Prediction of compound synthesis accessibility based on reaction knowledge graph

Baiqing Li[&,$,€], Hongming Chen[€,*]

[&]Guangdong Laboratory Animals Monitoring Institute, Guangzhou, Guangdong Provincial Key Laboratory of Laboratory Animals, 510663, P. R. China

[$]State Key Laboratory of Respiratory Disease, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, P. R. China

[€]Bioland Laboratory (Guangzhou Regenerative Medicine and Health - Guangdong Laboratory)，Guangzhou 510530, P. R. China

[*]Correspondence e-mail: chen_hongming@grmh-gdl.cn

Abstract

With the increasing application of deep learning based generative models for *de novo* molecule design, quantitative estimation of molecular synthetic accessibility becomes a crucial factor for prioritizing the structures generated from generative models. On the other hand, it is also useful for helping prioritization of hit/lead compounds and guiding retro-synthesis analysis. In current study, based on the USPTO and Pistachio reaction datasets, we created a refined chemical reaction network, in which a depth-first search was performed for identification of the reaction paths of product compounds. This reaction dataset was then used to build predictive model for distinguishing the organic compounds either as easy synthesize (ES) or hard-to synthesize (HS) classes. Three synthesis accessibility (SA) models were built using deep learning/machine learning algorithms. The comparison between our three SA scoring functions with other existing synthesis accessibility scoring schemes, such as SYBA, SCScore, SAScore were also carried out. and the graph based deep learning model outperforms those existing SA scores. Our results show that prediction models based on historical reaction knowledge could be a useful tool for measuring molecule complexity and estimating molecule SA.

Introduction

The fact that the drug-like chemical space[1–3] is around 10^{60}- 10^{100} makes the process of finding a compound which satisfies the plethora of criteria such as bioactivity, drug metabolism and pharmacokinetic (DMPK) profile, synthetic accessibility simultaneously, as difficult as finding a needle in a hay stack[4–6]. Hence both medicinal and computational chemists attempted to develop approaches to efficiently explore chemical space for the purpose of identifying the compounds with desirable pharmacological activities as well as ADMET properties[7–11]. Among these efforts, virtual library based de novo molecule design method represents an important computational paradigm.[12–15] The application of deep generative modelling for *de novo* molecule design has emerged in recent years. One major benefit of generative model method is that it can exhaustively explore a much larger chemical space comparing with virtual library based method. However, one of the big hurdle for structure generation in generative model is how to control the structural complexity of generated compounds. Ideally, the compounds designed by those generative models should be synthesized within in relative few synthesis steps.

The definition of molecular complexity is context-dependent and ambiguous, for example multifunctional, multi-ring containing or multi-chiral-centre compounds can be complex to synthesize. Herein, ‘synthesis accessibility’ (SA) was considered as the synonymous definition of complexity. In recent years, various metrics of SA was extensively used in virtual screening (VS) workflow.[16-20] For example, various metrics developed based on notions such as: assessing whether broken bonds between certain atomic types[16] are reasonable; whether individual building blocks[17,18] could be connected from the experiences of existing chemical reactions etc. Moreover, retro-synthetic rules were also integrated into VS process[19,20].

Simple SA assessment can be done by simply calculating some physicochemical properties such as number of atom, bond, ring and some unconventional hard-to-synthesis motifs, such as stereo-centre and macrocycles etc. SAScore[21], one of the first SA computation methods, was developed based on the frequency analysis of molecular ECFP4[22] fragment occurrence in PubChem database. It was proved to be a useful tool in many cheminformatics applications[23–26]. The rationale of the method is that correlating the SA of a molecule to the fragment occurring frequency. Each fragment is assigned a numerical SA score. The higher the fragment occurring frequency in PubChem database[27] is, the greater its SA score is. In addition, SAScore also took into account some complex structure motifs as penalty, such as stereo centers, spiro-rings and bridge ring atoms. But generally, the structural complexity and SA are not strongly correlated, partially because it does not incorporate the availability of the starting materials. For example, the total synthesis of a steroid is a tedious and challenging task, but if starting from readily prepared intermediates like cholesterol, the synthesis might only require very few reaction steps[28]. Therefore the molecular complexity based metrics could underestimate the SA of the molecules which can be easily synthesized from already existing precursors[29,30]. Hence, more general SA metric is needed.

A more realistic solution for estimating SA is to take the reaction route complexity[31] into account, which means that the more reaction steps is needed for synthesizing the compound, the lower the synthesis accessibility of the compound is. Although there are some domain-specific knowledge on what a good or reasonable synthetic route should be, in general, the synthesis complexity of compound becomes more large when more synthesis steps are needed[32,33]. Aligning with this principle, SCScore[28] was recently developed based on one simple premise: on average, reaction products are synthetically more complex than their corresponding reactants. By building 22 million reactant-product pairs from the commercial Reaxys database[34], a deep feed-forward neural network was trained to assign the synthetic accessibility score between 1 and 5. The main idea of the SCScore was to learn a synthetic complexity score which correlates with the number of reaction steps. But merely taking isolated reactant-product pairs into consideration and lack of consideration on the relationship among compounds cross different pairs will probably make the method not general enough for characterizing the SA. SYBA[35], another recent method for synthetic accessibility assessment, is a fragment-based method for classifying organic compounds as ES or HS. To quantify this, a Bernoulli naïve Bayes classifier was trained on ES molecules available from ZINC15 and corresponding HS molecule generated by Nonpher[36] algorithm. However, SYBA has the same problems as SCScore in terms of construction of dataset as one-to-one (ES and HS) pairs, both methods are lack of systematic comparison in a large chemical reaction database. It is worth mentioning that recently appeared RAscore[37], a feed forward neural network classifier, based on the AI driven computer aided synthesis planning (CASP) tool -- AiZynthFinder[38], could determine whether a synthetic route can be found for a particular compound, and how difficult (a retrosynthetic accessibility value) it may be to realize the route in the wet lab. In terms of dataset selection, it has little to do with the chemical reaction. ES dataset was generated by randomly sampling 200,000 compounds from ChEMBL, and HS dataset was collected from GDB17 database[39].

Gryzbowski et al[40] reported their work in constructing reaction knowledge graph aka NOC (network of chemistry), where large amount of compounds are inter-connected as either reactants or products. In current study, a similar knowledge graph based on reaction dataset of USPTO[41] and Pistachio[42] was constructed and, according to direction of the connected edges, nodes in the network were classified into two types of node: the node serving only as reactant (i.e. starting materials) and normal node which can serve either as reactant or product. The edge distance between the normal node and the reactant only node in the graph is used to represent possible reaction steps for synthesizing a compound. According to the number of synthesis step, an organic compound can be labeled as either ES or HS.

Instead of constructing data set in the pairwise manner like SCScore and SYBA, a dataset based on existing reaction evidence in the reaction knowledge graph can be curated. Various classification models based on machine learning, which includes graph neural network as well as the fully connected feed-forward neural network, were built to predict compound's SA. A comparison between our deep learning models and existing SA scoring functions like SYBA, SCScore and SAScore was also carried out. Our results show that the graph neural network model outperforms those available SA scoring functions.

