# Graph-Mining Algorithm for the Evaluation of Bond Formability

Frédéric Pennerath,*,† Gilles Niel,*,‡ Philippe Vismara,§ Philippe Jauffret,¶ Claude Laurenço,‡ and Amedeo Napoli||

Supélec, Metz campus, 2 rue Édouard Belin, 57070 Metz, France, Institut Charles Gerhardt Montpellier, UMR 5253, ENSCM, 8 rue de l'École Normale, 34296 Montpellier, France, LIRMM (UMII/CNRS), 161 rue Ada, 34392 Montpellier, France, Montpellier, SupAgro, place Pierre Viala, 34060 Montpellier, France, Unité de Gestion de la Chimiothèque Nationale (CNRS), ENSCM, 8 rue de l'École Normale, 34296 Montpellier, France, and Orpailleur Team, Loria, BP239, 54506, Vandoeuvre-lès-Nancy, France

Received October 5, 2009

The formability of a bond in a target molecule is a bond property related to the problem of finding a reaction that synthesizes the target by forming the bond: the easier this problem, the higher the formability. Bond formability provides an interesting piece of information that might be used for selecting strategic bonds during a retrosynthesic analysis or for assessing synthetic accessibility in virtual screening. The article describes a graph-mining algorithm called GemsBond that evaluates formability of bonds by mining structural environments contained in several thousand molecular graphs of reaction products. When tested on reaction databases, GemsBond recognizes most formed bonds in reaction products and provides explanations consistent with knowledge in organic synthesis.

## INTRODUCTION

Faced with the synthesis of a complex target molecule, organic chemists often reason using a retrosynthetic approach to build a synthesis plan. Retrosynthesis consists in transforming recursively a target into simpler molecules until a set of available starting materials has been recognized. Corey first formalized this approach and introduced the key concepts of transform and retron, the former being defined as the exact reverse of a synthetic reaction and the latter being defined as the minimal element in a target structure which enables the application of a transform to generate a synthetic precursor.[1] Because of the very large diversity of molecular structures targeted in both academic and industrial environments, only general strategies were proposed by Corey to analyze the target.[2] They are based on the recognition of structural patterns, such as topological, stereochemical, and functional group-based patterns, and also on the chemist's knowledge about synthesis methods, a knowledge necessary to identify potential transforms. However selecting the first bond or the first bond subset to be disconnected remains a complex and crucial task as it will govern all subsequent steps during the retrosynthesis.[3] Perceiving the right bond as strategic among other bonds depends on the chemist domain knowledge or on the availability of robust evaluation criterions.

Following Vleduts's initial advices[4] and the pioneering work of Corey about LHASA,[5] several synthesis design programs were developed for assisting chemists in conducting retrosynthesis. Our purpose is not to discuss herein all the related programs and the reader should consult some recent relevant reviews.[6−9] Although these synthesis design systems seem intellectually very attractive, most of them are no longer developed for various reasons and did not really emerge either in academic or in industrial world (SECS software was tested in the 1980s at Merck;[10] WODCA is always distributed by Molecular Networks (http://www.molecular-networks.com) and SYNGEN is available on the Web at http://syngen2.chem.brandeis.edu/syngen.html.). One probable reason is that these systems all rely on at least one knowledge base -including either transforms or generalized reactions, whose update is not yet automatable. Since no synthesis design system has been universally acknowledged to correctly define a retrosynthesis analysis, most organic chemists use available reaction databases as a primary source of chemical knowledge.

But the fundamental question about which bond should be first disconnected remains of utmost interest and there is a real need to precisely measure either the *breakability* of such a bond, that is, its ability to be disconnected in a given structural environment from a retrosynthesis point of view or equivalently its *formability*, that is, its ability to be formed from a synthesis point of view. All along this article, we will refer to this concept using the term of "formability". If we refer now to topological, stereochemical, and functional disconnective strategies formalized by Corey, the functional point of view is probably the most difficult one to be formalized. Indeed concerning the topological aspects, Bertz[11,12] already proposed a rigorous mathematical approach of Corey's strategic disconnections about cyclic and polycyclic systems, and the stereochemical point of view has efficiently been approached within CHIRON program.[13] However, estimation of the functionality influence on bond formability is more difficult for several reasons: (i) There is a great diversity of functional groups having various

* To whom correspondence should be addressed. E-mail: frederic.pennerath@supelec.fr (F.P.); gilles.niel@enscm.fr (G.N.).
† Supélec and Loria.
‡ Institut Charles Gerhardt.
§ LIRMM and SupAgro.
¶ CNRS UGCN.
|| Loria.

10.1021/ci9003909 
Published on Web 01/29/2010

electronic effects and even within a family of functional groups, for example, electron-withdrawing groups, the reactivity of a given bond may be altered by some structural changes on the functional group. (ii) This diversity strongly influences the bond formability since two or more functional groups may interact on the same bond. This causes a combinatorial explosion that must be controlled. A possible answer to this problem consists in building knowledge bases from the most used or famous synthesis methods. Unfortunately all these approaches cannot describe the whole diversity of the chemical reaction space. Therefore we tried to extract a broadest knowledge from available reaction databases. This work follows a precedent study on the perception of strategic bonds that was carried on a limited number of reaction examples.[14,15]

Thanks to the development of a graph data-mining algorithm called GemsBond,[16] we are now able to perceive bond formability from a greater number of reactions. This algorithm presented in this paper, mines a given set of reaction examples every time a new target molecule is analyzed. The structural environments of each bond of the target are compared to similar environments in the products of the reaction samples, in order to determine statistically whether the bond can be easily formed or not. GemsBond follows to the "transductive" learning paradigm, as defined by Vapnick's book[17] on pages 339−371: compared to a two-step learning method that first learns a model, for example, a set of classification rules, from training examples (i.e., induction step), and then applies this model to classify all test examples (i.e., prediction step), transductive learning considers test examples separately and for each of them, infers from training examples a model specifically designed for the classification of the test example. The extra computation cost of the transductive approach is justified here as determining a unique general model for bond formability is too complex (as it involves too many factors) to be reduced efficiently to a reasonably small set of rules.

This paper details the design of GemsBond and the various experimental results obtained when using this algorithm to evaluate bond formability. To this end, motivations and requirements of our algorithm are presented in the first part. The problem is then formalized in terms of discovery of formable bonds and of ranking these bonds according to their formability. In the following, we describe GemsBond as a heuristic search algorithm based on the notions of environment, confidence, and frequency of a formable bond. Experiment and results are then explained and discussed according to various parameters.

## METHOD

Description of GemsBond is developed as follows: First, we introduce the motivations and requirements underlying the problem of discovering formable bonds. Then we provide a formal definition of the problem and a first analysis, followed by details on GemsBond algorithm. Finally we discuss various ways to process the output of GemsBond for identifying formable bonds.

**Motivations and Requirements.** As already mentioned in the introduction, the formability of bonds is an interesting supplementary piece of information in the framework of retrosynthesis. Bond formability is also useful in the context of virtual screening to assess synthetic accessibility of molecules. Indeed virtual screening usually processes very large combinatorial libraries of molecules whereas a large amount of these virtual molecules are hard or even impossible to synthesize. Since a molecule whose all bonds have low formability is unlikely to be synthesized easily, computing the formability of bonds in these virtual molecules may help virtual screening to discard molecules with poor synthetic accessibility and to focus screening on remaining feasible molecules.

The formability of a bond in a target molecule is thus an interesting bond property related to the problem of finding a reaction that synthesizes the target by forming the bond: the easier this problem, the higher the formability. The notion of formable bonds is not a binary but a relative notion; many bonds of a target molecule can be formed by reactions, but depending on the considered bond, the problem of searching such a reaction might range from a trivial to a very hard problem. Deciding whether a given bond is formable consists mostly in deciding whether the bond has a structural environment that is favorable to its formation. However answering this question and formalizing the solution into an algorithm are not straightforward tasks. One obvious solution might consist in specifying manually every favorable environment. A bond is then classified as formable whenever it lies in such an environment. Nevertheless environments are not all equally favorable: some of them are more favorable than others and some of them are even unfavorable. A solution may consist in asking an expert to distribute environments over a "scale of formability" ranging from highly unfavorable to highly favorable. One way to decide whether a given bond is formable is then to determine the largest identified environment of this bond and to conclude according to the favorable or unfavorable status of this largest environment. Indeed, this solution raises serious problems: first the task of distributing these environments on a scale is subjective. Second it is very difficult, not to say impossible, for the expert to enumerate and label precisely every relevant environment.

For these reasons and contrasting with prior formal or expert systems, we have developed a machine learning method that evaluates bond formability from a large set of examples of environments without human supervision. The fundamental principle of our algorithm is to compare environments of a given bond with environments of bonds known to be formed by some reaction. This approach is supported by the existence of numerous examples of formed bonds available in databases of chemical reactions. Since those databases mostly specify molecules by their molecular graphs, the method focuses on information contained in the structural environment of this bond, that is, on subgraphs surrounding this bond. Our approach is related to recent development of graph-mining methods,[18−21] that is, data-mining methods dealing with data modeled by graphs, and to some extents to previous graph-learning approaches, such as the Subdue algorithm.[22] Before developing the details of our graph-mining algorithm, we first consider a list of requirements that such an algorithm should ideally meet:

*Granularity.* The algorithm must take into account the whole complexity of molecular graphs without prior reduction of the topological information since the formable character of a bond is sensitive to any change in its

environment; changing the type of only one atom or the multiplicity of only one bond in the proximity of the considered bond might drastically change the formability of the bond. This requirement differs from many problems in QSAR applications, where molecular graphs are first reduced to vectors of numerical descriptors that are then used as inputs to some numerical machine learning algorithm. For instance, the reduction of the topological information may consist in computing histograms of atom sequences[23] or molecular fragments of limited size[24] in molecular graphs.

*Autonomy.* The algorithm must autonomously learn from raw data in existing reaction databases, without requiring the intervention of an expert. Ideally the management operations must be minimized to addition or deletion of reactions to/from collection of selected examples.

*Robustness and Measurability.* Existing reaction databases contain some erroneous or incomplete entries. The algorithm should integrate statistics measures for managing uncertainty, for example, contradicting examples or unbalanced distribution of examples. This requirement discards algorithms that do not support contradiction between examples, like for instance, algorithms based on version spaces[25] using a graph-learning setting[26] or the early systems of inductive logic programming,[27] even if later some version space based algorithms have been improved to accept a tunable level of contradiction, like the MOLFEA system[28] that mines linear fragments of molecules. Moreover being formable for a bond is not a binary but a relative characteristic: a bond is more formable than others. To maximize accuracy and to rank bonds according to their formability, the algorithm must provide a multivalued score for each bond to be formable. Ideally, this score is a confidence index that is a bounded real number ranging continuously from 0 (the bond has no chance to be formable) to 1 (the bond has all chances to be formable).

*Explanatoriness and Readability.* The algorithm should justify the predicted level of formability of a bond with some explanations. These explanations must be interpretable by chemists and contribute to improve knowledge about formable bonds.

*Efficiency and Scalability.* The algorithm must be fast and scalable as the problem requires the search of a large number of bond environments in a large number of molecule examples. The method must be able to mine several thousands of graphs containing hundred thousand bonds. For this reason, the algorithm should have a linear complexity in the number of examples, contrasting with bottom-up graph-based learning methods[14,15,29] whose complexity is at least quadratic in the number of examples.

**Formalization of the Problem.** *Definition of the Problem.* The problem of discovering formable bonds is formalized as follows: A molecular graph $G$ is specified by a set $\mathbf{V}(G)$ of vertices representing atoms and a set $\mathbf{E}(G)$ of edges representing bonds. Every vertex is labeled at least by the chemical element and the electrical charge of the represented atom, and every edge is labeled at least by the type (single, double, triple, or aromatic) of the represented bond. The inputs of the problem of discovering formable bonds are then an *input molecule* represented by its molecular graph $G = (\mathbf{V}(G), \mathbf{E}(G))$, an *input bond* $b \in \mathbf{E}(G)$ of the input molecule, and a set $\mathcal{E}$ of *examples*, where an example is the molecular graph $g \in \mathcal{E}$ of a molecule associated to the subset of formable bonds of $g$, denoted $F(g)$.

The *discovery of formable bonds* consists in deciding whether the hypothesis "the input bond $b$ is formable" is true given the position of $b$ in $G$ and given the set $\mathcal{E}$ of examples. Closely related to this problem, the *ranking of formable bonds* consists in deciding which one of two input bonds $b_1$ and $b_2$ in the input molecule $G$ is the easiest to form. In other words, $b_1$ is to be more formable than $b_2$ if there are more known reactions that could synthesize $G$ by forming $b_1$ than $b_2$. The ranking of formable bonds amounts to associate to every input bond a real number as a *score* so that the higher the score, the more formable the bond. The ranking then consists in sorting the bonds in decreasing order of score. As already mentioned, the notion of formable bond is not binary so that the problem of ranking formable bonds is more relevant than the problem of discovering formable bonds. The latter problem is only introduced here as a convenient mean for evaluating the accuracy of our algorithm: since the discovery of formable bonds is a binary classification problem, the associated standard performance measures (ROC curve, F-measure) can be used to assess the quality of the algorithm.

