

Assessing Synthetic Accessibility of Chemical Compounds Using Machine Learning Methods

Yevgeniy Podolyan,[†] Michael A. Walters,[‡] and George Karypis^{*,†}

Department of Computer Science and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, and Institute for Therapeutics Discovery and Development, Department of Medicinal Chemistry, University of Minnesota, Minneapolis, MN 55455

Abstract

With *de novo* rational drug design, scientists can rapidly generate a very large number of potentially biologically active probes. However, many of them may be synthetically infeasible and, therefore, of limited value to drug developers. On the other hand, most of the tools for synthetic accessibility evaluation are very slow and can process only a few molecules per minute. In this study, we present two approaches to quickly predict the synthetic accessibility of chemical compounds by utilizing support vector machines operating on molecular descriptors. The first approach, RSsvm, is designed to identify the compounds that can be synthesized using a specific set of reactions and starting materials and builds its model by training on the compounds identified as synthetically accessible or not by retrosynthetic analysis. The second approach, DRsvm, is designed to provide a more general assessment of synthetic accessibility that is not tied to any set of reactions or starting materials. The training set compounds for this approach are selected from a diverse library based on the number of other similar compounds within the same library. Both approaches have been shown to perform very well in their corresponding areas of applicability with the RSsvm achieving receiver operator characteristic score of 0.952 in cross-validation experiments and the DRsvm achieving a score of 0.888 on an independent set of compounds. Our implementations can successfully process thousands of compounds per minute.

1 Introduction

Many modern tools available today for virtual library design or *de novo* rational drug design allow the generation of a very large number of diverse potential ligands for various targets. One of

[†]Computer Science and Computer Engineering

[‡]Institute for Therapeutics Discovery and Development

the problems associated with *de novo* ligand design is that these methods can generate a large number of structures that are hard to synthesize. There are two broad approaches to address this problem. The first approach is to limit the types of molecules a program can generate. For example, restriction of the number of fragment types and other parameters in the CONCERTS¹ program allows it to control the types of molecules being generated which results in a greater percentage of the molecules being synthetically reasonable. SPROUT² applies a complexity filter³ at each stage of a candidate generation by making sure that all fragments in the molecule are frequent in the starting material libraries. An implicit inclusion of synthesizability constraints is implemented in programs such as TOPAS⁴ and SYNOPSIS.⁵ The TOPAS uses only the fragments obtained from known drugs using a reaction-based fragmentation procedure similar to RECAP.⁶ Instead of using the small fragments, SYNOPSIS uses the whole molecules in the available molecules library as fragments. Both programs then use a limited set of chemical reactions to combine the fragments into a single structure thus increasing the likelihood that the produced molecule is synthetically accessible. However, a potential limitation of using reaction-based fragmentation procedures is that they can limit drastically the diversity of the generated molecules. Consequently such approaches may fail to design molecules for protein targets that are significantly different from those with known ligands.

The second approach is to generate a very large number of possible molecules and only then apply a synthetic accessibility filter to eliminate hard to synthesize compounds. A number of approaches have been developed to identify such structures that are based on structural complexity,⁷⁻¹² similarity to available starting materials,³ retrosynthetic feasibility,¹³⁻¹⁵ or a combination of those.¹⁶

Complexity-based approaches use empirically derived formulas to estimate molecular complexity. Various methods have been developed to compute structural complexity. One of the earliest approaches suggested by Bertz⁷ uses a mathematical approach to calculate the chemical complexity. Whitlock⁹ proposed a simple metric based on the counts of rings, nonaromatic unsaturations, heteroatoms and chiral centers. The molecular complexity index by Barone and Chanon¹⁰ considers only the connectivity of the atoms in a molecule and the size of the rings. In an attempt to overcome the high correlation of the complexity scores with the molecular weight, Allu and Oprea¹² devised a more thorough scheme that is based on relative atomic electronegativities and bond parameters, which takes into account atomic hybridization states and various structural features such as number of chiral centers, geminal substitutions, number of rings and the types of fused rings, etc. All methods computing molecular complexity scores are very fast because they do not need to perform any comparisons to other structures. However, this is also a deficiency in these methods because the complexity alone is only loosely correlated with the synthetic accessibility as many starting materials with high complexity scores are readily available.

One of the newer approaches based on molecular complexity³ attempts to overcome the problems associated with regular complexity-based methods by including the statistical distribution of various cyclic and acyclic topologies and atom substitution patterns in existing drugs or commercially available starting materials. The method is based on the assumption that if the molecule contains only the structural motifs that occur frequently in commercially available starting materials, then they are probably easy to synthesize.

The use of sophisticated retrosynthetic analysis is another approach to assess the synthetic feasibility. A number of methods such as LHASA,¹³ WODCA,¹⁵ and CAESA¹⁴ utilize this approach. These methods attempt to find the starting materials for the synthesis by iteratively breaking the

bonds that are easy to create using high-yielding reactions. Unlike the methods described above, which require explicit reaction transformation rules, the Route Designer¹⁷ automatically generates these rules from the reaction databases. However, these reaction databases are not always annotated with yields and/or possibly difficult reaction conditions, which would limit their use. In addition, almost all of the programs based on retrosynthetic analysis are highly interactive and relatively slow, often taking up to several minutes per structure. This renders these methods unusable for the fast evaluation of the synthetic accessibility of tens to hundreds of thousands of molecules that can be generated by the *de novo* ligand generation programs.

One of the most recent methods by Boda et. al¹⁶ attempts to evaluate the synthetic accessibility by combining molecular graph complexity, ring complexity, stereochemical complexity, similarity to available starting materials, and an assessment of strategic bonds where a structure can be decomposed to obtain simpler fragments. The resulting score is the weighted sum of the components, where the weights are determined using regression so that it maximizes the correlation with medicinal chemists' predictions. The method was able to achieve correlation coefficient of 0.893 with the average scores produced by five medicinal chemists. However, this method was only shown to work for the 100 compounds used in the regression analysis to select the weights for the individual components. In addition, this method uses a complex retrosynthetic reaction fitness analysis which was shown to have a very low correlation with the total score. The method is able to process only 200–300 molecules per minute.

In this paper we present methods for determining the relative synthetic accessibility of a large set of compounds. These methods formulate the synthetic accessibility prediction problem as a supervised learning problem in which the compounds that are considered to be synthetically accessible form the positive class and those that are not form the negative class. This approach is based on the hypothesis that the compounds containing fragments that are present in a sufficiently large number of synthetically accessible molecules are themselves synthetically accessible. The actual supervised learning model is built using Support Vector Machines (SVM) and utilizes a fragment-based representation of the compounds that allows it to precisely capture the compounds' molecular fragments and their frequency. Using this framework, we developed two complementary machine learning models for synthetic accessibility prediction that differ on the approach that they employ for determining the label of the compounds during training.

The first model, referred to as RSsvm, uses as positive instances a set of compounds that were retrosynthetically determined to be synthesizable in a small number of steps and as negative instances a set of compounds that required a large number of steps. The resulting machine learning model can then be used to substantially reduce the amount of time required by approaches based on retrosynthetic analysis by replacing retrosynthesis with the model's prediction. The advantage of this model is that by identifying the positive and negative instances using a retrosynthetic approach, it is trained on a set of compounds whose labels are reliable. However, assuming a sufficiently large library of starting materials, the resulting model will be dependent on the set of reactions used in the retrosynthetic analysis, and as such it may not be able to generalize to compounds that can be synthesized using additional reactions. Consequently, this approach is well-designed for cases in which the compounds need to be synthesized with a predetermined set of reactions as it is often the case in the pharmaceutical industry where only parallel synthesis reactions are used.

