SYBA: Bayesian estimation of synthetic accessibility of organic compounds

Milan Voršilák
Ustav molekularni genetiky Akademie Ved Ceske Republiky https://orcid.org/0000-0002-8923-1627
Michal Kolář
Ustav molekularni genetiky Akademie Ved Ceske Republiky https://orcid.org/0000-0002-4593-1525
Ivan Čmelo
University of Chemistry and Technology Prague https://orcid.org/0000-0001-7787-8653
Daniel Svozil (daniel.svozil@gmail.com)
University of Chemistry and Technology https://orcid.org/0000-0003-2577-5163

Research article

Keywords: synthetic accessibility, Bayesian analysis

Posted Date: February 4th, 2020

DOI: https://doi.org/10.21203/rs.2.22597/v1

Version of Record: A version of this preprint was published at Journal of Cheminformatics on May 20th, 2020. See the published version at https://doi.org/10.1186/s13321-020-00439-2.

SYBA: Bayesian estimation of synthetic accessibility of organic compounds

Milan Voršilák[1,2], Michal Kolář[3,4], Ivan Čmelo[1], Daniel Svozil[1,2,*]

[1] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, 166 28 Prague 6, Czech Republic.

[2] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics of the Czech Academy of Sciences, Vídeňská 1083, 142 20 Prague 4, Czech Republic.

[3] Laboratory of Genomics and Bioinformatics, Institute of Molecular Genetics of the Czech Academy of Sciences, Vídeňská 1083, 142 20 Prague 4, Czech Republic.

[4] Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, 166 28 Prague 6, Czech Republic.

[*] corresponding author, e-mail: Daniel.Svozil@vscht.cz

Email addresses:

MV: milan.vorsilak@img.cas.cz, ORCID: 0000-0002-8923-1627

MK: kolarmi@img.cas.cz, ORCID: 0000-0002-4593-1525

IČ: ivan.cmelo@vscht.cz, ORCID: 0000-0001-7787-8653

DS: daniel.svozil@vscht.cz, ORCID: 0000-0003-2577-5163

Abstract

SYBA (SYnthetic Bayesian Accessibility) is a fragment based method for the rapid classification of organic compounds as easy- (ES) or hard-to-synthesize (HS). SYBA is based on the Bayesian analysis of the frequency of molecular fragments in the database of ES and HS molecules. It was trained on ES molecules available in the ZINC15 database and on HS molecules generated by the Nonpher methodology. SYBA was compared with a random forest, that was utilized as a baseline method, as

well as with other two methods for synthetic accessibility assessment: SAScore and SCScore. When used with their suggested thresholds, SYBA improves over random forest classification, albeit marginally, and outperforms SAScore and SCScore. However, with thresholds optimized by the analysis of ROC curves, SAScore improves considerably and yields similar results as SYBA. Because SYBA is based merely on fragment contributions, it can be used for the analysis of the contribution of individual molecular parts to compound synthetic accessibility. Though SYBA was developed to quickly assess compound synthetic accessibility, its underlying Bayesian framework is a general approach that can be applied to any binary classification problem. Therefore, SYBA can be easily re-trained to classify compounds by other physico-chemical or biological properties. SYBA is publicly available at https://github.com/lich-uct/syba under the GNU General Public License.

Keywords
synthetic accessibility – Bayesian analysis

Background
Chemical space available for the generation of new molecules is huge [1-4], making the synthesis and testing of all possible compounds impractical. Therefore chemists, both experimental and computational, developed tools and approaches for the exploration of chemical space with the aim to identify new compounds with desirable physico-chemical, biological and pharmacological properties [5-12]. A major *in silico* method for chemical space exploration is *de novo* molecular design in which new virtual molecules are assembled from scratch [13-18]. An essential requirement for *de novo* designed compounds is their synthetic accessibility. Synthetic accessibility is commonly incorporated into *de novo* design programs by employing chemical strategies that guide an assembly process. Between these strategies belong, for example, the prevention of connections between certain atom types [19], the incorporation of retrosynthetic rules [20, 21] or the use of established chemical reactions to connect individual molecular building blocks [22, 23].

The latest development in *de novo* molecular design are molecular generators based on deep learning approaches [24-26]. These typically construct new molecules not by assembling the building blocks, but by producing chemically feasible SMILES strings [27-32]. The generators are able to produce

millions of virtual compounds, synthetic accessibility of which has to be quickly and efficiently assessed. Quick synthetic accessibility assessment can be based [33] on molecule's complexity that is typically calculated [34-37] from the number of atoms, bonds, rings, and/or hard-to-synthesize motifs, such as chiral centers or uncommon ring fusions. However, the definition of molecular complexity is ambiguous and context dependent [38, 39]. The structural complexity is not equivalent to the synthetic one as complexity-based metrics do not incorporate any information about starting materials and tend to remove molecules that can be synthesized from already existing complex precursors [40, 41]. A better way of synthetic accessibility assessment is to use the complexity of the synthetic route [42]. Based on this principle, SCScore, a data-driven metric designed to describe real syntheses, was developed recently [43]. SCScore is based on the idea that reaction products are synthetically more complex than reactants. To quantify this, a deep feed-forward neural network, that assigns a synthetic complexity score between 1 and 5, was trained on 22 million reactant-product pairs from the Reaxys database [44]. Using the hinge loss objective function, that supports the separation between scores in each reactant–product pair, the model learns synthetic complexity score that correlates with the number of reaction steps, but does not rely on the availability of reaction database or organic chemist ranking.

SAScore [45], another popular and rapid method for synthetic accessibility assessment, is based on the analysis of ECFP4 [46] fragments obtained from one million structures randomly selected from the PubChem database [47]. The main idea of SAScore is that when a molecular fragment occurs often in the PubChem database, it contributes to the synthetic accessibility of a molecule more than a less frequently occurring fragment. Each fragment is assigned a numerical score that is higher for frequent fragments and lower for infrequent ones. In addition to the fragment score, SAScore consists of a complexity penalty and symmetry bonus. These terms penalize nonstandard structural motives such as macrocycles, stereo centers, spiro and bridge atom, but reward the symmetry of a structure. SAScore acquires values between 1 and 10, where 6 is suggested by the authors [45] as a threshold to distinguish between easy- and hard-to-synthesize compounds. SAScore is a popular high-throughput measure and proved to be a very useful tool in many cheminformatics applications [27, 48-50].

In the present work, we further expand on main concepts of SAScore construction. We developed SYBA (SYnthetic Bayesian Accessibility), a rapid fragment-based score derived using Bayesian probabilistic modeling. Fragment contributions to SYBA are calculated not only from fragments present in synthetically accessible molecules, but also from fragments appearing in difficult-to-synthesize molecules. The SYBA score provides additional means of rapid synthetic accessibility evaluation that can complement other existing metrics.

Methods

SYBA score derivation

SYBA is derived from the Bayesian analysis of the frequency of molecular fragments that are present in the database of easy-to-synthesize (ES) and hard-to-synthesize (HS) molecules and under the assumption of the independence of molecular fragments. Though such assumption is bold, it was shown to provide surprisingly good results in many cheminformatics studies [51-55].

Each compound is represented by a binary fingerprint $\boldsymbol{F} = [f_1, f_2, \dots, f_M]$ of length M where f_i indicates the presence ($f_i = 1$) or absence ($f_i = 0$) of the specific fragment i in the compound. SYBA uses this fingerprint to assign the molecule to a class $C \in \langle \mathrm{ES}, \mathrm{HS} \rangle$. The calculation is based on the Bayes theorem

Equation 1 $$p(C|\boldsymbol{F}) = \frac{p(\boldsymbol{F}|C)\, p(C)}{p(\boldsymbol{F})},$$

where $p(C|\boldsymbol{F})$ is the posterior probability that a compound with a certain set of molecular fragments $\boldsymbol{F}$ belongs to the class C. The likelihood $p(\boldsymbol{F}|C)$ is the conditional probability that a compound from the class C contains a set of molecular fragments $\boldsymbol{F}$. The marginal probabilities $p(\boldsymbol{F})$ and $p(C)$ express our belief to observe a set of molecular fragments $\boldsymbol{F}$ and the molecule that belongs to the class C.

