

SOFTWARE

Open Access



# VSFlow: an open-source ligand-based virtual screening tool

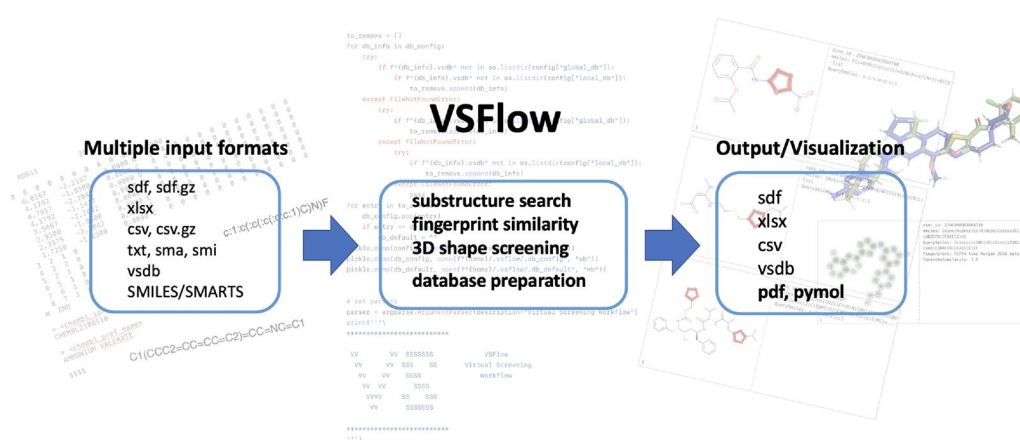
Sascha Jung<sup>1</sup>, Helge Vatheuer<sup>1</sup> and Paul Czodrowski<sup>2\*</sup>

## Abstract

Ligand-based virtual screening is a widespread method in modern drug design. It allows for a rapid screening of large compound databases in order to identify similar structures. Here we report an open-source command line tool which includes a substructure-, fingerprint- and shape-based virtual screening. Most of the implemented features fully rely on the RDKit cheminformatics framework. VSFlow accepts a wide range of input file formats and is highly customizable. Additionally, a quick visualization of the screening results as pdf and/or pymol file is supported.

**Keywords** Virtual screening, Substructure, Fingerprints, Shape, Python, RDKit

## Graphical Abstract



## Introduction

Virtual screening approaches are extensively used computational methods in modern drug discovery projects and they often replace or help to reduce more expensive and time-consuming high-throughput screenings nowadays [1]. There are two major categories of screening approaches: ligand-based and structure-based methods [2].

\*Correspondence:

Paul Czodrowski  
czodpaul@uni-mainz.de

<sup>1</sup> Department of Chemistry and Chemical Biology, TU Dortmund University, Otto-Hahn-Straße 6, 44227 Dortmund, Germany

<sup>2</sup> Department of Chemistry, Johannes Gutenberg University Mainz, Duesbergweg 10-14, 55128 Mainz, Germany



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Ligand-based methods are typically used if no X-ray structure of the target receptor is available. A single compound or a set of compounds known to bind to a specific target or to be active in a functional assay is typically used as the template to identify similar compounds in a large virtual database. In general, similarity can be evaluated on the basis of 2D and 3D molecular representations [3]. The classical 2D chemical similarity representations is based on molecular fingerprints (e.g. circular fingerprints, topological fingerprints, substructure fingerprints) transforming the molecular representation into a bit vector. The similarity between two vectors is then calculated with various similarity measures, most common is the Tanimoto coefficient. 3D similarity methods mainly consider the shape comparison of two molecules, typically extended by 3D pharmacophoric features, e.g. ROCS is considered the industry-leading commercial program for shape-based screenings [4].

Structure-based approaches, in most cases classical docking methods, are typically preferred if the target 3D structure information is available [5]. However, 2D ligand-based methods often require only a fraction of second for a single structure comparison task which allows to perform large screenings within a few hours even on a single, standard CPU. In contrast, docking methods are already considerably more resource demanding and time-consuming, not to mention more elaborated methods such as molecular dynamics simulations [6]. As a consequence, ligand-based methods are very attractive options for initial attempts to identify or filter relevant compounds in large and ultra-large virtual databases [7]. Furthermore, they are valuable tools to identify close analogues of known active compounds in a time efficient manner. In the last couple of years, several methods have been developed to screen non-enumerated chemical spaces up to  $10^{15}$  compounds and beyond in seconds to minutes on standard hardware [8]. The most elaborated technique for large space screening are chemical fragment spaces with corresponding connection rules, e.g. BioSolveIT's fragment spaces in connection with FTrees similarity implemented in their *infiniSee* software allows the screening of huge chemical spaces (e.g. Enamine REAL space) in seconds on standard hardware [9, 10].

There are many open-source web servers available for the screening of enumerated compound libraries using a variety of different structure- and ligand-based methods, recently reviewed by Singh et al. [11]. For example, many well-known databases such as ChEMBL, PubChem or ZINC include ligand-based similarity search functionalities with molecular fingerprints and/or substructure searches [12–14]. The web tool *SwissSimilarity* allows for the 2D fingerprint and 3D shape screening of common

public databases and compound libraries of most commercial vendors such as Enamine or ChemDiv [15, 16]. *Pharmit* additionally offers the possibility to screen large databases based on pharmacophore queries [17].

