

Supporting Information

LigQ: a WebServer to Select and Prepare Ligands for Virtual Screening

Leandro Radusky^{1,2}, Sergio Ruiz-Carmona³, Carlos Modenutti^{1,2}, Xavier Barril^{3,4}, Adrian G Turjanski^{1,2} and Marcelo A. Marti^{1,2}.

¹ Departamento de Química Biológica , Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires

² Insituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales (IQUIBICEN-CONICET), Pabellón II, Buenos Aires C1428EHA, Argentina

³ Department of Physical Chemistry, Faculty of Pharmacy and Institute of Biomedicine (IBUB), University of Barcelona, Avgda. Diagonal 643, Barcelona 08028, Spain

⁴ Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, Barcelona 08010, Spain

* Corresponding Author: marti.marcelo@gmail.com

Protein receptor	Ligands derived from PDB _a Seed I	Ligands derived from PFam PDBs _b Seed II	Ligands derived from BioAssays _c Seed III	Ligands derived from PFam BioAssays _d Seed IV
ace: Angiotensin-converting enzyme	16	23	0	21
ada: Adenosine deaminase	0	32	0	6
ampc: AmpC beta lactamase	52	80	156	157
dhfr: Dihydrofolate reductase	5	120	1	13
gart: glycnamide ribonucleotide transformylase	6	34	0	4
gbp: Glycogen phosphorylase beta	128	144	0	0
na: Neuraminidase	4	37	0	8
pnp: Purine nucleoside phosphorylase	19	89	0	6
tk: Thymidine kinase	27	29	0	1

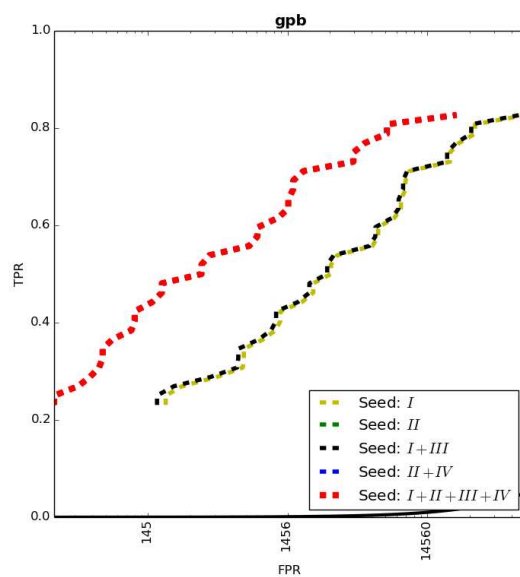
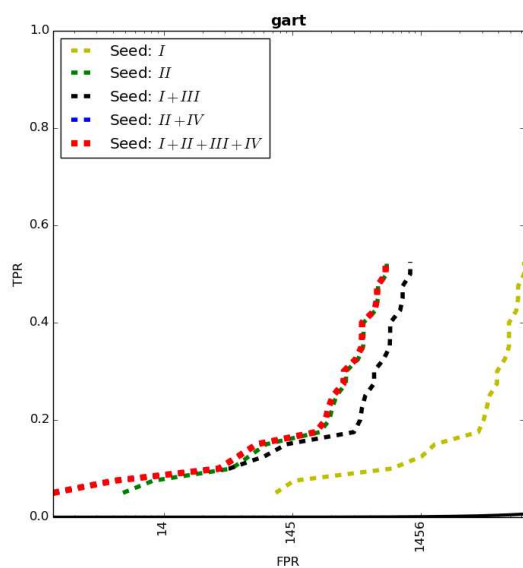
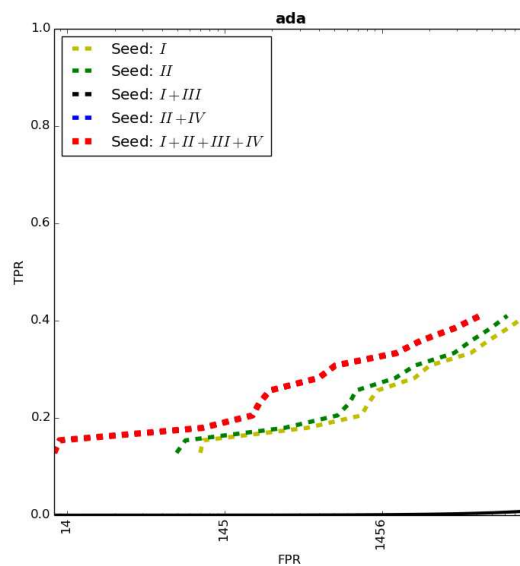
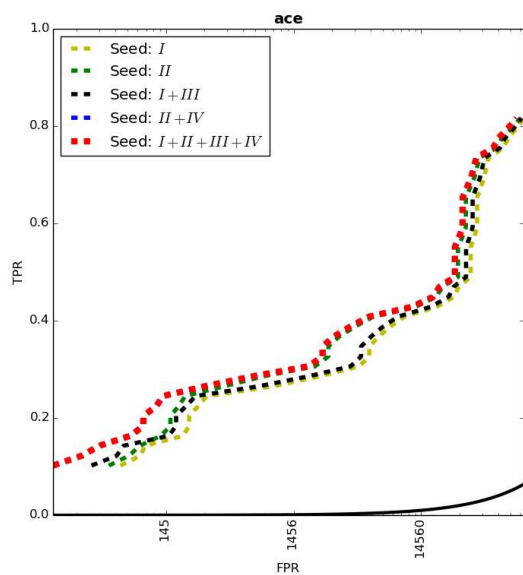
Table S1. Number of potential seed ligands derived from each source database, as described in Computational Methods section, for the different protein receptors taken from DUD database, taking only one for each different Pfam family.

