CReM: chemically reasonable mutations framework for structure generation

Pavel Polishchuk (pavel_polishchuk@ukr.net)

Software

Keywords: de novo structure generation, de novo design, matched molecular pairs

Posted Date: February 13th, 2020

DOI: https://doi.org/10.21203/rs.2.23402/v1

Version of Record: A version of this preprint was published at Journal of Cheminformatics on April 22nd, 2020. See the published version at https://doi.org/10.1186/s13321-020-00431-w.

Abstract

Structure generators are widely used in de novo design studies and their performance substantially influences an outcome. Approaches based on deep learning models and conventional atom-based approaches may result in invalid structures and did not address their synthetic feasibility issues. Conventional reaction-based approaches result in synthetically feasible compounds but novelty and diversity of generated compounds may be limited. Fragment-based approaches can provide better novelty and diversity of generated compounds but the issue of synthetic complexity of generated structure was not explicitly addressed before. Here, we developed a new fragment-based approach which results in chemically valid structures by design and gives flexible control over diversity, novelty, synthetic complexity and chemotypes of generated compounds. The approach was implemented as an open-source Python module.

Introduction

The drug-like chemical space is vastly enormous – its size estimates in ~10^{33} compounds [1]. In the nearest future it is impossible to enumerate this space or perform any kind of exhaustive search. Therefore, methods and strategies to explore this space effectively attract vivid research interest. One of popular strategies is de novo design – model-driven generation of new chemical structures with promising predicted properties [2-3]. Two major strategies of structure generation exist: i) iterative generation of structures to fit model predictions and ii) generation of structures having a desirable set of properties directly by machine learning (ML) models (e.g. inverse QSAR or generative neural networks). The former strategy is widely used and many studies describe different implementation schemes [4-9].

The interest in the second strategy rekindled recently due to the advances in deep learning and generative models [10-16]. Compound structure can be generated in unsupervised or supervised manner. In unsupervised approaches ML models (usually recurrent neural networks) are trained on structures of known compounds (usually represented by SMILES) and stochastically sample output structures. To make generation more focused a model trained on a large diverse set of compounds can be post-trained on a small subset of compounds active against a particular target. This creates a bias in the model and the model generates compounds more similar to the active ones [16]. In supervised approaches a model is optimized to find a combination of descriptors which correspond to active compounds and then structures of compounds are reconstructed from this descriptors set (so called “inverse QSAR” task). Conventional descriptors proved very difficult to reconstruct structures. Several approaches were developed in the past but gain low popularity due to many restrictions and limitations [17-22]. Recent advances in deep learning allowed to generate latent representation of input compounds, find an optimal set of the latent variables associated with desired property values and sample structures from this latent subspace of variables [23]. Nevertheless, no guarantee exists that the found optimal combinations of latent variables correspond to valid chemical structures. The percentage of valid structures generated by deep learning models can vary a lot depending on the deep learning model architecture from almost

100% to 4% [24]. Moreover, the currently developed approaches have no control over synthetic feasibility of generated structures which is usually estimated after compounds generation [23].

The general workflow of the first strategy for iterative chemical space exploration includes: i) generation of initial structures, ii) evaluation of generated structures by model(s) (QSAR, docking, pharmacophores, etc), iii) selection of the most promising candidates, iv) generation of new structures based on the selected ones and return to the step (ii). This procedure is repeated until compounds with desirable properties will be generated. Structure generation and property estimation steps are separated in this case. This makes it possible to use any combination of structure generation approaches and *in silico* models to predict properties of compounds. We can divide the conventional approaches in three groups: atom-based, fragment-based and reaction-based structure generators, each having their advantages and issues (Table 1).

Atom-based approaches represent "*ab initio"* methods among structure generators and use simple rules like "add/remove/replace atom/bond", "modify input structures" and "generate new ones" [25]. Theoretically it should be possible to generate every possible structure using these simple rules that can result in high novelty and diversity of structures. However, the cost is too many generation steps and consequently a combinatorial explosion. Therefore, atom-based approaches suit better for systematic exploration of a local chemical space. Chemical validity should be additionally controlled during structure generation to avoid erroneous structural changes. However, the main issue of atom-based approaches is synthetic feasibility, which cannot be controlled over the course of generation and may result in synthetically less accessible structures.

Reaction-based approaches generate new compounds by applying rules from a list of encoded chemical transformations to a library of reactants [7]. As it seems intuitively reaction-based approaches produce higher novelty and diversity in just a few generation steps compared to atom-based approaches, which may require many steps to achieve the same goal. Reaction-based approaches make large changes in structure during compounds generation and, therefore, seem more suitable for rough exploration of chemical space. With a comprehensive reactant library it would be also possible to enumerate close analogs of a reference compound to perform a local exploration of chemical space. Synthetic feasibility of generated compounds and an available synthetic route are the main advantages of reaction-based approaches. Applicability of this kind of approaches was demonstrated in several studies [7; 26-28]. Nevertheless, the limited number of rules and limited size of library may restrain these algorithms from exploring larger chemical space (therefore losing novelty and diversity of generated compounds).

Fragment-based approaches lie in between of atom-based and reaction-based ones [5; 9; 29]. The set of initial fragments directly determines novelty and diversity of generated compounds. So we expect that fragment-based approaches will outperform reaction-based ones (in terms of diversity and novelty) as it seems easier to collect a diverse library of fragments than a diverse library of reactants. One can also control exhaustiveness of chemical space exploration by varying the size of fragments. However, an accessible chemical space is smaller than for atom-based approaches. Chemical validity of enumerated

structures seems to be easier to control in case of fragments, but synthetic feasibility still presents an issue because linking of synthetically feasible fragments may result in synthetically infeasible molecules. Therefore, fragment linking should take into account chemical context of coupled fragments. Recently Liu et al. published an approach where they took into account types of atoms which can be linked to particular attachment points [30]. Their strategy resulted in generation of chemically valid structures but the authors did not study synthetic feasibility of generated structures and the context of one atom may be insufficient to guarantee generation of synthetically feasible molecules.

Table 1. Features of structure generation algorithms.

	atom-based	fragment-based	reaction-based
exhaustiveness of chemical space search	+++ suitable for systematic exploration of local chemical space	++ variable, controlled by the size and diversity of fragments	+ depends on diversity of a reactant library and a list of annotated reaction rules
structure novelty	+++ many steps to achieve high novelty	++	+
structure diversity	+++ many steps to achieve high diversity	++	+
chemically valid structures	-	-/+	+++
synthetically feasible structures	---	-/+	+++
time-consuming due to combinatorial explosion	+++	--	---

In this study a novel algorithm for fragment-based structure generation is suggested that is based on determination of interchangeable fragments from databases of known compounds to perform chemically reasonably mutations (CReM) of input structures. It generates chemically valid structures by design and allows to indirectly control synthetic feasibility of enumerated compounds as well as their chemotypes.

Implementation

The idea of interchangeable fragments – the core of the developed approach – is directly related to the matched molecular pairs approach considering their local context [31]. Interchangeable fragments are fragments that occur in the same local chemical context in structures of known compounds (Figure 1). Atoms within a particular radius around attachment points of a fragment represent this local chemical

context. We replace one fragment by another with the same chemical context, which should result in a chemically valid and feasible structure. Thus, by design chemical validity of generated structures is guaranteed. Intuitively it can be also expected that generated compounds are synthetically feasible.

