

ChemBioServer

# ChemBioServer 2.0

**George M. Spyrou, Evangelos Karatzas, Emmanouil Athanasiadis, Zoe Cournia, Juan Eiros Zamora**

**Biomedical Research Foundation  
Academy of Athens**

**The Cyprus Institute of Neurology and Genetics**

<https://chembioserver.vi-seem.eu/>

ChemBioServer: a web-based pipeline for filtering, clustering and visualization of chemical compounds used in drug discovery. Athanasiadis E., Cournia Z., Spyrou G. *Bioinformatics*. 2012 Nov 15;28(22):3002-3. Epub 2012 Sep 8. <http://dx.doi.org/10.1093/bioinformatics/bts551>

ChemBioServer 2.0: an advanced web server for filtering, clustering and networking of chemical compounds facilitating both drug discovery and repurposing. Karatzas E., Zamora J.E., Athanasiadis E., Dellis D., Cournia Z., Spyrou G. *Bioinformatics*. 2020 Apr 15;36(8):2602-4. Epub 2020 Jan 8. <http://dx.doi.org/10.1093/bioinformatics/btz976>

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Methodology</b>	<b>4</b>
2.1	Filtering . . . . .	4
2.2	Clustering . . . . .	6
2.3	Networking . . . . .	6
<b>3</b>	<b>Description of Program</b>	<b>8</b>
3.1	Filtering . . . . .	8
3.1.1	Browse Compounds . . . . .	8
3.1.2	Predefined Queries . . . . .	8
3.1.3	Combined Search . . . . .	10
3.1.4	Substructure . . . . .	11
3.1.5	Van der Waals . . . . .	13
3.1.6	Toxicity . . . . .	14
3.1.7	Re-ranking for Ensemble Docking . . . . .	16
3.1.8	Graphical representations of molecular properties . . . . .	17
3.2	Clustering . . . . .	18
3.2.1	Hierarchical . . . . .	18
3.2.2	Affinity Propagation . . . . .	19
3.3	Networking . . . . .	20
3.3.1	Structural Similarity Network Visualization . . . . .	20
3.3.2	Structural Similarity Network Analysis . . . . .	21
3.3.3	Combine two sdf files in a Network . . . . .	22
3.3.4	Attach similar-only nodes to Network . . . . .	23
3.3.5	Remove nodes from Network, based on similarity . . . . .	24
<b>4</b>	<b>Bibliography</b>	<b>25</b>

# 1 Introduction

Despite the improvements in available technologies to the pharmaceutical sector, the cost of commercializing a new drug doubles every 9 years (Scannell, et al., 2012). Designing novel organic compounds in a systematic fashion is a daunting task as it has been estimated that there can be up to 1060 molecules with drug-like properties (Polishchuk, et al., 2013). One of the initial stages in drug development is to explore this chemical space using libraries that attempt to capture its vastness with a small subset of very diverse molecules. Generating these libraries through exploration of this space is a challenge in itself, and several researchers have tackled the problem through different computational approaches, such as exhaustive search (Gómez-Bombarelli, et al., 2016), genetic algorithms (Virshup, et al., 2013) and very recently, deep neural networks (Gómez-Bombarelli, et al., 2018). Once a sufficiently large and diverse library of compounds is obtained (typically thousands of molecules), its components are virtually screened against a desired target to predict their energy and site of interaction (Lionta, et al., 2014). This initial prediction is of paramount importance in order to save both time and money, as the initial library is narrowed down to only the best scoring molecules that are selected for further screening using more detailed computational models and experimental assays.

One issue related to drug discovery is the problem of specificity. The complexity of a cell is still far beyond the reach of current simulation capabilities, and the real targets of drugs are never in isolation. Therefore, a compound that shows a strong affinity for a target could also have many off-target interactions, leading to undesired secondary effects. This is very often the case for protein families: groups of evolutionarily related proteins that share structural similarities.

On the other hand, already existing drugs might prove useful against a disease outside their initial target spectrum. Drugs with high structural similarity imply similar mode of action against similar targets. As it is highlighted in the study of Zhang et al., drug similarity analytics, including chemical structure similarity, aim to find drugs, which display similar pharmacological characteristics to the drug of interest (Zhang, et al., 2014). Drug repurposing studies and tools based on drug structural similarity have been already made (Gottlieb, et al., 2011; Li and Lu, 2012). A drug-drug network with nodes linked by their pairwise structural similarities shows direct association of compounds allowing the researcher to either choose or filter-out compounds based on these relations, as an additional virtual screening method.

