JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Expert System for Predicting Reaction Conditions: The Michael Reaction Case

G. Marcou,*[,†] J. Aires de Sousa,[‡] D. A. R. S. Latino,[‡] A. de Luca,[†] D. Horvath,[†] V. Rietsch,[§] and A. Varnek*[,†,∥]

[†]Laboratory of Chemoinformatics, University of Strasbourg, 1 rue B. Pascal, 67000 Strasbourg, France

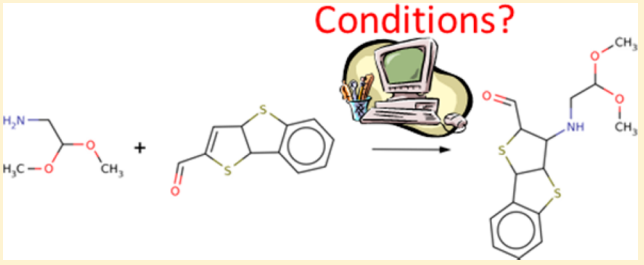
[‡]Departamento de Química and REQUIMTE, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa 2829-516 Caparica, Portugal

[§]eNovalys, 1 Rue Jean Sapidus, Batiment Pythagore, 67400 Illkirch, France

[∥]Laboratory of Chemoinformatics, Butlerov Institute of Chemistry, Kazan Federal University, Kazan 420008, Russia

Supporting Information

ABSTRACT: A generic chemical transformation may often be achieved under various synthetic conditions. However, for any specific reagents, only one or a few among the reported synthetic protocols may be successful. For example, Michael β-addition reactions may proceed under different choices of solvent (e.g., hydrophobic, aprotic polar, protic) and catalyst (e.g., Brønsted acid, Lewis acid, Lewis base, etc.). Chemoinformatics methods could be efficiently used to establish a relationship between the reagent structures and the required reaction conditions, which would allow synthetic chemists to waste less time and resources in trying out various protocols in search for the appropriate one. In order to address this problem, a number of 2-classes classification models have been built on a set of 198 Michael reactions retrieved from literature. Trained models discriminate between processes that are compatible and respectively processes not feasible under a specific reaction condition option (feasible or not with a Lewis acid catalyst, feasible or not in hydrophobic solvent, etc.). Eight distinct models were built to decide the compatibility of a Michael addition process with each considered reaction condition option, while a ninth model was aimed to predict whether the assumed Michael addition is feasible at all. Different machine-learning methods (Support Vector Machine, Naive Bayes, and Random Forest) in combination with different types of descriptors (ISIDA fragments issued from Condensed Graphs of Reactions, MOLMAP, Electronic Effect Descriptors, and Chemistry Development Kit computed descriptors) have been used. Models have good predictive performance in 3-fold cross-validation done three times: balanced accuracy varies from 0.7 to 1. Developed models are available for the users at http://infochim.u-strasbg.fr/webserv/VSEngine.html. Eventually, these were challenged to predict feasibility conditions for ~50 novel Michael reactions from the eNovalys database (originally from patent literature).



1. INTRODUCTION

Rationalizing chemical reactivity patterns in terms of the structure of the involved partners (reagent, solvent) represents the key goal of physical organic chemistry[1] and has major implications in all the branches of chemistry and biochemistry, from synthesis to the understanding of metabolic processes. However, computer-aided handling and, moreover, modeling of chemical processes[2] is a significantly more difficult task than storage and property predictions of individual molecules. This is due to the fact that reactions are involving several partners (not only the reagents, but also solvent—or, even worse, mixtures thereof—and catalysts) and "dynamical" breaking or forming bonds—not to mention transition states characterized by partially broken/formed bonds—which do not naturally fit the graph-based "atom=vertex/bond=edge" paradigm of molecular representation in chemoinformatics. Nevertheless, in view of the extraordinary importance of the problem, dedicated computational approaches were developed—see the above-cited review[2] for an exhaustive view of the literature in the field. In particular, computational modeling of reactivity is still a major challenge.

A large domain of physical organic chemistry deals with kinetic parameter estimations through the use of Hammet–Taft[3−5] and related linear free-energy relationships. For example, early works[6−8] applied this method to the reactivity of α,β-unsaturated compounds (mostly acrylonitrile) in the context of Michael-type reactions. Other approaches were developed. For example, topological descriptors[9] or quantum chemical descriptors[10] were related to reactivity parameters of acrylate monomers. The polymerization rate of (meth)acrylate monomers was investigated using both topological and

Received: November 23, 2014
Published: January 14, 2015

DOI: 10.1021/ci500698a
J. Chem. Inf. Model. 2015, 55, 239−250

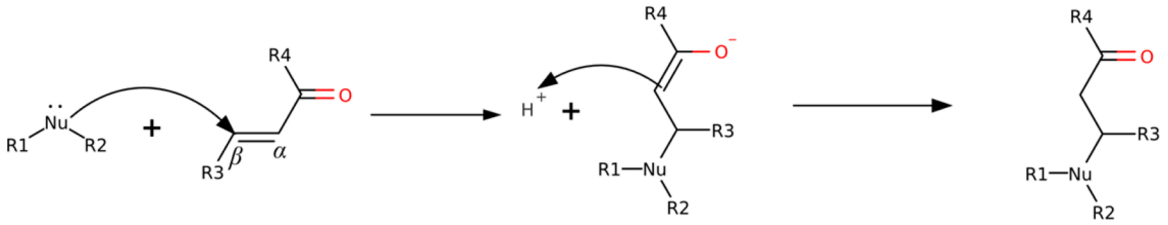


Figure 1. Mechanism of Michael addition or conjugate (1,4) addition—the addition of a nucleophile (Nu), the Michael donor, to the β carbon of an α,β-unsaturated carbonyl, the Michael acceptor, resulting in a product termed the Michael adduct.

quantum descriptors[11] and multilinear regression (MLR). The kinetic rate was also the target of a study using molecular descriptors of the resonance stabilization and of the steric hindrance of 66 Michael acceptors and MLR.[12] Wondrousch[13] developed site-specific quantum chemical parameters for quantifying the energy change associated with the gain or loss of electronic charge and applied the new parameters to predict the rates of reaction of a set of 31 α,β-unsaturated carbonyl compounds toward glutathione as a model nucleophile. More specifically, the two research groups involved in this collaborative work have each provided major contributions to the chemoinformatics of chemical reactions.

The MOLMAP[14] (molecular mapping of atom-level properties) reaction descriptor was developed for numerically encoding the transformations resulting from a chemical reaction. The method is based on the mapping of the chemical bonds on a self-organizing map (SOM)—the mapping of the chemical bonds of a molecule provides a pattern of the types of bonds available in that molecule. The reaction MOLMAP is obtained by the difference between the MOLMAPs of the products and the MOLMAPs of the reactants, which represents the bond changing in the reaction. This method, which does not require explicit assignment of the reaction center and avoids atom-to-atom mapping, is based on topological and physicochemical molecular features—related, in principle, to the reaction mechanism. The MOLMAP approach was applied to the genome-scale classification of enzymatic reactions[15−17] and automatic perception of chemical similarities between metabolic pathways.[18]

Latest, the development of Condensed Graphs of Reaction[19−21] (CGR) allowed chemical processes to be rendered as unique graphs, with special status given to forming and breaking bonds. The automation of the conversion process of regular reaction representations into CGRs involves the delicate stage of automated mapping, i.e. linking atoms in products to their correspondences in the reagent molecules.[22] This technology then allowed a generalization of the popular ISIDA fragment[23−26] count descriptor generator to handle CGRs in addition to usual molecular graphs and thus produce reaction-specific descriptors that were successfully applied to address the problem of automated recognition of reaction classes.[19]

The goal of the present paper is to apply these in-house developed reaction management tools to a further, and so-far rarely addressed challenge: prediction of condition-dependent reaction feasibility. Reaction conditions were previously considered as input parameters of a model, as for instance the quantitative structure−conditions−property relationship of the acid hydrolysis of esters.[27] A recent example[28] of reaction rate modeling, including reaction conditions as part of the model parameters, concerned the rate constant of bimolecular nucleophilic substitution reactions with neutral nucleophiles in various solvents. For this work they introduced a set of solvent descriptors to use together with reaction descriptors.

Often, a generic chemical transformation, corresponding to a given pattern of formed and broken bonds and labeled by a same name in organic chemistry may actually proceed under widely heterogeneous reaction conditions—sometimes via different mechanisms. Therefore, there may be no homogeneous recipe on how to perform such a process—required conditions (foremost catalyst and solvent) may vary from one to the other end of the spectrum of possible choices. However, the large domain of possible conditions of the "generic reaction" as perceived in "paper chemistry" does not mean that any particular representative thereof, based on specific reagents, could indistinctly be performed under either of the textbook-reported protocols. Discovering the appropriate synthesis protocol for a given process, out of the possible protocols known for that process class, is a time-consuming undertaking, often relying on the chemist's know-how and flair.

