

## CReM-dock: de novo design of synthetically feasible compounds guided by molecular docking

Guzel Minibaeva, Pavel Polishchuk\*

Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Hněvotínská 1333/5, 779 00 Olomouc, Czech Republic

pavlo.polishchuk@upol.cz

### Abstract

De novo generation of compounds is an attractive strategy allowing to explore much broader chemical space than virtual screening. Fragment-based approaches suffer from low synthetic accessibility of generated compounds. In this study we combined the previously developed fragment-based generator CReM and molecular docking to guide the exploration of chemical space. The developed approach allows to indirectly control synthetic accessibility of generated compounds and their diversity, augmentation of an objective function to generate compounds with more preferable physicochemical properties, control over preserving important protein-ligand interactions and ligand poses. The generated compounds demonstrated high novelty and were competitive to compounds generated by the state-of-the-art approaches. We demonstrated in different case studies flexibility of the developed approach and its applicability to de novo generation as well as fragment expansion tasks. The developed tool is open-source and available at <https://github.com/ci-lab-cz/crem-dock>.

### Scientific contribution

The developed tool, CReM-dock, solves tasks related to de novo design of promising molecules and expansion of co-crystallized ligands within a binding site guided by molecular docking. The key feature is integration of CReM structure generator and EasyDock. The former allows to generate chemically reasonable structures and indirectly control their synthetic feasibility. The latter supports different docking programs. Both provide great flexibility in exploration of chemical space.

**Keywords:** de novo structure generation, fragment-based design, fragment expansion, molecular docking

### Introduction

Since the size of drug-like chemical space is enormous,  $\sim 10^{36}$  compounds[1], it is unfeasible to fully enumerate it or enumerate a representative subset to perform virtual screening. De novo design is a promising strategy to discover new chemical entities in the vast chemical space while enumerating only a small portion of it. This is achieved by the combination of a structure generation tool and an objective function which is optimized in course of the exploration of chemical space and which focuses the generator to the most promising regions [2, 3]. Nowadays, there are multiple approaches to generate molecule structures from scratch [4, 5]. The main limitation of the majority of structure generative approaches is the limited synthetic accessibility of generated molecules[6]. Another restriction can come from an objective function used for searching of promising molecular structures. Application of machine learning models to guide the exploration of chemical space brings constraints related to a limited applicability domain of models. These models cannot reliably predict molecules which are too dissimilar to

training set compounds [7] and therefore restricts the relevant search space for de novo design [8]. Molecular docking has broader applicability and is less biased to particular chemotypes or structural patterns [9-11]. Therefore, docking score can be a good objective function in searching of new promising compounds.

There are many approaches where a structure generation tool was integrated with molecular docking to facilitate searching of new compounds. One group of approaches enumerates structures based on reaction rules. AutoCouple [12] was used to generate 70000 potential CBP bromodomain ligands using three selected reactions. Compounds were docked into multiple protein conformations. 53 top scored compounds were synthesized and several active compounds were identified. Hybridization of structural motifs of those actives resulted in design of a highly active ( $IC_{50} = 35$  nM) and selective (selectivity to BDR4 >10000) compound. Chevillard et al [13] used a reaction-based linking strategy to expand five previously identified fragment-sized ligands of  $\beta$ 2-adrenergic receptor. The goal was to target other binding pockets which were not occupied by the core fragments and, if the binding pose was suitable, link those building blocks to the core molecules by reductive amination reaction. Finally, eight compounds were synthesized and one demonstrated almost 40-fold improvement ( $K_i = 0.53$   $\mu$ M) relatively to the parent core fragment. Other examples of reaction-based tools coupled with docking are NAOMInext [14] and AutoGrow4 [15]. NAOMInext uses iterative growing of starting fragments by covalent coupling with suitable building blocks and evaluated the generated ligands using tethered docking. It was demonstrated that it was possible to reproduce some of previously identified ligands elaborated from smaller molecules co-crystallized with corresponding proteins. AutoGrow4 uses the genetic algorithm to design new compounds. It is based on growing of molecules obtained on a previous iteration using reaction rules and building block libraries and on merging of ligands having a common substructure moiety during the crossover operation. It was demonstrated that AutoGrow4 could generate compounds with docking score outperforming known ligands by a large margin. Although, reaction-based approaches should increase probability of synthetic accessibility of designed molecules, they do not guarantee that. For example, top scored molecules designed by AutoGrow4 are not synthetically feasible [15]. Other limitation is coverage of chemical space which depends on chosen reaction rules and libraries of building blocks.

Fragment-based approaches should result in better coverage of chemical space than reaction-based approaches due to larger fragment libraries and more freedom in fragment linking. However, they may suffer from lower synthetic accessibility of designed molecules. There are two the most common solution of this issue: i) compound filtration based on *post-hoc* evaluation of synthetic feasibility and ii) incorporation of a synthetic bias into the optimized objective function. LigBuilder [16] uses the former strategy. It implements growing or linking fragments using a limited set of pre-compiled small fragments and ring systems. There is no explicit or implicit synthetic bias and generated compounds are filtered post-hoc using the own retrosynthetic evaluation module. Graph-GA approach [17] uses the genetic algorithm on molecular graphs to generate new structures. Since this commonly results in poor synthetic accessibility of generated compounds the authors incorporated synthetic accessibility score [18] as a part of the objective function. This substantially improved the fraction of top scored molecules which were predicted synthetically accessible by the retrosynthetic analysis, from 2-29% to 76-91%. Other fragment-based approaches, like OpenGrowth [19] and FragExplorer [20], incorporated an implicit synthetic bias by applying restrictions on bonds which can be formed during fragment growing. OpenGrowth [19] uses Markov chain model which selects the next attaching fragment using probabilities of fragment linking in existing molecules. Pre-compiled SMARTS rules are applied during fragment growing to make creating connections more chemically reasonable. To further increase synthetic accessibility generated compounds are filtered by the pre-compiled set of 420 000 enumerated anti-patterns which were created by combination of 25 basic fragments and 26 rings from the fragment library and which do not occur in known drug molecules. However, synthetic accessibility of compounds generated by OpenGrowth estimated by Sylvania scores were worse than estimates obtained for drugs and compounds from ChEMBL [19]. FragExplorer [20] applies fragment growing and replacement guided by molecular interaction fields. It uses a precompiled fragment set of 63000 fragments with 0-

1 rotatable bonds, molecular mass within 40-175 Da and clogP < 4. A number of simple filters prevent the formation of undesirable bonds: heteroatom attachment points in fragments are not allowed to be joined to heteroatoms in the query molecule, positively charged nitrogen centers are not allowed to join to sp<sup>2</sup> carbons, aldehyde-aldehyde bonds are forbidden as well as thiol to amide bonds. However, the authors did not explore synthetic accessibility of generated molecules, therefore it is unclear whether those attempts improved synthetic accessibility or not. Post-hoc compound filtration by synthetic accessibility, like in LigBuilder, is not an optimal strategy because it may result in the small or even zero number of suitable compounds remained. Explicit biasing requires the selection of a fast and reliable synthetic accessibility scoring. Implicit biasing implemented in OpenGrowth and FragExplorer is quite simplistic. Checking just a pattern of a created bond ignoring further chemical context may be not enough to substantially increase the number of synthetically feasible compounds.

The third group of approaches uses machine learning generative models. OptiMol [21] includes a variational autoencoder trained on SELFIES to generate new molecules. Compounds generated on each iteration is docked and these results are used to fine-tune the model and thus to guide generation to more promising regions of chemical space. It was demonstrated that OptiMol can improve docking score over iterations, however, the number of novel molecules sharply decreased after 14 iterations. The generated molecules outperformed by docking scores molecules randomly selected from ZINC which were chosen as a baseline. The generated molecules had quantitative estimate of drug-likeness (QED) [22] and synthetic accessibility scores (SA) [18] not too far from ZINC molecules, indicating high quality of generated structures. DockStream is based on the REINVENT generative model, which is a recurrent neural network trained on SMILES [23]. It uses reinforcement learning to generate compounds with favorable docking scores and supports several docking programs. The special diversity filters were added to avoid stuck in local minima and repetitive generation of similar compounds. The objective function was augmented with QED to force generation of more drug-like structures. On several example DockStream could generate compounds with docking scores comparable or better than docking scores of known active molecules, however, synthetic accessibility of generated molecules was not discussed. SBMolGen [24] also uses a recurrent neural network trained on SMILES, but it implements Monte Carlo tree search for exploration of chemical space. To improve synthetic accessibility of final compounds the molecules obtained on each iteration which have SA score greater than 3.5 were assigned the reward score -1, thus, introducing a post-hoc bias in generation workflow. Synthetic accessibility could be also improved due to applied in-house structural filters based on frequency of occurrences of particular patterns in PubChem database, however, they were not disclosed in the paper. If a compound did not pass these filters, it was assigned the reward score -1. In four examples SBMolGen could generate molecules with docking scores better than those for known active molecules [24]. SampleDock [25] utilizes a variational autoencoder model pre-trained on SMILES. It iteratively samples molecules from the latent space using the greedy search by selection as a reference point in the latent space a compound from the previous iteration with the best docking score. SampleDock demonstrated that starting from benzene it could generate molecules with better scoring to CDK2 and SAR-CoV2 M<sup>Pro</sup> proteins than known actives. At the same time generated molecules shared substructures common with known actives and may have reduced novelty.

While fragment-based and machine learning-based approaches should provide better coverage of chemical space than reaction-based approaches, they lack synthetic feasibility of generated compounds. Biasing of the de novo generation toward more synthetically accessible compounds can be explicit or implicit. Explicit biasing is incorporation of synthetic accessibility score to an objective function. This requires choosing of an appropriate synthetic accessibility scoring, transformation/scaling of individual components of a multi-component objective function and its functional type, e.g., arithmetic or geometric mean. Implicit biasing may be simpler to implement, however, currently implemented approaches use too simplistic implicit biasing, which do not improve synthetic accessibility of generated compounds substantially and do not provide a fine-tuning control over it.

In this work we developed and implemented a de novo generation pipeline based on fragment-based generative approach CREM [26] and molecular docking. CREM provides indirect control over synthetic accessibility of generated molecules [27]. We integrated EasyDock module [28] in the pipeline to support different docking tools and enable customization with further docking programs if necessary. The pipeline implements the growing protocol which starts from a known X-ray structure of a protein-ligand complex or from a set of diverse starting fragments docked to a binding site which are further expanded by attaching new fragments in a chemically reasonable way. A user may specify whether it is necessary to keep the pose of a starting structure and/or keep specific protein-ligand contacts which are considered important for ligand binding. The developed pipeline is suitable for fragment expansion, scaffold decoration and de novo design.

## Methods

Similar to other conventional methodologies, our approach employs an iterative strategy for the exploration of chemical space by fragment growing consisting of several steps: i) docking of starting fragments/molecules, ii) selection of compound with the promising scores and satisfying a user-defined protein-ligand interaction fingerprint (the latter is optional), iii) growing of the selected molecules and iv) filtering molecules by physicochemical properties. All steps are repeated until physicochemical properties of compounds do not reach one of the user-defined thresholds.

There are two modes of structure generation: de novo design and fragment/molecule expansion. The first mode requires a set of starting fragments submitted as SMILES or 2D structures. The second mode takes 3D structures as input which correspond to actually observed or predicted binding poses. The difference between these modes is that 3D input molecules are passed directly to the growing step omitting docking and selection steps.

### *Docking*

Docking of molecules is performed using EasyDock [28] which currently supports Autodock Vina and Gnina (Smina scoring functions is supported through Gnina). EasyDock provides a simple interface to dock molecules and to integrate docking in external tools. It takes as input a protein structure and a configuration file where all required settings are specified (e.g. grid box coordinates and its size, search exhaustiveness, etc). We utilized the same database structure as in EasyDock to store all outputs of design and docking steps and all properties of generated compounds.

### *Molecule selection strategies*

We implemented three strategies to select molecules on each iteration: greedy, Pareto and clustering-based selection. Within the first strategy top N molecules with the highest docking scores are selected. While this may lead to highly scoring compounds their diversity may be lowered, therefore, we implemented two other protocols. In one protocol, Pareto ranking is applied to select molecules on the Pareto front with low molecular mass and high docking scores, promoting the growth of promising low molecular mass molecules. In the alternative strategy, molecules are clustered by K-means approach to the user-defined number of clusters and a specified number of top scored molecules are selected from each cluster. If some of top scored molecules in a cluster cannot be grown due to any reason (no hydrogens to replace, physicochemical properties reached the given thresholds, etc), a next molecule is selected until the specified number of molecules will be chosen from a cluster. Clustering protocol gives a more predictable number of selected compounds and runtime, while Pareto may result in a variable number of generated molecules and higher diversity of final solutions.

Optionally a user may specify additional criteria applying before actual selection. An RMSD threshold can be specified to keep the pose of a parent compound in successor compounds. If docking poses of generated compounds differ greater than the given threshold from their parent molecule, these compounds will not be considered for the selection. RMSD value is calculated between a parent and a successor molecule for heavy atoms of a maximum common substructure. This can be particularly useful in fragment expansion studies where the pose of the starting fragment is experimentally defined and it is expected that successor molecules should keep it. Another option is to specify a list of important protein-ligand contacts and a minimum similarity according to ligand-protein interaction fingerprints (PLIF). The fingerprints are calculated by ProLIF [29] and define H-bond donor/acceptors, hydrophobic, aromatic, positively or negatively charged contacts, metal centers. If a molecule does not meet the threshold, it will not be considered for selection. For example, if three contacts are specified and a threshold was set to 0.5, it means that at least two of these contacts should be observed in molecules (similarity 0.66), otherwise they will not be eligible to participate in the selection step. This is useful if important contacts are known *a priori* and the designed molecules should preserve them.

### *Molecule growing*

Chosen molecules undergo growing which is performed using CReM [26]. To replace hydrogens with larger fragments CReM uses a database of interchangeable fragments (CReM databases). These are fragments obtained from existing molecules by exhaustive fragmentation cutting up to four single bonds. For each fragment an environment is determined, which is a substructure comprising atoms with the distance of up to a given number of bonds (context radius) from attachment points of a fragment. Thus, fragments occurred in the same chemical context should be interchangeable and their replacement should result in synthetically feasible molecules. It was demonstrated previously that synthetic feasibility can be controlled indirectly and is improved by choosing a larger context radius and CReM databases composed from fragments of more synthetically accessible molecules [27].

