3D-e-Chem-VM: Cheminformatics Toolbox in a freely available Virtual Machine

Márton Vass[1], Stefan Verhoeven[2], Ross McGuire[3,4], Albert J. Kooistra[1,3], Lars Ridder[2], Tina Ritschel[3], Iwan de Esch[1], Rob Leurs[1], Gerrit Vriend[3], Chris de Graaf[1]

[1]Vrije Universiteit, Amsterdam, [2]Netherlands eScience center, Amsterdam

[3]Radboud University Medical Center, Nijmegen, [4]BioAxis Research, Oss.

The amount of publicly available experimental data in databases such as UniProt (protein sequences), PDB (protein structures), ChEMBL (small molecule bioactivities), and PubChem (compounds) is rapidly increasing.[1] However the vast number of possible protein-small molecule combinations is still far from covered. Therefore effective eScience technologies are needed for the integration of this large amount of heterogeneous data to enable development of predictive models of drug selectivity, polypharmacology, or off-target effects.

The 3D-e-Chem project

- Integration of large amounts of heterogeneous data from various sources
- Development of protein structure-based chemogenomics tools
- Creation of the 3D-e-Chem-VM[2] Virtual Machine with existing and novel cheminformatics tools[3] (KNIME[4] nodes) and chemogenomics workflows

VirtualBox machine image created by Packer

Collection of existing tools

Lubuntu OS, extensions via Ansible playbooks

Python and R configured to work within KNIME

KNIME Data Analytics with all free extensions

RDKit, Erlwood
CDK, Indigo
Vernalis
EMBL-EBI
OpenPhacts

Novel tools and workflows

Project hosted at GitHub
Wiki available inside VM

3D-e-Chem KNIME nodes and dependencies

KLIFS, GPCRdb
KRIPO
SyGMa

3D-e-Chem KNIME example workflows

GPCRdb, KRIPO, KLIFS
SyGMa, Chemdb4VS

Further integrated software
PyMOL, CAMB R package, Fpocket

Upgrade script fetches the latest Ansible playbooks and provisions the VM

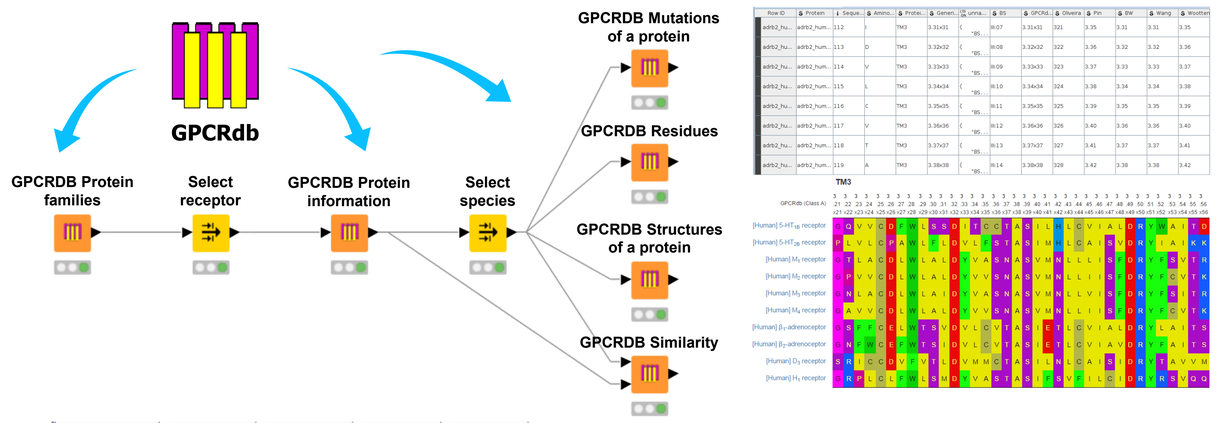
Vagrant box uploaded to Vagrant box catalog

