Journal of Cheminformatics

https://doi.org/10.1186/s13321-026-01195-5

Article in Press

CReM-pharm: de novo 3D pharmacophore-based design with synthetic accessibility awareness

Received: 31 Dec 2025

Accepted: 27 Mar 2026

Published online: 15 April 2026

Cite this article as: Denzler, A., Sriramulu, D., Pecha, J. *et al.* CReM-pharm: de novo 3D pharmacophore-based design with synthetic accessibility awareness. *J Cheminform* (2026). https://doi.org/10.1186/s13321-026-01195-5

Alina Denzler, Dinesh Sriramulu, Jozef Pecha & Pavel Polishchuk

CReM-pharm: de novo 3D pharmacophore-based design with synthetic accessibility awareness

Alina Denzler[1], Dinesh Kumar Sriramulu[1], Jozef Pecha[1], Pavel Polishchuk[1,*]

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

[*] Corresponding author: Pavel Polishchuk, pavlo.polishchuk@upol.cz

Abstract

De novo design methodologies have the potential to significantly enhance the exploration of chemical space in the search for promising ligands featuring novel chemotypes. This exploration can be directed through various computational strategies. 3D pharmacophore models, which represent the interaction patterns critical for protein-ligand recognition, can serve as valuable tools for the design of novel compounds. A common limitation of many generative approaches is the low synthetic feasibility of the generated molecular structures. In the present study, we developed a method capable of controllably generating compounds with a relatively high degree of synthetic accessibility by leveraging the CReM framework, while explicitly conforming to a specified 3D pharmacophore model. Evaluation of this approach across a diverse set of protein targets and pharmacophore models of varying complexity demonstrated its effectiveness and highlighted its advantages over the PGMG method, which employs a deep neural network architecture to generate ligands that may exhibit desired 3D geometries upon embedding. The proposed method has been implemented as an open-source tool, CReM-pharm, available at https://github.com/ci-lab-cz/crem-pharm.

Scientific Contribution

This study presents CReM-pharm, a software tool designed for the automated enumeration of molecular structures that explicitly conform to three-dimensional pharmacophore models. The approach specifically addresses challenges related to the synthetic accessibility of the generated structures and offers a means for predictable, indirect regulation of this aspect.

Keywords: de novo design, 3D pharmacophores, synthetic accessibility

Introduction

Pharmacophores encode electronic and steric features of a ligand that are essential for a biological response [1]. They facilitate clear interpretations and enable very fast screening of large compound libraries making them invaluable in the drug discovery process[2-5]. However, the screening of compound collections is constrained by the manageable number of compounds and the computational resources necessary for virtual screening. The largest databases of enumerated compounds contain approximately billions of entities[6], while the estimated size of drug-like chemical space is around 10^{36} molecules[7], making exhaustive screening of the entire chemical space impractical in the foreseeable future. To address this challenge, de novo approaches have been proposed, which dynamically enumerate compounds and adaptively explore chemical space, thereby allowing for the discovery of novel chemical entities in previously unexplored regions[8].

Current de novo methodologies predominantly rely on molecular docking or machine learning, with limited pharmacophore-based generation techniques available. Early strategies primarily utilized fragment-based generation, constructing molecules from small fragments[9-12]. Although this approach offered significant flexibility in generating novel compounds, it often resulted in poor synthetic accessibility, thereby limiting its practical application. This limitation arose from the fact that the linkage of synthetically accessible fragments did not consistently yield synthetically feasible molecules, as there was insufficient control over the bonds formed during generation. Consequently, de novo pharmacophore approaches were largely set aside until the advent of deep generative neural networks.

One of the first methods employing generative networks utilized the 3D shape of a reference molecule and identified pharmacophore features as additional constraints to train a shape autoencoder and a shape captioning network[13, 14]. Other researchers have employed reinforcement learning, using LigandScout as an agent to explicitly match a given 3D pharmacophore and guide model training[15]. The results indicated that while the generated molecules did not fully encompass all pharmacophore features, they significantly outperformed randomly sampled compounds in pharmacophore similarity scores. In another study, Imrie et al. proposed the use of graph convolution and 3D convolution networks to decorate scaffolds and generate linkers between two fragments using pharmacophore constraints[16]. They demonstrated that only 28% of the generated molecules exhibited at least 0.6 shape-based similarity to a given query. Building on Imrie's work, Hadfield et al.[17] developed the STRIFE approach, which extracts pharmacophoric information by calculating a fragment hotspot map[18] that identifies

regions of the binding pocket likely to contribute positively to binding affinity. STRIFE subsequently identifies pharmacophoric constraints that are likely to place a pharmacophore within a matching hotspot region and generates elaborations using a constrained graph variational autoencoder[17]. The efficacy of this approach has been validated through several retrospective case studies.

PGMG is another deep learning-based approach for de novo generation based on pharmacophore models[19]. In this approach, a pharmacophore model is represented as a complete graph, with nodes representing features and edges denoting distances, encoded by a gated graph convolution network, while molecules are embedded from SMILES. A notable aspect of PGMG is its ability to derive pharmacophore graphs from molecular graphs without requiring explicit 3D embedding. The transformer architecture is employed to learn the mapping between pharmacophore and molecular structure embeddings. The model was trained on compounds from ChEMBL (version 24), and using the pharmacophore graph as the sole input, PGMG could generate molecules that correspond to the respective pharmacophore. The authors demonstrated that PGMG could produce molecules with reasonable drug-like properties and docking scores superior to those of known active compounds. However, it is important to note that there is no assurance that all designed molecules will align with a query 3D pharmacophore model. The PGMG approach has certain limitations, including a restriction on the number of features in a model (not exceeding eight) and the inability to process pharmacophores with negatively charged centers. Furthermore, the synthetic accessibility of generated compounds was not explicitly addressed in the studies. Although it was shown that the range of synthetic accessibility scores[20] of generated compounds aligns with the reference ChEMBL space, the absence of explicit control raises concerns about the synthetic feasibility of the generated structures.

PhoreGen is a recently introduced deep learning-based method for pharmacophore-guided de novo molecular generation [21]. In this framework, a target pharmacophore model is used to guide the direct generation of complete 3D molecular structures. The approach is based on a diffusion model that applies asynchronous perturbation and denoising to both atomic coordinates and bond information, while a message-passing mechanism incorporates prior knowledge of ligand–pharmacophore mapping to maintain consistency between the evolving molecular structure and the pharmacophore constraints. A notable distinction from methods such as PGMG is that PhoreGen operates directly in 3D space and supports directed pharmacophore features, which may allow a more precise encoding of spatial interaction patterns. However, as with other generative approaches, the extent to which generated molecules consistently satisfy additional developability criteria may depend strongly on the specific pharmacophore model and generation settings [21].

In this study, we propose and implement a fragment-based de novo design approach for compounds guided by 3D pharmacophore models. We utilized a previously developed method for fragment-based structure generation, CReM, which allows for the growth of fragments while maintaining indirect but flexible control over the synthetic accessibility of the designed molecules[22]. The starting fragments match a selected subset of pharmacophore features, and matching fragments are sequentially expanded to match all features of a pharmacophore query. At each iteration, all structures are explicitly embedded in 3D space with the fixed coordinates of a parent part of a molecule, ensuring that the designed structures can adapt to the required conformation. The generation process is relatively rapid, as it does not necessitate the alignment of a molecule to a pharmacophore query at each iteration; instead, only the calculation of distances to the corresponding pharmacophore features is required to verify whether a molecule can match a query. Additionally, the approach generates molecules of minimal size sufficient to match a query, thereby accelerating the overall generation process and avoiding the creation of larger structures that do not contribute new features necessary for matching a query. There are no restrictions on the number of features in a pharmacophore model, and new feature types can be introduced by incorporating their SMARTS patterns into a configuration file.

Methods

De novo molecule generation using 3D pharmacophore queries

The workflow described herein employs an iterative growth strategy to expand initial molecular fragments that correspond to the starting features of a pharmacophore. This process aims to generate molecules that align with the maximum number of characteristics of a specified 3D pharmacophore. Pharmacophores are characterized by specific features, including hydrogen bond donors and acceptors, hydrophobic regions, aromatic, and positively or negatively charged centers, each defined by their respective coordinates in 3D space. Additionally, pharmacophore features may be accompanied with exclusion volumes, which delineate the boundaries of a binding site. In structure-based pharmacophore models, exclusion volumes are assigned to each atom of a protein that is located within 5 Å of the ligand's crystal structure. The radii of both the exclusion volumes and the pharmacophore features are determined on an individual basis. In this study, we utilized a feature radius of 1 Å and an exclusion volume radius of 2.2 Å to establish stringent parameters for the search process. Consequently, the input 3D pharmacophore is represented as a collection of spheres, encompassing both pharmacophore features and exclusion volumes within 3D space.

To expand molecules, we employed the CReM approach[22]. This methodology is based on the concept of interchangeable fragments, which are defined as fragments that exist within the same chemical context in existing molecules. To identify these fragments, we systematically fragmented existing molecules by cleaving up to four single bonds. Each fragment is linked to a corresponding chemical context of a given radius for all its attachment points. The context is determined by the remaining atoms of structure which are within up to a given number of bonds (context radius) from an attachment point of a fragment. To create a new molecule a chosen hydrogen atom is replaced with fragments from the CReM fragment database if their contexts are identical. As it was demonstrated previously this strategy resulted in novel molecules with reasonably high synthetic accessibility[22, 23].

The workflow consists of two major stages (Figure 1): (i) an initial screening phase (steps 1-7), in which starting fragments are identified by matching a subset of the pharmacophore features, and (ii) an iterative expansion phase (steps 8-19), in which these starting fragments are iteratively grown to satisfy all remaining features.

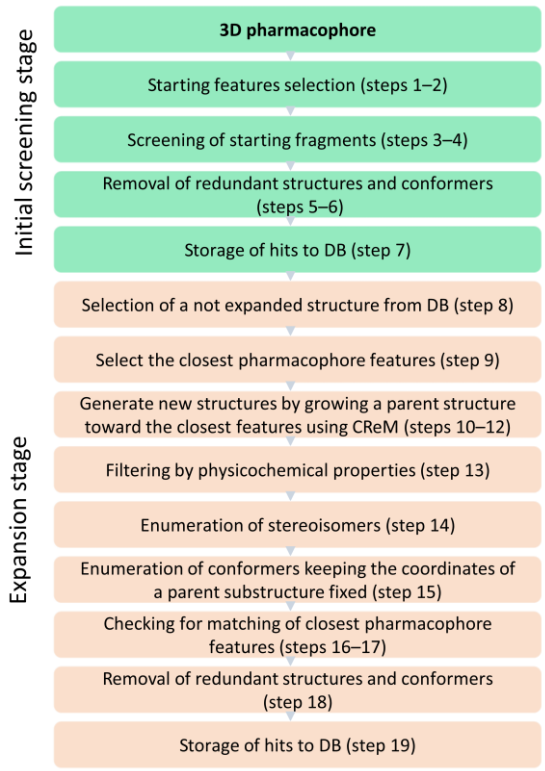


Figure 1. Overview of the generative pipeline. The expansion stage runs iteratively until all molecules will be expanded or the generation will be terminated.

1. The user first selects an initial set of pharmacophore features that are spatially proximal and likely to represent key interactions, such as those in the kinase hinge

region. As demonstrated below, we recommend selecting features located within 7 Å of one another; otherwise, suitable starting fragments may not be identified. In general, a more compact set of starting features is preferable, as it increases both the likelihood of finding matching fragments and the diversity of such matches. We further recommend selecting at least three non-collinear features. This imposes stricter constraints on fragment placement and thereby improves the robustness of the overall generation process. By contrast, selecting only two starting features, or features arranged linearly, necessitates sampling multiple possible orientations of the starting fragments around the axis connecting these features. As each resulting conformer must then be expanded independently, this substantially increases computational cost. It may also decrease the accuracy of the initial fragment placement and reduce the overall success of the generation process.

2. The remaining features of the input 3D pharmacophore are subsequently partitioned into non-overlapping groups of spatially proximal features, hereafter referred to as subpharmacophores. This partitioning is performed by agglomerative clustering of the 3D coordinates of the pharmacophore features using a user-defined distance threshold. The threshold should be chosen to minimize the number of singleton clusters. Singleton features may lead to greater variability in the orientation of the added fragments, and the resulting conformations may hinder further growth because of steric clashes between subsequently added fragments and exclusion volumes. The default threshold is 3 Å.
3. A collection of starting fragments is screened against the chosen subpharmacophore. This pharmacophore screening is facilitated by the pmapper and psearch Python modules, which were developed previously[24]. The starting fragments are transformed into a database of 3D conformers generated by Conforge[25], and pharmacophore features are assigned by pmapper. The screening process yields matched fragments that are aligned with the initial subpharmacophore. In cases where the starting subpharmacophore features are aligned along the same axis or are in close proximity, the coordinates of the matched fragments are further sampled by rotating the fragments around this axis in 10-degree increments. This additional sampling is performed to accelerate the matching process for subsequent steps, as the coordinates of already matched fragments will remain fixed.
4. If exclusion volumes are provided within the model, in the subsequent step, the potential collisions of the matching fragments or molecules with exclusion volumes are assessed. A molecular conformation is discarded if any atom is found to be closer than a specified threshold to any exclusion volume (with a default value of 2.2 Å). If all conformations of a molecule are discarded, the entire molecule is eliminated.

