polar pseudoenergy, the pericyclic (r, c) tuples should have low polar prediction values and high pericyclic prediction values, i.e., they should be clustered in the lower right of the plot. Similarly, the polar (r, c) tuples should have high polar predictions and low pericyclic predictions, i.e., they should cluster in the upper left of the plot. A plot of the maximum pericyclic and polar pseudoenergies using the final reaction predictors for all (r, c) tuples in the RP data set with a polar or pericyclic productive reaction is shown in Figure 12. While the predictions are not perfectly calibrated, note that there is some separation between the reaction types . Plots of the polar versus radical predictions and the radical versus pericyclic predictions show similar behavior.

The small amount separation in two-dimensional space suggests that fitting a decision curve using statistical learning techniques and more input features could yield an accurate reaction type predictor. In a 10-fold cross-validation scheme, we used the same training and testing splits for the polar, radical, and pericyclic ranking experiments described above to build and validate such a predictor. Using the individual ranking models trained only on the training data in a cross-validation fold, we calculate the pseudoenergies of the top three best ranked reactions for each reaction type. If no reaction is predicted for a particular reaction type and input (r, c) tuple, then all pseudoenergies are set to be one interquartile range less than the minimum energy of the top pseudoenergies of that reaction type over all (r, c) tuples in the current fold's training set. Then using the pseudoenergies for the training set (r, c) tuples, we train several different artificial neural network models. The artificial neural networks all have a single hidden layer of sigmoidal nodes and a single normalized exponential output node which provides a prediction of most probable reaction type for the input (r, c) tuple. Grid search using internal cross-validation is used to choose the number of hidden nodes and the L2 regularization parameter.

We train artificial neural networks with three different inputs. First, we train neural networks using only simple molecular features of the input reactants. The features are counts of lone pairs, empty orbitals, negatively charged atoms, positively charged atoms, π bonds, σ bonds, and peroxide bonds. We call the resulting neural network mol.nn. Second, we train neural networks with the top three energies for each reaction type, which we call energy.nn. Finally, we train neural networks combining both of these feature types, which we call mol.energy.nn.

The cross-validation results of the different artificial neural networks are shown in Table 8. Using only simple molecular features in mol.nn, the correct reaction type is chosen 95.7% of the time. Using the top three energies of each individual ranking models in energy.nn raises the accuracy to 98.5%. And finally, using both the molecular features and the energies in mol.energy.nn further raises the accuracy to 99.6%. Note the accuracies are approximately balanced across the different reaction types for all three models.

In Table 7, we show the results combining the separate ranking models from the overall validation experiment and the reaction type predictor. Note that the results in Table 7 are an upper bound on the separate ranking model results shown in Table 9. If the reaction type prediction is perfectly accurate, the ranking results will be equal to Table 7. However, if the reaction type prediction is incorrect for a particular (r, c) tuple, then the entire ranking for that tuple will be incorrect, and thus the overall ranking results will be weaker than the independent ranking model. One can see that there is very little degradation in the individual ranking model performances when reaction type prediction is taken into account. Furthermore, the overall performance of the complete ranking model is quite good, returning perfect rankings 78.2% of the time and recovering all reactions with a small number of errors 95.5% of the time.

Multistep Reaction Pathway Prediction. The ability of the ranking models to recover the vast majority of all productive reactions with at most a handful of errors and the high accuracy of the reaction type predictor suggests that building a mechanistic pathway predictor will be successful. We present a simple implementation of such a predictor using the trained reaction ranking models and reaction type predictor as a heuristic to guide a depth-first search. The pathway predictor takes a set of reactants, a description of reaction conditions, and

dx.doi.org/10.1021/ci3003039 | J. Chem. Inf. Model. 2012, 52, 2526−2540

a target compound. All inter- and intramolecular reactions are ranked using the trained models. The top-ranked reaction is applied, making one or more additions to the set of reactants. If the resulting products contain the target, the search is terminated. Otherwise, the reaction proposal, ranking, and exploration procedure is performed recursively. To constrain the search space, the procedure terminates when proscribed depth or breadth limits are exceeded and only reactions leading to new molecules are considered at each step.

We describe several applications of this pathway predictor to real mechanism problems in the Grossman[17] textbook. For all of these, we used ranking models trained without any of the productive reactions used in the correct pathway, i.e., they reflect careful validation experiments.

First, we show a three-step pathway for a the introduction of the FMOC protecting group on an amine in Figure 13. Here, each of the three steps is the top ranked. It is trivial for the tree-search to discover this pathway.

Next, we show an example of a heterogeneous pathway. The three-step mechanistic pathway shown in Figure 14 begins with a polar proton transfer reaction, then undergoes two successive radical reactions, beginning with the addition of a bromine radical and ending with a radical proton abstraction. Each of the shown reactions is the highest ranked at each step by their respective separate ranking model. The reaction type predictor successfully decides which reaction and ranking model to use at each step.