The SYBA score is defined as the logarithm of the ratio of the posterior probabilities that the molecule belongs to the ES and HS classes,

Equation 2 $$\mathrm{SYBA}(\boldsymbol{F}) = \ln\left(\frac{p(\mathrm{ES}|\boldsymbol{F})}{p(\mathrm{HS}|\boldsymbol{F})}\right).$$

Using Equation 1, the SYBA score can be expressed as

Equation 3 $$\mathrm{SYBA}(\boldsymbol{F}) = \ln\left(\frac{p(\mathrm{ES})}{p(\mathrm{HS})}\right) + \ln\left(\frac{p(\boldsymbol{F}|\mathrm{ES})}{p(\boldsymbol{F}|\mathrm{HS})}\right).$$

In the data set SYBA was derived from (further referred to as the training data set S), ES and HS compounds are represented evenly, the priors $p(\mathrm{ES})$ and $p(\mathrm{HS})$ are thus equal and the term $\ln\left(\frac{p(\mathrm{ES})}{p(\mathrm{HS})}\right)$ becomes zero:

Equation 4 $$\mathrm{SYBA}(\boldsymbol{F}) = \ln\left(\frac{p(\boldsymbol{F}|\mathrm{ES})}{p(\boldsymbol{F}|\mathrm{HS})}\right).$$

Assuming the independence of molecular fragments, the conditional probability $p(\boldsymbol{F}|C)$ factorizes to $p(\boldsymbol{F}|C) = \prod_{i=1}^{M} p(f_i|C)$ and the SYBA score simplifies to

Equation 5 $$\mathrm{SYBA}(\boldsymbol{F}) = \sum_{i=1}^{M} s_i(f_i)$$

where $s_i(f_i)$ is the score contribution from the fragment *i* (SYBA fragment score) given as

Equation 6 $$s_i(f_i) = \ln\left(\frac{p(f_i|\mathrm{ES})}{p(f_i|\mathrm{HS})}\right).$$

SYBA parameter estimation

In our model, we distinguish between the presence or absence of the fragment i in the molecule, which is expressed by $f_i = 1$, resp. $f_i = 0$. The collection of f_i values for all compounds of some class C forms a finite sequence of binary random variables f_i and can be modeled by a Bernoulli process [56]. This Bernoulli process is parametrized by the probability $w_{C,i}$ to find the fragment i in the molecule belonging to the class C. This parameter expresses our prior knowledge that the molecule belongs to the class C and it is a random variable per se. Therefore, it holds

Equation 7 $$p(f_i|C) = \int_0^1 p\left(f_i\middle|w_{C,i}, C\right) p(w_{C,i}|\, C)\, \mathrm{d}w_{C,i}.$$

Here, the prior $p(w_{C,i}|C)$ represents our knowledge that molecules from the class C contain the fragment i. The prior $p(w_{C,i}|C)$ is estimated from the training set S consisting of N_C molecules

belonging to the class C, $n_{C,i}$ of which contain the fragment i. In SYBA, a Bernoulli process conjugate prior, that enables analytic solution, is used. It can be expressed as the beta distribution [56]

Equation 8 $$p(w_{C,i}|C) = \frac{1}{\mathrm{B}(n_{C,i}+1, N_C - n_{C,i}+1)} {w_{C,i}}^{n_{C,i}} (1 - w_{C,i})^{N_C - n_{C,i}},$$

where B is the Euler beta function. The likelihood $p(f_i|w_{C,i}, C)$ in Equation 7 can be expressed as $p(f_i|w_{C,i}, C) = {w_{C,i}}^{f_i}(1 - w_{C,i})^{1-f_i}$. Thus, the posterior $p(f_i|C)$ in Equation 7 can be evaluated analytically using the beta function algebra:

Equation 9 $$P(f_i|C) = \int_0^1 \frac{1}{\mathrm{B}(n_{C,i}+1, N_C - n_{C,i}+1)} {w_{C,i}}^{n_{C,i}+f_i} \left(1 - w_{C,i}\right)^{N_C+1-n_{C,i}-f_i} \mathrm{d}w_{C,i} =$$

$$\frac{\mathrm{B}(n_{C,i} + f_i + 1, N_C + 1 - n_{C,i} - f_i + 1)}{\mathrm{B}(n_{C,i} + 1, N_C - n_{C,i} + 1)}$$

Since all variables are integers, Equation 9 can be rewritten in a factorial form and further simplified to

Equation 10 $$p(f_i|C) = \frac{1}{N_C+2} (n_{C,i} + 1)^{f_i} (N_C - n_{C,i} + 1)^{1-f_i}.$$

Plugging this estimate to Equation 6 yields

Equation 11 $$s_i(f_i) = \ln \frac{N_{\mathrm{HS}}+2}{N_{\mathrm{ES}}+2} + f_i \ln \frac{(n_{\mathrm{ES},i}+1)}{(n_{\mathrm{HS},i}+1)} + (1 - f_i) \ln \frac{(N_{\mathrm{ES}} - n_{\mathrm{ES},i}+1)}{(N_{\mathrm{HS}} - n_{\mathrm{HS},i}+1)}$$

where

- $s_i(f_i)$ is the score contribution from the fragment *i* (SYBA fragment score)
- f_i indicates the presence ($f_i = 1$) or absence ($f_i = 0$) of the specific fragment *i* in the compound
- N_{HS} is the number of HS and N_{ES} the number of ES molecules in the training data set S. If $N_{\mathrm{HS}} = N_{\mathrm{ES}}$, the first term in Equation 11 becomes zero.
- $n_{\mathrm{HS},i}$ is the number of HS molecules in the training data set S that contain the fragment *i*
- $n_{\mathrm{ES},i}$ is the number of ES molecules in the training data set S that contain the fragment *i*

Positive $s_i(f_i)$ means that the presence/absence of the fragment *i* is more probable in ES than in HS class and vice versa. The final SYBA score of a molecule is the sum of $s_i(f_i)$ contributions over all considered fragment types (Equation 5). Positive SYBA means that the compound belongs more likely to the ES class, while negative SYBA means that the compound belongs more likely to the HS class. The higher the absolute value of SYBA, the more evidence for the class membership is present in the molecule.

Training set construction

The training data set S consists of two subsets: S_+ contains ES structures and S_- contains HS structures (Figure 1, Additional file 1). While ES molecules can be readily obtained, for example, from the ZINC database of purchasable compounds [57, 58], no equivalent database of HS molecules exists. However, HS molecules can be designed by Nonpher [59], a method based on a molecular morphing approach [60]. In Nonpher, a starting molecule is gradually transformed into a more complex compound using small structural perturbations, such as the addition or removal of an atom or a bond. During this process, four complexity indices (Bertz [34], Whitlock [35], BC [36] and SMCM [37]) are monitored and if the generated compound exceeds at least one complexity threshold, it is considered to be HS. The thresholds of the complexity indices were determined [59] by the analysis of the complexity distribution of 22 723 223 ZINC compounds for eleven 50 Da wide molecular weight (*MW*) bins. In the given *MW* bin, a compound is considered to be HS if at least one complexity index is greater than its value for 99.9% ZINC compounds (Additional file 2 – Table S1).

Using Nonpher, 693 353 HS molecules were generated and they form the S_- data set. The S_+ data set containing ES compounds is formed by the same amount of molecules randomly chosen (excluding natural products) from the ZINC15 database [58] so that their distribution of the number of heavy atoms is the same as in the S_- data sets. Every S_+ and S_- molecule was fragmented using the Morgan fingerprint function in the RDKit toolkit [61]. Fragments with the radius of 2, corresponding to radial ECFP4 [46] fragments, were used. This type of fragments consists of a central atom and neighboring atoms connected to the central atom by one or two bonds.