Generation of a database of interchangeable fragments is a two-step procedure. On the first step structures of known compounds are exhaustively fragmented by cutting up to 4 non-cyclic single bonds (including bonds to hydrogen atoms). On the second step a context of a given radius is determined for attachment points of each fragment and encoded in a SMILES string. This SMILES string is canonicalized to get both a canonical numbering of attachment points and canonical SMILES representation of a context. Attachment points in a corresponding fragment are renumbered correspondingly. SMILES representation of a context of a given radius and an associated fragment are stored in a database table as a key-value pair for a subsequent search of interchangeable fragments (values) having an identical context (key) (Figure 1). If a context of two or more attachment points is identical, all possible permutations of these attachment points in a corresponding fragment are performed. The numbers of attachment points in the context are not changed because this will result in the same canonical SMILES representation. Fragments with alternative attachment point numbering are stored individually as key-value pairs (context in this case is the same). This situation is illustrated in Figure 2. The central fragment has three attachment points. Two of them, methylene groups linked by a single bond to a remaining part of a molecule, having numbers 1 and 2 are identical at the context of radius 1. Therefore, all possible permutations of attachment points in a fragment which are consistent with the attachment point numbering in a context are enumerated. In this case 1 and 2 are swapped and both fragments with different numbering are stored in a database. This is done to be able to make all possible replacements if some attachment points are equivalent.

To replace a fragment in a molecule its context of a given radius is determined and canonically encoded. The given SMILES string of a context is searched in a fragment database and fragments with the same context are retrieved and used for fragment replacement (Figure 2).

We implemented three modes of structure generation: MUTATE, GROW and LINK (Figure 3). Mutate is a replacement of an arbitrary fragment with another one. GROW is a special case of a MUTATE operation – replacement of a hydrogen with another fragment. LINK is a replacement of hydrogen atoms in two molecules to link them by an appropriate fragment. Additionally we provided an option to all these functions to protect particular atoms from modifications. In particular, this functionality can be useful for property/activity optimization studies to protect scaffold or pharmacophore features from changes.

There are several tuning parameters available:

1. structures of input compound used to create a database of interchangeable fragments;
2. radius of a considered molecular context;
3. frequency of occurrence of interchangeable fragments in the input database;
4. size of fragments which will replace each other;

5. maximum number of randomly chosen replacing fragments;
6. protection of selected atoms from modification.

The size of replaceable fragments can control exhaustiveness of chemical space exploration by increasing or decreasing search steps and depends on a goal of a particular study, thus, will not be investigated here. Structure optimization studies may require small steps to explore local chemical space around a parent compound whereas de novo design may require large steps in the beginning to quickly and coarsely explore larger chemical space and smaller steps in the end to finely tune generated structures.

Management of the content of the input compound database used for fragmentation gives indirect control over enumerated structures and provides additional flexibility. Selection of synthetically feasible input compounds may improve synthetic feasibility of generated compounds. The similar effect might be achieved by selection of frequently occurred context-fragment pairs which can be considered more synthetically feasible, which is similar to the synthetic accessibility score suggested by Ertl & Schuffenhauer [32]. At the same time pre-selection of compounds for fragment library enumeration may reduce diversity and novelty of generated structures. Increasing the radius of a considered molecular context will decrease appearance of new chemotypes in enumerated compounds and make replacements more conservative. With all these options the developed approach possesses great flexibility and control over generated structures. We will study tuning effects in the next section.

The maximum number of randomly selected replacements can speed up exploration of a chemical space because fragment databases generated can be very large and making all possible replacements can be costly. We will not investigate this parameter in this study.

The major limitation of the current implementation is that new ring systems cannot be created because rings are not cut during the fragmentation step and are replaced as a whole. Therefore, representativeness of ring systems in generated structures completely depends on an input compound database used for generation of a database of interchangeable fragments.

There are still no commonly used criteria to measure performance of structure generators and quality of virtually enumerated libraries. Recently several papers were published to address this issue [33-34]. In this study we will use Guacamol goal-directed benchmark to demonstrate general applicability of the CReM approach. Guacamol is a set of 20 tasks which goal is to rediscover known drugs, generate compounds similar to the reference ones, perform multi-objective optimization of properties of known drugs, or make scaffold hopping [33].

Additionally, we will simulate local exploration of DrugBank compounds in order to explore dependence of novelty, diversity and synthetic complexity of generated compounds on CReM tuning parameters: content of a fragment database, context radius and frequency of occurrence of fragment-context pairs. Novelty of generated compounds will be calculated as mean Tanimoto distance to a parent compound based on 2048-bit Morgan fingerprints of radius 2 calculated in RDKit. This will show how dissimilar the

generated compounds are from a parent compound (the higher score the better). Diversity of generated compounds will be calculated as mean Tanimoto distance based on 2048-bit Morgan fingerprints of radius 2 between all pairs of generated compounds. If the number of generated compounds will be large a random subset of 1000 compounds will be used to estimate diversity of generated structures. This procedure was repeated five times to estimate robustness of obtained value. Diversity will show how intrinsically diverse generated compounds are (the higher score the better). It is expected that novelty and diversity will be highly correlated because novelty as defined above can be interpreted as diversity relative to a reference compound. Synthetic complexity is predicted by the model developed by Coley et al [35]. The synthetic complexity score (SCScore) value is within the range from 1 to 5 where synthetically feasible compounds have score 1 whereas synthetically complex compounds are closer to score 5 (the lower score the better).

Results And Discussion

Fragment databases generation

ChEMBL database (version 22) was used as a source of structures for the databases of interchangeable fragments. 1 557 992 distinct structures containing only organic elements (C, N, O, S, P, F, Cl, Br, I, B) remained after curation procedure. The curation was performed using the previously developed protocol [36] which includes Chemaxon Stardardizer and Checker [37] and RDKit [38] sanitization checks.. To estimate the context radius' effect on the generated compounds we generated two databases with context radius from 1 to 5. The first database was generated from all ChEMBL compounds. The second one was generated from compounds without PAINS (1 464 907 compounds). The number of distinct context and fragment combinations grew linearly with context radius increase (Table 2).

Table 2. The number of distinct fragments and corresponding contexts in databases generated from the whole ChEMBL data set and its PAINS-less subset.

radius	ChEMBL	PAINS-less ChEMBL
1	35 833 160	34 240 810
2	41 676 473	39 800 734
3	51 730 960	49 403 630
4	62 821 316	59 971 431
5	74 168 168	70 717 996

To investigate the hypothesis that limiting synthetic complexity in fragmented structures improves synthetic accessibility of generated compounds we created fragment libraries from ChEMBL compounds having SCScore below a specified threshold (2, 2.5, 3, 3.5). The fixed context radius 3 was chosen as a reference because at this radius most functional groups are distinguishable and therefore structural replacements are reasonably specific. At smaller radius some functional groups cannot be distinguished, e.g. N-substituted amide (*-N-C(=O)) and amino (*-N-C-C) groups. Larger radius in this case would result

in too specific replacements because it would be able to distinguish amides differently substituted at α-carbon atom. The number of compounds and resulted fragment and context pairs substantially decreased with lowering the SCScore threshold (Table 3).

Table 3. Statistics of the initial ChEMBL data set and filtered ones by SCscore values and the number of resulted distinct fragment and context pairs.