A perfect illustration of this situation is provided by Michael addition processes. The Michael addition considered in this article consists in the β-addition of a nucleophile to an olefin activated by a directly bonded carbonyl (Figure 1). The reaction is also known as a conjugate addition,[29] and this particular generalization of the concept of the reaction is sometimes termed Michael-type additions.[30] The nucleophile, the electron-withdrawing group activating the olefin and the product are referred to as the Michael donor (MD), the Michael acceptor (MA), and the Michael adduct (MAd), respectively.

The reaction typically requires a basic catalyst in order to ionize the MD and improve the nucleophilicity and protic solvents such as methanol or ethanol to promote a rapid proton transfer to the anionic intermediate formed after the β-addition, leading to the MAd.[30] However, the reaction occurs also with acid catalysis. In this case, the acid coordinates the carbonyl of the MA and helps activating the olefin. Besides, aprotic solvents could also achieve high yields when they could better solubilize the MD, the MA, and the catalyst, precipitate the MAd, or inhibit side reactions. Concerning the specific MD concerned by this work, the ionization step is needed for thiols to produce thiolates in basic conditions which are the active species; however, amines are generally good nucleophiles themselves and need not to be ionized.

A Michael addition is often a crucial step of industrial processes for introducing enantioselectivity to the MAd,[31] for one-pot synthesis of complex structures through cyclization reactions[32] and polymer production.[30] Michael additions, and its reverse β eliminations, are also commonly found in biochemistry, e.g. in the metabolic pathways of fatty acids, catalyzed by the enoyl-CoA hydratase (EC 4.2.1.17),[33] or in the alkylation of biological macromolecules by sesquiterpene lactones.[34] Another example is the nucleophilic addition of

DOI: 10.1021/ci500698a
J. Chem. Inf. Model. 2015, 55, 239−250

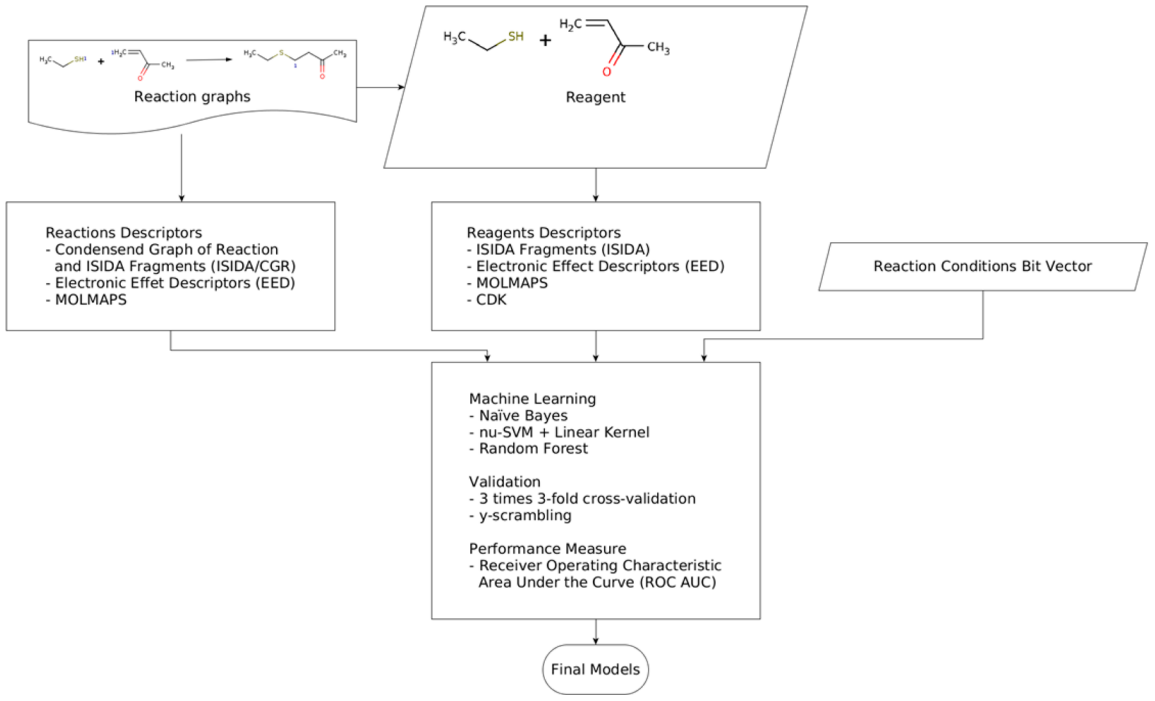


Figure 2. Workflow of the benchmark study. Molecular descriptors are either computed on reagents or on a representation of the reaction (either CGR or MOLMAP). Then several machine-learning methods are applied. Models built on a combination of a DS and a machine-learning method are compared based on their ROC AUC estimated by cross-validation.

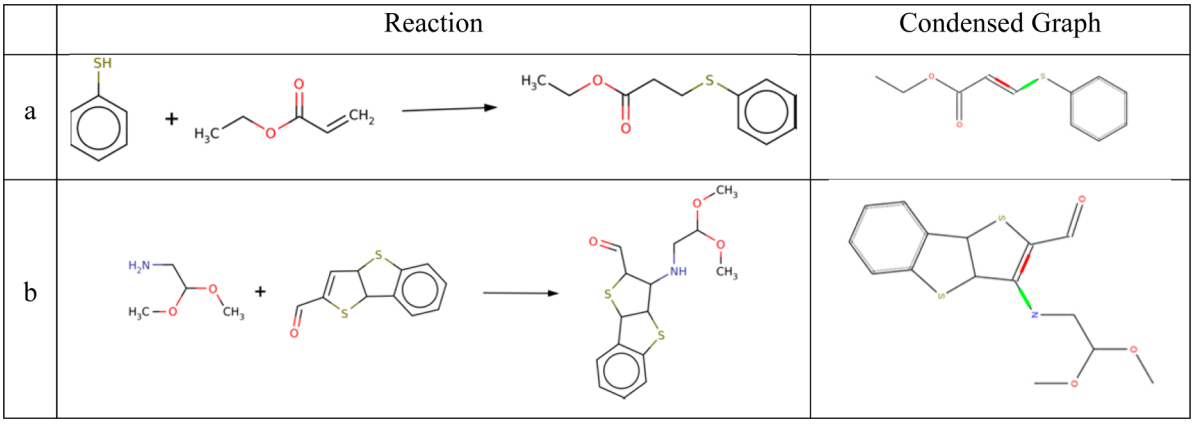


Figure 3. Instance of (a) occurring Michael addition with a basic catalysis in hydrophobic solvent and (b) a counterexample, a not occurring Michael addition. Reactions are represented as CGR: a red bond indicates a bond order decrease and a green bond indicates bond formation.

protein side chains to Michael acceptor xenobiotic as the triggering step of toxicological mechanisms (e.g., skin sensitization).[35]

Overall, three aspects characterize the Michael addition: (i) the diversity of the reaction conditions that can favor it, (ii) the wide scope of the reaction, and (iii) the interesting atom economy [ratio of the molecular mass of the product over the total molecular mass of the reactants] of the reaction. The first and third points are favorable from the industrial point of view: Michael addition is often a plausible solution in order to find green reaction conditions, tolerate a broad range of functional groups, and benefit from high reaction rates and conversion yields.[30] This versatility may also be a problem. For instance the Michael addition accounts for the propensity of undesired polymerization when the MAd remains a potent MA.

Yet, the reaction conditions can be drastically modulated in terms of solvent and catalyst, thus a rationalization of these conditions would help the efficient design of reaction protocols.

The aim of the herein developed QSRR (quantitative structure−reactivity relationships) models is to predict the status (occurring or not) of a Michael addition process with respect to each of the considered catalyst and solvent classes. A number of 2-classes classification models, each representing a particular reaction condition, have been built on a set of 198 thio- and aza-Michael reactions retrieved from the literature. Different machine-learning methods (Support Vector Machine, Naive Bayes, and Random Forest) in combination with different types of descriptors (ISIDA fragments issued from Condensed Graphs of Reactions, MOLMAP,[15] Electronic Effect Descriptors,[36] and Chemistry Development Kit[37]) have been used. Obtained models were investigated in terms of

DOI: 10.1021/ci500698a
J. Chem. Inf. Model. 2015, 55, 239−250

predictive performance in 3-fold cross-validation done three times.

An expert QSRR system has been eventually built: one can draw a putative Michael addition and apply the models. The result is a reaction condition profile, a set of binary values: 1 if the associated reaction condition is expected to favor the reaction and 0 otherwise. This tool was challenged to predict feasibility conditions for more than 50 additional Michael processes, previously imported from patent literature into in the reaction database of eNovalyst and not used for model training.

2. METHODS

The methodological aspects to be described here (see workflow in Figure 2) include data preparation, the various representations employed and the therefrom-derived reaction descriptors and, eventually, the applied machine-learning method.

2.1. Data Preparation. The set of 222 thio- and aza-Michael addition processes retrieved from literature, include an actual set of 198 reactions reported as chemically feasible. The remaining 24 *counterexample* Michael processes do not actually occur, because of predominance of the competing on-carbonyl addition reaction, followed by elimination (the nucleophile being either a primary amine or hydroxylamine, leading to Schiff base/oxime formation).