Independently of those two problems, the nature of considered examples may introduce two different problem settings:

*Manual Labeling.* An expert of organic synthesis has manually identified the formable bonds in selected examples of molecular graphs. Even if such a process relies on a subjective perception, it might also provide a very accurate data set while the number of examples does not exceed a few hundreds. However for larger set of examples (a few thousands), the labeling might be a tedious and error-prone task.

*Automated Labeling.* The examples are composed of the main products of chemical reactions extracted from available reaction databases. The examples of formable bonds are then the bonds formed in these extracted reactions and only them. This requires almost no human effort and no limitation in the number of reactions. However it also introduces a negative bias: as a target molecule might potentially be synthesized by more than one reaction, a bond that is not formed by the considered reaction and thus labeled as not formable might actually be formable. The consequence is an underestimate of the number of formable bonds in the examples.

Both settings have their own advantages and drawbacks. However, with regard to the requirements of autonomy and scalability, we focus on the case of automated labeling, assuming examples are directly extracted from available reaction databases.

Even if GemsBond is presented herein as an algorithm to evaluate bond formability, it is clear that the method addresses a more general classification problem introduced as the *problem of vertex (or edge) classification based on vertex (or edge) environment*.[16] In that sense, GemsBond may be used for predicting other bond or atom characteristics from examples. Because the ranking of bonds with respect to their formability can be used to decide which bonds are formable, as shown further in this section, the problem of computing a formability score for bonds comes first.

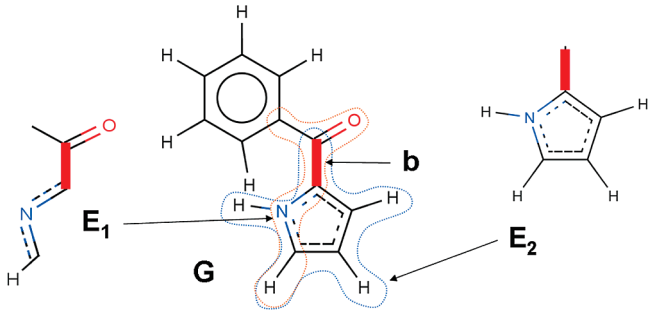



**Figure 1.** Input graph $G$ and two environments $E_1$ and $E_2$ of an input bond $b$ (thick bond).

**Analysis of the Problem of Ranking Formable Bonds.** As already stated, the problem of ranking formable bonds amounts to determine a score called *confidence index* and denoted by $\text{conf}_G(b)$ measuring the formability of an input bond $b$. This confidence is computed by searching in the examples the occurrences of structural environments that bond $b$ has in the input molecule. Formally, an *environment* $E$ of $b$ in an input molecular graph $G$ is a pair $E = (b, g)$ of the bond $b$ and any **connected** subgraph $g$ of $G$ containing $b$. Such a graph $g$ thus satisfies three conditions: (i) $b \in \mathbf{E}(g) \subseteq \mathbf{E}(G)$, (ii) $\mathbf{V}(g) \subseteq \mathbf{V}(G)$, and (iii) for every pair $(v_1, v_2)$ of vertices of $g$, there exists a path in $g$ (i.e., a sequence of incident edges of $g$) linking $v_1$ to $v_2$. In the following and for the sake of simplicity, the term *environment* mainly denotes subgraph $g$ even if it must always be understood as a pair $E = (b, g)$ or alternatively, as subgraph $g$ where bond $b$ is specifically annotated to be distinguished from other bonds. Figure 1 considers two different environments of the same input bond. The hypothesis of connectedness comes first from a combinatory requirement so that only a small proportion, but still a large number, of subgraphs of $G$ have to be considered. However connectedness is not a simplistic hypothesis: indeed groups of atoms, like functional groups, get most of their influence on the formable character of the input bond $b$, through bonds linking this group to $b$.

An *occurrence* of an environment $E$ in a example $g \in \mathcal{E}$ is then defined as an injective morphism $\mu$ from $\mathbf{V}(E)$ to $\mathbf{V}(g)$ that preserves bond-atom incidence relationship, atom labeling, and bond labeling. Formally if $\lambda_g^v$ and $\lambda_g^e$ denote the functions mapping vertices and edges, respectively, of graph $g$ to their label, then morphism $\mu$:$\mathbf{V}(E) \rightarrow \mathbf{V}(g)$ verifies

$$\forall \{v_1;v_2\} \in \mathbf{E}(E), \{\mu(v_1);\mu(v_2)\} \in \mathbf{E}(g) \text{ and } \lambda_g^e(\{\mu(v_1);\mu(v_2)\}) = \lambda_E^e(\{v_1;v_2\}) \quad (1)$$

$$\forall v \in \mathbf{V}(E), \lambda_g^v(\mu(v)) = \lambda_E^v(v) \quad (2)$$

Figure 2 shows three occurrences of the environment $E_1$ of Figure 1 in a set of three examples.

An occurrence of an environment can either credit or discredit the hypothesis "$b$ is formable": an occurrence of $E$ in $g$ is *positive* regarding hypothesis if the bond of $g$ mapped to the input bond $b$ is labeled as formable in the example, that is $\mu(b) \in F(g)$, otherwise it is *negative*. The number of positive (respectively negative) occurrences of $E$ in all examples of $\mathcal{E}$ is denoted $\text{occ}^+(E)$ (respectively $\text{occ}^-(E)$). Given one single environment $E$ of the input bond $b$, the probability that $b$ is formable given $E$ relatively to the distribution of bond environments in the set $\mathcal{E}$ of examples is

$$P(b \text{ is formable} | E \text{ is an environment of } b) = \frac{\text{occ}^+(E)}{\text{occ}^+(E) + \text{occ}^-(E)}$$

This conditional probability is called the *confidence* of environment $E$ in the hypothesis "bond $b$ is formable" and is denoted $\text{conf}(E)$. For example, in Figure 2, only example $e_3$ contains a negative occurrence as the bond mapped to $b$ is not formed by the considered reaction. The resulting confidence is therefore $\text{conf}(E) = 2/3$. The term confidence, like the subsequent term frequency, are chosen according to an analogy with the data-mining problem of frequent association rules.[30] However confidence is here defined as a ratio of numbers of environment occurrences, whereas confidence is usually defined in data-mining as a ratio of the pattern frequency within the subset of positive examples over the pattern frequency within the set of positive and negative examples. As a consequence, an environment occurring at two different places in one molecule is counted twice when computing its confidence. Usually, the bond $b$ should be classified as formable if its confidence is larger than 0.5. In practice several reasons make the problem not so straightforward: First, in the automated labeling case, the decision threshold is unknown and less than 0.5 since the number of formable bonds is underestimated, and this makes the confidence less than the real conditional probability. Second, the test assumes environment $E$ really occurs in the examples, that is, $\text{occ}^+(E) + \text{occ}^-(E) > 0$, for the confidence of $E$ to be defined. Third and more importantly, the input bond does not provide one but many different environments to be considered.

The second point raises the more general issue of environment representativeness: the more an environment occurs in various examples, the more the estimation of confidence is representative of examples and consequently is reliable. To assess the trust attached to a confidence value, the *frequency* $\text{freq}(E)$ of an environment $E$ of $b$ is introduced as the proportion of examples in $\mathcal{E}$ that contain at least one occurrence, positive or negative, of $E$. The higher the frequency of $E$, the more reliable the estimation of $\text{conf}(E)$. The frequency $\text{freq}(E)$ is used instead of the number $\text{occ}^+(E) + \text{occ}^-(E)$ of occurrences because frequency holds a property consistent with the intuitive idea of representativeness: frequency is a decreasing function with respect to environments. In other words, if two environments $E_1$ and $E_2$ of $b$ are such that $E_1$ is a subgraph of $E_2$ then $\text{freq}(E_1) \geq \text{freq}(E_2)$. Both confidence and frequency also have this other convenient property to range from 0 to 1.

In practice, the input bond $b$ does not have one but a large number of environments in the input molecular graph $G$. The question is then to decide which subset of these environments should be considered in order to compute the global confidence index $\text{conf}_G(b)$ on which the final classification of $b$ is based. It seems natural to think the closer to $G$ an environment $E$ of $b$, the closer confidences $\text{conf}_G(b)$ and $\text{conf}(E)$. As the environments that are the closest to $G$ are also the largest ones, the goal is to find the said "maximally occurring environment" $E_{\text{maxocc}}$, such that (i) $E_{\text{maxocc}}$ occurs in at least one example and (ii) every environment that contains $E_{\text{maxocc}}$ as a subgraph, does not occur in the examples. In general there is not a single but several environments $E_{\text{maxocc}}$ as illustrated by the simple example of Figure 3. In that example, input bond $b$ of input graph $G$ has two maximal environments $E_1$ and $E_2$ occurring in

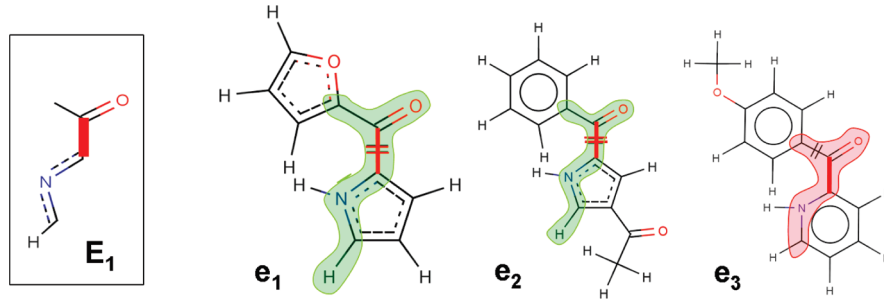



**Figure 2.** Some occurrences of environment $E_1$ in three examples. A formed bond is denoted by two parallel crossing strokes.

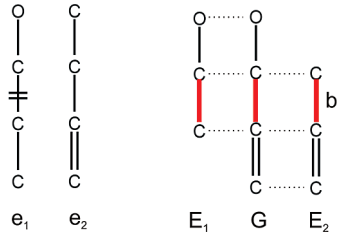



**Figure 3.** Maximally occurring environments.

examples $e_1$ or $e_2$. Environments $E_1$ and $E_2$ are maximal as augmenting either $E_1$ or $E_2$ by a bond of $G$ produces an environment that does not occur anymore in the examples. Moreover environments $E_1$ and $E_2$ respectively occur positively in $e_1$ and negatively in $e_2$. Their respective confidences of 1 and 0 give opposite advices, whereas environments $E_1$ and $E_2$ are hardly distinguishable in other respects because they have same frequency (1) and same size (2 bonds). This example shows that in the classification of an edge, wrt, its environment is a difficult problem in the general case. However it is possible to take advantage of the specific application for proposing an efficient heuristic solution as shown in the next section.

**GemsBond Algorithm for Ranking Formable Bonds.** GemsBond[16] is an algorithm based on a heuristic search of environment, using introduced notions of confidence and frequency.

*Heuristic.* As shown in the previous section, classification of a vertex or edge based on its environments is a hard problem. In order to overcome this difficulty, GemsBond algorithm is based on a heuristic specific to the problem of discovering formable bonds. This heuristic considers that the formable character of a bond $b$ mainly depends on the environment of $b$ that best supports the hypothesis that " $b$ is formable". This assumption formally means that the global confidence of bond $b$ is equal to the maximal value of confidence that can be reached by any environment of $b$

$$\mathrm{conf}_G(b) = \max_{E \text{ is an environment of } b} (\mathrm{conf}(E))$$

This maximal confidence is reached by the environment $E_{\max}$ called *explanatory environment* or simply *explanation* as it justifies the level $\mathrm{conf}_G(b)$ of confidence associated to input bond $b$. In case the maximal confidence is reached by several environments, the most representative environment, that is, the one with the highest frequency, is chosen as the explanation.

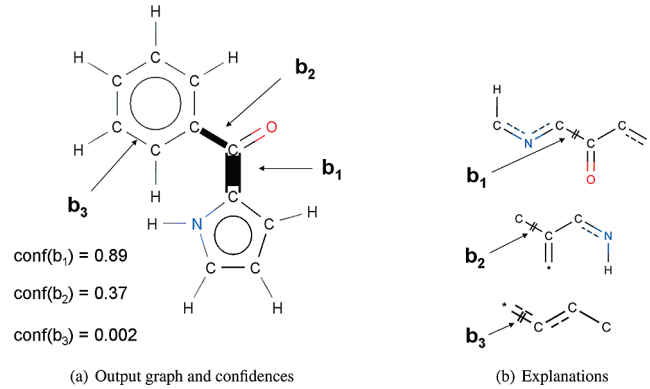
In other words, this heuristic introduces asymmetry in the problem by assuming that environments with low confidence have a negligible influence compared to environments with high confidence. This hypothesis results from the following argument: a bond environment with low confidence may be interpreted in two ways. Indeed an environment $E$ of bond $b$ has a low confidence either because the presence of $E$ may conflict with the formation of $b$ (i.e., some characteristics of $E$ inhibit the formation of $b$) or merely because $E$ is "neutral" with respect to formability of $b$ (i.e., $E$ does not bring any element that could make the formation of $b$ easier). However when one observes how environments with low confidence look like, one notices the "neutrality" of these environments as they do not exhibit any specific structural configuration that could inhibit the formation of their bond $b$. Unlike environments with low confidence, all environments with high confidence do include specific functional group configurations. Since environments with low confidence do not carry significant information, while those with high confidence do, the formability of $b$ mostly depends on the presence or absence of those environments with high confidence. This observation leads to estimate the formability of a bond based on the environment with the highest confidence, that is, to choose the aforementioned heuristic. Experiments described in the last section eventually confirm the soundness of this heuristic.