Several standalone tools focusing on enumerated 2D ligand-based screening approaches are available, most of which are commercial products [8]. Prominent examples are Schrödinger's *GPUSimilarity* integrated in their *LiveDesign* suite using a GPU-powered server in the background, *Arthor's NextMove* software with a SMARTS-based pattern matcher and Andrew Dalke's *chemfp* command line tool [18–20].

To the best of our knowledge, there is no open-source command line tool available which is similar to the *SwissSimilarity* or *Pharmit* web server and which allows for the comprehensive screening of databases and library files using different 2D and 3D ligand-based screening approaches, all combined in one tool.

In the following, we report an open-source command-line tool called “Virtual Screening WorkFlow” (*VSFlow*) written in Python and containing three different ligand-based screening modes. It relies on the open-source cheminformatics software *RDKit* [21]. *VSFlow* includes a substructure-based and fingerprint-based screening mode (2D) as well as a 3D shape-based screening mode (Fig. 1). Additionally, it possesses two tools for preparing and managing compound databases for virtual screening.

## Implementation

*VSFlow* is written in Python, is open-source and can be downloaded from <https://github.com/czodrowskilab/VSFlow>. It is licensed under the MIT license. As a prerequisite, a working installation of *Anaconda* or *Miniconda* is needed [22]. *VSFlow* including all dependencies can then be installed with the provided *yml* file as follows:

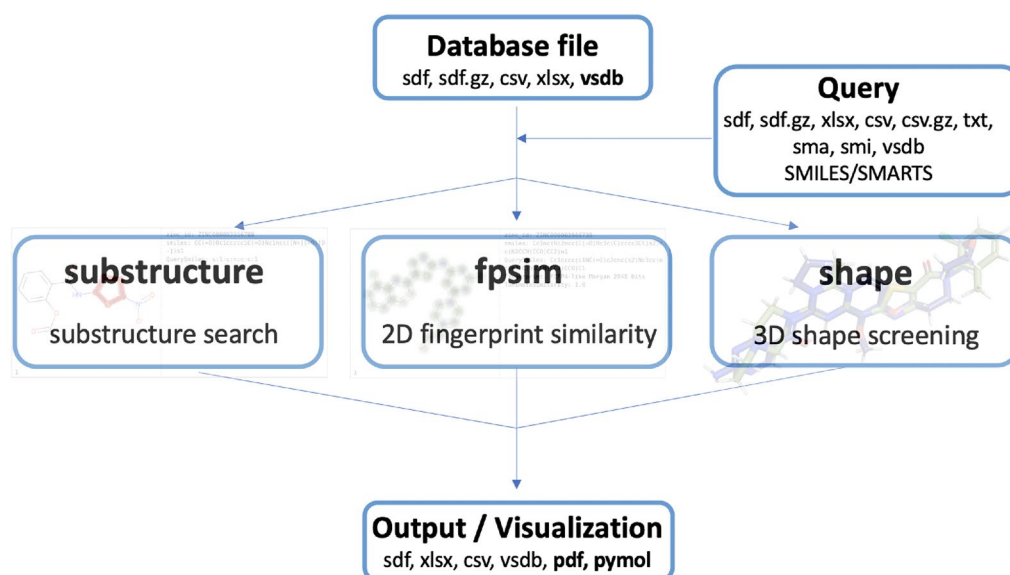
```
conda env create --quiet --force --file environment.yml
conda activate vsflow
pip install .
```

The Python dependencies are *rdkit*, *xlrd*, *xlswriter*, *pdfw*, *fpdf*, *pymol-open-source*, *molvs* and *matplotlib* [23, 24]. *VSFlow* requires Python version 3.7 or higher.

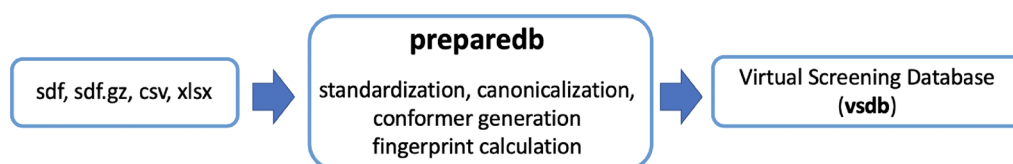
*VSFlow* includes 5 separate tools: *preparedb*, *substructure*, *fpsim*, *shape* and *managedb* (Fig. 1). All functionalities of *VSFlow* can also be run in parallel on multiple cores/threads. Parallelization is implemented via Python's built-in multiprocessing module.

### **preparedb: prepare databases**

*VSFlow* contains a tool to prepare compound libraries for virtual screening (*preparedb*). It allows for standardization of the molecules, generation of fingerprints



**Fig. 1** Different screening functionalities of VSFlow



**Fig. 2** Preparedb functionality of VSFlow: prepare compound libraries for virtual screening

and generation of multiple conformers (Fig. 2). The output file is a “virtual screening database” (.vsdb) file. The vsdb file is a Python pickle file containing all information in a special Python dictionary format which significantly enhances loading speed compared to SD files, particularly relevant for larger databases. Standardization is done on the basis of the MolVS rules and includes charge neutralization, salt removal and optionally tautomer canonicalization [23]. Fingerprints are generated with the RDKit chemistry framework. Conformers are generated with the RDKit ETKDGv3 method and optimized with the MMFF94 forcefield [25]. The following options are available:

- **standardize**: standardizes molecules, removes salts and associated charges
- **conformers**: generates multiple 3D conformers for database molecules
- **canonicalize**: adds the canonical tautomer to the database
- **fingerprint**: generates the respective fingerprint for each molecule and stores it in the database