Methods

Reaction dataset

The reaction dataset used in the present work contains the publicly available United States Patent Office extracts (USPTO[41]) ranging from 1976 to 2016, and the commercially available Pistachio[42] database provided by NextMove. After removing duplicates, all reactions were atom-mapped and classified using Filbert and HasELNut program provided by Nextmove, and 12,985,183 reaction items were remained. We then cleaned the data set using predefined filtering criteria described by Takkar et al[43] and removing some undesirable reactions (such as incomplete reactions, reactions which can't generate template), 9,041,882 valid reaction items were obtained. This data set was further processed with the workflow as shown in Figure 1 to prepare datasets for template extraction, knowledge graph generation and building of various predictive models.

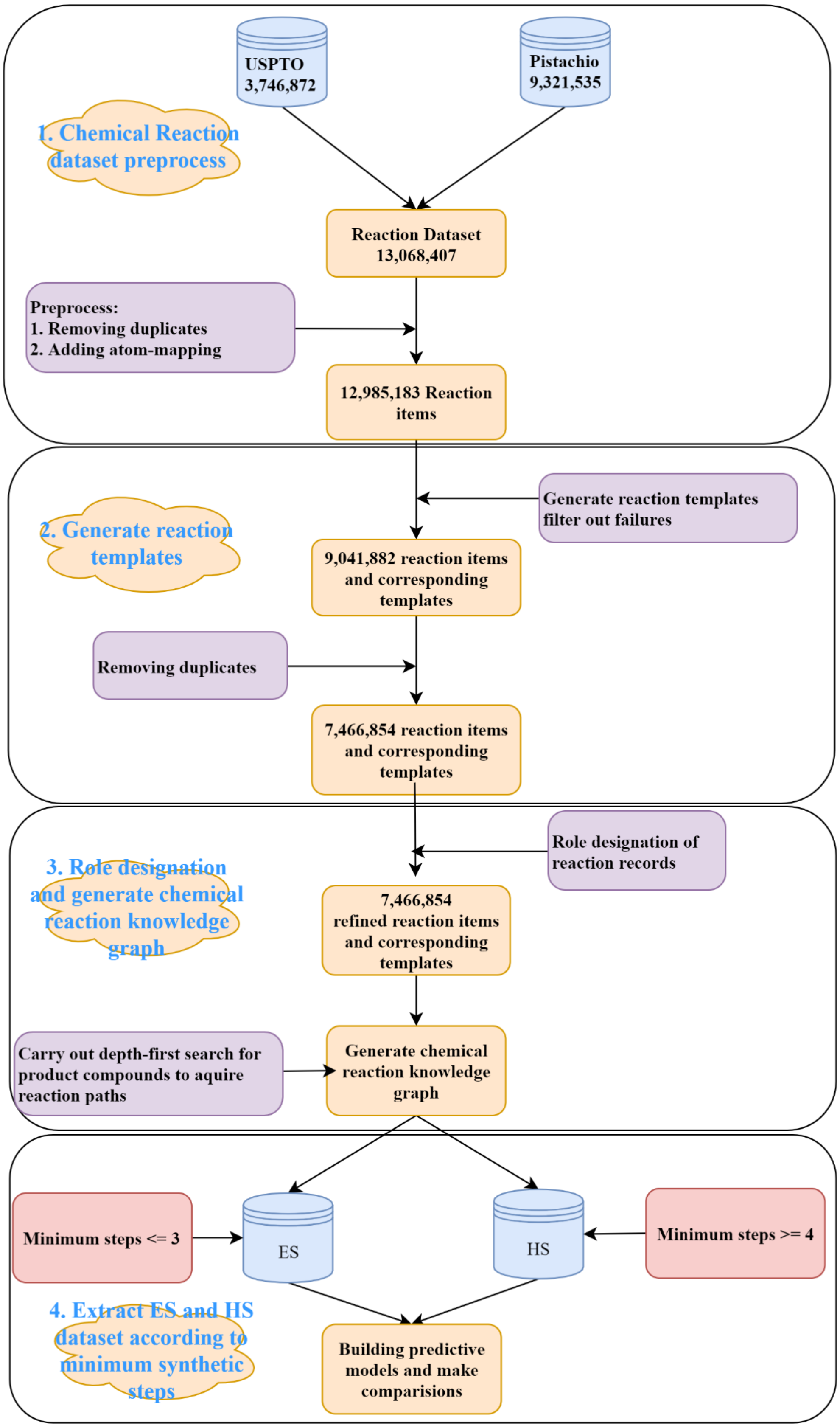

USPTO
3,746,872
Pistachio
9,321,535
1. Chemical Reaction dataset preprocess
Reaction Dataset
13,068,407
Preprocess:
1. Removing duplicates
2. Adding atom-mapping
12,985,183 Reaction items
Generate reaction templates filter out failures
2. Generate reaction templates
9,041,882 reaction items and corresponding templates
Removing duplicates
7,466,854 reaction items and corresponding templates
Role designation of reaction records
3. Role designation and generate chemical reaction knowledge graph
7,466,854 refined reaction items and corresponding templates
Carry out depth-first search for product compounds to aquire reaction paths
Generate chemical reaction knowledge graph
Minimum steps <= 3
ES
HS
Minimum steps >= 4
4. Extract ES and HS dataset according to minimum synthetic steps
Building predictive models and make comparisions

Figure. 1 The workflow of chemical reaction processing. It was divided into four steps, (1) dataset preparation; (2) generation of reaction templates; (3) role designation and generation of chemical reaction knowledge graph; (4) curation of ES and HS data set.

Template extraction and role designation

Coley et al[44,45] recently reported the development of a toolkit, RDChiral, for reaction template extraction. This toolkit not only can recognize the immediate neighborhood of reaction centers (red atoms in Figure 2), but also the required extended environment including special functional groups (green atoms in Figure 2) and neighboring carbon atoms as the extended motif (grey atoms in Figure 2) based on the user-defined radius (default value radius=1 used here). It is worth mentioning that Thakkar *et al*[43] added 70 additional functional groups and protecting groups besides the original 75 special groups that were included in RDChiral. The above mentioned valid reaction set was subsequently processed using RDKit and RDChiral[45] for template extraction. In total, 7,466,854 reaction templates were produced.