Apart from soundness, this heuristic also holds several advantages: first the explanation is easily analyzable because it is made of a single environment. Second the experiments show the explanatory environments selected by the heuristic are generally sufficiently frequent to be representative as they are small (typically composed of less than 10 atoms). Last, as the environments $E_{\max}$ are relatively small, their occurrences can be computed faster than those of larger environments like $E_{\mathrm{maxocc}}$.

*Search Algorithm.* The problem now consists in finding the environment $E_{\max}$ of $b$ that has the highest confidence. Because confidence does not hold any "useful property" (e.g., the antimonotonic property of frequency) that could restrict the search of $E_{\max}$, this optimization problem is considered as a general search problem in a state space guided by an evaluation function, as introduced by the artificial intelligence community and described in Chapter 3 of Russell and Norvig's book.[31] Here the state space is the set of environments of $b$, the initial state is the environment reduced to bond $b$, and the goal of the search is to reach $E_{\max}$. As the structure of the goal environment is unknown, the search potentially requires to enumerate all possible environments of $b$, to compute their confidence and to select the environment of highest confidence. However a systematic enumeration of all environments of $b$ is not tractable as the number of those environments is large and grows in average exponentially with the size of environment (cf., Figure 8). A pruning strategy is thus required to limit the search space.

As $E_{max}$ is expected to be rather small compared to the size of the input graph $G$, the idea is to search through the state space by making an environment $E$ iteratively growing from the input bond $b$. The initial state of the search is thus the environment reduced to the single bond $b$. The successive extensions of the current environment $E_{current}$ are guided by the growth of its confidence conf($E_{current}$), until every further extension makes the confidence decrease. The algorithm guarantees to find at least one local maximum but not necessarily the expected global maximum. At each step of the iterative algorithm, every possible extension of environment $E_{current}$ must be considered. Extensions consist in adding any bond to $E_{current}$ that is incident to an atom of $E_{current}$ but not already in $E_{current}$, or dually, in adding any atom to $E_{current}$ that is incident to a bond of $E_{current}$ but not already in $E_{current}$. These extensions can be enumerated efficiently in constant amortized time.

Some of the most common searching strategies have been tested in GemsBond from the simplest greedy search to more sophisticated variants of beam search strategy. As those more sophisticated strategies have not provided substantial improvements in quality of the results so far, only the simplest tested strategy, that is, greedy search, is presented: at each step the greedy algorithm only develops further extensions that locally provide the highest gain of confidence. The process is iterated until no extension further increases confidence. Greedy search is simple and fast but may converge toward a suboptimal local maximum of confidence. For this reason, a depth threshold $d_{min}$ is introduced as a parameter of GemsBond; while the exploration depth $d$ remains smaller than $d_{min}$, every extension of the current environment is developed regardless of its confidence. Since $d + 1$ is equal to the size of the current environment $E_{current}$ defined as the sum of the number of vertices and edges of $E_{current}$, the algorithm guarantees to develop at least all environments, whose size is less or equal to $d_{min} + 1$. To control the representativeness of found explanations, another parameter $f_{min} \in [0; 1]$ has been introduced for developing only environments whose frequency is greater than or equal to $f_{min}$. Pseudocode 1 summarizes GemsBond algorithm. The main loop enumerates every possible extension $e$ of the current environment $E_{current}$ in the input graph $G$ (line 1) before the confidence $c$ and frequency $f$ of the extended environment $e(E_{current})$ is evaluated (line 2). Only the environments that have a sufficient frequency and a maximal confidence (locally) or a size smaller than $d_{min}$ are further developed by recursive calls (line 6). Computing the confidence and frequency of $E_{current}$ (line 2) requires to count all positive and negative occurrences of $E_{current}$ in $E$. Graph mining algorithms using a depth-first search can efficiently compute the number of occurrences of the current graph pattern by using a fast-access and compact data structure called "embedding list" (see details in ref 21). This structure has been adapted to count simultaneously positive and negative occurrences in one single pass over the examples so that computing the confidence does not require more time than computing a number of occurrences. For sake of efficiency, extensions that produce an environment whose confidence is null or frequency is less than $f_{min}$ (line 3) are pruned and ignored by nested recursive calls to procedure *findEMax* (lines 5 and 7). This pruning saves processing time without modifying the results as both conditions conf($E_{current}$) $> 0$ and freq($E_{current}$) $\geq f_{min}$ are antimonotonic (i.e., false values for these conditions can never get true when extending the current environment).




(a) Output graph and confidences (b) Explanations

**Figure 4.** Output of GemsBond.

**Algorithm 1**: The greedy procedure findEMax($d$, $E_{current}$, $c_{current}$, $f_{current}$)

**Data**: input graph $G$, input bond $b$, example set $\mathscr{E}$, $d_{min}$, and $f_{min}$ thresholds
**Input**: current depth $d$, environment $E_{current}$, confidence $c_{current}$, and frequency $f_{current}$
**Result**: explanation $E_{max}$, confidence $c_{max}$, and frequency $f_{max}$ are global variables initialized resp. to the subgraph reduced to $b$, conf($E_{max}$), and freq($E_{max}$)
Set $C \leftarrow \emptyset$ of candidate extensions ;
$c_{\text{local max}} \leftarrow 0$ ;
Set $D$ of disabled extensions is a global variable initialized to $\emptyset$ ;
Set $D_e \leftarrow \emptyset$ of locally disabled extensions to enable ;
1 **forall** *extension* $e$ *of* $E_{current}$ *in* $G$ *but not in* $D$ **do**
2 $c \leftarrow \text{conf}(e(E_{current}), \mathscr{E})$ ; $f \leftarrow \text{freq}(e(E_{current}), \mathscr{E})$ ;
3 **if** $f \geq f_{min}$ *and* $c > 0$ **then**
**if** $d < d_{min}$ **then**
$C \leftarrow C \cup \{(e, c, f)\}$
**else**
4 **if** $c \geq c_{local\ max}$ *and* $c > c_{current}$ **then**
**if** $c > c_{local\ max}$ **then**
$c_{\text{local max}} \leftarrow c$ ; $C \leftarrow \emptyset$
$C \leftarrow C \cup \{(e, c, f)\}$
**else**
$D_e \leftarrow D_e \cup \{e\}$
**if** $c_{current} > c_{max}$ *or* $(c_{current} = c_{max}$ *and* $f_{current} > f_{max})$ **then**
$c_{max} \leftarrow c_{current}$ ; $f_{max} \leftarrow f_{current}$ ; $E_{max} \leftarrow E_{current}$
5 $D \leftarrow D \cup D_e$ ;
**forall** $(e, c, f) \in C$ **do**
6 findEMax($d + 1$, $e(E_{current})$, $c$, $f$)
7 $D \leftarrow D \setminus D_e$

*Output.* The output of GemsBond associates to every bond $b$ of the input molecule a confidence $\text{conf}_G(b) = \text{conf}(E_{max})$, an explanatory environment $E_{max}$, and a frequency freq($E_{max}$). The confidence is used to modulate the thickness of every bond in the drawing, i.e. the thicker, the more formable, so that the user may visually capture the output at a glance. The output for input graph of Figure 1 is given in Figure 4. The most formable bond $b_1$ coincides with the bond formed by the considered reaction.

In addition, every time GemsBond processes new data, resulting explanations are collected and merged to those resulting from previous processing. The resulting list of explanations is sorted by decreasing order of confidence and then frequency. The head of this list provides highly formable environments that might contribute to knowledge discovery from reaction databases. Figure 5 gives the six first environments in the list of explanations forming a carbon−carbon bond sorted by decreasing order of confidence and then frequency, after processing a hundred of input molecules based on several thousand examples. All those environments have a representative frequency ranging from 20 to 40, and

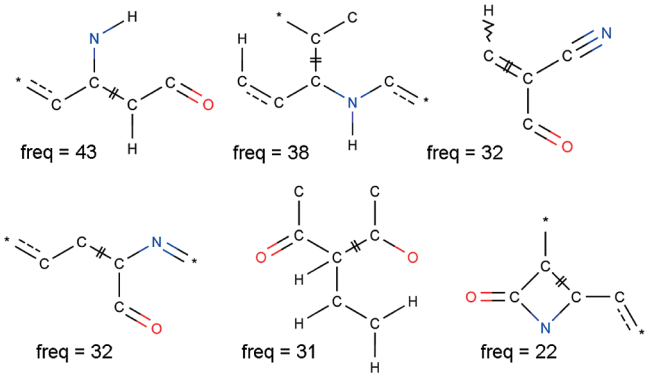



**Figure 5.** Six explanations of confidence 1 to form carbon−carbon bonds. Frequency values are absolute: frequency is the number of examples containing the environment.

because their confidence is 1, their numerous occurrences in the examples do not include any negative counterexample. All those environments that are characteristic of highly formable bonds, include combinations of heteroatoms and aromatic or multiple bonds as shown in the discussion section.

**Binary Classifiers to Predict Formable Bonds.** Once the global confidence index of every bond in a molecule has been computed, it is straightforward to address the problem of ranking formable bonds. To address the second problem of discovering formable bonds, which has been introduced to evaluate the performance of GemsBond, a binary classifier has to transform confidence values output by GemsBond into yes/no answers. Since the different types of bonds (single, double, triple, and aromatic) in the examples have their own distinct distributions, especially distributions of confidence, every type of bond requires a distinct classifier. Only three classifiers are developed subsequently even if many other classifiers could be considered:

*Thresholding on Confidence.* Given such a classifier for bond type $T$ parametrized by a threshold $c^T_{\min} \in [0; 1]$, a bond $b$ of type $T$ is classified as formable if $\text{conf}_G(b) \geq c^T_{\min}$.

*Thresholding on Rank.* Given such a classifier for bond type $T$ parametrized by a threshold $r^T_{\max} \in \mathbb{N}$, a bond $b$ of type $T$ in a input molecule $G$ is classified as formable if the rank $r^T(b)$ of $b$ is less or equal to a positive integer $r^T_{\max}$, in the list of bonds of type $T$ in $G$ sorted in decreasing order of confidence.

*Thresholding on Relative Rank.* Given such a classifier for bond type $T$ parametrized by a threshold $p^T_{\max} \in [0; 1]$, a bond $b$ of type $T$ in a input molecule $G$ is classified as formable if the relative rank of $b$ is less or equal to $p^T_{\max}$, where this relative rank is the fraction of the rank $r^T(b)$ divided by the number of bonds of type $T$ in $G$.

For instance, the eleven single bonds in molecule of Figure 4a are bond $b_1$ of confidence 0.89, bond $b_2$ of confidence 0.37, and eight C−H and one N−H bonds attached to an aromatic ring, all of confidence 0.06. To classify $b_1$ as the only formable single bond, classifier based on confidence thresholding (respectively rank and relative rank thresholding) must have its threshold $c^{\text{single}}_{\min}$ such that $0.37 < c^{\text{single}}_{\min} \leq 0.89$ ($r^{\text{single}}_{\max}$ equal to 1 and $p^{\text{single},}_{\max}$ such that $1/11 \leq p^{\text{single}}_{\max} < 2/11$, respectively).

Each of these three classifiers correspond to a different intuition: the intuition underlying the first classifier is that the formability of a bond $b$ only depends on the environments of $b$ irrespective of the formability of the other bonds in $G$. The second classifier assumes the number of formable bonds in a molecule is a constant number whatever the size of the molecule is. Finally the third classifier assumes the number of formable bonds is in average proportional to the number of bonds of the molecule $G$. The first and third classifiers intuitively make sense while the second does not. However in the automated labeling, the number of examples of formable bonds does not tend to be proportional to the size of $G$. It rather tends to be constant as a single reaction only form a small number of bonds (2 in average). Because of this artifact, the second classifier artificially makes more sense than the third one.

**Table 1.** Data Selection

| selection operation | remaining number of reactions |
|---|---|
| ChemInform RX and Reference Library | 1 202 174 |
| formation of C−C bond (any type) | 366 264 |
| yield ≥ 90% | 10 774 |
| monostep, monoproduct reaction | 10 170 |
| selection of main elements | 8256 |
| preprocessing | 7537 |

## EXPERIMENTAL SECTION

**Experiment Description.** Data were selected from two well-known reaction databases *ChemInform RX* and *Reference Library* provided by Symyx company. These two databases cover a long period and offer a large reaction diversity in terms of structural features such as substrate topology, stereochemical course, or substituent effects. Various classes of substances are represented as well as new reagents, catalysts, and experimental conditions. The amount of data to process was then reduced according to some selection procedure as summarized in Table 1. Reactions forming carbon−carbon bonds were selected first since these bonds are the most energy-demanding bonds within the skeleton of complex molecules. A random sample was built by setting a yield value to 90% then by selecting monostep reactions that lead to a single product. Finally, a selection of main elements (B, C, N, O, F, Si, P, S, Cl, Br, and I) allowed us to discard molecules containing ionic or organometallic bonds. The preprocessing of these 8256 reactions according to a previously published data preprocessing[32] led to 7537 reactions. This latter procedure discards reactions with incomplete or ambiguous atom mapping, saturates molecular graphs with hydrogen, checks consistency between the status of every bond (formed, broken, modified, or stable) and atom mapping, and extracts the main product. For this test, we assumed a broad definition of bond formability that does not depend on any specific experimental condition or catalyst so that we did not filter database entries based on catalyst and condition fields. However, it is worth noticing that the definition of bond formability may be conditioned by some specific families of reactions or experimental conditions; this only requires to select the entries of reaction databases according to the considered specific conditions.