To perform growing we replace only those hydrogen atoms which are at least at the distance of 2Å apart of any protein heavy atom. This will avoid growing in directions which do not have enough space to accommodate larger fragments. To control the number of generated compounds and make runtime more predictable one may specify the maximum number of randomly chosen replacements. If not specify, all possible replacements will be applied which can be very numerous for smaller context radiuses. We found that using 2000 random replacements works well and we used this number in all reported studies. Additionally, the size of steps in chemical space can be specified. By default, we attach fragments having up to 10 heavy atoms. Since CReM cannot create new cycle systems, it is required to choose this size large enough to enable rings addition to molecules.

By default, fragment selection in CReM employs a uniform distribution. However, to apply selective pressure and prioritize fragments with more desirable properties, the tool provides an option to customize the selection process. In one of the studies presented herein, fragment selection was weighted proportional to the squared fraction of Csp<sup>3</sup> atoms within the fragments. This should preferentially select fragments containing a higher proportion of saturated carbon atoms, thereby enriching the generated molecules with sp<sup>3</sup> carbon atoms in their scaffolds.

We implemented control over important physicochemical properties determining drug-likeness (molecular weight (MW), topological surface area (TPSA), lipophilicity (logP) and the number of rotatable bonds (RTB)) to restrict generation to mainly drug-like molecules. All of these parameters, except lipophilicity, increases or stay the same with an increasing number of atoms in a molecule and are mainly additive. If one of these parameters becomes equal or greater than a pre-defined threshold a molecule is discarded from further consideration. Due to the mainly additive nature of these properties, we were able to pre-filter CReM fragments on-the-fly and choose those ones for growing which unlikely will result in molecules exceeding pre-defined thresholds of physicochemical properties. This allows avoiding enumeration and docking of compounds with undesirable properties.

In all studies in the current work the following restrictions were applied to physicochemical properties of generated compounds: molecular mass  $\leq 450$  Da, the number of rotatable bonds  $\leq 5$ , lipophilicity  $\leq 4$ , topological polar surface area  $\leq 120\text{\AA}^2$ . These criteria satisfy Lipinski rule and keep some capacity for compound improvement.

After the generation step a stable tautomer may be generated by Chemaxon. This is an optional step, because sometimes the predicted stable tautomer may differ from the starting one, which can be known from an experiment, and this may break the whole generation pipeline resulting in wrong docking poses and scores and affect molecule selection on each step.

### *Implemented objective functions*

By default, molecules are ranked by docking score. However, the docking score can be augmented with further important properties to make generated molecules more balanced and closer to a desired property space. In particular, we augmented docking scores with quantitative estimate of drug-likeness (QED) [22]. First, we map docking scores to the range from 0 to 1 using formula  $s = (x - x_{min}) / (x_{max} - x_{min})$ , where  $x$  is a docking score of a molecule,  $x_{min}$  and  $x_{max}$  are minimum and maximum docking scores among compounds generated on a particular iteration and eligible for selection. Afterwards we multiply scaled docking scores and QED values for corresponding molecules to get a final score.

Another augmentation implemented was the calculation of the fraction of  $sp^3$  carbon atoms in Bemis-Murcko scaffolds (Csp<sup>3</sup>BM). The Csp<sup>3</sup>BM values ranging from 0 to 0.3 were linearly scaled to fit a range of 0 to 1, while values exceeding 0.3 were set to 1. The scaled Csp<sup>3</sup>BM values were subsequently squared to enhance selection pressure on this parameter and then multiplied by the docking score, which had been normalized to the range of 0 to 1 as described previously. The threshold of 0.3 was selected based on recommendations from the authors of the first CACHE challenge [30] in which we participated and possessed one of the top places [31].

There were also implemented other objective functions, for example based on docking efficiency (docking score divided on the number of heavy atoms), but we did not apply them in studies herein.

### *Protein preparation protocol*

To perform de novo generation, we prepared receptor structures using the Dock Prep protocol implemented in Chimera [32]. The preparation involved remodeling missed side chains and sequences utilizing the Dunbrack rotamer library [33] and MODELLER [34], respectively. Hydrogen atoms were added, considering pH of 7.4, and solvent molecules were removed. The structures were then converted to the PDBQT format using the *prepare\_receptor4.py* utility from AutoDock Tools. Grid boxes for docking were determined based on coordinates of native ligands. Specifically, the center of each grid box was calculated as the geometric center of a ligand, and the box size was set by adding 7 Å to the minimum and maximum coordinates of the ligand's heavy atoms. All prepared structures and grid box parameters were deposited to the repository - <https://github.com/ci-lab-cz/docking-files>.

### *Novelty assessment*

To assess the novelty of generated compounds calculated Tanimoto similarity to the closest compound from the set of known ones taken from ChEMBL (version 33). To calculate similarity, we used chemfp tool [35] and 2048-bit Morgan fingerprints of radius 2. The smaller the similarity, the greater the novelty of generated compounds. We estimated baseline similarity level using randomly selected pairs of compounds from ChEMBL to be used as reference. We randomly chose 10 000 compounds with MW  $\leq 500$  and calculated pairwise Tanimoto similarity. The mean similarity was 0.105, 95%-percentile – 0.178 and 99%-percentile – 0.230.



## UMAP and reference space

To analyze distribution of generated compounds relatively to reference ones we chose 100 000 random compounds with MW  $\leq$  500 from ChEMBL33 as a baseline reference set and separately we collected sets of actives for every individual target. Actives were selected from compounds tested in a single protein assay format and demonstrated pIC<sub>50</sub>, pKi or pK<sub>d</sub> equal or greater than 6. For visualization of chemical space we used UMAP [36] (umap-learn Python package) and 2048-bit Morgan fingerprints of radius 2. All parameters were set to default with the exception of the number of neighbors = 10 and metrics = "jaccard".

## Results

### Preparation of CReM fragment databases

For preparation of CReM fragment databases we used structures from ChEMBL22 [37]. Structures were curated according to the protocol based on Chemaxon Standardizer [38]: i) salts were removed and molecules were neutralized, ii) chemotypes were standardized, iii) duplicates were removed. We kept only compounds containing the following atoms: C, N, O, S and halogens. The collected initial dataset consisted of 1 554 260 structures. Further a subset of molecules was reduced by removing molecules matching at least one of structural alert from the set of BMS, Dundee, Glaxo, Inpharmatica and PAINS filters as implemented by Pat Walters - [https://github.com/PatWalters/rd\\_filters](https://github.com/PatWalters/rd_filters). As we demonstrated previously, removal of such molecules before fragment database creation guarantees generation of molecules having no such patterns if the size of patterns does not exceed the chosen context radius [26]. This reduced the data set size to 818 174 molecules. Further we defined subsets of molecules with restricted synthetic accessibility (SA) values as predicted by the approach of Ertl and Schuffenhauer [18]. We chose values 2 and 2.5 as reasonable thresholds while the average SA score for all ChEMBL22 compound was 3.0 and the median score was 2.73. This gave subsets with 67 970 and 338 422 molecules, respectively. These sets of molecules were exhaustively fragmented and converted to CReM fragment databases (Table 1).

Table 1. The number of fragmented molecules, fragments with maximum number of 10 heavy atoms and corresponding fragment-context pairs in created CReM databases.

CReM DB	n (fragmented molecules)	n (distinct fragments)	number of distinct fragment/context pairs for each radius				
			radius 1	radius 2	radius 3	radius 4	radius 5
ChEMBL	818 174	988 585	2 263 436	4 051 790	7 133 534	11 007 247	15 271 543
ChEMBL SA2.5 (SA $\leq$ 2.5)	338 422	272 988	671 140	1 263 268	2 319 377	3 752 375	5 419 544
ChEMBL SA2 (SA $\leq$ 2)	67 970	55 498	143 434	267 156	472 126	754 905	1 087 492

### Preparation of a starting fragment library

To prepare starting fragments we exhaustively fragmented 67 970 ChEMBL molecules having SA score less than 2. All attachment points were capped with hydrogens, resulted molecules were converted to canonical SMILES and duplicates were removed. The resulting 200 000 molecules we filtered according to their physicochemical properties:

- the number of heavy atoms is within the range 8-15

- the number of distinct H-bond donors and acceptors should be within the range 1-5. If an atom is labeled as an H-bond donor and an acceptor it was counted only once. This gives an estimate on the number of specific contacts.
- the number of rings is 1-3
- the number of fused ring systems 0-2
- the number of rotatable bonds is 0-2
- lipophilicity is less than 2
- topological polar surface area (TPSA) is greater than  $25A^2$
- the total number of halogen atoms (Cl, Br and I) is 0-1
- the maximum size of rings is 7

This was resulted in 20 164 molecules. For them we enumerated all stereoisomers with an RDKit script from the repository <https://github.com/DrrDom/rdkit-scripts> and tautomers using cxcalc Chemaxon utility [39]. Duplicates were checked and removed. Finally, we got 23 840 molecules which were used as starting fragments in de novo generation.

#### *Theoretical size of covered chemical space*

To estimate the number of molecules that can be generated using the CReM methodology from a given set of fragments, we considered a scenario in which four substituents are attached to each selected starting fragment simultaneously. Substituents replaced hydrogens were chosen from a CReM database, taking into account their chemical context of a specified radius, while ensuring that the total number of heavy atoms in resulting molecules did not exceed 36. This limit corresponds to a molecular weight of approximately 500, as demonstrated in our previous study [1].

The chemical environment for each non-equivalent hydrogen atom in an initial molecular fragment was identified using CReM. All possible combinations of four hydrogen atoms were analyzed. Restricting the analysis to non-equivalent hydrogens introduces an underestimation of the number of possible derivatives, as it excludes cases where substituents are attached to the same methyl group, for instance. This simplification arises from constraints within the current CReM implementation.

For each combination of four hydrogens, the total number of potentially enumerated compounds was computed as the product of the number of substituents available at each hydrogen position, under the constraint of a maximum total number of heavy atoms in generated molecules. The total number of molecules was determined by summing the number of enumerated compounds across all combinations of four substituents. To enhance computational efficiency, calculations were performed on a subset of 1000 randomly selected molecules. The final value was extrapolated by multiplying the result by a scaling factor of 23.84 (23840 starting fragments were in total). This provided an estimate under the assumption that fragments would only be attached to the initial starting fragment.

However, substituents can be attached not only to the starting fragment but also to previously introduced substituents. To account for this, the structure was considered as a tree where the five fragments (the starting fragment and four substituents) represent nodes. The potential number of combinations of connections between nodes was estimated using Cayley's formula,  $n^{n-2}$ , which predicts that five nodes can be connected in 125 distinct ways. This approach constitutes an additional simplification, leading to an overestimation of the number of derivatives, as not all linkage combinations are feasible due to chemical context constraints. These approximations



may compensate each other to obtain a rough estimate of the order of magnitude for the size of the covered chemical space.

A radius of 3 was chosen as it represents a sufficiently large default value to yield synthetically feasible molecules. For the largest fragment database, the estimated chemical space coverage was  $10^{17}$  (Table 2). This coverage decreased when using more restricted fragment databases:  $10^{16}$  compounds for SA2.5 and  $10^{15}$  for SA2. For the smallest fragment database (SA2), the chemical space size was also computed for radii of 2 and 4, yielding predictable changes. The coverage increased to  $10^{16}$  compounds for radius 2 and decreased to  $10^{13}$  for radius 4. These findings indicate that even under highly restricted conditions, the covered chemical space remains substantial.

Table 2. The estimated size of covered chemical space by starting fragment decorated with fragments from CReM databases.

CReM DB	radius	estimated size of covered chemical space
ChEMBL	3	$2.8 \times 10^{17}$
ChEMBL SA2.5	3	$4.2 \times 10^{16}$
ChEMBL SA2	3	$1.8 \times 10^{15}$
ChEMBL SA2	2	$8.4 \times 10^{16}$
ChEMBL SA2	4	$2.7 \times 10^{13}$

### De novo design of CDK2 inhibitors

To investigate influence of different settings on generation output we chose CDK2 kinase because it is a clinically relevant target, it has multiple X-ray protein-ligand complexes and it is frequently used in validation of modeling approaches. We chose CDK2 structure (PDB 2BTR) in complex with the inhibitor PNU-198873 ( $K_i = 95$  nM). This ligand forms an H-bond donor and an H-bond acceptor bonds with Leu83 residue from the hinge region. Since interaction with the hinge region is important for competitive kinase inhibitors, we set these two contacts as obligatory for all designed molecules. The search algorithm was clustering with 25 clusters and top two molecules were selected from each cluster. Thus, up to 50 molecules were selected for growing on each iteration. Every compound was grown to get up to 2000 new molecules. If the number of possible expansions was greater than 2000, random 2000 expansions were selected from the CReM database. We run generations for all three CReM databases and all five context radiuses. Every simulation was run three times to estimate robustness of the search, because it can be affected by stochasticity in choosing of growing fragments. For docking we set the same seed, therefore docking results were deterministic.

The total number of generated compounds was highly reproducible across runs and predictably decreased with choosing more restricted fragment databases and greater radiuses (Figure 1). The number of compounds which bound to the hinge region was 13-20% from the total number of generated compounds. From 23840 starting fragments only 1471 (6%) bound to the hinge region.