It is also possible to directly download the PDB ligands and the chembl database and store them as vsdb databases, e.g.

```
vsflow preparedb -d pdb -s -f ecfp -r 2 -o pdb_ligands.vsdb
```

The above command will download all pdb ligands, standardize the molecules (-s argument), calculate the ECFP2 fingerprint (-f and -r argument) for every molecule and store it along with the molecule in the database (-o argument). You can repeat this for the ChEMBL database, e.g. with a different fingerprint:

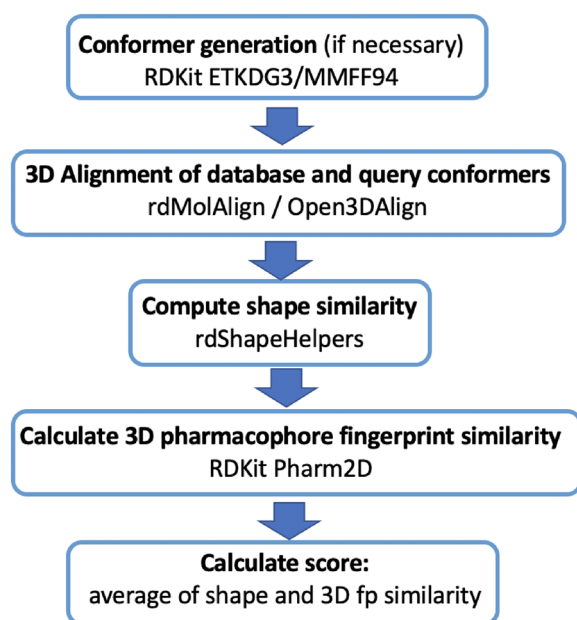
```
vsflow preparedb -d chembl -s -f fcfp -r 4 -o chembl_cpds.vsdb
```

#### **substructure: substructure search**

The substructure search (substructure) is performed based on the GetSubstructMatches() functionality available for RDKit Mol objects.

#### **fpsim: fingerprint similarity search**

The fingerprint generation relies on the RDKit framework. All fingerprints currently implemented in the



**Fig. 3** Different steps and RDKit functionalities which were combined to perform a screening based on pharmacophore alignment and shape similarity

RDKit (Morgan, RDKit, Topological Torsion and Atom Pairs fingerprint and MACCS keys) are supported and different similarity measures (Tanimoto, Tversky, Cosine, Dice, Sokal, Russel, Kulczynski and McConnaughey similarity) can be used.

#### shape: shape-based screening

Several functionalities of RDKit were combined to perform a screening based on a compounds' molecular shape (Fig. 3). First, generation of conformers (RDKit ETKDGv3 and MMFF94 forcefield) is done for 2D query structures. Conformers for database compounds can be generated using the preparedb functionality. Then, conformers of each query molecule are aligned to all conformers of each database molecule with the RDKit Open3DAlign functionality, either using MMFF94 force field parameters or Crippen atomic logP contributions (user-defined). In the next step, for every conformer pair the shape similarity is calculated (TanimotoDist, TverskyShape or ProtrudeDist) and the most similar conformer pair for every query/database molecule pair is selected (RDKit rdShapeHelpers). For the selected most similar conformer pair a 3D pharmacophore fingerprint is generated (RDKit Pharm2D) and the fingerprint similarity is calculated. By default, a combined score (combo score), the average of shape similarity and 3D fingerprint similarity, is used to rank the database molecules. The intended use case of the shape screening mode is to

screen a database of compounds with multiple conformers (prepared e.g. using the preparedb functionality of VSFlow) and to use a query ligand in a single, bioactive conformation, e.g. from the pdb database.

#### managedb: manage databases

The mode managedb is a convenience tool to update and manage compound databases which are integrated into VSFlow. A detailed description can be found in the VSFlow wiki [26].

#### Results and discussion

In the following section, the intended usage of VSFlow including some example commands are presented. A detailed description of the multiple possibilities to use VSFlow along with specific examples can be found in the VSFlow GitHub wiki [26].

In order to demonstrate the three main functionalities of VSFlow together with both its versatile input and output formats, we took the tyrosine-kinase inhibitor dasatinib as query molecule. As database, an SD file of the FDA-approved drugs generated from the ZINC database was used, comprising over 1600 molecules [14]. This database is also available in our GitHub repository.

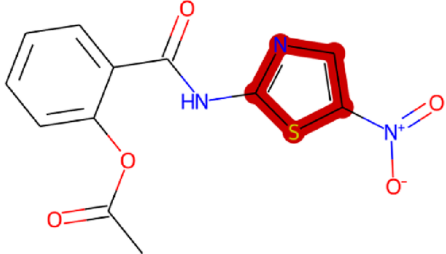
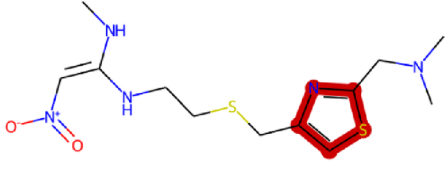
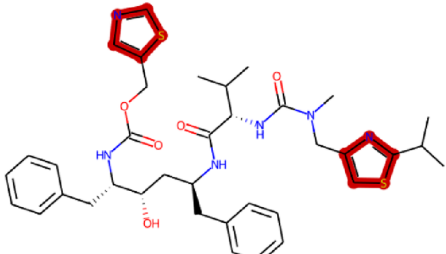
#### Substructure search