In other words, a *counterexample* (Figure 3b) example of Michael reaction is a β-addition process conceivable on paper, but too slow to actually occur against the competing carbonyl addition, irrespectively of conditions. These instances are rendered in the data set as Michael additions (as in Figure 3a) but marked to represent a not occurring reaction. Note that a yet different class of counterexamples may exist—unsaturated carbonyl/nucleophile pairs too deactivated to react either way, all reaction condition combinations being exhausted. Unfortunately, this training set does not include such examples, as rigorous negative results like these are very difficult to find and certify.

The actually feasible 198 reactions were classified both with respect to the nature of the required catalyst and to the nature of the required solvent. Reaction temperature and yield were not systematically reported, and therefore not subjected to specific scrutiny in this work. For the catalyst, the distinction is made between Brønsted acids (BA), Lewis acids (LA, including heavy metal ions), basic catalysis (B), and the absence of explicit catalyst (presumed autocatalysis, NA). Solvents are classified into hydrophobic (H), aprotic polar (A), and protic (P), with an extra class for processes not requiring a solvent (NA). The distribution of data across the different possible reaction conditions is summarized in Table 1.

Table 1. Number of Positive and (Presumed) Negative Instances Per Premise

	positive	negative
Solv:A	53	169
Solv:NA	52	170
Solv:P	103	119
Cat:LA	93	129
Cat:NA	61	161
Solv:H	40	182
Cat:BA	57	165
Cat:B	45	177
NO	25	198

Some of the processes were however reported to run under several (catalyst, solvent) scenarios, so there is no simple one-to-one assignment of a process into a (catalyst, solvent) pair of classes. Processes reported to occur under various (catalyst, solvent) conditions are simply more robust toward the reaction conditions. For example, a process running under both Brønsted and Lewis acid conditions requires some form of acid catalysis, while a reaction that may be performed either with Brønsted acids or with bases is even less specific.

A matrix like Table 2 can be constructed, featuring columns corresponding to catalyst and solvent conditions for each of the reactions reported on lines. The cells report the feasibility status of the process with respect to conditions: 1—feasible in those conditions, 0—not feasible (or, more precisely, never reported as feasible). The profile-based modeling strategy amounts to build, for each column, an independent model predicting this status. Each model will answer a punctual question such as "Is this process feasible with Brønsted acid catalysts?", "Is this process feasible in aprotic polar solvents?", etc.

Table 2. Relationship between Reaction Profiles and Multi-Class Classification[a]

	Catalyst					Solvent			Counterexample
Process ID	BA	LA	B	NA	H	A	P	NA	NO
1	1	0	0	0	0	0	1	0	0
2	0	1	0	0	0	1	0	0	0
3	0	0	1	0	0	0	0	1	0
4	0	0	0	1	1	0	0	0	0
5	1	1	0	0	0	1	1	0	0
6	1	0	1	0	1	0	1	0	0
7	1	1	1	0	0	1	0	1	0
8	1	0	0	1	0	1	1	1	0
9	0	0	0	0	0	0	0	0	1

[a]Brønsted acid, Lewis acid, basic catalyst, and no catalyst are labeled "Cat:BA", "Cat:LA", "Cat:B", and "Cat:NA", respectively. Hydrophobic, aprotic, protic solvent, and no solvent are labeled "Solv:H", "Solv:A", "Solv:P", and "Solv:NA", respectively. "NO" stands for counterexample processes (not occurring reactions: empty profiles).

2.2. Reaction Encoding Schemes. We used three encoding setups: (i) reagent-based, (ii) product-based and, when meaningful, (iii) a reaction-based setup. These various molecular and reaction descriptors are benchmarked in terms of their feasibility prediction propensities. In the first case, molecular descriptors are computed on the reagents only. The reaction-based setup is context-dependent: within the Condensed Graph approach, "reaction" descriptors are descriptors of the condensed graph, the pseudomolecular object formally representing the reaction. Reaction MOLMAPs represent changes of SOM landscapes of products vs reagents. For key-atom-based descriptors such as EED (vide infra for all these technical details) reaction descriptors represent a concatenation of reagent and product descriptor vectors. Since CDK terms are classical whole-molecule descriptors and do not include any information about key reaction centers, the "reaction-based" CDK description makes little sense and was skipped.

Each set of molecular descriptors will be referred to as a **Descriptor Space (DS)**.

Condensed Graphs of Reaction. Condensed Graphs of Reaction (CGR) represent a reaction by means of a fused "hypermolecular" graph regrouping all involved atoms and

DOI: 10.1021/ci500698a
J. Chem. Inf. Model. 2015, 55, 239−250

bonds, where a special dynamic bond status is assigned to key bonds broken, created, or of changing order.[21] In the case of Michael additions, the CGR can be straightforwardly obtained from the product graph, by altering the status of the nucleophile-β-carbon single bond to *create* and the one of the former carbon−carbon double bond to *double to single order change* (as exemplified in Figure 3). CGRs can thus be manipulated as any classical molecular entity.

CDK Descriptors. This class of molecular descriptors is used as a reference. They are computed on the reagents only, ignoring the special status of the nucleophilic and electrophilic reaction centers. 420 CDK descriptors were computed including physicochemical estimations like the logP, the polarizability, the mass, the charge; various atoms and bonds counts, BCUTs, WHIM, descriptors combining topological surface area and partial charge information, eccentricity, Kier and Hall, etc.

ISIDA Fragment Descriptors. The ISIDA/Fragmentor software was used to generate 566 classes of ISIDA fragment descriptors (basically colored fragment counts) from the CGRs. Naming of the different fragmentation classes is standardized, such as to be self-explanatory of the specific parameters of the associated fragmentation procedure. A fragment class name is a concatenation of codes for the fragment topology (sequence, circular fragments, etc.), coloration type (atom symbol, pharmacophore type, etc), bond order information inclusion, the user-defined fragment size ranges (in parentheses), counting type and options:[38] TopologicalFragmentationColorationType-BondInclusion-(LowerLength-UpperLength)-CountingType_Options

The fragment topology code is a roman number and corresponds to the following fragmentation: sequences (I), atom-centered fragments (II), and triplets (III). The code for coloration type is a chain of letters starting with a capital and followed by only lower case letters. In this work, only coloration based on the atom symbol (A) is used. The bond inclusion code (B) simply indicates the inclusion of bond orders in the string. The number of atoms to be included at minimum and maximum into a fragment is encoded by a range between a pair of parentheses. The counting type corresponds to the type of weight used to count the occurrences of fragments. In this work no weighting scheme was used. Finally, some options were considered: using atom pairs (P) and/or fragments restricted to those only containing dynamical bonds (OD). For instance IIAB(3-5)P-OD indicates an atom-centered fragmentation scheme including at least 3 and up to 5 atoms away from the center of the fragment, using atom pairs and including at least a dynamic bond.

In the following, R-ISIDA refers to ISIDA fragments of CGRs and Rg-ISIDA refer to concatenated vectors of ISIDA fragment counts in reagents (electrophile + nucleophile). The latter are, like CDK terms, plain molecular descriptors of the reagents, where reactive groups are not being given any special status.

MOLMAP Descriptors. MOLMAP descriptors (molecular map of atom-level properties) were developed to represent the chemical bonds existing in a molecule.[14] A Kohonen self-organizing map[39,40] (SOM) is trained with a diversity of bonds that are distributed over a 2D surface of neurons, according to similarities in their features. Topological and physicochemical bond features were used. The pattern of neurons activated by the bonds of a molecule is a representation of that molecule. In a chemical reaction, the difference between the MOLMAPs of the products and the MOLMAPs of the reactants was proposed to represent the structural changes occurring in the reaction.[14] It is a numerical fixed length code of a reaction that can be further processed by machine-learning methods.

SOMs with toroidal topology (and sizes 15 × 15, 20 × 20, and 25 × 25) were used in the experiments presented here. A linear decreasing triangular scaling function was used in a 50 cycle training with an initial learning rate of 0.1 and an initial learning span of half the size of the map. The learning span and the learning rate linearly decreased until zero. The winning neuron was selected using the minimum Euclidean distance between the input vector and the neuron weights. SOMs were implemented with in-house developed software derived from the JATOON Java applets.[41] Topological and physicochemical features represented the chemical bonds as described elsewhere.[15] The SOM was trained with all the chemical bonds of reactants and products in the data set.

Electronic Effect Descriptors. Electronic Effect Descriptors (EED) are atom-centric sums of topological distance-weighted properties of atoms described in detail elsewhere[36] and represent a generalization of previously published work.[42] The atom center(s) on which EED calculation focuses must be user-defined such as to match relevant reaction center(s), while the surrounding atoms contribute to different descriptor terms proportionally to their property (partial, sigma and pi charges, polarizability, electronegativity, nucleophilic energy index, electrophilic energy index, charge density index, hybridization index, and formal charge) associated with each term. The proportionality factor depends on the topological distance of the contributor to the center, such that remote atoms impact less. Some terms of the EED vector only stem from contributors within paths of alternating single/double or aromatic bonds, specifically aimed at capturing potential resonance effects. The EED vector characterizing the reaction center(s) of a molecule has 704 elements. For the Michael product, the EED descriptor focuses on the β-carbon as a reference atom (P-EED).