The second model, referred to as DRsvm, is based on the assumption that compounds belonging to dense regions of the chemical space were probably synthesized by a related route through combinatorial or conventional synthesis and as such, they are relatively easy to synthesize. Based

on that, this approach uses as positive instances a set of compounds that are similar to a large number of other compounds in a diverse library and as negative instances a set of compounds that are similar to only a few compounds (if any). The advantage of this approach is that assuming that the library from where they were obtained was sufficiently diverse, the resulting model is not tied to a specific set of reactions and as such it should be able to provide a more general assessment of synthetic accessibility. However, since the labels of the training instances are less reliable, the error rate of the resulting predictions can potentially be higher than those produced by RSsvm for compounds that can be synthesized by the reactions used to train RSsvm’s models.

We experimentally evaluated the performance of these models on datasets derived from the Molecular Libraries Small Molecule Repository (MLSMR) and a set of compounds provided by Eli Lilly and Company. Specifically, cross-validation results on MLSMR using a set of high-throughput synthesis reactions show that RSsvm leads to models that achieve an ROC score of 0.952, and on the Eli Lilly dataset, both RSsvm and DRsvm correctly position only synthetically accessible compounds among the 100 highest ranked compounds.

2 Methods

2.1 Chemical Compound Descriptors

The compounds are represented as a frequency vector of the molecular fragments that they contain. The molecular fragments correspond to the GF descriptors,¹⁸ which are the complete set of unique subgraphs between a minimum and maximum size that are present in each compound. The frequency corresponds to the number of distinct embeddings of each subgraph in a compound’s molecular graph. Details on how these descriptors were generated are provided in Section 3.3.

The GF descriptor-based representation of chemical compounds is well-suited for synthetic accessibility purposes for several reasons. (i) Molecular complexity (e.g., the number and types of rings, bond types, etc.) is easily encoded with such fragments. (ii) The fragments can implicitly encode chemical reactions through the new sets of bonds they create. (iii) The similarity to the starting materials can also be determined by the share of the common fragments present in the compound of interest and the compounds in the starting materials. (iv) Unlike descriptors based on physico-chemical properties,^{19,20} they capture elements of the compound’s molecular graph (e.g., bonding patterns, rings, etc.), which are more directly relevant to the synthesizability of a compound. (v) Unlike molecular descriptors based on hashed fingerprints (e.g., Daylight or extended connectivity fingerprints²¹) they precisely capture the molecular fragments present in each compound without the many-to-one mapping errors introduced as a result of hashing. (vi) Unlike methods utilizing a very limited set of predefined structural features (e.g., MACCS keys by Molecular Design Ltd.), the GF descriptors are complete and in conjunction with the use of supervised learning methods, can better identify the set of structural features that are important for predicting the synthetic accessibility of compounds.

2.2 Model Learning

We used Support Vector Machines (SVM)²² to build the discriminative model that explicitly differentiates between easy and hard to synthesize compounds. Once built, such a model can be

applied to any compound to assess its synthetic accessibility. The SVM machine learning method is a linear classifier of the form

$$f(\mathbf{x}) = \mathbf{w}^t \mathbf{x} + b, \quad (1)$$

where $\mathbf{w}^t \mathbf{x}$ is the dot product of the weight vector \mathbf{w} and the input vector \mathbf{x} which represents the object being classified. However, by applying a kernel trick,^{23,24} the SVM can be turned into a non-linear classifier. The kernel trick is applied by simply replacing the dot product in the above equation with the kernel function to compute similarity of two vectors. The SVM classifier can be rewritten as

$$f(\mathbf{x}) = \sum_{\mathbf{x}_i \in S^+} \lambda_i^+ K(\mathbf{x}, \mathbf{x}_i) - \sum_{\mathbf{x}_i \in S^-} \lambda_i^- K(\mathbf{x}, \mathbf{x}_i), \quad (2)$$

where S^+ and S^- are the sets of positive and negative class vectors, respectively, chosen as support vectors during the training stage; λ_i^+ and λ_i^- are the non-negative weights computed during the training; and K is the kernel function that computes the similarity between two vectors. The resulting value $f(\mathbf{x})$ can be used to rank the testing set instances, while the sign of the value can be used to classify the vectors into positive or negative class.

In this study, we used the Tanimoto coefficient²⁵ as the kernel function²⁶ as well as for computations of the similarity to the compounds in starting materials library. Tanimoto coefficient has the following form:

$$S_{A,B} = \frac{\sum_{i=1}^n x_{iA} x_{iB}}{\sum_{i=1}^n x_{iA}^2 + \sum_{i=1}^n x_{iB}^2 - \sum_{i=1}^n x_{iA} x_{iB}}, \quad (3)$$

where A and B are the vectors being compared, n is the total number of features, and x_{iA} and x_{iB} are the numbers of times an i -th feature occurs in vectors A and B , respectively. Tanimoto coefficient is the most widely used metric to compute similarity between the molecules encoded as vectors of various kinds of descriptors. It was recently shown²⁷ to be the best performing coefficient for the widest range of query and database partition sizes in retrieving complementary sets of compounds from MDL Drug Data Report database.

The Tanimoto coefficient for any two vectors is a value between 0 and 1 indicating how similar the compounds represented by those vectors are. Some examples of molecular pairs with different computed similarity scores (based on the GF descriptors) are presented in Figure 1.

2.3 Identifying the Training Compounds

In order to train a machine learning model, one needs to supply a training set consisting of positive and negative examples. In the context of synthetic accessibility prediction, the positive examples would represent the compounds that are relatively easy to synthesize and the negative examples would represent the compounds that are relatively hard to synthesize. The problem here lies in the fact that there are no catalogs or databases that would list compounds labeled as easy or hard to synthesize. The problem is complicated even further by the fact that even though the easy to synthesize molecules can be proven to be easy, the opposite is not true—one cannot prove with absolute certainty that the molecules classified as hard to synthesize are not actually easy to synthesize by some little-known or newly discovered method.

One way to create a training set of molecules would be to ask synthetic chemists to give a score to each molecule that would represent their opinion of the difficulty associated with the synthesis

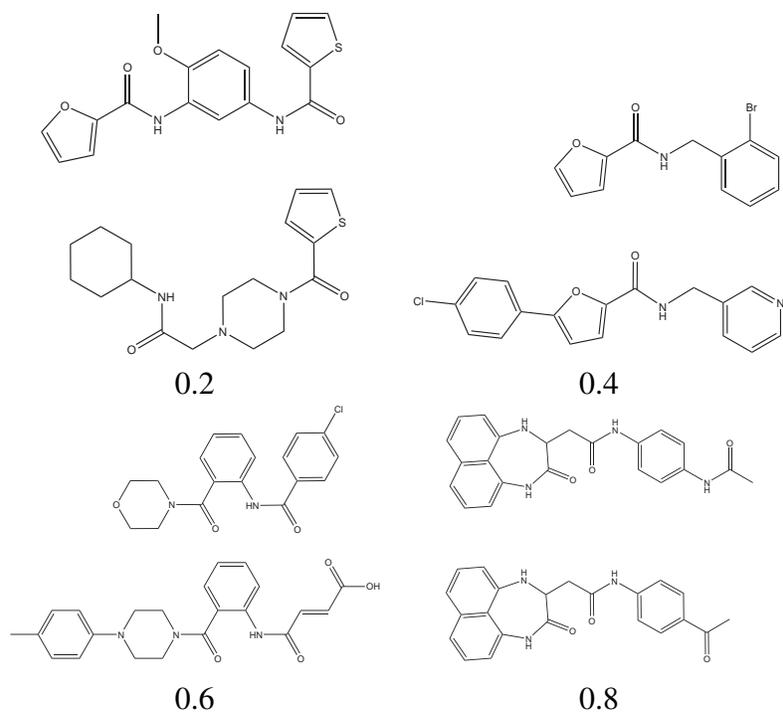


Figure 1: Examples of molecular pairs with given computed similarity scores.

of that molecule. One can then split the set into two parts, one with lower scores and another with higher scores. However, such an approach cannot lead to very large training sets, as there is a limit on the number of compounds that can be manually labeled. Moreover, the resulting scores are not objective as chemists often disagree with each other.¹⁶

In this paper, we developed two different approaches, described in the subsequent sections, that take as input a large diverse compound library and automatically identify the positive and negative compounds for training SVM models for synthetic accessibility prediction.