For the substructure search, a SMARTS representation of the thiazole function of dasatinib was taken as input to see how many other drugs might have that specific group. Besides the 36 hits (one of them, of course, dasatinib itself) in which the thiazole group was found, three molecules even have two thiazole groups, namely cefditoren, cobicistat and ritonavir. A pdf (supporting information) was generated displaying a table of the found hits with the 2D structures and the found substructure match highlighted in red as well as the information of the hit (e.g. ID, SMILES, Fig. 4). It should be mentioned that a pdf can only be generated in addition to an sdf, excel or csv file.

```
vsflow substructure -sma "s:1:c:n:c:c:1" -d fda.sdf
-o substructure.sdf --pdf
```

#### Fingerprint similarity

For the fingerprint similarity function fpsim, a SMILES input of the molecule was used with default parameters, i. e. an FCFP4-like Morgan 2048 bits of radius 2 for which the Tanimoto coefficient was calculated. A pdf file was selected as output format as well as an Excel file. The simmap parameter will generate a similarity map that visualizes the contribution of the specific atoms to

<p>1</p> 	<p>zinc_id: ZINC000003956788  smiles: <chem>CC(=O)Oc1ccccc1C(=O)Nc1ccc([N+](=O)[O-])s1</chem>  QuerySmiles: <chem>s:1:c:n:c:c:1</chem></p>
<p>2</p> 	<p>zinc_id: ZINC000001530737  smiles: <chem>CN/C(=C/[N+](=O)[O-])NCCSCc1csc(CN(C)C)n1</chem>  QuerySmiles: <chem>s:1:c:n:c:c:1</chem></p>
<p>3</p> 	<p>zinc_id: ZINC000003944422  smiles: <chem>CC(C)c1nc(CN(C)C(=O)N[C@H](C(=O)N[C@@H](Cc2ccccc2)C[C@H](O)[C@H](Cc2ccccc2)NC(=O)N(C)C)C)C</chem>  QuerySmiles: <chem>s:1:c:n:c:c:1</chem></p>

**Fig. 4** Exemplary page of the pdf file generated after substructure search. The left column shows the hits with the substructure matches highlighted in red, the right column the ID of the hits as well as the SMILES and the query SMARTS

the similarity between the molecules in the database and dasatinib (Fig. 5) [27].

```
vsflow fpsim -d fda.sdf -o fingerprint.xlsx --pdf --simmap
-smi "Cc1cccc(c1NC(=O)c2cnc(s2)Nc3cc(nc(n3)C)N4CCN(CC4)CCO)C1"
```

### Shape similarity

In order to perform a shape screening, a new database, containing a maximum of 20 conformers, was generated with the `-c` argument because the original database only had one conformer per compound.

```
vsflow preparedb -i fda.sdf -np 8 -c 20
-o fda_multiple_confs.vsd
```

Since that is a rather resource-intensive step, multiprocessing was carried out with the help of the `-np` parameter. The following shape search, also multiprocessed, was then done with the previously prepared vsdb pickle file using the instance coordinates of dasatinib in complex with tyrosine protein kinase ABL1 (PDB: 2GQG).

```
vsflow shape -i 2gqg_C_1N1.sdf -np 8
-d fda_multiple_confs.vsd -o shape.sdf --pymol
```

More than half of the top 10 hits were other kinase inhibitors. By default, the shape functionality creates two sd files, one with the query molecule (`shape_1_query.sdf`) and the found hits as a second file (`shape_1.sdf`). Additionally, a PyMOL session file was generated (`--pymol` parameter) so that the aligned structures could be visually inspected directly (Fig. 6).

The RMSD spread of the conformer generation process (ETKDG3 followed by MMFF94 minimization) is given in Fig. 7). It shows a clear upwards trend: the more rotatable bonds, the larger the RMSD.

### Runtime performance

To give the user an idea of the expected runtime performance, we performed a substructure and 2D similarity search in the pdb and ChEMBL28 database [12, 28]. We performed the searches on up-to-date standard notebook hardware, namely a 12th Gen Intel(R) Core(TM) i7-12700 H with 2.70 GHz and 20 cores and 32GB RAM running Windows 11. To get an idea of the performance on your own system, you may execute the following commands accordingly. Both ChEMBL and pdb database can be downloaded and prepared directly within VSFlow:

```
vsflow preparedb -d pdb -o pdb_ecfp4 -f ecfp -np 6
vsflow preparedb -d chembl -o chembl_ecfp4 -f ecfp -np 6
```