However, the reaction is a multicomponent object in which each protagonist needs to be encoded. In terms of reagents, the EED were centered (1) on the nucleophilic atom and (2) on the β-carbon of the electrophilic α,β-unsaturated carbonyl compound, respectively. The reaction can thus be fully described by the concatenation of the three EED vectors—in arbitrary, but preserved order: nucleophilic reagent + electrophilic reagent + product. This 3 × 704 = 2112-dimensional reaction vector spans the Global EED reaction space (R-EED). The reaction space concatenating the two reagent EED vectors (Rg-EED) is 1408-dimensional.

2.3. Modeling Strategies. This section briefly describes the modeling algorithm employed here.

Random Forest.[43] Each forest was composed of 500 random trees combined by bagging. At each node, m variables are selected at random out of the full set and the best split on these m variables is used to split the node (m was used with the default value of $\log_2$(total number of variables) + 1). Each tree was fully developed without pruning. Models were evaluated using 3-fold cross-validation performed three times using the ROC AUC as a performance measure. The Weka package[44] was used for implementing Random Forests. For each reaction condition, all ISIDA DS were iteratively scanned. The DS that performed significantly worse than the best one, according to a student t test with 95% confidence, were discarded. The

DOI: 10.1021/ci500698a
J. Chem. Inf. Model. 2015, 55, 239−250

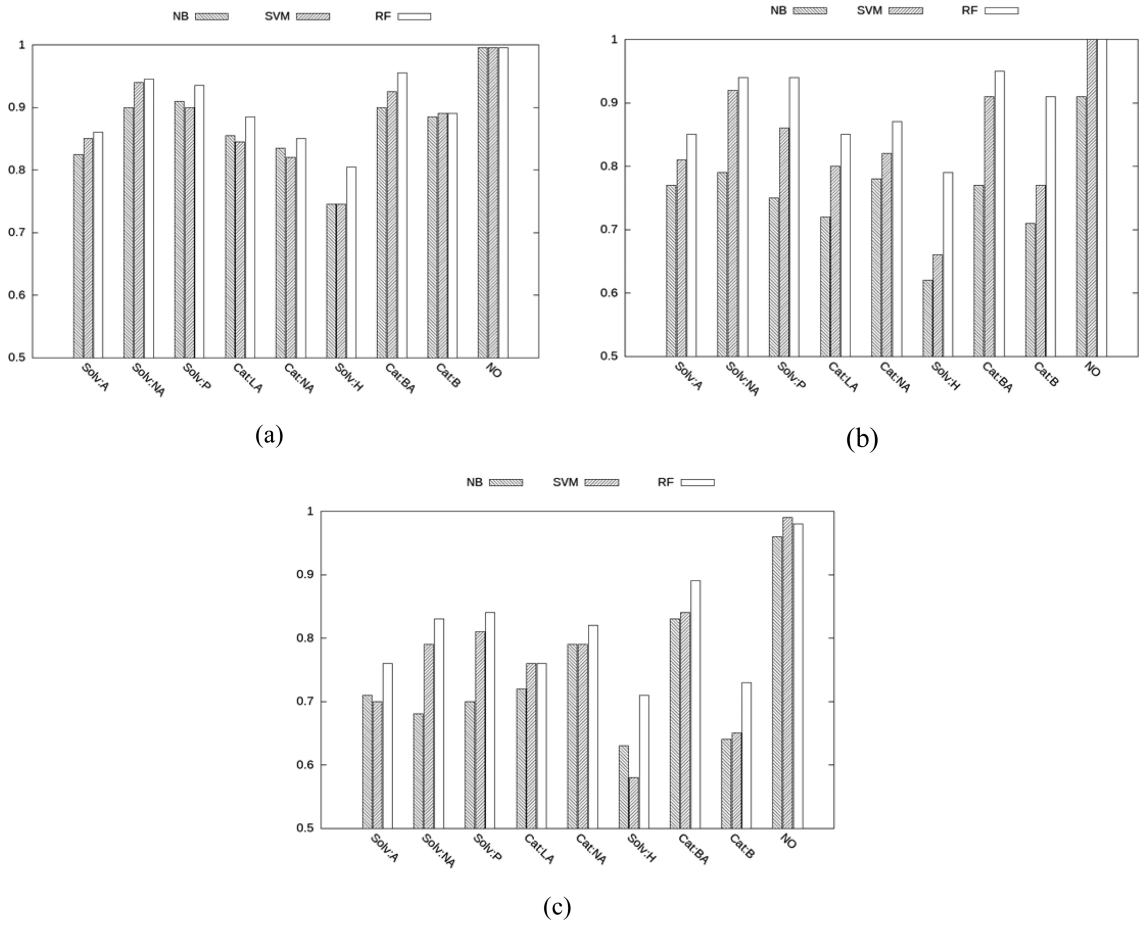


Figure 4. Average of the ROC AUC measured in 3-fold cross-validation performed three times for models built using (a) CGR based ISIDA fragment descriptors, reaction based (b) EED, and (c) MOLMAPS descriptors. ν-SVM, Random Forest, or Naive Bayes were used as machine-learning method.

comparisons were performed on the average of nine ROC AUC obtained by 3-fold cross-validation performed three times.

Support Vector Machine Classification. Support Vector Machine[45−47] (SVM) models were comprehensively produced with the Weka software package for all DS and validated using three repetitions of 3-fold cross-validation. The ν-SVM algorithm was used with linear kernel exclusively. The parameter ν was scanned from 0 to 1 by step of 0.025 and chosen based on the performances (ROC AUC) of the model through a nested 3-fold cross-validation. Therefore, in the main cross-validation loop, the test set was not used for tuning the parameter. The final value of the ν parameter was chosen as the median of the recorded optimal values during the validation stage.

Finally, ISIDA DS were ranked according to their performances for each modeling task. Performances were measured as the average of the nine ROC AUC values obtained using 3-fold cross-validation done three times. For each modeling task, a DS was kept if the performances of the corresponding model were not significantly worse from those of the best model according to a student t test with a confidence of 95%.

Naive Bayes.[48,49] In this setup we used a log odds ratio score. For the ith reaction described by a set of descriptors $\{x^i\}$, the score $S(C|\{x^i\})$ compares the Bayesian estimate of the log-likelihood that the reaction belong to the class C rather than to alternative class, $\overline{C}$ (the complement of C):

$$S(C|\{x^i\}) = \log(P(C)) + \sum_j \log(P(x_j^i|C)) - \log(P(\overline{C})) - \sum_j \log(P(x_j^i|\overline{C})) \quad (1)$$

$P(C)$ (respectively $P(\overline{C})$) is the a priori probability of the class C (respectively $\overline{C}$) estimated by its frequency. The terms $P(x_j^i|C)$ and $P(x_j^i|\overline{C})$, respectively, are estimates of the probability that the descriptor j takes the value x_j^i as observed in the ith reaction considering that the reaction is of class C (respectively $\overline{C}$).

The probability estimates are based on different assumptions depending on the nature of the descriptors considered. In the case of ISIDA descriptors, the x_j^i are counts of particular fragments. Therefore, a descriptor value is assumed to follow a binomial law. This is analogue to the Multinomial Naïve Bayes algorithm.[50] The total number of fragments observed in a CGR, n, is identified as the trial number parameter of the distribution. The success rate parameter of the distribution, p_j^C (respectively $p_j^{\overline{C}}$), is estimated as the observed frequency of the fragment j in the population of reactions of class C (respectively $\overline{C}$). These estimates are eventually modified using a Laplacian smoothing.

In the case of EEDs or CDK, the descriptors are taking continuous values and the estimates are based on a normal distribution assumption with mean and variance computed for