Test set construction

SYBA performance could have been assessed using a test set created in a similar way as the training set S, i.e. using HS compounds generated by Nonpher. However, such test set would be biased towards chemical space covered by Nonpher. Therefore, two test sets were constructed in conceptually different manner. First test set, further denoted as T_{MC}, was manually curated from the literature, second test set, referred to as T_{CP}, was computationally picked from the ZINC15 [58] and GDB17 databases [62].

HS compounds in T_{MC} (denoted as T_{MC-}) were obtained by the analysis [59] of 296 published compounds assessed by experienced medicinal chemists [41, 45, 63, 64]. Based both on original chemists' scores, as well as on scores given by SAScore [45], FA4 [64], SYLVIA [63] and RASA [41] methods, the final T_{MC-} data set of 40 HS compounds was assembled. A complementary T_{MC+} data set consists of 40 ES compounds selected from the ZINC15 database [58] in such a way that the distribution of the number of their heavy atoms is the same as in the T_{MC-} data set. Because small T_{MC} size may bias the results, 30 different T_{MC} data set instances were generated using the same 40 T_{MC-} compounds, but different 40 T_{MC+} compounds (Additional file 3).

HS compounds in the T_{CP} test set (Additional file 4), denoted as T_{CP-}, were obtained by the analysis of the publicly available subset of 50M molecules from the GDB-17 database [62]. Only molecules exceeding thresholds (Additional file 2 – Table S1) of all 4 monitored complexity indices (Bertz [34], Whitlock [35], BC [36] and SMCM [37]) were considered to be HS. In total, 3 581 molecules form the T_{CP-} data set. A complementary T_{CP+} data set consists of the same number of compounds randomly selected from the ZINC15 database [58] that follow the same size distribution as HS compounds and that, in addition, do not exceed any of the aforedescribed complexity indices. Data sets used in the present work are summarized in Figure 1.

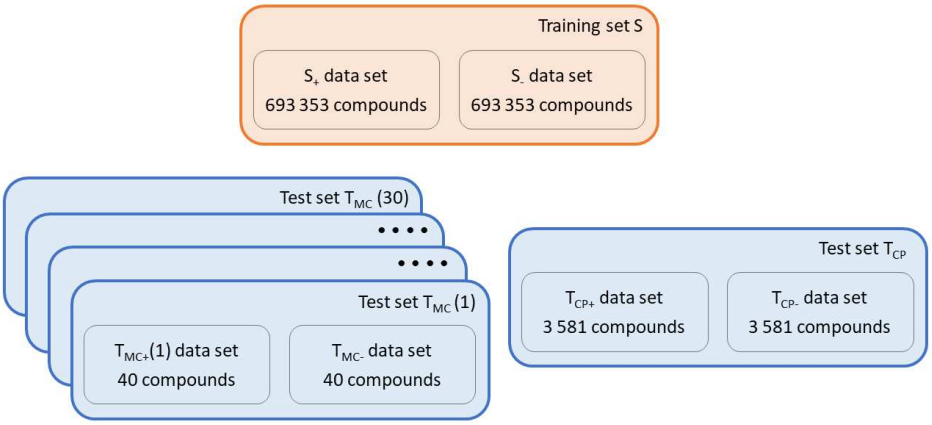

Figure 1 Data sets summary. Training set was used to derive SYBA scores, as well as to train a random forest classifier. Training set consists of 693 353 molecules randomly selected from the ZINC15 database [58] that are considered to be ES (S_+ data set) and of same number of HS molecules generated by Nonpher [59] (S_- data set). Two test sets were used to compare the performance of SYBA, a random forest, SAScore [45] and SCScore [43]. Manually curated test set (T_{MC}) contains 40 compounds (T_{MC-} data set) considered to be HS by experienced medicinal chemists [59] suplemented by 40 ES compounds randomly selected from the ZINC15 database (T_{MC+} data set). 30 T_{MC} data set instances differing in T_{MC+} compounds were constructed. Computationally picked test set (T_{CP}) consists of 3 581 HS compounds that were obtained from the GDB-17 database [62] (T_{CP-} data set) complemented by same number of compounds randomly selected from the ZINC15 database (T_{CP+} data set).

Performance evaluation

The performance of classification models studied in the present work was assessed by four different metrics: the classification accuracy (*Acc*), sensitivity (*SN*), specificity (*SP*) and area under the ROC curve (*AUC*). *Acc* gives the percentage of correctly classified samples regardless of their class.

Equation 12 $$\text{Accuracy } (Acc) = \frac{TP+TN}{TP+TN+FN+FP}$$

where true positives (*TP*) are ES compounds predicted by a model to be ES, true negatives (*TN*) are HS compounds predicted to be HS, false positives (*FP*) are HS compounds predicted to be ES and false negatives (*FN*) are ES compounds predicted to be HS. The accuracy can also be evaluated for positive and negative classes independently leading to *SN* and *SP*. *SN* is the percentage of a correctly predicted positive class compounds, while the percentage of a correctly predicted negative class compounds is known as *SP*.

Equation 13 $$\text{Sensitivity } (SN) = \frac{TP}{TP+FN}$$

Equation 14 $$\text{Specificity } (SP) = \frac{TN}{TN+FP}$$

SN and *SP* can be combined in the receiver operating characteristic (ROC) curve that is the graphical representation of the trade-off between true positive rate (given as *SN*) and false positive rate (given as $1 - SP$) over all possible thresholds (Figure 2). The area under the ROC curve (*AUC*) is the quantitative measure of the performance of a classifier and is equal to the probability that a classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. A random classifier has *AUC* of 0.5, while *AUC* for a perfect classifier is equal to 1.

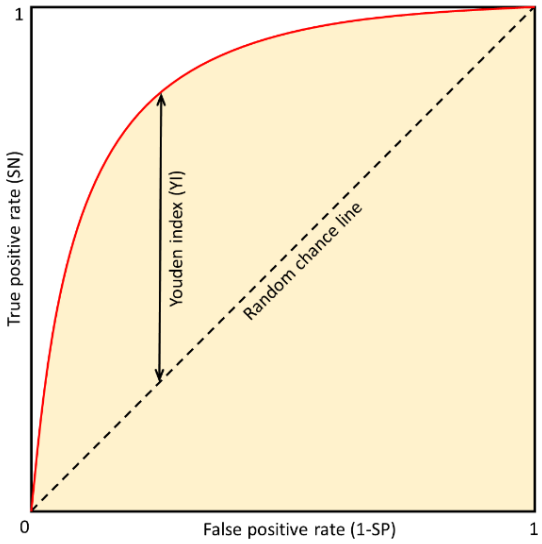

Figure 2 ROC curve and Youden index. The ROC curve (red line) is the dependency of true positive rate (it equals to *SN*) on false positive rate (it equals to $1 - SP$) at various thresholds. The random chance line represents a classifier that assigns examples into individual classes randomly. Orange shaded area represents the area under the ROC curve (*AUC*). The larger the *AUC*, the better is the overall performance of the classifier. Youden index (*YI*) is the point on the ROC curve that is farthest from the random chance line along the *SN* axis.

Random forest classification, SAScore and SCScore

Because of its wide adoption in various cheminformatics projects [59, 65-67], random forest (RF) classification was used as a baseline method with which SYBA, SAScore [45] and SCScore [43] were compared. In RF, compounds were encoded by 1024-bits long Morgan fingerprint with radius 2 that corresponds to the ECFP4 fingerprint [46]. The RF classifier was implemented in Scikit-learn [68] and consisted of 100 trees [65] with Gini index used as a purity criterion to split a node. RF was trained using the training set S (Figure 1). SAScore was calculated using its implementation in the RDKit toolkit [61]. SCScore code was downloaded from the public GitHub repository [69].