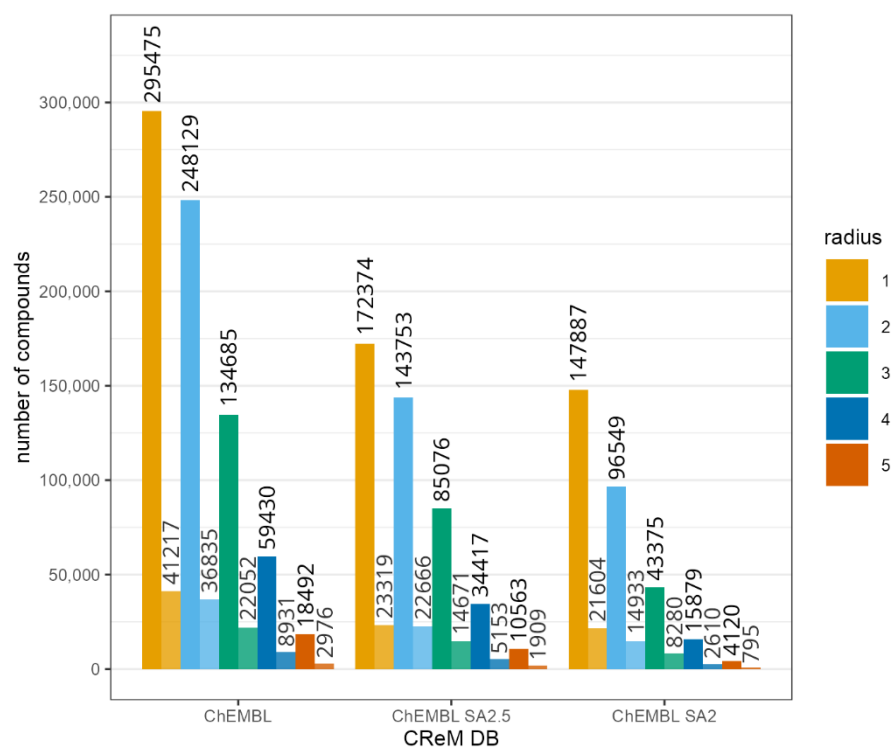


Figure 1. Average number of generated compounds for different generation settings (fragment CReM databases and context radii) across three runs. Darker color denotes the total number of generated compounds, lighter colors – the total number of generated compounds which satisfy the required ligand-protein interactions (H-bond donor and acceptor with Leu83).

### *Docking and synthetic accessibility scores of generated compounds*

For further analysis from each run we chose top 100 compounds, which satisfied the required protein-ligand interactions (hinge region binding). For those molecules average docking and SA scores were calculated. As expected, a clear trade-off between SA and docking scores was observed (Figure 2a). Generations for all three CReM databases and radii 1 and 2 achieved similar docking scores between -13 and -12.3, while SA scores varied in a large range from 4.15 for a full fragment database to 2.9 for ChEMBL SA2 database. Further increase of the context radius improved SA scores less pronounced. The minimum SA value 2.37 was achieved for ChEMBL SA2 database and radius 5. However, this radius increase resulted in worsening of average docking scores to almost -10.5. A clear dependence between chosen fragment databases and SA scores was observed (Figure 2b). The variance across runs was small. Thus, the choice of a fragment database and a radius predictably changes SA scores of generated compounds. This creates another feature of CReM approach fine-tune control over synthetic accessibility of generated structures. The only outlier was the run for the full CReM database and radius 4, which resulted in the average SA score around 4, while two other runs with the same settings gave average SA scores below 3.

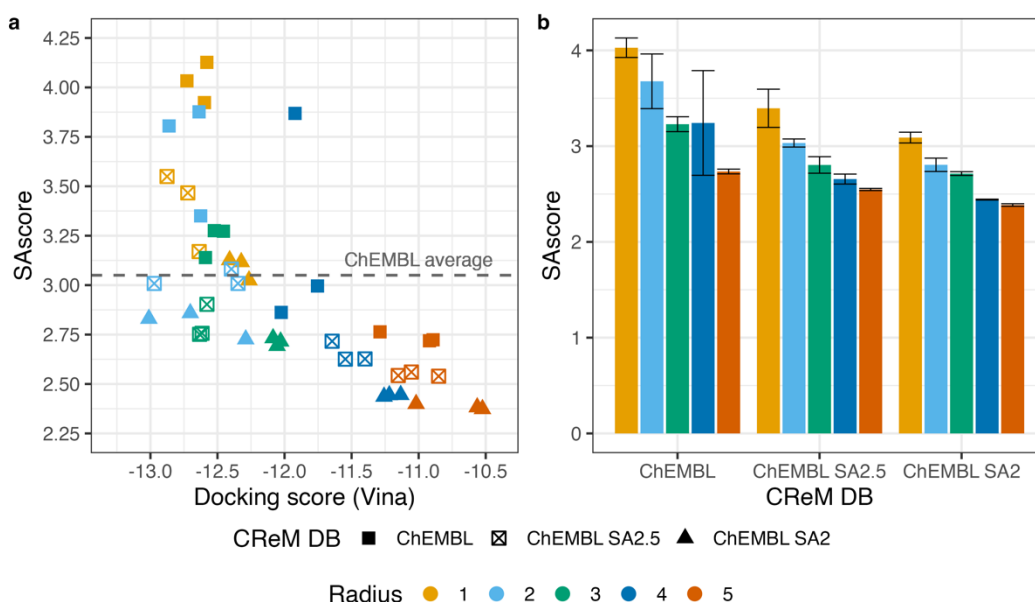


Figure 2. Statistics of top 100 designed compounds bound to the hinge region of CDK2. (a) Average docking and SA scores for top 100 compounds. (b) Average SA scores and standard deviation across three runs for top 100 compounds.

#### *Scaffold diversity and reproducibility of runs*

We analyzed reproducibility of molecular structures in independent runs. The number of identical molecules among top 100 compounds bound to the hinge region was the lowest for radius 1 and 2 (0-14%). For larger radii the number of identical compounds was substantially increased to 27-83% (Figure 3a). Therefore, repetitive runs may be less reasonable for generations with radius 3 and greater.

Analysis of diversity of top 100 generated compounds was performed based on Murcko scaffolds. The average number of distinct Murcko scaffolds was small (12-33) for smaller context radii from 1 to 3 (Figure 3b). This indicates that in these conditions some scaffolds may result in multiple successful successors which outperform others. For a larger context radius, 4 and 5, the average number of distinct scaffolds was much larger (from 29 to 64 out of maximum possible 100). The effect of larger radii was more pronounced for more restricted fragment databases. This can be explained by the limited number of expansions of each molecule on each iteration and therefore the smaller number of successive compounds are generated for each scaffold decreasing probability that a single scaffold will be overrepresented and outperform the others. Despite higher diversity of scaffolds for context radii of 4 and 5, these scaffolds were frequently reproducible even across different fragment databases (Figure S2).

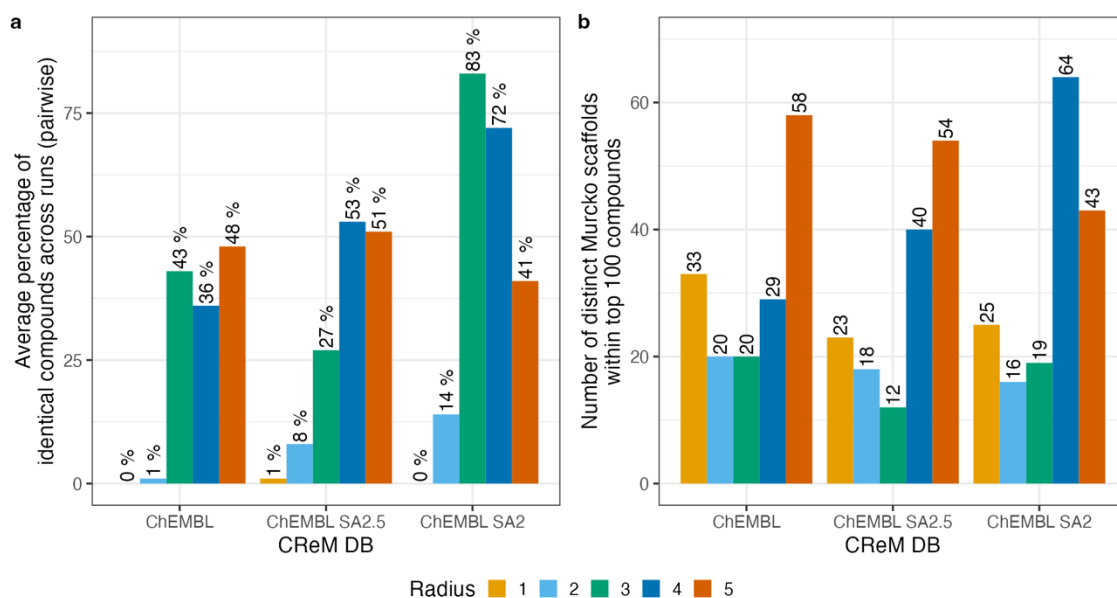


Figure 3. Statistics for top 100 compounds bound to the hinge region of CDK2. (a) Average percentage of identical structures among top 100 compounds between pairs of three independent runs for every set of generation settings. (b) The average number of distinct Murcko scaffolds among top 100 compounds across three independent runs.

#### Novelty of designed compounds

To assess novelty of designed compounds we calculated Tanimoto similarity using 2048-bit Morgan fingerprints of radius 2 to all compounds from ChEMBL33 (2.37 million, Figure 4) and to the subset of ChEMBL33 which demonstrated activity to CDK2 ( $pIC_{50}$ ,  $pK_i$  or  $pK_d \geq 6$ , 1001 compounds). The majority of top scored compounds has maximum similarity below 0.5 to any compound from ChEMBL and below 0.3 to any known CDK2 inhibitors. This confirms that generated compounds are structurally different from previously explored chemical space.

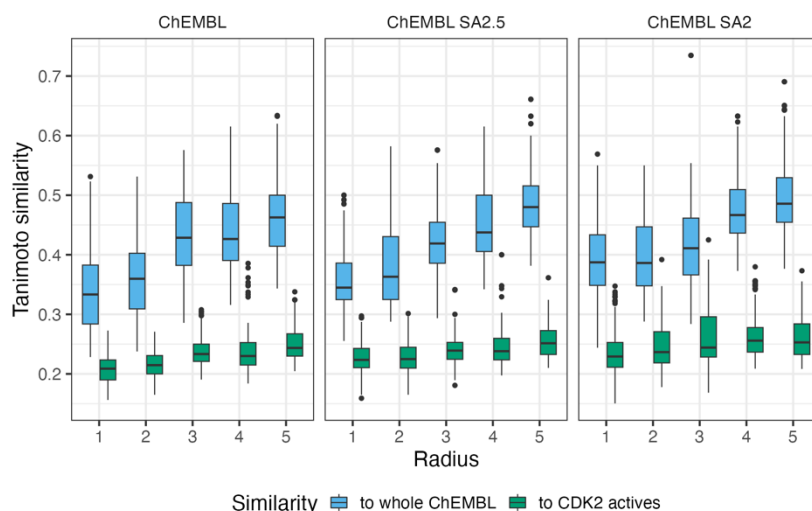


Figure 4. Tanimoto similarity (2048-bit Morgan fingerprints with radius 2) of top 100 compounds, which bind to the hinge region of CDK2, generated in individual runs for particular settings to the most similar compounds from the whole ChEMBL33 database and to the most similar known CDK2 inhibitors ( $pK_i/pK_d/pIC_{50} \geq 6$ ).

### Optimal de novo generation settings

From the experiments and the analysis performed above it is obvious that there is a Pareto front of possible solutions with regards to docking and SA scores. It is not reasonable to use the small radius (1-2) in combination with the full fragment database or SA2.5 database for compound generation. The generated compounds have high docking scores but they are much more synthetically complex (Figure 2a). Using larger radiuses (4-5) result in compounds which may be better synthetically accessible, but with poorer docking scores. The nearly optimal settings may be combinations of a fragment database SA2 and radius 2 or a fragment database SA2.5 and radius 3, which result in the highest docking scores (-13 – -12.3) and a reasonably good SA scores (~2.75). The combination of a fragment database SA2 and radius 2 gives a greater number of distinct compounds and scaffolds within top 100 molecules relatively to the fragment database SA2.5 and radius 3 (Figure 3, Figure 4). Therefore, we supposed the former settings as an optimal one and used them in all further experiments.

### Comparison with docking of random ZINC compounds

To evaluate the performance of identification compounds with high docking scores we performed docking of a random subset of molecules from ZINC [40]. We selected compounds which satisfied the same physicochemical criteria as we used for de novo structure generation: molecular mass  $\leq 450$  Da, the number of rotatable bonds  $\leq 5$ , lipophilicity  $\leq 4$ , topological polar surface area  $\leq 120\text{\AA}^2$ . The number of selected compounds (120 000) was approximately equal to the average number of compounds docked during the de novo generation using ChEMBL SA2 fragment database and radius 2 (96 549 generated molecules + 23 840 starting fragments). Only 617 ZINC compounds (0.51%) could establish the required contacts with the hinge region, that is much lower than among de novo generated compounds (13-20%). Docking scores of top scored compounds were also worse than for de novo generated molecules irrespective the ability of compounds to bind to the hinge region (Figure 5).

This shows, that de novo generation can achieve better docking scores than conventional virtual screening using comparable computational resources. De novo generation with explicit biasing towards preferable interactions also outperformed conventional docking in the number of identified compounds satisfying a pre-defined protein-ligand interaction pattern.

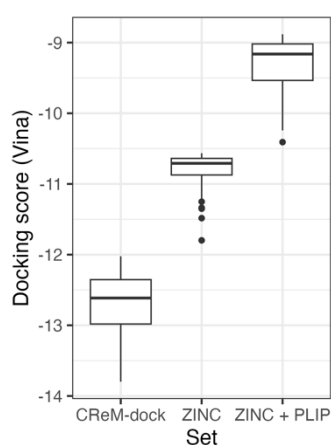


Figure 5. Distribution of docking scores of top 100 de novo generated compounds bound to the hinge region and top 100 compounds from a random subset of ZINC that do not take binding to the hinge region into account and those which take it into account.

### *Application of different selection strategies*

We evaluated implemented selection strategies (greedy, clustering and Pareto) in three independent runs using the settings determined above as optimal ones. Within the greedy strategy we selected top 50 compounds with the highest docking scores. For the clustering strategy we varied the number of clusters and the number of selected compounds from each cluster in that way that the total number of compounds selected on each iteration was 50. In the case of the Pareto strategy, it was impossible to control the number of compounds selected for growing on each iteration.