DOI: 10.1021/ci500698a
J. Chem. Inf. Model. 2015, 55, 239−250

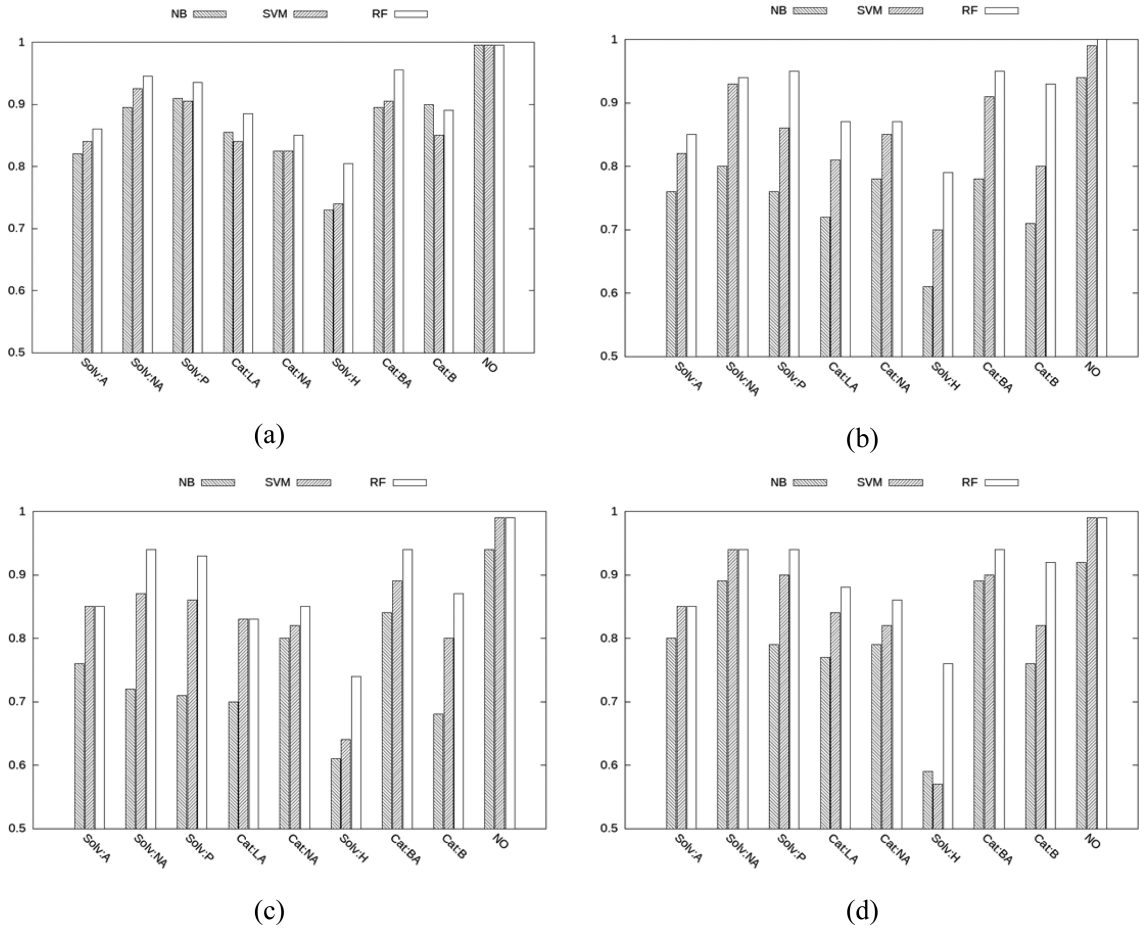


Figure 5. Average of the ROC AUC measured in 3-fold cross-validation performed three times for models built using (a) reagent-based ISIDA fragment descriptors, (b) reagent-based EED, (c) MOLMAPS, and (d) CDK descriptors. ν-SVM, Random Forest, or Naive Bayes were used as the machine-learning method.

each descriptor j on the subsets of reactions of class C (respectively $\overline{C}$).

Therefore, for each reaction condition premise and for each DS, a Naive Bayes model is built. All models are validated using 3-fold cross-validation three times. The success of the models is measured using the ROC AUC (ROC area under the curve) obtained during the validation. Thus, for each modeling setup, nine ROC AUC are available. All DS are ranked according to the ROC AUC. For each reaction condition, a DS is considered relevant for modeling if the observed distribution of ROC AUC cannot be statistically demonstrated to be worse than the best one using a student t test with 95% confidence.

2.4. Model Validation. A model is a combination of a chosen DS and a parametrized algorithm (PA) operating on this DS. Cross-validation allows the selection of the best model according to a more robust success criterion than fitting quality and, thus, to control overfit. For each premise, each DS and each modeling algorithm, 90 y-scrambling experiments were performed.[51] Each model based on the randomized data set was built and cross-validated following the same procedure as for nonrandomized data. The collected cross-validated ROC AUC were used to compute the upper bound of an interval covering 99% of the distribution of the scrambled ROC AUC. A model is considered as not validated if its cross-validation ROC AUC is below this bound for nonrandomized data.

Eventually, details about the external validation using the web-based expert system will be directly reported in the Results section below.

3. RESULTS

3.1. Model Performance As a Function of Descriptors and Learning Methods. The performances of the benchmark of different descriptors are summarized in Figures 4 and 5. For ISIDA descriptors, the statistics reported in the tables in the Supporting Information give the range of performances for those DS that were selected as described in section 2.3. The number of DS found appropriate to model the respective feasibilities with the RF method varied from 8 to 83, depending of the targeted reaction condition. When SVM was used, 16 to 164 DS were kept, while 7 to 50 DS "survived" the t-test relevance challenge when Naive Bayes was the employed machine-learning tool. The average ROC AUC measured on those DS is reported in bar charts.

The first observation is that some feasibility challenges are significantly "easier" than others (as measured by associated average ROC AUC scores). The easiest task is to detect counterexamples (NO)—the nonoccurring Michael additions. The second easiest tasks are modeling the Brønsted acid catalysis (Cat:BA), the protic solvent (Sol:P), and the absence of solvent (Sol:NA) feasibilities. Modeling the aprotic solvent (Sol:A), Lewis acid catalyst (Cat:LA), absence of catalyst (Cat:NA), and basic catalyst (Cat:B) conditions were rather

DOI: 10.1021/ci500698a
J. Chem. Inf. Model. 2015, 55, 239−250

challenging, and the most difficult task was to model the hydrophobic solvent (Sol:H). Concerning the feasibility in hydrophobic solvent, models were of rather poor quality, sometimes at the limit of significance when compared to scrambled data (see the Supporting Information).

The good cross-validation propensity of the counterexample predictors is not surprising, because of the limited set of counterexamples, all conforming to a same few, obvious structural "clichés". Because of such bias, these models have great statistics but are not very useful. They recognize the specific patterns, such as a selenium atom (Figure 6) or the

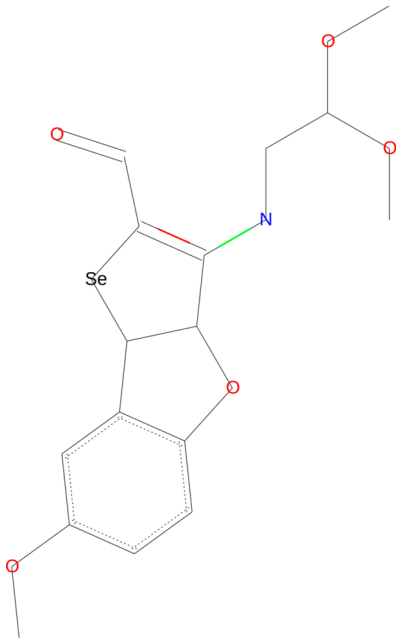


Figure 6. Representative CGR of a nonoccurring Michael reaction.

unsaturated aldehyde with a heavily substituted, sterically hindered, often halogenated β carbon. Therefore, they would likely fail as a general feasibility model, predicting if a presumed Michael addition can be eventually realized or not, if confronted with a diverse set of counterexamples.

Another observation is that the performances of the models relative to the reaction condition are rather independent of the modeling procedure. Nevertheless, some weak trends can be evidenced. In general the Naive Bayes models are the least performing and the Random Forest are the best, sometimes challenged by some SVM models. Yet, the Naive Bayes models have no parametrization and are merely representative of the statistic contents of the data set. Beside they are by far the fastest models to train and to apply. Thus, they are ideal candidates for real-time applications.

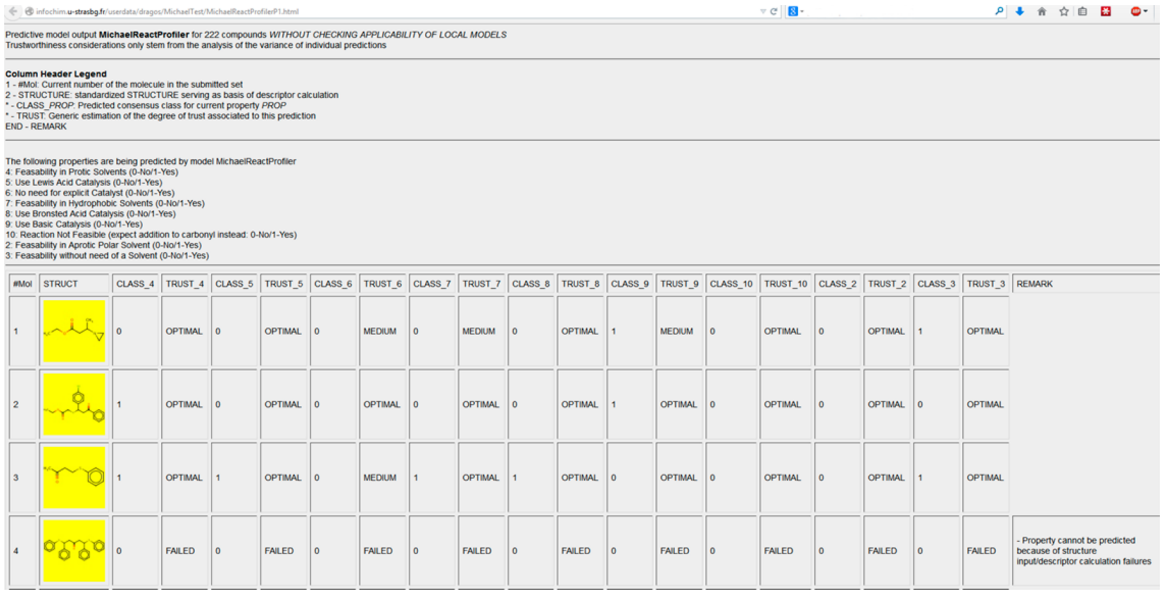
The dependence on the chosen DS is more complex. First, MOLMAPS computed on the reagents are much more efficient than those computed on the reaction itself (Figures 4c and 5c). Reaction MOLMAPS are obtained from the difference of products and reagents MOLMAPS, which erases information on bonds adjacent to the reaction center and beyond it. These are expected to be relevant here, particularly because the reaction center is the same across the whole data set. Except for MOLMAPS, the reagent based DS yield equivalent results to the reaction based DS, especially for Random Forest models. EED descriptors (Figures 4b and 5b) computed on the entire reaction are largely redundant with those computed on the reagents only (a subset of the former), so it was expected that the two DS perform similarly. But this was unexpected for ISIDA descriptors computed on reactions, based on the CGR technology and hence clearly marking the key atoms of the transformation (Figure 4a), and ISIDA descriptors computed on the reagents, with no highlighting of key atoms (Figure 5a). The latter count fragments also covering atoms and bonds outside the reaction center. The excellent performances of the CDK (Figure 5d) and reagent based ISIDA (Figure 5a) DS signals that, within the considered training set, reaction-condition specific structural "patterns" or "signatures" can be established even in absence of explicit knowledge of the reaction center itself.