Classification thresholds

In SYBA, more positive value means a higher probability that the compound is ES and more negative value indicates a higher probability that the compound is HS (Equation 4). The threshold value of zero is used to distinguish between ES and HS compounds. For SAScore, the recommended value of 6.0 [45] was used as a threshold. In RF, the final prediction is based on a number of decision trees that predict either of classes. Here, 0.5 is used as a threshold, i.e., if more decision trees predict ES than HS class, a compound is classified as ES and vice versa. For SCScore [43], no threshold was suggested by the authors. In such case, the threshold can be identified by the analysis of the ROC curve. A frequently used measure that enables the selection of an optimal threshold is the Youden index (*YI*) [70, 71]. *YI* is defined as

Equation 15 $$YI = \max(SN + SP - 1)$$

and ranges between 0 and 1 (Figure 2). The optimal threshold value is selected by maximizing *YI,* i.e., by maximizing the sum of *SN* and *SP*.

Results and discussion

SYBA identifies fragments typical for ES and HS compounds

In total, 224 659 ECFP4 fragments were obtained for ES compounds and 4 820 791 different fragments for HS compounds. 129 013 fragments are common for both S_+ and S_- subsets. 37.4 % of S_+ fragments and 78.8 % of S_- fragments are present only once in the whole data set S (singletons). Typical ES and HS fragments are shown in Figure 3 and Figure 4.

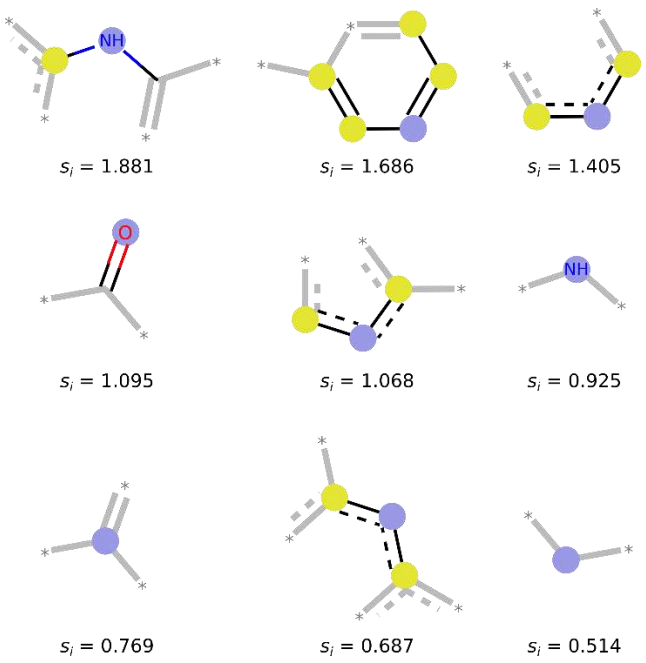

Figure 3 ES fragments enriched in the S_+ data set. Nine fragments that are most frequent in the S_+ data set and, at the same time, least frequent in the S_- data set. s_i is SYBA fragment score. Blue circle represents fragment's central atom, yellow circle represents an aromatic atom. Fragment images were generated by the RDKit function DrawMorganEnvs().

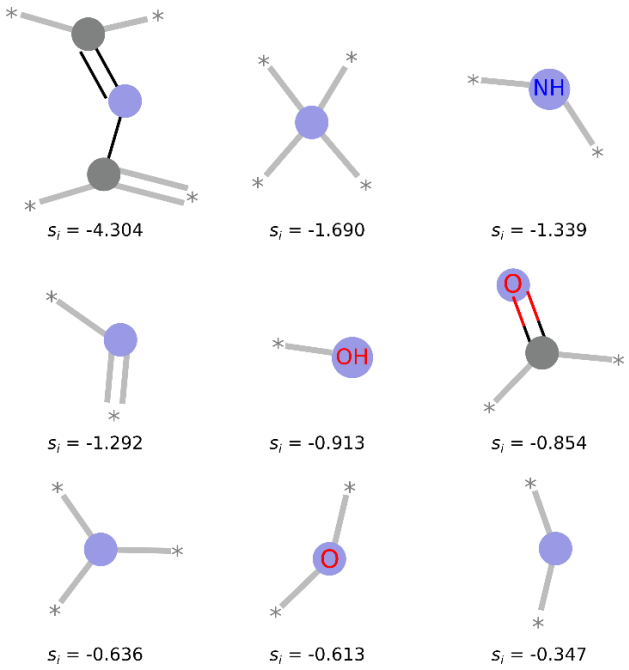

Figure 4 HS fragments enriched in the S- data set. Nine fragments that are most frequent in the S- data set and, at the same time, least frequent in the S_+ data set. s_i is SYBA fragment score. Blue circle represents fragment's central atom, gray circle represents aliphatic ring atom. Fragment images were generated by the RDKit function DrawMorganEnvs().

SYBA also enables the visualization and interpretation of fragment score contributions. Each SYBA fragment score can be projected to fragment root atom and this projection can be used to analyze which fragments contribute unfavorably to the ease of molecule's synthetic availability (Figure 5).

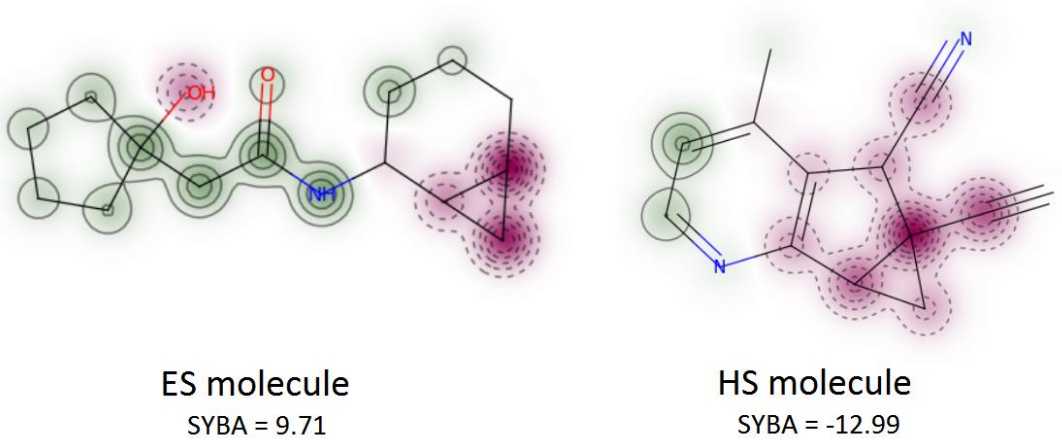

Figure 5 SYBA fragment score visualization. Fragment score is projected on fragment root atom and the whole molecule is visualized as a similarity map [72]. The more frequent the fragment is in the S_+ data set compared to the S_- data set, the greener is its central atom. Similarly, the more frequent the fragment is in the S_- data set compared to the S_+ data set, the redder is its central atom. This visualization enables to analyze the contributions of the individual parts of the molecule to its synthetic accessibility. In the HS molecule, the quaternary carbon is most problematic. Another substructure decreasing compound synthetic accessibility is a fused cyclopropane ring as can be observed both in ES and HS compounds.

SYBA outperforms current methods on the manually curated test set

The quality measures of the classification of the compounds in the manually curated T_{MC} test set (Figure 1), averaged over 30 T_{MC} instances, are summarized in Table 1, confusion matrices are reported in Additional file 2 – Panel S1 and Panel S2 and ROC curves are shown in Figure 6. When classifiers are used with their default thresholds, the best performing model, in terms of *Acc*, is SYBA followed by RF and SAScore. While SYBA and RF sensitivity and specificity are well balanced, SAScore shows high sensitivity (*SN* = 0.934, i.e., on average 93.4 % of *ES* compounds are predicted as *ES*), while its specificity (i.e., the ability to correctly classify *HS* compounds) is rather low (*SP* = 0.300). The observed high sensitivity of SAScore is not surprising as only 0.2 % of ZINC structures have SAScore greater than 6.0 and out of these, only lower units were selected into the T_{MC+} set. When thresholds are optimized to maximize *YI*, the most accurate models are SYBA, RF and SAScore followed afar by SCScore. Notable is the improvement of SAScore *SP* by 0.619 compared to the default threshold. The increase in SAScore *SP* comes, however, at the cost of *SN* that decreases by 0.135. The worst performing model, SCScore, is only slightly better (*AUC* = 0.528) than a random model. However, because T_{MC+} and T_{MC-} data sets consist of only 40 compounds each, the results must be interpreted with caution as small changes in confusion matrices lead to relatively large changes in reported metrics.