Target	Family	DUD Ligands	Seed compounds	Extended Compounds	Structures generated
AmpC beta lactamase (ampc)	PF00144	21	237	695	2913
Dihydrofolate reductase (dhfr)	PF00186	410	133	512	4147

Table S2. Number of compounds computed in each module for the targets AmpC and DHFR.

Method \ Avg values for DUD	Enrichment Factor 1%	AUC	Hit Rate 1%
LIGSIFT	20.8	0.79	59
mRaise	20.2	0.76	55.5
LigQ	17.1	0.71	51.5

Table S3. Average calculations over the whole DUD dataset compared to other methods.



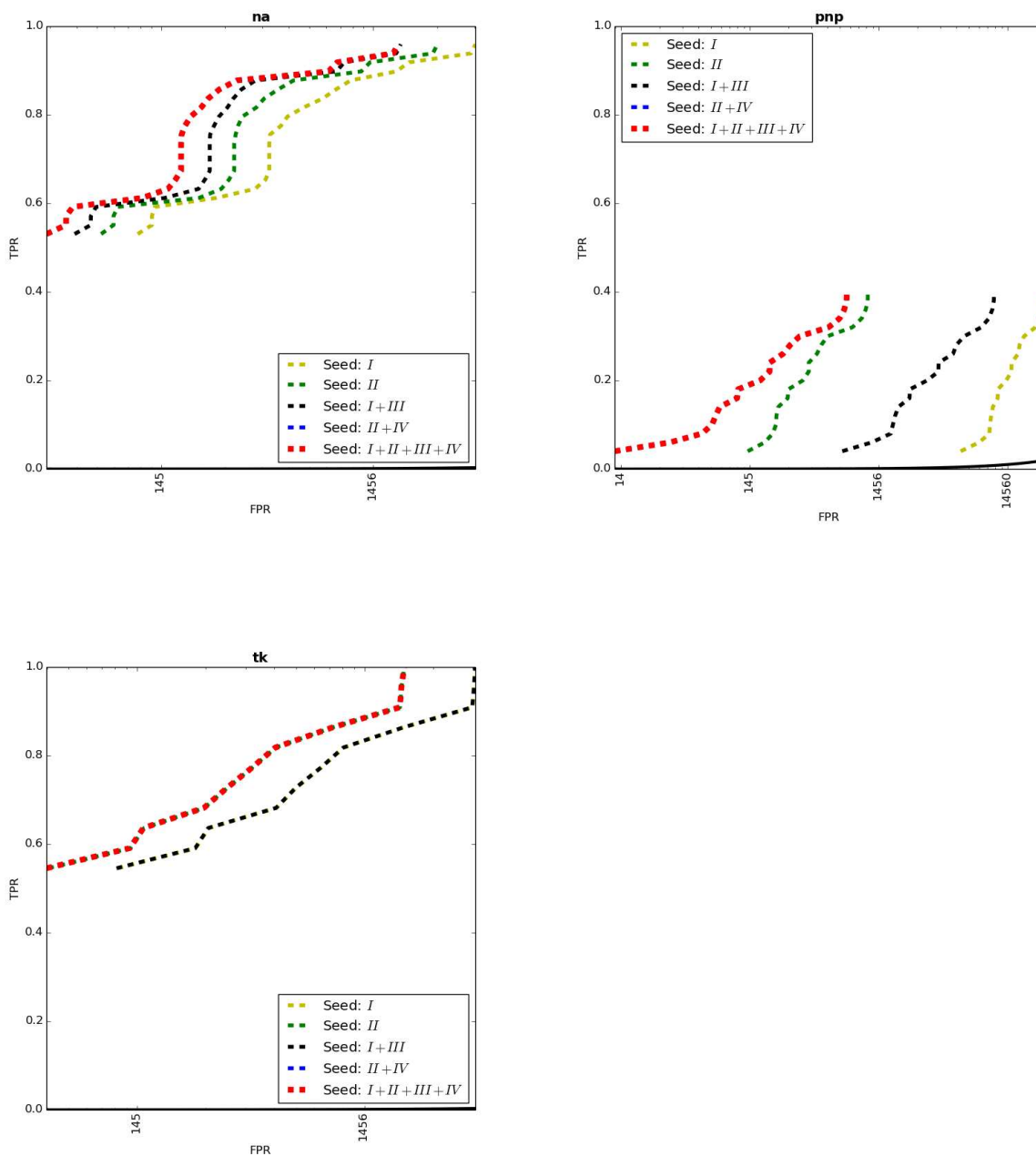


Figure S1. Plots of the semilogarithmic ROC curves for all tested proteins, except AmpC and dhfr analyzed in Figure 1. TPR is defined as the number of retrieved true binders relative to the total number of true binders. The FPR is defined as the number of total retrieved compounds with respect to the whole database size (ca. 1.4 million compounds). FPR label indicates the

actual number of retrieved compounds for clarity purposes. Different lines correspond to different seed compound groups. Different lines correspond to different seed compound groups, as described in Computational Methods. finally red lines use as seed all ligands together.