ChemBioServer (Athanasiadis, et al., 2012; Karatzas, et al., 2020) is a very successful application that has been continuously supported by our Groups and is gaining attention from the scientific community (for the last 11 months it has an average of 8749 hits per month). We have updated the initial version of this server with:

- (a) a functionality that re-ranks virtual screening results based on screening the same compound library against different protein members of the same family, selecting only those compounds that score high for the protein of interest,
- (b) a group of networking tools in order to allow researchers to create networks of compounds and provide useful network metrics,
- (c) a functionality that infers potential drug repurposing based on structural similarity,
- (d) a filtering functionality to filter out compounds that are similar to unwanted substances (e.g. failed drugs).

## 2 Methodology

### 2.1 Filtering

The “Filtering” section of ChemBioServer allows researchers to browse and filter compounds based on intra-ligand steric clashes, unwanted toxicophores, and desirable or undesirable chemical moieties or physicochemical properties. In this update, the functionality “Docking Re-ranking” has been added to this group of actions. Very often users need to select compounds that rank high for their target of interest but low for evolutionarily related proteins with similar binding sites (e.g. in a set of protein kinases) in order to avoid potential side effects. Thus, they employ cross-docking virtual screening in multiple receptor structures to identify compounds that will be predicted to bind only to the receptor of interest and not to receptors of the same protein family. ChemBioServer 2.0 can post-process cross-docking results and automatically re-rank virtual screening output to reveal compounds that rank high for the protein of interest in seconds. First, the user uploads virtual screening results for the target(s) of interest using the “Upload target file(s)”. Multiple file upload is allowed as users may choose to dock a chemical library in multiple conformations of a given protein. Next, the

user uploads virtual screening results of the same chemical library that has been performed in protein structures users want to filter against. Again, multiple file upload is allowed. ChemBioServer 2.0 then re-ranks compounds and outputs to the user those compounds that rank high for the target of interest and low for undesired targets (based on the provided docking scores).

The re-ranking algorithm is equipped with three methods to define compound selectivity for the target protein: automatic, manual or based on minimum desired docking score difference of the compound set. In all three methods, the user has to specify the minimum number of compounds that should be retrieved from the re-ranking procedure. The automatic method detects high-scoring docked compounds for the target of interest that have a low docking score for the undesired protein targets. It thus starts by defining low and high docking score cutoffs as the top 1% best scoring compounds for the target(s) and the top 1% worst scoring compounds for the rest of the proteins, respectively. These cutoffs are iteratively relaxed using 1% increments until the minimum number of compounds desired by the user meets the filter conditions. The manual method provides more flexibility, as the user manually specifies the low and high docking scores as cutoffs and a direct search is performed. The third method provides an alternative way to define compound specificity for a given protein target. Often, the absolute values of docking scores as cutoffs might not be as important as the actual predicted free energy difference (docking score) between the compounds for each protein. The larger this difference, the more selective the compounds will be. Therefore, with the "Score Difference" selection from the Method Selection tab the user can specify a desired level of energy difference, and the program will proceed in a similar fashion to the automatic procedure. It will start by defining the top 1% lowest scoring compounds for the target protein and the second cutoff will be set above by given score difference. While the number of compounds that pass this filter is below the minimum number of compounds specified, the low energy cutoff will be gradually increased by 1% steps, and the high energy cutoff will always be at least above the set score difference (in kcal/mol). These two last methods are not guaranteed to succeed, as there might be no compounds that meet the selection criteria defined by the user. In such a case, the program falls back to the automatic method. After the filtered compounds are obtained in a data frame, they are written to an Excel file, which is available for download. This format was chosen to make it more accessible to a general scientific user base with no knowledge of programming. The algorithm uses the Pandas Python package API<sup>7</sup>, conveniently allowing for data processing. The linchpin of this library is the Data Frame object, which is used to store data in memory by read-

ing CSV files. These objects support Boolean indexing and have multiple methods implemented in C, which are faster than conventional Python ‘for’ loops. One of the three methods can be chosen and corresponding input boxes appear dynamically using JavaScript. The input files are stored in the server and analyzed by calling a Python script through PHP. The results are stored for 24 hours and a link to download them is displayed after successful finishing of the analysis.