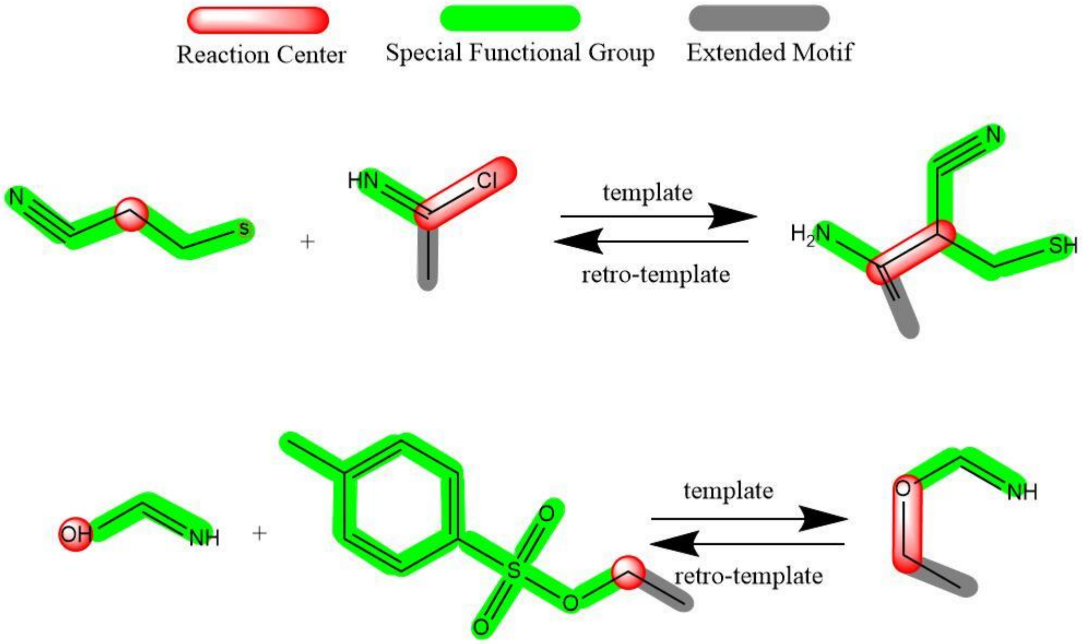


Figure 2. Examples of template extraction, in which colored sites are added to reaction template. The red atom refers to the reaction center where the reaction occurs; The green atom refers to the special functional group around the reaction center; The grey atom corresponds to the extended motif which is the carbon atoms within

certain distance to the reaction center.

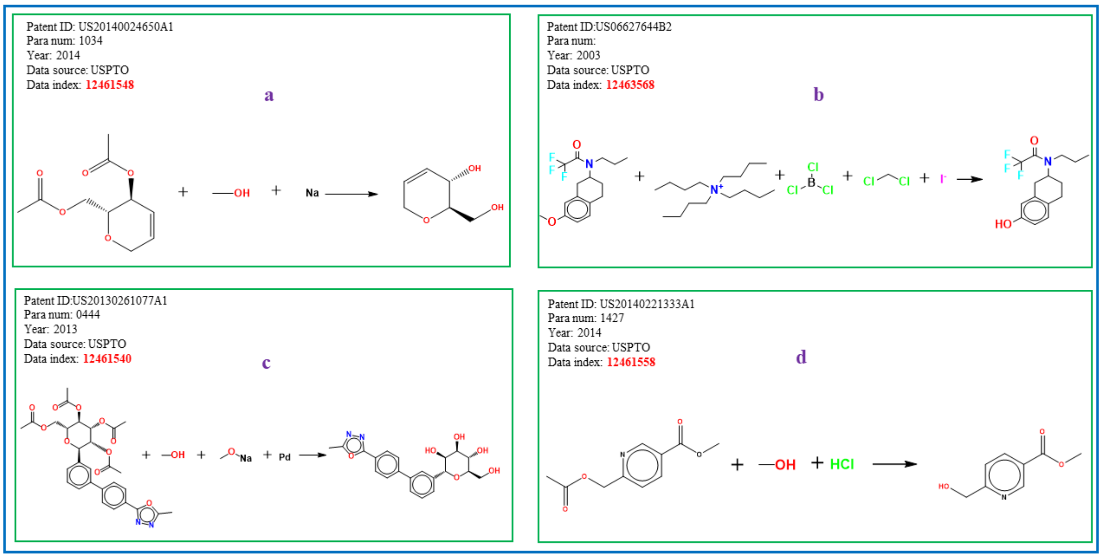


Figure 3. Four examples of original flawed chemical reaction records.

It was found that, in the obtained reaction records going through the previous procedure, the reagent part in reaction SMILES very often was misplaced into the reactant part. For example, in Figure 4a, it was obviously wrong to put the Sodium (Na) into reactant list and methanol should be solvent; In Figure 4b, Iodide ion should not be in this reaction, methylene dichloride should be solvent. The exact role of Boron trichloride is unknown, but it's clear that boron trichloride and the ammonium ion does not contribute mass to the product and shouldn't belong to reactant; In Figure 4c, Pd should be catalyst, sodium methanolate should serve as base, and methyl alcohol should be solvent; In Figure 4d, methyl alcohol should be solvent, the HCl provides acidic environment and should be assigned to reagent. In those problematic reaction records, the reagents/solvents in the reaction SMILES were mixed with the reactants and this could make troubles in identifying the correct reactant structures for later creating the reaction network. This misplacement though did not affect the template extraction process, as this operation treated all non-product parts equally. Extracted reaction templates above were used to identify the reaction roles (mainly the reactants and reagents). A Knime[46] workflow was developed to clean up the reactant list of each reaction and following steps were carried out to identify reagent components:

(1) Before extracting templates, along with reactants, all reagent components (solvent and catalyst) of original reaction records were treated as reactants and formed original reactants (ORs)

(2) Template extraction procedure was carried out based on the reorganized reaction SMILES, both normal and inverse reaction templates were generated. The product structures were then put into the inverse reaction template to generate the predicted reactants (PRs) by using RDChiral functions.

(3) For each reaction record, the generated PRs were then compared with ORs. For each PR component, any OR component which matched exactly or had highest pairwise Tanimoto similarity with it was regarded as the true reactant corresponding to the PR. After all PRs were compared, the remaining ORs were then designated as reagents. In general, the structure similarity between reactants and reagents is quite low, reagents can be easily distinguished and removed from the reactant list.