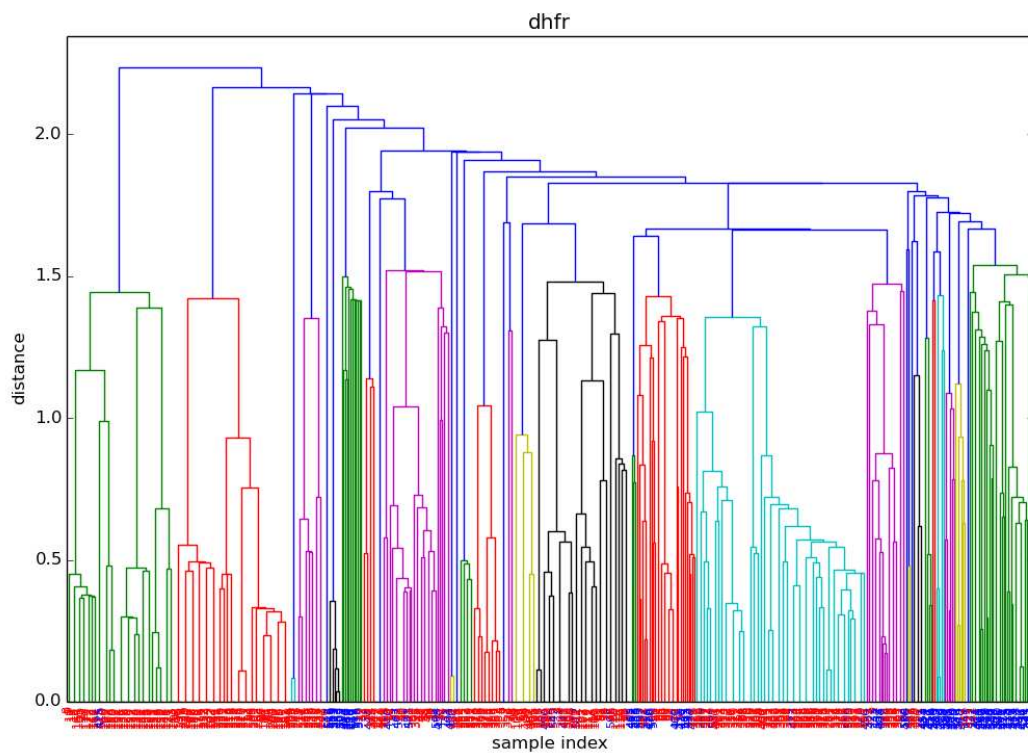
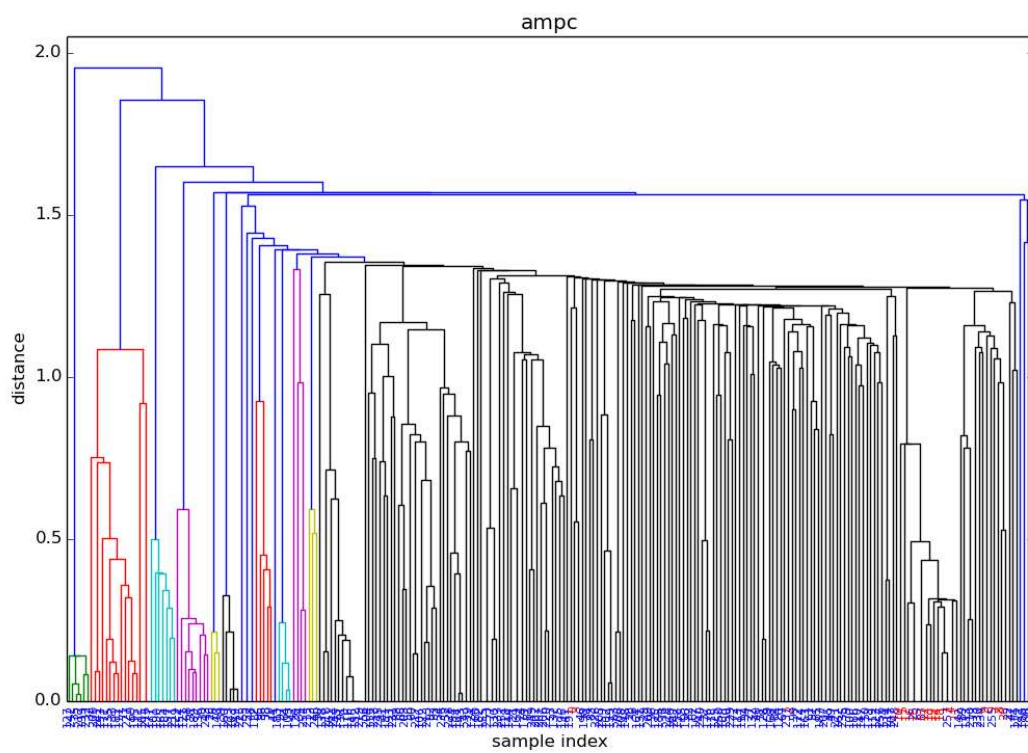


Figure S2. Dendrogram of the distance matrix of the DUD ligands plus the seed compounds for a) AmpC b) DHFR. In the x axis, DUD ligands are highlighted in red, and in blue are those compounds coming from the seed set.

Module Tutorials

Pocket Detection Module (PDM)

This module aim is to find -for the desired target- possible ligand binding pockets and rank them according to their druggability (a measure of their tendency to bind drug-like compounds compounds).

As shown in Figure S3, the PDM accepts as input either a UniProt or PDB accession code.

When a uniprot is entered PDM automatically searches for possible structures, and if no structure is available module builds -if possible- an homology based model. Pockets and druggability are computed using fpocket software, as implemented in our previous works (see Radusky et. al. 2014 in references).

User should enter his/her email to receive notifications (like job finished). Once the job is



The image shows a web form for the Pocket Detection Module (PDM). It contains the following elements:

- A label "Define your protein target code:" followed by a text input field with placeholder text "target ID (6 characters of Uniprot, 4 if PDB)".
- A label "Database of the input protein code:" followed by two radio buttons. The first is labeled "Uniprot" and is selected (indicated by a filled circle). The second is labeled "PDB" and is unselected (indicated by an empty circle).
- A label "E-mail (optional)" followed by a text input field with placeholder text "example@mail.com".
- A "Submit Job" button at the bottom.

correctly launched, server will show “work in progress” page.

Figure S3. The input form for the Pocket Detection Module.