## 2.2 Clustering

ChemBioServer 2.0 still features the two clustering methods that were initially included under the “Clustering” labeled section; hierarchical and affinity propagation clustering. Both methods return structural clusters of the input compounds to the users together with their distance matrix as well as a graphical visualization. The affinity propagation clustering also returns exemplar compounds for each cluster.

## 2.3 Networking

The “Networking” section of ChemBioServer features all similarity-based network-related actions that have been added to this update. Similarity networks present a visualization of the strongest connections between substances based on their structural similarity. Nodes that are close to each other imply similar mode of action in a pharmaceutical setting. Apart from the holistic type of visualization, network analysis offers insights regarding the neighborhood of each node and the topology of the network reveals nodes that may connect distinct subnetworks of compounds, inferring multiple modes of action for some compounds. Moreover, key drug players can be highlighted based on network properties such as degree, strength or betweenness, as structural representatives of a highly connected group of compounds. Usually, researchers need to discover new uses for existing drugs against diseases, hence lowering the cost of new drug creation (i.e. drug repositioning). Structural drug repurposing is a form of drug repositioning where predicted drugs target the same proteins as drugs structurally similar to them. For this reason, fast screening of drug lists is important in order to bring together test molecules with seemingly suitable substances based on their similarity. On the other hand, chemical substances might be deemed inappropriate for further studies based on structural criteria such as similarity to toxic substances

or previously failed drugs from clinical trials. The similarity edge lists derived from ChemBioServer's networking actions can be further explored via network analytics applications. Five networking functionalities are implemented and labeled "Structural Similarity Network Visualization", "Structural Similarity Network Analysis", "Combine two sdf files in a Network", "Attach similar-only nodes to Network" and "Remove nodes from Network, based on similarity" realise the aforementioned needs. In "Structural Similarity Network Visualization" the user uploads an sdf file and after choosing a similarity metric between "Tanimoto", "Euclidean", "Cosine", "Dice" and "Hamming" and a value cutoff threshold for the edges (based on similarity values) can visualize the network and download the similarity matrix between all input compounds. This matrix is returned through the call of the function `calcDrugFPSim` from the `Rcpi` package which calculates the drug molecules' similarity derived from their molecular fingerprints. The graph is drawn in the user interface via the javascript library `vis.js`. "Structural Similarity Network Analysis" uses the same type of input values and the calculated similarity matrix is used as an adjacency matrix in order to create a graph using the `igraph` package in R. Node metrics "Degree", "Betweenness" and "Strength" are then presented in a sortable table after execution.

The "Combine two sdf files in a Network" action allows the user to test an sdf file against another reference sdf set and paints the two groups of compounds in different colors, as well as allows the user to download the initial similarity matrix between the compounds of both input sets. In the "Attach similar-only nodes to Network" tab, a main network is created for the reference set with a given edge threshold and then compounds from the test set are attached to the main network via another edge threshold (e.g. stricter connections). Then the user can download the upper triangular adjacency matrix of the whole network, as well as the edge list of the reference - test edges. Finally, in the "Remove nodes from Network, based on similarity" tab, a main network is created for the reference set with a given edge threshold and then compounds similar to ones from the test set (second edge threshold input) are removed, together with their edges, from the network. Once again, the user can download the upper triangular adjacency matrix of the new network, as well as the edge list of the reference - test edges that accounted for the removal of the reference nodes.

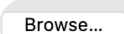
## 3 Description of Program

### 3.1 Filtering

#### 3.1.1 Browse Compounds

In this section you can upload a file and browse its compounds.

- Information on the compounds of the sdf is presented.  
A link for each compound that leads to an external applet ([Jmol](#)) is also provided and a 3D representation of the molecule according to the provided in the sdf file x-y-z coordinates is available.

- **Step 1.**  No file selected.

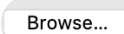
Please, Upload files\* in either ".sdf", or ".mol" format.

**\*Warning:**

- Uploaded filename **should not contain** special characters ^ ()\_
- Files are temporary saved on the server and deleted after processing.
- **Maximum** allowed upload size is **100MB** .

- **Final Step.** 

#### 3.1.2 Predefined Queries

- **Step 1.**  No file selected.

Please, Upload files\* in either ".sdf", or ".mol" format.

**\*Warning:**

- Uploaded filename **should not contain** special characters ^ ()\_
- Files are temporary saved on the server and deleted after processing.
- **Maximum** allowed upload size is **100MB** .