Even without perfect ranking predictions, the pathway predictor can discover complex pathways. The power of this is seen in the prediction of the full eight-step Robinson Annulation reaction pathway shown in Figure 15. Here, not all of the shown mechanistic steps are the highest ranked for the set of reactants at each step. However, the lowest rank of the correct reaction at any step is three. Thus, the tree-search procedure is able to discover this pathway with a maximum branching factor of 3.

An example of a multistep radical pathway is an anti-Markonikov addition of bromine to an alkene shown in Figure 16. Again, this pathway contains reactions which are not the top ranked. However the lowest rank of any step in this pathway is two, making a tree-search feasible. Finally, an example of a correctly predicted two-step pericyclic pathway is shown in Figure 17. Here, again the reactions are not the top ranked, but are close enough to the top to make the search feasible.

This is the first mechanistic pathway prediction system ever published. Its potential uses include: use as a validation module for retrosynthetic proposals, as a tool for synthetic chemists to help understand results of reactions, as an annotation tool for existing, nonmechanistically defined reaction databases, and as a pedagogical tool for students to explore possible mechanistic routes.

A webserver implementation of single step predictors and the pathway predictor (with restricted depth and breadth parameters to not overload the server) is available through the chemoinformatics portal at http://cdb.ics.uci.edu/ under **ReactionPredictor**.

DISCUSSION

Being able to predict the mechanism and outcome of chemical reactions is a fundamental scientific problem. Ultimately, we hope that a reaction prediction system can recapitulate and even surpass the ability of trained human chemists. In previous work, we introduced a prototype data-driven mechanistic reaction predictor covering a small subset of polar organic chemistry. In the current work, we take a larger step toward the ultimate goal of an expert human-level reaction prediction system by expanding the machine learning method to a much broader range of chemistry and by implementing a pathway search algorithm that can accurately choose the type of applicable chemistry.

As a largely untapped area, there is of course room for improving reaction prediction approaches. One key result of our work is that treating reaction prediction as a machine learning problem is effective in the presence of enough data and with appropriate implementations. We curate reactions from two graduate-level organic chemistry textbooks, but the prediction accuracy and breadth of coverage is sure to improve with further expansion of the data sets from other sources. Overall, the machine learning performance described here is quite good, but further experiments and analyses could be performed to try to optimize each of the individual machine learning components: feature engineering, feature selection, classification, and ranking. In particular, with proper amounts of data, slight changes to the reaction models and feature representations, and improvements in selection methods, stereospecific reaction prediction should be an achievable goal.

We present the first pathway prediction algorithm. As a combinatorial search problem, there are many areas where the current implementation could be improved. Currently, the search only uses the trained reaction models as a heuristic for exploration. However, one could conceivably incorporate knowledge of the target molecule in the exploration decisions. For example, one could prune branches of the search space where the reacting molecules show significant divergence from the target molecule by molecular fingerprint similarity.

One can also envision turning the current work on reaction prediction toward the inverse problem of retro-synthesis. The reaction models to propose reactions could be inverted to propose retro-reactions. Similar machine learning representations, filtering models, and ranking models could then potentially be used to learn a retro-synthetic predictor that works at the level of mechanisms.

Finally, using the mechanistic pathway predictor to annotate reaction databases will provide several benefits. First, mechanistic annotations could provide an easy method to atom-map reactions. Most reactions contained in these databases are not atom-mapped, and this is a significant impediment for large-scale chemoinformatics approaches. Second, mechanistic annotations could provide a source of new training data for reaction prediction. Reaction databases, such Reaxys,[29] CAS,[30] or SPRESI,[31] or reactions culled from other sources such as patents that are not mechanistically annotated are currently not usable for training our machine learning predictors. However, the use of the pathway predictor as an annotation tool immediately suggests a semisupervised approach to reaction prediction: build a pathway predictor using the available training data, then mechanistically annotate a reaction database, and then retrain the pathway predictor using the new annotated reactions.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pfbaldi@ics.uci.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Andrew Brethorst for work on curating reactions. We thank Drs. David van Vranken, Steve Hannessian, and James Nowick for useful discussions. We thank OpenEye and ChemAxon for free academic software licenses. Finally, we acknowledge Jordan Hayes for systems support and webserver deployment assistance.

■ REFERENCES

(1) Fooshee, D.; Nguyen, T.; Nizkorodov, S.; Laskin, J.; Laskin, A.; Baldi, P. COBRA: A Computational Brewing Application for Predicting the Molecular Composition of Organic Aerosols. *Environ. Sci. Technol.* **2012**, *46*, 6048−6055.

(2) Kayala, M. A.; Azencott, C. A.; Chen, J.; Baldi, P. Learning to Predict Chemical Reactions. *J. Chem. Inf. Model.* **2011**, *51*, 2209−2222.

(3) Corey, E. J.; Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **1969**, *166*, 178−192.

