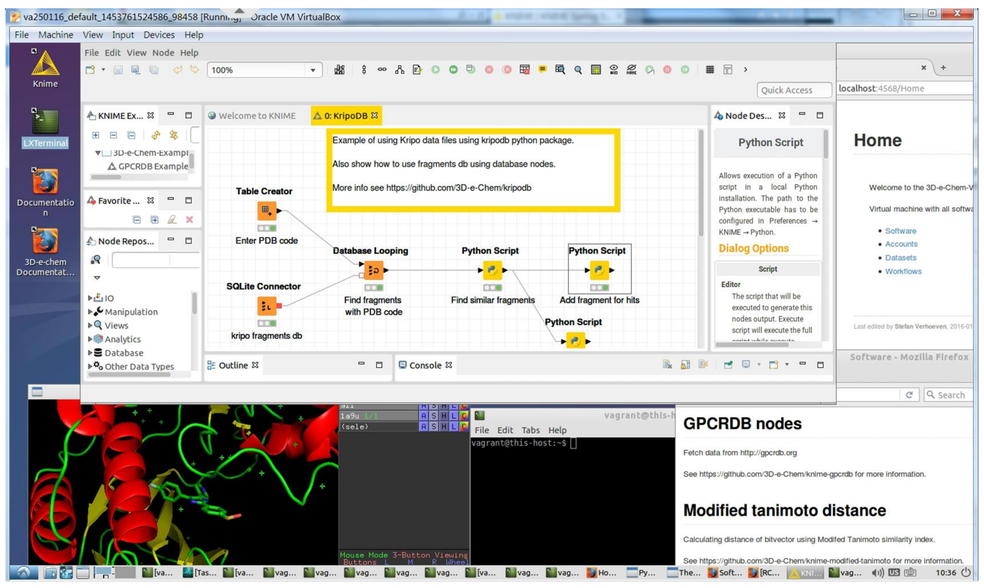
3-D-eChem VM: Cheminformatics research infrastructure in a downloadable Virtual Machine

Stefan Verhoeven[1], Marton Vass[2], Iwan De Esch[2], Rob Leurs[2], Scott J. Lusher[1,3], Gerrit Vriend[3], Tina Ritschel[3], Chris De Graaf[2], Ross McGuire[3,4]

[1]Netherlands eScience center, Amsterdam, [2] Vrije Universiteit, Amsterdam, [3]Radboud University Medical Center, Nijmegen, [4]BioAxis Research, Oss.

Abstract

3D-e-Chem VM[1] is freely available Virtual Machine (VM) encompassing tools, databases & workflows, including new resources[2] developed for ligand binding site comparisons and GPCR research. The VM contains a fully functional cheminformatics infrastructure consisting of a chemistry enabled relational database system (PostgreSQL[3] + RDKit[4]) with a data analytics workflow tool (KNIME[5]) and additional cheminformatics capabilities. Tools, workflows and reference data sets are made available. The wide range of cheminformatics functionalities are provided in the downloadable 3D-e-Chem VM allowing immediate use in research and education.

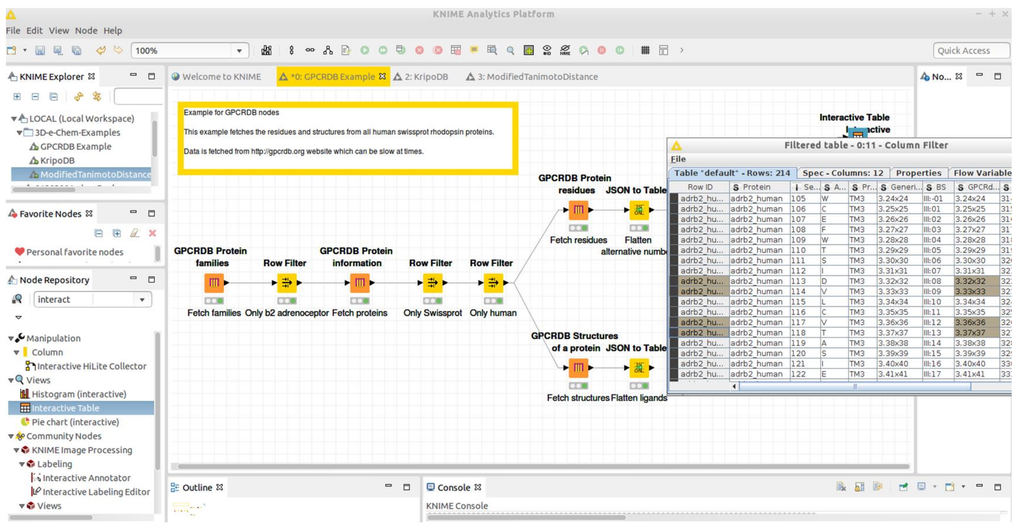


Snapshot of 3D e-Chem Virtual Machine

- Functionality demonstrates KNIME in the field of macromolecular structure chem- and bio-informatics – more than 'just' small molecule cheminf
- Tools exploit, integrate chem- & bio-informatics knowledge of G Protein-Coupled Receptors, important targets of current & new drugs
- New freely available KNIME nodes: automated, customizable integration of heterogeneous GPCR-ligand interaction data, covering experimental bioactivity, protein-ligand structure and nodes to utilize KRIPO[6] sub-pocket similarity fingerprints