As previously we analyzed top 100 compounds from each run with best docking scores and which bind to the hinge region (Figure 6a). The greedy selection resulted in the highest docking scores while SA scores were comparable to outputs of the clustering strategy. The Pareto selection resulted in docking scores comparable to the clustering approach, but synthetic complexity of generated compounds was somewhat greater.

As expected, application of the greedy selection strategy gave highly reproducible molecular structures in independent runs which were also characterized by low scaffold diversity (Figure 6b). The clustering selection resulted in moderate diversity of scaffolds with low reproducibility across runs (Figure 6b). At the same time more than a half of scaffolds generated within greedy strategy were reproduced by the clustering approach. The Pareto selection resulted in the highest scaffold diversity with low reproducibility across independent runs, similarly to clustering (Figure 6b).

We specifically studied the effect of clustering settings on generated molecules. The results showed that decreasing the number of clusters led to improvement of docking scores of top 100 compounds, but they were synthetically more complex. Increasing the number of clusters and simultaneous decreasing of the number of selected compounds from each cluster results in the opposite trend (Figure 6c). We hypothesize that this could be a result of distribution of molecules with high docking scores in individual clusters. There may be a situation that there are only few clusters comprising highly scoring molecules and top scoring molecules in other clusters have moderate docking scores. Then the greater number of clusters will lead to the greater number of compounds with moderate scores that may decrease performance to some extent. However, improvement of synthetic accessibility may compensate this drop in docking scores. It should be also noted that choosing a greater number of clusters resulted in increased diversity of Murcko scaffolds among top 100 compounds. The average number of distinct scaffolds was 8, 16 and 46 for 5, 25 and 50 cluster setups, respectively.

The percentage of generated compounds, which bind to the hinge region, was the highest for the Pareto strategy (23-28%) followed by the greedy approach (20-21%) and clustering (15-22%). The number of generated and docked molecules was also varied a lot across different strategies. The lowest number for compounds was generated for the clustering strategy irrespective the chosen number of clusters (90 000-99 000). For the greedy strategy the number of generated molecules was 103 000-106 000. Whereas for the Pareto strategy it was much greater – 175 000-201 000 compounds (Table S1).

While Pareto suggests the highest diversity of scaffolds of generated compounds its runtime is less predictable than for other strategies because a variable number of compounds is selected on each iteration and the number of iterations can be greater. The greedy search results in highly reproducible outputs with molecules having high docking scores. Therefore, it is not necessary to run it multiple generation. However, the diversity of generated molecules is relatively low. The clustering strategy suggests a balanced approach, which has a predictable runtime, and a user may increase diversity of generated molecules by increasing the number of clusters.



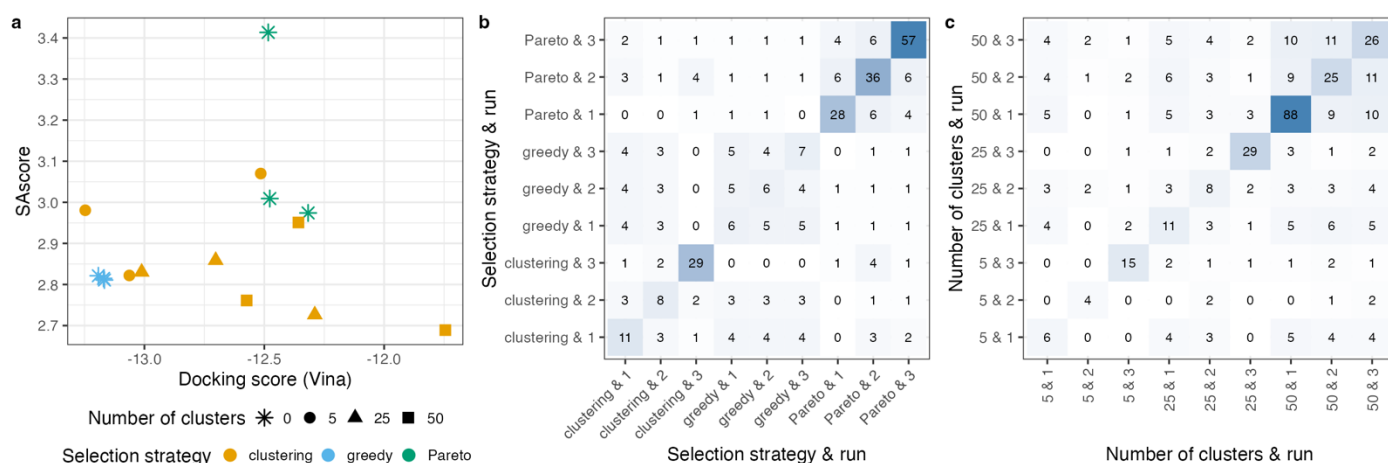


Figure 6. Statistics for top 100 molecules bound to the hinge region from three independent runs using different selection strategy (greedy, clustering and Pareto) and cluster settings (all runs used ChEMBL SA2 fragment database and radius 2). (a) Average docking and SA scores for top 100 molecules. (b) The number of distinct Murcko scaffolds among top 100 compounds for different selection strategies. (c) The number of distinct Murcko scaffolds among top 100 compounds for different clustering settings.

### Augmentation of a docking scoring function

The objective function, which is a docking score by default, can be augmented with additional parameters important for particular projects. Augmentation of docking scores with drug-likeness (QED) substantially improves drug-likeness of generated compounds. However, docking scores of top 100 molecules were somewhat worse relatively to the runs based exclusively on docking scores. Synthetic complexity of generated compounds was comparable to those generated with docking score alone but more variable (Figure 7).

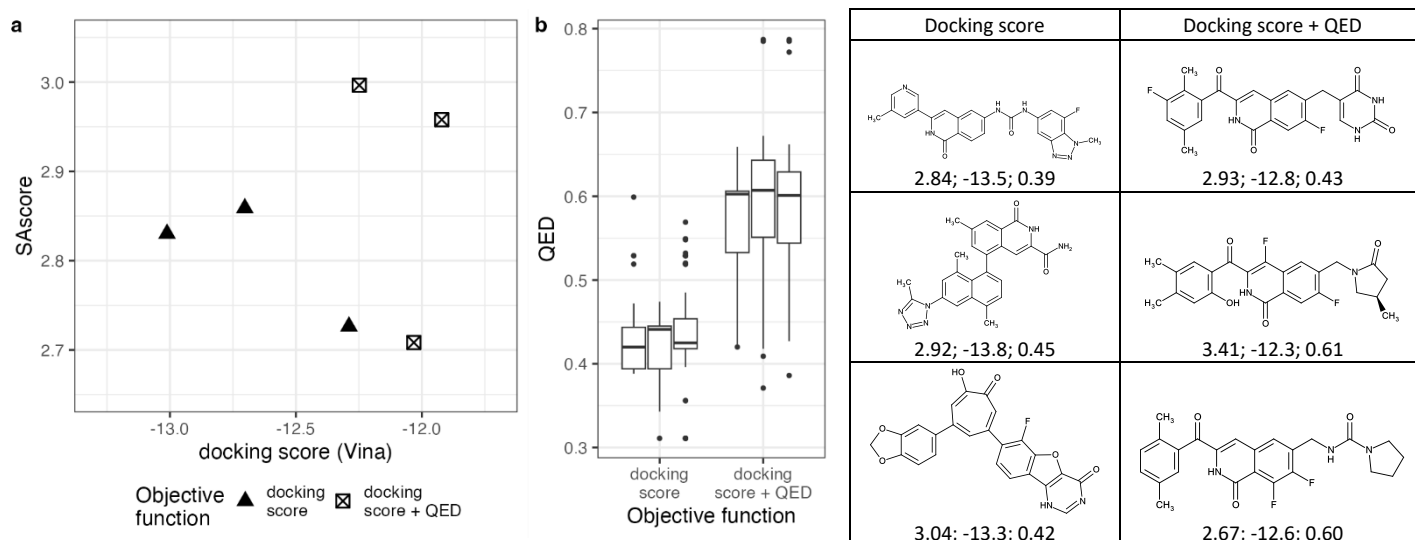


Figure 7. Statistics for top 100 molecules bound to the hinge region from three independent runs using docking score as an objective function and docking score augmented with drug-likeness (QED) (all runs used ChEMBL SA2 fragment database, radius 2 and the clustering selection strategy with 25 clusters and top 2 compounds selected from each). (a) Average docking and SA scores for top 100 molecules. (b) Distribution of drug-likeness for top 100 compounds for individual runs. Top scored structures from individual runs are shown on the left. The numbers are SA score, docking score and QED, respectively.

Previous studies have demonstrated that molecules with a higher degree of saturation are more likely to advance through various clinical development stages [41]. In this work, several strategies were employed to bias molecular generation towards compounds with at least 30%  $sp^3$  carbon atoms in their scaffolds. Additionally, the CReM fragment databases were pre-filtered to exclude fragments containing rings larger than six atoms. Retaining these larger rings in databases significantly increased the frequency of such ring systems in generated molecules, because they disproportionately contributed to the  $sp^3$  carbon fraction.

All simulations were performed in triplicate, and the statistical results were averaged for simplicity. The first strategy involved augmenting the docking score with the fraction of  $sp^3$  carbon atoms in Bemis-Murcko scaffolds (Csp<sup>3</sup>BM). This approach yielded only a marginal improvement in the proportion of generated molecules meeting the desired criteria (Figure 8a). A more effective strategy involved employing a custom sampling function that selected fragments from the CReM database in proportion to the squared fraction of  $sp^3$  carbon atoms.

The most substantial improvement was achieved by pre-filtering the starting fragments to include only those with Csp<sup>3</sup>BM values of 0.3 or higher. This filtering reduced the number of starting fragments from 23840 to 2851 but significantly increased the proportion of generated molecules with Csp<sup>3</sup>BM values meeting the threshold, from 15-22% to 63-66% (Figure 8a). This filtering also resulted in higher synthetic accessibility (SA) scores for top-scoring compounds that satisfied PLIP and had Csp<sup>3</sup>BM  $\geq$  0.3, while maintaining docking scores comparable to those results obtained using the full fragment set (Figure 8b).

To further investigate, an alternative set of starting fragments was prepared using the same protocol outlined in the Methods section, starting from the CReM SA2.5 database. This yielded 27802 starting fragments enriched in  $sp^3$  carbon atoms. Using this set, a similarly high proportion of compounds meeting the desired criteria (57-65%) was achieved. The top-scoring molecules from this dataset exhibited slightly higher docking scores and comparable SA scores. Representative examples of these top-ranked molecules are provided in Table 3.

In conclusion, the widely employed strategy of augmenting docking scores with additional parameters to steer molecular generation towards a desired region of chemical space proved inefficient in this context, likely due to the need for fine-tuning the augmented objective function. A more effective approach was direct control over the composition of starting fragments and the fragments used during molecule growth.

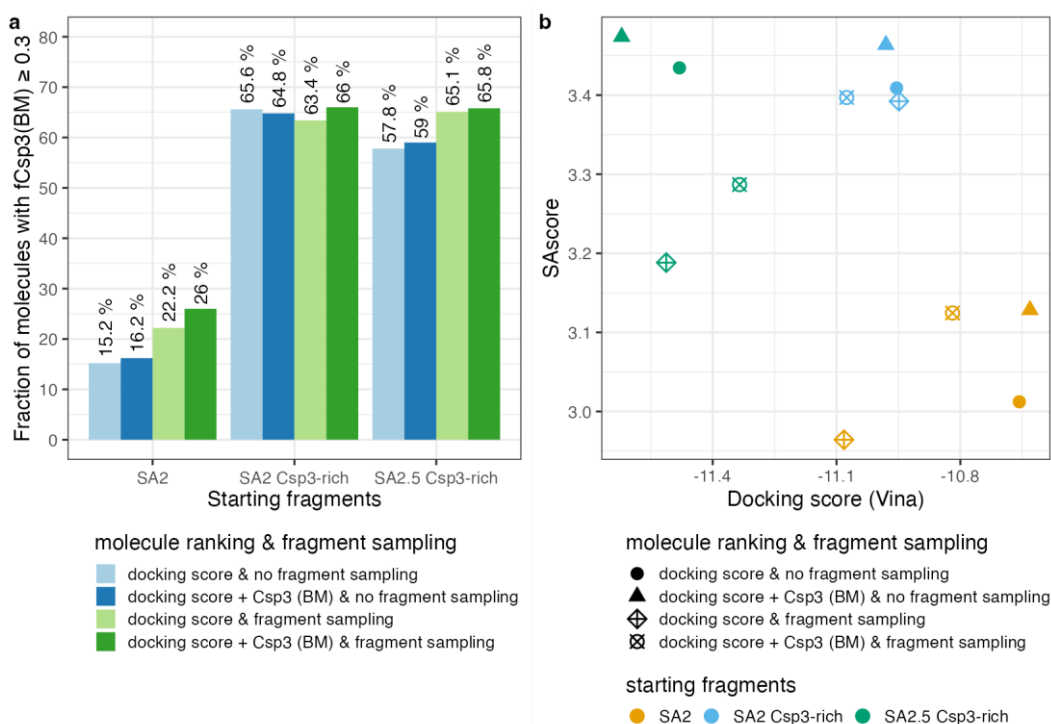
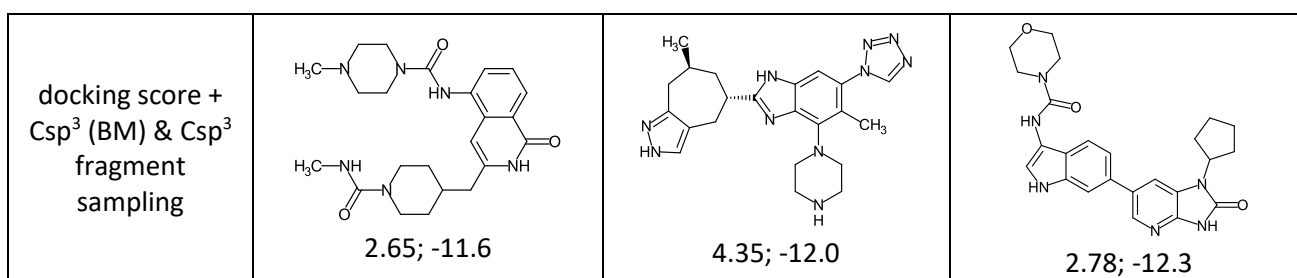


Figure 8. (a) The average fraction of molecules having at least 30% of  $sp^3$  carbon atoms in their Bemis-Murcko scaffolds, which were generated in three independent runs. (b) Average docking and SA scores for top 100 molecules across three independent runs for different setups. The top 100 molecules were selected among those which satisfied PLIP and had  $Csp^3BM \geq 0.3$ .