5. From the list of matched fragments, only those with the smallest substructure are retained, while all superstructures are removed. This step aims to minimize redundancy and the number of molecules under consideration. For instance, if both benzene and toluene match the same pharmacophore features, benzene, possessing all necessary features, is retained while toluene is deemed redundant. This process ultimately results in the generation of molecules with the smallest structure that still match the maximum number of pharmacophore features.
6. To reduce redundancy among the molecules that pass the preceding steps, only representative conformers are retained. The conformers are clustered by RMSD using agglomerative clustering with complete linkage and a user-defined threshold (default: 0.25 Å). From each cluster, the conformer with the lowest average RMSD to all other conformers in that cluster is selected as the representative.
7. conformers exhibiting a root mean square deviation (RMSD) greater than a specified threshold (default value of 0.25 Å) relative to other conformers are retained to further reduce redundancy. This is achieved by agglomerative clustering and selection of representative conformer from each cluster
8. Compounds that successfully passed steps 3 through 6 are stored in an output database. This database includes all matching conformers of each molecule, along with corresponding lists of pharmacophore features that were tested and matched. It is mandated that starting fragments match all features, while partial matching is permitted in later stages. This limits the enumeration of potential structures and decrease computational expenses. The structure of an output database was made compatible with EasyDock[26] to be able to retrieve structures using the same scripts.
9. The main cycle commences with the selection of molecules designated for growth. Multiple molecules may be expanded concurrently, contingent upon the availability of CPU cores. To maintain a balance between exploration and exploitation, at least one molecule that was successfully generated in a previous run is selected, supplemented by a requisite number of unexpanded molecules to fully utilize all available cores. Consequently, the search aims to identify diverse molecules that match the maximum number of pharmacophore features from the very beginning, allowing the user to interrupt the search at any time once a satisfactory number of candidates has been generated. If not interrupted the search will run until all possible structures will be enumerated.
10. For the molecule designated for expansion (the parent molecule), the closest subpharmacophore, whose features have not yet been tested, is selected. This selection dictates the direction of the subsequent growth step.
11. Within the chosen parent molecule, the shortest distance (D) between any heavy atom with at least one hydrogen and any feature of the selected subpharmacophore

is determined. Subsequently, all heavy atoms of the parent molecule (which possess hydrogens) located within D + 2Å of any feature of the subpharmacophore are additionally designated as growth points.

12. All selected hydrogens are independently replaced. The fragments for these replacements are sourced from a CReM fragment database. To streamline the selection of fragments and enhance the efficiency of the overall process, additional filters are applied within the grow function of CReM:
    a. Fragments are selected only if they contain at least the required number of each feature type defined by the subpharmacophore, ensuring that all specified features can, in principle, be matched. For example, if a subpharmacophore includes a single hydrogen bond donor feature, it would be illogical to attempt to append a phenyl group, as it is not bearing this function and would not yield a match irrespective of a conformation.
    b. Fragments are further filtered based on their physicochemical properties to favor the generation of drug-like molecules. Users may specify maximum limits for molecular mass, topological surface area, the number of rotatable bonds, and lipophilicity. The first three parameters either increase or remain constant with the increasing size of the molecule, necessitating that fragments possess these parameters not exceeding the difference between the threshold value and the corresponding value for the parent molecule. Lipophilicity (logP) is mainly an additive characteristic; therefore, fragments are retained if their lipophilicity is less than the difference between the desired logP and that of the parent molecule, plus a correction factor of 0.5 to account for non-additivity.
13. The CReM grow function is applied, replacing the hydrogens at the selected atoms with the chosen fragments. The maximum size of the attached fragment is restricted to 12 heavy atoms, which is deemed sufficient to match adjacent subpharmacophores.
14. The generated molecules undergo user-defined physicochemical filtering (including molecular weight, topological polar surface area, the number of rotatable bonds, and lipophilicity) to retain only those that meet all specified criteria.
15. For the remaining molecules, all stereoisomers at undefined stereocenters are enumerated, with each stereoisomer treated as a distinct molecule in subsequent steps.
16. For each stereoisomer, a restricted generation of conformers is performed. The part of the molecule corresponding to the parent structure is fixed, maintaining the same coordinates as in the parent conformer, while other atoms are treated as flexible. Their coordinates are stochastically sampled using Conforge[25], selected for its

high speed and accuracy in generating biologically relevant conformations. If the parent molecule possesses multiple conformers, all are utilized as individual templates for conformer generation. This step is the most time-intensive. Other conformer generators, RDKit and Openbabel, can be also invoked instead of Conforge, but they are much slower and did not show advantages.

17. The generated conformers are evaluated to ensure they do not overlap with exclusion volumes. Conformers that pass this evaluation are then checked for matching pharmacophore features. This process is fast as it does not necessitate additional alignment; the coordinates of each conformer are already aligned with those of the template parent conformer from the previous step. Consequently, the closest distances between pharmacophore features of the same type in a conformer and the query are calculated. If this distance is less than a user-defined threshold, the corresponding query feature is designated as matched.
18. For each molecule, conformers that match the maximum number of query features are retained. Conformers of the same compound derived from different template conformers of the parent molecule are merged and filtered based on RMSD to minimize redundancy.
19. From the list of remaining molecules, those that are superstructures of other molecules in the list are removed (identically as performed in step 6). This action further reduces redundancy and preserves the smallest molecules that match the same pharmacophore features.
20. Molecules and their conformers that successfully pass all preceding steps are stored in the output database.
21. The procedure returns to step 8 and is reiterated until all molecules in the database have been expanded. The process can be interrupted at any time by the user, and generation can be automatically resumed upon restarting the program. If an output database already exists, generation will continue from the last checkpoint.

The current implementation supports only undirected pharmacophore features, owing to limitations of the pmapper and psearch tools. As a consequence, generated structures may occasionally contain incorrectly oriented hydrogen-bond donors or acceptors, potentially compromising the plausibility of the resulting structures. Nevertheless, this limitation is not expected to represent a major issue if exclusion volumes are employed. By constraining the placement and orientation of fragments, exclusion volumes should help preserve the appropriate orientation of hydrogen-bond donors and acceptors. This issue is specifically addressed in the study presented below.

Database of interchangeable fragments

To create the database for interchangeable fragments, we utilized molecules sourced from ChEMBL (version 22), standardized their structural representations, and eliminated those exhibiting structural alerts as defined by BMS, Dundee, Glaxo, Inpharmatica, or PAINS, following the implementation of Pat Walters (https://github.com/PatWalters/rd_filters). Subsequently, we conducted a comprehensive fragmentation of the molecules by severing single bonds. Each fragment, along with its corresponding contexts within a radius of 1 to 5, was systematically cataloged in a CReM fragment database. Furthermore, additional CReM fragment databases were constructed from subsets of molecules characterized by high synthetic accessibility, as assessed by the synthetic accessibility (SA) score proposed by Ertl and Schuffenhauer[20]. Prior research[23] has indicated a strong correlation between the SA score and the predictions generated by the retrosynthetic reaction prediction tool AiZynthFinder[27]. We created two supplementary databases by fragmenting molecules with SA scores of at most 2 and 2.5, respectively. Consequently, the database with an SA score of ≤ 2 was a subset of the database with an SA score of ≤ 2.5, which in turn was a subset of the full fragment database. The quantity of fragments and fragment/context pairs increased correspondingly (Table 1).

Selected fragments containing up to 12 heavy atoms could span pharmacophore features located as far as 13.8 Å from the attachment point. In practice, however, fragments of such size are relatively rare. The median maximum distance from the attachment point is 6.1 Å, and the third quartile is 7.1 Å, indicating that the effective reach of most fragments is substantially smaller. This may become problematic if a pharmacophore model comprises feature groups separated by a relatively large distance, as the available fragments may be insufficient to bridge these gaps. The limit of 12 heavy atoms was selected to constrain the size of the fragment databases, which increases rapidly with fragment size; however, the database can be extended to include larger fragments if necessary.

Table 1. Statistics of CReM fragment databases obtained from ChEMBL molecules with no structural alerts. Only fragments having one attachment points and consisting of at most 12 heavy atoms were counted in statistics, because only these fragments will be used for expansion of starting molecules.

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

full	818 174	131 754	235 980	453 461	981 275	1 836 212	2 707 664
SA ≤ 2.5 (SA2.5)	338 422	56 052	104 820	200 829	402 465	712 096	1 015 257
SA ≤ 2 (SA2)	67 970	17 152	34 044	61 454	111 830	187 317	267 906

Database of starting fragments

To prepare starting fragments we extracted all fragments from CReM SA2 database, capped all attachment points with hydrogens, converted molecules to canonical SMILES and removed duplicates. The resulting molecules were 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 H-bond donor and acceptor it was counted only once. This gives an estimate on the number of polar centers in a fragment.
- the number of rings is 1-3
- the number of fused ring systems 0-2
- the maximum size of a single ring is 7
- 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

This process yielded a total of 20,164 structures. For these structures, we enumerated all stereoisomers utilizing an RDKit script sourced from the repository at https://github.com/DrrDom/rdkit-scripts, as well as tautomers through the cxcalc utility provided by Chemaxon. Duplicates were checked and removed. Ultimately, we obtained 23,840 distinct structures, which will serve as the starting fragments for de novo generation. We generated up to 50 conformers for each molecule using Conforge and subsequently applied a filtering criterion based on RMSD of 0.5Å to reduce redundancy.

The described protocol was applied to select relatively conformationally rigid fragments containing a sufficient number of pharmacophore centers, including at least one hydrogen-bond donor or acceptor. The starting structures should remain relatively small to allow subsequent expansion while keeping the final molecules within Lipinski-compliant size limits. The selected fragments containing up to 15 heavy atoms are capable of spanning features separated by as much as 11.4 Å (the maximum pairwise feature

distance observed in this set). However, selecting an initial set of features that are too far apart is not recommended, as fragments capable of matching such distant features are relatively uncommon and the initial stage of the generation process may therefore fail. In this fragment set, the third quartile of the maximum pairwise feature distance is 8.0 Å and the median is 7.0 Å; these values may thus serve as practical upper bounds for the distance between selected starting features.

Targets selected for evaluation of de novo generation approach

For the validation of CReM-pharm, we selected targets from various protein families that have been previously utilized in other studies for the assessment of de novo generator tools[19, 28, 29] (Table 2). The ligands within the complexes were optimized using the MMFF method, and feature-based three-dimensional pharmacophores were generated using LigandScout [30]. All directed features were transformed into undirected representations (spheres), as CReM-pharm currently supports only undirected features (Figure 2). This approach is less restrictive and may enhance the quantity and diversity of the designed molecules. The pharmacophores comprised between four and ten features of varying types, including hydrogen bond donors and acceptors, hydrophobic, aromatic, and both positively and negatively charged elements. Some pharmacophore models, such as 2BTR, were compact, while others, like 7ONT and 8DV7, exhibited complexity with features positioned at considerable distances from one another, potentially posing additional challenges for de novo generation.

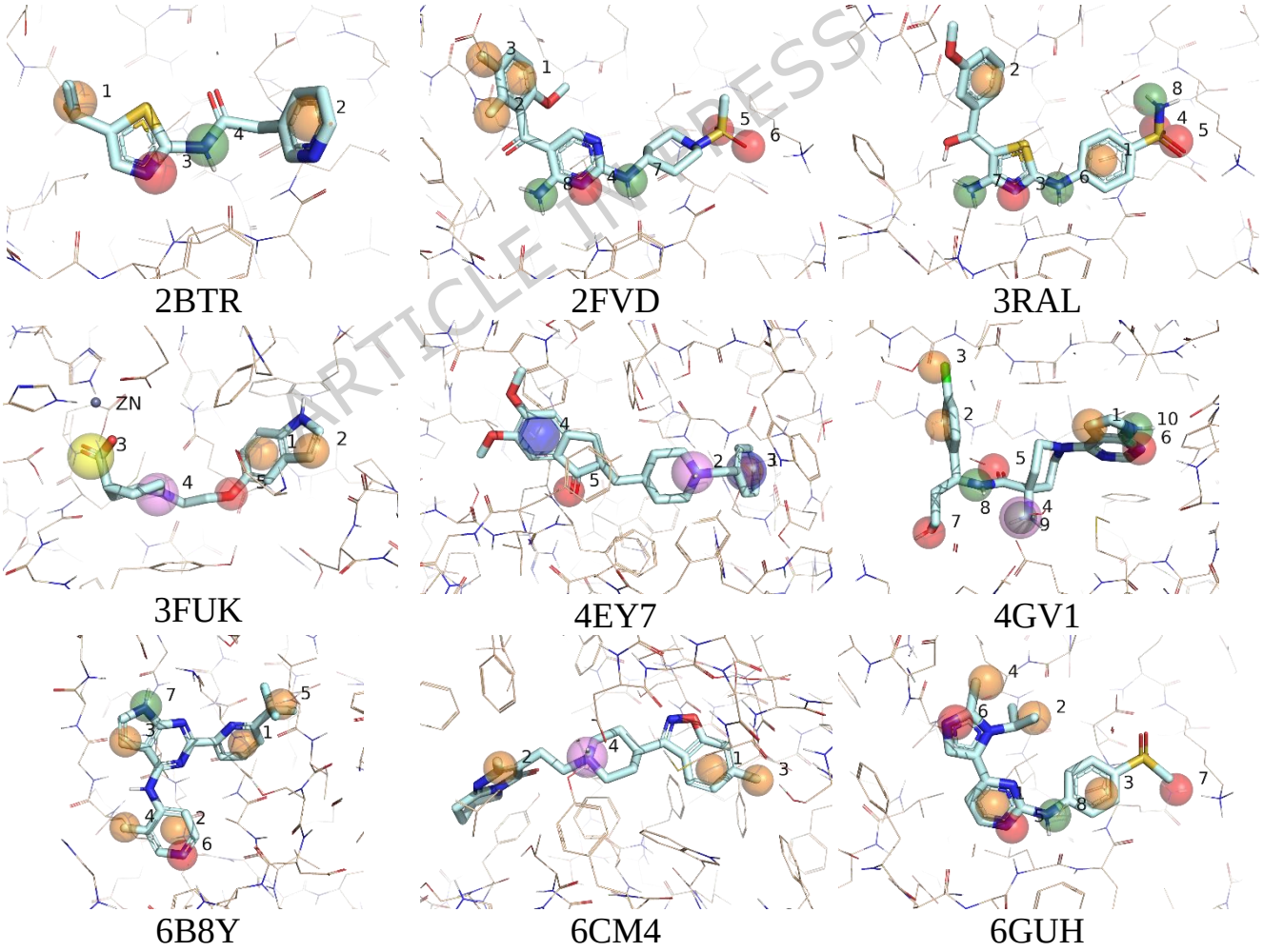
Table 2. Targets selected for validation of de novo generation based on 3D pharmacophores