Model	*AUC*	*Acc*	*SN*	*SP*	Threshold
Default threshold					
SYBA	**0.905 (0.029)**	**0.859 (0.025)**	0.893 (0.050)	**0.825 (0.000)**	0.0
SAScore	0.865 (0.032)	0.617 (0.020)	**0.934 (0.040)**	0.300 (0.000)	6.0
RF	0.875 (0.027)	0.819 (0.026)	0.863 (0.051)	0.775 (0.000)	0.5
Optimized threshold					
SYBA	**0.905 (0.029)**	**0.878 (0.022)**	**0.887 (0.057)**	0.869 (0.040)	10.1 (9.3)
SAScore	0.865 (0.032)	0.859 (0.026)	0.799 (0.038)	**0.919 (0.036)**	3.9 (0.2)
SCScore	0.528 (0.036)	0.601 (0.032)	0.707 (0.077)	0.496 (0.053)	3.7 (0.2)
RF	0.875 (0.027)	0.842 (0.031)	0.855 (0.047)	0.828 (0.030)	0.6 (0.05)

Table 1 The performance of classification models for the manually curated T_{MC} test set. Quality measures *AUC*, *Acc*, *SN* and *SP*, as well as thresholds, are reported as their average values over 30 T_{MC} instances. Standard deviations are given in the parenthesis.

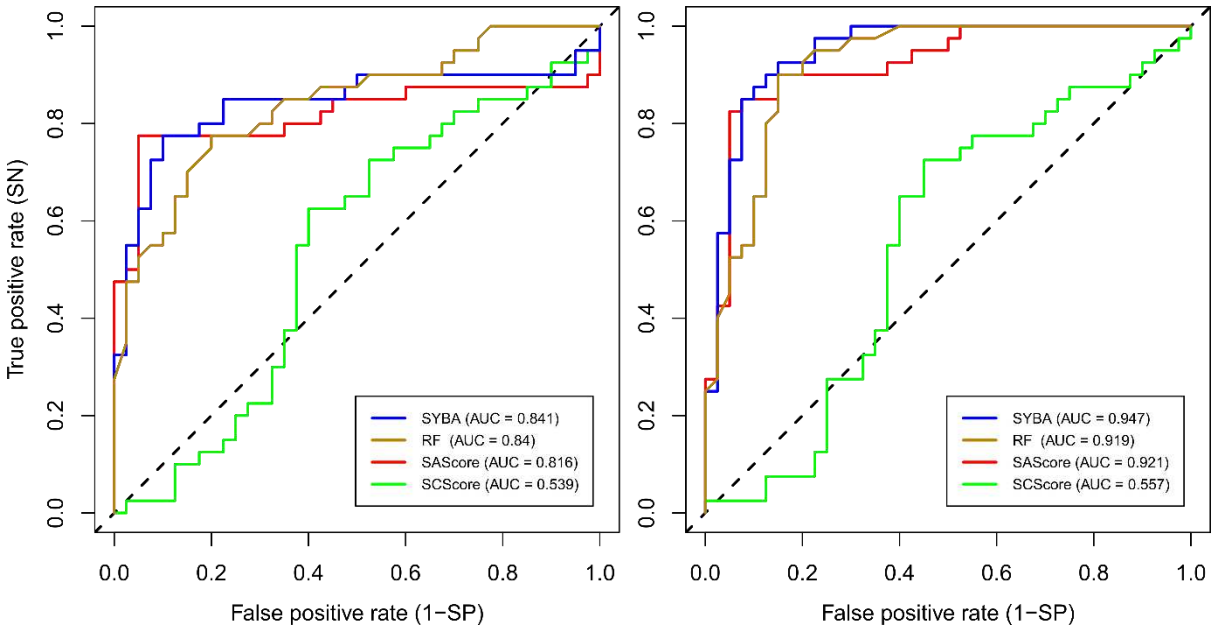

Figure 6 The ROC curves of classification models for the manually curated T_{MC} test set. Out of 30 possible T_{MC} instances, ROC curves of the T_{MC} test set with the smallest (left) and largest (right) SYBA *AUC* are shown.

SYBA is on par with SAScore on computationally picked test set

Stronger evidence is provided by the classification of compounds in the larger computationally picked T_{CP} test set (Figure 1). When used with their default thresholds, both SYBA and RF are more accurate

than SAScore (Table 2, Additional file 2 – Panel S3). Low SAScore accuracy (*Acc* = 0.658) is caused by its low specificity when almost 70 % of HS compounds are predicted to be ES (Table 2). However, SAScore specificity increases by 0.677 when the classification threshold is shifted to its optimal value of 4.5 (Figure 7). At this threshold, SAScore is on par with SYBA and RF methods (Table 2). However, SYBA retains its high performance over much broader range of threshold values than SAScore (Additional file 2 – Figure S1).