The tested compounds must pass ALL of the following criteria:

#### Lipinski Rules

- Molecular Weight:  $< 500$
- Hydrogen Bond Donors:  $\leq 5$
- Hydrogen Bond Acceptors:  $\leq 10$
- Partition coefficient  $\log P \leq 5$

#### Veber Rules

- Polar surface Area  $\leq 140 \text{ \AA}^2$
- Rotatable Bonds  $\leq 10$

#### Ghose Filters

- Total Atoms  $\geq 20$  AND  $\leq 70$
- Molar Refractivity  $\geq 40$  AND  $\leq 130$

- **Final Step.** 

*\*After the next successful execution your downloadable file links will be:*

1. *Compounds that Pass searching criteria:* [http://chembioserver.vi-seem.eu/upload/rietCUaBTO\\_pass.sdf](http://chembioserver.vi-seem.eu/upload/rietCUaBTO_pass.sdf)
2. *Compounds that Fail searching criteria:* [http://chembioserver.vi-seem.eu/upload/rietCUaBTO\\_fail.sdf](http://chembioserver.vi-seem.eu/upload/rietCUaBTO_fail.sdf)

*Results are stored for a week.*

### 3.1.3 Combined Search

- **Step 1.**  No file selected.

Please, Upload files\* in either ".sdf", or ".mol" format.

**\*Warning:**

- Uploaded filename **should not contain** special characters ^ ()\_
- Files are temporary saved on the server and deleted after processing.
- **Maximum** allowed upload size is **100MB** .

- **Optional Step.**  No file selected.

- Upload searchings parametres file.

In this optional step, the user is able to upload a file with comma-separated custom criteria values.

- If optional step was skipped, choose custom parametres:

- Molecular Weight:
- Number of Charges:
- Number of C Atoms:
- Number of RNH2:
- Number of R2NH
- Number of R3N
- Number of ROPO3
- Number of ROH
- Number of RCHO
- Number of RCOR
- Number of RCOOH
- Number of RCOOR
- Number of ROR
- Number of RINGS
- Number of AROMATIC
- Hydrogen Bond Donors
- Hydrogen Bond Acceptors
- Partition coefficient log P
- Polar surface Area

- **Final Step.**

*\*After the next successful execution your downloadable file links will be:*

1. Searching parameters: [http://chembioserver.vi-seem.eu/upload/JVjivizoni\\_parametres.txt](http://chembioserver.vi-seem.eu/upload/JVjivizoni_parametres.txt)
2. Compounds that Pass searching criteria: [http://chembioserver.vi-seem.eu/upload/JVjivizoni\\_pass.sdf](http://chembioserver.vi-seem.eu/upload/JVjivizoni_pass.sdf)
3. Compounds that Fail searching criteria: [http://chembioserver.vi-seem.eu/upload/JVjivizoni\\_fail.sdf](http://chembioserver.vi-seem.eu/upload/JVjivizoni_fail.sdf)

Results are stored for a week.

### 3.1.4 Substructure

This Assessment examines whether File 1 contains "same" compounds with File 2 or not.

- **Step 1.** Please, Upload files\* in either ".sdf", or ".mol" format.

- Compounds file 1:  No file selected.

- Compounds file 2:  No file selected.

**\*Warning:**

- Uploaded filename **should not contain** special characters ^ ()\_
- Files are temporary saved on the server and deleted after processing.
- **Maximum** allowed upload size is **100MB** .

An sdf [Sample file](#) Dataset with 1459 molecules which are commercial fragments extracted by FDA approved drugs. This fragments have been taken from [ChemInformatic Tools and Databases](#).

An sdf [Sample file](#) that contains 5 common unwanted fragments with undesired functional properties is also available to users.

1. c1ccc2c(c1)OCO2 - benzo-dioxane.
2. c1(c(ccs1)C(=O)C)N - 2-amino-3-carbonyl thiophene.
3. c1cc(c(cc1)O)O - catechol.
4. s1c(ncc1)N - aminothiazole.
5. S1C(=S)NC(=O)C1 - rhodanine.