Some more unexpected observations were made. For instance, since thiols are rather acidic and thiolates are good nucleophiles, it was expected to observe thio-Michael additions frequently in the presence of a basic catalyst. This is not the case in the current state of the database, and no preference is given to sulfur for Michael additions occurring in basic conditions. For aza donors, the impact of acidic catalysis is nuanced—a certain concentration of acid may favor the activation of the carbonyl, but excess thereof might completely consume the nucleophilic amine, by protonation. These models cannot mimic such complex mechanistic effects: though it might be expected that acidic catalysts would favor aza-Michael additions, the present data and the models do not support it. The aza-Michael additions of this training set are most often reported to take place in aprotic solvents.

Finally, it is difficult to interpret results of Michael additions occurring in hydrophobic conditions. After the counter-examples, they are the least represented situation in the data set, the most diverse and their feasibility predictors are of poor quality. The models tend to disfavor polar groups and are sensitive to nonspecific patterns such as aromatic rings or fatty chains—a meaningful trend, since the prerequisite of feasibility in hydrophobic solvents is the solubility of reagents therein.

3.2. Web-Based Expert System for Estimating Michael Addition Reaction Condition Profile. The web-based prediction server http://infochim.u-strasbg.fr/webserv/VSEngine.html required significant technical development in order to support reaction feasibility predictions in a context so-far dedicated to individual compound properties. The system expects, upon input, to be provided with one or several putative Michael addition product structures, drawn as regular molecules. No user-provided marking of key atoms is required. First, the compounds undergo preprocessing and stand-ardization (salt strip-of, conversion to major tautomeric form, split-charge rendering of nitroxides, etc), like any compound submitted to single-molecule QSPR predictions (see available models on the server). However, the Michael addition feasibility predictor (click "MichaelReactProfiler") operates differently from standard molecule property predictors. First, it calls a ChemAxon API-based java substructure search tool in order to detect, and mark, the key atom (the carbon atom representing the β addition center) in the product molecules. It is recognized by matching against a SMARTS[52] query *[#7, #16][C:2]CC =, #[N, O] designing a carbon directly connected to a nitrogen or sulfur (these being the only two classes of nucleophilic partners covered by the training set) and situated at two saturated bonds from a carbonyl (C═O), imine (C═N), or nitrile (C#N) carbon. The addition center herewith receives the label "2". The hereby marked product

DOI: 10.1021/ci500698a
J. Chem. Inf. Model. 2015, 55, 239−250


Predictive model output **MichaelReactProfiler** for 222 compounds *WITHOUT CHECKING APPLICABILITY OF LOCAL MODELS*
Trustworthiness considerations only stem from the analysis of the variance of individual predictions

Column Header Legend
1 - #Mol: Current number of the molecule in the submitted set
2 - STRUCTURE: standardized STRUCTURE serving as basis of descriptor calculation
* - CLASS_*PROP*: Predicted consensus class for current property *PROP*
* - TRUST: Generic estimation of the degree of trust associated to this prediction
END - REMARK

The following properties are being predicted by model MichaelReactProfiler
4: Feasability in Protic Solvents (0-No/1-Yes)
5: Use Lewis Acid Catalysis (0-No/1-Yes)
6: No need for explicit Catalyst (0-No/1-Yes)
7: Feasability in Hydrophobic Solvents (0-No/1-Yes)
8: Use Bronsted Acid Catalysis (0-No/1-Yes)
9: Use Basic Catalysis (0-No/1-Yes)
10: Reaction Not Feasible (expect addition to carbonyl instead: 0-No/1-Yes)
2: Feasability in Aprotic Polar Solvent (0-No/1-Yes)
3: Feasability without need of a Solvent (0-No/1-Yes)

#Mol	STRUCT	CLASS_4	TRUST_4	CLASS_5	TRUST_5	CLASS_6	TRUST_6	CLASS_7	TRUST_7	CLASS_8	TRUST_8	CLASS_9	TRUST_9	CLASS_10	TRUST_10	CLASS_2	TRUST_2	CLASS_3	TRUST_3	REMARK
1		0	OPTIMAL	0	OPTIMAL	0	MEDIUM	0	MEDIUM	0	OPTIMAL	1	MEDIUM	0	OPTIMAL	0	OPTIMAL	1	OPTIMAL	
2		1	OPTIMAL	0	OPTIMAL	0	OPTIMAL	0	OPTIMAL	0	OPTIMAL	1	OPTIMAL	0	OPTIMAL	0	OPTIMAL	0	OPTIMAL	
3		1	OPTIMAL	1	OPTIMAL	0	MEDIUM	1	OPTIMAL	1	OPTIMAL	0	OPTIMAL	0	OPTIMAL	0	OPTIMAL	1	OPTIMAL	
4		0	FAILED	0	FAILED	0	FAILED	0	FAILED	0	FAILED	0	FAILED	0	FAILED	0	FAILED	0	FAILED	- Property cannot be predicted because of structure input/descriptor calculation failures

Figure 7. Sample output of the web-based expert system predictor of Michael addition feasibility.

structures can be used to generate EED terms focusing on the marked atoms, as provided for the training compounds and thus required for prediction by the herein developed models. Next, the nucleophilic center is recognized and labeled as atom "1" using a dedicated SMARTS definition: *[#7, #16:1][C:2]-CC =, #[N, O]. Eventually, the retrosynthesis generating the reagents from the product is simulated using ChemAxon's Standardizer[53] driven by an appropriate SMIRKS[54] description of the desired transformation, i.e. breaking of the single bond between the two marked atoms, and conversion of the single bond adjacent to the carbonyl group into a double one: *[#7, #16:1][C:2]CC =, #[N, O] ≫ [#7, #16:1] − *.[C:2]=CC =, #[N, O]. This produces separated and already marked output files of the nucleophilic and electrophilic reagents, ready to use for input to the calculation of EED descriptors. Should this virtual retrosynthetic step fail due to structural ambiguities—as, for example, in compounds that are products of more than one putative Michael additions, involving different moieties—then the particular item will be discarded from the list of predictions (more precisely, it will be formally replaced by a water molecule with a prediction status of FAILED, in order to ensure that the prediction output file follows, conversion errors notwithstanding, the order of items entered by the user). Each description scenario using EED terms (EEDs of products, of reagents, and of the reaction, respectively) will then be exploited. For each item, its corresponding descriptors are input to the associated SVM model predicting feasibility associated with each of considered reaction conditions. Therefore, feasibility for each reaction condition will be estimated thrice—based on product, reagents, and reaction EEDs—and the predominant "vote" will be output, associated with a trust level of OPTIMAL (all three predictions agree on the current feasibility status) or MEDIUM (current feasibility status adopted by a 2:1 majority). So far, the predictor tool does not include any additional applicability domain (AD) assessment,[55] since the success of the preprocessing and retrosynthesis phase already represents, per se, a partial check that the input product is eligible to be predicted by this tool. However, since the options for both *Solvent* and *Catalyst* are rather comprehensive, we suggest, as an additional trust criterion, to discard prediction in which feasibility is granted with respect to one or more solvent categories, but to none of the catalysts, and vice versa. Indeed, suppose that a reaction is predicted to be "feasible under Lewis catalysis", but, on the other hand, it is also predicted to be unfeasible in either protic, or aprotic polar, or hydrophobic solvent, or in the absence of solvent. As there is no other real option left for the choice of solvation, this contradicts the assumed efficacy of Lewis catalysts—there seems to be no appropriate environment in which these should be applied. Reciprocally, if solvent options appear as feasible, but none of the catalyst choices seem to work, a similar contradiction arises. Therefore, we suggest that only profiles with at least one "1" result for solvent and another for catalyst classes to be considered. This "meta-rule" of applicability is not enforced on the web page, but should be considered by the end user at the result interpretation stage.