Table 3. Top scored generated compounds across three independent runs for each combination of settings. Compounds bind to the hinge region and have the fraction of  $sp^3$  carbon atoms in scaffolds equal or greater than 0.3.

	Starting fragments		
	SA2	SA2 Csp <sup>3</sup> -rich	SA2.5 Csp <sup>3</sup> -rich
docking score & no fragment sampling	 3.32; -11.7	 4.22; -11.8	 4.39; -12.3
docking score & Csp <sup>3</sup> fragment sampling	 3.56; -11.8	 4.19; -11.7	 3.62; -12.5
docking score + Csp <sup>3</sup> (BM) & no fragment sampling	 2.51; -11.6	 3.78; -12.3	 3.62; -12.5



### The effect of protein conformation

A protein in complexes with different ligands may have different conformations of binding site residues and we explored the effect of protein conformations on the output of the generative pipeline. We chose three other complexes of CDK2 (PDB: 2FVD, 3RAL, 6GUH) and performed three independent runs using the optimal parameters defined above. The results show that while SA scores vary in a narrow range of values, docking scores were affected more pronounced (Figure 9a). Diversity of Murcko scaffolds within individual runs was similar and was not dependent on protein conformation: there were 16-22 distinct scaffolds among top 100 compounds in average. Reproducibility of scaffolds for the same protein conformations was low and scaffolds were almost not reproduced across different protein conformations (Figure 9b). The generated compounds poorly overlap with the reference ChEMBL space and frequently created distant clusters. However, compounds generated for different protein conformations keep certain level of similarity. According to UMAP compounds generated for 2BTR and 3RAL are often closely clustered, there should be also some similarity between compounds from 2BTR and 6GUH (Figure 10). This is confirmed by checking the best scored molecules. Top scored compounds for 2BTR and 3RAL frequently possess the same 1(2H)-isoquinolinone core. Some molecules generated for 6GUH also have this core (Table 4).

These results suggest that for real applications it would be more reasonable to use different conformations of a protein to design compounds with higher docking score and better fitting to the shape of a protein binding site. These can be conformations from X-ray of protein-ligand complexes as well as conformations sampled from molecular dynamics simulation of complexes. However, the latter hypothesis we did not verify in this study.

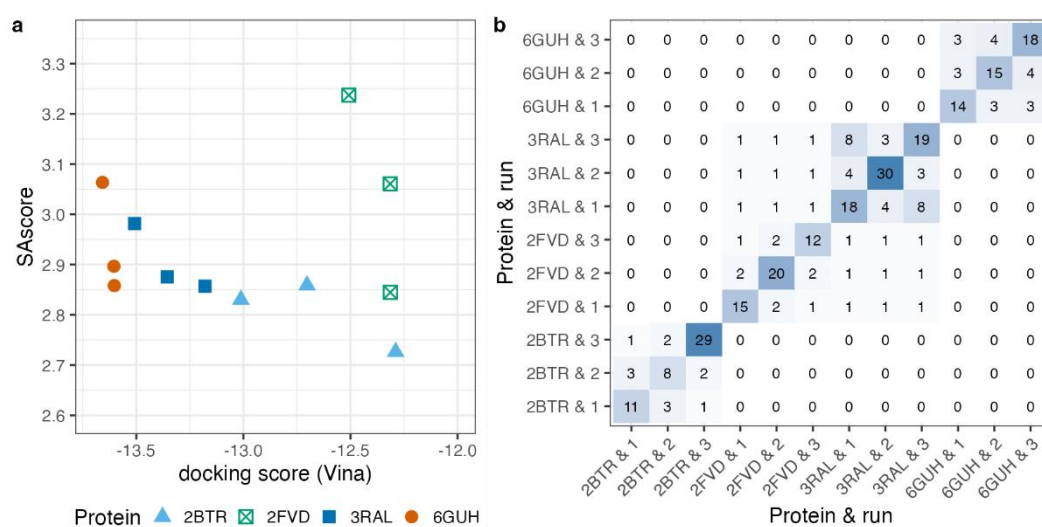
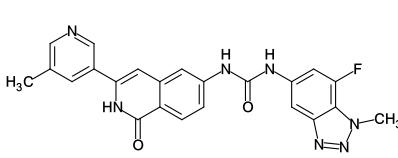
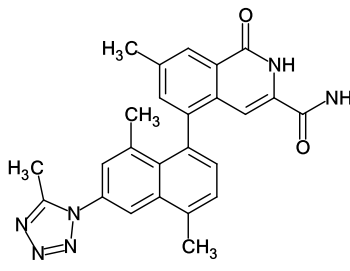
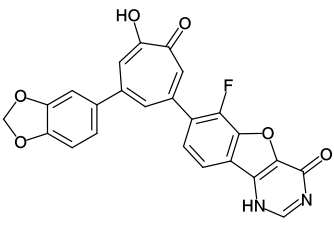
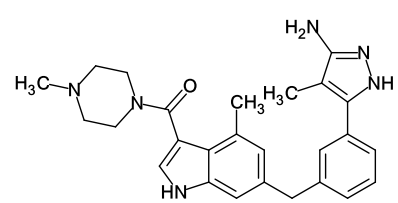
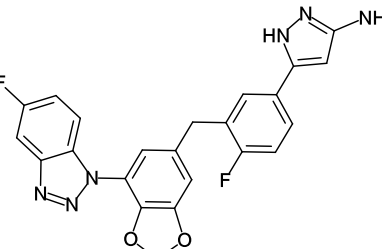
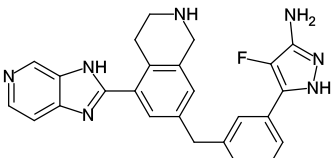
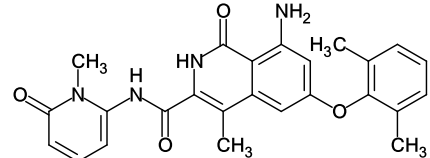
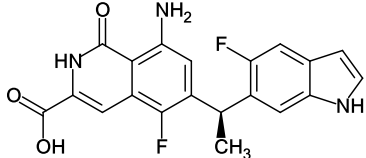
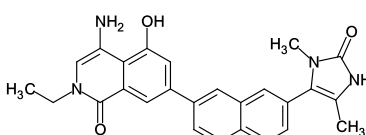
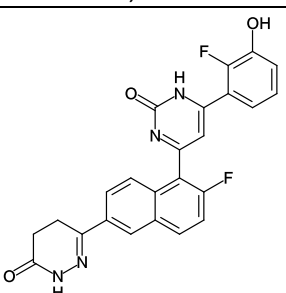
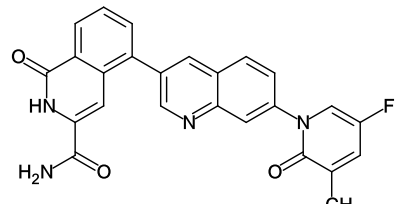
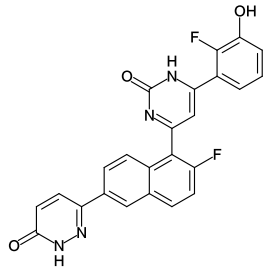


Figure 9. (a) Average docking and SA scores for top 100 generated molecules bound to the hinge region from three independent runs for each of CDK2 protein structures. (b) The number of distinct Murcko scaffolds for individual runs and across runs.

Table 4. Compounds with the highest docking score generated in individual runs for particular conformations of CDK2. The numbers below structures are SA and docking scores.

	Run1	Run 2	Run 3
2BTR	 <p>2.84; -13.5</p>	 <p>2.92; -13.8</p>	 <p>3.04; -13.3</p>
2FVD	 <p>2.78; -12.6</p>	 <p>2.98; -12.2</p>	 <p>3.12; -12.6</p>
3RAL	 <p>2.86; -12.7</p>	 <p>3.57; -11.2</p>	 <p>3.00; -11.9</p>
6GUH	 <p>3.06; -14.3</p>	 <p>2.81; -14.0</p>	 <p>2.92; -14.2</p>

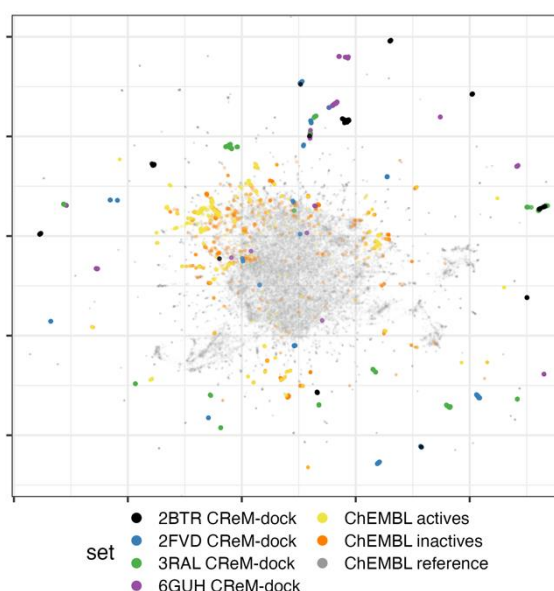


Figure 10. UMAP (number of neighbors is 10) of top 100 generated compounds from three independent runs for four different conformations of CDK2 protein (1200 compounds in total), known CDK2 inhibitors from ChEMBL33 ( $pK_i/pK_d/IC_{50} \geq 6$ ) and a random set of 50000 compounds from ChEMBL as a reference space.

#### *Post hoc evaluation of possible selectivity of designed CDK2 inhibitors*

We investigated possible selectivity of generated compounds to highly related protein targets. Therefore, we selected several kinases which share high sequence similarity to CDK2 (39-66% of identity) and had X-ray structures of protein-ligand complexes: CDK1 (6GU2), CDK5 (4AU8), CDK6 (6OQO), CDK7 (8P4Z), CDK16 (5G6V), MAPK7 (5BYZ) and MAPK13 (5EKO). For each kinase we determined protein-ligand interaction patterns encoding the hinge region and collected inhibitors having  $pIC_{50}/pK_i/pK_d \geq 6$  from ChEMBL33 as reference compounds. Known inhibitors were docked into corresponding proteins and we chose the median docking score as an activity threshold for further assessments (Table S2).

Top 100 compounds from each of the previous 45 generations of CDK2 ligands for 2BTR protein conformation (3 runs  $\times$  3 fragment databases  $\times$  5 radiuses) were selected for analysis. These compounds were docked to all of selected kinases and we counted the number of compounds having docking scores better than the median docking scores for reference active compounds. The designed compounds mainly achieved docking scores better than the median score of known inhibitors (Figure 11). The only notable drop was observed for CDK6. Docking scores of designed compounds, which bind to the hinge region of corresponding kinases, were frequently better than docking scores of known actives in all cases (Figure S2), but the number of such compounds is very low (Figure 11).

Thus, while designed compounds possess high docking scores to structurally related protein targets, they may not form important protein-ligand interactions and, therefore, may still be selective. Explicit inclusion of other proteins as anti-targets into the generation pipeline may solve this issue, but this is not implemented. However, this will proportionally increase required computing capacity and still will not guarantee generation of selective compounds, because these proteins may also exist in different conformational states while ligands may bind only to some of them and inclusion of all conformations of all anti-targets would be unfeasible.



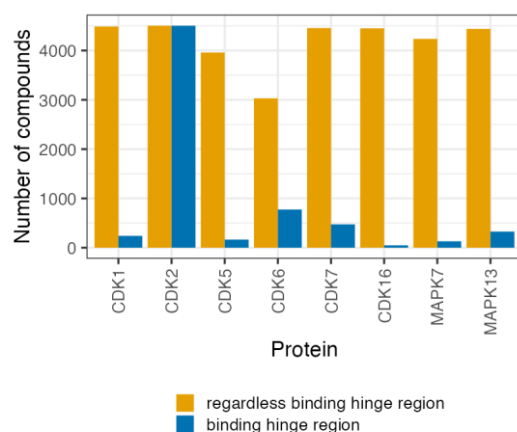


Figure 11. The number of designed compounds which achieved docking scores better than the median docking scores of corresponding actives for each kinase. Yellow bars denote compounds without consideration of hinge region binding and blue bars denote compounds which additionally satisfy at least 2 out of 3 hinge region contacts.

### De novo design of ligands for different protein families

To demonstrate wider applicability of the developed pipeline we chose typical human proteins from different protein classes (Table 5), which have experimental 3D structures and a large number of known ligands. For every structure we define a set of important protein-ligand contacts (H-bond donor or acceptor, metal acceptor or cationic interactions) which were used as a restriction during the generation process (Table S2). In all cases the threshold for similarity of protein-ligand interaction pattern (PLIP) was set to 0.6. This means that for patterns with one or two contacts all contacts should be found to make a compound eligible to be chosen for the next iteration. For patterns with three contacts, that meant the at least two contacts should be satisfied.

One simulation per target was executed using the optimal settings determined above. The total number of enumerated and docked compounds varied from 110 000 to 147 000 for individual targets. The percentage of compounds matched the require protein-ligand interaction patterns within 0.6 threshold was very variable, from 2.3% for ESR1 to 53.5% for HDAC2 (Table 5).

Table 5. Statistics on the number of generated compounds and diversity of Murcko scaffolds.