Data set	number of compounds	number of distinct fragments & contexts of radius 3
ChEMBL	1 557 992	51 730 960
SCscore <= 3.5	552 162	20 514 883
SCscore <= 3	284 461	10 661 179
SCscore <= 2.5	111 365	4 091 634
SCscore <= 2	27 916	951 993

Parent compounds selection

DrugBank compounds were selected as a source of parent compounds for simulation of a local exploration of a chemical space. 6002 compounds were left after curation of the whole database using the same protocol mentioned above. They were ranked according to their SCScore values and each twelfth compound was selected to create a subset of 500 compounds.

The MUTATE operation was applied to the selected 500 compounds. We set the minimum size of the replaced fragment to 0 to enable replacement of hydrogens. The maximum size of the replaced fragment was up to 10 heavy atoms. The size of a replacing fragment could be smaller or larger that the replaced fragment by at most ±2 heavy atoms.

Influence of a context radius on generated compound sets

The database of fragments generated from the whole ChEMBL data set was used to study influence of a context radius on generated sets of compound. Starting from the selected 500 DrugBank compounds the corresponding number of data sets was generated; after that average novelty, diversity and synthetic complexity were calculated for each data set. As we expected the number of generated compounds substantially dropped with increase of context radius because a smaller number of matching context-fragment pairs were found in a fragment database. Average novelty and diversity decreased more smoothly and generated data sets still covered a large range of these values. Synthetic complexity seemed to be the least affected but it also decreased with increase of context radius (Figure 4).

The general distributions of data set parameters depicted in Figure 4 struggle to give a complete picture of properties of the generated data sets. Therefore, we calculated difference plots to emphasize the changes of parameters depending on a context radius for each data set. We selected data sets generated at the context radius 3 as reference points and calculated differences between corresponding parameters of corresponding data sets generated at other context radius and the reference data sets (Figure 5). The

results revealed the same trend in a more pronounced way. Despite of substantial drop in the number of generated compounds up to 1 million, novelty, diversity and synthetic complexity mainly were not decreased much with increasing of the context radius from 1 to 3. Increase in the context radius to 4 or 5 resulted in a continuation of decrease of the number of generated compounds on up to 100 000 and slight decrease of average novelty, diversity and synthetic complexity.

Control over synthetic complexity

We examined two possible strategies to reduce synthetic complexity of generated compounds. The first one is based on an idea that constructing structures from frequently occurring fragments results in more synthetically feasible compounds [32]. We used the whole ChEMBL fragment database limiting replacements by the fragment occurrence in the database: no limits or minimum 10, 100 or 1000 occurrences. The second strategy is based on fragment databases generated from more synthetically accessible compounds selected according to SCScore [35]. Here we used databases generated from compounds having SCScore value not greater than 2, 2.5, 3 or 3.5. No restrictions on fragment occurrence were applied in this case. 500 DrugBank compounds were used as a starting structures and the corresponding number of compound sets were generated using each of these strategies.

Table 4 shows the number of times each generation strategy resulted in a data set with best average synthetic complexity, novelty and diversity. The strategy, which used whole ChEMBL fragment database with no restriction, (ChEMBL & 0) resulted in the highest novelty and diversity of enumerated data sets in about 70% of cases, whereas the lowest average synthetic complexity was observed in only 13 cases (2.6%). According to synthetic complexity of generated compounds the top strategy with the greatest number of wins (205 cases or 41%) was based on the fragment database generated from compounds with SCScore $\leq$ 2. Limitation of fragment replacements based on fragment occurrence improved synthetic complexity of generated compounds, but substantially decreased the number of output compounds. The second best strategy resulted in 169 data sets (33.8%) with lowest synthetic complexity used fragments occurred at least 1000 times in the whole ChEMBL database. However, the average number of output compounds was 15 in this case whereas the strategy used the fragment database generated from compounds with SCScore $\leq$ 2 resulted in 430 compounds in average. The latter strategy outperformed all strategies based on fragment occurrence limitation by the number of generated compounds and their synthetic complexity.

Table 4. Comparison of different generation strategies.

fragment database & occurrence	wins (by lower mean SCscore)	wins (by higher mean novelty)	wins (by higher mean diversity)	mean/median number of compounds
ChEMBL & 0	13	357	347	7812 / 3800
ChEMBL & 10	5	2	47	695 / 323
ChEMBL & 100	17	1	7	98 / 66
ChEMBL & 1000	169	0	4	19 / 15
SCScore ≤ 3.5 & 0	16	55	61	5331 / 2540
SCScore ≤ 3 & 0	19	29	24	4031 / 1898
SCScore ≤ 2.5 & 0	56	16	9	2473 / 1122
SCScore ≤ 2 & 0	205	40	1	1040 / 430
total number of compounds:	500	500	500	

Synthetic complexity reduced more pronounced in cases where fragment databases generated from more synthetically accessible compounds were used (Figure 6). In the case of the fragment database created from compounds with SCScore ≤ 2 the average decrease was 0.26 that is comparable to the value 0.25 imposed by the authors of SCScore as an objective during optimization to separate reactants and products. Therefore, it might be expected that compounds having SCScore value lower by 0.25 would require one less step to be synthesized.

Control over chemotypes of generated compounds

Radius of considered fragments context implicitly determines the chemotypes of generated structures. Within the developed approach we cannot create the fragments that did not occur in the input database and have the size equal or less than a chosen radius. Here size means the longest distance between two atoms in a fragment, where the distance is the shortest path between two atoms. So, by increasing the radius one can make a generation of structures more conservative relatively to chemotypes of compounds used for generation of a fragment database. This can be useful in a scenario, where compounds with undesirable patterns are removed from the input database and a fragment database is generated with a reasonably large radius to avoid generation of compounds with these undesirable fragments (chemotypes). At the same time increasing the radius can decrease the number of generated compounds and their diversity and novelty. Therefore, the choice of the radius is a trade-off between these options.

We simulated unrestricted iterative stochastic exploration of a chemical space starting from a benzene molecule to demonstrate this ability. On each iteration the MUTATE operation was applied to input compounds to perform all possible replacements. Generated compounds with molecular mass greater than 500 were discarded. The remaining compounds were ordered by molecular mass and split into 5 bins. One compound was randomly selected from each bin. The selected five compounds passed to the next iteration. Totally 100 iterations were executed.

This procedure was run once per each context radius from 1 to 5 using the PAINS-less fragment database. Then compounds generated on all iterations of each run were collected and examined to contain PAINS fragments. As expected, the number of emerged PAINS patterns decreased with increase of context radius: 102 distinct PAINS patterns for radius 1 were detected, 52 for radius 2, 28 for radius 3, 26 for radius 44 and 1 for radius 5. The last one is dialkyl aniline moiety having a methoxy substituent at the position 4 (Figure 7). The longest distance in this pattern equals to 6 bonds between nitrogen atom and methoxy carbon atom, therefore the context of radius 5 could not fully cover it and it was reconstructed during structural mutations. Only one PAINS pattern was generated due to stochastic nature of the simulation but other PAINS fragments having the size greater than 5 may appear in structures generated with context radius 5. However, no smaller patterns were detected. This demonstrated that one can implicitly control chemotypes of generated compounds by increase of context radius. The full list of determined PAINS patterns is provided in Additional file 1.

Guacamol benchmarking

To demonstrate general applicability of the CReM approach we used Guacamol benchmarks. The Guacamol tasks are diverse and might require diverse search strategies – as it is difficult to expect that one strategy will demonstrate the optimal performance at every task. We implemented a single iterative algorithm and applied it to all Guacamol tasks not tuning it to separate tasks.