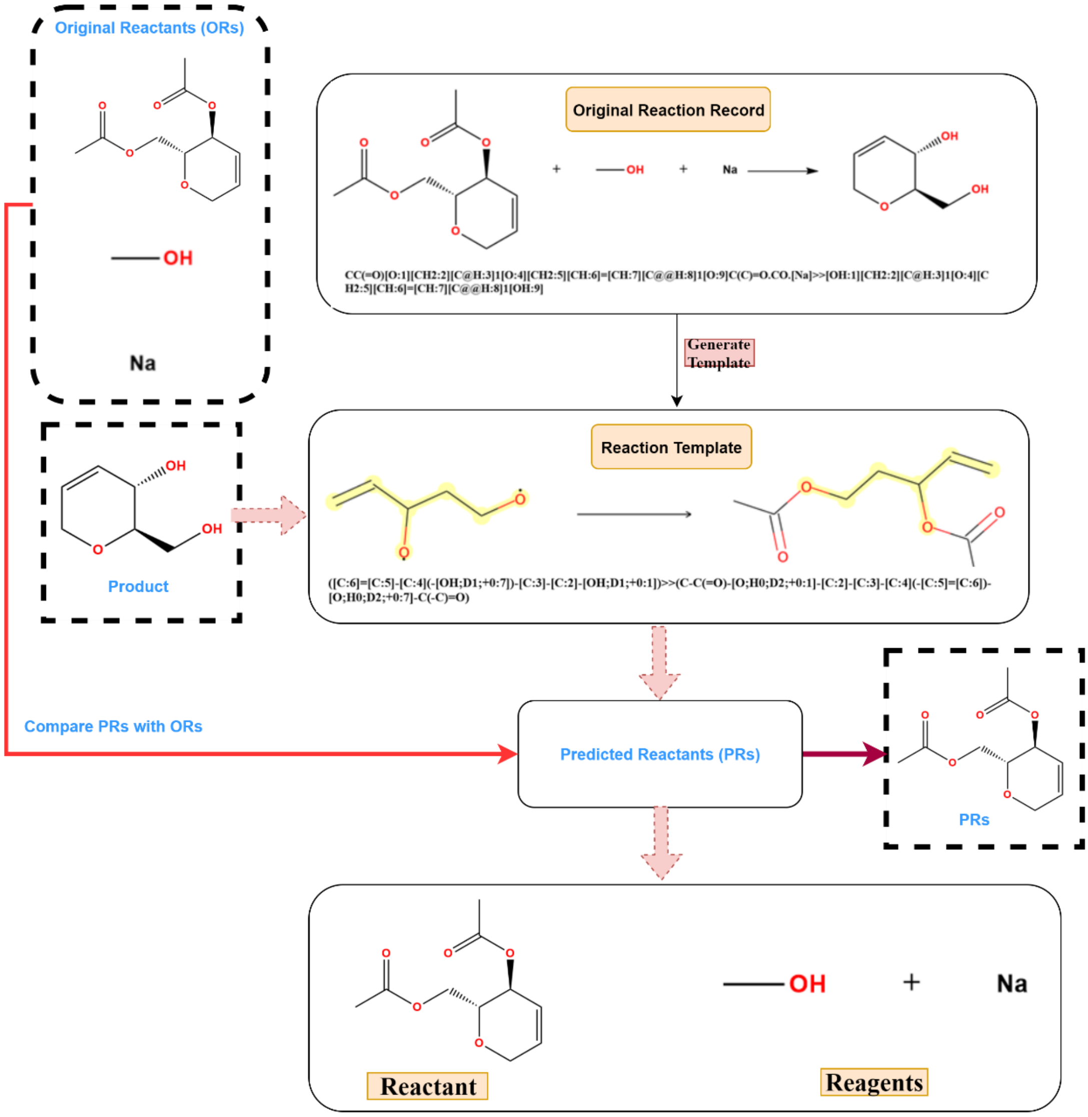

Original Reactants (ORs)
OH
Na
Original Reaction Record
OH
Na
CC(=O)[O:1][CH2:2][C@H:3]1[O:4][CH2:5][CH:6]=[CH:7][C@@H:8]1[O:9]C(C)=O.CO.[Na]>>[OH:1][CH2:2][C@H:3]1[O:4][CH2:5][CH:6]=[CH:7][C@@H:8]1[OH:9]
Generate Template
Product
Reaction Template
([C:6]=[C:5]-[C:4](-[OH;D1;+0:7])-[C:3]-[C:2]-[OH;D1;+0:1])>>(C-C(=O)-[O;H0;D2;+0:1]-[C:2]-[C:3]-[C:4](-[C:5]=[C:6])-[O;H0;D2;+0:7]-C(-C)=O)
Compare PRs with ORs
Predicted Reactants (PRs)
PRs
OH
Na
Reactant
Reagents

Figure 4. The process of role designation. (1) Rearrange ORs; (2) Generate inverse reaction templates; (3) Match original product with inverse templates and get PRs; (4) Compare PRs with ORs

Reaction knowledge graph

Gryzbowski et al[40] constructed directed complex networks using known organic chemical reactions, in which the nodes refer to the chemical substances (either reactants or products) and the directed edges correspond to the chemical reactions where the substances involve in. In current study, a similar network was constructed based on the combined Pistachio and USPTO datasets. After going through the above mentioned cleaning process, substances that don’t directly contribute to the reactions such as solvents, catalysts were removed and finally a refined chemical reaction network containing 2,192,740 nodes was constructed (as shown in Figure 5) and its details are listed in Table 1. The network file (GraphML format) generated from USPTO data set after clearing can be found in https://github.com/jidushanbojue/YaSAScore/data.

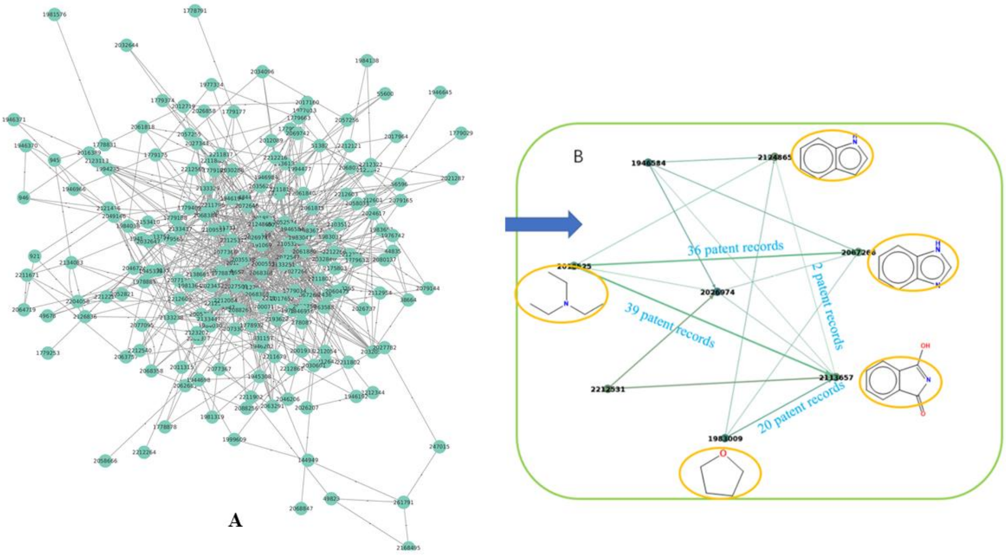


Figure 5. Some example nodes of the reaction network. (A) Part of the reaction network. (B)Details of some exemplified nodes. The width of the line represents the number of chemical patents involved in the reaction.

As shown in Figure 5, the starting nodes of the directed edge are reactants of the reaction and the destination nodes are the products. There are two types of node in the reaction knowledge graph, one type is the node which doesn't connect to any in-flux edge and only connect to out-flux edges, it is called terminal node which only serves as reactant and there are 488,220 terminal nodes existed in the graph. The terminal nodes may be limited by the data set which was used to build the knowledge graph and are not necessary starting materials. Here we used a set of commercial available building block molecules[38] from ZINC database[47] to identify the starting materials among the terminal nodes and in total 38,664 terminal nodes were identified as starting material. The other type is the normal node which connects to both in-flux and out-flux edges and there are 691,830 nodes in total. The normal node can be recognized as either the product of the starting materials, or the reactant for other products. The path (if exists) length between a starting material (terminal node) and a product (normal node) on the graph can be referred as possible reaction steps (RS) for synthesizing a compound. From a practical point of view, the route with minimum step to the product can be regarded as the best reaction route for synthesizing the product. Some example structures are shown in Figure 6.