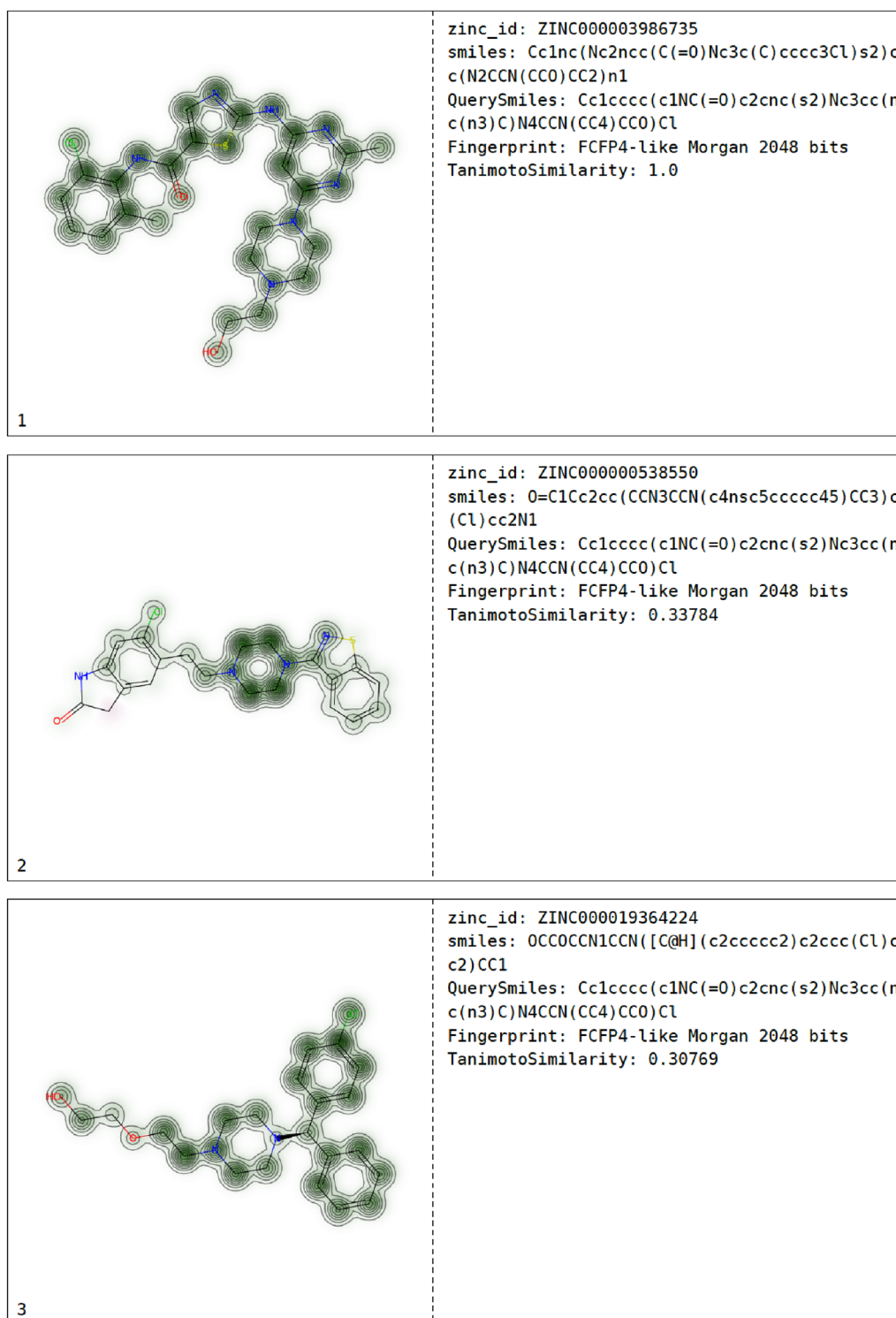
With the above calls, the pdb and chembl databases are downloaded into VSFlow and 2048-bit ECFP4 fingerprints are generated for each compound and stored within the output vsdb file. Preparation of the pdb database (containing 36,796 unique compounds at 22/05/2022) took 11 s on our system, preparation of the chembl28 database (2066377 compounds) took 511 s. Now, we performed a substructure and similarity screening using a SMILES as query, once in single-core mode and once on 6 cores:

```
vsflow substructure -smi "C(C1=CC=CC=C1)C1=CC=NC=C1" -d pdb_ecfp4.vsd
-o pdb_subsearch.sdf
vsflow substructure -smi "C(C1=CC=CC=C1)C1=CC=NC=C1" -d pdb_ecfp4.vsd
-o pdb_subsearch.sdf -np 6
vsflow fpsim -smi "C(C1=CC=CC=C1)C1=CC=NC=C1" -d pdb_ecfp4.vsd
-o pdb_fpsim.sdf -f from_db
vsflow fpsim -smi "C(C1=CC=CC=C1)C1=CC=NC=C1" -d pdb_ecfp4.vsd
-o pdb_fpsim.sdf -f from_db -np 6
vsflow substructure -smi "C(C1=CC=CC=C1)C1=CC=NC=C1" -d chembl_ecfp4.vsd
-o chembl_subsearch.sdf
vsflow substructure -smi "C(C1=CC=CC=C1)C1=CC=NC=C1" -d chembl_ecfp4.vsd
-o chembl_subsearch.sdf -np 6
vsflow fpsim -smi "C(C1=CC=CC=C1)C1=CC=NC=C1" -d chembl_ecfp4.vsd
-o chembl_fpsim.sdf -f from_db
vsflow fpsim -smi "C(C1=CC=CC=C1)C1=CC=NC=C1" -d chembl_ecfp4.vsd
-o chembl_fpsim.sdf -f from_db -np 6
```