If the list of seed structures was empty the seed structures were chosen randomly from the list of SMILES supplied with Guacamol and represented the whole ChMEBL database. The size of a population selected on each iteration was set be equal to the size of the output population but not less than 10 compounds. To make the search adaptive we adjusted fragment size of replacement according to the current score of the population. If the score was equal or less than 0.3 (far from the goal) the replacing fragment can differ at most on ±10 heavy atoms from the replaced one. If the score was greater than 0.8 (close to the goal) the replacing fragment can differ at most on ±4 heavy atoms from the replaced one. Intermediate fragment sizes (5-9) were chosen if the score was within 0.3 - 0.8 range. This allows to quickly explore chemical space in the beginning and better tune structures at the end of generation. For each compound in a population up to 1000 randomly chosen replacements were applied. Compounds, which were already used for structure generation, were stored in a separate list and removed from the list of generated structures. Remaining top scored compounds were selected for the next iteration.

Since the implemented optimization procedure is local and can get stuck in local optima we implemented three levels of “patience”. At the first level if the best score was not improved after three consecutive iterations the fragment size was increased on ±1 and the number of randomly chosen replacements on 100 irrespectively to the current score. This makes small stepwise increase in chemical space exploration. If after 10 iterations no improvement was observed larger changes were applied: the fragment size was increased on ±10 and the number of replacements on 500. This would enable rougher exploration of a chemical space around best candidates. At the third level, if after 33 iteration no improvement was observed new seed compounds were randomly selected to restart the search but best found candidates

were kept. This procedure was not applied if the seed structure was supplied with the task. The list of already visited compounds was cleared after any change of generator parameters whether this was caused by improving of the best score or by exceeding of one of “patience” levels.

Some of the target benchmark compounds contain complex ring systems. Therefore, due to current limitation of the implemented CReM approach to generate new ring systems the whole ChEMBL fragment database was used in this study. Maximum execution time of each task was set to 5 hours or maximum 1000 iterations were allowed.

The results demonstrated that the implemented search algorithm based on CReM approach compared well with the published reference approaches by achieving the highest score in 16 out of 20 tasks. However, the total score was slightly lower than the total score of Graph GA approach, which uses genetic algorithm on molecular graphs. This is mainly due to the considerable advantage demonstrated by Graph GA approach (0.891) over CReM-based approach (0.763) in the task of generation of molecules, which were as structurally dissimilar to sitagliptin as possible but had similar lipophilicity and topological polar surface area. Interestingly, the other reference approaches performed even worse in this task. Output results and tuning parameters are available in Additional files 2 and 3.

Table 5. Results for Guacamol benchmarks.

task	SMILES LSTM*	SMILES GA*	Graph GA*	Graph MCTS*	CReM
Celecoxib rediscovery	**1.000**	0.732	**1.000**	0.355	**1.000**
Troglitazone rediscovery	**1.000**	0.515	**1.000**	0.311	**1.000**
Thiothixene rediscovery	**1.000**	0.598	**1.000**	0.311	**1.000**
Aripiprazole similarity	**1.000**	0.834	**1.000**	0.380	**1.000**
Albuterol similarity	**1.000**	0.907	**1.000**	0.749	**1.000**
Mestranol similarity	**1.000**	0.79	**1.000**	0.402	**1.000**
C11H24	**0.993**	0.829	0.971	0.410	0.966
C9H10N2O2PF2Cl	0.879	0.889	**0.982**	0.631	0.940
Median molecules 1	**0.438**	0.334	0.406	0.225	0.371
Median molecules 2	0.422	0.38	0.432	0.170	**0.434**
Osimertinib MPO	0.907	0.886	0.953	0.784	**0.995**
Fexofenadine MPO	0.959	0.931	0.998	0.695	**1.000**
Ranolazine MPO	0.855	0.881	0.92	0.616	**0.969**
Perindopril MPO	0.808	0.661	0.792	0.385	**0.815**
Amlodipine MPO	0.894	0.722	0.894	0.533	**0.902**
Sitagliptin MPO	0.545	0.689	**0.891**	0.458	0.763
Zaleplon MPO	0.669	0.413	0.754	0.488	**0.770**
Valsartan SMARTS	0.978	0.552	0.990	0.04	**0.994**
Deco Hop	0.996	0.970	**1.000**	0.590	**1.000**
Scaffold Hop	0.998	0.885	**1.000**	0.478	**1.000**
total score	17.341	14.398	**17.983**	9.011	17.919

* results were taken from the ref [35]

Conclusion

The developed CReM approach of structural transformation generates chemically valid structures by design. It also enables one to indirectly influene generation outcome by customizing of an input compound database, which is used for generation of a database of interchangeable fragments. Ability to select more synthetically accessible input compounds for fragmentation can result in more synthetically accessible generated compounds. The performed experiments showed that even with a small library of synthetically feasible compounds (27916 ChEMBL compounds with SCscore $\leq$ 2) one may generate rather diverse sets of structures with better predicted synthetic feasibility. However, there is always a trade-off between the size of a fragment database and novelty and diversity of generated structures. End user can also choose a more conservative generation strategy by increasing the value of considered context radius. This allows to avoid generation of new structural motifs with size lesser than the chosen radius value. Combination of a custom input compound database and chosen context radius gives flexible control over the chemotypes of generated structures, their number, diversity and synthetic feasibility. The major limitation of the current implementation is inability to create new ring systems. Therefore, diversity of ring systems in generated compounds completely depends on their representativeness in the input compound database. The developed approach can be used in combination with any modeling tools to iteratively explore chemical space and optimize compound properties.

Implementation details

The developed software relies on SMILES representation on RDKit. Changes in SMILES representation will affect generated fragment databases. Therefore, databases generated with RDKIt 2017.09 version are not compatible with newer versions of RDKit.

Supplementary Information

Additional file 1. The lists of PAINS patterns found in generated structures based on the PAINS-less ChEMBL fragment database.

Additional file 2. Output of Guacamol benchmarking software.

Additional file 3. Tuning parameters used to generate compounds based on CReM framework for Guacamol tests.

Declarations

Availability and requirements

Project name: CReM

Project home page: http://www.qsar4u.com/pages/crem.php

GitHub: https://github.com/DrrDom/crem

Operating system(s): cross-platform

Programming language: Python 3

Other requirements: RDKit 2017.09 or higher

License: BSD 3-clause

Any restrictions to use by non-academics: no

Acknowledgements

The author thanks Guzel Mindubaeva for implementation and testing of some functions and Dr. Olena Mokshyna for critical reading of the manuscript.

Author's contribution

The author read and approved the final manuscript.

Funding

This research was funded by the Ministry of Education, Youth and Sports of the Czech Republic within the INTER-EXCELLENCE LTARF18013 project (agreement number MSMT-5727/2018-2).

Competing interests

The author declares no competing interests.