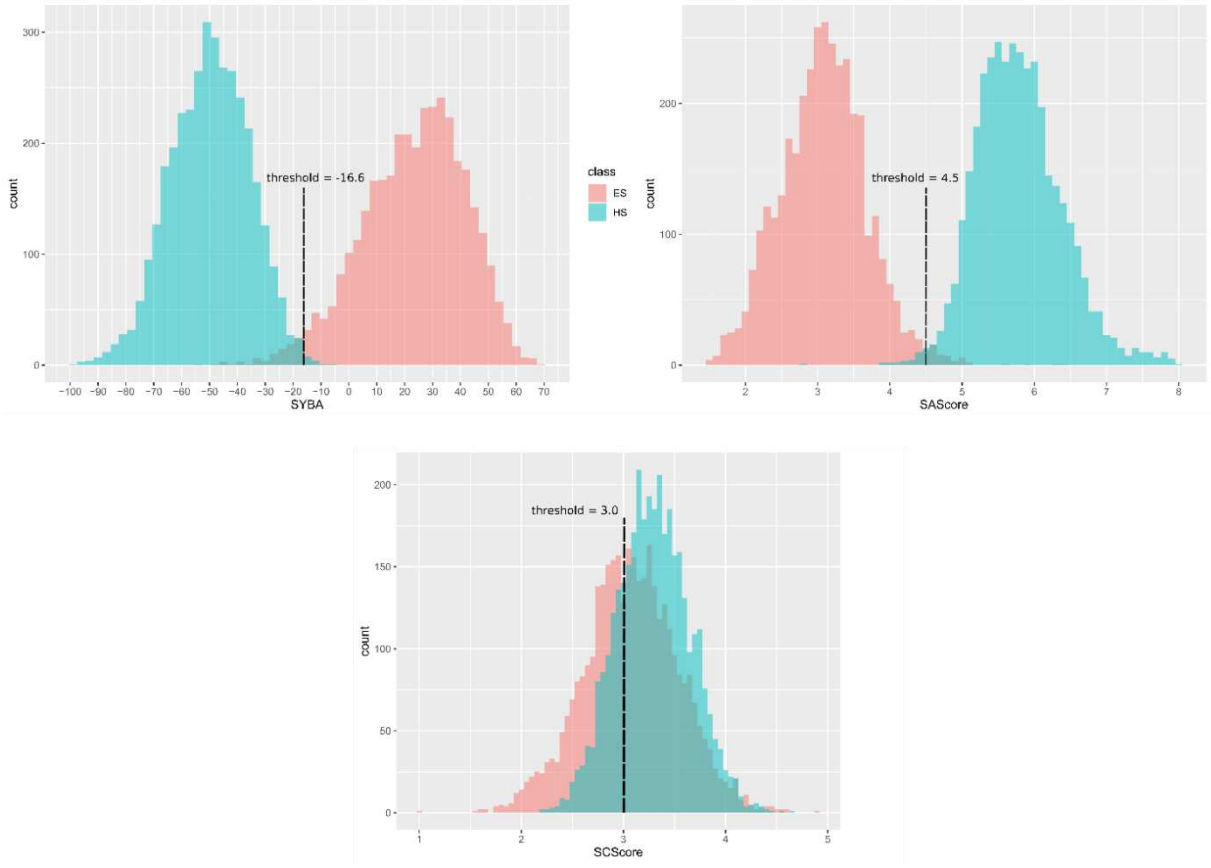

Figure 7 SYBA, SAScore and SCScore histograms of ES and HS compounds in the computationally picked T_{CP} test set. The positions of optimal thresholds are shown. SAScore recommended threshold of 6.0 leads to a large number of FP (Additional file 2 – Panel S3). If the threshold is moved to its optimal value of 4.5, SAScore specificity increases from 0.317 to 0.994, i.e. by 0.677.

High performance of SYBA, RF and SAScore is also evident from their *AUC* that is close to one (Figure 8, Table 2). On the other hand, SCScore fails to distinguish between ES and HS compounds as can be deduced from its ROC curve (Figure 8). In its optimal threshold of 3.0, SCScore predicts a majority of T_{CP} compounds as HS (Figure 7, Additional file 2 – Panel S3).

Model	*AUC*	*Acc*	*SN*	*SP*	Threshold

Default threshold					
SYBA	0.998	**0.949**	0.898	**1.000**	0.0
SAScore	**0.999**	0.658	**0.999**	0.317	6.0
RF	0.996	0.902	0.804	0.999	0.5
Optimized threshold					
SYBA	0.998	0.986	0.977	**0.996**	-16.6
SAScore	**0.999**	**0.989**	**0.985**	0.994	4.5
SCScore	0.590	0.581	0.357	0.805	3.0
RF	0.996	0.975	0.967	0.983	0.2

Table 2 The performance of classification models for the computationally picked T_{CP} test set.

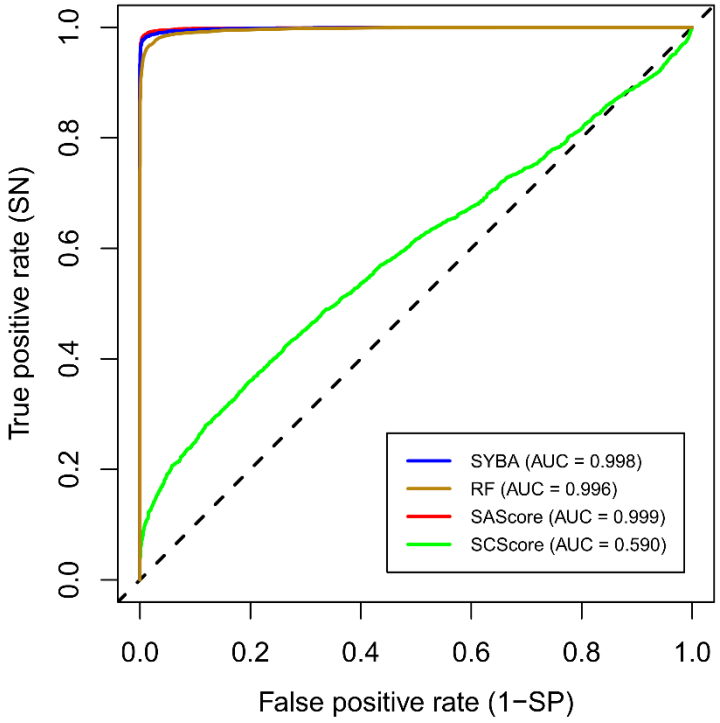

Figure 8 ROC curves of classification models for the T_{CP} test set.

In addition to the T_{MC} and T_{CP} test sets, the performance of SAScore and SCScore was also assessed using the training set S, as this data set was not used for their parametrization. Classification results are shown in Table 3 and Figure 9, confusion matrices are available in Additional file 2 – Panel S4.

Model	*AUC*	*Acc*	*SN*	*SP*	Threshold
Default threshold					
SAScore	0.981	0.767	0.998	0.536	6.0
Optimized threshold					
SAScore	**0.981**	**0.933**	**0.935**	**0.932**	4.4
SCScore	0.667	0.623	0.564	0.682	3.7

Table 3 The performance of classification models for the training set S.

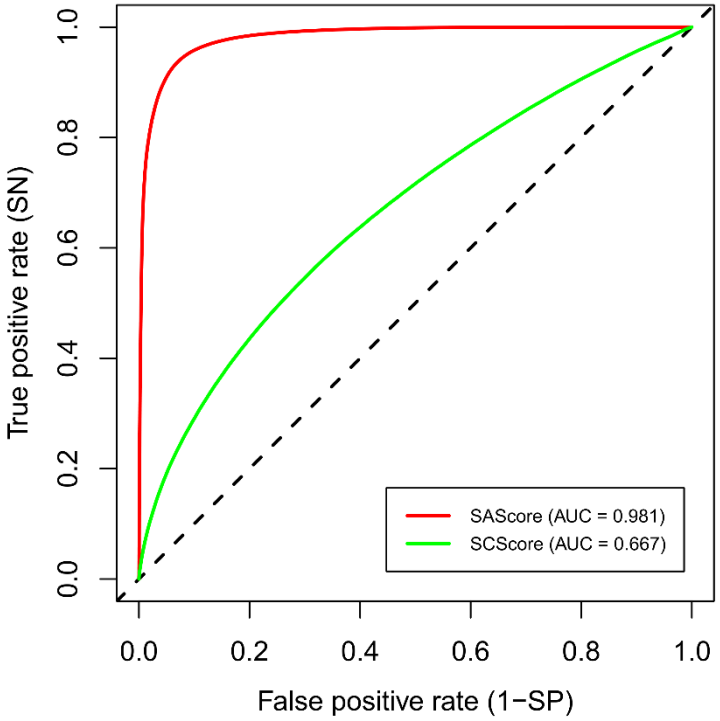

Figure 9 ROC curves of classification models for the training set S.

In agreement with previous experiments on the T_{MC} and T_{CP} test sets, SAScore is able to distinguish between ES and HS compounds more accurately than SCScore. However, to achieve best performance, SAScore classification threshold must be shifted from its default value of 6.0 to the optimal value of 4.4. At this threshold, SAScore is both highly sensitive and specific, while SCScore is, using its optimal threshold of 3.7, sensitive and specific only moderately.

The observed poor performance of SCScore in all data sets may follow from the fact that SCScore differs conceptually from other methods tested in the present work. In SCScore, the problem of

predicting synthetic complexity is reformulated as the analysis of reactions consisting of reactant-product pairs and SCScore correlates with a number of reaction steps. In SCScore derivation, each molecule is analyzed as a whole in the context of all molecules and reactions as they appear in the Reaxys database[44]. Thus, SCScore is likely biased [43] by the types of reactants and products in the Reaxys database. Therefore, we hypothesize that the unsatisfactory results of SCScore are caused by the fact that HS compounds in our data sets are out of its applicability domain.

Conclusions

In the present work, SYBA method for the classification of organic compounds as easy- and hard-to-synthesize is described. SYBA is an additive fragment-based approach meaning that the compound is decomposed into individual substructure fragments, each fragment is assigned its respective SYBA fragment score and these are summed to obtain the final SYBA score. The fragment scores were derived by the Bayesian analysis of the frequency of ECFP4 fragments occurring in the database of ES compounds, that were randomly chosen from the ZINC15 database [58], and HS compounds, that were generated using the Nonpher approach [59]. If the SYBA score is positive, the compound is considered to be ES and vice versa. While SYBA score can theoretically acquire values between plus and minus infinity, a majority of compounds will have SYBA score between -100 and +100 in real applications. It must be stressed that the absolute value of the SYBA score is the measure of the confidence of prediction and not of the degree of synthetic accessibility.