The main results page for the PDM is shown in Figure S4. The results consists of a list of all druggable pockets (DS score > 0.5) found for the desired target. Each pocket properties can be further explored (view properties) and they can be also analyzed in the protein structure context on-line (View - see below). Results can be also downloaded for local analysis using stand alone protein structure visualization software (Download Button).

Job Summary

Analyzed protein [O60885](#): Bromodomain-containing protein 4
 Gene: BRD4, Organism: Homo sapiens
 Download the candidate active site grid for rDock software: [Download](#)
 Download the pocket software results: [Download](#)

O60885

Chain	Pockets	Visualize												
Pdb: 4UYD Chain: A	<table border="1"> <thead> <tr> <th>Pocket Number</th> <th>Druggability Score</th> <th>All Properties</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>0.3980</td> <td>View Properties</td> </tr> <tr> <td>4</td> <td>0.9146</td> <td>View Properties</td> </tr> <tr> <td>3</td> <td>0.5044</td> <td>View Properties</td> </tr> </tbody> </table>	Pocket Number	Druggability Score	All Properties	1	0.3980	View Properties	4	0.9146	View Properties	3	0.5044	View Properties	View 3D Visualizer
Pocket Number	Druggability Score	All Properties												
1	0.3980	View Properties												
4	0.9146	View Properties												
3	0.5044	View Properties												

Druggable pockets are starred

Figure S4. The main results page for the Pockets Detection Module.

An example of pocket visualization is shown in figure S5. Visualization is performed using GLMol plug-in. Both pockets (shown as white spheres in the present case) and protein (shown as green ribbons) can be displayed in several ways, and other structural features such as ligands/solvent atoms (like the Sulfate ion shown as sticks) found in the structure can be highlighted.

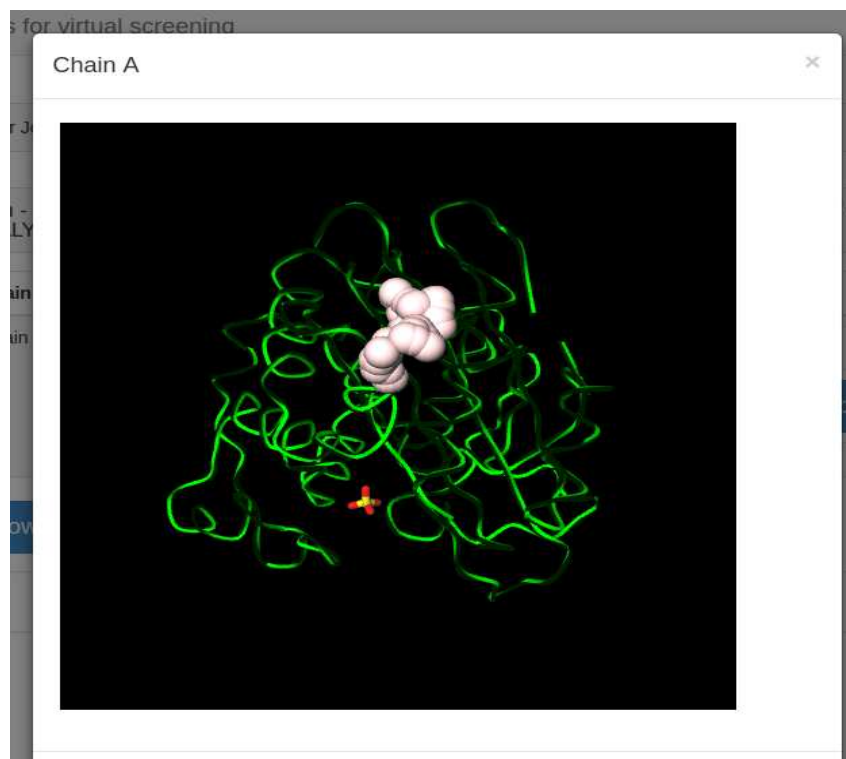


Figure S5. The druggable pockets found over the structure can be both downloaded and visualized online with a GLMol modal viewer.

Ligand Detection Module (LDM)

This module aim is to retrieve a set of compounds that a are potential binders to the desired target(s). As explained in the main text the LDM first assigns each target protein domain, to a PFAM domain, and also to any match for these domains in the ChEMBL bioactivity database. Ligands corresponding to the matching entries are retrieved and classified in four groups according to the linking evidence (see below).