References

1. 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., 27:675-679. doi:http://dx.doi.org/10.1007/s10822-013-9672-4
2. Schneider P, Schneider G (2016) De Novo Design at the Edge of Chaos. J.Med.Chem., 59:4077-4086. doi:10.1021/acs.jmedchem.5b01849
3. Schneider G (2017) Automating drug discovery. Nature Reviews Drug Discovery, 17:97. doi:10.1038/nrd.2017.232
4. Böhm H-J (1992) The computer program LUDI: A new method for the de novo design of enzyme inhibitors. J. Comput.-Aided Mol. Des., 6:61-78. doi:10.1007/bf00124387
5. Wang R, Gao Y, Lai L (2000) LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. Molecular modeling annual, 6:498-516. doi:10.1007/s0089400060498

6. Brown N, McKay B, Gilardoni F, Gasteiger J (2004) A Graph-Based Genetic Algorithm and Its Application to the Multiobjective Evolution of Median Molecules. J. Chem. Inform. Comput. Sci., 44:1079-1087. doi:10.1021/ci034290p
7. 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 Computational Biology, 8:e1002380.
8. Firth NC, Atrash B, Brown N, Blagg J (2015) MOARF, an Integrated Workflow for Multiobjective Optimization: Implementation, Synthesis, and Biological Evaluation. Journal of Chemical Information and Modeling, 55:1169-1180. doi:10.1021/acs.jcim.5b00073
9. Chéron N, Jasty N, Shakhnovich EI (2016) OpenGrowth: An Automated and Rational Algorithm for Finding New Protein Ligands. J.Med.Chem., 59:4171-4188. doi:10.1021/acs.jmedchem.5b00886
10. Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. Journal of Cheminformatics, 9:48. doi:10.1186/s13321-017-0235-x
11. Popova M, Isayev O, Tropsha A (2017) Deep Reinforcement Learning for De-Novo Drug Design.
12. Yuan W, Jiang D, Nambiar DK, Liew LP, Hay MP, Bloomstein J, Lu P, Turner B, Le Q-T, Tibshirani R, Khatri P, Moloney MG, Koong AC (2017) Chemical Space Mimicry for Drug Discovery. Journal of Chemical Information and Modeling, 57:875-882. doi:10.1021/acs.jcim.6b00754
13. Li Y, Zhang L, Liu Z (2018) Multi-Objective De Novo Drug Design with Conditional Graph Generative Model.
14. Polykovskiy D, Zhebrak A, Vetrov D, Ivanenkov Y, Aladinskiy V, Mamoshina P, Bozdaganyan M, Aliper A, Zhavoronkov A, Kadurin A (2018) Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery. Molecular Pharmaceutics, 15:4398-4405. doi:10.1021/acs.molpharmaceut.8b00839
15. Putin E, Asadulaev A, Ivanenkov Y, Aladinskiy V, Sanchez-Lengeling B, Aspuru-Guzik A, Zhavoronkov A (2018) Reinforced Adversarial Neural Computer for de Novo Molecular Design. Journal of Chemical Information and Modeling doi:10.1021/acs.jcim.7b00690
16. Segler MHS, Kogej T, Tyrchan C, Waller MP (2018) Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Central Science, 4:120-131. doi:10.1021/acscentsci.7b00512
17. Skvortsova MI, Baskin II, Slovokhotova OL, Palyulin VA, Zefirov NS (1993) Inverse problem in QSAR/QSPR studies for the case of topological indexes characterizing molecular shape (Kier indices). J. Chem. Inform. Comput. Sci., 33:630-634. doi:10.1021/ci00014a017
18. Faulon J-L, Churchwell CJ, Visco DP (2003) The Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences. J. Chem. Inform. Comput. Sci., 43:721-734. doi:10.1021/ci020346o
19. Faulon J-L, Visco DP, Pophale RS (2003) The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. J. Chem. Inform. Comput. Sci., 43:707-720. doi:10.1021/ci020345w

20. Miyao T, Arakawa M, Funatsu K (2010) Exhaustive Structure Generation for Inverse-QSPR/QSAR. Molecular Informatics, 29:111-125. doi:10.1002/minf.200900038
21. Miyao T, Kaneko H, Funatsu K (2016) Inverse QSPR/QSAR Analysis for Chemical Structure Generation (from y to x). Journal of Chemical Information and Modeling, 56:286-299. doi:10.1021/acs.jcim.5b00628
22. Miyao T, Funatsu K (2017) Finding Chemical Structures Corresponding to a Set of Coordinates in Chemical Descriptor Space. Molecular Informatics, 36:1700030-n/a. doi:10.1002/minf.201700030
23. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A (2018) Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Central Science, 4:268-276. doi:10.1021/acscentsci.7b00572
24. Elton DC, Boukouvalas Z, Fuge MD, W. CP (2019) Deep learning for molecular generation and optimization - a review of the state of the art. arxiv
25. Hoksza D, Škoda P, Voršilák M, Svozil D (2014) Molpher: a software framework for systematic chemical space exploration. Journal of Cheminformatics, 6:7. doi:10.1186/1758-2946-6-7
26. Szymkuć S, Gajewska EP, Klucznik T, Molga K, Dittwald P, Startek M, Bajczyk M, Grzybowski BA (2016) Computer-Assisted Synthetic Planning: The End of the Beginning. Angewandte Chemie International Edition, 55:5904-5937. doi:10.1002/anie.201506101
27. Batiste L, Unzue A, Dolbois A, Hassler F, Wang X, Deerain N, Zhu J, Spiliotopoulos D, Nevado C, Caflisch A (2018) Chemical Space Expansion of Bromodomain Ligands Guided by in Silico Virtual Couplings (AutoCouple). ACS Central Science, 4:180-188. doi:10.1021/acscentsci.7b00401
28. Merk D, Grisoni F, Friedrich L, Gelzinyte E, Schneider G (2018) Computer-Assisted Discovery of Retinoid X Receptor Modulating Natural Products and Isofunctional Mimetics. J.Med.Chem., 61:5442-5447. doi:10.1021/acs.jmedchem.8b00494
29. Kutchukian PS, Lou D, Shakhnovich EI (2009) FOG: Fragment Optimized Growth Algorithm for the de Novo Generation of Molecules Occupying Druglike Chemical Space. Journal of Chemical Information and Modeling, 49:1630-1642. doi:10.1021/ci9000458
30. Liu T, Naderi M, Alvin C, Mukhopadhyay S, Brylinski M (2017) Break Down in Order To Build Up: Decomposing Small Molecules for Fragment-Based Drug Design with eMolFrag. Journal of Chemical Information and Modeling, 57:627-631. doi:10.1021/acs.jcim.6b00596
31. Dalke A, Hert J, Kramer C (2018) mmpdb: An Open-Source Matched Molecular Pair Platform for Large Multiproperty Data Sets. Journal of Chemical Information and Modeling, 58:902-910. doi:10.1021/acs.jcim.8b00173
32. Ertl P, Schuffenhauer A (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of Cheminformatics, 1:8. doi:10.1186/1758-2946-1-8
33. Brown N, Fiscato M, Segler MHS, Vaucher AC (2019) GuacaMol: Benchmarking Models for de Novo Molecular Design. Journal of Chemical Information and Modeling doi:10.1021/acs.jcim.8b00839

34. Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Sergey Nikolenko, Alan Aspuru-Guzik, Zhavoronkov A (2019) Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. arxiv
35. Coley CW, Rogers L, Green WH, Jensen KF (2018) SCScore: Synthetic Complexity Learned from a Reaction Corpus. Journal of Chemical Information and Modeling, 58:252-261. doi:10.1021/acs.jcim.7b00622
36. Structure sanitization workflow (2019). https://bitbucket.imtm.cz/projects/STD/repos/std/browse.
37. JChem 19.2.0 (2019). ChemAxon http://www.chemaxon.com.
38. RDKit: Open-Source Cheminformatics Software 2017.09 (2017). http://rdkit.org/.
39. Schomburg K, Ehrlich H-C, Stierand K, Rarey M (2010) From Structure Diagrams to Visual Chemical Patterns. Journal of Chemical Information and Modeling, 50:1529-1535. doi:10.1021/ci100209a