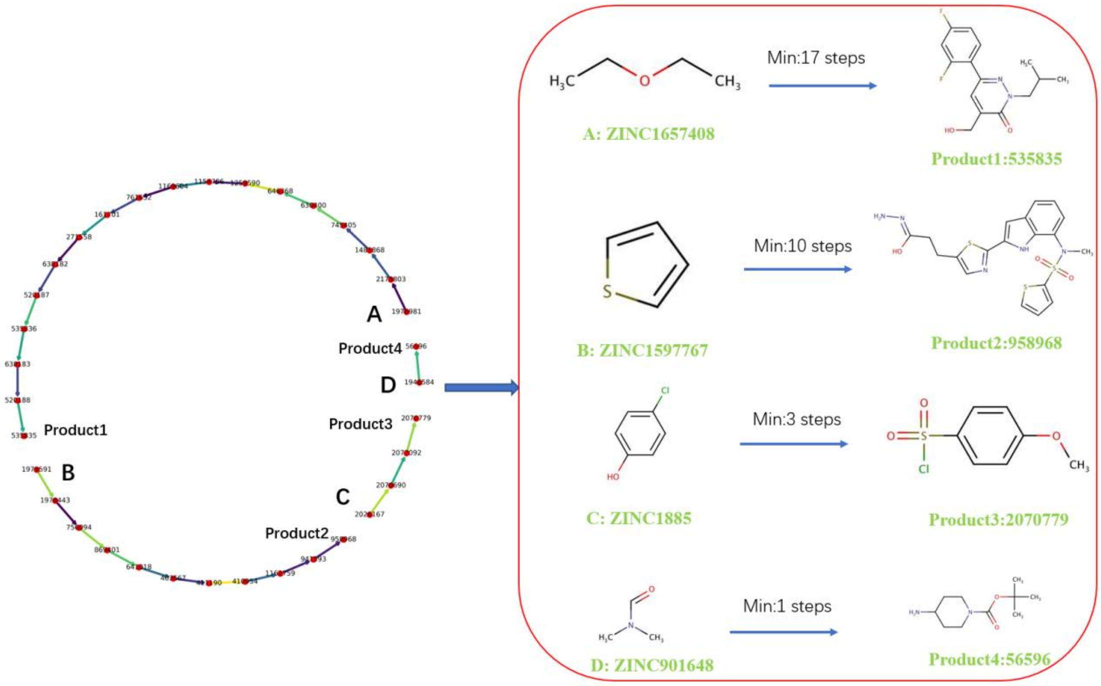


Figure 6. The reaction steps of four example structures in the graph. A, B, C, and D are all ZINC starting materials. It takes at least 17 steps from A to Product 1, at least 10 steps from B to Product 2, at least 3 steps from C to Product 3 and 1 step from D to Product 4.

Table 1. Statistics in the reaction knowledge graph

Products	1,368,588
Terminal node	488,220
Normal node	691,830
Starting material	38,664
Reactant + Product (network node)	2,192,740
Agent (not network node)	20,286
Reactant + Product + Agent	2,213,026

Synthesis accessibility prediction model

The main goal of current study is to use the reaction step (RS) data which was obtained from the reaction knowledge graph as the surrogate of SA to build classification models of compound SA. Given that a compound in the graph has multiple paths to reach the starting materials, the minimum step was chosen to represent SA. The distribution of the minimum RS can be seen in Figure 7, it seems that distribution of RS is quite uneven, most of the compounds' RS is less than 3 steps. Accordingly, we considered compounds whose minimum RS is less than or equal to 3 as ES class (753,842 compounds in total) and compounds whose RS is great than or equal to 4 as HS class (122,248 compounds in total). In order to create a balanced data set for model training, a structural clustering analysis was done on the ES class compounds to select a diverse ES compound set which also had roughly the same number of compounds to the HS class. The oetoolkit[48] based program Flush[49] was used for clustering and the Tanimoto similarity threshold that was calculated based on Foyfi fingerprints[50] was set as 0.62. The seed of each cluster was selected and a data set containing 123,837 compounds was curated, in which the ratio of ES and HS compound is roughly at 1:1. This balanced data set was then split into training, validation and test set with the ratio of 8:1:1. In addition, a full test set was also composed by adding those remained ES class compounds into the balanced test set for evaluating the model performance on all compounds not included in the training set.

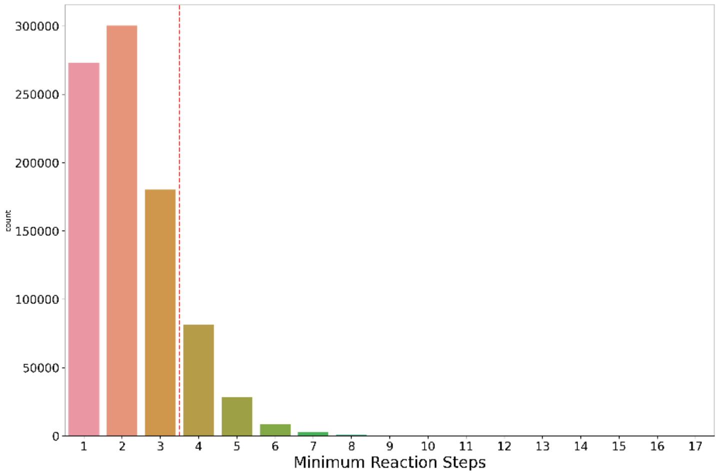


Figure 7. The distribution of the minimum RS. Compounds whose minimum RS is less than or equal to 3 as ES class (753,842 compounds in total) and compounds whose RS is great than or equal to 4 as HS class (122,248 compounds in total)

The fully connected deep neural network (DNN) models using molecular fingerprint as input were built, in which the 2048-length bit string Extended-Connectivity Fingerprints (ECFPs) by setting the maximum searching depth to 2 was generated using RDkit (https://www.rdkit.org/). Graph convolution neural network based model using molecular 2D structure directly as input were also built. Here the modified message passing neural network, CMPNN[51], was employed for constructing the graph neural network. The DNN classifier were trained using Keras with Tensorflow as the back end, the RMSprop optimizer was used and binary cross-entropy was chosen as the loss function. The learning rate was decayed on plateau by a factor of 0.5. The optimal combination of parameters for the model was searched based on the model performance on the validation set. The CMPNN classifier was trained using default parameters as in the original literature[51]. Additionally, the performance of several existing SA models like SYBA, SCScore and SAScore were also examined on our data set for comparison purpose. The hyper-parameters of those models can be found in Supporting Information Table S1.

Performance evaluation

The performance of the SA prediction models was evaluated by three different indicators: the classification accuracy (ACC), Matthews correlation coefficient (MCC) and area under the ROC curve (AUC).

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$MCC = \frac{TP*TN-FN*FP}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \quad (2)$$

Where true positive (TP) refers to the true ES, and true negative (TN) refers to the true HS. ACC represents the percentage of correctly classified samples regardless of their predicted classes. ROC-AUC and MCC[52] are two important metrics which make trade-off between TP rate and FP rate over the whole possible thresholds. MCC, ACC and ROC-AUC are commonly used metrics for measuring the performance of binary classification model.

Results and Discussion

Analysis of chemical space of the data set

In order to examine the chemical space of the ES and HS class compounds, a PCA (principal component analysis) analysis was carried out on the whole ES and HS compound sets based on six physicochemical descriptors which was calculated using RDkit package, i.e. MW (molecular weight), TPSA (topological polar surface area), RTB (number of rotatable bonds), HBD (number of H bond donors), HBA (number of H bond acceptors). The dimensionality of the input space was reduced by PCA to the top 2 components that explained 80% of the variance in the data. Figure 8 showed that the chemical spaces of ES and HS compound are basically identical. The distribution of individual properties can be seen in Figure 9 and they are also identical. These results demonstrate that using these physicochemical properties either alone or collectively is difficult to separate ES and HS molecules and a dedicated model is needed to predict SA.