Table 2 describes the statistical distribution of bonds in the resulting set of molecular graphs. Since retrosynthesis is mostly interested in forming carbon−carbon bonds and since most bonds incident to an hydrogen atom are missing

**Table 2.** Statistical Distributions of Bonds in Examples

| bond family | bond type | single | double | triple | aromatic | total |
|---|---|---|---|---|---|---|
| hydrogen incident bonds | $N$ | 145823 | | | | 145823 |
| | $N/N_{tot}$ | 47% | | | | 47% |
| | $N_c/N$ | 2.6% | | | | 2.6% |
| carbon−carbon bonds | $N$ | 54146 | 4226 | 349 | 59209 | 117930 |
| | $N/N_{tot}$ | 17% | 1.3% | 0.11% | 19% | 38% |
| | $N_c/N$ | 12.5% | 20.4% | 2.3% | 1.6% | 7.3% |
| other bonds | $N$ | 32696 | 9455 | 732 | 5382 | 48265 |
| | $N/N_{tot}$ | 10% | 3.0% | 2.3% | 1.7% | 15.5% |
| | $N_c/N$ | 5.6% | 2.2% | 0.8% | 10% | 5.3% |
| all bonds | $N$ | 232665 | 13681 | 1081 | 64591 | 312018 |
| | $N/N_{tot}$ | 75% | 4.4% | 0.3% | 21% | 100% |
| | $N_c/N$ | 5.3% | 7.8% | 1.3% | 2.3% | 4.8% |

in molecular graphs, bonds have been split into three families: carbon−carbon bonds, bonds incident to an hydrogen atom obtained after saturation of molecular graphs with hydrogen, and remaining bonds (i.e., carbon−heteroatom or heteroatom−heteroatom). For every pair of bond type and family, $N$ denotes the number of these bonds, $N/N_{tot}$ the ratio within the total number $N_{tot} = 312\,018$ of bonds in the examples, and $N_c/N$ the proportion of formed bonds within the category.

## RESULTS

Various tests have been performed to assess the effects of the different variables of the problem such as the different types of bonds, the different binary classifiers, the different values for $d_{min}$ and $f_{min}$ parameters and some of the different models of molecules. The principle of those tests is always the same: every test runs GemsBond on a set $\mathscr{S}$ of input molecular graphs, whose subset $F_i$ of formable bonds is known (as the formed bonds in main reaction products). By comparison of the output subset $F_o$ of formable bonds predicted by classifier with the input subset $F_i$, bonds are split into four classes of *true positive* (i.e., $F_o \cap F_i$), *true negative* (i.e., $F_o^C \cap F_i^C$, where $E^C$ denotes the complement of set $E$ within the set of bonds), *false positive* (i.e., $F_o \cap F_i^C$), and *false negative* (i.e., $F_o^C \cap F_i$) bonds. If their respective numbers in $\mathscr{S}$ are denoted TP, TN, FP, and FN, then the *specificity*, *sensitivity*, and *accuracy* of a classifier are respectively computed as SP = TN/(TN + FP), SE = TP/(TP + FN), and ACC = (TP + TN)/(TP + FP + TN + FN). Given a binary classifier, like thresholding based on either confidence, rank, or relative rank, for a given bond type $T$, values for the classifier parameter $p$ (i.e., $c_{min}^T$, $r_{max}^T$, or $p_{max}^T$) are sampled to draw the ROC curve (SP($p$), SE($p$)) in the unit square.[33]

The numbers of formed and nonformed bonds are imbalanced: formed bonds are 19 times less frequent than not formed bonds, all types of bond being taken into account. This imbalance leads to the well-known "accuracy paradox": the "blind" classifier predicting every bond as not formable achieves a very high accuracy equal to the average ratio of not formed bonds that is 95%. Instead, the area under the ROC Curve[33] or AUC and the F-measure[34] are provided as the most common criteria to assess the quality of a binary classifier. AUC is useful for evaluating the quality of ranking formable bonds as AUC is proved to be equal to the probability for a randomly selected formed bond to get a better score (i.e., a higher confidence or a lower rank depending on the considered classifier) than a randomly selected not formed bond. $F$-measure is the harmonic mean of the precision PREC = TP/(TP + FP) and sensitivity SE (i.e., $F_{mes} = (2 \cdot \text{PREC} \cdot \text{SE})/(\text{PREC} + \text{SE})$). $F$-measure was originally designed for information retrieval problems whose purpose is not to maximize accuracy but the chance to find some truly positive documents thus relaxing the penalty of retrieving some false positive documents. Consequently, $F$-measure is adapted to the problem of discovering formable bonds as the purpose here is to find some formable bonds rather than minimizing the error rate. Since tests take their input molecules in excerpts of reaction databases (automated labeling), the underestimate of formable bonds in $\mathscr{S}$ induces an overestimate of false positive errors FP and, thus, an underestimate of the specificity and finally of AUC and $F$-measure. The results displayed are thus a pessimistic bound of the actual performance of GemsBond. All tests use cross-validation: the 7537 reaction products are split into 75 subsets of 100 molecules each. Confidences of bonds within each subset are computed using the 7437 remaining products as examples. Finally results are averaged over the 75 tests. The remaining part of this section describes these results in terms of classification accuracy and processing time that have been obtained for a reference test. Then the section studies the influence of the various factors relatively to this reference test.

**Reference Test.** The reference test uses a thresholding on confidence applied to the set of examples saturated with hydrogen atoms and with both parameters $d_{min}$ set to 5 and $f_{min}$ set to 5/7537 = 0.07%. Figure 6 displays for each type of bonds, the normalized distributions of confidence for both sets of formed and not formed bonds. Except for the triple bond that is not sufficiently represented in the examples, GemsBond computes higher values of confidence for formed bonds than for not formed ones on average so that a thresholding on confidence is able to separate efficiently both distributions. Accordingly, the associated ROC curves shown in Figure 7 are clearly above the diagonal except for triple bonds. The performance of the classifier is summarized in Table 3.

According to $F$-measure, the bonds whose formability is the easiest to be predicted, are single bonds, followed by double, aromatic, and triple bonds in this order. Bad performance for triple bonds was expected since examples have not been selected to focus on the particularly rare event of triple bond formation: examples contain only 11 (0.004%) formed triple bonds out of only 1081 (0.3%) triple bonds.

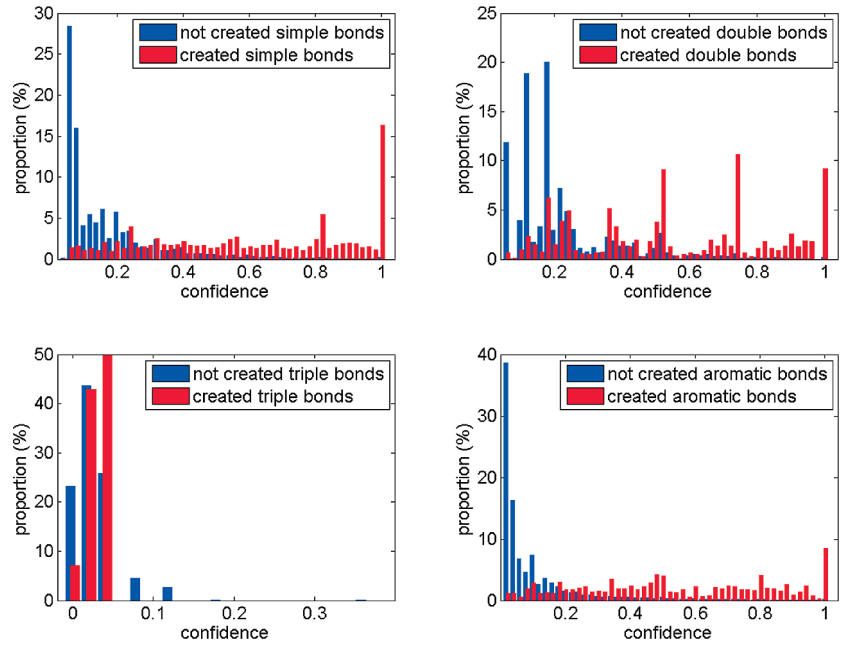



**Figure 6.** Normalized distributions of formed and not formed bonds.

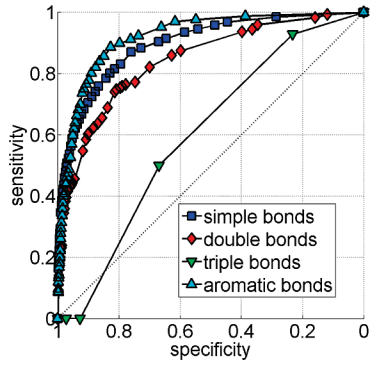



**Figure 7.** ROC curves when thresholding on confidence with $d_{min} = 5$. All curves, but the one for triple bonds, are clearly above the diagonal.

On the contrary, excellent performance in predicting formable aromatic bonds was unexpected as aromatic bonds were not the focus of the data selection either. But contrarily to triple bonds, aromatic bonds are very common (21% of bonds).

For the sake of a fair evaluation, these results may be assessed by computing a prediction gain ranging from 0 to 1. This quality coefficient defined over a measure (i.e., AUC, maximal accuracy, or maximal *F*-measure) is chosen so that it is equal to zero for random classifiers and equal to one for the perfect classifier (i.e., without any prediction error). Measures related to the family of random classifiers with probability α of classifying a bond as formable are an AUC of 0.5, a maximal accuracy of $1 - N_c/N$ reached for $\alpha = 0$ and a maximal *F*-measure of $(2 \cdot N_c)/(N_c + N)$ for $\alpha = 1$, where ratio $N_c/N$ denotes the ratio of formed bonds. Given then a measure *M* (i.e., AUC, maximal accuracy, or maximal *F*-measure), the prediction gain on *M* is computed as

$$G_M = \frac{M_{\text{gemsbond}} - M_{\text{random}}}{M_{\text{perfect}} - M_{\text{random}}} = \frac{M_{\text{gemsbond}} - M_{\text{random}}}{1 - M_{\text{random}}}$$

where $M_{\text{perfect}}$, $M_{\text{gemsbond}}$, and $M_{\text{random}}$ are the maximal value of *M* provided by the perfect classifier, GemsBond (for the optimal choice for $c_{min}$), and the random classifier (for the optimal choice for α), respectively. Table 4 provides these prediction gains of GemsBond relatively to the family of random classifiers: As explained earlier, gain of accuracy randomly fluctuates whereas significant gains are observed for AUC and F-measure. The accuracy paradox clearly appears for aromatic bonds as GemsBond gets for this bond type, the highest predictive gain (+84% on *AUC*) and in the meantime, the lowest gain on accuracy (−30%), that is even lower than the accuracy of the random classifier.

**Influence of the Various Parameters.** *Influence of the Type of Binary Classifier.* Table 5 compares performance of the three proposed binary classifiers (i.e., thresholding on confidence, rank, and relative rank), other things being equal to the reference test. Thresholding on confidence provides the best performance. This observation confirms that the formability of a bond mostly depends on its environment independently of the formability of the other bonds in the molecule. Otherwise there is no difference of performance between classifiers based on rank and relative rank, whereas, as explained earlier, a theoretical advantage was given to thresholding on rank.

*Influence of the Minimal Depth of Search $d_{min}$.* GemsBond has to mine every environment whose size, defined as the total number of atoms and bonds of the environment, is less than threshold $d_{min}$. The larger $d_{min}$, the longer the processing time but the higher the chance of finding an environment $E_{max}$ of maximal confidence. This last effect might in turn influence the classification accuracy. Figure 8 gives the influence of $d_{min}$ on processing time, AUC and *F*-measure. Without surprise, the processing time appears to grow exponentially with threshold $d_{min}$ since the number of environments an edge has in average, grows exponentially with the size of the environment. AUC also tends to grow with $d_{min}$ but very slowly and to a varying extent: the $d_{min}$ parameter growth improves prediction for aromatic bonds and for single bonds to a smaller extent but it does not improve prediction for double bonds for which a greedy search is sufficient. Therefore a good default value for $d_{min}$ is 5 where the four AUCs are almost optimal while the

**Table 3.** Results of Reference Test

| | | | accuracy | | *F*-measure | |
|---|---|---|---|---|---|---|
| bond family | bond type T | AUC | maximal value | associated $c^T_{min}$ | maximal value | associated $c^T_{min}$ |
| all bonds | single | 0.90 | 0.96 | 0.82 | 0.50 | 0.64 |
| | double | 0.84 | 0.93 | 0.90 | 0.46 | 0.66 |
| | triple | 0.60 | 0.99 | 0.38 | 0.04 | 0.04 |
| | aromatic | 0.92 | 0.97 | 0.94 | 0.36 | 0.66 |
| | any type | 0.90 | 0.96 | NR | 0.47 | NR |
| carbon−carbon bonds | single | 0.84 | 0.89 | 0.86 | 0.51 | 0.64 |
| | double | 0.77 | 0.83 | 0.74 | 0.52 | 0.64 |
| | triple | 0.50 | 0.98 | 0.38 | 0.05 | 0.04 |
| | aromatic | 0.92 | 0.98 | 0.94 | 0.35 | 0.66 |
| | any type | 0.88 | 0.94 | NR | 0.43 | NR |

**Table 4.** Prediction Gain Relatively to the Family of Random Classifiers

| bond family | bond type | gain of AUC | gain of maximal accuracy | gain of maximal *F*-measure |
|---|---|---|---|---|
| all bonds | single | 80% | 25% | 44% |
| | double | 68% | 10% | 37% |
| | triple | 20% | 23% | 1% |
| | aromatic | 84% | −30% | 33% |
| | any type | 80% | 17% | 42% |
| carbon−carbon bonds | single | 68% | 12% | 37% |
| | double | 54% | 17% | 27% |
| | triple | 0% | 13% | 1% |
| | aromatic | 84% | −25% | 33% |
| | any type | 76% | 18% | 34% |

**Table 5.** AUC and *F*-Measure of Classifiers

| | thresholding on confidence | | | thresholding on rank | | | thresholding on relative rank | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *F*-measure | | | *F*-measure | | | *F*-measure | |
| bond type T | AUC | max value | assoc. $c^T_{min}$ | AUC | max value | assoc. $r^T_{max}$ | AUC | max value | assoc. $p^T_{max}$ |
| single | 0.90 | 0.50 | 0.64 | 0.88 | 0.47 | 3 | 0.88 | 0.45 | 0.14 |
| double | 0.84 | 0.46 | 0.66 | 0.69 | 0.22 | 2 | 0.69 | 0.22 | 0.12 |
| triple | 0.60 | 0.37 | 0.04 | 0.47 | 0.03 | 3 | 0.48 | 0.03 | 0.56 |
| aromatic | 0.92 | 0.36 | 0.66 | 0.72 | 0.10 | 3 | 0.73 | 0.10 | 0.20 |
| any | 0.90 | 0.47 | NR | 0.84 | 0.38 | NR | 0.84 | 0.37 | NR |

average processing time of a bond does not exceed two tenths of a second.