Figures

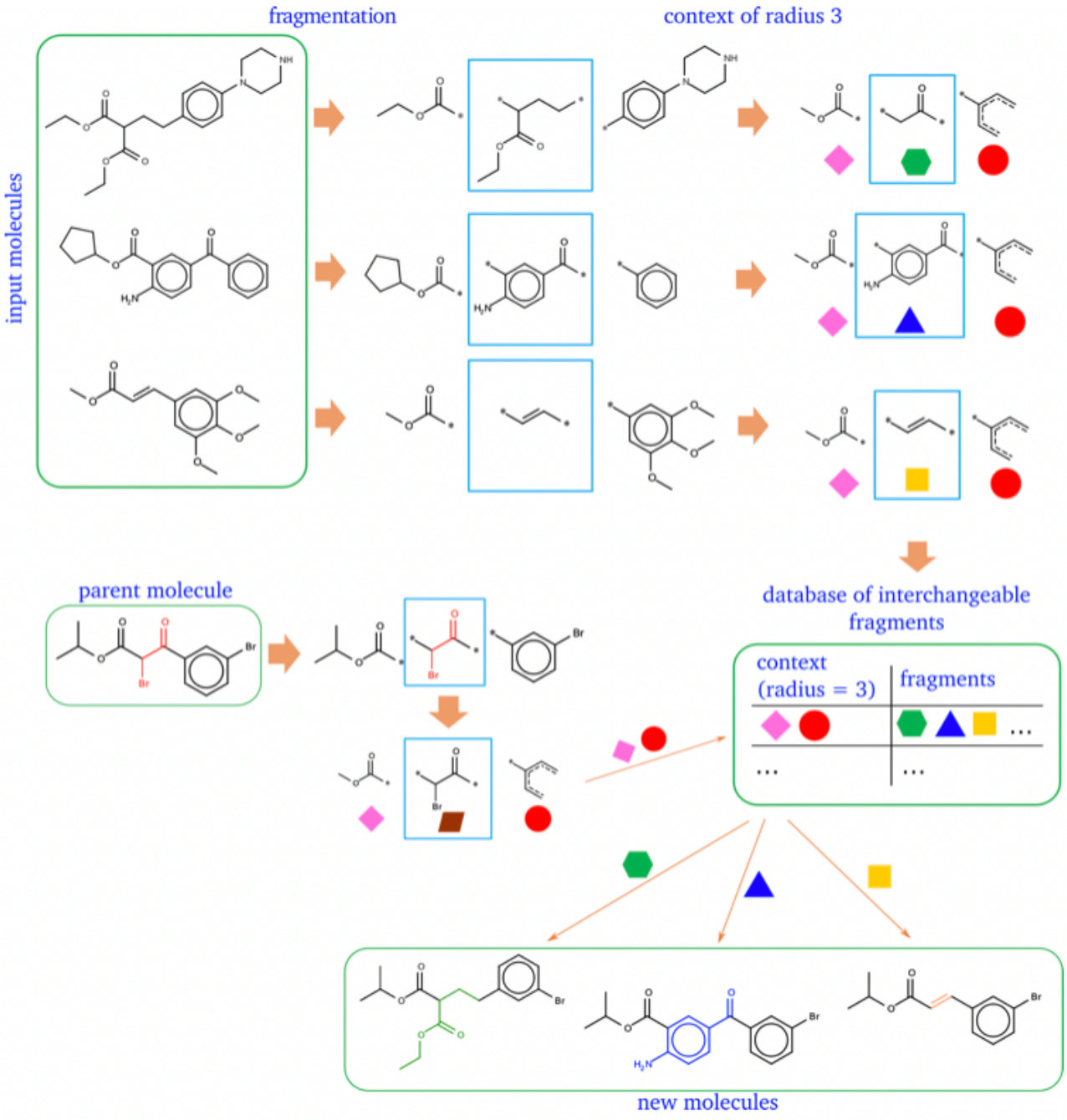

Figure 1

Generation of a database of interchangeable fragments and new molecules.

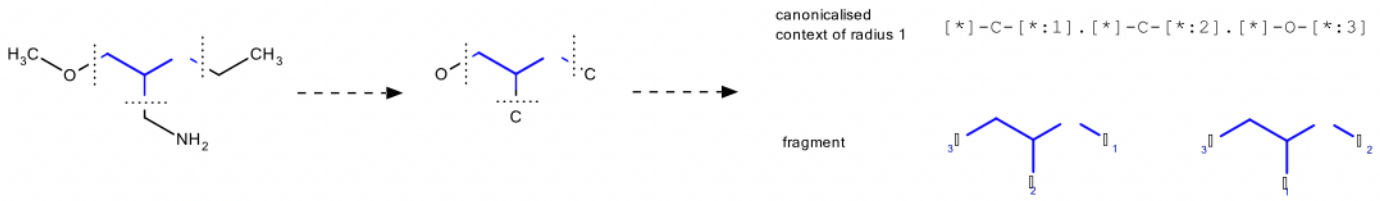

Figure 2

Canonicalization of attachment point numbers in contexts and fragments.

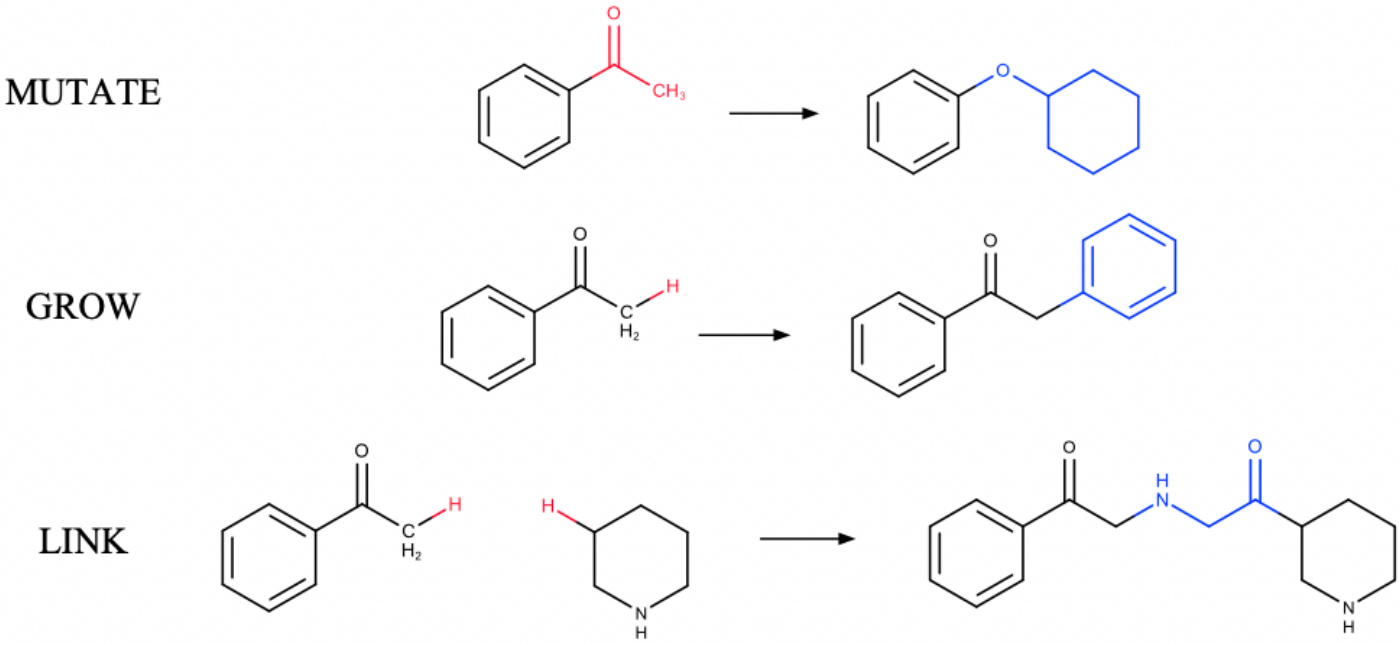

Figure 3

Structure generation modes.