2.3.1 Retrosynthetic Approach (RS_{SVM} Model)

This approach identifies the subsets of compounds that are easy or hard to synthesize by using a retrosynthetic decomposition approach^{28,29} to determine the number of steps required to decompose each compound into easily available compounds (e.g., commercially available starting materials). Those compounds that can be decomposed in a small number of steps are considered to be easy to synthesize, whereas those compounds that either required a large number of steps or they could not be decomposed are considered to be hard to synthesize.

This approach requires three inputs: (i) the compounds that need to be decomposed, (ii) the reactions whose reverse application will split the compounds, and (iii) the starting materials library. The decomposition is performed in the following way. First, each reaction is applied to a given compound to check whether it is applicable to this particular compound. All applicable reactions, i.e., reactions that break bond(s) that is/are present in a given molecule, are applied to produce two or more smaller compounds. Each of the resulting compounds is checked for an exact match in

the starting materials library. Those compounds not found among the starting materials are further broken down into smaller compounds. The process is done in a breadth-first search fashion to find the shortest path to the starting materials. The search is halted when all parts of the compound along one of the reaction paths in this search tree are found in the starting materials library or when no complete decomposition is found after performing a certain number of decomposition rounds.

The schematic representation of the decomposition process is shown in Figure 2, where the squares represent compounds (dashed squares represent compounds found in the starting materials library) and circles represent reactions (dashed circles represent reactions that are not applicable to the chosen compound). One can see that the height of the decomposition tree for compound A is two since it can be decomposed with two rounds of reactions. Using reaction R_4 , compound A can be broken down into compounds D and E. Since D is already in the starting materials library, it does not need to be broken down any further. The compound E can be broken down into H and C, both of which are in the starting materials library, with another reaction R_4 . The retrosynthetic analysis stopped after two rounds because the path $A \rightarrow R_4 \rightarrow (D)E \rightarrow R_4 \rightarrow (H)(C)$ terminates with all compounds found in the starting materials library.

In this paper, we considered all compounds that have decomposition search trees with one to three rounds of reactions (i.e., have a height of one to three measured in terms of reactions) to be easy to synthesize and compounds that cannot be synthesized in at least five rounds of reactions to be hard to synthesize. Note that this method assumes that the difficulty associated with performing these reactions is the same. However, the overall approach can be easily extended to account for reactions with different levels of difficulty by using variable depth cut-offs based on the reactions used. These easy and hard to synthesize compounds are used as the sets of positive and negative instances for training the RSsvm models and will be denoted by RS^+ and RS^- , respectively. The molecules with a decomposition search tree height of four or five were not used in training or testing the RSsvm models.

The advantage of this approach is that it provides an objective way of determining the synthetic accessibility of a compound based on the input reactions and library of starting materials. Moreover, even though for a large number of reactions, retrosynthetic analysis is computationally expensive, the RSsvm model requires this step to be performed only once while generating its training set, and subsequent applications of the RSsvm model does not require retrosynthetic analysis. Thus, the RSsvm model can be used to dramatically reduce the computational complexity of retrosynthetic analysis by using it to prioritize the compounds that will be subjected to such type of analysis.

2.3.2 Dense Regions Approach (DRsvm Model)

A potential limitation with the retrosynthetic approach is that the resulting RSsvm model will be biased towards the set of reactions used to identify the positive and negative compounds. As a result, it may fail to properly assess the synthetic accessibility of compounds that can be decomposed using a different set of reactions. To address this problem, we developed an alternate method for selecting the sets of positive and negative training instances from the input library. The assumption underlying this approach is that if a compound has a large number of other compounds in the library that are significantly similar to it, then this compound and many of its neighbors were probably synthesized by a related route through combinatorial or conventional synthesis and are, therefore, synthetically accessible.

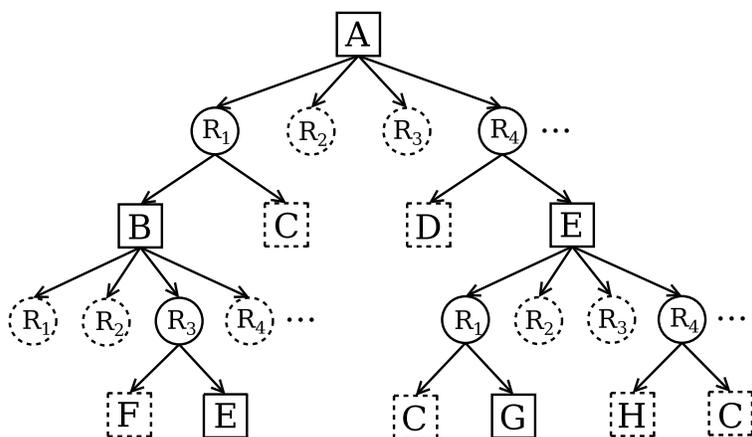


Figure 2: Example of compound decomposition. The squares represent compounds (dashed squares represent compounds found in the starting materials library). The circles represent reactions (dashed circles represent reactions that are not applicable to the chosen compound). The height of the decomposition tree for compound A is 2 because it can be decomposed with a minimum of 2 rounds of reactions.

In this approach, the positive training set, denoted by DR^+ , is generated by selecting from the input library the ones with at least κ^+ other compounds that have a similarity of at least θ . These compounds represent the dense area of the library. Conversely, the negative training set, denoted by DR^- , is generated by selecting those library compounds that have no more than κ^- other compounds that have a similarity of at least θ . Note that $\kappa^- \ll \kappa^+$ and that the similarity between compounds is determined using the Tanimoto coefficient of their GF descriptor representation. The DR^+ and DR^- sets are used for building the DR_{SVM} model for synthetic accessibility prediction. In our study, we used $\kappa^+ = 20$, $\kappa^- = 1$, and $\theta = 0.6$, as we found them to produce a reasonable number of positive and negative training compounds.

2.3.3 Baseline Approaches

In order to establish a basis for the comparison of the above models, we also developed two simple approaches for synthetic accessibility prediction. The first, denoted by $SMNP_{SVM}$, is an SVM-based model that is trained on the starting materials as the positive class and natural products as the negative class. This approach is motivated by the facts that (i) most commercially available materials are easily synthesized and (ii) natural products are considered to be hard to synthesize.³⁰ As a result, an SVM-based model that uses starting materials as the positive class and natural products as the negative class may be used to prioritize a set of compounds with respect to their synthetic accessibility. The second, denoted by MAX_{SMSIM} , is a scheme that ranks the compounds whose synthetic accessibility needs to be determined based on their highest similarity to a library of starting materials. Specifically, for each unknown compound, its Tanimoto similarity to each of the compounds in a starting materials library is computed, and the maximum similarity value is used as its synthetic accessibility score. The final ranking of the compounds is then obtained by sorting them in descending order based on those scores. The use of the maximum similarity to

starting materials is motivated by the empirical observation that the molecules in the positive set have, on average, higher maximal similarity to starting materials than the molecules in the negative set (Table 3).