The LDM takes as input one target to search seed sets, or a list of targets either as Uniprot or PDB accession codes. User email is required for notifications, but not mandatory. The figure S3 also works to figure how this module input is.

The results page for the LDM is shown in Figure S6, where all the retrieved compounds are listed. For each compound the reported target, the compound ID (using ZINC codes) and a 2D representation of the molecules is shown. More details on the compound can be found using properties link, while all assays reported for the compound can be listed in there are in the assays column. A search over the results can be performed with the search input up to the right.

In the top of the job's results page, a link to download all compounds and the summarization statistics of the job can also be found.

Job Summary

Protein O60885: Bromodomain-containing protein 4
 Gene: BRD4, Organsim: Homo sapiens
 Seed set cardinality: 163.
 Download the whole codes list: [Download](#)
 Download the whole properties list: [Download](#)

Found Compounds

Showing 1 to 25 of 163 rows 25 records per page

Search

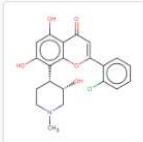
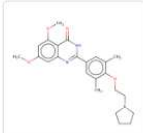
Compound	Targets	Properties	Image (click to enlarge)	Assays	Similar Purchasable Compounds
CPB	4O71 <i>Crystal target or Uniprot target</i>	Properties			explore... <i>Similar purchasable compounds on ZINC database</i>
5GD	5DW2	Properties			explore...

Figure S6. The main results page for the Ligands Detection Module.


Extend Compound Set Module (ECS)


The ECS module, is the core of the ligQ application. Using as input a series of compounds - which we call the seed set- it will query the ligQ “purchasable” compounds database looking for

similar compounds as those provided by the seed and thus retrieving a new list of potential binders.

As shown in figure S7 the ECS module accepts as input, a list of ZINC codes, PDB compound codes, InChis or SMILES -which usually are derived from the previously described LDM.


Define your target molecules:


 A list of newline-separated compounds

Type of input provided: ☒ Code ☐ SMILES ☐ InChI  The input format

Define extension type: ☒ Get most similar compounds ☐ Get compounds by Tanimoto Threshold

Insert Tanimoto threshold or Number of similar compounds to retrieve:

 Number of similar compounds to retrieve
Or Tanimoto Index threshold to search in DB

Filter by physicochemical properties  Set properties thresholds

E-mail (optional)

Figure S7. The input page for the Extend Compounds Set Module.

The search can be performed and/or filtered using several criteria -as shown in Figure S8- such as fingerprints or chemical properties, and the user can specify the amount of retrieved compounds or a chemical similarity threshold (which is defined using Tanimoto Coefficient)

Filter by physicochemical properties

☒ Filter by Volume range 0 1000

☒ Filter by LogP range 0 10

☒ Filter by HB Donors range 0 50

☒ Filter by HB Acceptors range 0 50

Figure S8. The extension of the ligand set can be done defining similarity criteria and applying filters over the extended compounds. All this parameters are defined in the input page.

The results given by this module are quite similar to the results page in the LDM output (Figure S6): a list of compounds, showing 2D representation, and several compound properties (MW; LogQ, Charge, etc.). User can also access and view the corresponding InChI, SMILES and download all (or a selected group) in SDF format, and also look for similar purchasable compounds in the ZINC database.

Ligand Structure Generation (LSG) Module

This module aim is given a list of compounds -which could be derived from the LDM or ELS modules- to built all possible 3D structures needed to perform a Virtual Screening procedure. The LSG accepts as input -figure S9- a list of compounds in any of the following formats (ZINC, PDBid, InChi, SMILES, SDF).

Define your target molecules:

Type of input provided: ☒ Code ☐ SMILES ☐ InChI

Define extension type: ☒ Most probable geometries ☐ Wider geometries set (slower)

E-mail (optional)

Submit Job

Figure S9. The input page for the for the Ligand Structure generation Module.

The module is able to deploy the structures in either SDF or PDB format that can be downloaded (figure S10).

Job Summary

Download the input list: [Download](#)

Geometries set cardinality: **573**

Download all the geometries in InChI format: [Download](#)

Download all the geometries in PDB format: [Download](#)

Download all the geometries in SDF format: [Download](#)

Job summary and downloads

Search

Showing 1 to 25 of 573 rows records per page

Download in distinct formats


Compound id	Derived from input compound:	Image (click to enlarge)	Downloads	See structure
CHEMBL422897_1	CHEMBL422897		SDF PDB InChI	View Structure 3D Visualizer

Figure S10. The main results page for the Ligand Structure Generation Module.