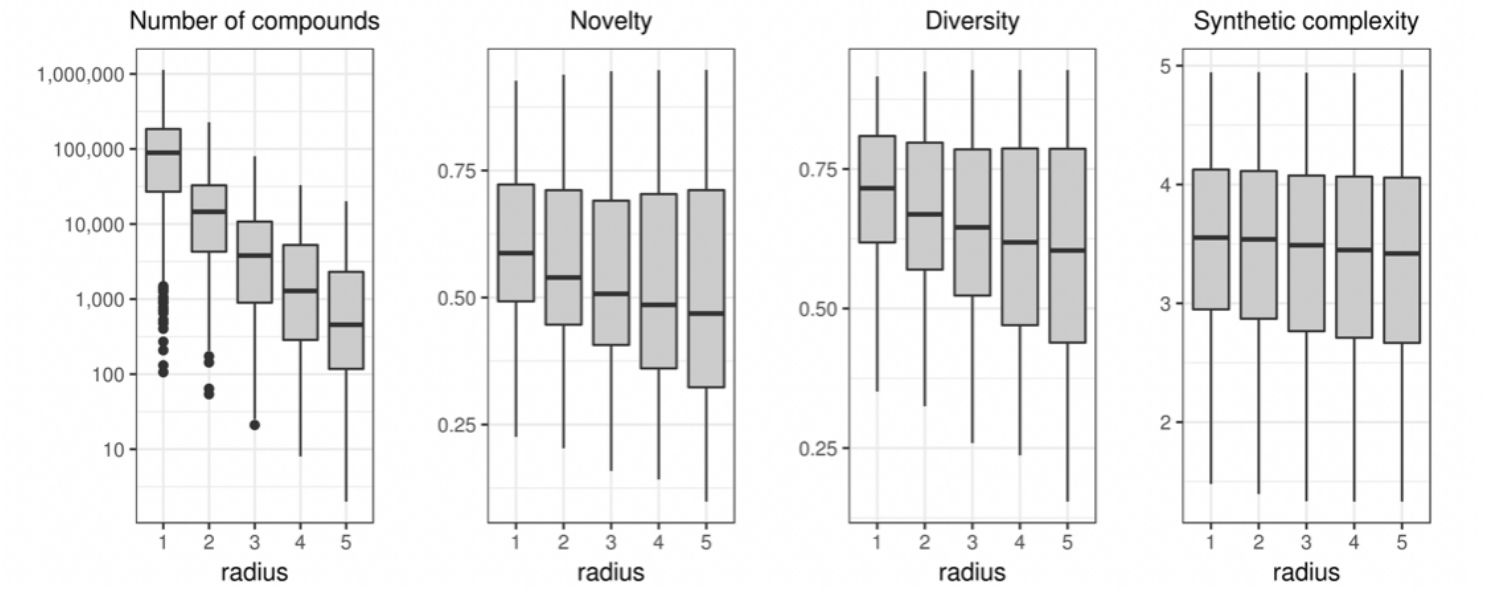

Figure 4

Distributions of the number of generated compounds and average novelty, diversity and synthetic complexity based on 500 compound sets generated from DrugBank compounds using whole ChEMBL fragment database with different context radius.

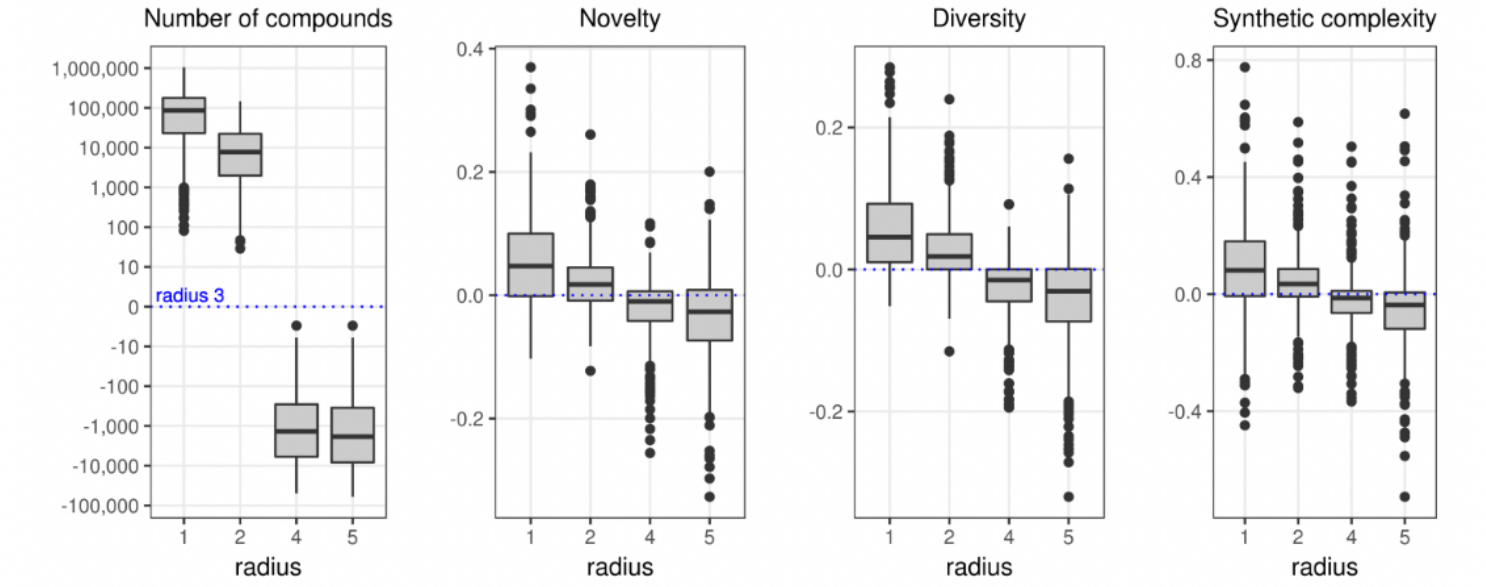

Figure 5

Distribution of differences in the number of compounds and average novelty, diversity and synthetic complexity between data sets generated using context radius of 3 (reference) and others. Positive values demonstrate that parameter values of a data set are greater than for data sets generated at radius 3 and vice versa.