SYBA was compared with other two recent classification methods, SAScore [45] and SCScore [43]. As a baseline for the comparison, RF classifier was used due to its wide adoption in many cheminformatics applications. All methods were assessed using accuracy, sensitivity, specificity and area under the ROC curve. While SYBA and RF provide similar performance, we recommend to use SYBA due to its smaller complexity, lower computational demands and more straightforward analysis of its individual contributions. SYBA outperforms commonly used SAScore, however, if SAScore threshold is moved from its implicit value of 6.0 [45] to the value of 4.5, the results become comparable. Compared to SAScore, SYBA is more robust with respect to the change of the threshold value. The advantage of SYBA score over SAScore is that SYBA is based purely on fragment

contributions and no other ad hoc heuristics, such as complexity or symmetry penalty, are used. Therefore, SYBA fragment scores can be mapped[72] onto a molecule and used for the analysis of the contribution of its individual substructures to the overall synthetic accessibility. SYBA, RF and SAScore substantially outperform SCScore. Weak performance of SCScore can be, in our opinion, attributed to the fact that HS compounds in our test sets lie outside its applicability domain.

SYBA can be used to quickly rank large molecular data sets that originate, for example, in *de novo* molecular design. However, SYBA is conceptually based on the notion that a compound can be categorized as easy- and hard-to-synthesize. As the synthetic accessibility is a vaguely defined term, SYBA's simplifying approach, though accurate enough, cannot compete with more sophisticated synthetic path-reconstruction methods that enable the incorporation of other factors such as the availability of starting materials, reaction yields or a price aspect. At the end, the definitive assessment of synthetic accessibility is in the hands of experienced organic chemists.

In conclusion, we would like to stress that Bayesian framework used in SYBA derivation is a general approach that can be applied to any binary classification problem. Thus, SYBA software can be used beyond its original purpose of predicting synthetic accessibility. SYBA can be re-trained and applied for the classification of compounds by other physico-chemical or biological properties that can be related to any set of descriptors the distribution of which can be obtained from sets of positive and negative examples.

Authors' contributions

MV, MK and DS conceptualized the problem. MV was responsible for method development, implementation and validation. MV also maintains SYBA GitHub repository. MK derived SYBA and IČ took part in method testing. DS supervised the study and prepared the manuscript with the active participation of MV, MK and IČ. All authors read and approved the final manuscript.

Acknowledgements

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported from the Ministry of Education of the Czech Republic (RVO 68378050-KAV-NPUI and LM2018130). IČ research was further supported by a specific university research (MSMT No. 20/2015). MK was supported by the project "Center for Tumor Ecology - Research of the Cancer Microenvironment Supporting Cancer Growth and Spread" (reg. No. CZ.02.1.01/0.0/0.0/16_019/0000785) within the Operational Programme Research, Development and Education.

Availability of data and materials

The code used to train and benchmark SYBA model is available from https://github.com/lich-uct/syba repository. Nonpher is available from https://github.com/lich-uct/nonpher.

Abbreviations

Acc – accuracy

AUC – area under the ROC curve

ES – easy-to-synthesize

HS – hard-to-synthesize

MW – molecular weight

RF – Random forest

ROC – receiver operating characteristic

S – training set

SN – sensitivity

SP – specificity

SYBA – SYnthetic Bayesian Accessibility

T_{CP} – computationally picked test set

T_{MC} – manually curated test set

YI – Youden index

Additional files

Additional file 1 – Training set S. It consists of 693 353 ES compounds selected from the ZINC15 database and of 693 353 ES compounds generated by Nonpher.

Additional file 2 – The supporting document contains threshold values of complexity indices, confusion matrices of the classification of T_{MC}, T_{CP} and S data sets and additional figures.

Additional file 3 – Manually curated test set (T_{MC}). It consists of 40 HS compounds manually selected from scientific papers and of 30 ES sets, each of them contains 40 compounds selected from the ZINC15 database.

Additional file 4 – Computationally picked test set (T_{CP}). It consists of 3 581 HS compounds that were obtained from the GDB-17 database complemented by the same number of compounds randomly selected from the ZINC15 database.

References

1. Bohacek RS, McMartin C, Guida WC (1996) The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev 16(1):3-50.
2. Polishchuk PG, Madzhidov TI, Varnek A (2013) Estimation of the size of drug-like chemical space based on GDB-17 data. J Comput Aided Mol Des 27(8):675-679.
3. Ertl P (2003) Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. J Chem Inf Comput Sci 43(2):374-380.
4. Reymond JL, van Deursen R, Blum LC, Ruddigkeit L (2010) Chemical space as a source for new drugs. Medchemcomm 1(1):30-38.
5. Llanos EJ, Leal W, Luu DH, Jost J, Stadler PF, Restrepo G (2019) Exploration of the chemical space and its three historical regimes. Proc Natl Acad Sci U S A 116(26):12660-12665.
6. Karlov DS, Sosnin S, Tetko IV, Fedorov MV (2019) Chemical space exploration guided by deep neural networks. Rsc Advances 9(9):5151-5157.
7. Gromski PS, Henson AB, Granda JM, Cronin L (2019) How to explore chemical space using algorithms and automation. Nature Reviews Chemistry 3(2):119-128.
8. Walters WP (2019) Virtual Chemical Libraries. J Med Chem 62(3):1116-1124.
9. Franzini RM, Neri D, Scheuermann J (2014) DNA-encoded chemical libraries: advancing beyond conventional small-molecule libraries. Acc Chem Res 47(4):1247-1255.
10. Lopez-Vallejo F, Caulfield T, Martinez-Mayorga K, Giulianotti MA, Nefzi A, Houghten RA, Medina-Franco JL (2011) Integrating virtual screening and combinatorial chemistry for accelerated drug discovery. Comb Chem High Throughput Screen 14(6):475-487.
11. Hoffmann T, Gastreich M (2019) The next level in chemical space navigation: going far beyond enumerable compound libraries. Drug Discov Today 24(5):1148-1156.
12. van Hilten N, Chevillard F, Kolb P (2019) Virtual Compound Libraries in Computer-Assisted Drug Discovery. J Chem Inf Model 59(2):644-651.