Furthermore, we are fully aware that the limited training set size would implicitly limit the applicability of the approach, which is, at this point, rather a proof-of-concept tool. Eventually, predicted feasibility classes will be output, together with their trust markers, as labeled columns (legend relates numbering to the actual conditions provided)—see Figure 7.

3.3. External Validation Challenge. Here 4802 novel, putative Michael reaction products (matching the above-mentioned structural pattern) were extracted from the eNovalys database and submitted to the web server, for feasibility prediction. Out of these, the following cases were discarded:

a. The web server failed to unambiguously break the assumed Michael addition product into nucleophilic and electrophilic reaction partners, as mentioned above.
b. The web server found a way to break the product into expected reaction partners—however, the herein postulated reaction did not correspond to the synthetic process present in the database.
c. The retrosynthesis proposed by the web server did indeed match the Michael reaction from the database—however, the feasibility profiles did not respect the empirical rule of at least one feasible solvation and one feasible catalytic condition (see previous paragraph).

DOI: 10.1021/ci500698a
J. Chem. Inf. Model. 2015, 55, 239−250

Table 3. Predicted Feasibility Conditions vs Experimentally Used Setups for the 52 External Michael Reactions

solvent				catalyst			
predicted	used	no. instances	no. total	predicted	used	no. instances	no. total
P	P	18	22	NA	NA	11	24
P, H	P	1		B	B	6	
P, A	P, A	1		LA, NA	NA	3	
P, A	P	1		LA, B	B	2	
A	A	1		BA	BA	1	
P, H	A	2	30	LA, B	oxidant, B	1	
P	NA	15		NA	B	1	28
P	H	1		LA, NA	radical initiator	1	
P	A	6		LA, NA	B	2	
H	A	2		LA	NA	12	
A	NA	3		LA	BA	1	
A	H	1		LA	B	9	
				BA	NA	1	
				B	NA	1	

Only 52 reactions passed the filters a−c. Two data files, available as Supporting Information, report web server predictions vs experimental conditions listed in the database (ExtVal.xlsx), and, respectively, the SMILES of the concerned transformations (CoherentMichaelUniq.smi). Note: the ID columns present in both files can be used to establish a match between these entries.

The processes included two atypical instances using oxidants as catalysts in the database—thus likely driven by radical mechanisms (which does not preclude the feasibility by the classical heterolytic 1,4 addition as predicted by the web server). As can be seen from Table 3, instances where the predicted conditions matched the experimentally used setups, clearly count as predictive successes (upper part of the table, on colored background). However, only 8 cases out of the 52 witnessed the simultaneously matching predictions of both catalyst and solvent.

The lower part of the table does however not feature predictions that are necessarily wrong. These are, certainly, at least partly faulty, because of the failure to highlight the actually used experimental conditions as feasible. While there was no failure at all to recognize all the processes compatible with a protic solvent, only 2 out of the 12 processes realized with aprotic polar solvents were predicted feasible under these circumstances. Processes run in hydrophobic environments, or without the need of additional solvents were never recognized as such. With respect to the catalyst, only three classes (no catalyst, Brønsted acids, and bases) figured among practically used choices: 14 out of the 26 catalyst-free processes and 8 of the 19 base-catalyzed were recognized as such. Out of the two reactions needing Brønsted acids, one was acknowledged as such.

It is not known to this point whether the conditions recommended by the predictor would actually be appropriate or not. Note that most of the processes are predicted as compatible with the use of a Lewis acid catalyst—however, this choice was not even once exploited in practice. It is unclear whether this is due to the actual failure to perform the reactions under Lewis acid catalysis, because the option has never been tested, or avoided for extraneous reasons (postsynthesis purification, following steps, etc). Unfortunately, experimental assessment of this problem is not feasible by our teams.

4. CONCLUSION

This work presented the description and modeling of condition-specific feasibility of the Michael addition. Reaction conditions are described by four classes of solvent, four classes of catalyst, and a counterexample category (not occurring Michael additions). Different representations of the reactions were put under scrutiny: ISIDA, MOLMAP, CDK, and EED descriptors, alternatively derived on the basis of reagents only, products only or reaction-based. Definition of reaction-based descriptors is context-dependent. It makes no sense for classical, molecule-oriented terms like CDK, relies on the Condensed Graph "trick" to render a reaction as a pseudomolecule for the ISIDA Fragmentor, and is given by the difference (MOLMAPS) or, respectively, concatenation (EED) of product and reagent descriptors.

Naive Bayes, SVM and Random Forest algorithms were used to build the models and it was observed that the best models were obtained either using Random Forest or SVM.

It was observed that model performances where only marginally dependent on the description strategies. Reagent-based descriptors were often as potent as reaction-based ones, even in situations (ISIDA) when the former did not explicitly focus on the reaction centers, but treated them indiscriminately from all the other reagent atoms. Since key reaction centers are however straightforward to describe on the basis of atom nature and connectivity information implicitly present in all the used descriptors, this is not surprising.

Yet, this study was an opportunity to build QSRR models on a rarely studied problem, namely reaction conditions dependence of chemical feasibility. First, we demonstrated that the reaction conditions concept can be efficiently formulated as a multiple task problem. Second, predictive models were obtained for each of the defined reaction condition. Models are accessible on the Web site http://infochim.u-strasbg.fr/webserv/VSEngine.html.

However, the external validation exercise highlighted a key problem in modeling feasibility conditions: absence of reliable "negative" information for training and testing. Since feasibility conditions are *not* mutually exclusive, a reaction reported to successfully occur under given conditions is not necessarily unfeasible under not (yet) envisaged choices of catalyst and solvent. As the zeros in the training profiles typically stand for "do not know" rather than "unfeasible", the fact that robust, cross-validated separation of classes was readily achievable is, in

DOI: 10.1021/ci500698a
J. Chem. Inf. Model. 2015, 55, 239−250

itself, quite remarkable and is doubtlessly "enhanced" by unavoidable biases of the training data. It should not be forgotten that reported processes were carried out by experienced chemists, choosing reaction conditions according to their know-how. Therefore, specific structural signatures would be preferentially assigned to specific conditions. There is no direct incentive for experimentalists to scan for additional feasibility domains in condition space, unless their initial "educated guess" was proven wrong (unfortunately, there is little chance for such information to be published or captured in public databases). Accordingly, the interpretation of the external validation exercise is less than obvious, as the only objective criterion is the count of reported feasibility instances that failed to be recognized as such by the prediction. By contrast, when the expert system predicted a process as feasible under conditions differing from the listed ones, it remains open and subject to actual experimental validation (unfortunately, an effort not envisageable by our teams). While this article represents a successful proof-of-concept in modeling reaction feasibility, building of robust, practically useful models is conditioned by massive, automated screening of feasibility conditions for sets of diverse reactions, in order to generate a not-yet-available critical mass for machine learning.

■ ASSOCIATED CONTENT

Supporting Information

All statistics for all experiments, including descriptor calculations on the products of the reactions (that are too similar to those obtained on reagents to be presented into the article). The lists of the top outliers according to each DS and modeling method are given. The scrambling statistics are also detailed. Finally, it includes also a document containing the data set (reactions and their corresponding reaction condition profiles) and the external validation result files cited in section 3.3. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: g.marcou@unistra.fr.
*E-mail: varnek@unistra.fr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the French-Portuguese PESSOA program 28728QF and the Region Alsace for funding. A.d.L. thanks the Region Alsace for her Ph.D. fellowship. F. Hoonakker and A. Wagner are acknowledged for fruitful discussion. Financial support from Fundação para a Ciência e a Tecnologia (FCT) Portugal, under Projects PEst-C/EQB/LA0006/2013 and grant SFRH/BPD/63192/2009 (D. Latino) are greatly appreciated. A.V. thanks the Russian Scientific Foundation (Agreement No 14-43-00024, dated by October 1, 2014) for support. The authors thank the referees for their helpful remarks.

■ REFERENCES

(1) Anslyn, E. V.; Dougherty, D. A. *Modern Physical Organic Chemistry*; University Science, 2006.

(2) Warr, W. A. A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Inf* **2014**, *33* (6−7), 469−476.

(3) McDaniel, D. H.; Brown, H. C. An Extended Table of Hammett Substitutent Constants Based on the Ionization of Substituted Benzoic Acids. *J. Org. Chem.* **1958**, *23* (3), 420−427.

(4) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59* (1), 96−103.

(5) Hansch, C.; Leo, A.; Taft, R. W. A Survey of Hammett Substituent Constants and Resonance and Field Parameters. *Chem. Rev.* **1991**, *91* (2), 165−195.

(6) Friedman, M.; Wall, J. S. Additive Linear Free-Energy Relationships in Reaction Kinetics of Amino Groups with α,β-Unsaturated Compounds. *J. Org. Chem.* **1966**, *31* (9), 2888−2894.

(7) Friedman, M.; Cavins, J. F.; Wall, J. S. Relative Nucleophilic Reactivities of Amino Groups and Mercaptide Ions in Addition Reactions with α,β-Unsaturated Compounds. *J. Am. Chem. Soc.* **1965**, *87* (16), 3672−3682.