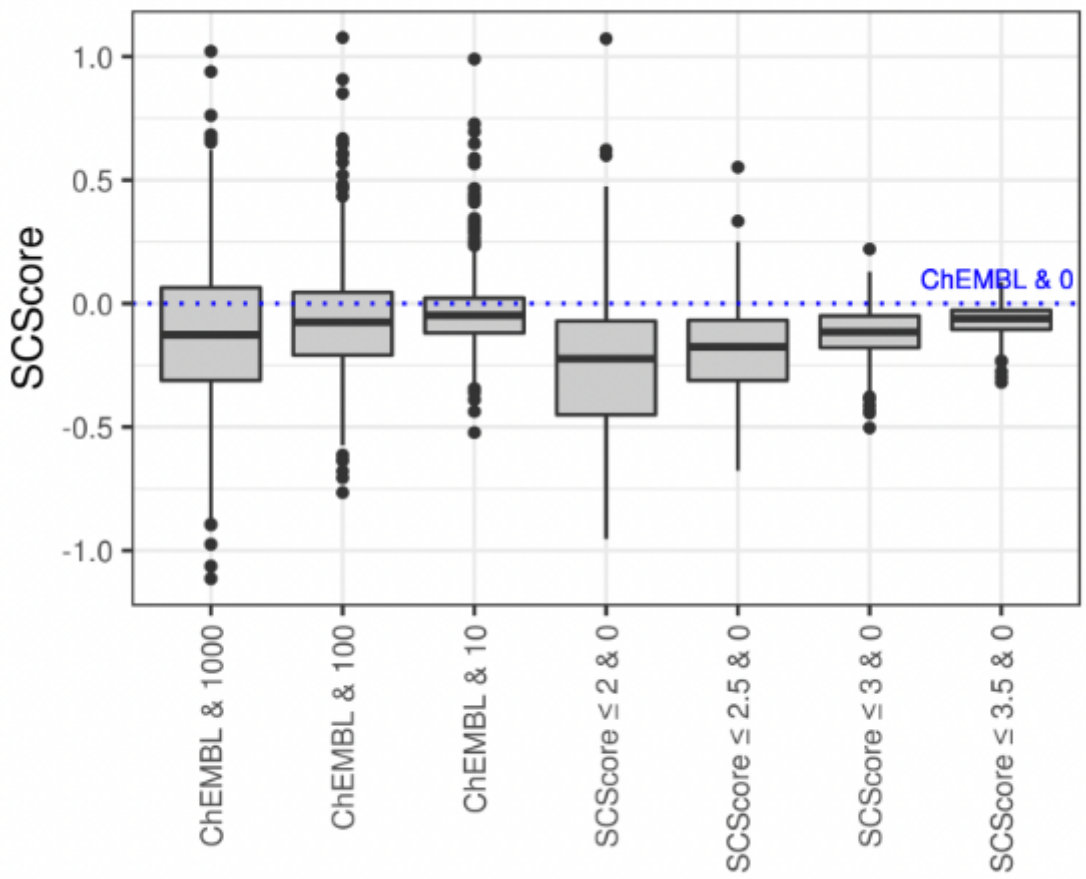

Figure 6

Changes in average synthetic complexity of data sets generated from 500 DrugBank compounds used different restricted strategies relatively to unrestricted generation used the whole ChEMBL fragment database. Replaced fragment occurrence is given after an ampersand symbol.

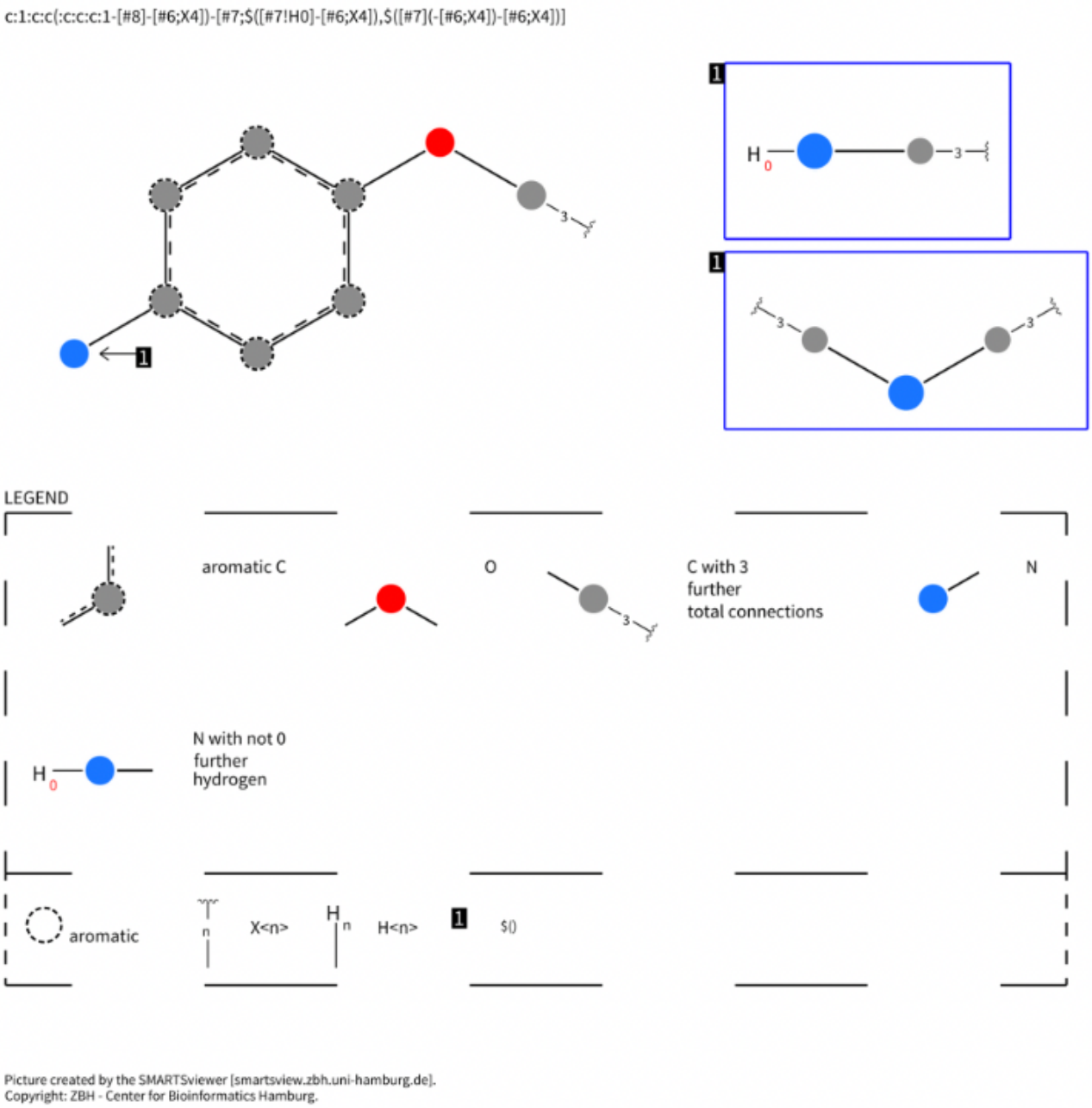

Figure 7

The structure of PAINS anil_di_alk_C(246). The image was produced with SMARTSviewer [39].

Supplementary Files

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

- goaldirectedresults.json
- goaldirectedparams.json

- foundpainspatterns.docx