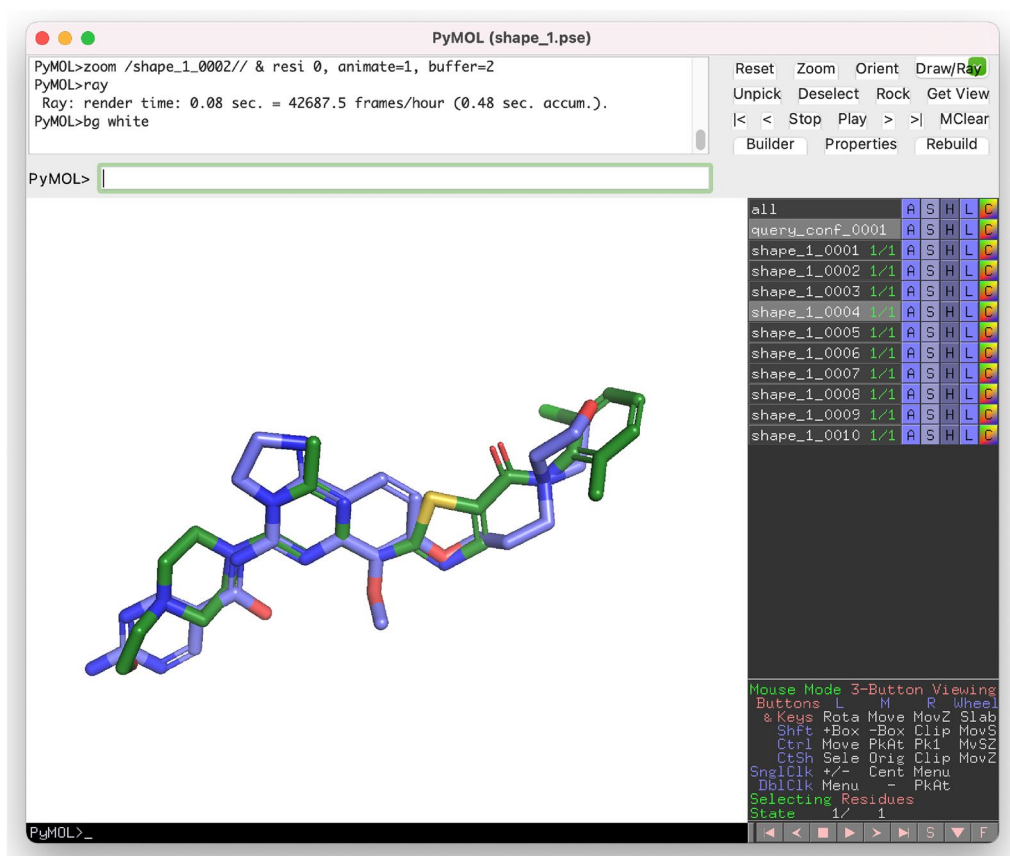
Table 1 summarizes the overall runtime for each call, e.g. it contains the loading time for the database file, the substructure or similarity search and the generation of the output file.

### Virtual screening performance

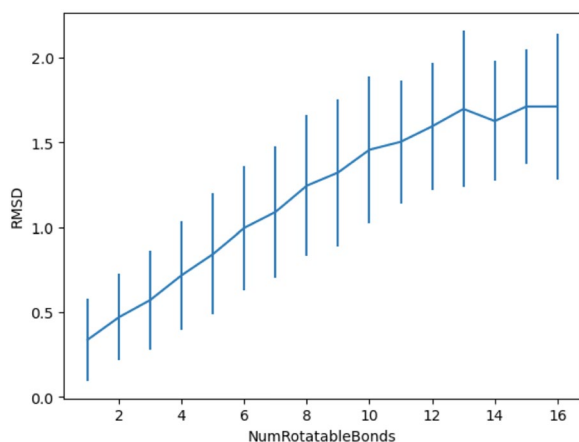
To give the user an idea about the performance of the tool in virtual screening practice, i.e. whether it could identify active compounds, we did some basic simulated screenings using the maximum unbiased validation (MUV) dataset [29]. The MUV dataset is based on PubChem bioactivity data and consists of 17 targets, each with 30 actives and 15,000 decoys. The choice of actives and decoys is done based on confirmatory and primary screens, which makes the dataset very difficult for virtual screening methods. We performed sample screenings based on 2D fingerprint and 3D shape similarity (mode `fpsim` and `shape`). The general performance of 2D fingerprints implemented in RDKit has been studied extensively before, with the MUV dataset being part of a larger evaluation set [30]. We adapted a simplified version of the workflow described before by Rohrer [29] and Riniker [30]. In short, for each of the 17 subsets in the MUV dataset, one of the 30 active compounds was selected as query molecule and the remaining 29 actives were pooled together with the



**Fig. 5** Exemplary page of the pdf file generated after fpsim search. The fingerprint similarity (FCFP4-like Morgan 2048 bits) of the molecules with the query molecule dasatinib is visualized in the left column, the right column shows IDs of the molecule as well as the search parameters and the calculated Tanimoto similarity



**Fig. 6** Screenshot from the PyMOL session file generated after shape similarity screening. By default, the first ten hits (one of them shown here in blue) are aligned with the query molecule dasatinib (green)



**Fig. 7** RMSD spread of the conformer generation process (ETKDG3 followed by MMFF94 minimization) for the search of the bioactive conformation (Platinum data set)

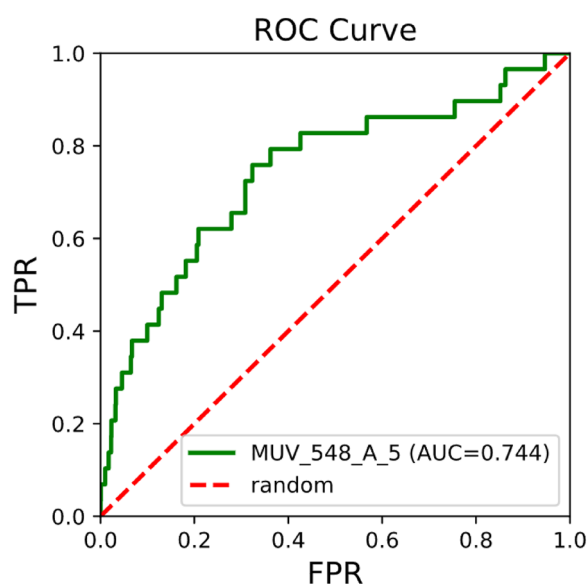
15,000 decoys and used as validation set. This query/validation split was done for all 30 actives. For the resulting 30 query/validation test splits per subset the

