JOURNAL OF CHEMICAL INFORMATION AND MODELING

Application Note

pubs.acs.org/jcim

3D-e-Chem-VM: Structural Cheminformatics Research Infrastructure in a Freely Available Virtual Machine

Ross McGuire,*[,†,‡,#] Stefan Verhoeven,*[,§,#] Márton Vass,[#,∥] Gerrit Vriend,[†] Iwan J. P. de Esch,[∥] Scott J. Lusher,[†,§] Rob Leurs,[∥] Lars Ridder,[§] Albert J. Kooistra,[†,∥] Tina Ritschel,[†] and Chris de Graaf*[,∥]

[†]Centre for Molecular and Biomolecular Informatics (CMBI), Radboudumc, 6525 GA Nijmegen, The Netherlands

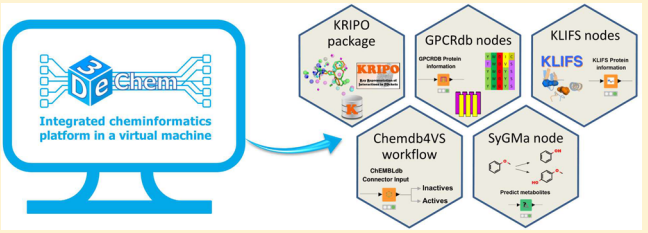
[‡]BioAxis Research, Pivot Park, 5349 AE Oss, The Netherlands

[§]Netherlands eScience Center, 1098 XG Amsterdam, The Netherlands

[∥]Division of Medicinal Chemistry, Faculty of Sciences, Amsterdam Institute for Molecules, Medicines and Systems (AIMMS), Vrije Universiteit Amsterdam, 1081 HZ Amsterdam, The Netherlands

Supporting Information

ABSTRACT: 3D-e-Chem-VM is an open source, freely available Virtual Machine (http://3d-e-chem.github.io/3D-e-Chem-VM/) that integrates cheminformatics and bioinformatics tools for the analysis of protein−ligand interaction data. 3D-e-Chem-VM consists of software libraries, and database and workflow tools that can analyze and combine small molecule and protein structural information in a graphical programming environment. New chemical and biological data analytics tools and workflows have been developed for the efficient exploitation of structural and pharmacological protein−ligand interaction data from proteomewide databases (e.g., ChEMBLdb and PDB), as well as customized information systems focused on, e.g., G protein-coupled receptors (GPCRdb) and protein kinases (KLIFS). The integrated structural cheminformatics research infrastructure compiled in the 3D-e-Chem-VM enables the design of new approaches in virtual ligand screening (Chemdb4VS), ligand-based metabolism prediction (SyGMa), and structure-based protein binding site comparison and bioisosteric replacement for ligand design (KRIPOdb).