• **Final Step.** **Process Data**

*\*After the next successful execution your downloadable file links will be:*

1. *Results in txt form:* [http://chembioserver.vi-seem.eu/upload/ygCkcAecAD\\_Similarity\\_Results.txt](http://chembioserver.vi-seem.eu/upload/ygCkcAecAD_Similarity_Results.txt)
2. *Unique compounds of file 1:* [http://chembioserver.vi-seem.eu/upload/FEc5qp4uCZ\\_unique\\_data.sdf](http://chembioserver.vi-seem.eu/upload/FEc5qp4uCZ_unique_data.sdf)
3. *Common compounds of file 1:* [http://chembioserver.vi-seem.eu/upload/FEc5qp4uCZ\\_common\\_data.sdf](http://chembioserver.vi-seem.eu/upload/FEc5qp4uCZ_common_data.sdf)

*Results are stored for a week.*

### 3.1.5 Van der Waals

• **Step 1.**  No file selected.

Please, Upload files\* in either ".*sdf*", or ".*mol*" format.

**\*Warning:**

- Uploaded filename **should not contain** special characters ^ ()\_.
- Files are temporary saved on the server and deleted after processing.
- **Maximum** allowed upload size is **100MB** .

• **Step 2.** Please, Select vdW Parametres.

• **van der Waals Energy Threshold:**

• **van der Waals Radii Tolerance:**

• **Final Step.**

*\*After the next successful execution your downloadable file links will be:*

1. Results in txt form: [http://chembioserver.vi-seem.eu/upload/s9NO7lKufb\\_vdW\\_Results.txt](http://chembioserver.vi-seem.eu/upload/s9NO7lKufb_vdW_Results.txt)
2. Compounds that pass the vdW test: [http://chembioserver.vi-seem.eu/upload/s9NO7lKufb\\_vdW\\_pass.sdf](http://chembioserver.vi-seem.eu/upload/s9NO7lKufb_vdW_pass.sdf)
3. Compounds that fail the vdW test: [http://chembioserver.vi-seem.eu/upload/s9NO7lKufb\\_vdW\\_fail.sdf](http://chembioserver.vi-seem.eu/upload/s9NO7lKufb_vdW_fail.sdf)

*Results are stored for a week.*

### 3.1.6 Toxicity

• **Step 1.**  No file selected.

Please, Upload files\* in either ".sdf", or ".mol" format.

**\*Warning:**

- Uploaded filename **should not contain** special characters ^ ()\_
- Files are temporary saved on the server and deleted after processing.
- **Maximum** allowed upload size is **100MB** .

1. N#N dinitrogen
2. C(=O)F formyl fluoride-Michael acceptor
3. C(=O)Cl formyl chloride-Michael acceptor
4. C(=O)Br formyl bromide-Michael acceptor
5. O1CC1 oxirane
6. C/N=N/C diazene
7. c1ccc2c(c1)cc1c(c2)cccc1 anthracene
8. C1=CC(=O)C=CC1=O quinone
9. c1cc(ccc1O)O hydroquinone
10. C=CC(=O)C butenone-Michael acceptor
11. CCOOCC O-O heteroatom
12. CCNNCC hydrazine-N-N heteroatom
13. CCNOCC N-O heteroatom
14. C=C(Cl)C chloroethane-Michael acceptor
15. C=C(F)C fluoroethane-Michael acceptor
16. C=C(Br)C bromoethane-Michael acceptor
17. C=CC#N acrylonitrile-Michael acceptor

18. C=C[N+](=O)[O-] nitroethene-Michael acceptor
19. CCSSCC disulfane-S-S heteroatom
20. c1ccc2c(c1)OCO2 benzo-dioxane
21. N(C(=S)NC)C thiourea
22. c1(c(ccs1)C(=O)C)N 2-amino-3-carbonyl thiophene
23. c1cc(c(cc1)O)O catechol
24. s1c(ncc1)N aminothiazole
25. S1C(=S)NC(=O)C1 rhodanine

• **Final Step.**  Process Data

*\*After the next successful execution your downloadable file links will be:*

1. *Results in txt form:* [http://chembioserver.vi-seem.eu/upload/SFpp1v7d5N\\_Toxicity\\_Results.txt](http://chembioserver.vi-seem.eu/upload/SFpp1v7d5N_Toxicity_Results.txt)
2. *Compounds that pass the toxic test:* [http://chembioserver.vi-seem.eu/upload/SFpp1v7d5N\\_tox\\_pass.sdf](http://chembioserver.vi-seem.eu/upload/SFpp1v7d5N_tox_pass.sdf)

*Results are stored for a week.*

### 3.1.7 Re-ranking for Ensemble Docking

This page will find those molecules which have a low energy of binding for a target protein whilst having a high energy of binding for others. Please refer to the [tutorial](#) to learn how to re-rank your docking results.