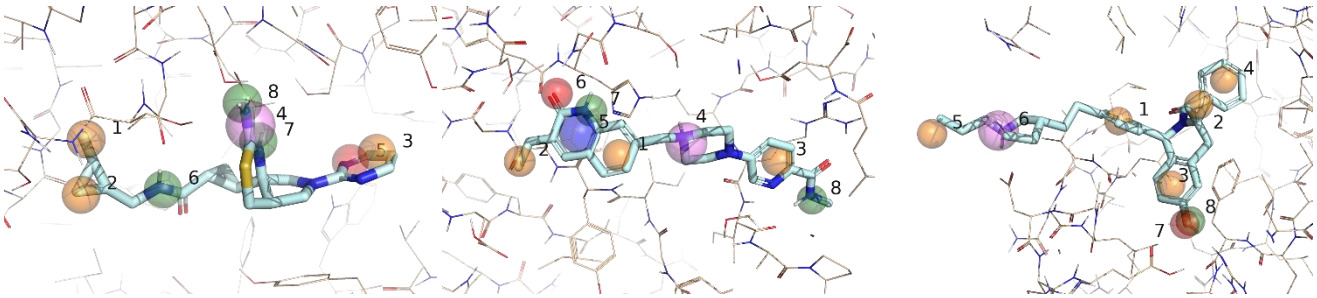
Protein target	Protein target name and family	Protein ChEMBL ID	PDB	Number of features	Type and the number of features*	Protein-ligand interaction pattern	Minimum PLIF similarity
CDK2	Cyclin-dependent kinase 2 (kinase)	301	3RAL	8	3A,3D,2H	Glu81 (HB donor); Leu83 (HB donor); Leu83 (HB acceptor)	0.6

			2BTR	4	1A,1D,2H	Leu83 (HB donor); Leu83 (HB acceptor)	1
			2FVD	8	3A,2D,3H	Glu81 (HB donor); Leu83 (HB donor); Leu83 (HB acceptor)	0.6
			6GUH	8	3A,1D,4H	Leu85 (HB donor); Leu85 (HB acceptor)	1
BACE1	Beta-secretase (protease)	4822	6UWP	8	1A,3D,1P,3H	Asp228 (HB donor); Asp32 (HB donor)	1
DRD2	Dopamine D2 receptor (GPCR)	217	6CM4	4	1P,3H	Asp114 (cationic)	1
ESR1	Estrogen receptor (nuclear receptor)	206	8DV7	8	1A,1D,1P,5H	Glu353 (HB donor); Arg394 (HB acceptor)	0.6
PARP1	Poly [ADP-ribose] polymerase 1 (transferase)	3105	7ONT	8	1A,2D,1P,3H,1a	Gly863 (HB donor); Gly863 (HB acceptor)	1
TGFR1	TGF-beta receptor type-1 (transferase)	4439	6B8Y	7	1A,1D,5H	Asp351 (HB donor)	1

AKT1	RAC-alpha serine/threonine-protein kinase (kinase)	4282	4GV1	10	3A,3D,1P,3H	Ala230 (HB acceptor); Glu228 (HB donor)	1
ACHE	Acetylcholinesterase (hydrolase)	220	4EY7	5	1A,1P,1H,2a	Tyr337 (cation-pi)	1
LTA4H	Leukotriene A-4 hydrolase (hydrolase)	4618	3FUK	5	1A,1N,1P,2H	Tyr267 (cation-pi); Zn701 (metal acceptor)	1

*A – H-bond acceptor, D – H-bond donor, H – hydrophobic, a – aromatic, P – positively charged center, N – negatively charged center.





6UWP 7ONT 8DV7

Figure 2. 3D structure-based pharmacophore models used for evaluation of de novo generation tools. Features of a smaller radius: green – H-bond donors, red – H-bond acceptors, orange – hydrophobic. Features of a larger radius: blue – aromatic, violet – positively charged, yellow – negatively charged.

Physicochemical property restrictions applied to generation of molecules

In order to generate drug-like compounds with potential for further optimization, we established the following criteria: molecular weight (MW) should be less than or equal to 450, lipophilicity (logP) should not exceed 4, topological polar surface area (TPSA) should be less than or equal to 120 Å², and the number of rotatable bonds (RTB) should be limited to 7. These criteria were uniformly applied across all tasks. It is noteworthy that not all native ligands from the selected protein-ligand complexes (Figure 2) conformed to these criteria. A significant outlier was the 8DV7 ligand, which exhibited a molecular mass of 495 Da, a logP of 6.4, and a TPSA of 147 Å². The substantial size of this ligand resulted in a pharmacophore characterized by distantly positioned features, which may pose additional challenges in the generation of compounds that both meet the established criteria and match all pharmacophore characteristics.

Post-processing of generated molecules

Molecules generated by fragment-based approaches may have unfavorable or unstable tautomeric forms. Therefore, before docking of generated molecules we generated a major stable tautomer by means of Chemaxon cxcalc utility. This affected approximately 20% of generated molecules.

Molecular docking

To estimate the ability of molecules to bind specific proteins we employed molecular docking as implemented in EasyDock[26]. Autodock Vina was used as a docking engine. Protonation of ligands was performed by Chemaxon cxcalc utility at pH 7.4.

Results and discussion

During further validation of the developed tool, we addressed the following questions:

1) the impact of various fragment libraries and context radii on the quantity of generated structures, as well as their corresponding docking and SA scores;
2) the influence of the selection of starting features located in different positions of the same pharmacophore model on the properties and similarity of the generated structures
3) the properties and similarity among structures generated using distinct pharmacophore models for the same protein
4) the properties and novelty of structures generated from pharmacophore models of proteins across different families, along with a comparative analysis of these structures against those produced by the state-of-the-art tool PGMG;
5) the reproducibility of conformations of the generated structures within the context of molecular docking;
6) the relationship between docking scores of generated compounds and the number of pharmacophore features matched;
7) the description of the web-application serving to enhance users' familiarity with the tool.

De novo generation of potential CDK2 ligands

Evaluation of influence of generation setting on generated structures and choosing optimal settings

In order to evaluate various configurations and select optimal parameters for subsequent studies, we employed the 3RAL pharmacophore model, which comprises eight features positioned in close proximity to one another. This arrangement facilitates multiple iterations for the expansion of initial fragments and presents a manageable challenge due to the proximity of the features. The initial features selected included two hydrogen bond donors and one hydrogen bond acceptor, which are indicative of binding interactions within the hinge region of the CDK2 kinase (specifically feature numbers 3, 6, and 7, as illustrated in Figure 2). We assessed two primary settings that significantly influence the generated structures: the context radii utilized for replacement (ranging

from 1 to 5) and the CReM fragments database, which was derived from the complete set of ChEMBL compounds as well as from subsets of compounds deemed more synthetically feasible (Table 1).

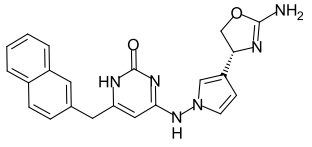
All simulations were conducted on computational nodes equipped with dual Intel Xeon CPU E5-2650 processors operating at 2.00 GHz, each supporting 16 threads and 48 GB of RAM. Given that the generation process is CPU-intensive, the overall performance is constrained by processor speed. Each task was allotted a maximum duration of 12 hours, after which it was terminated if not completed. For smaller radii (1-3), the generation processes predominantly reached the maximum time limit due to the substantial number of structures and their conformations generated, ranging from 800,000 to 2.8 million compounds. Conversely, for larger radii, the generation times varied from several minutes to a few hours, resulting in a reduced number of enumerated compounds (Table 3, Table S4). Notably, for the largest CReM database with a radius of 1, the execution of the program was halted after approximately 2 hours, leading to incomplete results for this particular run; however, several compounds that matched all specified features were nonetheless identified.

Table 3. The number of enumerated compounds for each set of settings, the number of compounds matched at least some features of a pharmacophore and running time of individual generation tasks.

PDB model	starting features	CReM DB	radius	total number of generated structures matching at least some features of pharmacophore model	total number of enumerated and embedded structures	generation time
3RAL	3,6,7	all	1	3375	894601	126m (interrupte
3RAL	3,6,7	all	2	12354	1882607	720m
3RAL	3,6,7	all	3	15833	823964	720m
3RAL	3,6,7	all	4	4049	139716	253m
3RAL	3,6,7	all	5	332	7500	7m 11s
3RAL	3,6,7	SA2.5	1	18153	2864314	720m
3RAL	3,6,7	SA2.5	2	19282	1812272	720m
3RAL	3,6,7	SA2.5	3	10195	428929	720m
3RAL	3,6,7	SA2.5	4	2544	77357	90m
3RAL	3,6,7	SA2.5	5	252	4518	4m 58s
3RAL	3,6,7	SA2	1	20563	1720462	720m
3RAL	3,6,7	SA2	2	16291	845327	720m
3RAL	3,6,7	SA2	3	2479	98708	79m 50s
3RAL	3,6,7	SA2	4	678	21015	11m 30s

3RAL	3,6,7	SA2	5	139	944	3m 25s
3RAL	1,6	SA2.5	3	5222	441756	720m
3RAL	4,5,8	SA2.5	3	1748	241622	161m 46s
2BTR	3,4	SA2.5	3	6561	1244015	720m
2FVD	4,7,8	SA2.5	3	177	28414	29m 22s
6GUH	1,5,8	SA2.5	3	26527	1276803	720m
3FUK	3,4	SA2.5	3	376	95904	102m 8s
4EY7	4,5	SA2.5	3	2585	179798	502m
4GV1	1,6,10	SA2.5	3	731	89710	81m 12s
6B8Y	3,7	SA2.5	3	6379	2075364	720m
6CM4	1,3	SA2.5	3	13469	450532	720m
6UWP	1,2,6	SA2.5	3	86	1669	2m 10s
7ONT	1,5,6,7	SA2.5	3	6565	252633	720m
8DV7	3,7,8	SA2.5	3	7098	717505	390m 42s

We conducted docking of all generated compounds against the 3RAL protein structure. From this analysis, we selected the top 100 compounds based on their docking scores for the computation of synthetic accessibility (SA) scores[20]. A notable trade-off was observed between docking scores and SA scores, exhibiting a trend consistent with our previous findings[23, 29]. Specifically, a smaller context radius yielded superior docking scores; however, this was accompanied by diminished SA scores (Figure 3). In alignment with prior observations, the utilization of CReM fragment databases derived from more synthetically feasible compounds resulted in the generation of molecules that were more synthetically accessible. To facilitate a comparative analysis, we computed SA scores for all compounds sourced from ChEMBL 33, which served as a reference dataset. The average SA score for all ChEMBL compounds was determined to be 3.05, with a median SA score of 2.78. In numerous instances, the average SA scores of the top-scoring generated compounds fell below these thresholds, thereby reinforcing the conclusion that our approach produces synthetically reasonable molecules. We identified the optimal parameters as the SA2.5 CReM fragment database and a radius of 3, as these settings yielded high docking scores while maintaining SA scores at a reasonably low level, below the average for the entire ChEMBL dataset. These parameters were subsequently applied to all further generations in this study.

	Top scored examples (docking / SA scores)
	

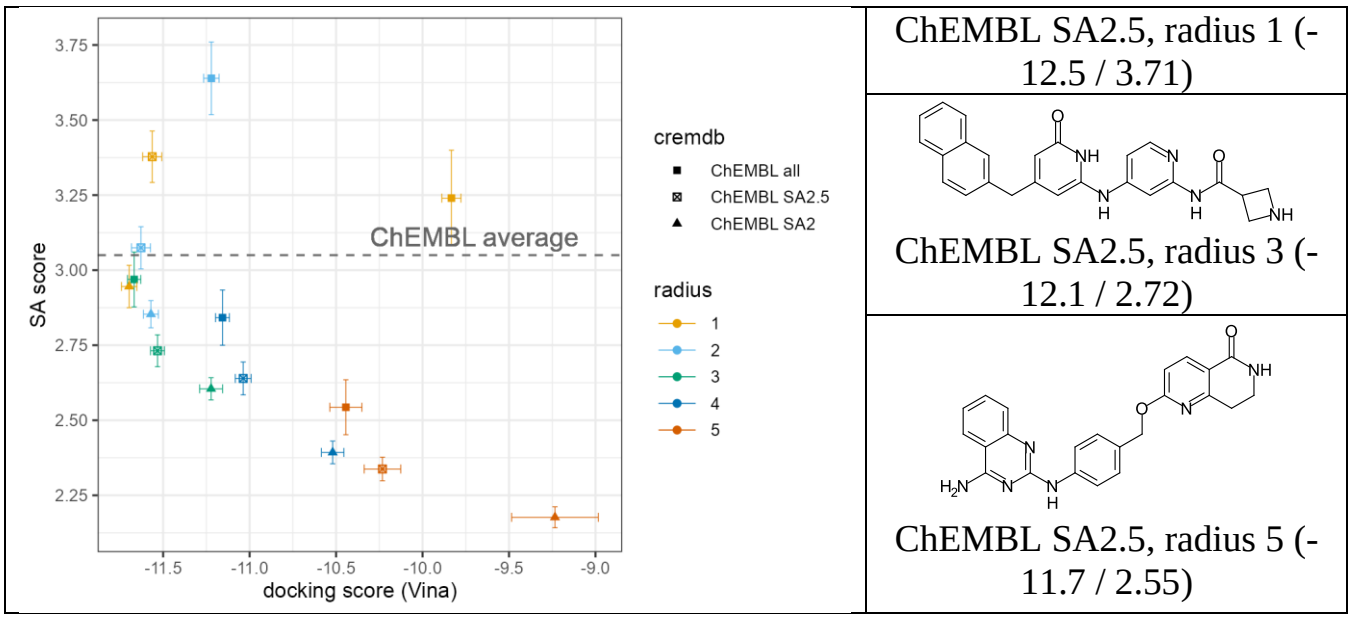


Figure 3. Average docking and SA scores for top 100 compounds selected by docking scores from every generation run. Error bars correspond to 95% confidence intervals calculate by t-test. Lower values of docking and SA scores indicate more favorable outcomes.