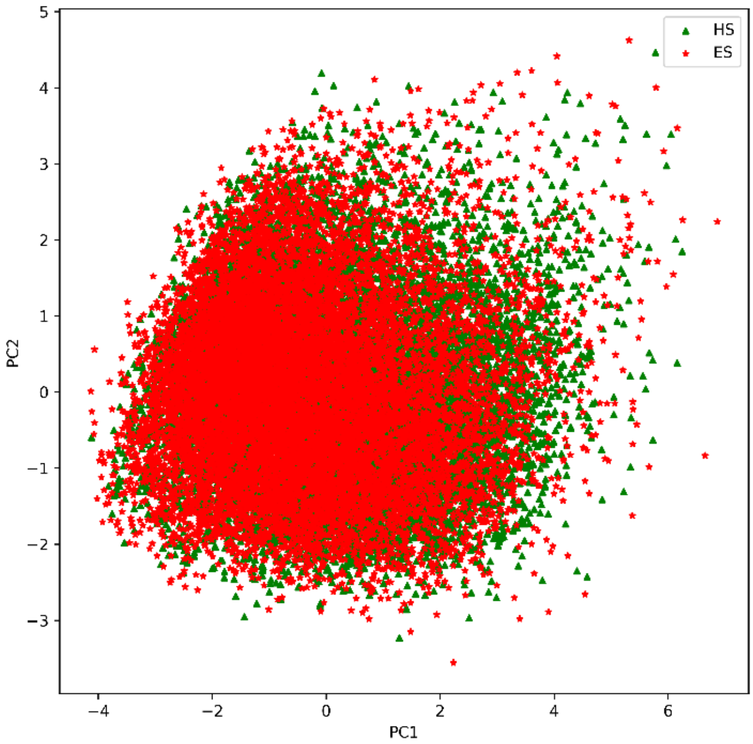


Figure. 8 PCA analysis on the physicochemical descriptors of the ES and HS datasets.

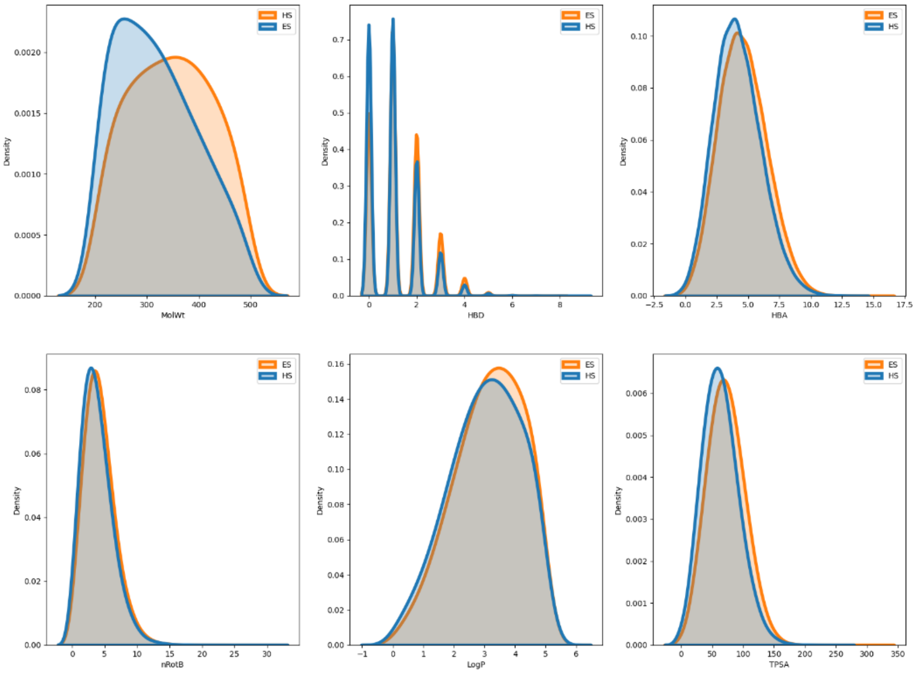


Figure. 9 The distribution of physicochemical descriptors on the ES and HS datasets.

Evaluation of model performance

Table 2 the performance of different models on the balanced test set

MODEL	ROC-AUC	ACC	MCC
CMPNN	0.791	0.715	0.434
DNN-ECFP	0.749	0.685	0.371
SYBA	0.465	0.497	-0.012
SYBA-2[1]	0.76	0.69	0.382
SAScore	0.513	0.498	-0.011
SCScore	0.621	0.582	0.167

SYBA-2[1]: retrained the SYBA model on our own dataset

Various predictive models were built on the data set. For CMPNN and DNN model, the optimal parameters were determined based the performance on the validation set. For DNN models, the ECFP fingerprint was used as the input descriptor. Table 2 shows that, among all models, CMPNN model achieved best results. Those existing SA models (SYBA, SAScore and SCScore) performed worse than both DNN-ECFP and CMPNN models. It is worth noting that SYBA-2 is a model which was retrained on our own training set using SYBA algorithm and the performance of SYBA-2 is still worse than that of CMPNN model.

Table 3 the performance of different models on AllTestSet

MODEL	ROC-AUC	ACC	MCC
CMPNN	0.741	0.582	0.096
DNN-ECFP	0.734	0.646	0.098
SYBA	0.569	0.973	0.017
SYBA-2	0.694	0.651	0.079
SAScore	0.569	0.972	0.009
SCScore	0.584	0.459	0.035

Because the number of ES (753,842) compound is much greater than the number of HS (122,248), after clustering analysis, around 630K ES molecules were left out. To gain a full picture of the model performance, those remained ES compounds was also evaluated. In this case, all the remaining compounds from the clustering analysis were added into the test set and the predictions were done on all compounds in the test set. The full test set results for each model were shown in Table 3. In this much larger and unbalanced data set, CMPNN ROC-AUC value was almost the same to that of the previous balanced test set and it was still the best model, while DNN-ECFP model ranked as the second best model. The retrained SYBA-2 model performed worse when the predictions on the remaining ES compounds were taken into account. It is worth mentioning that the SYBA and SAScore model have extremely high ACC score, but their performance on ROC-AUC is very poor, which suggests these models always tend to classify compounds as ES class, and have difficulty in identifying HS molecules.

These results suggest that the structural complexity of molecule doesn't correlate well the actual reaction step data. Models built with actual reaction step data may better reflect the SA. Conceptually, SAScore and SYBA methods are different to the method used in the present work. The SCScore model was built on the reaction data but it only consider the relationship between reactant-product pairs, while our model is built on the true reaction step data. Overall, our model at some extent should reflect the true synthesis accessibility instead of only structural complexity.

Conclusion

In the present work, we have developed predictive models for quantifying synthesis accessibility based directly on a refined chemical reaction network constructed on the USPTO and Pistachio reaction datasets. In contrast to existing SA methods which was built based on compound complexity, we used the minimum synthesis step of a product compound, which was obtained by carrying out the depth-first search of a chemical reaction network, as the surrogate of the synthesis accessibility. Compounds was designated as either ES or HS classes depending on their minimum synthesis step and three SA prediction models were built using deep learning/machine learning algorithms.

We compared these SA scoring functions with existing SA scoring schemes, such as SYBA, SCScore, SAScore. The graph convolution neural network model outperforms those existing SA scores. Our analysis of reaction knowledge graph is still at the early stage. We expect building SA prediction models based on historical reaction data could be an interesting future direction for quantitively assessing molecular SA. With more reaction data and bigger reaction network, SA prediction model could be further improved.