- **Step 1.** Upload target file(s).

These should be docking results for your target protein. Multiple file upload is allowed (different PDBs for a given protein).

No files selected.

- **Step 2.** Upload the rest of file(s).

These should be the rest of docking results for other proteins. These are the structures you want to filter against. Multiple file upload is allowed.

No files selected.

**\*Warning:**

- Only **CSV** formatted files are accepted.
- Please name your files like so: PDB-ProteinID.csv (For example, 1PY5-ALK5.csv)
- Uploaded filename **should not contain any** special character i.e. @%\$ \_ ^ . - ~
- Files are temporary saved on the server and deleted after processing. The results will be available for 24 h.
- The CSV files should at least have two columns named '**docking score**' and '**Unique SMILES Stereo**'.
- Each CSV file should be <100MB. The combined size of all files should be <3.2GB.

- **Step 3.** Please, select parameters.

- **Method Selection:**

1. Automatic   
Min compounds

- **Final Step.**

*\*After the next successful execution your downloadable file links will be:*

1. Results in csv form: <http://chembioserver.vi-seem.eu/results/3sl9plZEVj.csv>

*Results are stored for a week.*



### 3.1.8 Graphical representations of molecular properties

In this Step the following graphical representations are presented:

**PCA2 vs PCA1** Principal Component Analysis (PCA) first component (PCA1) against the second component (PCA2), based on the tanimoto coefficient (distance).

**PSA vs logP** Logarithm of the calculated Partition coefficient (logP) against the Polar Surface Area (PSA).

**PSA vs MW** Molecular Weight (MW) against the Polar Surface Area (PSA).

**logP vs MW** Molecular Weight (MW) against Logarithm of the calculated Partition coefficient (logP).

Plots are created by using the [Raphaël javascript library](#).

• **Step 1.**  No file selected.

Please, Upload files\* in either ".sdf", or ".mol" format.

**\*Warning:**

- Uploaded filename **should not contain any** special character i.e. @%\$^.-~
- Files are temporary saved on the server and deleted after processing.
- File should contain **more than two (2) compounds** in order to work.
- **Maximum** allowed upload size is **100MB** .

• **Final Step.**

## 3.2 Clustering

### 3.2.1 Hierarchical

• **Step 1.**  No file selected.

Please, Upload files\* in ".sdf", ".mol", or binary fingerprint in ".txt" format.

**\*Warning:**

- Uploaded filename **should not contain** special characters ^ ()\_
- Files are temporary saved on the server and deleted after processing.
- **Maximum number of compounds** that can be easily distinguished in the plot area is approximately **1000**.
- **Maximum** allowed upload size is **100MB** .
- **fingerprint.txt** format should contain two parts separated by a comma (,) character, the binary fingerprint and the name of the compound i.e:

000001100000010100001111011000000000000111111111111110, Compound 0001

001111111111111111110000000000000000011011100000000000, Compound 0002

0101111001100100011010010010101110000000101111001101, Compound 0003

Binary part should not contain any space before comma. (see [Example Data set 7](#). for more information)

Please notice that a [PDF reader](#) program should be installed in order to display clustering results.

• **Step 2.** Please, select Parametres

• **Distance Selection:**

Select Distances... 

• **Clustering Linkage Selection:**

Select Linkage... 

• **Clustering Threshold:**

Select Clusters... 

• **Final Step.**

*\*After the next successful execution your downloadable file links will be:*

1. *Distance matrix in csv format:* [http://chembioserver.vi-seem.eu/upload/FBshDpmyll\\_dist.txt](http://chembioserver.vi-seem.eu/upload/FBshDpmyll_dist.txt)
2. *Results in txt format:* [http://chembioserver.vi-seem.eu/upload/FBshDpmyll\\_groups.txt](http://chembioserver.vi-seem.eu/upload/FBshDpmyll_groups.txt)
3. *Clusters in Zip format:* <http://chembioserver.vi-seem.eu/upload/FBshDpmyll.zip>

*Results are stored for a week.*

### 3.2.2 Affinity Propagation

The affinity propagation algorithm takes as input a set of pairwise similarities among compound fingerprints, considering them as potential representative compounds (exemplars). The clusters and their corresponding representative compounds are calculated by exchanging messages between data points until a maximization process converge. Thus, exemplars for each cluster are proposed to the researcher for further investigation.