*Influence of the Minimal Frequency Threshold $f_{min}$.* Figure 9 depicts the influence of the minimal frequency threshold $f_{min}$. Processing time is a decreasing function of $f_{min}$ as the investigated environments are pruned when $f_{min}$ increases. The processing time even appears to be a function of type $t = a + b \cdot \log(1/f_{min})$ (cf., Figure 9d), expressing the fact that every extension of the current environment decreases in average its frequency by a constant factor. AUC (Figure 9a) and *F*-measure (Figure 9b) remain stable, while $f_{min}$ is kept sufficiently low. Once $f_{min}$ is raised, AUC and *F*-measure jump down as important explanatory environments are pruned by the constraint of minimal frequency. When $f_{min}$ tends to 1, both AUC and F-measure tend to values of random classifiers.

*Influence of Hydrogen Atoms.* Examples and input molecules have been saturated by hydrogen atoms before they have been processed by GemsBond because hydrogen atoms may play a role in the environment $E_{max}$. Table 6 gives the AUC and *F*-measure for predicting formable carbon−carbon bonds with and without hydrogen saturation. As expected, the hydrogen atoms play a role in the formable character of carbon−carbon bonds even if their influence is globally weak. This influence is confirmed by the presence of

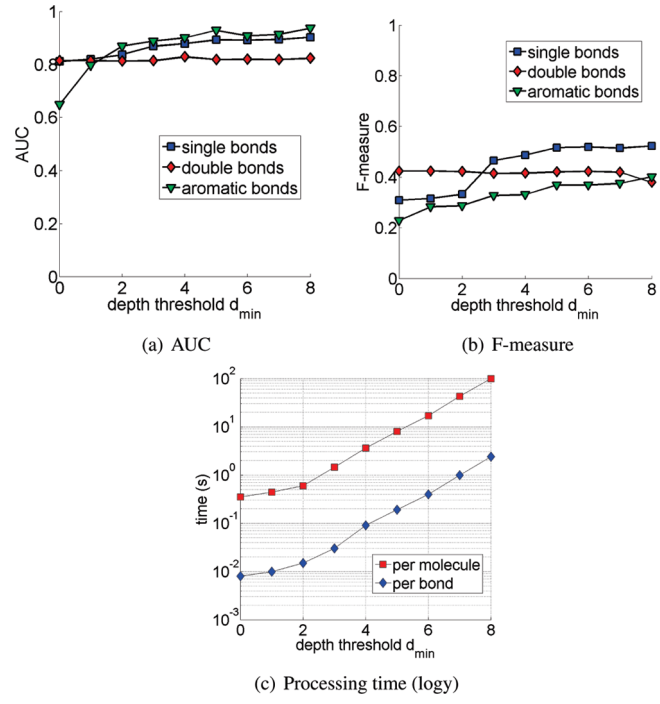
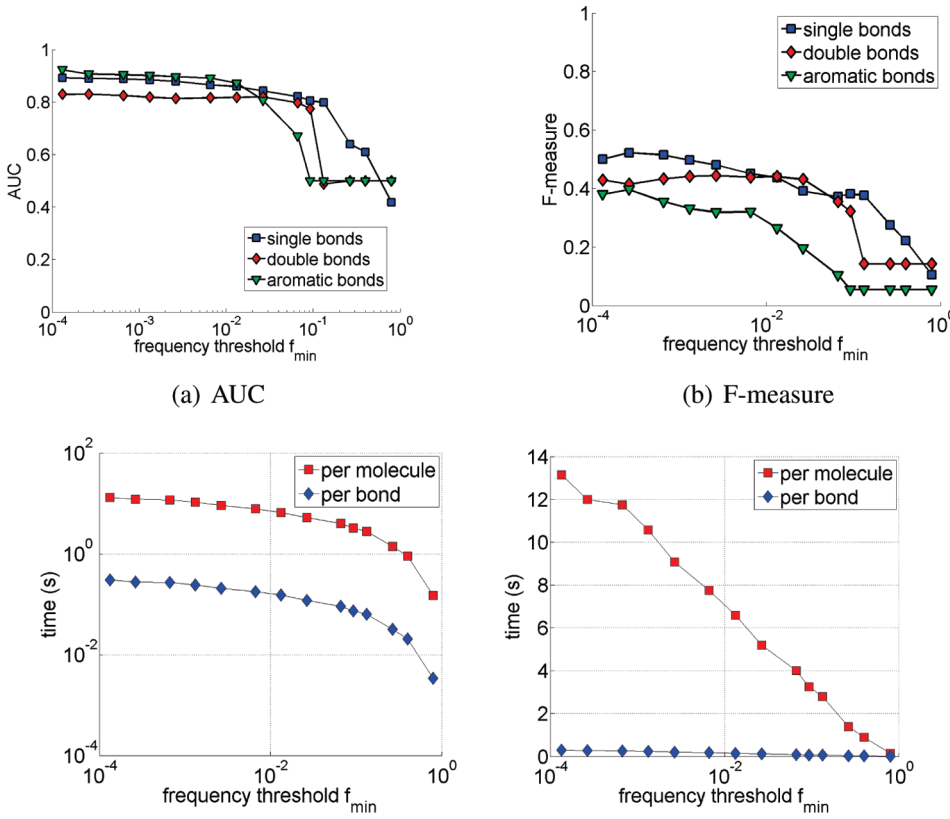



**Figure 8.** Influence of $d_{min}$.




(a) AUC (b) F-measure

(c) Processing time (loglog) (d) Processing time (logx)

**Figure 9.** Influence of $f_{min}$.

**Table 6.** Influence on Carbon−Carbon Bonds of Saturation by Hydrogen Atoms

| | with hydrogen saturation | | without hydrogen saturation | |
|---|---|---|---|---|
| bond type | AUC | *F*-measure | AUC | *F*-measure |
| single | 0.84 | 0.50 | 0.81 | 0.47 |
| double | 0.77 | 0.52 | 0.77 | 0.52 |
| triple | 0.50 | 0.05 | 0.50 | 0.04 |
| aromatic | 0.92 | 0.35 | 0.93 | 0.32 |
| any | 0.88 | 0.43 | 0.87 | 0.40 |

**Table 7.** Influence on Processing Time of Saturation by Hydrogen Atoms for Different $d_{min}$ Values

| | processing time (s) per molecule | |
|---|---|---|
| $d_{min}$ value | without hydrogen | with hydrogen |
| 0 | 0.4 | 0.4 |
| 3 | 0.7 | 1.4 |
| 5 | 1.7 | 7.2 |
| 7 | 4.7 | 49 |

hydrogen atoms in explanations having high confidence. However this improvement is costly as shown in Table 7, since the large number of added hydrogen atoms substantially increase the number of possible environments to mine and consequently the processing time when $d_{min}$ increases.

## DISCUSSION

As previously mentioned, GemsBond produces the confidence $\text{conf}_G(b)$ of an input bond $b$, the explanatory environment or explanation $E_{max}$, whose confidence $\text{conf}(E_{max})$ is equal to $\text{conf}_G(b)$, and the frequency $\text{freq}(E_{max})$. The explanation is the structural environment relevant to justify the formability level of bond $b$. As a chemical graph, an explanation can be directly visualized and interpreted by an organic chemist. In this section, we examine to which extent the information produced by various explanations is chemically relevant. We compare the explanations with either structural objectives of synthesis methods or retrons of transformations corresponding to these methods, keeping in mind that GemsBond has no formal knowledge about chemical transformations or retrons. In this discussion, only formed carbon−carbon bonds of all types are taken into account since they were the focus of the data selection method at the early phase of the data-mining process. A first question to be solved is to determine from which confidence threshold a bond can be considered as effectively formable by organic chemists. To answer this question, compound **1** in Figure 10 was chosen as a prototypical example because it contains cyclic and noncyclic parts and two remote functions without any influence on each other. [We define a *function* as a connected molecular substructure, which comprises exclusively carbon−carbon multiple (including aromatic) bonds, carbon−heteroatom bonds, or heteroatom−heteroatom bonds. An atom is *functional* if it belongs to a function.] The results returned by the application of GemsBond to compound **1** given in Figure 11 can be interpreted as follows:

(1) Most bonds of compound **1** show confidence values lower than 0.28. None of these bonds contains any functional atom, except for the double bond (C21, C22). The explanation graphs related to these bonds are small-sized because GemsBond determines that a larger environment does not

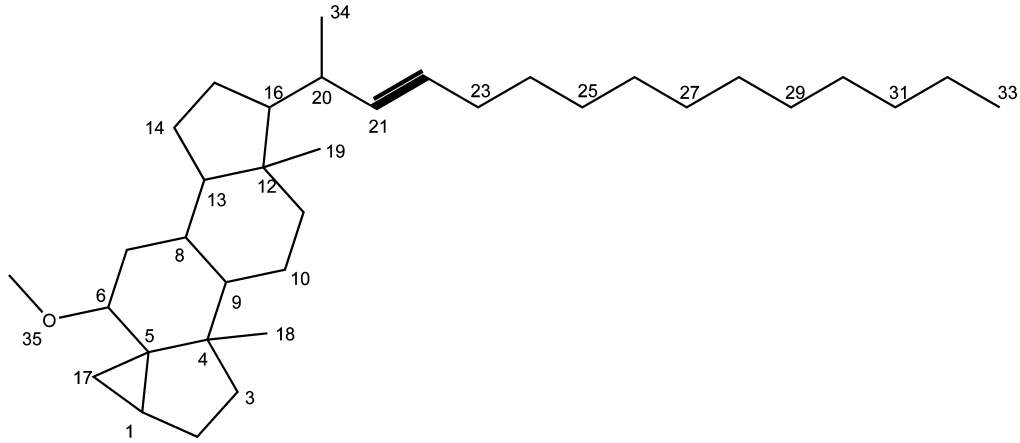



**Figure 10.** Compound **1**. The bold bond specifies the bond formed during the reaction.

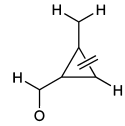
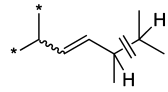
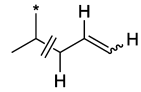
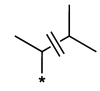
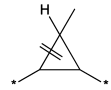
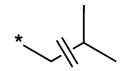
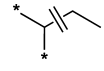
| Bond(s) $b$ | Conf. $\text{conf}_G(b)$ | Explanation $E_{max}$ | Freq. of $E_{max}$ | Bond $b$ | Conf. $\text{conf}_G(b)$ | Explanation $E_{max}$ | Freq. of $E_{max}$ |
|---|---|---|---|---|---|---|---|
| *C*1,*C*17 | 0.78 |  | 12 | *C*7,*C*8<br>*C*15,*C*16 | 0.27 |  | 1182 |
| *C*16,*C*20 | 0.70 |  | 109 | *C*21,*C*22 | 0.25 |  | 670 |
| *C*23,*C*24 | 0.60 |  | 554 | *C*1,*C*2<br>*C*3,*C*4<br>*C*4,*C*5<br>*C*11,*C*12<br>*C*12,*C*13<br>*C*13,*C*14 | 0.24 |  | 1770 |
| *C*5,*C*17 | 0.46 |  | 75 | *C*20,*C*21 | 0.21 |  | 1838 |
| *C*1,*C*5 | 0.39 |  | 111 | *C*2,*C*3<br>*C*10,*C*11<br>*C*14,1*C*5<br>*C*24,*C*25<br>...<br>*C*31,*C*32 | 0.20 |  | 2535 |
| *C*20,*C*34 | 0.35 |  | 1181 | *C*22,*C*23 | 0.17 |  | 691 |
| *C*5,*C*6 | 0.32 |  | 411 | *C*4,*C*18<br>*C*12,*C*19 | 0.16 |  | 2479 |
| *C*6,*C*7 | 0.29 |  | 884 | *C*32,*C*33 | 0.15 |  | 3533 |
| *C*4,*C*9<br>*C*8,*C*9<br>*C*8,*C*13<br>*C*12,*C*16 | 0.28 |  | 506 | | | | |

**Figure 11.** Bonds of compound **1** sorted in decreasing order of confidence. In the explanation columns, the bond the environment refers to is displayed as a formed bond (i.e., with two parallel crossing strokes); a wiggly stroke means the bond is of undefined stereochemistry; the asterisk denotes an atom of any type.

bring any further information to discriminate between these bonds. They have low confidences and high frequencies, and from a retrosynthesis point of view, a domain expert may consider that all these nonfunctional bonds should be preserved rather than disconnected. On the other hand it may be surprising to observe a low confidence for the double bond (C21, C22), which is a very common structural pattern in organic synthesis. First this double bond is far from any other function present in the molecule and such an isolated double bond is generally more difficult to form than an activated double bond. Second, though double bonds are more frequently formed than single bonds (see $N_c/N$ in Table 2), they are less frequent than single bonds (see $N/N_{tot}$ in Table 2). Therefore a direct comparison between their respective confidences is irrelevant. Third, filtering the data set during the selection operation with a minimal yield value of 90% introduces a bias in the results. We verified this point by querying ChemInform and RefLib databases using the graph substructure of the explanation related to double bond (C21, C22) as query substructure. Without any further constraint the proportion of hits obtained with a 90% yield is 3.7%. If the double bond is isolated, that is, without any functional atom included for example in an electron-withdrawing group, the proportion of hits obtained with a 90% yield is 1.1%.