Protein target	Protein target name and family	PDB	Total number of generated compounds	Number of compounds satisfying PLIP at the level of 0.6	Number of distinct Murcko scaffolds in top 100 generated compounds satisfying PLIP
BACE1	Beta-secretase (protease)	6UWP	111 232	11 277 (10.1%)	8
DRD2	Dopamine D2 receptor (GPCR)	6CM4	125 005	36 695 (29.4%)	10
ESR1	Estrogen receptor (nuclear receptor)	8DV7	147 125	3 449 (2.34%)	40
HDAC2	Histone deacetylase 2 (epigenetic regulator)	7ZZT	112 085	59 994 (53.5%)	39
PARP1	Poly [ADP-ribose] polymerase 1 (transferase)	7ONT	110 931	27 328 (24.6%)	19

For the further analysis we selected top 100 compounds with the highest docking scores and satisfying PLIP requirements. In the majority of cases docking scores of top 100 generated compounds were better than those

scores of known active ligands ( $pK_d/pK_i/pIC_{50} \geq 6$ ) deposited in ChEMBL33 (Figure 12a). Synthetic accessibility scores were mainly below 3 and comparable with SA scores of compounds from ChEMBL (Figure 12b). One exception was DRD2 where SA scores of top 100 designed compounds were greater than SA scores of ChEMBL compounds ( $\sim 3.5$  vs.  $\sim 2.6$ ). Other exception was BACE1, where designed compounds had better SA scores than compounds from ChEMBL ( $\sim 2.8$  vs.  $\sim 4$ ). High SA scores for ChEMBL compounds can be explained by presence of many polycyclic and spirocyclic chiral compounds among actives. Novelty of generated compounds was high. Tanimoto similarity calculated for every generated compound based on 2048-bit Morgan radius 2 fingerprints to the closest neighbor from the whole ChEMBL33 and from the subset of known actives was mainly below 0.5 and 0.3, respectively, indicating high novelty of generated compounds (Figure 12c).

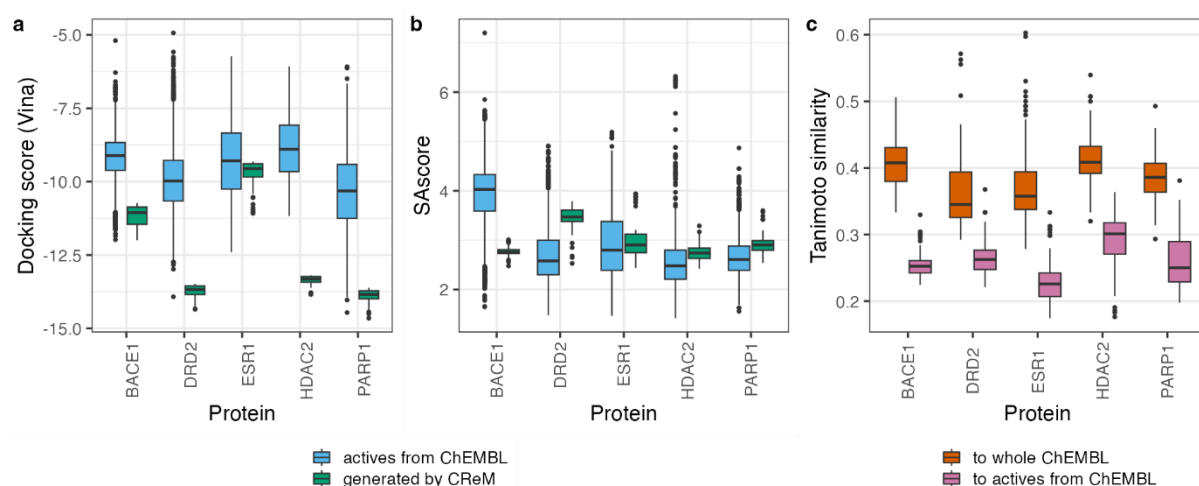


Figure 12. (a) Docking scores for top 100 generated compounds satisfying PLIP requirements and active compounds from ChEMBL33 ( $pK_i/pIC_{50} \geq 6$  and  $MW \leq 500$ ) regardless their protein-ligand interaction patterns. (b) SA scores for the same compounds as on plot (a). (c) Novelty of generated compounds expressed as Tanimoto similarity of top 100 generated compounds to the closest neighbor from the whole ChEMBL33 and from known actives.

### Comparison with state-of-the-art approaches

For comparison purposes we chose REINVENT4 [42] (<https://github.com/MolecularAI/REINVENT4>) which includes the previously developed DockStream approach [23] for de novo design guided by molecular docking. REINVENT is a recurrent neural network model trained on SMILES of ChEMBL structures. DockStream implements reinforcement learning using molecular docking, optionally augmented, as a reward function. As targets we used CDK2 (PDB 2BTR), BACE1 (PDB 6UWP), DRD2 (PDB 6CM4), ESR1 (PDB 8DV7), HDAC2 (PDB 7ZZT), PARP1 (PDB 7ONT), which were used in other studies [23, 43, 44]. To achieve better comparison results we adjusted REINVENT4 settings in accordance with CReM-dock setup, but we tried to keep them close to optimal ones suggested by the authors (Table S3). For every case we run 400 REINVENT iterations to achieve approximately the same number of docking events as in the case of CReM-dock approach. To make results more compatible we introduced to REINVENT4 the same version of AutoDock Vina 1.2.5 as we used for CReM-dock experiments. The exhaustiveness parameter was set to 8 as this is the default value in REINVENT4 and it could not be changed in settings. The same value was applied for CReM-dock generations. We used an objective function suggested by the REINVENT4 authors which is a geometric mean of docking scores scaled by the sigmoid function and drug-likeness (QED). This should make generated structures more drug-like. For CReM-dock we chose to use not only the docking score as a single objective but also a geometric mean of docking scores and QED as described above. The latter should make results more comparable. Since reinforcement learning optimizes a single agent, this may result in highly similar output structures. Therefore, we enabled the Murcko scaffold diversity filter suggested by the REINVENT authors, that

should explicitly control diversity of solutions and keep it high. Support of PLIP is not implemented in REINVENT4, therefore we performed generation without these restrictions. REINVENT4 and CReM-dock were run once for each target. Radical and charge states of compounds generated by REINVENT4 were fixed by an in-house script, otherwise calculated physicochemical properties and SA scores were incorrect as well as PLIP detection. A small number of compounds (~500 across all generations) was discarded as the script was unable to fix them.

The number of REINVENT compound docked to each target varied from 141 000 to 152 000. The percentage of REINVENT compounds satisfying the same physicochemical restrictions applied to CReM-dock varied from 34% to 51% (Table S4), that was lower than expected. The main issue was high lipophilicity of generated compounds. Top scored compounds had average lipophilicity above 5 almost for all targets (Table 6, Figure S4). Thus, augmentation of a docking score with QED did not fully solve the issue of generation of unfavorable molecules. The explanation could be that such highly lipophilic molecules have high docking score which compensate their poor drug-likeness. The explicit filtering of compounds by physicochemical properties implemented in CReM-dock looks more favorable because it avoids generation and docking of compounds with unfavorable properties and wasting computational resources.

The percentage of REINVENT compounds satisfying PLIP was expectably low for the majority of targets (0.01-2.7%) with the exception of HDAC2 for which 55.4% of molecules satisfied PLIP (Table S4). The small number of satisfying compounds can be explained by absence of explicit biasing of the REINVENT objective function with protein-ligand interaction patterns. The large number of generated HDAC2 ligands satisfying PLIP could be explained by the small number of required contacts – there was only one contact with Zn ion. However, in the case of DRD2 target, where the PLIP also consisted of only a single interaction with Asp114 (cationic), the number of generated compounds establishing this contact was low: 2274 compounds or 2.7% from the total number of generated molecules. This contact is considered canonical for dopamine D2 and many other GPCRs [45, 46] and highly likely it should be present in protein-ligand interactions of active compounds. This result highlights the importance of taking into account key protein-ligand interactions explicitly in order to generate a greater number of compounds preserving specific contacts. Otherwise, generation may be too broad and only a small portion of generated molecules will be able to establish these interactions.

Consideration of highly lipophilic compounds in real applications may be not reasonable, therefore we selected top 100 REINVENT molecules satisfying the same physicochemical rules and, additionally, PLIP constraints as applied to CReM-dock generations (Table 6). Simultaneous application of both these criteria resulted in even less than 100 compounds for some targets (BACE1 – 29 compounds and ESR1 – 4, Table S4). Applying these criteria decreased docking scores of the remaining top scored compounds, but improved their drug-likeness, while synthetic accessibility scores were remained almost the same (Figure 13). CReM-dock resulted in comparable docking scores of top scored compounds relatively to REINVENT compounds filtered by physicochemical properties. However, applying the PLIP filter for REINVENT compounds substantially reduced their number and correspondingly docking scores and even known actives outperformed molecules generated by REINVENT. This result was not very surprising because there was no bias by PLIP in the REINVENT objective function and therefore the number of analyzed compounds was much smaller than for CReM-dock. Synthetic accessibility of compounds generated by REINVENT and CReM-dock were comparable for the most targets. The notable difference was observed for DRD2 and HDAC2, where CReM-dock resulted in more complex molecules (Figure 13).

Novelty of compounds generated by REINVENT was somewhat lower than for CReM-dock compounds. The generated REINVENT compounds were more similar to ChEMBL molecules, because the latter were used for training the REINVENT model (Figure 14). CReM-dock using fragments from the ChEMBL database generated compounds which were less similar to the reference space. CReM-dock compounds overlap with ChEMBL reference space to the lesser extent and sometimes create separate distant clusters (Figure S5). However, we found that

there were a reasonable number of known actives from ChEMBL which were similar to some of generated compounds (Table S5), that confirms that CReM-dock also explores relevant chemical space.

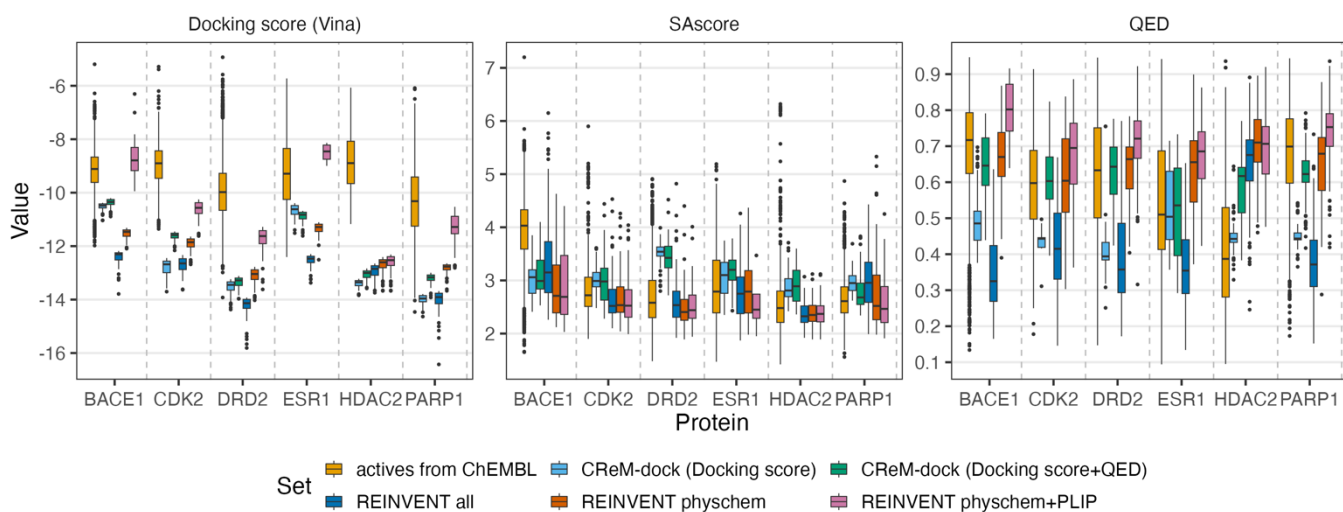


Figure 13. Distribution of properties for top 100 compounds (with the highest docking scores) generated by CReM-dock and REINVENT in comparison with known actives from ChEMBL33. REINVENT molecules were additionally filtered by physicochemical properties and PLIP as CReM-dock settings.