Data and Software Available

Pistachio dataset was commercial chemical reaction database and used with permissions. Filbert, NameRxn and HazelNut were used for atom-mapping and classification under license from NextMove software. The detailed hyperparameters of all models can be found in supplementary materials. The scripts of templates extraction, generation of chemical reaction network, and the training, predicting and analysis process of each model can be found in the GitHub repository https://github.com/jidushanbojue/YaSAScore. The KNIME workflow of chemical reaction role designation, and the refined network file (GraphML format) of open-source USPTO could also be found in the GitHub repository, https://github.com/jidushanbojue/YaSAScore.

Author Information

Corresponding Author

Hongming Chen, E-mail: chen_hongming@grmh-gdl.cn

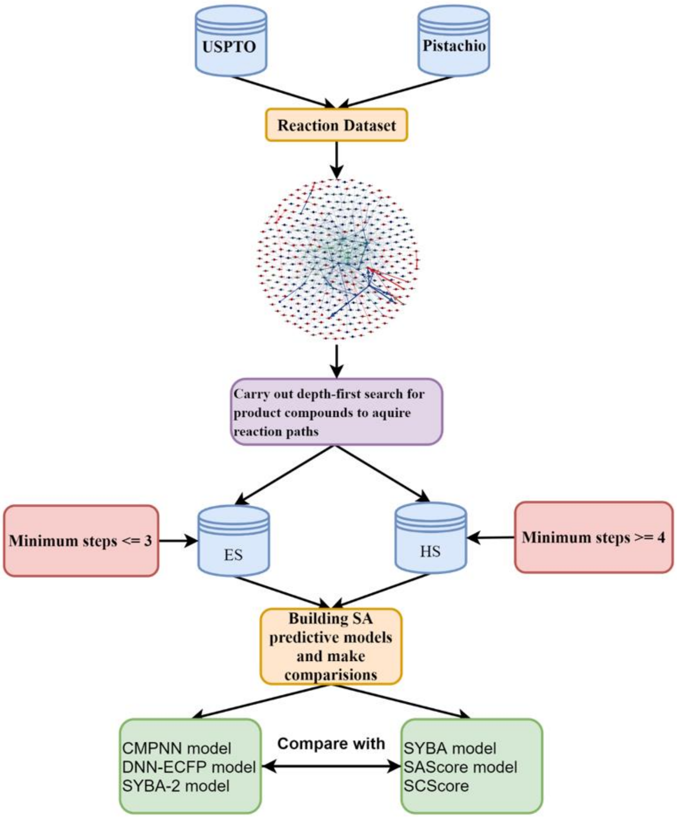
Competing interests

The authors declare that they have no competing interests.

Abbreviations

PRs, Predicted Reactants; ORs, Original Reactants; ES, Easy to Synthesize; HS, Hard to Synthesize; NOC, network of chemistry; RS, Reaction Steps.

TOC figure



TOC figure. We developed predictive models for quantifying synthesis accessibility based directly on a refined chemical reaction network constructed on the USPTO and Pistachio reaction datasets. And the minimum synthesis step of a product compound, which was obtained by carrying out the depth-first search of the chemical reaction network, was used as the surrogate of the synthesis accessibility. Compounds was designated as either ES or HS classes depending on their minimum synthesis step. Then three SA prediction models (CMPNN, DNN-ECFP, and SYBA-2) were built using deep learning/machine learning algorithms and compared with SYBA, SAScore, SAScore.

Reference

(1) Reymond, J.-L.; van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical Space as a Source for New Drugs. *Medchemcomm* **2010**, *1* (1), 30. https://doi.org/10.1039/c0md00020e.

(2) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput. Aided. Mol. Des.* **2013**, *27* (8), 675–679. https://doi.org/10.1007/s10822-013-9672-4.

(3) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16* (1), 3–50. https://doi.org/10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6.

(4) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; John Hart, A.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science (80-.).* **2019**, *365* (6453). https://doi.org/10.1126/science.aax1566.

(5) Green, C. P.; Engkvist, O.; Pairaudeau, G. The Convergence of Artificial Intelligence and Chemistry for Improved Drug Discovery. *Future Med. Chem.* **2018**, *10* (22), 2573–2576. https://doi.org/10.4155/fmc-2018-0161.

(6) Plowright, A. T.; Johnstone, C.; Kihlberg, J.; Pettersson, J.; Robb, G.; Thompson, R. A. Hypothesis Driven Drug Design: Improving Quality and Effectiveness of the Design-Make-Test-Analyse Cycle. *Drug Discov. Today* **2012**, *17* (1–2), 56–62. https://doi.org/10.1016/j.drudis.2011.09.012.

(7) Llanos, E. J.; Leal, W.; Luu, D. H.; Jost, J.; Stadler, P. F.; Restrepo, G. Correction for Llanos et Al., Exploration of the Chemical Space and Its Three Historical Regimes. *Proc. Natl. Acad. Sci.* **2019**, *116* (29), 14779–14779. https://doi.org/10.1073/pnas.1910465116.

(8) Gromski, P. S.; Henson, A. B.; Granda, J. M.; Cronin, L. How to Explore Chemical Space Using Algorithms and Automation. *Nat. Rev. Chem.* **2019**, *3* (2), 119–128. https://doi.org/10.1038/s41570-018-0066-y.

(9) Hoffmann, T.; Gastreich, M. The next Level in Chemical Space Navigation: Going Far beyond Enumerable Compound Libraries. *Drug Discov. Today* **2019**, *24* (5), 1148–1156. https://doi.org/10.1016/j.drudis.2019.02.013.

(10) van Hilten, N.; Chevillard, F.; Kolb, P. Virtual Compound Libraries in Computer-Assisted Drug Discovery. *J. Chem. Inf. Model.* **2019**, *59* (2), 644–

651. https://doi.org/10.1021/acs.jcim.8b00737.

(11) Simm, G. N.; Vaucher, A. C.; Reiher, M. Exploration of Reaction Pathways and Chemical Transformation Networks. *J. Phys. Chem. A* **2018**, *123* (2), 385–399. https://doi.org/10.1021/acs.jpca.8b10007.

(12) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat. Rev. Drug Discov.* **2005**, *4* (8), 649–663. https://doi.org/10.1038/nrd1799.

(13) Loving, K.; Alberts, I.; Sherman, W. Computational Approaches for Fragment-Based and De Novo Design. *Curr. Top. Med. Chem.* **2010**, *10* (1), 14–32. https://doi.org/10.2174/156802610790232305.

(14) Kutchukian, P. S.; Shakhnovich, E. I. De Novo Design: Balancing Novelty and Confined Chemical Space. *Expert Opin. Drug Discov.* **2010**, *5* (8), 789–812. https://doi.org/10.1517/17460441.2010.497534.

(15) Medina-Franco, J. L.; Martinez-Mayorga, K.; Meurice, N. Balancing Novelty with Confined Chemical Space in Modern Drug Discovery. *Expert Opin. Drug Discov.* **2014**, *9* (2), 151–165. https://doi.org/10.1517/17460441.2014.872624.