(2) Three bonds of compound **1** have confidence values ranging between 0.60 and 0.78. First, bond (C1, C17) shows the highest confidence, while its related explanation graph contains more specific information, for example, oxygen atom O35, than explanations related to bonds (C5, C17) and (C1, C5), which have lower confidences. GemsBond not only discriminates the environments of the three bonds of the cyclopropyl moiety but also assigns them higher confidences than those of nonfunctional bonds. This result is in agreement with domain knowledge, especially because small-sized ring bonds are more often formed than aliphatic chain bonds. Bonds (C16, C20) and (C23, C24) respectively have a confidence equal to 0.70 and 0.60 and are both located in a $\beta$-position to a double bond. We observe that bonds located in a $\beta$-position to a double bond, even nonfunctional ones, generally have a confidence higher than 0.35. That is, for example, the case of bond (C20, C34) whose related explanation graph is rather simple. We also notice that the higher is the number of atoms or branched atoms present in an explanation, the higher is the confidence of the corresponding bond and the lower is the related frequency. This is illustrated by the difference between the confidences of bonds (C16, C20) and (C23, C24).

(3) The ether function has a weak but significant influence on the confidences of the bonds (C5, C6) and (C6, C7) located in α-position. This α-effect, that is, the increase of the confidence computed for a bond located in α-position to a functional atom, is generally observed as discussed later in this section.

These results lead to a first ranking of formable bonds. Single bonds having confidences higher than 0.60 can be considered as easily formable while those having confidences lower than 0.30 can be considered as hardly formable. Single bonds having confidences ranging from 0.30 to 0.60 are considered as moderately formable. This interpretation may be refined by examining the whole results produced by GemsBond. For each of the 7537 reaction products being considered, at least one bond per product is assigned the best confidence. Table 8 displays the distribution of reaction products as a function of the best confidence. We observe that only 1.7% of reaction products show a best confidence lower than 0.40, while 85.8% show a best confidence higher than 0.60. Furthermore GemsBond generated 312 018 explanations that correspond to the total number of bonds included in the 7537 reaction products and only 5% of these 312 018 bonds have been assigned a confidence higher than 0.60. Thus a confidence threshold of 0.60 seems to be able to discriminate at least one formable bond per target molecule in 85.8% of cases.

**Table 8.** Reaction Product Distribution as a Function of the Best Confidence

| | reaction products | |
|---|---|---|
| best confidence | number | ratio |
| $c < 0.2$ | 45 | 0.6% |
| $0.2 \le c < 0.4$ | 85 | 1.1% |
| $0.4 \le c < 0.6$ | 946 | 12.6% |
| $0.6 \le c < 0.8$ | 1727 | 22.9% |
| $0.8 \le c < 1$ | 2542 | 33.7% |
| $c = 1$ | 2192 | 29.1% |

Having this confidence threshold in mind, let us examine formable bonds to understand their structural environments by studying their related explanations. The main observation lies in variable α- or $\beta$-effects induced by the presence of a functional atom in α- or $\beta$-position to this bond. This is examplified in Figures 12 and 13 for some usual isolated functions. The alcohol function in compound **2**, as well as the ether function in compound **3**, has a weak influence on the formability of α-bonds and no effect on that of $\beta$-bonds (Figure 12). The influence of the amine function in compound **4**, is stronger on the formability of α-bonds which have confidences ranging from 0.41 to 0.44. Compounds **5** and **6**, bearing a thioether and a sulfone function, respectively, show an analogous behavior with slight differences on confidences related to their α-bonds. The halogen functions present in compounds **7**, **8**, and **9**, lead to medium-range α- and $\beta$-effects. If the chloro and the bromo compounds present similar confidences, the confidence of the α-bond of compound **9** (conf = 0.15) is clearly lower than those corresponding to α-bonds of compounds **7** and **8**. Moreover the explanation related to this α-bond does not display any iodine atom and is identical to the explanation related to the C32−C33 of compound **1**. Indeed, such primary iodo compounds are usually prepared by radical or nucleophilic substitution or by addition on a double bond, both processes that do not form any α-bond.

In Figure 13, more examples are given to study how confidence is modified by functional groups containing multiple bonds. In most cases a bond located in $\beta$-position to a multiple bond obtains a higher confidence than most bonds located in α-position. For example, all compounds **10**−**13** show the same explanation for the $\beta$-bond whatever the nature of the carbonyl or carboxyl group. Among these groups, the aldehyde function is the unique function allowing a confidence increase of its α-bond, a fact consistent with domain knowledge because this function is frequently introduced through a one-carbon chain elongation. If we consider now the aromatic ring of compound **14**, first a significant $\beta$-effect (conf = 0.50) is observed, second all

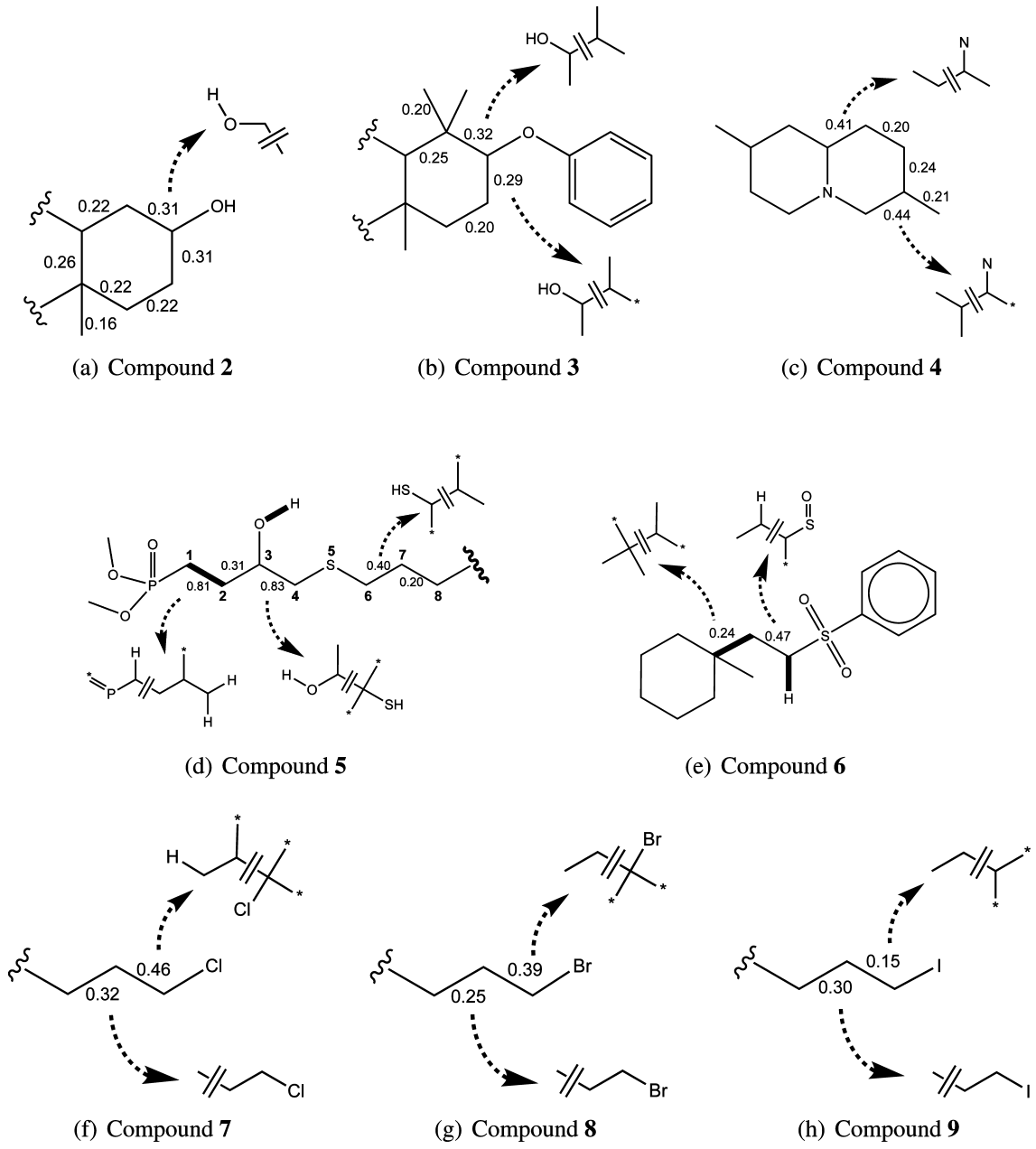
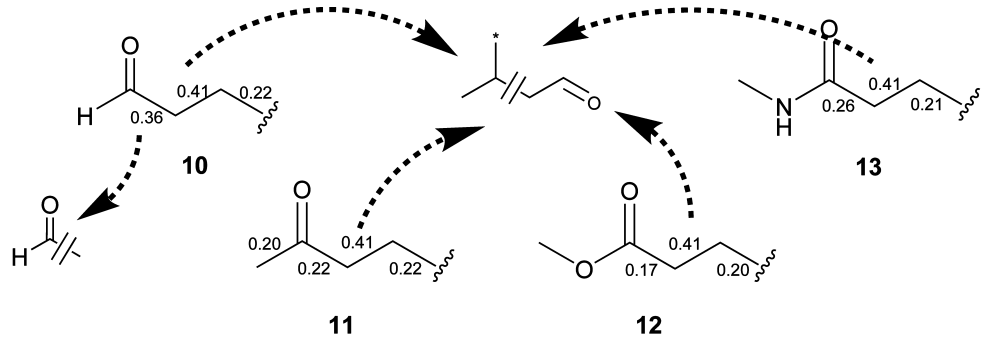


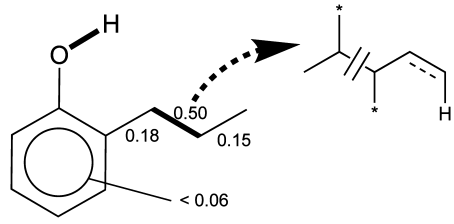

**Figure 12.** Meaningful confidences and explanations related to compounds **2**−**9**. A wiggle stroke means the graph is a substructure of the studied molecule; bold bonds specify the bonds formed during the reaction.

aromatic bonds show a very low confidence. This seems consistent with the fact that aromatic bonds are very rarely formed during a synthesis. Imine, nitrile, and acetylenic functions of derivatives **15**, **16**, and **17**, respectively, show medium-range effects on both α- and $\beta$-bonds. We observe that the $\beta$-bond confidences of **15** does noticeably vary depending on the position of the imine nitrogen atom. The two bonds located in $\beta$-position to the nitrile function of **16** have different environments and consequently show different confidences (conf = 0.51, conf = 0.64). The confidence of the α-bond to the nitrile function is also medium-range revealing a partial formability character as in the case of aldehyde **10**. Finally, the acetylenic function of **17** has a higher influence on the confidence of the α-bond than on the one of the $\beta$-bond; this information is consistent with the frequent use of synthesis methods forming such carbon−carbon bonds. The acetylenic bond, which is usually rarely formed, shows here a low confidence as in the case of the aromatic bonds of **14**. More generally, the explanations related to bonds in compounds **2**−**17** provide interesting pieces of information as they highlight the presence of a function responsible for either an α- or a $\beta$-effect. But no $\gamma$-effect is detected for these compounds. On the other hand, except for the easily formable bonds located in $\beta$-position to double bonds (conf $\geq$ 0, 60), most formable bonds of compounds **2**−**17** have medium-range confidence values.