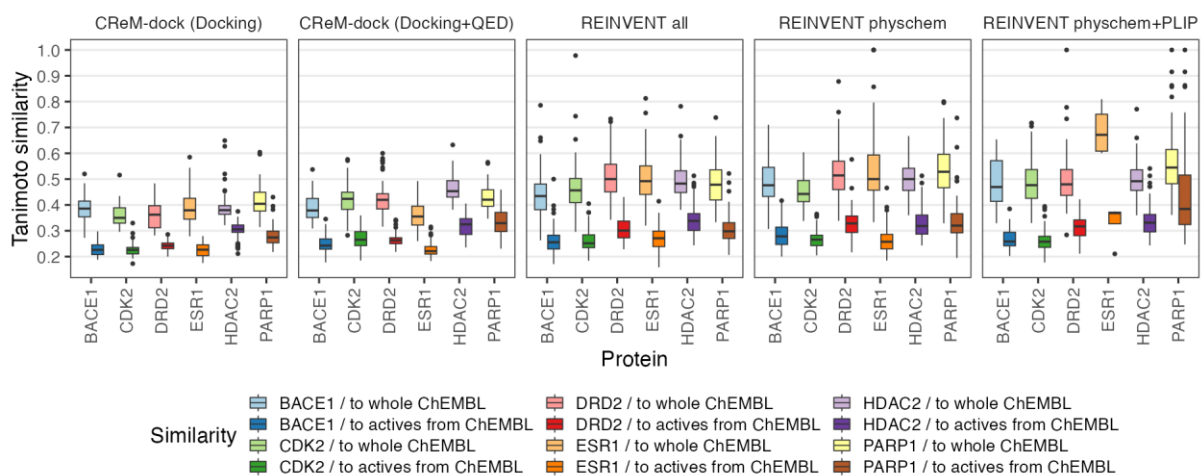
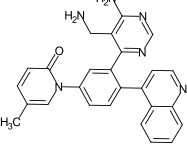
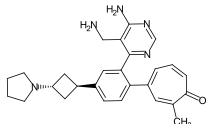
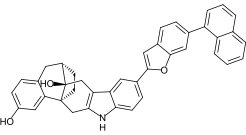
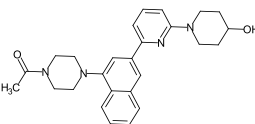
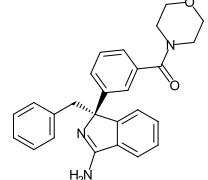
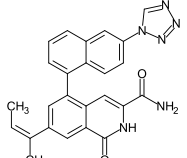
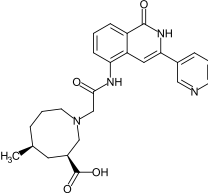
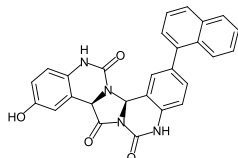
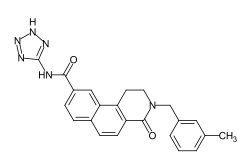
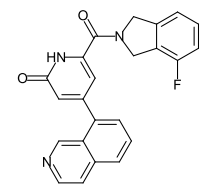
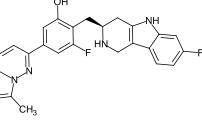
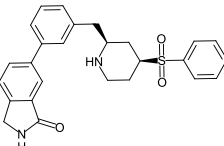
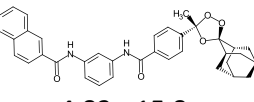
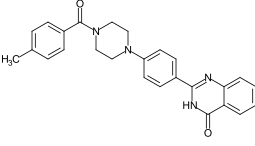
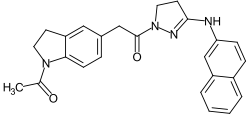
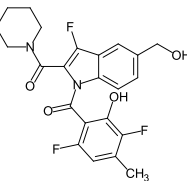
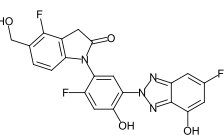
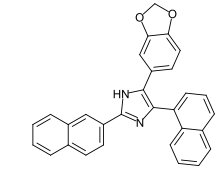
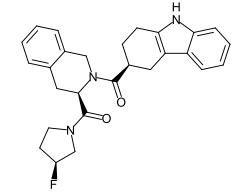
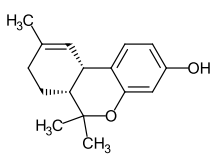
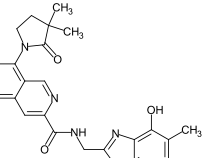
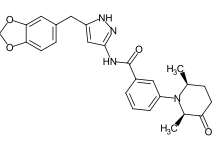
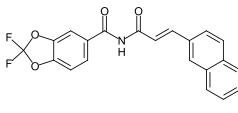
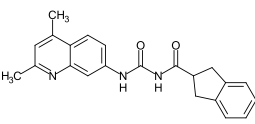
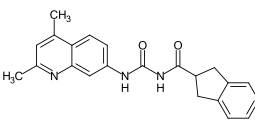
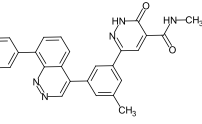
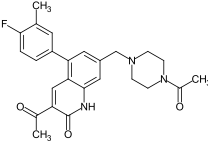
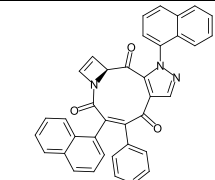
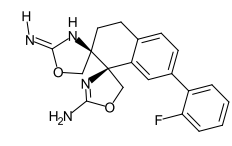
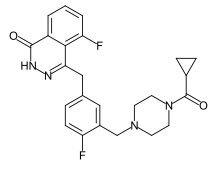


Figure 14. Novelty of top 100 compounds generated by CReM-dock and REINVENT for different protein targets.

Table 6. Top scoring structures generated by CReM-dock and REINVENT4. The numbers below structures are SA and docking scores, respectively.

	CReM-dock (Docking)	CReM-dock (Docking + QED)	REINVENT4	REINVENT4 (physicochemical filters)	REINVENT4 (physicochemical filters + PLIP)
BACE1	 2.73; -11.0	 2.88; -10.9	 4.71; -13.8	 2.40; -12.1	 3.03; -10.0
CDK2	 2.95; -13.7	 3.22; -12.2	 3.63; -13.6	 2.52; -12.7	 2.59; -11.6
DRD2	 3.60; -14.4	 3.39; -14.0	 4.82; -15.8	 2.11; -14.0	 2.66; -13.6
ESR1	 2.99; -11.5	 3.27; -11.6	 2.35; -13.4	 3.62; -12.5	 3.47; -9.00
HDAC2	 2.94; -13.8	 3.53; -13.6	 2.41; -13.7	 2.3; -13.7	 2.3; -13.7
PARP1	 2.75; -14.6	 2.48; -13.9	 3.90; -16.4	 5.15; -13.7	 2.55; -12.8

### Fragment expansion study

CReM-dock approach based on fragment growing perfectly suits to tasks where a smaller ligand should be expanded within a binding site to generate a larger molecule better fill the cavity. To validate this ability of CReM-dock we chose pairs of compounds from the previous work of Malhotra and Karanicolas [47], who collected a large set of pairs of smaller and larger ligands co-crystallized with the same protein and which are available in PDB database. As criteria for ligand pairs selection we used: (i) the number of heavy atoms in a larger ligand is less than

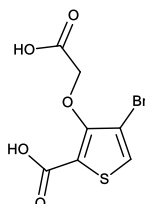
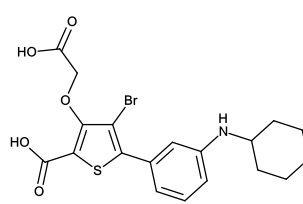
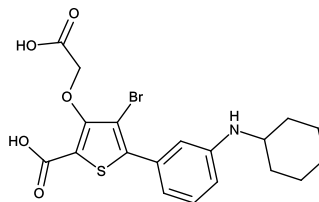
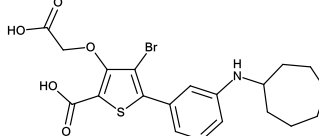
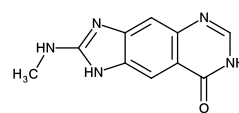
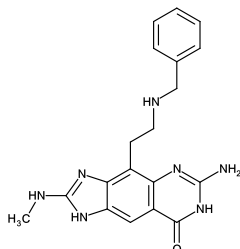
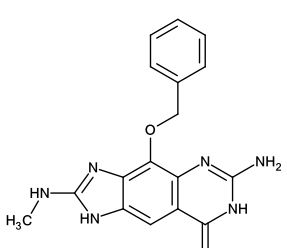
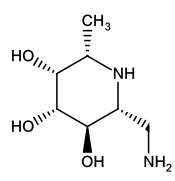
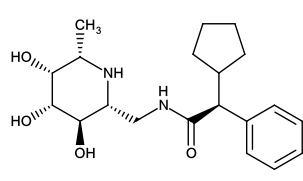
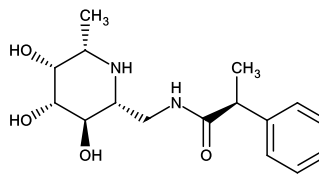
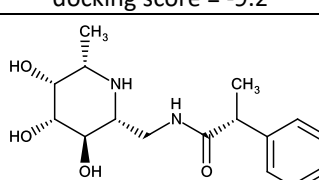
36 to roughly satisfy  $MW \leq 500$ , (ii) smaller and larger ligands should have high volume overlap indicating that the pose is not substantially changed upon expansion, (iii) RMSD between binding pockets of corresponding protein conformations was low indicating that there were no major changes in positions of side chain residues and (iv) the difference in activity between smaller and larger ligand was greater than two orders of magnitude. Finally, we chose three pairs of ligands satisfying these criteria (Table 7).

To run simulations, we adjusted some settings to not going search too far and save computational resources, because molecules upon growing increase their molecular weight, it was unreasonable to continue generation after reaching the molecular mass of a desired ligand. The threshold values of physicochemical parameters were set to the corresponding values of a larger ligand + 15% (MW, logP, TPSA) and RTB was set to the corresponding value of a larger ligand + 1. To further restrict the generation, we chose RMSD threshold to 1 Å and set minimum PLIP similarity 0.6 for 2ZWZ and 3S1G and 0.5 for 2HB1 (Table S6). This should guide generations towards compounds preserving the binding pose of a starting fragment and its interactions with a corresponding protein. For generation we chose ChEMBL SA2 fragment library and radius 2. As a selection strategy we used clustering with 100 clusters and selection of top 1 compounds from every cluster. This was done to make search more exhaustive and cover a larger chemical space. This did not increase computational costs substantially, because we started from relatively large fragments and the target molecules should be generated in a few steps. Overall, for each of three targets from 7600 to 23600 molecules were generated.

From the pool of generated compounds we selected most similar ones to the target compounds using Tanimoto similarity on 2048-bit Morgan fingerprints of radius 2. In all cases the most similar generated ligands possessed the same binding mode and preserved important protein-ligand interactions (Table 7, Figure 15). In the case of 2HB1-2QBS pair the pipeline was able to find a compound identical to the larger ligand 2QBS. There was also another compound with similarity 1, which contained a cycloheptyl residue instead of a cyclohexyl. Both compounds had very similar binding modes relatively to the target ligand. In the case of 3S1G-3GC4 pair the most similar generated ligand did not contain a positively charged secondary amine center and, thus, could not establish a contact with Asp280 like the target ligand. However, the generated compound contained an amino group strengthen the binding to Asp156 and a hydrophobic residue filling the pocket similarly as the target compound. In the case of 2ZWZ-2ZX9 pair the most similar generated ligand fills the hydrophobic pocket highly similar to the target compound, however, it contained a methyl group instead of a cyclopentyl residue. The most similar compound is represented by two enantiomers. While both enantiomers could adopt the binding site and preserve the contacts, S-enantiomer having the same configuration of the corresponding chiral center as the target compound had a slightly better docking score and RMSD value than R-enantiomer. In all cases the generated compounds could at least approach the chemical space represented by a target molecule and fit the binding site preserving the pose and contacts (Figure 15). This proves the ability of the approach to explore not only novel chemical space, but also relevant space which was previously experimentally confirmed.



Table 7. The pairs of smaller (starting) and larger (target) ligands and designed compounds the most similar to the corresponding target one.

Starting ligand	Target ligand	Similarity of starting and target molecules	Generated molecules most similar to the target one	Similarity of a generated molecule to the target ligand	RMSD of a generated ligand relatively to the starting one
 <b>2HB1</b> $K_i = 160 \mu\text{M}$	 <b>2QBS</b> $K_i = 210 \text{ nM}$	0.36	 1	1	1.25
			 1	1	1.52
 <b>3S1G</b> $K_i = 6500 \text{ nM}$	 <b>3GC4</b> $K_i = 25 \text{ nM}$	0.32	 0.63	0.06	
 <b>2ZWZ</b> $K_i = 16.3 \text{ nM}$	 <b>2ZX9</b> $K_i = 0.054 \text{ nM}$	0.32	 docking score = -9.2	0.69	0.86
			 docking score = -9.05	0.69	1.03

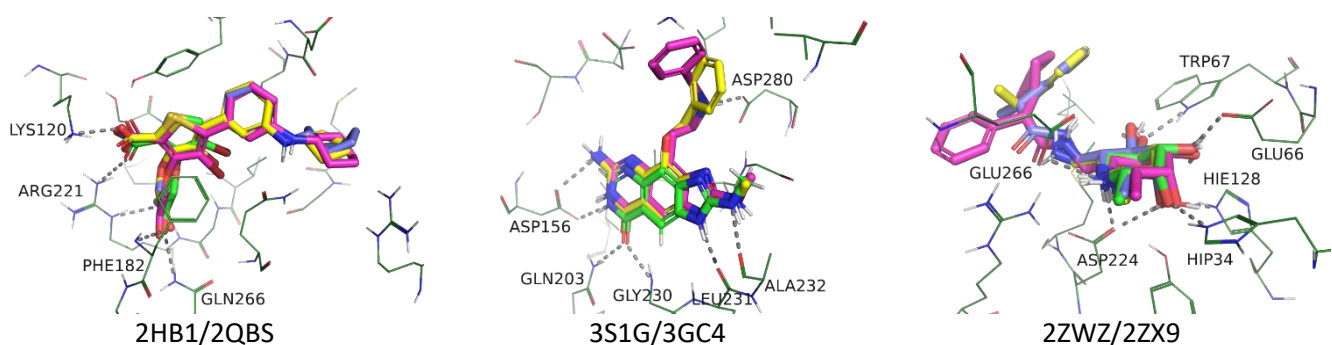


Figure 15. Binding poses of smaller (green) and larger (magenta) ligands as well as generated compounds (yellow/blue) the most similar to the target (larger) ligand.

## Discussion

Practical tips:

- Use larger context radiuses and/or fragments obtained from more synthetically accessible molecules
- For larger radiuses and smaller fragment databases there is no need to make repeated runs, results are highly reproducible
- For higher diversity of solutions choose Pareto selection, for more robust runtime – selection based on clustering
- Use several protein conformations, if possible
- Apply restrictions to protein-ligand interactions, if possible
- Augmentation of an objective function works for drug-likeness
- Use fragments enriched with sp<sup>3</sup> carbon atoms to generate corresponding molecules, augmentation of the objective function works poorly in this case

Analyzing outputs of CReM-dock we noticed that there were many top scored compounds bearing many methyl groups or halogen atoms. Those atoms are mainly added at the later iterations and are unlikely critically important for ligand-protein recognition and binding. However, introduction of these groups can make structures more complex and less synthetically feasible. To overcome this issue, we may suggest to specify the minimum size of the attached fragment greater than 1, which is used by default.

A fragment expansion strategy may be applied in an alternative context. Specifically, an anchor fragment that demonstrates binding to a protein can be derived from a co-crystallized ligand. This fragment can subsequently be expanded using CReM-dock while preserving its binding pose and interactions. Compared to de novo design, this approach offers potential advantages, as the binding pose of at least one fragment is known, thereby increasing the likelihood of generating active molecules.