• **Step 1.**  No file selected.

Please, Upload files\* in ".sdf", ".mol", or binary fingerprint in ".txt" format.

**\*Warning:**

- Uploaded filename **should not contain** special characters ^ ()\_
- Files are temporary saved on the server and deleted after processing.
- **Maximum** allowed upload size is **100MB**.
- **fingerprint.txt** format should contain two parts separated by a comma (,) character, the binary fingerprint and the name of the compound i.e:

0000011000000101000011110110000000000000111111111111110, Compound 0001

0011111111111111111111000000000000000011011100000000000, Compound 0002

010111001100100011010010010101110111000000101111001101, Compound 0003

Binary part should not contain any space before comma.

(see [Example Data set 7](#). for more information)

• **Step 2.** Please, Select Distance Parametres.

• **Distance Selection:**

• **Final Step.**

*\*After the next successful execution your downloadable file links will be:*

1. *SDF with exemplar compounds:* [http://chembioserver.vi-seem.eu/upload/o4sDmDBFeE\\_exemplars.sdf](http://chembioserver.vi-seem.eu/upload/o4sDmDBFeE_exemplars.sdf)

*Results are stored for a week.*

## 3.3 Networking

### 3.3.1 Structural Similarity Network Visualization

The structural similarity network of compounds is created through the similarity matrix of drugs. This matrix is derived from the function calcDrugFPsim from the Rcp package which calculates the drug molecules' similarity derived by their molecular fingerprints. Choose the Similarity Metric as well as the cutoff threshold in [0, 1] for the edge drawing on the network. The graph is drawn through vis.js.

• **Step 1.**  No file selected.  
Please, Upload files\* in ".sdf" format.

**\*Warning:**

- Uploaded filename **should not contain** special characters ^ ()\_
- Files are temporary saved on the server and deleted after processing.
- **Maximum** allowed upload size is **100MB** .

• **Step 2.** Similarity Metric:

• **Step 3.** Edge Threshold [0-1]:

• **Final Step.**

*\*After the next successful execution your downloadable file links will be:*

1. Adjacency matrix of network: <http://chembioserver.vi-seem.eu/results/40FltTaHpv.tsv>

*Results are stored for a week.*

### 3.3.2 Structural Similarity Network Analysis

The similarity matrix from the calcDrugFPSim of Rcpis is used as an adjacency matrix in order to create a graph using the igraph package. Through network analysis, node metrics are presented in a table after execution.

• **Step 1.**  No file selected.

Please, Upload files\* in **".sdf"** format.

**\*Warning:**

- Uploaded filename **should not contain** special characters ^ ()\_
- Files are temporary saved on the server and deleted after processing.
- **Maximum** allowed upload size is **100MB** .

• **Step 2.** Similarity Metric:

• **Step 3.** Edge Threshold [0-1]:

• **Final Step.**

### 3.3.3 Combine two sdf files in a Network

The structural similarity network of compounds is created through the similarity matrix of drugs. This matrix is derived from the function calcDrugFPsim from the Rcpic package which calculates the drug molecules' similarity derived by their molecular fingerprints. Choose the Similarity Metric as well as the cutoff threshold in [0, 1] for the edge drawing on the network. The graph is drawn through vis.js. One sdf file might act as the reference network, while the other is composed of the compounds which the user wants to query against the reference set in order to find the closest structural neighbors.

• **Step 1.**  No file selected.

Please, Upload **reference set** \* in **".sdf"** format.

• **Step 2.**  No file selected.

Please, Upload **test set** \* in **".sdf"** format.

**\*Warning:**

- Uploaded filename **should not contain** special characters ^ ()\_
- Files are temporary saved on the server and deleted after processing.
- **Maximum** allowed upload size is **100MB** .

• **Step 3.** Similarity Metric:

• **Step 4.** Edge Threshold [0-1]:

• **Final Step.**

*\*After the next successful execution your downloadable file links will be:*

1. *Adjacency matrix of network:* <http://chembioserver.vi-seem.eu/results/zAtNhQBRZJ.tsv>

*Results are stored for a week.*

### 3.3.4 Attach similar-only nodes to Network

This utility requires two input sdf files as input; the first sdf is used to create the base network "A" of compounds and the second sdf is parsed in order to find and attach structurally similar nodes "B" to the base network. Two different similarity thresholds are given by the user; one for the connectivity strength of the base network nodes and one indicating the strength of the connections between the nodes of A and B.