3 Materials

3.1 Data Sets

In this study we used four sets of compounds. (i) A library of starting materials (SM) that is needed for the retrosynthetic analysis. This set was compiled from the Sigma-Aldrich's Selected Structure Sets for Drug Discovery, Acros, and Maybridge databases. After stripping the salts and removing the duplicates, the starting materials library contained 97,781 compounds including the multiple tautomeric forms for some compounds. (ii) A library of natural products (NP) obtained from InterBioScreen Ltd, which is used to train the SVM-based baseline model (Section 2.3.3). This library contains 41,746 natural compounds, their derivatives and analogs that are offered for high-throughput screening programs. (iii) The compounds in the Molecular Libraries Small Molecule Repository (MLSMR), which is available from PubChem.³¹ This library was used as the diverse library of compounds for training the RSsvm and DRsvm models and evaluating the performance of the RSsvm models. MLSMR is designed to collect, store and distribute compounds for high-throughput screening (HTS) of biological assays that were submitted by the research community. Since a large portion of the deposited compounds were synthesized via parallel synthesis, one can assume that many of them should be relatively straightforward to synthesize. The size of the MLSMR library that we used contained 224,278 compounds. (iv) A set of 1,048 molecules obtained from Eli Lilly and Company that is used for evaluating the performance of the RSsvm and DRsvm models that were trained on the MLSMR library, and as such providing an external way of evaluating the performance of these methods. These compounds were considered to be hard to synthesize by Eli Lilly chemists and their synthesis was outsourced to an independent company specializing in the chemical synthesis. The set contains 830 compounds that were eventually synthesized and 218 compounds for which no feasible synthetic path was found by chemists. We will denote the set of synthesized compounds as EL^+ and the set of compounds that were never synthesized as EL^- .

3.2 Reactions

For the retrosynthetic analysis, we used a set of 23 reactions that are employed widely in the high-throughput synthesis. The set included parallel synthesis reactions such as reductive amination; amine alkylation; nucleophilic aromatic substitution; amide, sulfonamide, urea, ether, carbamate, and ester formation; amido alkylation; thiol alkylation; and Suzuki and Buchwald couplings. Note that this set can be considerably expanded to include various bench-top reactions and as such lead to potentially more general RSsvm models for synthetic accessibility prediction. However, because we are interested in assessing the potential of the RSsvm models in the context of high-throughput synthesis, we only used parallel synthesis reactions.

Table 1 shows some statistics of the retrosynthetic decomposition of the MLSMR, natural products, and Eli Lilly datasets based on this set of reactions. The compounds in these datasets that

Table 1: Characteristics of the retrosynthetic decomposition of different datasets.

	MLSMR		Natural Products		Eli Lilly Data Set			
	Count	% of total	Count	% of total	EL ⁺		EL ⁻	
	Count	% of total	Count	% of total	Count	% of total	Count	% of total
Found in SM	7,890	3.5	459	1.1	0	0.0	0	0.0
Depth 1	26,074	11.6	1,751	4.2	56	6.7	0	0.0
Depth 2	20,025	8.9	1,387	3.3	28	3.4	0	0.0
Depth 3	2,522	1.1	155	0.4	4	0.5	0	0.0
Depth 4	188	0.1	20	<0.1	0	0.0	0	0.0
Depth 5	5	<0.1	11	<0.1	0	0.0	0	0.0
Depth >5 [†]	167,574	74.7	37,963	90.9	742	89.4	218	100.0
Total	224,278	100.0	41,746	100.0	830	100.0	218	100.0

[†] Compounds with no identifiable path with depth up to 5.

were found in the starting materials library (first line in the table) were not used in this study. These results show that a larger fraction of the MLSMR library (21.6%) can be synthesized in three or fewer steps (easy to synthesize) than the corresponding fractions of the NP library (7.9%) and Eli Lilly dataset (10.6%). These results indicate that the compounds in the Eli Lilly dataset are considerably harder to synthesize than those in MLSMR. Moreover, since the retrosynthetic decomposition of all the compounds in EL⁻ is greater than five, these results provide an independent confirmation on the difficulty of synthesizing the EL⁻ compounds. Note that one reason for the relatively small fraction of the MLSMR and Eli Lilly compounds that are determined to be synthetically accessible is due to the use of a small number of parallel synthesis reactions and the number will be higher if additional reactions are used.

In addition, Table 2 provides some information on the role of the different reactions in decomposing the easy to synthesize compounds in MLSMR. Specifically, for each reaction, Table 2 shows the number of compounds that used that reaction at least once and the number of compounds that had their decomposition tree height increased when the reaction was excluded from the set of available reactions. These statistics indicate that not all reactions are equally important. Some of the reactions, such as reactions 4 and 11 (amide formation from amines with acids and acyl halides, respectively), are ubiquitous. Reaction 4 was used in the decomposition of 21,495 molecules, whereas reaction 11 was used in the decomposition of 12,392 molecules. On the other hand, several reactions have been used in the decomposition of fewer than 100 molecules. In addition, some of the reactions are replaceable, in the sense that their elimination will not cause significant changes in the decomposition trees of the compounds. For instance, the elimination of reaction 11 increased the synthesis path for 68 out of the 12,392 molecules that used that reaction in their shortest decomposition paths. On the other hand, the elimination of reaction 17 (thiol alkylation with alkyl halides) caused 96% of the 3,467 molecules that used that reaction, to become hard to synthesize (i.e., their depth of the decomposition tree became greater than 5). Similarly, reaction 19 (the synthesis of the imidazo-fused ring systems from aminoheterocycles and α -haloketones), while used only by 161 molecules, was absolutely essential to their decomposition as the decom-

Table 2: The use of each reaction in the decomposition of 48,781 easy to synthesize MLSMR molecules and the effect of each reaction exclusion on the height of the shortest decomposition path.

Reaction	Molecules using reaction [†]	Molecules with decompos. changed to given depth			Description [‡]
		4	5	> 5	
1	3989	5	0	1351	Reductive amination of aldehydes and ketones
2	4721	9	0	2855	Amine alkylation with alkylhalides
3	1055	24	28	844	NAS of haloheterocycles with amines
4	21495	125	179	17872	Amide formation with amines and acids
5	56	0	0	41	Urea formation with isocyanates and amines
6	844	0	0	729	Suzuki coupling
7	2620	0	0	318	Phenol alkylation with alkyl halides
8	644	0	0	451	Ether formation from phenols and secondary alcohols
9	7224	1	0	7169	Sulfonamide formation with sulfonyl chlorides and amines
10	2544	0	0	478	N–N'-Disuccinimidyl carbonate mediated urea formation
11	12392	0	0	68	Amide formation with amines and acyl halides
12	3622	45	0	827	Amido alkylation with alkyl halides
13	68	0	0	0	Urea formation with ethoxycarbonyl-carbamyl chlorides and amines
14	2131	0	0	4	Ester formation with acyl halides and alcohols
15	3411	0	0	3093	Ester formation with acids and alcohols
16	400	0	0	368	NAS of haloheterocycles with alcohols and phenols
17	3467	0	0	3345	Thiol alkylation with alkyl halides
18	236	0	0	235	Benzimidazole formation with phenyldiamines and aldehydes
19	161	0	0	161	Synthesis of imidazo-fused ring systems from aminoheterocycles and α -haloketones
20	429	0	0	302	Carbamate synthesis with primary alcohols and secondary amines
21	1061	12	0	108	Reductive amination of carbonyl compounds followed by acylation with carboxylic acids
22	40	0	0	0	NAS of di-haloheterocycles with amines
23	332	0	0	0	Ether formation from phenolic aldehydes followed by reductive amination with amines

[†] The number of molecules whose shortest decomposition search tree contains a given reaction.