## Conclusions

The suggested fragment-based approach to structure generation guided by molecular docking demonstrated promising results. It solves to some extent the issue of synthetic accessibility of generated compounds and proposes indirect, but fine-tuning control over it. It was demonstrated that choosing of a larger context radius and a fragment database created by fragmenting more synthetically accessible molecules improves synthetic accessibility of generated structures. There is an option to augment the docking score with different physicochemical parameters to generate more favorable compounds. While this worked well in the case of drug-likeness it was not an optimal solution for generating of compounds with a large fraction of sp<sup>3</sup> carbon atoms. In the latter case the better solution was explicit biasing by selecting Csp<sup>3</sup>-enriched starting fragments or fragments used for growing. Further flexibility of the approach comes from different selection strategies which allow to control diversity of created molecules. It was also demonstrated that using different protein conformations taken from different complexes results in diverse structures and different docking scores of generated compounds and, therefore, the reasonable strategy would be to use multiple available protein conformations to generate compounds and analyze the combined output of all runs. The ability to preserve particular protein-ligand contacts during the generation is important if these contacts were previously determined as essential. This greatly increases the number of compounds possessing the corresponding interaction patterns. It is possible to apply an additional restriction by setting an RMSD threshold relatively to the pose of a starting fragment to enable preserving the binding pose of generated compounds. We demonstrated in three retrospective studies that this strategy resulted in generation of compounds which were

identical or highly similar to the known target molecules which were designed by medicinal chemists in the course of optimization of starting fragments. The developed tool is competitive to the state-of-the-art REINVENT4 approach using recurrent neural network to generate structures and reinforcement learning with a sophisticated objective function to guide generation. The structures generated by CReM-dock had frequently higher novelty and comparable docking and synthetic accessibility scores. The developed tool proposes great flexibility to adapt it to particular needs and address de novo generation as well as fragment expansion or scaffold decoration tasks.

### **Availability and requirements**

Project name: CReM-dock

GitHub: <https://github.com/ci-lab-cz/crem-dock>

Operating system(s): Linux

Programming language: Python 3

Other requirements: RDKit, EasyDock

License: BSD-3

Any restrictions to use by non-academics: no

### **Data deposition**

Structures of all molecules generated in the course of this study are accessible by this link - <https://doi.org/10.5281/zenodo.14577996>.

### **Acknowledgements**

The authors thank Guzel Mindubaeva for her contribution on early stages of the project, Aleksandra Ivanova helping implementation some features and Veincent Yap for his contribution to the EasyDock project to better sample ring conformations for more accurate docking.

### **Author contribution**

G.M. software development, design of the study, simulations, analysis, draft manuscript writing. P.P. design of the study, analysis, draft manuscript writing and editing, project supervision and funding.

### **Competing interests**

The author declares no competing interests.

### **Funding**

The work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through INTER\_EXCELLENCE II grant LUAUS23262 and the e-INFRA CZ (ID:90254), and partially by ELIXIR-CZ (LM2023055) and CZ-OPENSOURCE (LM2023052).



## 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. <http://dx.doi.org/10.1007/s10822-013-9672-4>
- (2) Schneider G, Fechner U (2005) Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery* 4:649-663. 10.1038/nrd1799
- (3) Hartenfeller M, Schneider G (2011) Enabling future drug discovery by de novo design. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1:742-759. doi:10.1002/wcms.49
- (4) Meyers J, Fabian B, Brown N (2021) De novo molecular design and generative models. *Drug Discovery Today*. <https://doi.org/10.1016/j.drudis.2021.05.019>
- (5) Bilodeau C, Jin W, Jaakkola T, Barzilay R, Jensen KF (2022) Generative models for molecular discovery: Recent advances and challenges. *WIREs Computational Molecular Science* n/a:e1608. <https://doi.org/10.1002/wcms.1608>
- (6) Gao W, Coley CW (2020) The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* 60:5714-5723. 10.1021/acs.jcim.0c00174
- (7) Domenico G, Giuseppe Felice M, Marco C, Angelo C, Orazio N (2016) Applicability Domain for QSAR Models: Where Theory Meets Reality. *International Journal of Quantitative Structure-Property Relationships (IJQSPR)* 1:45-63. 10.4018/ijqspr.2016010102
- (8) Ochi S, Miyao T, Funatsu K (2017) Structure Modification toward Applicability Domain of a QSAR/QSPR Model Considering Activity/Property. *Mol. Inf.* 36:1700076-n/a. 10.1002/minf.201700076
- (9) Li J, Fu A, Zhang L (2019) An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdisciplinary Sciences: Computational Life Sciences* 11:320-328. 10.1007/s12539-019-00327-w
- (10) Pinzi L, Rastelli G (2019) Molecular Docking: Shifting Paradigms in Drug Discovery. *International Journal of Molecular Sciences* 20:4331.
- (11) Shen C, Ding J, Wang Z, Cao D, Ding X, Hou T (2020) From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *WIREs Computational Molecular Science* 10:e1429. <https://doi.org/10.1002/wcms.1429>
- (12) Batiste L, Unzue A, Dolbois A, Hassler F, Wang X, Deerein N, Zhu J, Spiliotopoulos D, Nevado C, Caflisch A (2018) Chemical Space Expansion of Bromodomain Ligands Guided by in Silico Virtual Couplings (AutoCouple). *ACS Cent. Sci.* 4:180-188. 10.1021/acscentsci.7b00401
- (13) Chevillard F, Rimmer H, Betti C, Pardon E, Ballet S, van Hilten N, Steyaert J, Diederich WE, Kolb P (2018) Binding-Site Compatible Fragment Growing Applied to the Design of  $\beta$ 2-Adrenergic Receptor Ligands. *J. Med. Chem.* 61:1118-1129. 10.1021/acs.jmedchem.7b01558
- (14) Sommer K, Flachsenberg F, Rarey M (2019) NAOMInext – Synthetically feasible fragment growing in a structure-based design context. *European Journal of Medicinal Chemistry* 163:747-762. <https://doi.org/10.1016/j.ejmech.2018.11.075>
- (15) Spiegel JO, Durrant JD (2020) AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *J. Cheminf.* 12:25. 10.1186/s13321-020-00429-4
- (16) Yuan Y, Pei J, Lai L (2020) LigBuilder V3: A Multi-Target de novo Drug Design Approach. *Frontiers in Chemistry* 8. 10.3389/fchem.2020.00142
- (17) Steinmann C, Jensen JH (2021) Using a genetic algorithm to find molecules with good docking scores. *PeerJ Physical Chemistry*.
- (18) Ertl P, Schuffenhauer A (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* 1:8. 10.1186/1758-2946-1-8
- (19) Chéron N, Jasty N, Shakhnovich EI (2016) OpenGrowth: An Automated and Rational Algorithm for Finding New Protein Ligands. *J. Med. Chem.* 59:4171-4188. 10.1021/acs.jmedchem.5b00886
- (20) Cross S, Cruciani G (2022) FragExplorer: GRID-Based Fragment Growing and Replacement. *J. Chem. Inf. Model.* 10.1021/acs.jcim.1c00821
- (21) Boitreaud J, Mallet V, Oliver C, Waldispühl J (2020) OptiMol: Optimization of Binding Affinities in Chemical Space for Drug Discovery. *J. Chem. Inf. Model.* 60:5658-5666. 10.1021/acs.jcim.0c00833
- (22) Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. *Nature Chemistry* 4:90. 10.1038/nchem.1243

- (23) Guo J, Janet JP, Bauer MR, Nittinger E, Giblin KA, Papadopoulos K, Voronov A, Patronov A, Engkvist O, Margreitter C (2021) DockStream: a docking wrapper to enhance de novo molecular design. *J. Cheminf.* 13:89. 10.1186/s13321-021-00563-7
- (24) Ma B, Terayama K, Matsumoto S, Isaka Y, Sasakura Y, Iwata H, Araki M, Okuno Y (2021) Structure-Based de Novo Molecular Generator Combined with Artificial Intelligence and Docking Simulations. *J. Chem. Inf. Model.* 10.1021/acs.jcim.1c00679
- (25) Xu Z, Wauchope OR, Frank AT (2021) Navigating Chemical Space by Interfacing Generative Artificial Intelligence and Molecular Docking. *J. Chem. Inf. Model.* 10.1021/acs.jcim.1c00746
- (26) Polishchuk P (2020) CReM: chemically reasonable mutations framework for structure generation. *J. Cheminf.* 12:28. 10.1186/s13321-020-00431-w
- (27) Polishchuk P (2020) Control of Synthetic Feasibility of Compounds Generated with CReM. *J. Chem. Inf. Model.* 60:6074-6080. 10.1021/acs.jcim.0c00792
- (28) Minibaeva G, Ivanova A, Polishchuk P (2023) EasyDock: customizable and scalable docking tool. *J. Cheminf.* 15:102. 10.1186/s13321-023-00772-2
- (29) Bouysset C, Fiorucci S (2021) ProLIF: a library to encode molecular interactions as fingerprints. *J. Cheminf.* 13:72. 10.1186/s13321-021-00548-6
- (30) Ackloo S, Al-awar R, Amaro RE, Arrowsmith CH, Azevedo H, Batey RA, Bengio Y, Betz UAK, Bologna CG, Chodera JD, Cornell WD, Dunham I, Ecker GF, Edfeldt K, Edwards AM, Gilson MK, Gordijo CR, Hessler G, Hillisch A, Hogner A, Irwin JJ, Jansen JM, Kuhn D, Leach AR, Lee AA, Lessel U, Morgan MR, Moulton J, Muegge I, Oprea TI, Perry BG, Riley P, Rousseaux SAL, Saikatendu KS, Santhakumar V, Schapira M, Scholten C, Todd MH, Vedadi M, Volkamer A, Willson TM (2022) CACHE (Critical Assessment of Computational Hit-finding Experiments): A public-private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nature Reviews Chemistry* 6:287-295. 10.1038/s41570-022-00363-z
- (31) Li F, Ackloo S, Arrowsmith CH, Ban F, Barden CJ, Beck H, Beránek J, Berenger F, Bolotokova A, Bret G, Breznik M, Carosati E, Chau I, Chen Y, Cherkasov A, Corte DD, Denzinger K, Dong A, Draga S, Dunn I, Edfeldt K, Edwards A, Eguida M, Eisenhuth P, Friedrich L, Fuerll A, Gardiner SS, Gentile F, Ghiabi P, Gibson E, Glavatskikh M, Gorgulla C, Guenther J, Gunnarsson A, Gusev F, Gutkin E, Halabelian L, Harding RJ, Hillisch A, Hoffer L, Hogner A, Houlston S, Irwin JJ, Isayev O, Ivanova A, Jacquemard C, Jarrett AJ, Jensen JH, Kireev D, Kleber J, Koby SB, Koes D, Kumar A, Kurnikova MG, Kutlushina A, Lessel U, Liessmann F, Liu S, Lu W, Meiler J, Mettu A, Minibaeva G, Moretti R, Morris CJ, Narangoda C, Noonan T, Obendorf L, Pach S, Pandit A, Perveen S, Poda G, Polishchuk P, Puls K, Pütter V, Rognan D, Roskams-Edris D, Schindler C, Sindt F, Spiwok V, Steinmann C, Stevens RL, Talagayev V, Tingey D, Vu O, Walters WP, Wang X, Wang Z, Wolber G, Wolf CA, Wortmann L, Zeng H, Zepeda CA, Zhang KYJ, Zhang J, Zheng S, Schapira M (2024) CACHE Challenge #1: Targeting the WDR Domain of LRRK2, A Parkinson's Disease Associated Protein. *J. Chem. Inf. Model.* 64:8521-8536. 10.1021/acs.jcim.4c01267
- (32) Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* 25:1605-1612. <https://doi.org/10.1002/jcc.20084>
- (33) Shapovalov M, Dunbrack R (2011) A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* 19:844-858. <https://doi.org/10.1016/j.str.2011.03.019>
- (34) Webb B, Sali A (2016) Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics* 54:5.6.1-5.6.37. <https://doi.org/10.1002/cpbi.3>
- (35) Dalke A (2019) The chemfp project. *J. Cheminf.* 11:76. 10.1186/s13321-019-0398-8
- (36) McInnes L, Healy J (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv 1802.03426.
- (37) Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños María P, Mosquera Juan F, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux Chris J, Segura-Cabrera A, Hersey A, Leach Andrew R (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* 47:D930-D940. 10.1093/nar/gky1075
- (38) *standardizer version 22.19.0, ChemAxon* (<https://www.chemaxon.com>); (accessed).
- (39) *cxcalc version 22.19.0, ChemAxon* (<https://www.chemaxon.com>); (accessed).
- (40) Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, Moroz YS, Mayfield J, Sayle RA (2020) ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* 60:6065-6073. 10.1021/acs.jcim.0c00675

- (41) Lovering F, Bikker J, Humblet C (2009) Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J. Med. Chem.* 52:6752-6756. 10.1021/jm901241e
- (42) Loeffler HH, He J, Tibo A, Janet JP, Voronov A, Mervin LH, Engkvist O (2024) Reinvent 4: Modern AI-driven generative molecule design. *J. Cheminf.* 16:20. 10.1186/s13321-024-00812-5
- (43) García-Ortegón M, Simm GNC, Tripp AJ, Hernández-Lobato JM, Bender A, Bacallado S (2022) DOCKSTRING: Easy Molecular Docking Yields Better Benchmarks for Ligand Design. *J. Chem. Inf. Model.* 62:3486-3502. 10.1021/acs.jcim.1c01334
- (44) Thomas M, O'Boyle NM, Bender A, De Graaf C (2024) MolScore: a scoring, evaluation and benchmarking framework for generative models in de novo drug design. *J. Cheminf.* 16:64. 10.1186/s13321-024-00861-w
- (45) Mansour A, Meng F, Meador-Woodruff JH, Taylor LP, Civelli O, Akil H (1992) Site-directed mutagenesis of the human dopamine D2 receptor. *European Journal of Pharmacology: Molecular Pharmacology* 227:205-214. [https://doi.org/10.1016/0922-4106\(92\)90129-J](https://doi.org/10.1016/0922-4106(92)90129-J)
- (46) Józwiak K, Płazińska A. Structural Insights into Ligand—Receptor Interactions Involved in Biased Agonism of G-Protein Coupled Receptors. In *Molecules*, 2021; Vol. 26.
- (47) Malhotra S, Karanicolas J (2017) When Does Chemical Elaboration Induce a Ligand To Change Its Binding Mode? *J. Med. Chem.* 60:128-145. 10.1021/acs.jmedchem.6b00725