**Table 1** Runtime performance of substructure and similarity search on 12th Gen Intel(R) Core(TM) i7-12700 H with 2.70 GHz and 20 cores and 32GB RAM running Windows 11

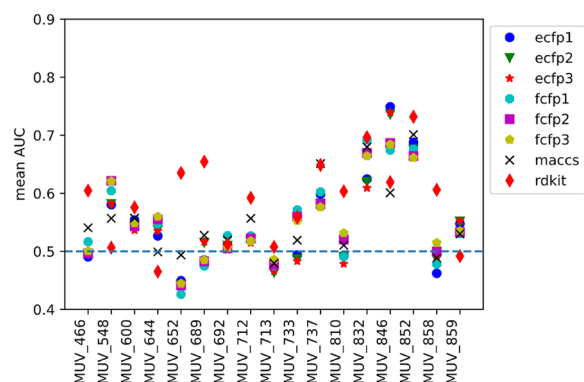
	pdb database	ChEMBL database	cores
Substructure	1 s	162 s	1
	0.8 s	89 s	6
Similarity	1 s	157 s	1
	0.75 s	77 s	6

virtual screening performance was measured by the area under the receiver operating curve (AUC, example curve shown in Fig. 8) and the mean value was calculated for each subset (mean AUC). The screening consisted of two steps: (1) generation of a vsdb database with standardized molecules and pre-computed fingerprints or conformers for the validation set; (2) 2D or 3D similarity screening of the validation set against the query molecule.





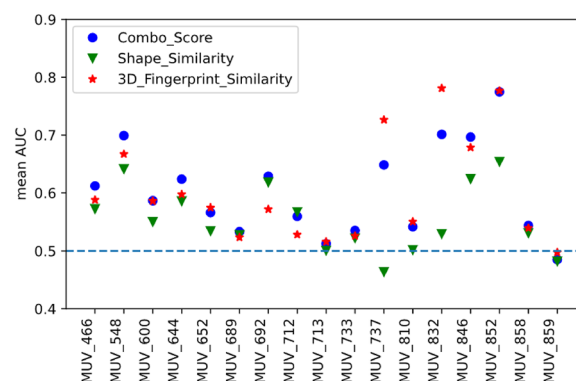
**Fig. 8** Example of a receiver operating curve (ROC) obtained for a query/validation test split after a 2D similarity screening with ECFP2 fingerprint. The MUV subset MUV\_548 was the validation set, the query was compound MUV\_548\_A\_5. The area under the curve (AUC) is 0.744. FPR = false positive rate, TPR = true positive rate



**Fig. 9** Results of virtual screening validation with the MUV dataset for 2D fingerprint similarity. The expectation of mean AUC of 0.5 for random rankings is indicated by the blue dashed line

The results for 2D similarity screening with various descriptors is summarized in Fig. 9. They follow, in general, the trend observed by Riniker et al. for 2D fingerprints on the MUV dataset [30]. For some targets, a significant enrichment of actives (e.g. meanAUC = 0.74 for ECFP3 fingerprint for target FactorXIa [MUV\_846]) is observed, whereas for other targets no enrichments could be observed based on simple 2D similarity calculations.

Fig. 10 summarizes the results for the 3D shape-based virtual screenings. Best performance is observed when



**Fig. 10** Results of virtual screening validation with the MUV dataset for 3D shape-based screenings. The expectation of mean AUC of 0.5 for random rankings is indicated by the blue dashed line

using the combo score for result ranking for most MUV subsets. However, for MUV\_737 (estrogen receptor alpha) and MUV\_832 (cathepsin G) scoring with 3D fingerprint yields a better overall enrichment.

## Conclusions

VSFlow is a versatile command-line tool to perform ligand-based virtual screenings in large compound databases on the basis of the RDKit cheminformatics framework. It allows to perform a substructure search, a 2D fingerprint-based and a 3D shape-based similarity search based on the respective functionalities implemented in RDKit. Screenings can be easily parallelized to multiple cores and the screening results can be directly visualized as pdf or pymol file. The integration of VSFlow in existing virtual screening setups is straightforward because the entire code is open source.

## Availability and requirements

- Project name: VSFlow - Virtual Screening Workflow
- Project home page: <https://github.com/czodrowski-lab/VFlow>
- Operating system(s): Platform independent
- Programming language: Python
- Other requirements: Anaconda or Miniconda
- License: MIT
- Any restrictions to use by non-academics: no.

## Acknowledgements

We thank Juliana Gretz for an early version of VSFlow.

## Author contributions

SJ wrote the code and the manuscript. HV contributed the case studies. PC supervised the work by SJ and HV. All authors read and approved the final manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Availability of data and materials

The source code and the files needed to reproduce the examples from this manuscript can be found at <https://github.com/czodrowskilab/VFlow>.

## Declarations

### Competing interests

The authors declare that they have no competing interests.