• **Step 1.**  No file selected.

Please, Upload **reference set** \*in ".sdf" format.

• **Step 2.**  No file selected.

Please, Upload **test set** \*in ".sdf" format.

**\*Warning:**

- Uploaded filename **should not contain** special characters ^ ()\_
- Files are temporary saved on the server and deleted after processing.
- **Maximum** allowed upload size is **100MB** .

• **Step 3.** Similarity Metric:

• **Step 4.** Edge Threshold for base Network[0-1]:

• **Step 5.** Edge Threshold for A - B inference links[0-1]:

• **Final Step.**

*\*After the next successful execution your downloadable file links will be:*

1. Upper Triangular Adjacency matrix of network: <http://chembioserver.vi-seem.eu/results/DWcEKZqqR9.tsv>
2. Edgelist between A - B: [http://chembioserver.vi-seem.eu/results/DWcEKZqqR9\\_AB\\_edgelist.tsv](http://chembioserver.vi-seem.eu/results/DWcEKZqqR9_AB_edgelist.tsv)

*Results are stored for a week.*

### 3.3.5 Remove nodes from Network, based on similarity

This utility requires two input sdf as input; the first sdf is used to create the base network "A" of compounds and the second sdf is parsed in order to find and remove nodes from AB\_edgelist that are structurally similar to nodes "B". Two different similarity thresholds are given by the user; one for the connectivity strength of the base network nodes and one indicating the similarity threshold for deletion of nodes from A that are similar to B.

• **Step 1.**  No file selected.  
Please, Upload **reference set** \* in **".sdf"** format.

• **Step 2.**  No file selected.  
Please, Upload **test set** \* in **".sdf"** format.

**\*Warning:**

- Uploaded filename **should not contain** special characters ^ ()\_
- Files are temporary saved on the server and deleted after processing.
- **Maximum** allowed upload size is **100MB** .

• **Step 3.** Similarity Metric:

• **Step 4.** Edge Threshold for base Network[0-1]:

• **Step 5.** Edge Threshold for A - B, for removal of A nodes[0-1]:

• **Final Step.**

*\*After the next successful execution your downloadable file links will be:*

1. Upper Triangular Adjacency matrix of network: <http://chembioserver.vi-seem.eu/results/IZbIgE2pE2.tsv>
2. Edgelist between A - B: [http://chembioserver.vi-seem.eu/results/IZbIgE2pE2\\_AB\\_removed\\_edgelist.tsv](http://chembioserver.vi-seem.eu/results/IZbIgE2pE2_AB_removed_edgelist.tsv)

*Results are stored for a week.*



## 4 Bibliography

1. Athanasiadis, E., Cournia, Z. and Spyrou, G. ChemBioServer: a web-based pipeline for filtering, clustering and visualization of chemical compounds used in drug discovery. *Bioinformatics* 2012;28(22):3002-3003.
2. Gómez-Bombarelli, R., et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature materials* 2016;15(10):1120.
3. Gómez-Bombarelli, R., et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* 2018;4(2):268-276.
4. Gottlieb, A., et al. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology* 2011;7(1):496.
5. Karatzas, E., et al. ChemBioServer 2.0: an advanced web server for filtering, clustering and networking of chemical compounds facilitating both drug discovery and repurposing. *Bioinformatics* 2020;36(8):2602-2604
6. Li, J. and Lu, Z. A new method for computational drug repositioning using drug pairwise similarity. In, 2012 IEEE International Conference on Bioinformatics and Biomedicine. IEEE; 2012. p. 1-4.
7. Lionta, E., et al. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current topics in medicinal chemistry* 2014;14(16):1923-1938.
8. Polishchuk, P.G., Madzhidov, T.I. and Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of computer-aided molecular design* 2013;27(8):675-679.
9. Scannell, J.W., et al. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature reviews Drug discovery* 2012;11(3):191.
10. Virshup, A.M., et al. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *Journal of the American Chemical Society* 2013;135(19):7296-7303.

11. Zhang, P., et al. Towards personalized medicine: leveraging patient similarity and drug similarity analytics. *AMIA Summits on Translational Science Proceedings* 2014;2014:132.