[‡] NAS = Nucleophilic aromatic substitution.

position paths of height up to five could not be discovered for all 161 molecules when the reaction was omitted.

3.3 Descriptors

The GF descriptors of the compounds were generated with the AFGen 2.0 program.^{32,33} AFGen treats all molecules as graphs with atoms being vertices and bonds being edges in the graph. The atom types serve as vertex labels while bond types serve as the edge labels. We used fragments of length from four to six bonds. The fragments contained cycles and branch points in addition to linear paths. Since most chemical reactions create only one or just a few new bonds, such a fragment length will be sufficient to implicitly encode the reactions that may have been used in the compound synthesis. Our experiments with the maximum length of the fragments used as GF descriptors have shown that increasing it beyond six does not improve the classification accuracy but significantly increases the computational requirements of the methods.

In order to compare the performance of the SVM models based on the GF descriptors to those using fingerprints, we generated the latter with ChemAxon’s GenerateMD.³⁴ The produced fingerprints are similar to the Daylight fingerprints. The 2048-bit fingerprints were produced by hashing the paths in the molecules of the length up to six bonds. Note that both the GF descriptors and the hashed fingerprints contained only heavy atoms, i.e, all atoms except hydrogen.

3.4 Implementation Details

The retrosynthetic decomposition has been implemented in C++ programming language using OpenEye’s OEChem software development kit.³⁵ The kit, which is a library of classes and functions, allows among other things to apply a given reaction (encoded as Daylight’s SMIRKS patterns describing molecular transformations) to a set of reactants and obtain the products if the reaction is applicable. If the same reaction can be applied in multiple ways then multiple sets of products will be obtained. The reactions were encoded in reverse, i.e. in the direction from products to reactants. The *SVM^{light}*³⁶ implementation of the support vector machines was used to perform the learning and classification. All input parameters to the *SVM^{light}* with the exception of the kernel (see Section 2.2) were left at their default values. All experiments were performed on a Linux workstation with a 3.4 GHz Intel Xeon processor.

3.5 Training and Testing Framework

The MLSMR library was used to evaluate the performance of the RSsvm model in the following way. Initially, in order to avoid molecules that are very similar within the positive and the negative sets as well as across training and testing sets, a subset of MLSMR compounds was selected from the entire MLSMR set such that no two compounds had a similarity greater than 0.6 (the similarity was computed using the Tanimoto coefficient of GF-descriptor representation). A greedy maximal independent set (MIS) algorithm was used to obtain this set of compounds. The algorithm works in the following way. Initially, a graph is built in which vertices represent compounds and edges are placed between vertices that have similarity greater than 0.6. After the graph is constructed, a vertex is picked from the graph, placed in the independent set and then removed with all of its neighbors from the graph. The process is repeated until the graph is empty. In order to maximize

Table 3: Characteristics of molecules in the different positive and negative datasets.

	MLSMR				Eli Lilly		Natural Products	Starting Materials
	RS ⁺	RS ⁻	DR ⁺	DR ⁻	EL ⁺	EL ⁻		
Median number of unique features	190	268	276	207	257	346	332	163
Mean max. similarity to SM	0.74	0.65	0.69	0.63	0.63	0.52	0.70	0.88 [†]

[†] Similarity to compounds in SM library other than itself.

the size of the independent set in general and the positive subset (RS⁺) in particular, the order in which the vertices were picked from the graph was based on the label and the degree (number of neighbors) of the vertex. For that purpose, all vertices in a graph were sorted first by their label (positively labeled vertices were placed in a queue before the negatively labeled ones) and then by the degree (vertices with fewer neighbors were placed ahead of the ones with a larger number of neighbors). The order of vertices with the same label and degree was determined randomly. The application of this MIS algorithm with a subsequent classification using retrosynthetic analysis resulted in 12,113 easy to synthesize compounds (RS⁺) and 35,580 hard to synthesize compounds (RS⁻). The variations in the sizes of the sets and in the results of the performance studies due to the random ordering of the vertices with the same label and degree were found to be insignificant and will not be discussed. Both, RS⁺ and RS⁻ sets, were randomly split into five subsets of similar size, leading to five folds each containing a subset of positive and a subset of negative compounds. The RSsvm model was evaluated using five-fold cross validation, i.e. the model was learned using four folds and tested on the fifth fold (repeating the process five times for each of the folds). The baseline models were also evaluated on each of the five test folds. In addition, an RSsvm model was trained on all RS⁺ and RS⁻ compounds without performing the MIS selection. This model was then used to determine the synthetic accessibility of the compounds in the Eli Lilly dataset; thus, providing an assessment of the RSsvm’s model on an independent dataset.

The performance of the DRsvm-based approach was also assessed by using the method described in Section 2.3.2 to identify the DR⁺ and DR⁻ compounds in MLSMR, train the DRsvm model, and then use it to predict the synthetic accessibility of the compounds in the Eli Lilly dataset. The DR⁺ and DR⁻ training sets included 89,073 and 15,802 compounds, respectively, with DR⁺ containing 23.1% and DR⁻ containing 9.9% of compounds identified as easy by a retrosynthetic analysis. Note that we did not attempt to evaluate the performance of the DRsvm-based models on the RS⁺ and RS⁻ compounds of the MLSMR dataset, because the labels of these compounds are derived from the small number of reactions used during the retrosynthetic decomposition, and as such they do not represent the broader sets of synthetically accessible compounds that the DRsvm-based approach is designed to capture.

Some characteristics of the above MLSMR and Eli Lilly sets of compounds, such as median number of unique features (graph fragments) and mean maximal similarity to the compounds in the starting materials library, are given in Table 3.

3.6 Performance Evaluation

We used the receiver operating characteristic (ROC) curve to measure the performance of the models. The ROC measures the rate at which positive class molecules are discovered compared to the rate of the negative class molecules.³⁷ The area under the ROC curve, or simply ROC

Table 4: Cross-validation performance on the MLSMR dataset.

Method	ROC	ROC50	PA100 [†] (Enrichment)
RS _{SVM} (GF descriptors)	0.952	0.276	99.0 (3.9)
RS _{SVM} (2048-bit fingerprints)	0.948	0.208	97.0 (3.8)
SMNP _{SVM}	0.567	0.004	27.4 (1.1)
MAX _{SMSIM}	0.679	0.007	39.2 (1.5)
Random	0.500	0.003	25.4 (1.0)

[†] The number of true positives among the 100 highest-ranked test compounds.

score, can indicate how good a method is. For example, in the ideal case, when all of the positive class molecules are discovered before (i.e., classified with higher score than) the molecules of the negative class, the ROC score will be 1. In a completely random selection case, the ROC score will be 0.5. In addition to the ROC score, we also report two measures designed to give an insight into the highest-ranked compounds. The first measure is the ROC50 score, which is the modified version of the ROC score that indicates the area under the ROC curve up to the point when 50 negatives have been encountered.³⁸ The second measure is the percentage of the 100 highest-ranked test compounds that are true positive molecules. We will denote this measure by PA100 (positive accuracy in the top 100 compounds). Since in a completely random selection case ROC50 score depends on the actual number of negative class test instances and PA100 depends on the ratio of positive to negative class instances, the expected random values for these measures are given in each table. We also provide the enrichment factor in the top 100 compounds, which is calculated by dividing the actual PA100 by the expected random PA100 value. Note that in the case of cross-validation experiments, the results reported are the averages over five different folds.