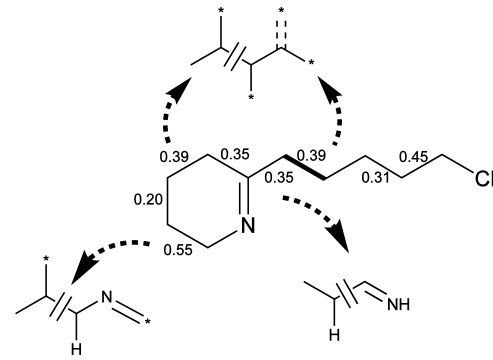
If we examine now the compounds possessing at least one bond having a confidence higher than 0.60, that is, about 86% of studied compounds, we notice that formability of such a bond very often depends on conjugated α- and $\beta$-effects. Some representative examples are given in Figure 14. In compound **18**, bond (C3, C4) shows a confidence equal to 0.89, a noticeably higher value than previously observed for a bond in α-position to an alcohol function or in $\beta$-position to an alkene function, in compounds **2** and **5** in Figure 12. The explanation related to this bond actually



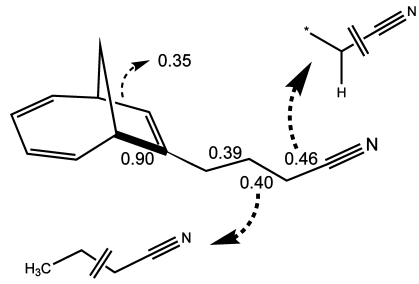

(a) Compounds **10** to **13**



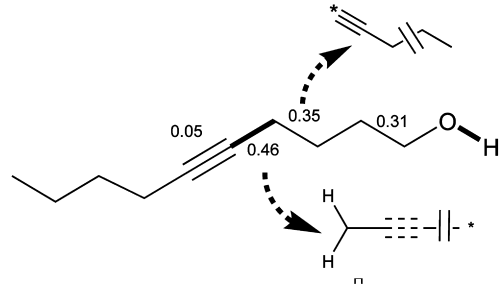

(b) Compound **14**




(c) Compound **15**




(d) Compound **16**




(e) Compound **17**

**Figure 13.** Meaningful confidences and explanations related to compounds **10**−**17**. Dotted lines represent aromatic bonds; a wiggle stroke means the graph is a substructure of the studied molecule; bold bonds specify the bonds formed during the reaction.

includes both alcohol and alkene functions, and involves the retron of the transformation corresponding to addition of an allyl organometallic species to a carbonyl derivative. We observe comparable values for bonds (C2−C3) of compounds **19** and **20**, whose related explanation graphs involve retrons of transformations corresponding to addition of an ester enolate to an aldehyde and to addition of a ketone enolate to an imine, respectively. In the case of compound **21**, three bonds have a confidence higher than 0.60: (C2, C3), (C1, C3′), and (C2′, C3′). The highest confidence reached by bond (C1, C3′) is not surprising because connecting a carbonyl group to an aromatic cycle is a very standard process. Actually the related explanation includes both carbonyl and aromatic functions. From an organic synthesis point of view, it is understandable to retrieve double bond (C2, C3) among the formable bonds since it may result from a classic olefination reaction. More surprising is the computed confidence for the bond (C1, C1′) of compound **22** because no α- or $\beta$-effect justifies such a high value (conf = 1.00) by comparison with the corresponding bond of **18** and **19** (conf = 0.18). The related explanation graph contains both alkene and phenyl functions and a noticeably high number of atoms and bonds, and thus, it has a low frequency (freq = 8). By contrast, bonds (C1, C2) and (C3, C4) show lower confidences (conf = 0.50, conf = 0.60), but their related explanation graphs (freq = 866, freq = 554) contain a single function. Finally, five distinct bonds of compound **23** have a confidence higher than 0.60 and can thus be considered as formable bonds. The two best confidences (conf = 0.74, conf = 0.75) concern the bonds, (C7, C8) and (C9, C10), respectively, incident to the alkyne function. These two bonds show higher confidences than the one observed for the analogous bond of compound **16** (conf = 0.46) because of the inclusion of the neighboring double bond in both explanations. In terms of organic synthesis, these confidences are in agreement with known Csp−Csp$^2$ cross-coupling methods.

Among the synthesis methods enabling the formation of six-membered rings, Diels−Alder reaction is probably the

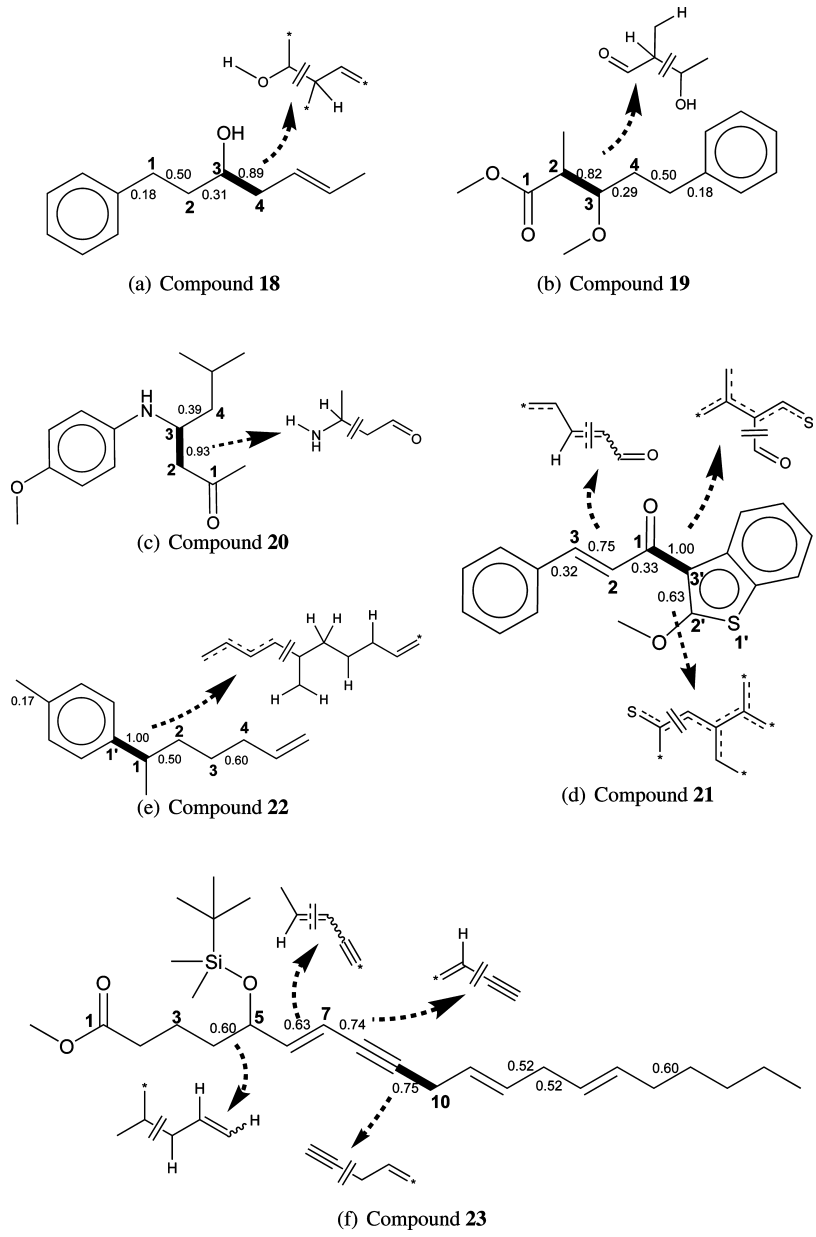



**Figure 14.** Meaningful confidences and explanations related to compounds **18**−**23**. Dotted lines represent aromatic bonds; bold bonds specify the bonds formed during the reaction.

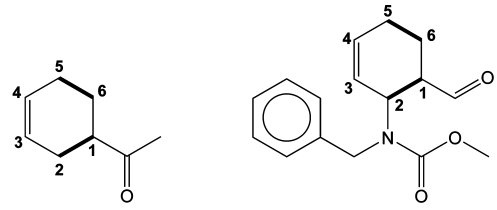
most famous. We studied therefore reaction products **24**−**28**, resulting from this synthesis method (Figure 15), and especially the confidences of bonds (C1, C2) and (C5, C6), which are formed during the reaction.

We noticed in each case high confidences for these two bonds. The explanations related to the formation of bond (C1, C2) include, for each studied compound, both alkene and carbonyl functions that constitute a part of the expected favorable environment for a Diels−Alder reaction. For bond (C5, C6), the explanations in the case of compounds **24**−**26** are identical to the explanation related to a bond located in a $\beta$-position to a double bond, see compounds **5**, **22**, **23**, or the bonds (C2, C7) and (C5, C7) of compounds **26** and **27**. Interestingly, in the case of compounds **27** and **28**, the explanations include the retron of the Diels−Alder transformation and especially the six-membered ring. Recall that GemsBond has no a priori chemical knowledge about chemical transformations and synthesis methods and that it computes a confidence for each bond without considering the simultaneous formation of two bonds as in the case of cycloaddition reaction products. This point must be considered in future developments of our approach that already discriminates the two formable bonds involved in this cycloaddition.

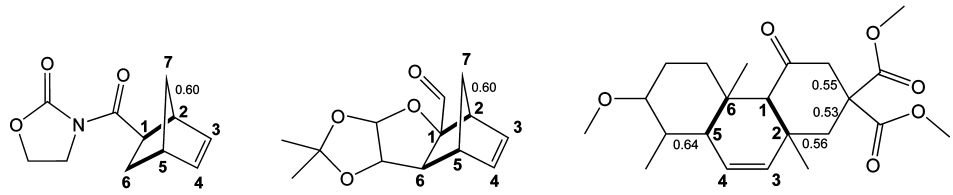
Briefly, GemsBond recognizes at least one formable bond in 86% of studied bonds. It produces an explanation including in many cases the retron of a known transformation, as well as the functions that constitute a necessary environment for the bond to be formed. This study only considers carbon−carbon formed bonds and should be extended to carbon−heteroatom bonds or heteroatom−heteroatom bonds to take into account the vast domain of heterocyclic compounds.

## RELATED WORK

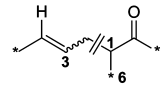
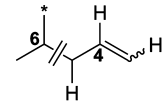
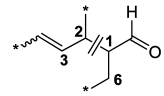
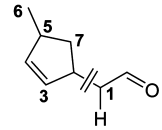
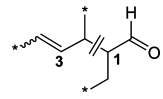
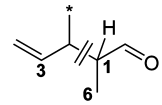
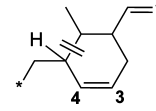
Discovery of formable bonds can be seen as a special treatment of the problem of discovering strategic bonds. The fact of being a strategic bond is rather a fuzzy notion that



(a) Compound **24** (b) Compound **25** (c) Diels-Alder reaction retron



(d) Compound **26** (e) Compound **27** (f) Compound **28**

| Compound number | $(C1,C2)$ bond Confidence | Explanation | $(C5,C6)$ bond Confidence | Explanation |
|---|---|---|---|---|
| 24 | 0.69 |  | 0.60 |  |
| 25 | 1.00 |  | as above | |
| 26 | 1.00 |  | as above | |
| 27 | 1.00 |  | 0.97 |  |
| 28 | 0.81 |  | 0.87 |  |

(g) Explanations of the different bonds $(C1,C2)$ and $(C5,C6)$

**Figure 15.** Meaningful confidences and explanations related to compounds **24**−**28**. Dotted lines represent aromatic bonds; bold bonds specify the bonds formed during the reaction.

may lead to different formal definitions, and we focus on recent approaches dealing with retrosynthetically important bonds. Apart from Bertz's[11] and Hanessian's[13] studies cited in the introduction that deal with the topological and stereochemical strategic disconnections, two other formal approaches were recently developed to find retrosynthetically important bonds.[35,36] The former proposes a new index, called molecular centrality, based on the concept of convergent synthesis and computed from quadratic shortest path distances.[35] This index is mainly topological and proves to be moderately useful in case of highly functionalized molecules. The latter can be perceived as an improvement of the former as it includes the molecular centrality and two additional parameters: the bond dissociation energy and the number of chiral centers.[36] From these three parameters and through a logistic regression analysis, authors introduce a new scoring function assessing the retrosynthetic importance of bonds by studying reaction centers from two available reaction databases. Though the results are rather satisfactory, only one bond is considered as the analysis output whereas synthetic chemists often consider several bonds as retrosynthetically interesting according to the molecule functionality.

A problem very close to bond formability is the evaluation of synthetic accessibility of a compound set,[37] although this problem was rarely approached through a retrosynthesis-based assessment. Two programs are noticeable in this domain: LCOLI[38] derived from LHASA and CAESA developed in Johnson's group.[39] CAESA applies retrosynthesis-based rules to a molecule set and compares the resulting fragments to available starting materials and to a fragment database. These fragments derive from generalized structures from functional groups. The synthetic accessibility

of a target was also considered in Gasteiger's group from a retrosynthesis point of view.[40] They developed a scoring function based on four criteria: topological complexity indexes, a stereochemical complexity index, comparison to structural patterns from a starting material database, and computation of a property called retrosynthetic reaction fitness. This property consists in disconnecting the target at strategic bonds that belong to the extended reaction centers, that is, that include neighboring α-atoms. The resulting patterns are then compared to those extracted from the Theilheimer reaction database. This process is supported by the key idea that a high proportion of reference patterns found in the target enables to suppose that the probability of retrieving an existing synthesis is also high.