Received: 24 May 2022 Accepted: 18 February 2023

Published online: 31 March 2023

## References

- Maia EHB, Assis LC, de Oliveira TA, da Silva AM, Taranto AG (2020) Structure-based virtual screening: from classical to artificial intelligence. *Front Chem*. <https://doi.org/10.3389/fchem.2020.00343>
- Gimeno A, Ojeda-Montes MJ, Tomás-Hernández S, Cereto-Massagué A, Beltrán-Debón R, Mulero M, Pujadas G, García-Vallvé S (2019) The light and dark sides of virtual screening: what is there to know? *Int J Mol Sci*. <https://doi.org/10.3390/ijms20061375>
- Maggiore G, Vogt M, Stumpfe D, Bajorath J (2014) Molecular similarity in medicinal chemistry. *J Med Chem* 57(8):3186–3204. <https://doi.org/10.1021/jm401411z>
- ROCS 3.4.3.0: OpenEye Scientific Software, Santa Fe, NM (2022) <http://www.eyesopen.com>. Accessed 7 Apr 2022
- Torres PHM, Sodero ACR, Jofily P, Silva-Jr FP (2019) Key topics in molecular docking for drug design. *Int J Mol Sci*. <https://doi.org/10.3390/ijms20184574>
- Pinzi L, Rastelli G (2019) Molecular docking: shifting paradigms in drug discovery. *Int J Mol Sci*. <https://doi.org/10.3390/ijms20184331>
- Gentile F, Yaacoub JC, Gleave J, Fernandez M, Ton AT, Ban F, Stern A, Cherkasov A (2022) Artificial intelligence-enabled virtual screening of ultra-large chemical libraries with deep docking. *Nat Protoc* 17:672–697. <https://doi.org/10.1038/s41596-021-00659-2>
- Warr WA, Nicklaus MC, Nicolaou CA, Rarey M (2022) Exploration of ultralarge compound collections for drug discovery. *J Chem Inf Model* 62:2021–2034. <https://doi.org/10.1021/acs.jcim.2c00224>
- Lessel U, Wellenzohn B, Lillenthal M, Claussen H (2009) Searching fragment spaces with feature trees. *J Chem Inf Model* 49:270–279. <https://doi.org/10.1021/ci800272a>
- infiniSee version 4.0.0; BioSolveIT GmbH, Sankt Augustin, Germany (2022) [www.biosolveit.de/infiniSee](http://www.biosolveit.de/infiniSee)
- Singh N, Chaput L, Villoutreix BO (2021) Virtual screening web servers: designing chemical probes and drug candidates in the cyberspace. *Brief Bioinform* 22:1790–1818. <https://doi.org/10.1093/bib/bbaa034>
- Gaulton A, Hersey A, Nowotka ML, Bento AP, Chambers J, Mendez D, Motow P, Atkinson F, Bellis LJ, Cibrian-Uhalte E, Davies M, Dedman N, Karlsson A, Magarinos MP, Overington JP, Papadatos G, Smit I, Leach AR (2017) The ChEMBL database in 2017. *Nucleic Acids Res* 45:945–954. <https://doi.org/10.1093/nar/gkw1074>
- Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE (2021) Pubchem in 2021: new data content and improved web interfaces. *Nucleic Acids Res* 49:1388–1395. <https://doi.org/10.1093/nar/gkaa971>
- Sterling T, Irwin JJ (2015) Zinc 15 - ligand discovery for everyone. *J Chem Inf Model* 55(11):2324–2337. <https://doi.org/10.1021/acs.jcim.5b00559>
- Bragina ME, Daina A, Perez MAS, Michielin O, Zoete V (2022) The swiss-similarity 2021 web tool: Novel chemical libraries and additional methods for an enhanced ligand-based virtual screening experience. *Int J Mol Sci*. <https://doi.org/10.3390/ijms23020811>
- Zoete V, Daina A, Bovigny C, Michielin O (2016) Swiss-similarity: a web tool for low to ultra high throughput ligand-based virtual screening. *J Chem Inf Model* 56:1399–1404. <https://doi.org/10.1021/acs.jcim.6b00174>
- Sunseri J, Koes DR (2016) Pharmit: interactive exploration of chemical space. *Nucleic Acids Res* 44:442–448. <https://doi.org/10.1093/nar/gkw287>
- Dalke A (2019) The chemfp project. *J Cheminform*. <https://doi.org/10.1186/s13321-019-0398-8>
- <https://github.com/schrodinger/gpsimilarity>. Accessed 22 May 2022
- <https://www.nextmovesoftware.com/arthor.html>. Accessed 22 May 2022
- RDKit: Open-Source Cheminformatics Software (2022). <https://www.rdkit.org>. Accessed 7 Apr 2022
- Anaconda Software Distribution. Anaconda Inc (2020). <https://docs.anaconda.com/>
- <https://github.com/mcs07/MolVS>. Accessed 22 May 2022
- The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC (2022). <https://pymol.org>
- Wang S, Witek J, Landrum GA, Riniker S (2020) Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *J Chem Inf Model* 60(4):2044–2058. <https://doi.org/10.1021/acs.jcim.0c00025>. (PMID: 32155061)
- <https://github.com/czodrowskilab/VFlow/wiki>. Accessed 22 May 2022
- Riniker S, Landrum GA (2013) Similarity maps—a visualization strategy for molecular fingerprints and machine-learning methods. *J Cheminform* 5(1):43. <https://doi.org/10.1186/1758-2946-5-43>
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. *Nucleic Acids Res* 28:235–42
- Rohrer SG, Baumann K (2009) Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data. *J Chem Inf Model* 49:169–184. <https://doi.org/10.1021/ci8002649>
- Riniker S, Landrum GA Open-source Platform to Benchmark Fingerprints for Ligand-based Virtual Screening. <http://www.jchemin.com/content/5/1/26>

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