4 Results

4.1 Cross-Validation Performance on the MLSMR Dataset

Table 4 shows the five-fold cross validation performance achieved by the RS_{SVM} models on the MLSMR dataset. Specifically, the results for two different RS_{SVM} models are presented that differ on the descriptors that they use for representing the compounds (GF descriptors and 2048-bit fingerprints). In addition, this table shows the performance achieved by the two baseline approaches SMNP_{SVM} and MAX_{SMSIM} (Section 2.3.3), and an approach that simply produces a random ordering of the test compounds. To ensure that the performance numbers of all schemes are directly comparable, the numbers reported for the baseline and random approaches were obtained by first evaluating these methods on each of the test-folds that were used by the RS_{SVM} models, and then averaging their performance over the five folds.

These results show that the two RS_{SVM} models are quite effective in predicting the synthetic accessibility of the test compounds and that their performance is substantially better than that achieved by the two baseline approaches across all three performance assessment metrics. The performance difference of the various schemes is also apparent by looking at their ROC curves

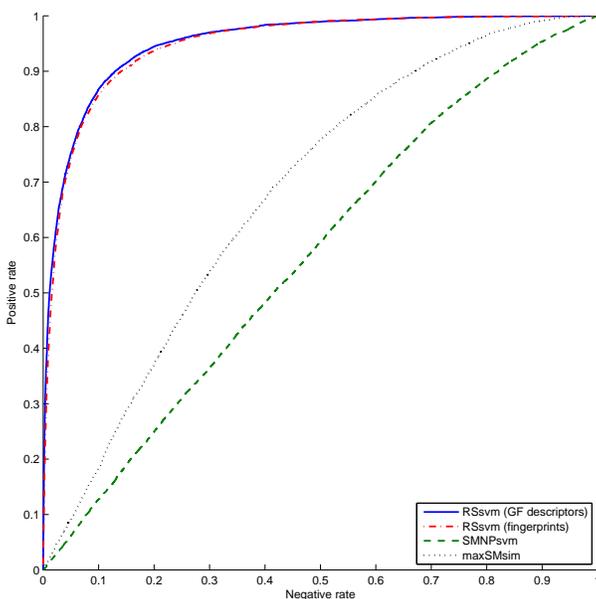


Figure 3: ROC curves for ranking easy to synthesize MLSMR (RS^+) and hard to synthesize MLSMR compounds (RS^-) using different methods/models.

shown in Figure 3. These plots show that both of the RS_{svm} models are able to rank over 85% of the RS^+ compounds same or higher than the top 10% ranked RS^- compounds, which is substantially better than the corresponding performance of the baseline approaches. Comparing the two RS_{svm} models, the results show that GF descriptors outperform the 2048-bit fingerprints. The performance difference between the two schemes is more pronounced in terms of the ROC50 metric (0.276 vs 0.208) and somewhat less for the other two metrics.

Comparing the performance of the two baseline approaches, we see that they perform quite differently. By using starting materials as the positive class and natural products as the negative class, the $SMNP_{svm}$ model fails to rank the compounds in RS^+ higher than the RS^- compounds and achieves an ROC score of just 0.567. The reason for this poor performance may lie in the fact that natural products have very complex structures and the model easily learns how to differentiate between natural products and everything else instead of easy and hard to synthesize compounds. On the other hand, by ranking the molecules based on their maximum similarity to the starting materials, $maxSM_{sim}$ performs better and achieves an area under the ROC curve of 0.679. The $maxSM_{sim}$ approach also produced better results in terms of ROC50 and PA100. Nevertheless, the performance of neither approach is sufficiently good for being used as a reliable synthetic accessibility filter, as they achieve very low ROC50 and PA100 values.

4.2 Performance on the Eli Lilly Dataset

Table 5 shows the performance achieved by the DR_{svm} , RS_{svm} , and the two baseline approaches on the Eli Lilly dataset, whereas Figure 4 shows the ROC curves associated with these schemes.

Table 5: Performance on the Eli Lilly dataset.

Model	ROC	ROC50	PA100 (Enrichment)
DR _{SVM}	0.888	0.713	100.0 (1.3)
RS _{SVM}	0.782	0.537	100.0 (1.3)
SMNP _{SVM}	0.374	0.058	65.0 (0.8)
MAXSM _{SIM}	0.711	0.378	100.0 (1.3)
MAXDR _{SIM}	0.744	0.443	97.0 (1.2)
Random	0.500	0.115	79.2 (1.0)

As discussed in Section 3.5, the DR_{SVM} and RS_{SVM} models were trained on compounds that were derived from the MLSMR dataset, and consequently these results represent an evaluation of these methods on a dataset that is entirely independent from that used to train the respective models. All the results in this table were obtained using the GF descriptor representation of the compounds. In addition, the performance of the random scheme is also presented, in order to establish a baseline for the ROC50 and PA100 values.

These results show that both DR_{SVM} and RS_{SVM} perform well and lead to ROC and ROC50 scores that are considerably higher than those obtained by the baseline approaches. The DR_{SVM} model achieves the best overall performance and obtains an ROC score of 0.888 and an ROC50 score of 0.713. Among the two baseline schemes, MAXSM_{SIM} performs the best (as it was the case on the MLSMR dataset) and achieves ROC and ROC50 scores of 0.711 and 0.378, respectively. The fact that both RS_{SVM} and MAXSM_{SIM} achieve a PA100 score of 100 but the former achieves a much higher ROC50 score, indicate that MAXSM_{SIM} can easily identify some of the positive compounds, but its performance degrades considerably when the somewhat lower ranked compounds are considered.

In order to show that the SVM itself contributes to the performance of the DR_{SVM} method as opposed to the training set only, we also computed the performance of a simple method in which the score of each compound is computed as the difference of the maximum similarity to the compounds in the DR⁺ and DR⁻ sets. The method is labeled MAXDR_{SIM} in the Table 5. The lower performance of this method compared to the DR_{SVM} (and even RS_{SVM}) indicates the advantage of using SVM as opposed to the simple similarity to the training sets. Note that we also computed the performance of the method based on the maximum similarity to the positive class (DR⁺) alone but found its performance to be slightly worse than that of the MAXDR_{SIM}.

The RS_{SVM} model achieves the second best performance, but in absolute terms its performance is considerably worse than that achieved by DR_{SVM} and also its own performance on the MLSMR dataset. These results should not be surprising as the RS_{SVM} model was trained on positive and negative compounds that were identified via a retrosynthetic approach based on a small set of reactions. As the decomposition statistics in Table 1 indicate, only 10.6% of the EL⁺ compounds can be decomposed in a small number of steps using that set of reactions. Consequently, according to the positive and negative class definitions used for training the RS_{SVM} model, the Eli Lilly dataset contained only 10.6% positive compounds, which explains RS_{SVM}'s performance degradation. However, what is surprising is that despite this, RS_{SVM} is still able to achieve good prediction performance, indicating that it can generalize beyond the set of reactions that were used to derive its training set.

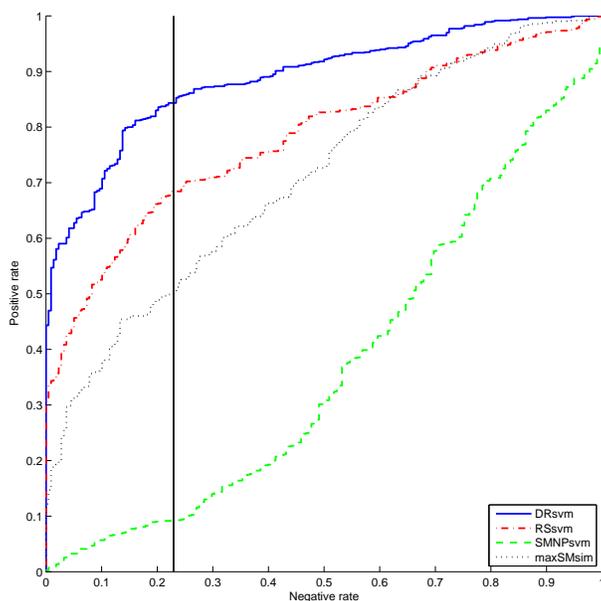


Figure 4: ROC curves for ranking the compounds in the Eli Lilly dataset using different methods (vertical line indicates the edge where ROC50 is calculated).