Finally, to help chemists in building global synthesis plans, a retrosynthesis package was designed recently and jointly by Johnson's group and Pfizer company.[41] To this purpose, a rule database is built by clustering elemental reactions from Beilstein Crossfire (http://www.crossfirebeilstein.com/) or Accelrys MOS (http://accelrys.com/) reaction databases. Then a rule-based and exhaustive retrosynthesis is performed on the target to sort the proposed routes according to their ease of access and to their structural analogies with starting materials. The authors tested this web-based software, marketed by the SymBioSys Inc. company (http://www.simbiosys.ca/archem/) during a short period. This program appears to be very interesting from an education point of view but lacks important disconnections for common rings, such as the seven-membered rings, or for heterocyclic rings.

From a computer science perspective, the notion of graph frequency used by GemsBond originates in recent development of graph-mining methods,[18−21] whose purpose is to produce every graph pattern with a frequency (i.e., the number of examples the pattern occurs in) larger than a threshold. Those methods have found applications in chemistry, in particular to predict molecule activity.[28,42,43] For instance, Borgelt and Berthold[42] search for molecular substructures that are simultaneously frequent in a set of positively classified molecules and unfrequent in a set of negatively classified compounds. These discriminating substructures can then be reused as features to classify a molecule according to some activity. However and contrasting with those algorithms, GemsBond does not address the problem of classifying molecules but the classification of bonds (or possibly atoms) and uses reaction databases instead of molecule data sets. Another major difference is that GemsBond is not a complete algorithm searching for all frequent patterns but is rather based on a heuristic transductive search. In that sense, GemsBond is related to Subdue algorithm[22] that also heuristically searches for graph patterns in a set of graphs.

To the best of authors' knowledge, only one method, namely, CNN,[14,15] has previously attempted to learn formable bonds from examples. CNN determines formable bonds from their environments by processing a set of molecular graphs where examples of formable bonds are known. However the way the problem is addressed is different: CNN uses a sophisticated inductive learning method computing maximal common subgraphs (MCSGs) intersecting the input graph and the examples, and then iterates the process over those MCSGs. Finally these intersecting environments vote for the current hypothesis according to a complex polling algorithm. The main limitation of CNN is scalability because (i) CNN makes an intensive use of costly MCSGs computation (that is a NP-hard problem) and (ii) the complexity of CNN is quadratic with the number $n$ of examples (whereas the complexity of GemsBond is in $O(n)$). In terms of prediction accuracy, comparison is more difficult. The only reported results of CNN are in concern with a jack-knife test over the 694 carbon−carbon single bonds contained in 75 examples of molecules. Moreover atoms and bonds of the examples have been manually annotated by additional information like inductive and mesomeric effects. Accuracy is slightly better than for GemsBond but definitive conclusions cannot be drawn since test conditions are very different. The higher scalability of GemsBond makes this algorithm a good candidate to mine large and diverse reaction databases.

## CONCLUSION AND PERSPECTIVES

The graph-mining approach used for GemsBond provides an accurate, fast, and scalable method for evaluating the formability of bonds in molecules. GemsBond may, as such, be a valuable tool to help synthesis planning and virtual screening. Many perspectives emerge from this work. On the side of computer science, the search algorithm can be improved to become more accurate without compromising speed and scalability. In particular, heuristics might be improved to take into account environments that are unfavorable for a bond to be formable. On the side of chemoinformatics, an effort must be undertaken to better understand how the choice of examples might influence the quality of the results. Even if GemsBond is scalable comparatively to previous methods, it remains clear that GemsBond is not able to process the millions of reactions available in databases. It is thus important to define a methodology for building a restricted set of few thousands reactions that are representative of various and useful synthesis methods. The integration of higher level knowledge in molecular graphs, for example, by decomposing molecular graphs into functional groups or by enriching the description of atoms and bonds, might also improve the quality of outputs produced by GemsBond. This should not only increase accuracy but also provide more meaningful explanations, closer to environments chemists would have provided. Finally GemsBond could be in the future applied to other problems of classification of bonds or atoms, by adapting heuristics to these problems.

## ACKNOWLEDGMENT

We gratefully acknowledge partial financial support provided by Centre National de la Recherche Scientifique (CNRS) through a PEPS interdisciplinary project funding.

## REFERENCES AND NOTES

(1) Corey, E.; Long, A.; Rubenstein, S. Computer-Assisted Analysis in Organic Synthesis. *Science* **1985**, *228*, 408–418.
(2) Corey, E.; Cheng, X. *The Logic of Chemical Synthesis*; John Wiley & Sons: New York, 1989.
(3) Smit, W.; Bochkov, A.; Caple, R. *Organic Synthesis: The Science Behind the Art*; Royal Society of Chemistry: Cambridge, U.K., 1998.
(4) Vleduts, G. E. Concerning one System of Classification and Codification of Organic Reactions. *Inf. Storage Retr.* **1963**, *1*, 117–146.
(5) Corey, E. Computer-Assisted Analysis of Complex Synthetic Problems. *Q. Rev. Chem. Soc.* **1971**, *25*, 455–482.

(6) Pfoertner, M.; Sitzmann, M. Computer-Assisted Synthesis Design by WODCA (CASD). In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; pp 1457−1507.
(7) Ott, M. Cheminformatics and Organic Chemistry. Computer-Assisted Synthetic Analysis. *Cheminformatics* **2004**, *1*, 83–109.
(8) Hanessian, S. Man, Machine and Visual Imagery in Strategic Synthesis Planning: Computer-Perceived Precursors for Drug Candidates. *Curr. Opin. Drug Discovery Dev.* **2005**, *8*, 798–819.
(9) Chen, W. L. Chemoinformatics: Past, Present, and Future. *J. Chem. Inf. Model.* **2006**, *46*, 2230–2255.
(10) Gund, P.; Grabowski, E. J. J.; Hoff, D. R.; Smith, G. M.; Andose, J. D.; Rhodes, J. B.; Wipke, W. T. Computer-Assisted Synthetic Analysis at Merck. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 88–93.
(11) Bertz, S. H.; Sommer, T. J. Rigorous Mathematical Approaches to Strategic Bonds and Synthetic Analysis Based on Conceptually Simple New Complexity Indexes. *Chem. Commun.* **1997**, *24*, 2409–2410.
(12) Ruecker, C.; Ruecker, G.; Bertz, S. H. Organic Synthesis—Art or Science. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 378–386.
(13) Hanessian, S. Total Synthesis of Natural Products: The"Chiron"Approach. In *Organic Chemistry Series*; Pergamon Press: Oxford, U.K., 1983; Vol. 3.
(14) Régin, J.-C.; Gascuel, O.; Laurenço, C. Machine learning of strategic knowledge in organic synthesis from reaction databases. *Proceedings of the First European Conference on Computational Chemistry*; (ECCC-1), Nancy, France, May 23−27; AIP Press: Woodbury, NY, 1995; pp 618−623.
(15) Régin, J.-C. Ph.D. thesis, Université des Sciences et Techniques du Languedoc, Montpellier, 1995.
(16) Pennerath, F.; Polaillon, G.; Napoli, A. A Method for Classifying Vertices of Labeled Graphs Applied to Knowledge Discovery from Molecules. *Proceedings of the 18th European Conference on Artificial Intelligence*; (ECAI'08), Patras, Greece; IOS Press: Amsterdam, 2008; pp 147−151.
(17) Vapnik, V. N. *Statistical Learning Theory*; Wiley: New York, 1998.
(18) Inokuchi, A.; Washio, T.; Motoda, H. An a Priori-Based Algorithm for Mining Frequent Substructures from Graph Data. *PKDD*; Springer: Heidelberg, Germany, 2000; pp 13−23.
(19) Kuramochi, M.; Karypis, G. Frequent Subgraph Discovery. *Proceedings of the 2001 IEEE International Conference on Data Mining*, 29 Nov−2 Dec 2001, San Jose, CA; IEEE Computer Society: Los Alamitos, CA, 2001; pp 313−320.
(20) Yan, X.; Han, J. gSpan: Graph-Based Substructure Pattern Mining. *ICDM*; IEEE Computer Society: Los Alamitos, CA, 2002; pp 721−724.
(21) Nijssen, S.; Kok, J. N. A Quickstart in Frequent Structure Mining Can Make a Difference. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Seattle, WA, August 22−25, 2004; ACM Press: New York, 2004; pp 647−652.
(22) Cook, D. J.; Holder, L. B. Substructure Discovery Using Minimum Description Length and Background Knowledge. *J. Artif. Intell. Res.* **1994**, *1*, 231–255.
(23) Mahé, P.; Ueda, N.; Akutsu, T.; Perret, J. L.; Vert, J. P. Graph Kernels for Molecular Structure-Activity Relationship Analysis with Support Vector Machines. *J. Chem. Inf. Model.* **2005**, *45*, 939–951.
(24) Fröhlich, H.; Wegner, J. K.; Sieker, F.; Zell, A. Optimal Assignment Kernels for Attributed Molecular Graphs. *Machine Learning*; Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7−11, 2005; ACM Press: New York, 2005; pp 225−232.
(25) Mitchell, T. M. Version Spaces: A Candidate Elimination Approach to Rule Learning. *Proceedings of the 5th International Joint Conference on Artificial Intelligence*; Cambridge, MA, August 1977; William Kaufmann: San Mateo, CA, 1977; pp 305−310.
(26) Ganter, B.; Grigoriev, P. A.; Kuznetsov, S. O.; Samokhin, M. V. Concept-Based Data Mining with Scaled Labeled Graphs. *ICCS*; Springer: Heidelberg, Germany, 2004; pp 94−108.
(27) Muggleton, S.; Srinivasan, A.; King, R. D.; Sternberg, M. J. E. Biochemical Knowledge Discovery Using Inductive Logic Programming. *DS '98: Proceedings of the First International Conference on Discovery Science*; Springer: Heidelberg, Germany, 1998; pp 326−341.
(28) Helma, C.; Cramer, T.; Kramer, S.; Raedt, L. D. Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1402–1411.
(29) Jauffret, P.; Hanser, T.; Tonnelier, C.; Kaufmann, G. Machine Learning of Generic Reactions: I) Scope of the Project. *Tetrahedron Comput. Methodol.* **1990**, *3*, 323–333.
(30) Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; Verkamo, A. I. Fast Discovery of Association Rules. In *Advances in Knowledge Discovery and Data Mining*; AAAI/MIT Press: Menlo Park, CA, 1996; pp 307−328.
(31) Russell, S. J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, 2002.
(32) Pennerath, F.; Polaillon, G.; Napoli, A. Prétraitement des bases de données de réactions chimiques pour la fouille de schémas de réactions. *Proceedings of the 8th Conference on Extraction et gestion des connaissances (EGC'2008)*; Sophia-Antipolis, France, January 2008; 2008; Editions Cépaduès: Toulouse, France, pp 547−558.
(33) Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.
(34) van Rijsbergen, C. J. *Information Retrieval*, 2nd ed.; Butterworths: London, 1979.
(35) Tanaka, A.; Kawaia, T.; Fujiia, M.; Matsumotoa, T.; Takabatakea, T.; Okamotob, H.; Funatsu, K. Molecular Centrality for Synthetic Design of Convergent Reactions. *Tetrahedron* **2008**, *64*, 4602–4612.
(36) Tanaka, A.; Kawai, T.; Matsumoto, T.; Fujii, M.; Takabatake, T.; Okamoto, H.; Funatsu, K. Construction of a Statistical Evaluation Model Based on Molecular Centrality to Find Retrosynthetically Important Bonds in Organic Compounds. *Eur. J. Org. Chem.* **2008**, *2008*, 5995–6007.
(37) Baber, J. C.; Feher, M. Predicting Synthetic Accessibility: Application in Drug Discovery and Development. *Mini-Rev. Med. Chem.* **2004**, *4*, 681–692.
(38) Long, A.; Chen, R.; Marby, C. A.; Sukharevsky, A. P.; Ohm, K. From Synthesis Planning to Combinatorial Chemistry: Applications of the Lhasa Suite, Presented at 228th ACS National Meeting, Philadelphia, PA, August 22−26, 2004; http://acscinf.org/docs/meetings/228nm/presentations/228nm52.pdf (accessed Dec 17, 2009).
(39) Gillet, V.; Myatt, G.; Zsoldos, Z.; Johnson, P. SPROUT, HIPPO, and CAESA: Tools for De Novo Structure Generation and Estimation of Synthetic Accessibility. *Perspect. Drug Discovery Des.* **1995**, *3*, 34–50.
(40) Boda, K.; Seidel, T.; Gasteiger, J. Structure and Reaction Based Evaluation of Synthetic Accessibility. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 311–325.
(41) 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*, 593–602.
(42) Borgelt, C.; Berthold, M. R., Mining Molecular Fragments: Finding Relevant Substructures of Molecules. *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM '02)*; 9−12 December 2002, Maebashi City, Japan; IEEE Computer Society: Los Alamitos, CA, 2002; pp 51−58.
(43) Fischer, I.; Meinl, T., Graph Based Molecular Data Mining—An overview. *Proceedings of the IEEE International Conference on Systems, Man & Cybernetics*; The Hague, Netherlands, 10−13 October 2004; IEEE Computer Society: Los Alamitos, CA, 2004; pp 4578−4582.

CI9003909