GPCRdb nodes:

- GPCRDB Protein Families: Connection to GPCRDB[7], and extraction of protein family information
- GPCRDB Protein Information: Retrieval of e.g. source, species, sequence data from Uniprot identifiers or protein family slug ID
- GPCRDB Protein Residues: Retrieval of residue & numbering -generic numbering and/or alternate numbering schemes
- GPCRDB Structures of a Protein: Retrieval of GPCR structures (names, literature references, pdb code and ligands



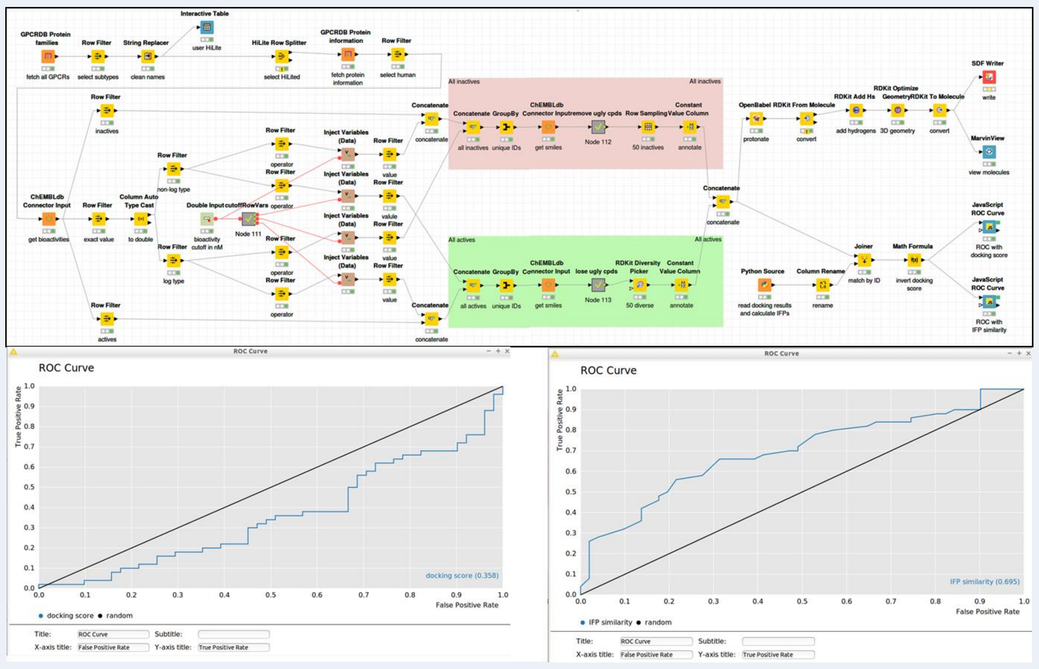
Simple workflow to connect to GPCRDB and extract information on residues & ligands

The 3D-e-Chem VM is available in Atlas[8a] the Vagrant[8b] box catalog of HashiCorp. Ansible[8c], a configuration management engine uses SSH to provision the VM via playbooks: The following software and enhancements are installed:

- LXDE, a lightweight desktop environment with graphical user interface exposing a command line prompt, database and website
- Knime + extensions + plugins, RDKit, Indigo[9], Erlwood[10], OpenPhacts[11]
- R[12], Python[13] configured for Knime + CAMB[14], as R package
- Fpocket[15], protein cavity detection program + PyMol[16]
- Github wiki & copy in VM including guide to create local ChEMBL
- Virtualbox guest additions – ease of use & file shares.
- Upgrade script, VM can be upgraded by re-download or upgrade script.

Chemdb4VS Workflow

Chemdb4VS: Extraction of GPCR ligand datasets from CHEMBLdb, design of focused decoy sets with similar physical-chemical properties, and preparation of their three-dimensional molecular structures for virtual screening studies.