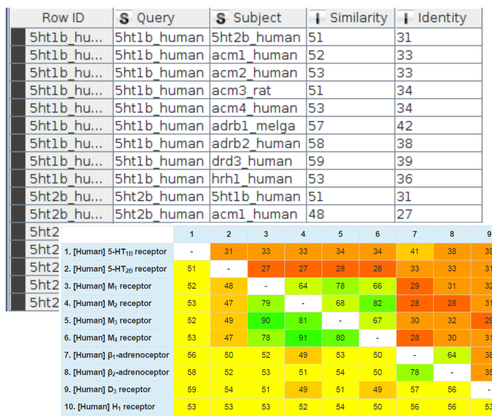
Installation

```
> vagrant init nlesc/3d-e-chem
> vagrant up
```

GPCRdb nodes

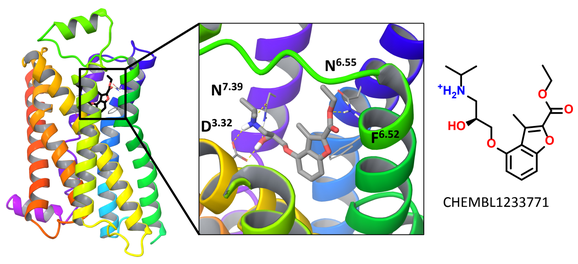
GPCRdb[5] is a database that contains annotated data, and web-based analysis and interactive visualization tools for G protein-coupled receptors.





Row ID	Query	Subject	Similarity	Identity
5ht1b_hu...	5ht1b_human	5ht2b_human	51	31
5ht1b_hu...	5ht1b_human	acm1_human	52	33
5ht1b_hu...	5ht1b_human	acm2_human	53	33
5ht1b_hu...	5ht1b_human	acm3_rat	51	34
5ht1b_hu...	5ht1b_human	acm4_human	53	34
5ht1b_hu...	5ht1b_human	adrb1_melga	57	42
5ht1b_hu...	5ht1b_human	adrb2_human	58	38
5ht1b_hu...	5ht1b_human	drd3_human	59	39
5ht1b_hu...	5ht1b_human	hrh1_human	53	36
5ht2b_hu...	5ht2b_human	5ht1b_human	51	31
5ht2b_hu...	5ht2b_human	acm1_human	48	27

5ht2	1	2	3	4	5	6	7	8	9	10
1. [Human] 5-HT$_{1B}$ receptor	-	31	33	33	34	34	41	38	39	36
2. [Human] 5-HT$_{2B}$ receptor	51	-	27	27	28	28	33	33	31	30
3. [Human] M$_1$ receptor	52	48	-	64	78	66	29	31	32	38
4. [Human] M$_2$ receptor	53	47	79	-	68	62	28	28	31	34
5. [Human] M$_3$ receptor	52	49	90	81	-	67	30	32	29	35
6. [Human] M$_4$ receptor	53	47	78	91	80	-	28	30	31	35
7. [Human] β_1-adrenoceptor	56	50	52	49	53	50	-	64	38	32
8. [Human] β_2-adrenoceptor	58	52	53	51	54	50	78	-	35	33
9. [Human] D$_2$ receptor	59	54	51	49	51	49	57	56	-	33
10. [Human] H$_1$ receptor	53	53	53	52	54	50	56	56	53	-