Built structures can also be analyzed pressint the “View Structure” button and a pop up will display the structure using the Jmol plugin (Figure S11).

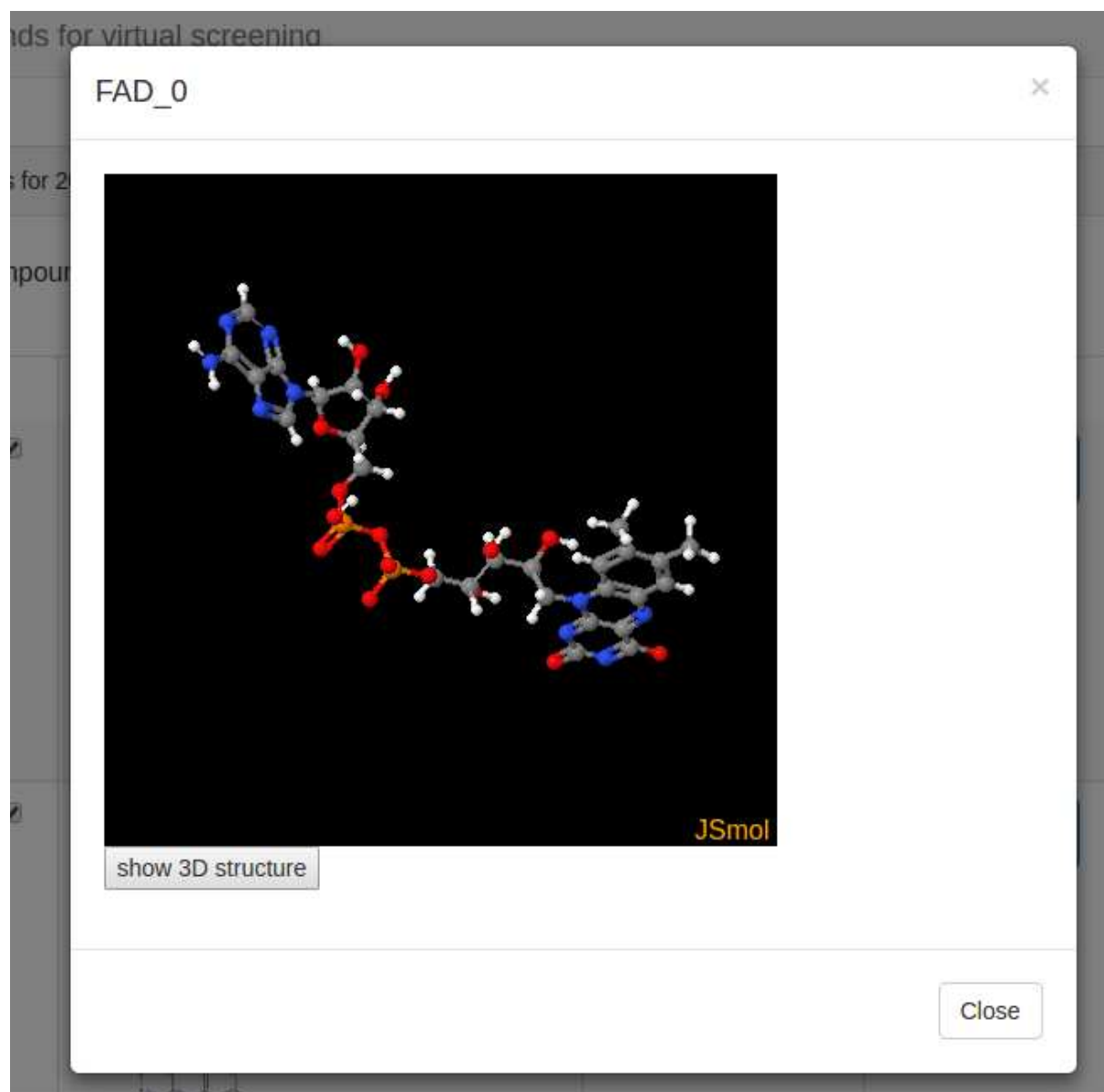


Figure S11. All the computed geometries can be both download on the selected file format (SDF or PDB) and visualized online with a JSmol modal viewer.