Study the influence of the starting features on generated compounds

To assess the impact of initial pharmacophore features, we conducted generations utilizing an H-bond donor and a hydrophobic feature (features 1 and 6, Figure 2), along with two H-bond acceptors and one H-bond donor that corresponded to a sulfonamide group situated at the terminal region of the 3RAL ligand (features 4, 5, and 8, Figure 2). The number of structures generated was lower in comparison to the previous generation that employed starting features 3, 6, and 7. Specifically, 5,222 and 1,748 structures were produced for the feature sets 1,6 and 4,5,8, respectively, whereas the feature set 3,6,7 yielded 10,195 structures (Table 3). In all cases, a subset of the compounds conformed to all pharmacophore features. Consequently, the selection of terminal feature groups for structure generation can still yield structures that align with all pharmacophore features, albeit with a reduced number of such structures and lower docking scores (Figure 4). Other physicochemical properties, including SA scores, were largely consistent across the different runs (Figure 4).

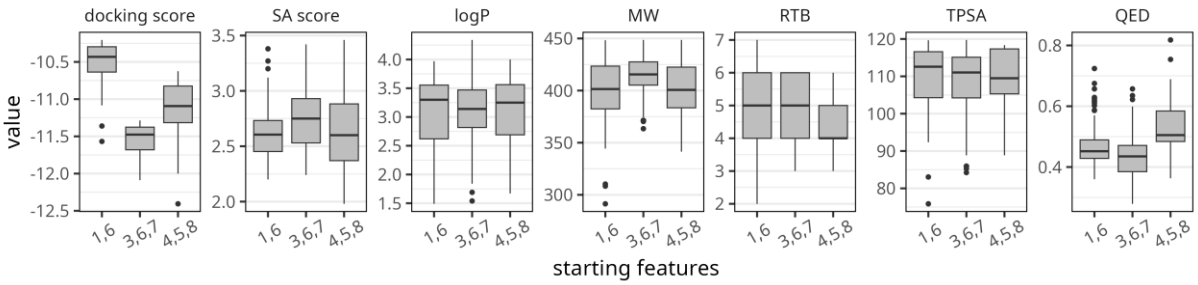


Figure 4. Distribution of physicochemical properties, docking and SA scores for top 100 compounds selected by best docking scores for individual runs started from different sets of pharmacophore features. Lower values of docking and SA scores indicate more favorable outcomes.

The generated structures exhibited low similarity across different runs, with only a limited number of compounds achieving a Tanimoto similarity exceeding 0.4, as determined by Morgan fingerprints of radius 2 (Figure 5). This finding suggests that employing distinct sets of initial features may enhance the diversity of the resultant structures set. Furthermore, the analysis indicates that the structures generated within a single run are also diverse. The structures formed several small clusters, which aligns with expectations given the utilization of a growing strategy; it is plausible that some of the top-scoring compounds share the same parent substructures.

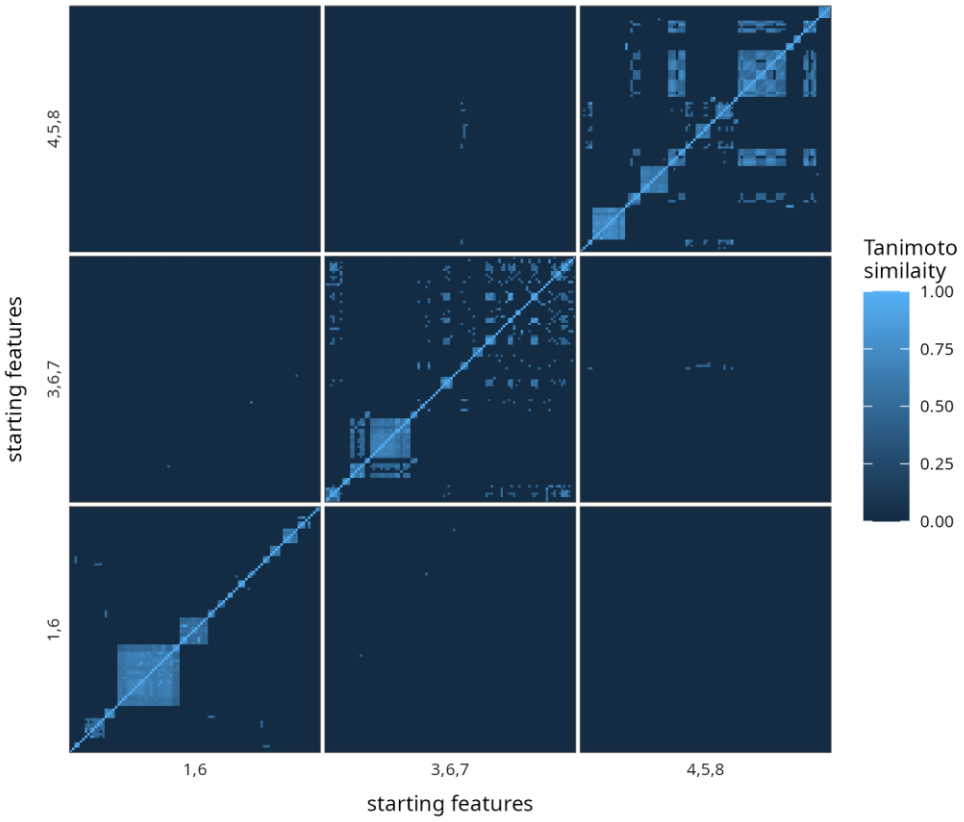


Figure 5. Pairwise similarity for top 100 compounds selected by docking scores for individual generation runs started from different sets of pharmacophore features of the 3RAL pharmacophore model. Similarities below 0.4 were nullified to better highlight intersection of analyzed sets of molecules.

Comparison of compounds generated for different pharmacophore models of CDK2

In addition to the 3RAL complex, we selected three additional CDK2-ligand complexes (2BTR, 2FVD, and 6GUH) to develop corresponding pharmacophore models. The complexity of these models varied; the 2BTR model comprised four features, while the other two models contained eight features each. For all models, we initially selected features that represented interactions with the hinge region: 3,4 features for 2BTR, 4,7,8 features for 2FVD, and 1,5,8 features for 6GUH (Figure 2). The number of structures generated differed significantly across the models: 6,561 compounds were produced for 2BTR, 177 for 2FVD, 26,527 for 6GUH, and 10,195 for 2BTR (Table 3). In the case of the 2FVD model, the maximum number of matched features achieved by the structures was six out of eight, while for the other models, a fraction of compounds matched all features.

For the analysis, we selected the top 100 compounds based on docking scores from each model. The compounds generated from the 2BTR pharmacophore model were smaller and exhibited lower docking scores compared to those generated from the 3RAL and 6GUH models (Figure 6). This outcome was anticipated, as the generation protocol was designed to produce minimal-sized structures that matched all pharmacophore features, and the 2BTR pharmacophore model was relatively small and simple. Consequently, the generated compounds had a lower molecular weight than those produced from the other larger models. Nevertheless, these compounds can be further decorated using basic CReM functions or by employing CReM-dock[29], which facilitates the growing of molecules within a binding site through molecular docking.

The compounds generated from the 2FVD pharmacophore model did not achieve the maximum number of matched features, resulting in a limited total number of structures generated. Thus, it was not unexpected that the docking scores of the top 100 compounds were poor (Figure 6). Overall, the physicochemical properties and synthetic accessibility (SA) scores of the generated compounds were comparable across the models, with SA scores generally not exceeding 3.05, which aligns with the average SA score for compounds in the ChEMBL database. These findings confirm the stability of the implemented generation protocol.

The overlap among the top 100 compounds generated from the different pharmacophore models of the CDK2 protein was minimal, with only a few compounds exhibiting a Tanimoto similarity greater than 0.4. Additionally, the diversity of the top-scoring compounds within individual runs was notably high (Figure S1). Therefore, employing various pharmacophore models for the same target can enhance both the quantity and diversity of generated molecules. The examples of top scored structures are given in Table 4.

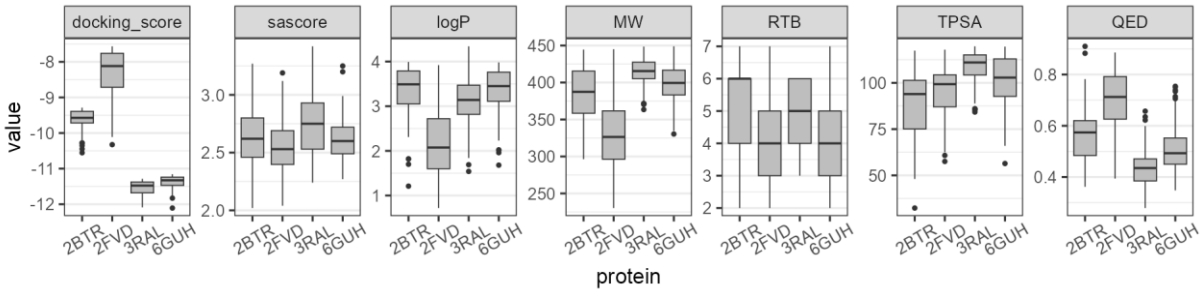


Figure 6. Distribution of physicochemical properties, docking and SA scores of compounds generated for different pharmacophore models of CDK2. Lower values of docking and SA scores indicate more favorable outcomes.

Comparison of de novo generation approaches CReM-pharm and PGMG using pharmacophore models for targets from different protein families

To achieve a more comprehensive validation of the proposed approach, we selected targets from various protein families with different levels of complexity (Table 2, Figure 2). For each target, we identified between two to four initial features that were positioned in close proximity to one another and situated centrally within the model (Table S1). Structures were generated utilizing the ChEMBL SA2.5 fragment database with a radius of 3, as these parameters were determined to be optimal for generating ligands targeting CDK2. Throughout all CReM-pharm runs, we imposed uniform constraints on the physicochemical properties, specifically: molecular weight (MW) $\leq$ 450, logP $\leq$ 4, topological polar surface area (TPSA) $\leq$ 120, and rotatable bonds (RTB) $\leq$ 7.

For comparative analysis, we selected PGMG [19] and PhoreGen [21]. PGMG is an approach that enumerates SMILES representations of compounds supposed to align with a specified three-dimensional pharmacophore query. All pharmacophore models were converted into the requisite format, and up to 20,000 molecules were requested for generation via the web application provided by the authors (https://www.csuligroup.com/PGMG). It is noteworthy that certain pharmacophore

models were deemed unsuitable for de novo generation due to inherent restrictions within PGMG. Specifically, the 4GV1 model featured a negatively charged feature, which is not acceptable by PGMG, while the 3FUK model comprised ten features, exceeding the maximum allowable limit of eight features in PGMG. Consequently, no compounds were generated for these particular models by PGMG. It should also be noted that PGMG approach does not use exclusion volumes and restrictions by physicochemical properties is not possible.

PhoreGen is a pharmacophore-guided framework for de novo 3D molecular generation based on diffusion modeling. The method begins from a noisy molecular representation and iteratively denoises both atomic coordinates and bond/graph information in an asynchronous manner, while message-passing operations maintain structural consistency throughout the generation process. A central feature of the approach is the incorporation of prior information on ligand–pharmacophore feature mapping into the denoising procedure, thereby steering generation toward molecules that satisfy the desired pharmacophore arrangement while preserving chemical validity and structural diversity. In contrast to CReM-pharm and PGMG, PhoreGen employs directed pharmacophore features. At the same time, similarly to CReM-pharm, it uses exclusion volumes, which may contribute to more accurate ligand placement. Explicit control of the physicochemical properties of the generated structures is not available in PhoreGen. For the present study, LigandScout pharmacophore models were converted into the required input format using in-house scripts. PhoreGen was then executed in five independent replicates with different random seeds, generating 2,000 structures per run, for a total of 10,000 structures for each pharmacophore model.

The number of structures generated by CReM-pharm was very variable from 86 structures for 6UWP to 26,527 for 6GUH (Table 3). This includes all generated structures matching at least some of pharmacophore features. For many pharmacophore models structures matching all pharmacophore features were successfully generated. In the extreme case of 6UWP the maximum number of 5 out of 8 features were matched by generated structures and no further extension was possible. This can be attributed to two primary factors. Firstly, we imposed a limitation on the number of rotatable bonds, capping it at seven; in some cases, further extensions resulted in the formation of structures that surpassed this threshold. Secondly, many compounds contained a hydroxyl group, which matched a hydrogen bond donor feature of the model. This hydroxyl group was utilized in subsequent extensions, leading to its conversion into either an ester or ether group. Consequently, this transformation resulted in the loss of the hydrogen bond donor characteristic in generated structures and their exclusion from further consideration.