(4) Todd, M. Computer-aided organic synthesis. *Chem. Soc. Rev.* **2005**, *34*, 247−266.

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

(6) Tanaka, A.; Okamoto, H.; Bersohn, M. Construction of functional group reactivity database under various reaction conditions automatically extracted from reaction database in a synthesis design system. *J. Chem. Inf. Model.* **2010**, *50*, 327−338.

(7) Pennerath, F.; Niel, G.; Vismara, P.; Jauffret, P.; Laurenço, C.; Napoli, A. Graph-mining algorithm for the evaluation of bond formability. *J. Chem. Inf. Model.* **2010**, *50*, 221−239.

(8) Yuan, Y.; Pei, J.; Lai, L. LigBuilder 2: A Practical de Novo Drug Design Approach. *J. Chem. Inf. Model.* **2011**, *51*, 1083−1091.

(9) Huang, Q.; Li, L.-L.; Yang, S.-Y. RASA: a rapid retrosynthesis-based scoring method for the assessment of synthetic accessibility of drug-like molecules. *J. Chem. Inf. Model.* **2011**, *51*, 2768−2777.

(10) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **2000**, *113*, 9901−9904.

(11) Jorgensen, W. L. CAMEO: a program from the logical prediction of the products of organic reactions. *Pure Appl. Chem.* **1990**, *62*, 1921−1932.

(12) 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. Model.* **2000**, *40*, 482−494.

(13) Chen, J. H.; Baldi, P. Synthesis explorer: a chemical reaction tutorial system for organic synthesis design and mechanism prediction. *J. Chem. Educ.* **2008**, *85*, 1699−1703.

(14) Chen, J. H.; Baldi, P. No electron left behind: a rule-based expert system to predict chemical reactions and reaction mechanisms. *J. Chem. Inf. Model.* **2009**, *49*, 2034−2043.

(15) Röse, P.; Gasteiger, J. Automated derivation of reaction rules for the EROS 6.0 system for reaction prediction. *Anal. Chim. Acta* **1990**, *235*, 163−168.

(16) Borghini, A.; Crotti, P.; Pietra, D.; Favero, L.; Bianucci, A. Chemical reactivity predictions: Use of data mining techniques for analyzing regioselective azidolysis of epoxides. *J. Comput. Chem.* **2010**, *31*, 2612−2619.

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

(18) Carey, F. A.; Sundberg, R. J. *Advanced Organic Chemistry, Part A: Structure and Mechanisms*, 5th ed.; Springer: New York, NY, 2007.

(19) James, C. A.; Weininger, D.; Delany, J. Daylight Theory Manual, 2004. http:// daylight.com/dayhtml/doc/theory/index.html (accessed May 2012).

(20) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity - A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(21) Sheridan, R. P.; Korzekwa, K. R.; Torres, R. A.; Walker, M. J. Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *J. Med. Chem.* **2007**, *50*, 3173−3184.

(22) Swamidass, S. J.; Chen, J.; Bruand, J.; Phung, P.; Ralaivola, L.; Baldi, P. Kernels for small molecules and the prediction ofmutagenicity, toxicity and anti-cancer activity. *Bioinformatics* **2005**, *21* (Suppl 1), i359−368.

(23) Azencott, C.-A.; Ksikes, A.; Swamidass, S. J.; Chen, J. H.; Ralaivola, L.; Baldi, P. One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical, and biological properties. *J. Chem. Inf. Model.* **2007**, *47*, 965−974.

(24) Hähnke, V.; Hofmann, B.; Grgat, T.; Proschak, E.; Steinhilber, D.; Schneider, G. PhAST: pharmacophore alignment search tool. *J. Comput. Chem.* **2009**, *30*, 761−771.

(25) Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: criteria of maxdependency, max-relevance, and min-redundancy. *IEEE Trans Patt Anal Mach Intell* **2005**, *27*, 185−205.

(26) Cieslak, D. A.; Chawla, N. V. Learning Decision Trees for Unbalanced Data. *Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I*, Berlin, Heidelberg, September 15−19, 2008; pp 241−256.

(27) Neuneier, R.; Zimmermann, H.-G. How to train neural networks. In *Neural Networks: Tricks of the Trade*; Orr, G. B., Müller, K.-R., Eds.; Springer-Verlag: Heidelberg, Germany, 1998; pp 373−423.

(28) Järvelin, K.; Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* **2002**, *20*, 422−446.

(29) Ridley, D. D. Searching for chemical reaction information. In *The Beilstein Online Database*; Heller, S. R., Ed.; ACS: Washington, DC, 1990; Vol. *436*, pp 88−112.

(30) Blake, J. E.; Dana, R. C. CASREACT: more than a million reactions. *J. Chem. Inf. Model.* **1990**, *30*, 394−399.

(31) Roth, D. L. SPRESIweb 2.1, a selective chemical synthesis and reaction database. *J. Chem. Inf. Model.* **2005**, *45*, 1470−1473.