13. Schneider G, Fechner U (2005) Computer-based de novo design of drug-like molecules. Nat Rev Drug Discov 4(8):649-663.
14. Loving K, Alberts I, Sherman W (2010) Computational approaches for fragment-based and de novo design. Curr Top Med Chem 10(1):14-32.
15. Medina-Franco JL, Martinez-Mayorga K, Meurice N (2014) Balancing novelty with confined chemical space in modern drug discovery. Expert Opin Drug Discov 9(2):151-165.
16. Schneider P, Schneider G (2016) De Novo Design at the Edge of Chaos. J Med Chem 59(9):4077-4086.
17. Kutchukian PS, Shakhnovich EI (2010) De novo design: balancing novelty and confined chemical space. Expert Opinion on Drug Discovery 5(8):789-812.
18. Hartenfeller M, Schneider G (2011) De novo drug design. Methods Mol Biol 672:299-323.
19. Hartenfeller M, Proschak E, Schuller A, Schneider G (2008) Concept of combinatorial de novo design of drug-like molecules by particle swarm optimization. Chem Biol Drug Des 72(1):16-26.
20. Schneider G, Lee ML, Stahl M, Schneider P (2000) De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. J Comput Aided Mol Des 14(5):487-494.
21. Fechner U, Schneider G (2006) Flux (1): a virtual synthesis scheme for fragment-based de novo design. J Chem Inf Model 46(2):699-707.
22. Vinkers HM, de Jonge MR, Daeyaert FF, Heeres J, Koymans LM, van Lenthe JH, Lewi PJ, Timmerman H, Van Aken K, Janssen PA (2003) SYNOPSIS: SYNthesize and OPtimize System in Silico. J Med Chem 46(13):2765-2773.
23. Hartenfeller M, Zettl H, Walter M, Rupp M, Reisen F, Proschak E, Weggen S, Stark H, Schneider G (2012) DOGS: reaction-driven de novo design of bioactive compounds. PLoS Comput Biol 8(2):e1002380.
24. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. Drug Discov Today 23(6):1241-1250.
25. Hessler G, Baringhaus KH (2018) Artificial Intelligence in Drug Design. Molecules 23(10).
26. Xu Y, Lin K, Wang S, Wang L, Cai C, Song C, Lai L, Pei J (2019) Deep learning for molecular generation. Future Med Chem 11(6):567-597.
27. Popova M, Isayev O, Tropsha A (2018) Deep reinforcement learning for de novo drug design. Sci Adv 4(7):eaap7885.
28. Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. J Cheminform 9(1):48.
29. Blaschke T, Olivecrona M, Engkvist O, Bajorath J, Chen H (2018) Application of Generative Autoencoder in De Novo Molecular Design. Mol Inform 37(1-2).
30. Segler MHS, Kogej T, Tyrchan C, Waller MP (2018) Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Cent Sci 4(1):120-131.
31. Gupta A, Muller AT, Huisman BJH, Fuchs JA, Schneider P, Schneider G (2018) Generative Recurrent Networks for De Novo Drug Design. Mol Inform 37(1-2).
32. Merk D, Friedrich L, Grisoni F, Schneider G (2018) De Novo Design of Bioactive Small Molecules by Artificial Intelligence. Mol Inform 37(1-2).
33. Mendez-Lucio O, Medina-Franco JL (2017) The many roles of molecular complexity in drug discovery. Drug Discov Today 22(1):120-126.
34. Bertz SH (1981) The first general index of molecular complexity. Journal of the American Chemical Society 103(12):3599-3601.
35. Whitlock HW (1998) On the Structure of Total Synthesis of Complex Natural Products. The Journal of Organic Chemistry 63(22):7982-7989.
36. Barone R, Chanon M (2001) A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. Journal of Chemical Information and Computer Sciences 41(2):269-272.
37. Allu TK, Oprea TI (2005) Rapid Evaluation of Synthetic and Molecular Complexity for in Silico Chemistry. Journal of Chemical Information and Modeling 45(5):1237-1243.
38. Selzer P, Roth HJ, Ertl P, Schuffenhauer A (2005) Complex molecules: do they add value? Curr Opin Chem Biol 9(3):310-316.

39. Sheridan RP, Zorn N, Sherer EC, Campeau LC, Chang CZ, Cumming J, Maddess ML, Nantermet PG, Sinz CJ, O'Shea PD (2014) Modeling a crowdsourced definition of molecular complexity. J Chem Inf Model 54(6):1604-1616.
40. Gillet VJ, Myatt G, Zsoldos Z, Johnson AP (1995) SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. Perspectives in Drug Discovery and Design 3:34-50.
41. Huang Q, Li L-L, Yang S-Y (2011) RASA: A Rapid Retrosynthesis-Based Scoring Method for the Assessment of Synthetic Accessibility of Drug-like Molecules. J Chem Inf Model 51(10):2768-2777.
42. Li J, Eastgate MD (2015) Current complexity: a tool for assessing the complexity of organic molecules. Org Biomol Chem 13(26):7164-7176.
43. Coley CW, Rogers L, Green WH, Jensen KF (2018) SCScore: Synthetic Complexity Learned from a Reaction Corpus. Journal of Chemical Information and Modeling 58(2):252-261.
44. Reaxys. https://www.reaxys.com. Accessed 24th January 2020.
45. Ertl P, Schuffenhauer A (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J Cheminf 1:1-11.
46. Rogers D, Hahn M (2010) Extended-Connectivity Fingerprints. J Chem Inf Model 50(5):742-754.
47. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B *et al* (2019) PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 47(D1):D1102-D1109.
48. Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang XP, Norval S, Sassano MF, Shin AI, Webster LA *et al* (2012) Automated design of ligands to polypharmacological profiles. Nature 492(7428):215-220.
49. Yang X, Zhang J, Yoshizoe K, Terayama K, Tsuda K (2017) ChemTS: an efficient python library for de novo molecular generation. Sci Technol Adv Mater 18(1):972-976.
50. Chevillard F, Kolb P (2015) SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability. J Chem Inf Model 55(9):1824-1835.
51. Clark AM, Dole K, Coulon-Spektor A, McNutt A, Grass G, Freundlich JS, Reynolds RC, Ekins S (2015) Open Source Bayesian Models. 1. Application to ADME/Tox and Drug Discovery Datasets. J Chem Inf Model 55(6):1231-1245.
52. Xia X, Maliski EG, Gallant P, Rogers D (2004) Classification of kinase inhibitors using a Bayesian model. J Med Chem 47(18):4463-4470.
53. Bender A (2011) Bayesian methods in virtual screening and chemical biology. Methods Mol Biol 672:175-196.
54. Vogt M, Bajorath J (2007) Introduction of an information-theoretic method to predict recovery rates of active compounds for Bayesian in silico screening: theory and screening trials. J Chem Inf Model 47(2):337-341.
55. Koutsoukas A, Lowe R, Kalantarmotamedi Y, Mussa HY, Klaffke W, Mitchell JB, Glen RC, Bender A (2013) In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naive Bayes and Parzen-Rosenblatt window. J Chem Inf Model 53(8):1957-1966.
56. Klenke A (2008) Probability Theory, A Comprehensive Course. London, Springer, London.
57. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) ZINC: A Free Tool to Discover Chemistry for Biology. J Chem Inf Model 52(7):1757-1768.
58. Sterling T, Irwin JJ (2015) ZINC 15 – Ligand Discovery for Everyone. J Chem Inf Model 55(11):2324-2337.
59. Voršilák M, Svozil D (2017) Nonpher: computational method for design of hard-to-synthesize structures. J Cheminf 9(1):1-20.
60. Hoksza D, Skoda P, Vorsilak M, Svozil D (2014) Molpher: a software framework for systematic chemical space exploration. J Cheminf 6:1-13.
61. RDKit: Open-source cheminformatics. http://www.rdkit.org. Accessed 24th January 2020.

62. Ruddigkeit L, van Deursen R, Blum LC, Reymond J-L (2012) Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. J Chem Inf Model 52(11):2864-2875.
63. Boda K, Seidel T, Gasteiger J (2007) Structure and reaction based evaluation of synthetic accessibility. J Comput-Aided Mol Des 21(6):311-325.
64. Fukunishi Y, Kurosawa T, Mikami Y, Nakamura H (2014) Prediction of Synthetic Accessibility Based on Commercially Available Compound Databases. J Chem Inf Model 54(12):3259-3267.
65. Sheridan RP (2013) Using random forest to model the domain applicability of another random forest model. J Chem Inf Model 53(11):2837-2850.
66. Kensert A, Alvarsson J, Norinder U, Spjuth O (2018) Evaluating parameters for ligand-based modeling with random forest on sparse data sets. J Cheminform 10(1):49.
67. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43(6):1947-1958.
68. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V *et al* (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research (J Mach Learn Res) 12:2825-2830.
69. SCScore GitHub. https://github.com/connorcoley/scscore. Accessed 24th January 2020.
70. Youden WJ (1950) Index for rating diagnostic tests. Cancer 3(1):32-35.
71. Fluss R, Faraggi D, Reiser B (2005) Estimation of the Youden Index and its associated cutoff point. Biom J 47(4):458-472.
72. Riniker S, Landrum GA (2013) Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. J Cheminform 5(1):43.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile3.xlsx
- Additionalfile1.xlsx
- Additionalfile4.xlsx
- Additionalfile2.docx