The number of structures generated by PGMG did not exceed 10,000 in all cases (Table S2). The majority of them were very large and lipophilic. The simple search resulted in 13,256 molecules out of all 80,624 generated structures, which contained a polymethylene chain of at least 10 CH_2 groups. This can be explained by the absence of reasonable restrictions in the generative model, where the main issue could be the inability to consider exclusion volumes. The percentage of structures satisfying some drug-like criteria (MW <= 500, logP <=5 and TPSA <= 120) was relatively low, from 0.8% to 46.5% and 16.8% in average (Table S2). If we consider the criteria applied to CReM-pharm generations except the number of rotatable bonds (MW <= 450, logP <=4 and TPSA <= 120), the percentage of PGMG structures passed filters further decreased and varied from 0.1% to 24.2% with the average value 6% (Table S2). This indicates that PGMG by default generates mainly non-drug-like compounds.

Although PhoreGen generated 10,000 structures for each pharmacophore model, up to 37% of these were duplicates and were therefore removed (Table S5). The proportion of structures meeting broad drug-like criteria (MW $\leq$ 500, logP $\leq$ 5, and TPSA $\leq$ 120) substantially varied across pharmacophore models, ranging from 99.5% to 5.7%. As expected, the proportion satisfying the more stringent CReM-like criteria (MW $\leq$ 450, logP $\leq$ 4, and TPSA $\leq$ 120) was lower, ranging from 95.5% to 0.5% (Table S5). The lowest yields of drug-like structures were observed for the 3RAL, 4GV1, and 8DV7 models, which contain numerous features separated by relatively large distances. However, other structurally complex pharmacophore models performed more favorably. These findings indicate that the drug-likeness of PhoreGen-generated structures is strongly dependent on the underlying pharmacophore model and that the fraction of drug-like outputs is difficult to predict in advance.

We conducted an additional analysis to assess whether the conformers generated by CReM-pharm and PhoreGen are structurally plausible and whether molecules generated by PGMG are capable of matching the corresponding 3D pharmacophore models. To this end, up to 25 low-energy conformers were generated for each structure using ConfGen, followed by virtual screening in LigandScout. Screening was performed with pharmacophore models containing directed features as originally generated in LigandScout. In addition, all features were converted to undirected representations and the screening was repeated to evaluate the importance of feature directionality. All models included exclusion volumes defined previously in LigandScout. Screening was performed while allowing the omission of features, down to models retaining only a single feature, in order to count the number of structures matching a given number of pharmacophore features. Such exhaustive screening was computationally infeasible for the complete set of PGMG-generated structures because many of these molecules were very large, for example containing long polymethylene chains, and possessed numerous pharmacophore

features, resulting in a prohibitively large number of possible feature combinations. To reduce computational complexity, PGMG-generated structures with molecular masses above 600 Da were excluded in this analysis.

As expected, the fraction of PGMG-generated structures capable of matching at least two pharmacophore features was low for several models, including 2BTR, 2FVD, 3RAL, and 8DV7, whereas the corresponding reduction for CReM-pharm and PhoreGen was relatively modest (Figure 7). This likely reflects the absence of exclusion volumes in the PGMG generation pipeline, which may prevent even structures of moderate molecular weight from fitting the pharmacophore models. It was anticipated that both PGMG and PhoreGen would produce structures capable of matching all features of the pharmacophore models. For PhoreGen, at least a small fraction of such structures was observed for nearly all models, irrespective of whether directed or undirected pharmacophore representations were used. By contrast, PGMG-generated structures matched all features only for the 6CM4 four-point pharmacophore (Figure S3, Figure S4).

Structures generated by CReM-pharm generally matched more pharmacophore features than those generated by PGMG and, in many cases, fewer than those generated by PhoreGen. However, it should be emphasized that CReM-pharm was not designed to generate only structures matching the complete pharmacophore, as all molecules matching at least some feature subsets were retained in the output. A close agreement was observed between the distribution of CReM-pharm-generated structures matching a given number of features and the expected distribution (Figure 8). This suggests that the conformers produced within the CReM-pharm pipeline are structurally reasonable and can be reproduced by unconstrained conformer generation. Notably, the numbers of structures matching the directed and undirected pharmacophore models were very similar, indicating that the use of directed features may not be essential. This observation further suggests that exclusion volumes are sufficient to guide fragment orientation appropriately, such that the directions of hydrogen-bond donors and acceptors in the generated structures largely correspond to the expected arrangements.

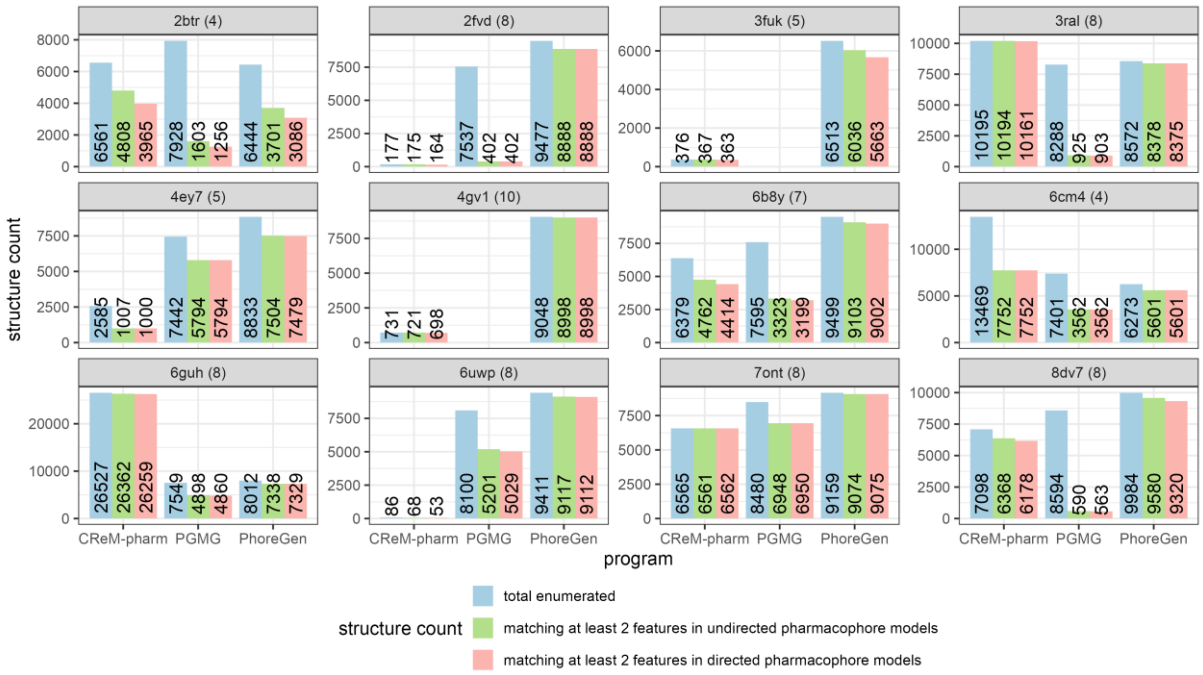


Figure 7. The number of structures generated by each tool and the number of structures which could match at least two features in the independent virtual screening runs using directed and undirected 3D pharmacophore models by means of LigandScout. The numbers in brackets near the model code is the number of features in the undirected models. There was one more feature in 4GV1 and 7ONT directed models due to double H-bond donor centers, which are reduced to a single feature in undirected models.

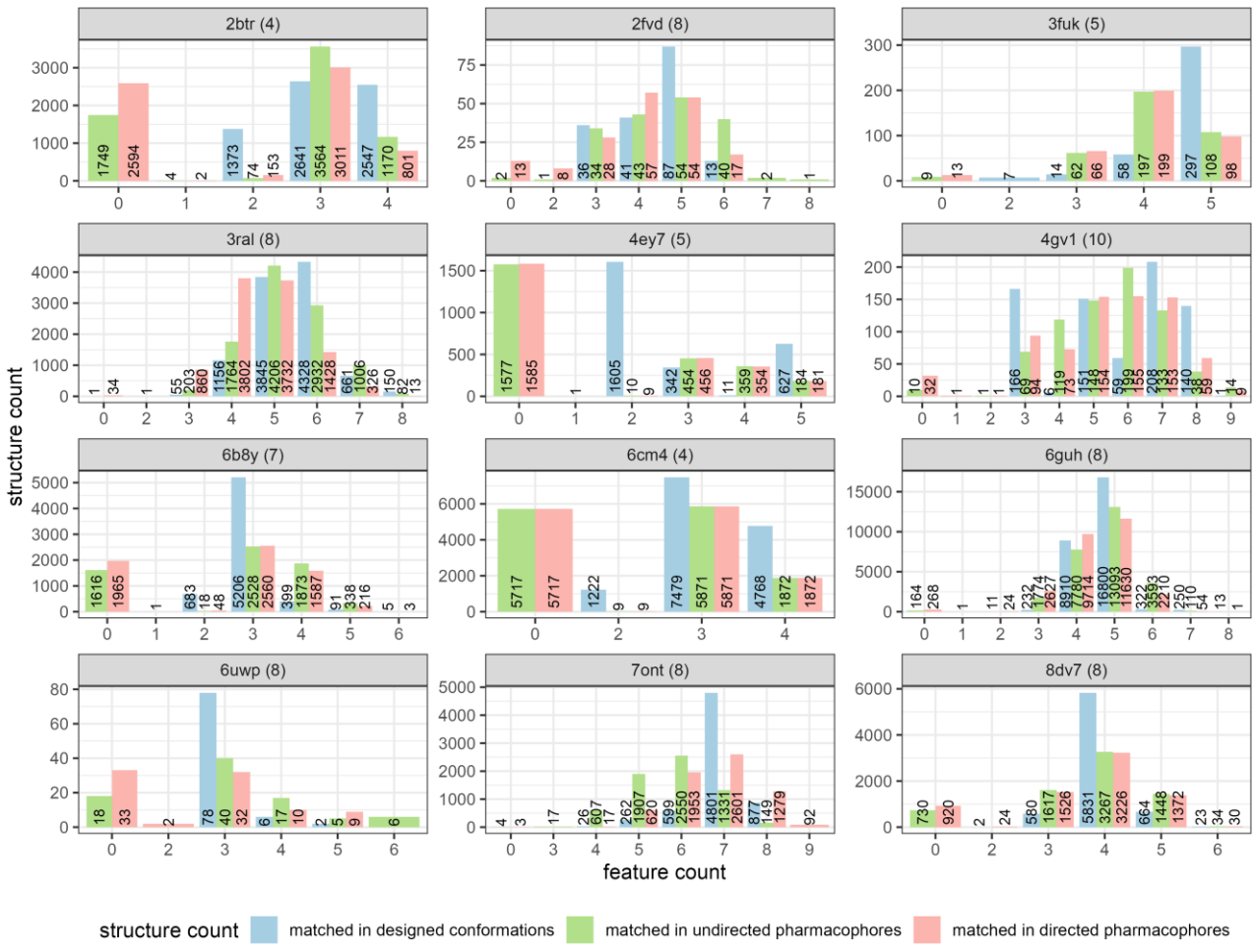


Figure 8. The number of CReM-pharm structures matching specific numbers of features in the originally designed conformers and in the course of independent virtual screening by directed and undirected pharmacophore models.

For the further analysis we selected PGMG and PhoreGen structures satisfying criteria aligning to CReM-pharm generations (MW <= 450, logP <=4 and TPSA <= 120, except of RTB constrain). Since PGMG generates structures with undefined configuration of stereocenters we enumerated up to 8 stereoisomers for each of them before docking. Full enumeration of all stereoisomers and their docking was infeasible because there were a large number of structures with a large number of chiral centers, that would produce too many stereoisomers. For the subsequent analysis we chose a stereoisomer with the best docking score as a representative one of each compound.

As a baseline we selected 50,000 random compounds from ZINC[31] which meet the same physicochemical criteria as CReM-pharm generated structures. This number of compounds was selected because computational resources required to dock these compounds will be comparable to computational resources spent on generation and docking of the largest set of molecules generated by CReM-pharm for 6GUH. Thus, this

is the output from conventional virtual screening pipeline with the same computational efforts as de novo generation. As an additional reference, we docked active ChEMBL33 compounds ($pK_i/pK_d/pIC_{50} \geq 6$) having molecular mass below 500 Da for each target.

For this analysis, the top 100 compounds ranked by docking score were selected for each target and each protocol, namely CReM-pharm, PGMG, PhoreGen, and ZINC. In several cases - specifically 2FVD, 7ONT, and 8DV7 for PGMG, and 8DV7 for PhoreGen—fewer than 100 structures satisfied the applied physicochemical criteria (Table S2, Table S5). For six targets (3RAL, 4GV1, 6CM4, 6GUH, 7ONT, and 8DV7), structures generated by CReM-pharm achieved better docking scores than those produced by the other generative methods. For five targets (2BTR, 2FVD, 4EY7, 6GUH, and 6UWP), PGMG-generated compounds yielded the best docking scores, whereas PhoreGen performed best for three targets (2FVD, 3FUK, and 6B8Y) (Figure 9). It should be noted that for two pharmacophore models, 3FUK and 4GV1, PGMG did not generate any structures because of either an excessive number of features (more than eight) or the presence of unsupported features, such as a negatively charged center. In general, the generated structures only rarely outperformed known ChEMBL actives in terms of docking score. By contrast, the top-ranked ZINC compounds showed docking scores comparable to, or better than, those of the best designed molecules.