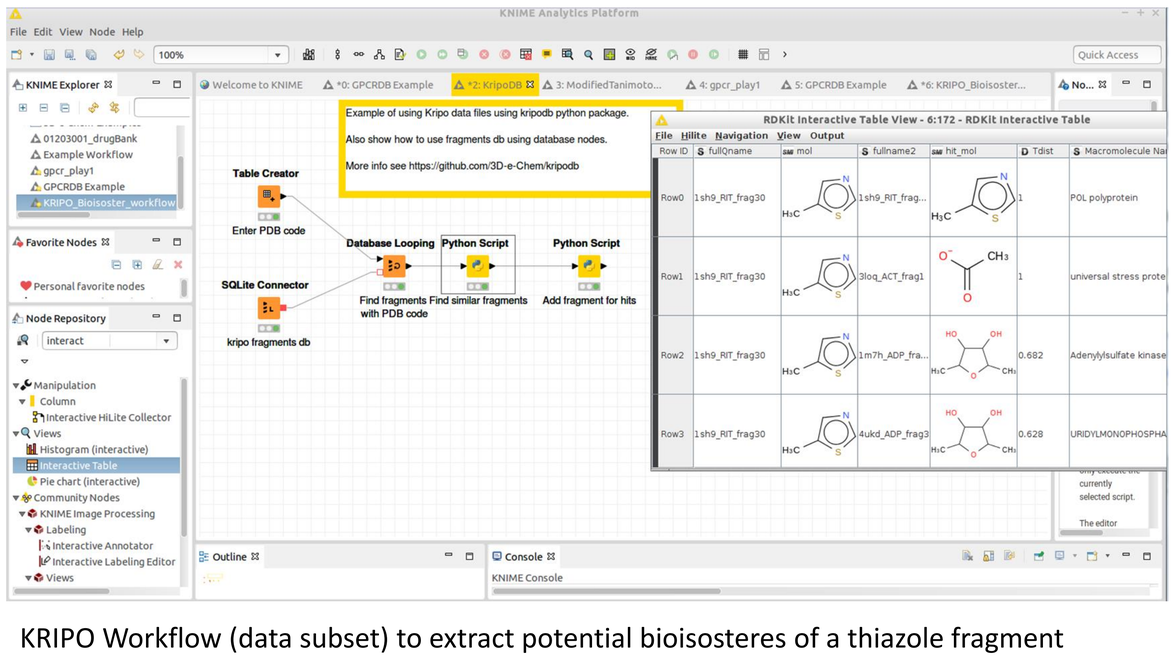
Provides a quick, customizable method to assemble experimentally supported ligand data sets for validation of VS methods:

- Selection of the target of interest
  - GPCRDB nodes access GPCR target lists & identifiers, download ortholog data and retrieve Uniprot identifiers
- Fetching and processing bioactivity data from the ChEMBL database
  - ChEMBLdb Connector Input node followed by Row Filter nodes for log & non-log activity types. User definable bioactivity cutoff
- Preparation of the ligand data sets
  - SMILES downloaded, filtered for unusual elements, druglikeness, number. Representative actives selected by RDKit Diversity Picker, ligand protonation & RDKIT 3D geometry optimization.
- Evaluation of the virtual screening efficiency after import
  - VS campaign results i using the ligand sets imported to KNIME and ROC curves plotted using the JavaScript ROC Curve node.

KRIPO nodes:

Customised Python script, Bit Vector Distance and Database Connector nodes:

- An SQLite DB contains an all-v-all comparison of Kripo fingerprint similarity (using a modified Tanimoto distance metric) of pdb ligands/ligand fragments.
- KRIPO similarity uses 3-point pharmacophores derived from fragment binding site environments –bioisosteres via protein-ligand binding site similarity



KRIPO Workflow (data subset) to extract potential bioisosteres of a thiazole fragment

References

1. https://github.com/3D-e-Chem/3D-e-Chem-VM , http://dx.doi.org/10.5281/zenodo.45266
2. http://dx.doi.org/10.5281/zenodo.45990 , http://dx.doi.org/10.5281/zenodo.45265 , http://dx.doi.org/10.5281/zenodo.45270
3. http://www.postgresql.org/ ,
4. RDKit: Open-source cheminformatics; http://www.rdkit.org
5. 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
6. 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, 8, 2031-2043
7. 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. **2016**, *Nucleic Acids Res.*, 44, D356-D364.
8. a) https://www.hashicorp.com/atlas.html b) https://www.vagrantup.com/ c) http://www.ansible.com/
9. https://github.com/ggasoftware/indigo
10. https://tech.knime.org/community/erlwood
11. Drug discovery FAQs: workflows for answering multidomain drug discovery questions. Chichester, C.; Digles, D.; Siebes, R.; Loizou, A.; Groth, P.; Harland, L. **2015** Drug Discov. Today 20: 399–405
12. R Development Core Team **2008**. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.
13. Python Software Foundation. Python Language Reference,. Available at http://www.python.org
14. Chemically Aware Model Builder (camb): an R package for property and bioactivity modelling of small molecules J Cheminform. **2015** 7:45 Murrell, D.S.; Cortes-Ciriano, I.; van Westen, G.J.; Stott, I. P.; Bender, A.; Malliavin, T.E.; Glen R.C.
15. Fpocket: An O.S. platform for ligand pocket detection BMC Bioinformatics, **2009**, 10:168. Le Guilloux, V.; Schmidtke, P.; Tuffery P.
16. The PyMOL Molecular Graphics System, Version 1.7.0.0, Schrödinger, LLC