4.3 Generalization to New Reactions

A key parameter of the RSsvm model is the set of reactions that were used to derive the sets of positive and negative training instances during the retrosynthetic decomposition. As the discussion in Section 4.2 alluded, one of the key questions on the generality of the RSsvm-based approach is its ability to correctly predict the synthetic accessibility of compounds that can be synthesized by the use of additional reactions. To answer this question, we performed the following experiment, which is designed to evaluate RSsvm’s generalizability to new reactions.

First, we selected 20 out of our 23 reactions whose elimination had an impact on the decomposition tree depth of the MLSMR compounds and randomly grouped them into four sets, each having five reactions. Then for each training and testing dataset of the 5-fold cross validation framework used in the experiments of Section 4.1, we performed the following operations that are presented graphically in Figure 5. For each group of reactions, we eliminated its reactions from the set of 23 reactions and used the remaining 18 reactions to determine the shortest synthesis path of each RS^+ compound in the training set (RS^+ -train) based on retrosynthetic decomposition (Section 2.3.1). This retrosynthetic decomposition resulted in partitioning the RS^+ -train compounds into three subsets, referred to as P^+ -train, P^x -train, and P^- -train. The P^+ -train subset contains the compounds that, even with the reduced set of reactions, still had a synthesis path of length up to three, the P^x -train subset contains the compounds whose synthesis path became either four or five, and the P^- -train subset contains the compounds whose synthesis path became higher than five. Thus, the elimination of the reactions turned the P^- -train compounds into hard to synthesize compounds from a retrosynthetic analysis perspective. We then built an SVM model using as pos-

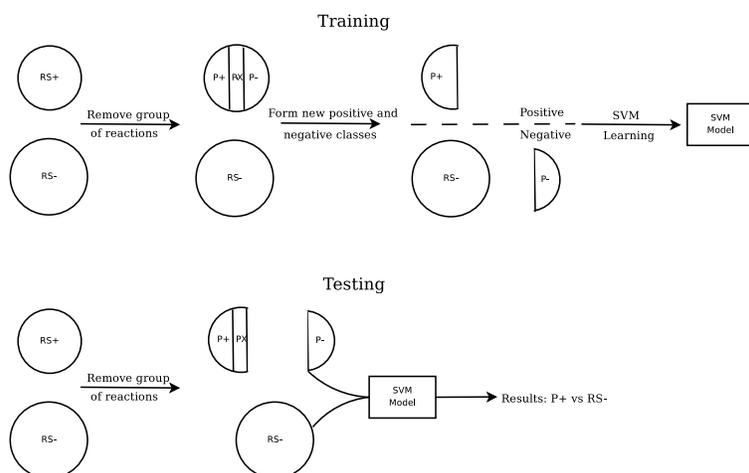


Figure 5: Schematic representation of the training and testing with a group of reactions eliminated. (P^+ , P^- and P^x are the subsets of RS^+ that are easy to synthesize (path ≤ 3), hard to synthesize (path > 5) or in between, respectively, after a group of reactions is eliminated.)

Table 6: Performance when extra reactions are used to determine the positive test compounds.

Group	Reactions in group	% molecules [†] moved to P^- -train	RSSVM			MAXSMSIM			Random	
			ROC	ROC50	PA100 (Enrich.)	ROC	ROC50	PA100 (Enrich.)	ROC50	PA100
1	8,9,10,20,21	17.3%	0.557	0.005	22.2 (1.3)	0.673	0.002	14.2 (0.8)	0.004	16.8
2	1,3,6,14,15	12.2%	0.752	0.006	27.6 (1.7)	0.668	0.007	34.2 (2.0)	0.004	16.7
3	2,7,16,17,18	14.2%	0.690	0.006	32.6 (1.7)	0.649	0.005	22.2 (1.2)	0.004	18.7
4	4,5,11,12,19	38.6%	0.717	0.004	41.2 (1.2)	0.631	0.004	37.4 (1.1)	0.004	33.6

[†] The percentages refer to all 48,621 easy to synthesize compounds. The proportions for the RS^+ set after selecting maximal independent set are slightly different.

itive instances only the molecules in P^+ -train and as negative instances the union of the molecules in RS^- -train and P^- -train. The P^x -train subset, which comprised less than 1% of the RS^+ , was discarded. The performance of this model was assessed by first using the above approach to split the RS^+ compounds of the testing set into the corresponding P^+ -test, P^x -test, and P^- -test subsets, and then creating a testing set whose positive instances were the compounds in P^- -test and whose negative instances were the compounds in RS^- -test. Note that since the short synthesis paths (i.e., ≤ 3) of the compounds in P^- -test require the eliminated reactions, this testing set allows us to assess how well the SVM model can generalize to compounds whose short synthesis paths depend on reactions that were not used to determine the positive training compounds.

Table 6 shows the results of these experiments. This table also lists the reactions in each group and the fraction of all 48,621 easy to synthesize molecules in RS^+ that became hard to synthesize after a particular group of reactions was eliminated in the retrosynthetic decomposition. For comparison purposes, this table also shows the performance of the best-performing baseline method (MAXSMSIM) and that of the random scheme. Note, that since a maximal independent set of RS^+ is used for training and testing, the percentages of the molecules that became negative in the latter are slightly different from those for the whole RS^+ set, which explains why the PA100 values for the random scheme are not proportional to the percentage of molecules that became negative.

These results show that, as expected, the relative performance of the RSSVM model on these test sets decreases when compared to the results reported in Table 4. This reflects the considerably

harder nature of this classification problem as it designed to test how well RSsvm generalizes to compounds that can only be synthesized when additional reactions are used. However, despite this, for three out of the four reaction groups, RSsvm is still able to achieve reasonably good performance, obtaining ROC scores ranging from 0.690 to 0.752. Also, its performance compares favorably to the MAXSMSIM scheme for three out of the four subsets, as it achieves better ROC, ROC50, and PA100 scores. Note that we also assessed the performance of these models on testing sets whose positive compounds were the P⁺-test subsets. The performance achieved by both RSsvm and MAXSMSIM was comparable to that reported earlier.

4.4 Computational Requirements

An advantage of the SVM-based synthetic accessibility approach over a similar approach that is based entirely on retrosynthetic decomposition is its prediction speed. For example, in our study, during the cross-validation experiments with the MLSMR dataset, the RSsvm models were processing about 6,700 molecules per minute. This is significantly better than the retrosynthetic approaches which process only several molecules per minute or less. In general, the amount of time required by the SVM-based approach to rank the compounds is proportional to the size of the model (i.e., the number of support vectors), which in turn depends on the number of the training compounds and the separability of the positive and negative classes. Even though its absolute computational performance will change based on the training set, we expect it to remain considerably faster than the retrosynthetic-based approach. The time required to train SVM models depends on the number of vectors in the training sets. Thus, the RSsvm model required 86 minutes for training while the DRsvm model, which is trained on a larger number of molecules, required almost 9 hours. The classification of the MLSMR molecules into DR⁺ and DR⁻ sets required 69 minutes to complete. Note that training set preprocessing and SVM model training are operations that need to be performed only once.

5 Discussion & Conclusions

In this paper we presented and evaluated two different approaches for predicting the synthetic accessibility of chemical compounds that utilize support vector machines to build models based on the descriptor representation of the compounds. These two approaches exhibit different performance characteristics and are designed to address two different application scenarios. The first approach, corresponding to the RSsvm-based models, is designed to identify the compounds that can be synthesized using a specific set of reactions and starting materials library, whereas the second approach, corresponding to the DRsvm-based models, is designed to provide a more general assessment of synthetic accessibility that is not tied to any set of reactions or starting materials databases.