SA scores of top generated structures by CReM-pharm were mainly within the narrow range 2.5-3, while for top Phoregen and PGMG structures SA scores varied substantially larger frequently exceeding 4 (Figure 9). In the majority of cases structures generated with CReM-pharm had better or comparable SA scores to PhoreGen and PGMG structures. SA scores for ChEMBL actives were mainly below 3 with the exception of 6UWP and, thus, were aligned with CReM-pharm and ZINC outputs. In the majority of cases structures generated with CReM-pharm also had favorable drug-like properties (Figure 9). These results show the promising nature of compounds generated by CReM-pharm. They have low SA scores, high drug-likeness and docking scores are comparable to ChEMBL actives and ZINC compounds (Table 4). It may also be noted that some structures generated by PhoreGen do not have required features, as an example is the structure having the best docking score for 6CM4 (Table S6), which does not have a center of a positive charge.

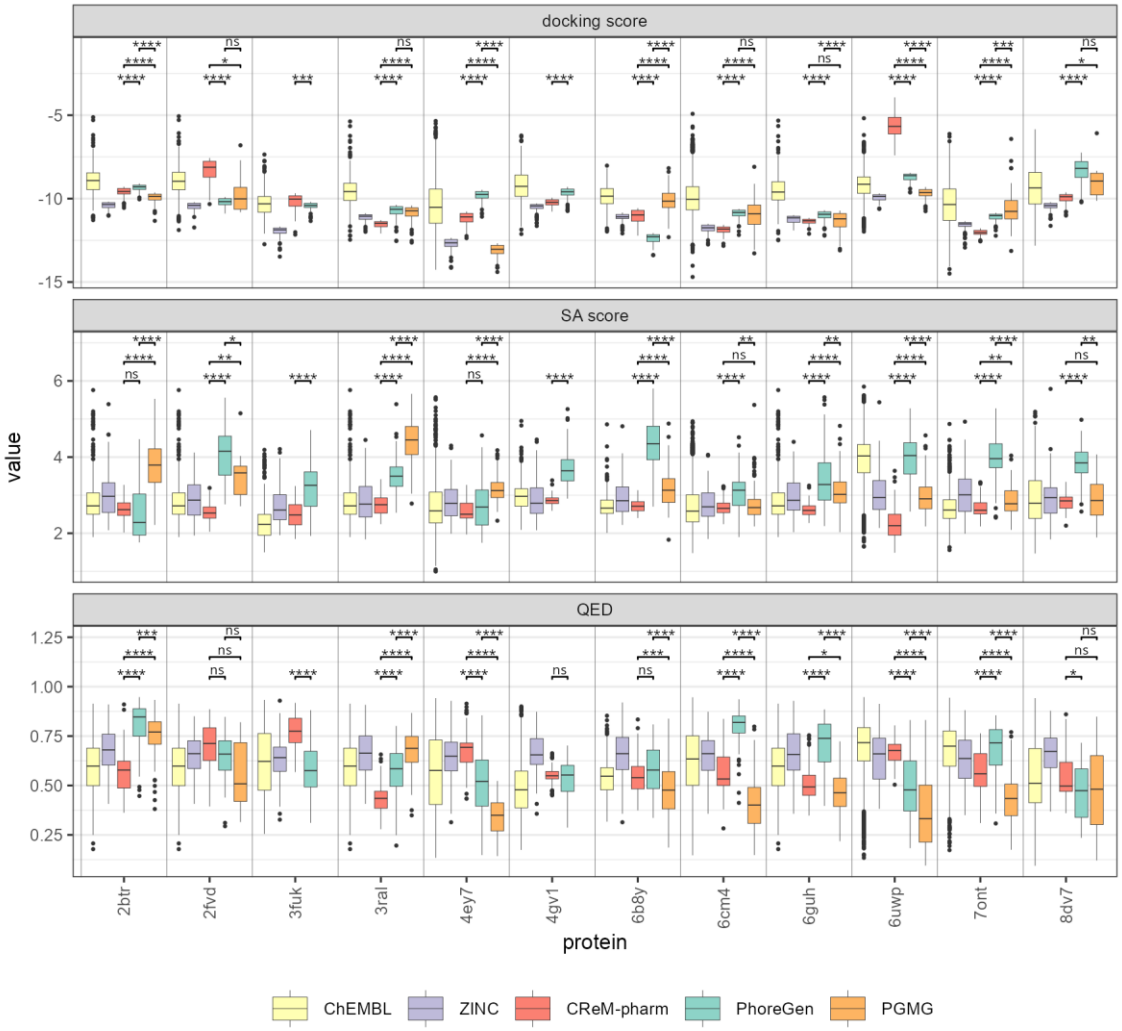
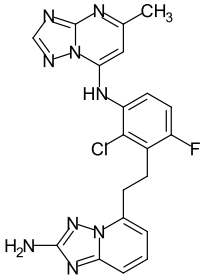
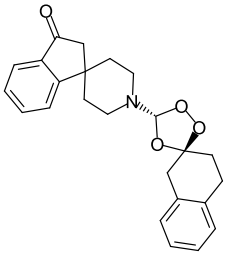
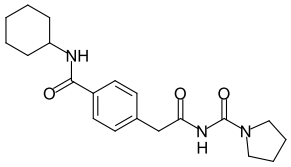
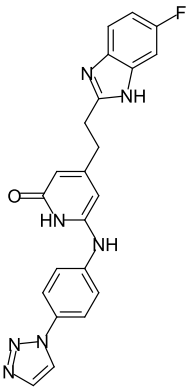
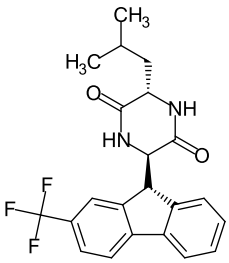
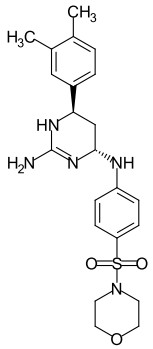
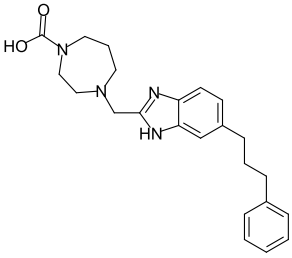
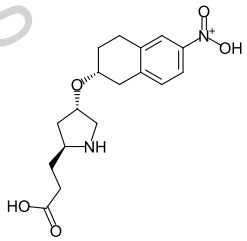
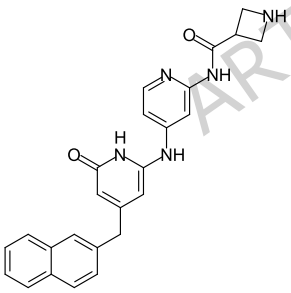
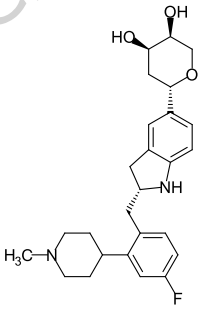
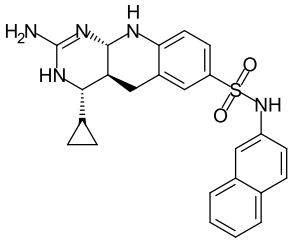
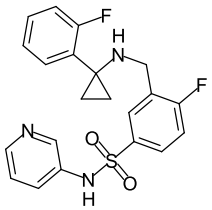
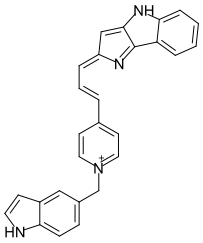
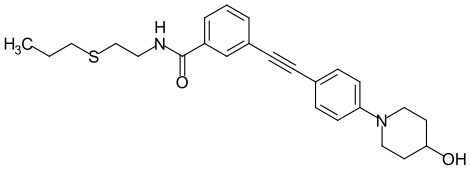
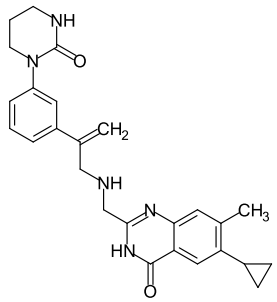
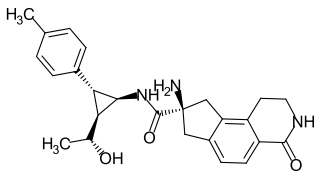
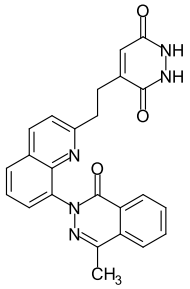
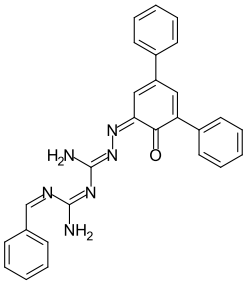
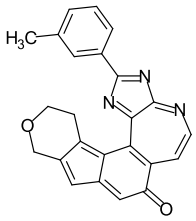
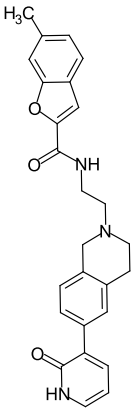
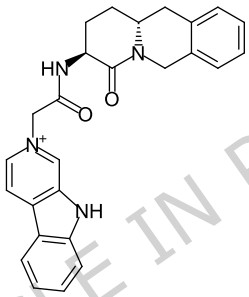
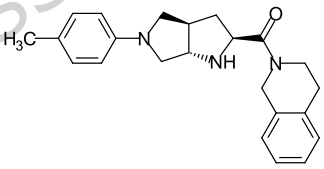
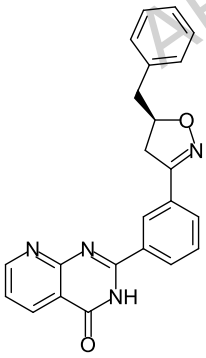
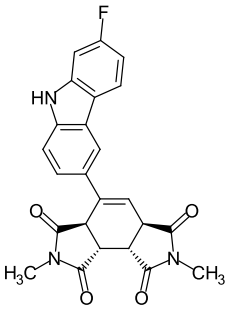
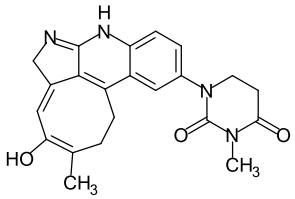
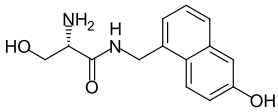
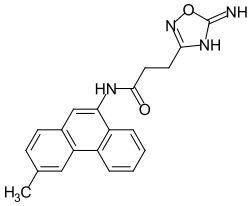
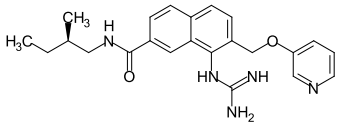


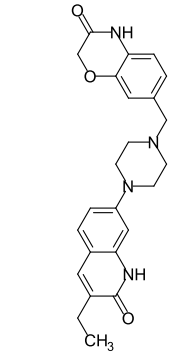
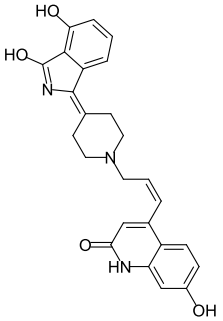
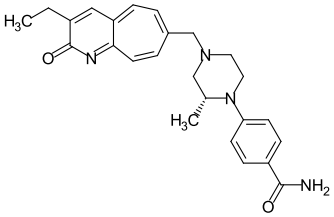
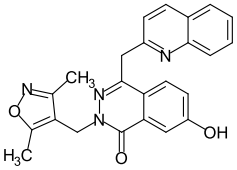
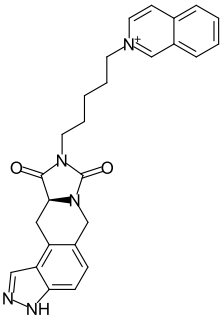
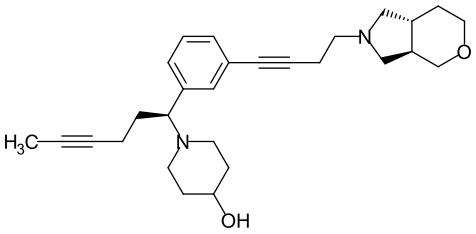
Figure 9. Distribution of docking and SA scores of top 100 compounds by docking scores for individual proteins and programs. PGMG and PhoreGen compounds were preliminary filtered by MW <= 450, logP <= 4 and TPSA <= 120 but not by rotatable bonds to align to physicochemical restrictions applied to CReM-pharm. Lower values of docking and SA scores indicate more favorable outcomes. Statistical significance was calculated by t-test: ns - not significant ($p \geq 0.05$), ** - $p < 0.01$, *** - $p < 0.001$, **** - $p < 0.0001$.

Table 4. Top scored compounds from generative tools CReM-pharm, PGMG and PhoreGen. The numbers designate docking scores and SA scores, for which lower values indicate better performance. Absent structures for PGMG are for pharmacophore models having more than 8 features or containing negatively charged centers which cannot be treated by PGMG.