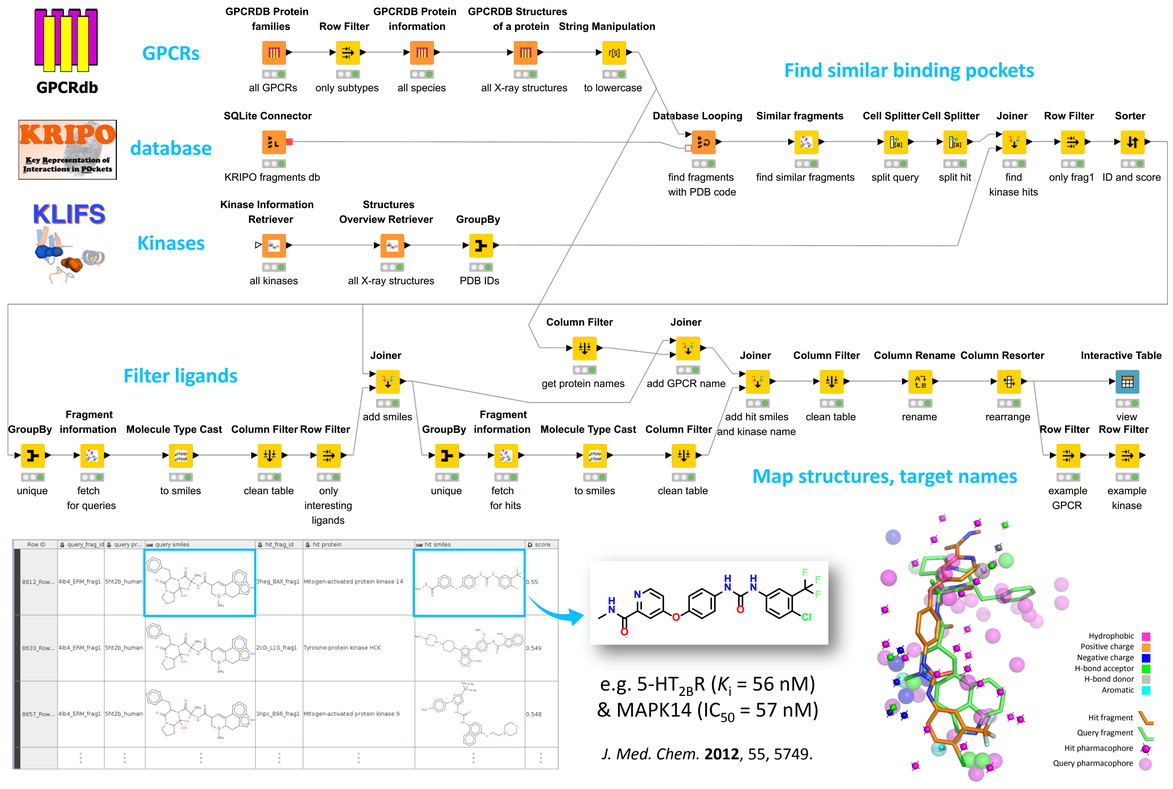
CHEMBL1233771

- Protein families: connection to GPCRdb, fetching hierarchic GPCR tree
- Protein information: name, species, source, sequence, UniProt ID
- Protein residues: generic and specific numbering schemes, segments
- Protein similarity: sequence similarity of specified segments and receptors
- Structures of a protein: PDB code, metadata, literature reference, ligands
- Structure-ligand interactions: interacting amino acids and interaction types
- Mutations of a protein: annotated mutational data, ligands, references

KRIPO nodes

KRIPO[6] is a pharmacophore-based binding pocket similarity assessment tool derived from the PDB and used to identify bioisosteres or off-targets.

- Similar fragments: find fragments bound to similar binding pockets from the pre-calculated similarity matrix, returns fragment IDs
- Fragment information: fragment structure, and associated protein identifiers



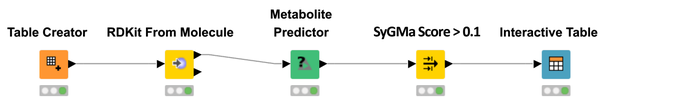
KLIFS nodes

KLIFS[7] is a structural chemogenomics database of kinase-ligand interaction information to assess determinants of kinase-ligand binding and selectivity.

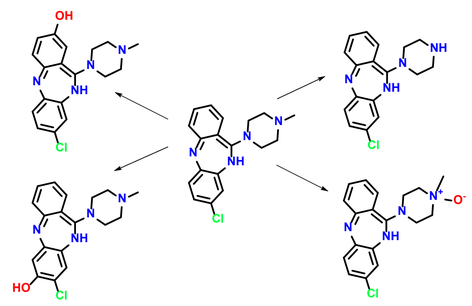
- Information nodes: retrieve kinases, names, UniProt ID, sequences
- Interaction nodes: interacting amino acids, interaction types, segments
- Ligand nodes: retrieve ligand name, ligand ID, ligand structure, and InChIKey
- Structure nodes: retrieve & download structure, PDB code, and annotations

SyGMa node

SyGMa is a rule-based method for Systematic Generation of potential Metabolites. Generates Phase I & II metabolites from the input molecule.