The results presented in Section 4.1 showed that the RSsvm-based models are very effective in identifying the compounds that can be synthesized in a small number of steps using the set of reactions and starting materials under consideration. As a result, these models can be used to prioritize the compounds that will need to be retrosynthetically analyzed in order to confirm their synthetic accessibility and identify the desired synthesis path, and as such significantly reduce the amount of time required by retrosynthetic analysis approaches. In addition, the experiments

showed that even when additional reactions and/or starting materials are required for the synthesis of a compound, the RSsvm-based models are still able to prioritize a set of compounds based on their synthetic accessibility (Sections 4.2 and 4.3). However, as expected, the quality of the predictions are lower in such cases.

The results presented in Section 4.2 showed that the approach utilized by the DRsvm-based model, which automatically identifies from a large diverse library a set of positive and negative training instances based on their similarity to other compounds, lead to an effective synthetic accessibility prediction model. This approach outperforms both the baseline approaches and the RSsvm model, since the latter is biased towards the set of reactions and starting materials used for identifying its training instances.

Acknowledgement

This work was supported by NSF ACI-0133464, IIS-0431135, NIH R01 LM008713A, and by the Digital Technology Center at the University of Minnesota. We would like to thank Dr. Ian Watson at Eli Lilly and Company who graciously agreed to evaluate our models with their compounds.

References

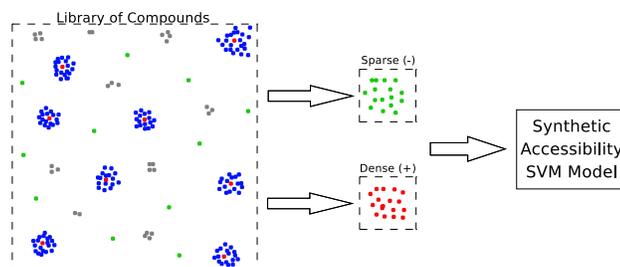
- (1) Pearlman, D. A.; Murcko, M. CONCERTS: Dynamic collection of fragments as an approach to de novo ligand design. *J. Med. Chem.* **1996**, *39*, 1651–1663
- (2) Gillet, V.; Johnson, A.; Mata, P.; Sike, S.; Williams, P. SPROUT: A program for structure generation. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 127–153
- (3) Boda, K.; Johnson, A. P. Molecular Complexity Analysis of de Novo Designed Ligands. *J. Med. Chem.* **2006**, *49*, 5869–5879
- (4) Schneider, G.; Lee, M.-L.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drugderived building blocks. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 487–494
- (5) Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Aken, K. V.; Janssen, P. A. J. SYNOPSIS: SYNthesize and OPTimize System in Silico. *J. Med. Chem.* **2003**, *46*, 2765–2773
- (6) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP—retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522
- (7) Bertz, S. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599–3601
- (8) Hendrickson, J.; Toczko, P. H. A. Molecular complexity: a simplified formula adapted to individual atoms. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 63–67
- (9) Whitlock, H. On the Structure of Total Synthesis of Complex Natural Products. *J. Org. Chem.* **1998**, *63*, 7982–7989
- (10) Barone, R.; Chanon, M. A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 269–272
- (11) Rücker, C.; Rücker, G.; Bertz, S. Organic Synthesis - Art or Science? *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 378–386

- (12) Allu, T.; Oprea, T. Rapid Evaluation of Synthetic and Molecular Complexity for in Silico Chemistry. *J. Chem. Inf. Model.* **2005**, *45*, 1237–1243
- (13) Johnson, A.; Marshall, C.; Judson, P. Starting material oriented retrosynthetic analysis in the LHASA program. 1. General description. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 411–417
- (14) Gillet, V.; Myatt, G.; Zsoldos, Z.; Johnson, A. SPROUT, HIPPO and CAESA: Tool for de novo structure generation and estimation of synthetic accessibility. *Perspect. Drug Discovery Des.* **1995**, *3*, 34–50
- (15) Pförtner, 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
- (16) Boda, K.; Seidel, T.; Gasteiger, J. Structure and reaction based evaluation of synthetic accessibility. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 311–325
- (17) 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
- (18) Wale, N.; Watson, I.; Karypis, G. Indirect Similarity Based Methods for Effective Scaffold-Hopping in Chemical Compounds. *J. Chem. Inf. Model.* **2008**, *48*, 730–741
- (19) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000
- (20) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, Germany, 2009
- (21) Rogers, D.; Brown, R.; Hahn, M. Using Extended-Connectivity Fingerprints with Laplacian-Modified Bayesian Analysis in High-Throughput Screening. *J. Biomol. Screen.* **2005**, *10*, 682–686
- (22) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, 1995
- (23) Aizerman, M.; Braverman, E.; Rozonoer, L. Theoretical foundations of the potential function method in pattern recognition learning. *Autom. Rem. Cont.* **1964**, *25*, 821–837
- (24) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A training algorithm for optimal margin classifiers. *5th Annual ACM Workshop on COLT*, Pittsburgh, PA, 1992; pp 144–152
- (25) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996
- (26) Geppert, H.; Horvath, T.; Gartner, T.; Wrobel, S.; Bajorath, J. Support-Vector-Machine-Based Ranking Significantly Improves the Effectiveness of Similarity Searching Using 2D Fingerprints and Multiple Reference Compounds. *J. Chem. Inf. Model.* **2008**, *48*, 742–746
- (27) Chen, J.; Holliday, J.; Bradshaw, J. A Machine Learning Approach to Weighting Schemes in the Data Fusion of Similarity Coefficients. *J. Chem. Inf. Model.* **2009**, *49*, 185–194
- (28) Corey, E. J. Retrosynthetic Thinking — Essentials and Examples. *Chem. Soc. Rev.* **1988**, *17*, 111–133
- (29) Corey, E. J.; Cheng, X.-M. *The Logic of Chemical Synthesis*; Wiley: New York, 1995
- (30) Bajorath, J. Chemoinformatics methods for systematic comparison of molecules from natural and synthetic sources and design of hybrid libraries. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 431–439
- (31) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucl. Acids Res.* **2009**, *37*, W623–W633

- (32) Wale, N.; Karypis, G. Acyclic Subgraph-based Descriptor Spaces for Chemical Compound Retrieval and Classification. *IEEE International Conference on Data Mining (ICDM)*, 2006
- (33) AFGen 2.0. <http://glaros.dtc.umn.edu/gkhome/afgen/overview> (accessed July 1, 2009)
- (34) GenerateMD, v. 5.2.0, ChemAxon. <http://www.chemaxon.com/jchem/doc/user/GenerateMD.html> (accessed July 1, 2009)
- (35) OEChem TK, OpenEye Scientific Software. <http://www.eyesopen.com/products/toolkits/oechem.html> (accessed July 1, 2009).
- (36) Joachims, T. Making large-scale support vector machine learning practical. In *Advances in kernel methods: support vector learning*; Schölkopf, B., Burges, C. J. C., Smola, A. J., Eds.; MIT Press: Cambridge, MA, USA, 1999; pp 169–184
- (37) Fawcett, T. An introduction to ROC analysis. *Pattern Recog. Lett.* **2006**, *27*, 861–874
- (38) Gribskov, M.; Robinson, N. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.* **1996**, *20*, 25–33

For Table of Contents Only

Assessing Synthetic Accessibility of Chemical Compounds
Using Machine Learning Methods



Yevgeniy Podolyan, Michael A. Walters, and George Karypis