	CReM-pharm	PGMG (physicochemical filters)	PhoreGen (physicochemical filters)

2BTR	 -10.6 / 3.10	 -11.3 / 4.54	 -10.0 / 2.01
2FVD	 -10.3 / 2.85	 -10.8 / 3.56	 -10.9 / 3.49
3FUK	 -12.2 / 2.21	-	 -11.3 / 4.06
3RAL	 -12.1 / 2.72	 -12.6 / 4.09	 -12.5 / 3.83
4EY7	 -12.4 / 2.31	 -14.4 / 3.41	 -11.1 / 2.46

4GV1	 -10.8 / 2.99	-	 -11.7 / 4.37
6B8Y	 -12,2 / 2.78	 -12.3 / 3.13	 -13.4 / 3.31
6CM4	 -12.8 / 2.54	 -13.3 / 3.53	 -12.2 / 3.38
6GUH	 -12.1 / 2.99	 -13.1 / 4.03	 -12.2 / 3.64
6UWP	 -7.4 / 2.48	 -10.8 / 2.72	 -9.63 / 3.05

7ONT	 -12.6 / 2.49	 -13.1 / 3.31	 -12.2 / 2.85
8DV7	 -11.0 / 2.61	 -10.1 / 3.46	 -11.9 / 4.83

Novelty of generated structures and coverage of known chemical space

To assess the novelty of the top 100 structures, we conducted a search for the most similar compounds within the whole ChEMBL33, as well as among the active compounds in ChEMBL33 corresponding to specific proteins, utilizing chemfp[32] with 2048-bit Morgan fingerprints of radius 2. The active compounds were sourced from ChEMBL33, defined as those with pK_i/pK_d/pIC_{50} values of 6 or higher. The top 100 structures identified by all programs exhibited low similarity to the established active space, with Tanimoto similarity coefficients primarily below 0.3 for PGMG and PhoreGen structures and below 0.4 for CReM-pharm (Figure 10). In contrast, similarity to the entire ChEMBL database was higher, reaching values between 0.4 and 0.6 for all programs. Notably, in five cases (6B8Y, 6CM4, 6GUH, 7ONT, and 8DV7), the PGMG generator successfully reproduced six known structures from ChEMBL, which was anticipated given that ChEMBL compounds were utilized in training the PGMG model. Additionally, CReM-pharm generated 14 known ChEMBL compounds for the case of 6UWP, which was also expected, as the structures generated for 6UWP matched a maximum of five features and were relatively small, with a molecular mass below 200 Da. Nevertheless, these previously untested compounds may be considered for repurposing studies or fragment-based screening aimed at identifying new ligands for the corresponding targets. In summary, the top-scoring structures produced by both programs can be regarded as predominantly novel.

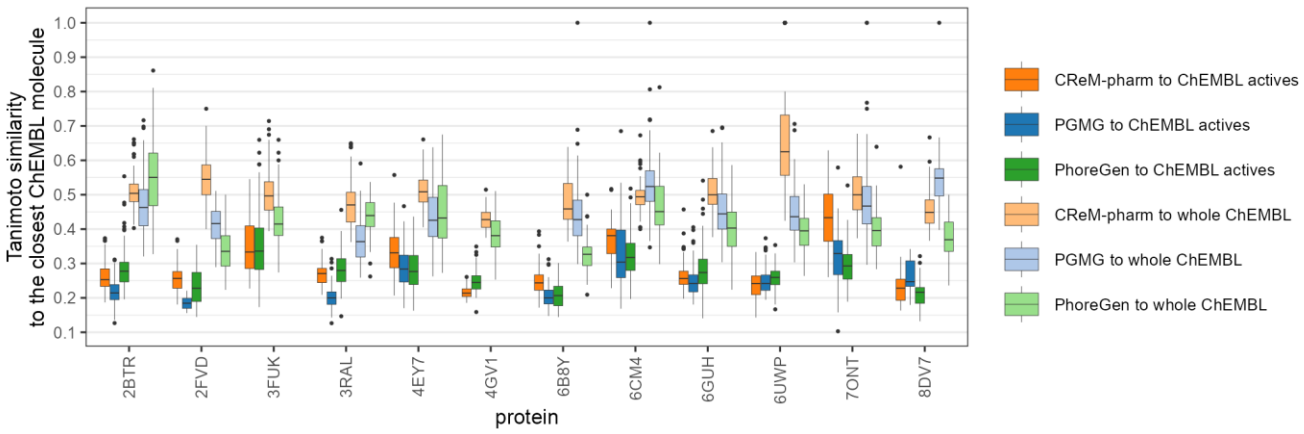


Figure 10. Tanimoto similarity (2048-bit Morgan fingerprints with radius 2) of top 100 generated compounds selected by docking scores to the closest neighbor from the whole ChEMBL database and from the set of known actives.

Although the similarity between the top-scoring generated structures and known active compounds was found to be low, it remains important to assess the overall coverage of the known active chemical space. This evaluation may demonstrate that the approach not only produces unique structures with high docking scores but also effectively explores relevant chemical space. To assess this capability, we conducted a search for the closest generated structures to each active compound from the corresponding set and count the number of ChEMBL actives that had generated structures with Tanimoto similarity exceeding the thresholds of 0.5 and 0.6. This analysis aims to illustrate the extent to which the generated compounds align with the known active chemical space (Table 5).

Our findings revealed that only 3 out of 10 targets had at least some known actives achieving a similarity score of 0.6 with PGMG compounds, whereas 8 out of 12 targets exhibited this level of similarity with CReM-pharm compounds and 10 out of 12 for PhoreGen. Furthermore, the number of known active compounds that displayed high similarity with the generated compounds was also greater for CReM-pharm and PhoreGen. These results suggest that CReM-pharm and PhoreGen, in general, more frequently encompasses previously explored chemical space while maintaining a high degree of novelty among the top-scoring generated compounds.

Table 5. The number of known actives from ChEMBL having the closest neighbors among generated compounds within the specified similarity

protein	CReM-pharm,	CReM-pharm,	PGMG,	PGMG,	PhoreGen,	PhoreGen,

	Tanimoto ≥ 0.5	Tanimoto ≥ 0.6	Tanimoto ≥ 0.5	Tanimoto ≥ 0.6	Tanimoto ≥ 0.5	Tanimoto ≥ 0.6
2BTR	4	0	1	0	26	12
2FVD	0	0	0	0	3	0
3FUK	45	7			42	15
3RAL	99	16	0	0	11	1
4EY7	21	2	35	26	14	0
4GV1	0	0			23	4
6B8Y	20	2	0	0	0	0
6CM4	252	28	98	8	144	27
6GUH	23	2	1	0	43	7
6UWP	0	0	2	0	0	0
7ONT	48	11	6	0	7	1
8DV7	53	7	3	1	11	0

Reproducibility of poses of structures generated by CReM-pharm in docking

It was of particular interest to analyze whether the top scored poses found by docking correspond to conformations generated during the de novo generation and 3D embedding within the CReM-pharm pipeline. It was found that the only a small portion of structures have poses within 2Å from the conformation created during de novo generation, up to 13.8%. The only exception was 7ONT pharmacophore model for which almost a half of molecules (48.6%) reproduced conformations found by de novo generation (Table 6). Poor reproducibility of poses may be explained by several factors. Molecules matching only a small number of total features are usually small and therefore are less restricted by the shape and the volume of the binding site and can change their pose in docking. However, some pharmacophore models have a small number of features (e.g. 6CM4 contains 4 features) which are span over a large distance and therefore generated structures is also large. While the binding site of 6CM4 is deep the molecules could flip or shift towards outside resulting in RMSD values greater than 2Å (Figure 11). Among pharmacophore models containing more than 5 features low RMSD were mainly achieved for more rigid molecules having a smaller number of rotatable bonds (Figure S2, 7ONT). However, in the case of CDK2 receptor (3RAL) despite of a large number of generated structures only a tiny portion of them could reproduce the poses. This may be explained by the wide binding site which may adopt ligand of different sizes (Figure 11). PARP1 binding site has another specific structure, which guides the ligands to dock in the poses similar to those from de novo generation. It has narrow and wide opposite sides. Therefore, the designed ligands cannot flip, because the part of a ligand matching a wider cavity of the binding site cannot fit into the smaller side of the site (Figure 11). Other possible reasons of difference in poses could be i) specific tautomeric forms used for

docking due to generation of stable tautomers during post-processing and ii) possible errors in protonation states, due to which some H-bond donors may disappear and as well as positively or negatively charged centers. Structures having reproduced docking poses do not outperform the remaining molecules by docking scores. However, these structures may have priority for selection for further stages, because consistency between pharmacophore and docking modeling may indicate a more reliable result.

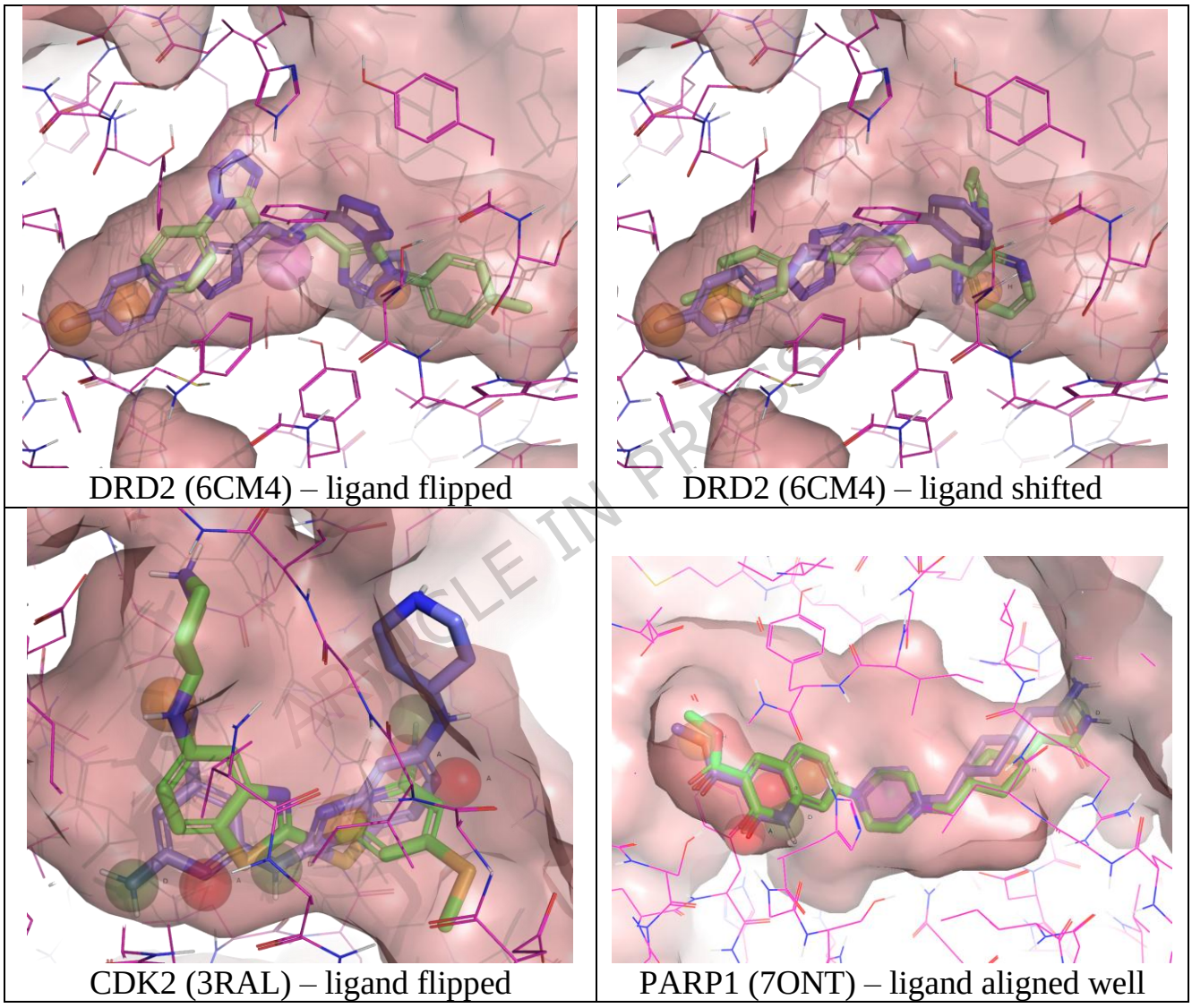


Figure 11. Poses of molecules from de novo pharmacophore generation (blue) and molecular docking (green).

Table 6. The total number of generated structures by CReM-pharm and the number of structures whose conformation embedded during de novo design was within 2Å from the top docking pose.

protein	number of molecules	number of molecules with RMSD ≤ 2Å	percentage of molecules with RMSD ≤ 2Å
2BTR	6561	523	8.0%
2FVD	177	0	0%
3FUK	376	4	1.1%
3RAL	10195	81	0.8%
4EY7	2585	60	2.3%
4GV1	731	51	7.0%
6B8Y	6379	13	0.2%
6CM4	13469	1290	9.6%
6GUH	26527	233	0.9%
6UWP	86	2	2.3%
7ONT	6565	3193	48.6%
8DV7	7098	981	13.8%

Dependence of docking scores of compounds on the number of matched features

We conducted an analysis of the distribution of docking scores among structures characterized by different numbers of matching pharmacophore features. Our findings generally indicate that structures matching a greater number of pharmacophore features tend to exhibit better docking scores, which may be attributable to their higher molecular weight (Figure 12). However, there exists a significant probability that structures failing to match all features may still yield high docking scores, thereby warranting their consideration for further evaluation. This observation suggests that it may not be essential for molecules to match all pharmacophore features in order to attain high docking scores. Consequently, it may be beneficial to optimize pharmacophore models prior to their generation by eliminating less critical features. This approach could potentially enhance the diversity of generated compounds while maintaining high docking score levels.