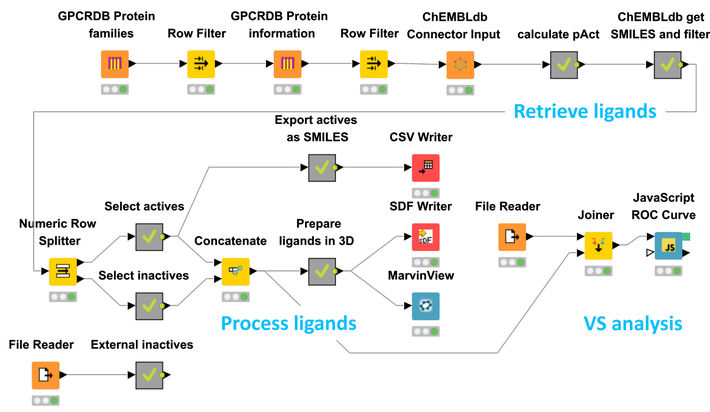


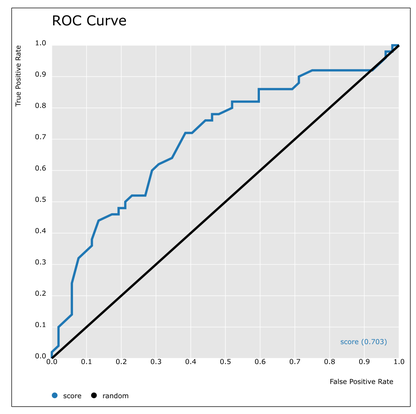
Row ID	Drug	smiles	RDKit molecule	SyGMa_metabolite	SyGMa_pathway	SyGMa...
1	Clozapine	CN1CCN...			N-demethylation_(R-N(CR)CH3);	0.418
2	Clozapine	CN1CCN...			N-oxidation_(tertiary_NCH3);	0.19



Chemdb4VS workflow

Quick and customizable method to assemble experimentally supported data sets for validation of virtual screening (VS) methods. Extraction of known active and inactive GPCR ligands from CHEMBLdb[1], and preparation of their three-dimensional molecular structures for VS.





References

1. http://www.uniprot.org/ , http://www.rcsb.org/ , https://www.ebi.ac.uk/chembl/ , https://pubchem.ncbi.nlm.nih.gov/
2. https://github.com/3D-e-Chem/3D-e-Chem-VM
3. http://dx.doi.org/10.5281/zenodo.58374 , http://dx.doi.org/10.5281/zenodo.58372 , http://dx.doi.org/10.5281/zenodo.58369
4. KNIME the Konstanz Information Miner, Studies in Classification, Data Analysis, and Knowledge Organization. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R; Kotter, T,; Meinl T.; Ohl, P.; Sieb C.; Thiel, K.; Wiswede, B., **2007**, ISBN 978-3-540-78239-1
5. GPCRDB: an information system for G protein-coupled receptors. Isberg, V.; Mordalski, S.; Munk, C; Rataj, K.; Harpsøe, K.; Hauser, A.S.; Vroling, B.; Bojarski, A.J.; Vriend, G.; Gloriam., D.E. *Nucleic Acids Res.* **2016**, *44*, D356-64.
6. http://dx.doi.org/10.5281/zenodo.58374 , Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. Wood, D.J.; de Vlieg, J.; Wagener, M.; Ritschel, T. *J. Chem. Inf. Mod.* **2012**, *52*, 2031-43.
7. http://klifs.vu-compmedchem.nl/ , KLIFS: a structural kinase-ligand interaction database. Kooistra, A.J.; Kanev, G.K.; van Linden, O.P.; Leurs, R.; de Esch, I.J.; de Graaf, C. *Nucleic Acids Res.* **2016**, *44*, D365-71.
8. SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites. Ridder, L.; Wagener, M. *ChemMedChem.* **2008**, *3*, 821-32.