(8) Friedman, M.; Wall, J. S. Application of a Hammett-Taft Relation to Kinetics of Alkylation of Amino Acid and Peptide Model Compounds with Acrylonitrile. *J. Am. Chem. Soc.* **1964**, *86* (18), 3735−3741.

(9) Toropov, A.; Kudyshkin, V. O.; Voropaeva, N. L.; Ruban, I. N.; Rashidova, S. S. QSPR modeling of the reactivity parameters of monomers in radical copolymerizations. *Journal of Structural Chemistry* **2004**, *45* (6), 945−950.

(10) Yu, X.; Yi, B.; Wang, X. Quantitative structure-property relationships for the reactivity parameters of acrylate monomers. *Eur. Polym. J.* **2008**, *44* (12), 3997−4001.

(11) Morrill, J. A.; Biggs, J. H.; Bowman, C. N.; Stansbury, J. W. Development of quantitative structure-activity relationships for explanatory modeling of fast reacting (meth)acrylate monomers bearing novel functionality. *Journal of Molecular Graphics and Modelling* **2011**, *29* (5), 763−772.

(12) Schwöbel, J. A. H.; Wondrousch, D.; Koleva, Y. K.; Madden, J. C.; Cronin, M. T. D.; Schüürmann, G. Prediction of Michael-Type Acceptor Reactivity toward Glutathione. *Chem. Res. Toxicol.* **2010**, *23* (10), 1576−1585.

(13) Wondrousch, D.; Böhme, A.; Thaens, D.; Ost, N.; Schüürmann, G. Local Electrophilicity Predicts the Toxicity-Relevant Reactivity of Michael Acceptors. *J. Phys. Chem. Lett.* **2010**, *1* (10), 1605−1610.

(14) Zhang, Q.-Y.; Aires-de-Sousa, J. Structure-Based Classification of Chemical Reactions without Assignment of Reaction Centers. *J. Chem. Inf. Model.* **2005**, *45* (6), 1775−1783.

(15) Latino, D. A. R. S.; Aires-de-Sousa, J. Assignment of EC Numbers to Enzymatic Reactions with MOLMAP Reaction Descriptors and Random Forests. *J. Chem. Inf. Model.* **2009**, *49* (7), 1839−1846.

(16) Latino, D. A. R. S.; Zhang, Q.-Y.; Aires-de-Sousa, J. Genome-scale classification of metabolic reactions and assignment of EC numbers with self-organizing maps. *Bioinformatics* **2008**, *24* (19), 2236−2244.

(17) Latino, D. A. R. S.; Aires-de-Sousa, J. Genome-Scale Classification of Metabolic Reactions: A Chemoinformatics Approach. *Angew. Chem., Int. Ed.* **2006**, *45* (13), 2066−2069.

(18) Latino, D. A. R. S.; Aires-de-Sousa, J. Automatic Perception of Chemical Similarities Between Metabolic Pathways. *Mol. Inf.* **2012**, *31* (2), 135−144.

(19) de Luca, A.; Horvath, D.; Marcou, G.; Solovev, V.; Varnek, A. Mining Chemical Reactions Using Neighborhood Behavior and Condensed Graphs of Reactions Approaches. *J. Chem. Inf. Model.* **2012**, *52* (9), 2325−2338.

(20) Varnek, A. *Chemoinformatics and Computational Chemical Biology*; Springer, 2010.

(21) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. Condensed graph of reaction: considering a chemical reaction as one single pseudo molecule. In *The 19th International Conference on Inductive Logic Programming*, Heverlee, Belgium, July 2−4, 2009; http://lsiit.u-strasbg.fr/Publications/2009/HLVW09.

(22) Muller, C.; Marcou, G.; Horvath, D.; Aires-de-Sousa, J. o.; Varnek, A. Models for Identification of Erroneous Atom-to-Atom

DOI: 10.1021/ci500698a
J. Chem. Inf. Model. 2015, 55, 239−250

Mapping of Reactions Performed by Automated Algorithms. *J. Chem. Inf. Model.* **2012**, *52* (12), 3116−3122.
(23) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. Isida Property-labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29* (12), 855−868.
(24) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. v.; Marcou, G. Isida - Platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4* (3), 191−198.
(25) Varnek, A.; Fourches, D.; Solov'ev, V.; Klimchuk, O.; Ouadi, A.; Billard, I. Successful "in silico" design of new efficient uranyl binders. *Solvent Extr. Ion Exch.* **2007**, *25* (4), 433−462.
(26) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural Fragments: an Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput.-Aided Mol. Des.* **2005**, *19* (9−10), 693−703.
(27) Halberstam, N.; Baskin, I.; Palyulin, V.; Zefirov, N. Quantitative structure-conditions-property relationship studies. Neural network modelling of the acid hydrolysis of esters. *Mendeleev Commun.* **2002**, *12* (5), 185−186.
(28) Madzhidov, T. I.; Polishchuk, P. G.; Nugmanov, R. I.; Bodrov, A. V.; Lin, A. I.; Baskin, I. I.; Varnek, A. A.; Antipin, I. S. Structure-reactivity relationships in terms of the condensed graphs of reactions. *Russian Journal of Organic Chemistry* **2014**, *50* (4), 459−463.
(29) McMurry, J. *Fundamentals of Organic Chemistry*; Brooks/Cole, 2011.
(30) Mather, B. D.; Viswanathan, K.; Miller, K. M.; Long, T. E. Michael addition reactions in macromolecular design for emerging technologies. *Prog. Polym. Sci.* **2006**, *31* (5), 487−531.
(31) Mortreux, A.; Petit, F. *Industrial Applications of Homogeneous Catalysis*; Springer Science and Business Media, 1988; Vol. *10*.
(32) Balme, G.; Bouyssi, D.; Monteiro, N. In *Metal Catalyzed Cascade Reactions*; Springer: Berlin Heidelberg, 2006; pp 115−148.
(33) Engel, C. K.; Kiema, T. R.; Hiltunen, J. K.; Wierenga, R. K. The crystal structure of enoyl-CoA hydratase complexed with octanoyl-CoA reveals the structural adaptations required for binding of a long chain fatty acid-CoA molecule. *J. Mol. Biol.* **1998**, *275* (5), 847−859.
(34) Schmidt, T. J. Structure-activity relationships of sesquiterpene lactones. *Studies in Natural Products Chemistry* **2006**, *33* (M), 309−392.
(35) Roberts, D. W.; Natsch, A. High Throughput Kinetic Profiling Approach for Covalent Binding to Peptides: Application to Skin Sensitization Potency of Michael Acceptor Electrophiles. *Chem. Res. Toxicol.* **2009**, *22* (3), 592−603.
(36) Elhabiri, M.; Sidorov, P.; Cesar-Rodo, E.; Marcou, G.; Lanfranchi, D. A.; Davioud-Charvet, E.; Horvath, D.; Varnek, A. Electrochemical Properties of Substituted 2-Methyl-1,4-Naphtoquinones: Redox Behavior Predictions. *Chem.—Eur. J.* **2014**, DOI: 10.1002/chem.201403703.
(37) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 493−500.
(38) Laboratoire de Chemoinformatique Strasbourg. *Nomenclature of ISIDA Fragments*; 2012.
(39) Kohonen, T. *Self-Organizing Maps*; Springer, 2001.
(40) Kohonen, T. *Self-Organization and Associative Memory*; Springer: Heidelberg, 1984.
(41) Aires-de-Sousa, J. JATOON: Java tools for neural networks. *Chemometrics and Intelligent Laboratory Systems* **2002**, *61* (1−2), 167−173.
(42) Braban, M.; Pop, I.; Willard, X.; Horvath, D. Reactivity Prediction Models Applied to the Selection of Novel Candidate Building Blocks for High-Throughput Organic Synthesis of Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 1119−1127.
(43) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5−32.
(44) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter* **2009**, *11* (1), 10−18.
(45) Ivanciuc, O. *Applications of Support Vector Machines in Chemistry*; Wiley-VCH: Weinheim, 2007; Vol. *23*.
(46) Smola, A. J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Statistics and Computing* **2004**, *14* (3), 199−222.
(47) Schölkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, and London, England, 2002.
(48) Watson, P. Naive Bayes classification using 2D pharmacophore feature triplet vectors. *J. Chem. Inf. Model.* **2008**, *48* (1), 166−178.
(49) Chick, S. E. Subjective Probability and Bayesian Methodology. *Handbooks in Operations Research and Management Science: Simulation*; Elsevier, 2006; pp 225−257.
(50) Rennie, J.; Shih, L.; Teevan, J.; Karger, D. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning*, Washington, D.C., Aug 21−24, 2003; pp 616−623.
(51) Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47* (6), 2345−2357.
(52) DayLight SMARTS. http://www.daylight.com/dayhtml/doc/theory.smarts.html (accessed August 18, 2014).
(53) ChemAxon Standardizer. http://www.chemaxon.com/jchem/doc/user/standardizer.html (accessed Feb 2009).
(54) *SMIRKS*; Daylight Chemical Information Systems: 2007; Vol. *2007*.
(55) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *ATLA Alternatives to Laboratory Animals* **2005**, *33* (5), 445−459.

DOI: 10.1021/ci500698a
J. Chem. Inf. Model. 2015, 55, 239−250