(16) Hartenfeller, M.; Proschak, E.; Schüller, A.; Schneider, G. Concept of Combinatorial De Novo Design of Drug-like Molecules by Particle Swarm Optimization. *Chem. Biol. Drug Des.* **2008**, *72* (1), 16–26. https://doi.org/10.1111/j.1747-0285.2008.00672.x.

(17) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: Reaction-Driven de Novo Design of Bioactive Compounds. *PLoS Comput. Biol.* **2012**, *8* (2), e1002380. https://doi.org/10.1371/journal.pcbi.1002380.

(18) 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.; Van Aken, K.; Janssen, P. A. J. SYNOPSIS: SYNthesize and OPtimize System in Silico. *J. Med. Chem.* **2003**, *46* (13), 2765–2773. https://doi.org/10.1021/jm030809x.

(19) Schneider, G.; Lee, M. L.; Stahl, M.; Schneider, P. De Novo Design of Molecular Architectures by Evolutionary Assembly of Drug-Derived Building Blocks. *J. Comput. Aided. Mol. Des.* **2000**, *14* (5), 487–494. https://doi.org/10.1023/a:1008184403558.

(20) Fechner, U.; Schneider, G. Flux (1): A Virtual Synthesis Scheme for Fragment-Based de Novo Design. *J. Chem. Inf. Model.* **2006**, *46* (2), 699–707. https://doi.org/10.1021/ci0503560.

(21) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform.* **2009**, *1* (1), 1–11. https://doi.org/10.1186/1758-2946-1-8.

(22) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Acs.org* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(23) Yang, X.; Zhang, J.; Yoshizoe, K.; Terayama, K.; Tsuda, K. ChemTS: An Efficient Python Library for de Novo Molecular Generation. *Sci. Technol. Adv. Mater.* **2017**, *18* (1), 972–976. https://doi.org/10.1080/14686996.2017.1401424.

(24) Besnard, J.; Ruda, G. F.; Setola, V.; Abecassis, K.; Rodriguiz, R. M.; Huang, X. P.; Norval, S.; Sassano, M. F.; Shin, A. I.; Webster, L. A.; Simeons, F. R. C.; Stojanovski, L.; Prat, A.; Seidah, N. G.; Constam, D. B.; Bickerton, G. R.; Read, K. D.; Wetsel, W. C.; Gilbert, I. H.; Roth, B. L.; Hopkins, A. L. Automated Design of Ligands to Polypharmacological Profiles. *Nature* **2012**, *492* (7428), 215–220. https://doi.org/10.1038/nature11691.

(25) Chevillard, F.; Kolb, P. 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.* **2015**, *55* (9), 1824–1835. https://doi.org/10.1021/acs.jcim.5b00203.

(26) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for de Novo Drug Design. *Sci. Adv.* **2018**, *4* (7), 1–15. https://doi.org/10.1126/sciadv.aap7885.

(27) Http://pubchem.ncbi.nlm.nih.gov/. The PubChem Database.

(28) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58* (2), 252–261. https://doi.org/10.1021/acs.jcim.7b00622.

(29) Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson, A. P. SPROUT, HIPPO and CAESA: Tools for de Novo Structure Generation and Estimation of Synthetic Accessibility. *Perspect. Drug Discov. Des.* **1995**, *3* (1), 34–50. https://doi.org/10.1007/BF02174466.

(30) Huang, Q.; Li, L.-L.; Yang, S.-Y. RASA: A Rapid Retrosynthesis-Based Scoring Method for the Assessment of Synthetic Accessibility of Drug-like Molecules. *J. Chem. Inf. Model.* **2011**, *51* (10), 2768–2777. https://doi.org/10.1021/ci100216g.

(31) Li, J.; Eastgate, M. D. Current Complexity: A Tool for Assessing the Complexity of Organic Molecules. *Org. Biomol. Chem.* **2015**, *13* (26), 7164–7176. https://doi.org/10.1039/c5ob00709g.

(32) Heifets, A. Automated Synthetic Feasibility Assessment: A Data-Driven Derivation of Computational Tools for Medicinal Chemistry. **2014**.

(33) Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103* (12), 3599–3601. https://doi.org/10.1021/ja00402a071.

(34) Https://new.reaxys.com/. Reaxys.

(35) Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. SYBA: Bayesian Estimation of Synthetic Accessibility of Organic Compounds. *J. Cheminform.* **2020**, *12* (1), 35. https://doi.org/10.1186/s13321-020-00439-2.

(36) Voršilák, M.; Svozil, D. Nonpher: Computational Method for Design of Hard-to-Synthesize Structures. *J. Cheminform.* **2017**, *9* (1), 1–7. https://doi.org/10.1186/s13321-017-0206-2.

(37) Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic Accessibility Score (RAscore) – Rapid Machine Learned Synthesizability Classification from AI Driven Retrosynthetic Planning. *Chem. Sci.* **2021**, *12* (9), 3339–3349. https://doi.org/10.1039/D0SC05401A.

(38) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: A Fast, Robust and Flexible Open-Source Software for Retrosynthetic Planning. *J. Cheminform.* **2020**, *12* (1), 70. https://doi.org/10.1186/s13321-020-00472-1.

(39) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864–2875. https://doi.org/10.1021/ci300415d.

(40) Grzybowski, B. A.; Bishop, K. J. M.; Kowalczyk, B.; Wilmer, C. E. The “wired” Universe of Organic Chemistry. *Nat. Chem.* **2009**, *1* (1), 31–36. https://doi.org/10.1038/nchem.136.

(41) Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature. **2012**. https://doi.org/10.17863/CAM.16293.

(42) Software, N.; Centre, I.; Science, C.; Milton, P.; Cb, C. Pistachio-Release 02 Mar 2020. **2020**.

(43) Thakkar, A.; Kogej, T.; Reymond, J. L.; Engkvist, O.; Bjerrum, E. J. Datasets and Their Influence on the Development of Computer Assisted Synthesis Planning Tools in the Pharmaceutical Domain. *Chem. Sci.* **2019**, *11* (1), 154–168. https://doi.org/10.1039/c9sc04944d.

(44) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3* (5), 434–443. https://doi.org/10.1021/acscentsci.7b00064.

(45) Coley, C. W.; Green, W. H.; Jensen, K. F. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. *J. Chem. Inf. Model.* **2019**, *59*, 2529–2537. https://doi.org/10.1021/acs.jcim.9b00286.

(46) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl,

P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*; 2008; pp 319–326. https://doi.org/10.1007/978-3-540-78246-9_38.

(47) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–182. https://doi.org/10.1021/ci049714+.

(48) Santa Fe N, U. www. eyesopen. co. OpenEye Scientific Software.

(49) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (4), 747–750. https://doi.org/10.1021/ci9803381.

(50) Blomberg, N.; Cosgrove, D. A.; Kenny, P. W.; Kolmodin, K. Design of Compound Libraries for Fragment Screening. *J. Comput. Aided. Mol. Des.* **2009**, *23* (8), 513–525. https://doi.org/10.1007/s10822-009-9264-5.

(51) Song, Y.; Zheng, S.; Niu, Z.; Fu, Z.; Lu, Y.; Yang, Y. Communicative Representation Learning on Attributed Molecular Graphs. **2020**, No. July, 2831–2838. https://doi.org/10.24963/ijcai.2020/392.

(52) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta - Protein Struct.* **1975**, *405* (2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9.