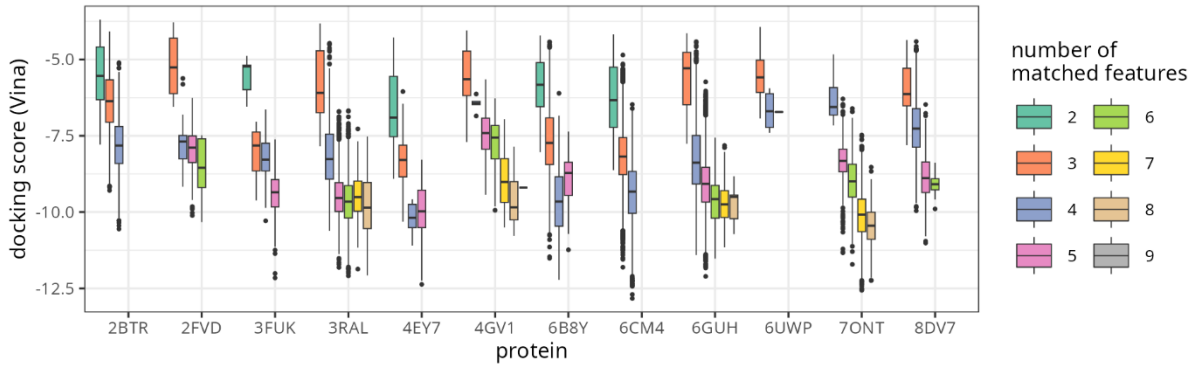


Figure 12. Distribution of docking scores of compounds generated by CReM-pharm and matching different number of features. Lower values of docking scores indicate more favorable outcomes.

Web-application

The software developed has been released as open-source under the GPL-3 license. To facilitate its evaluation, a web application has been created, accessible at the following link: https://crem-pharm.k8s.imtm.cz. This application employs simplified, hard-coded settings to allow for the rapid generation of structures based on a provided pharmacophore. Users can input a pharmacophore model in the form of an xyz-file or as a text, select the initial pharmacophore features, and impose restrictions on the physicochemical properties of the generated structures. Upon submission, the generation process is limited to a maximum duration of five minutes to prevent server overload. Given that the approach prioritizes exploration, this time frame is deemed sufficient to yield a diverse set of output structures. The generation process utilizes the highly restricted CReM fragment database, which was derived from fragments of ChEMBL33 molecules with SA score of 2 or lower. During the execution of the process, users will be able to monitor progress, including the number of structures that align with the pharmacophores and the number of structures pending expansion. The output will consist of a table displaying the generated structures, which can be downloaded as an sdf-file containing the 3D conformations of the structures that correspond to the pharmacophore model.

Conclusions

The iterative growth of molecular fragments to construct structures that align with features of pharmacophore models has demonstrated its utility across various case studies involving different protein families. The approach performs exhaustive de novo generation; however, its feasibility is enhanced by several key features: i) the selection of a more restricted fragment database and a larger context radius, which collectively improve the synthetic accessibility of the generated structures; ii) the retention of only those compounds that meet a minimal size requirement to align with pharmacophore model features, thereby preventing unnecessary expansions and enumeration; iii) the imposition of constraints on the physicochemical properties of the generated structures, which further improves their drug-likeness; and iv) the incorporation of the rapid and reliable conformer generator, Conforge, which greatly speeds up exploration of chemical space relatively to other open-source conformer generator. Importantly, the approach is

designed to maintain a balance between exploration and exploitation, allowing for interruption at any point once a desired number of structures has been generated, with the capability to automatically resume generation from the last checkpoint by executing the same command.

Although CReM-pharm incrementally designs conformers by fixing the parent part of the molecule, which could in principle lead to unreasonable geometries, additional validation by unconstrained conformer generation and pharmacophore-based virtual screening demonstrated that the resulting conformers are structurally plausible and largely consistent with the intended 3D pharmacophore models. Moreover, despite the fact that CReM-pharm employs only undirected pharmacophore features, the validation results were highly similar for models containing directed and undirected features. This observation suggests that explicitly directed features may not be essential for pharmacophore models that include exclusion volumes, as these appear sufficient to guide fragment placement and to ensure the appropriate orientation of hydrogen-bond donors and acceptors.

The synthetic accessibility of the generated molecules is primarily influenced by the selected CReM fragment database and context radius, rather than the composition or complexity of the pharmacophore model, and can be efficiently and predictably controlled by user settings. The synthetic accessibility (SA) scores of structures generated by CReM-pharm for various proteins and pharmacophore models were found to be within a similar range and predominantly below the average SA score of compounds in the ChEMBL database, indicating favorable synthetic accessibility of the generated structures.

It was observed that choosing the initial features of a pharmacophore model or employing different pharmacophore models of the same protein yields distinct generated structures with low mutual similarity. This variability can be leveraged to enhance the quantity and diversity of enumerated structures. The structures generated by CReM-pharm exhibited novelty in relation to the reference ChEMBL space, while also partially overlapping with previously explored active compound spaces for specific targets. Consequently, the tool is capable of generating not only novel compounds but also relevant ones.

CReM-pharm presents certain advantages over the current state-of-the-art approaches, PGMG and PhoreGen. Specifically, CReM-pharm allows for the restriction of drug-like compound generation, whereas PGMG predominantly produces non-drug-like compounds and PhoreGen likewise yields such compounds in some cases. The synthetic accessibility of compounds generated by PGMG and PhoreGen is not controllable and spans a broad range of values, often exceeding 4, necessitating more extensive post-processing of the structures produced and reducing their effective number.

The most time-intensive aspect of the developed approach is the embedding of conformers, an issue that will be addressed in future developments. The tool enhances the arsenal available to researchers in drug design and complements other tools built upon the CReM structure generator, such as CReM-dock[29], which can be utilized to further refine structures generated by CReM-pharm to optimize binding site occupancy and establish additional interactions not captured by a pharmacophore model, thereby improving the binding affinity of the designed ligands.

Contributions

Conceptualization – P.P.; formal analysis – A.D., J.P., P.P.; funding acquisition – P.P.; investigation – A.D., J.P., P.P.; methodology – A.D., P.P.; software – A.D., D.S., P.P.; validation – A.D., J.P., P.P., visualization – P.P.; writing – original draft, review & editing – P.P.

Competing interests

No competing interests are declared

Availability and requirements

Project name: CReM-pharm

Project home page: https://github.com/ci-lab-cz/crem-pharm

Operating system(s): platform independent

Programming language: Python

Other requirements: RDKit, CDPKit, CReM, networkx, dask, pmapper, psearch

License: GPL3

The source code of CReM-pharm is available at https://github.com/ci-lab-cz/crem-pharm. 3D pharmacophore models used in the study, structures of all generative runs of CReM-pharm, PGMG and PhoreGen as well as ZINC compounds and active compounds from ChEMBL are accessible at https://doi.org/10.5281/zenodo.17174627. Pre-compiled CReM fragments databases are available at https://doi.org/10.5281/zenodo.16909328.

Acknowledgement

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the INTER-EXCELLENCE II program (LUAUS23262) and through e-INFRA CZ (ID:90254). The authors also acknowledge Inte:Ligand for the license for LigandScout and Andrew Dalke for the license for chemfp.

References

(1) Wermuth CG. Pharmacophores: Historical Perspective and Viewpoint from a Medicinal Chemist. In *Pharmacophores and Pharmacophore Searches*, 2006; pp 1-13.
(2) Schuster D, Nashev LG, Kirchmair J, Laggner C, Wolber G, Langer T, Odermatt A (2008) Discovery of Nonsteroidal 17β-Hydroxysteroid Dehydrogenase 1 Inhibitors by Pharmacophore-Based Screening of Virtual Compound Libraries. J. Med. Chem. 51:4188-4199. 10.1021/jm800054h
(3) Hinsberger S, Hüsecken K, Groh M, Negri M, Haupenthal J, Hartmann RW (2013) Discovery of Novel Bacterial RNA Polymerase Inhibitors: Pharmacophore-Based Virtual Screening and Hit Optimization. J. Med. Chem. 56:8332-8338. 10.1021/jm400485e
(4) Krautscheid Y, Senning CJÅ, Sartori SB, Singewald N, Schuster D, Stuppner H (2014) Pharmacophore Modeling, Virtual Screening, and in Vitro Testing Reveal Haloperidol, Eprazinone, and Fenbutrazate as Neurokinin Receptors Ligands. J. Chem. Inf. Model. 54:1747-1757. 10.1021/ci500106z
(5) Polishchuk PG, Samoylenko GV, Khristova TM, Krysko OL, Kabanova TA, Kabanov VM, Kornylov AY, Klimchuk O, Langer T, Andronati SA, Kuz'min VE, Krysko AA, Varnek A (2015) Design, Virtual Screening, and Synthesis of Antagonists of αIIbβ3 as Antiplatelet Agents. J. Med. Chem. 58:7681-7694. http://dx.doi.org/10.1021/acs.jmedchem.5b00865
(6) Hoffmann T, Gastreich M (2019) The next level in chemical space navigation: going far beyond enumerable compound libraries. Drug Discovery Today 24:1148-1156. https://doi.org/10.1016/j.drudis.2019.02.013
(7) 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
(8) Schneider G, Fechner U (2005) Computer-based de novo design of drug-like molecules. Nature Reviews Drug Discovery 4:649-663. 10.1038/nrd1799
(9) Tschinke V, Cohen NC (1993) The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypotheses. J. Med. Chem. 36:3863-3870. 10.1021/jm00076a016
(10) Gillet VJ, Newell W, Mata P, Myatt G, Sike S, Zsoldos Z, Johnson AP (1994) SPROUT: Recent developments in the de novo design of molecules. J. Chem. Inf. Comput. Sci. 34:207-217. 10.1021/ci00017a027

(11) Huang Q, Li L-L, Yang S-Y (2010) PhDD: A new pharmacophore-based de novo design method of drug-like molecules combined with assessment of synthetic accessibility. J. Mol. Graphics Modell. 28:775-787. https://doi.org/10.1016/j.jmgm.2010.02.002
(12) Lippert T, Schulz-Gasch T, Roche O, Guba W, Rarey M (2011) De novo design by pharmacophore-based searches in fragment spaces. J. Comput.-Aided Mol. Des. 25:931-945. 10.1007/s10822-011-9473-6
(13) Skalic M, Jiménez J, Sabbadin D, De Fabritiis G (2019) Shape-Based Generative Modeling for de Novo Drug Design. J. Chem. Inf. Model. 59:1205-1214. 10.1021/acs.jcim.8b00706
(14) Skalic M, Sabbadin D, Sattarov B, Sciabola S, De Fabritiis G (2019) From Target to Drug: Generative Modeling for the Multimodal Structure-Based Ligand Design. Molecular Pharmaceutics 16:4282-4291. 10.1021/acs.molpharmaceut.9b00634
(15) Yoshimori A, Kawasaki E, Kanai C, Tasaka T (2020) Strategies for Design of Molecular Structures with a Desired Pharmacophore Using Deep Reinforcement Learning. Chemical and Pharmaceutical Bulletin 68:227-233. 10.1248/cpb.c19-00625
(16) Imrie F, Hadfield TE, Bradley AR, Deane CM (2021) Deep generative design with 3D pharmacophoric constraints. Chem. Sci. 12:14577-14589. 10.1039/d1sc02436a
(17) Hadfield TE, Imrie F, Merritt A, Birchall K, Deane CM (2022) Incorporating Target-Specific Pharmacophoric Information into Deep Generative Models for Fragment Elaboration. J. Chem. Inf. Model. 10.1021/acs.jcim.1c01311
(18) Radoux CJ, Olsson TSG, Pitt WR, Groom CR, Blundell TL (2016) Identifying Interactions that Determine Fragment Binding at Protein Hotspots. J. Med. Chem. 59:4314-4325. 10.1021/acs.jmedchem.5b01980
(19) Zhu H, Zhou R, Cao D, Tang J, Li M (2023) A pharmacophore-guided deep learning approach for bioactive molecular generation. Nature Communications 14:6234. 10.1038/s41467-023-41454-9
(20) 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
(21) Peng J, Yu J-L, Yang Z-B, Chen Y-T, Wei S-Q, Meng F-B, Wang Y-G, Huang X-T, Li G-B (2025) Pharmacophore-oriented 3D molecular generation toward efficient feature-customized drug discovery. Nature Computational Science 5:898-914. 10.1038/s43588-025-00850-5
(22) Polishchuk P (2020) CReM: chemically reasonable mutations framework for structure generation. J. Cheminf. 12:28. 10.1186/s13321-020-00431-w
(23) Polishchuk P (2020) Control of Synthetic Feasibility of Compounds Generated with CReM. J. Chem. Inf. Model. 60:6074-6080. 10.1021/acs.jcim.0c00792
(24) Kutlushina A, Khakimova A, Madzhidov T, Polishchuk P (2018) Ligand-Based Pharmacophore Modeling Using Novel 3D Pharmacophore Signatures. Molecules 23:3094. https://doi.org/10.3390/molecules23123094
(25) Seidel T, Permann C, Wieder O, Kohlbacher SM, Langer T (2023) High-Quality Conformer Generation with CONFORGE: Algorithm and Performance Assessment. J. Chem. Inf. Model. 63:5549-5570. 10.1021/acs.jcim.3c00563

(26) Minibaeva G, Ivanova A, Polishchuk P (2023) EasyDock: customizable and scalable docking tool. J. Cheminf. 15:102. 10.1186/s13321-023-00772-2
(27) Genheden S, Thakkar A, Chadimová V, Reymond J-L, Engkvist O, Bjerrum E (2020) AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. J. Cheminf. 12:70. 10.1186/s13321-020-00472-1
(28) 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
(29) Minibaeva G, Polishchuk P (2024) CReM-dock: de novo design of synthetically feasible compounds guided by molecular docking. ChemRxiv.
(30) Wolber G, Langer T (2005) LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. J. Chem. Inf. Model. 45:160-169. 10.1021/ci049885e
(31) 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
(32) Dalke A (2019) The chemfp project. J. Cheminf. 11:76. 10.1186/s13321-019-0398-8