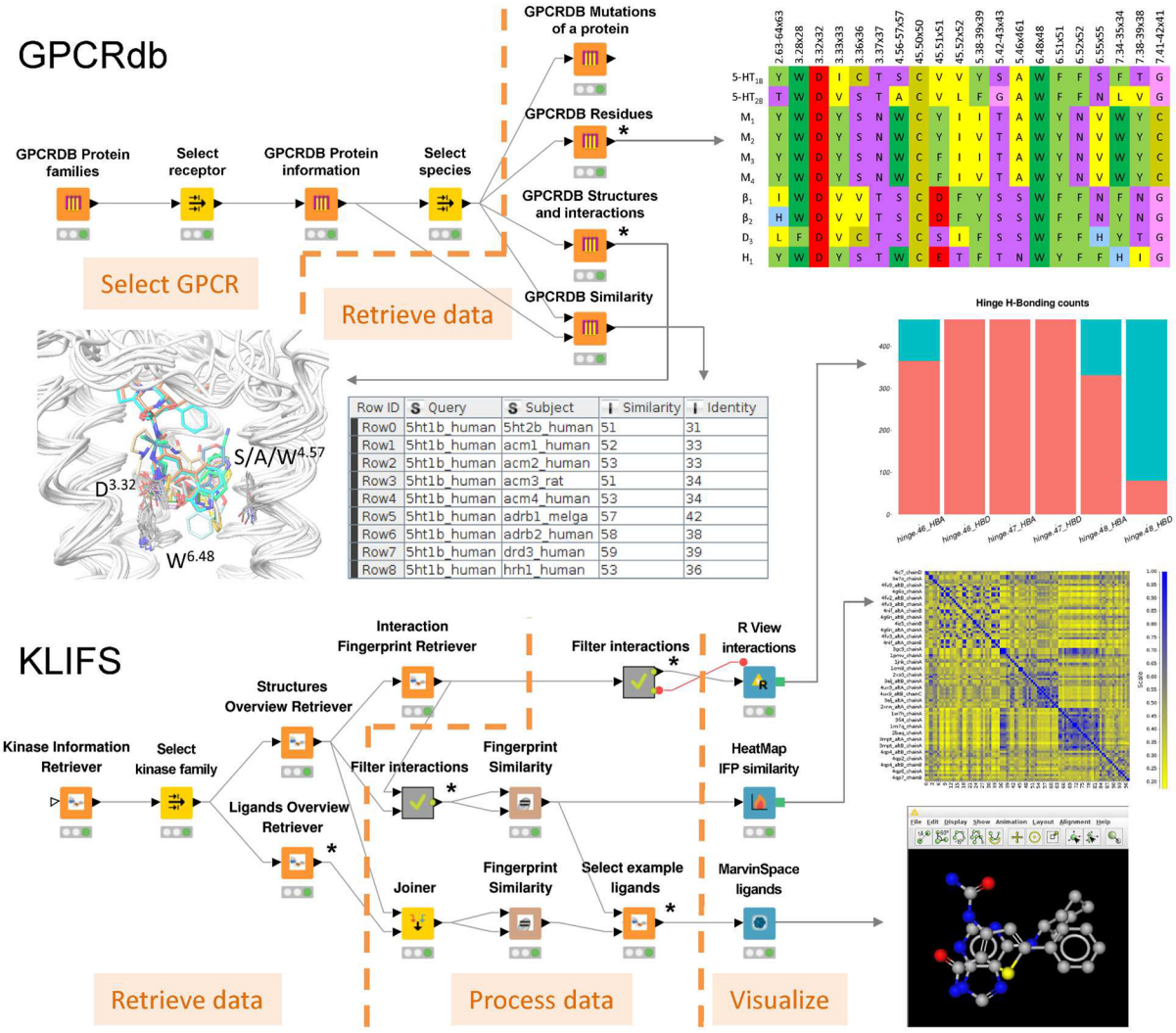
INTRODUCTION

In the postgenomic era, data generation in the pharmaceutical sciences has massively accelerated and new analytical eScience approaches are needed to adequately exploit this new chemical and biological information.[1,2] Open source cheminformatics tools are available to generate, annotate, and visualize structures of small molecules and calculate chemical descriptors and fingerprints for their comparison and the identification of structure−property or structure−activity relationships.[3−12] These tools are available in various forms, often as libraries or extensions to widely used environments such as R,[13] Python,[14] or Java.[15] Data analytics platforms such as KNIME[16] allow the combination of bioinformatics and cheminformatics tools[17,18] and integration of the growing amount of publically available chemical, structural, and biological data from ChEMBL,[19] PubChem,[20] BindingDB,[21] and PDB.[22] KNIME has emerged as a widely used open source data mining tool, and the KNIME repository contains configurable nodes to perform a wide variety of functions that can be combined in customizable data analytics workflows.[16−18] The standard KNIME nodes, together with those supplied by the user community,[18] allow access to the functionality of several cheminformatics tools including RDKit,[3] CDK,[4,10] ChemAxon,[7] Erlwood,[18] Indigo,[8] and OpenBabel.[9] The EMBL-EBI[23] and Vernalis nodes,[18] provide access to ChEMBL and PDB, respectively, and the OpenPhacts[24] (ChemBioNavigator,[25] PharmaTrek[26]) nodes allow the mining of yet more heterogeneous data.

The majority of the aforementioned KNIME nodes concentrate on small molecule cheminformatics. We have developed new cheminformatics and bioinformatics tools that provide detailed information on the structural interactions between small molecule ligands and their biological macromolecular targets (http://3d-e-chem.github.io) and incorporated these tools in an open source Virtual Machine, 3D-e-Chem-VM, that makes use of the KNIME infrastructure. 3D-e-Chem-VM consists of software libraries, workflow tools, and databases that allow interoperability of different chemical and biological data formats, enabling the analysis and integration of small molecule and protein structural information in the graphical programming environment of KNIME. The VM facilitates efficient implementation and updating of installation prerequisites and dependencies. The new cheminformatics tools, KNIME nodes, and data analytics workflows enable efficient data mining from established structural (PDB[22]) and bioactivity (ChEMBL[19]) databases as well as customized G protein-coupled receptor (GPCRdb[27]) and protein kinase (KLIFS[28,29]) focused data resources. The cheminformatics toolbox allows the design of

Received: November 16, 2016
Published: January 26, 2017

DOI: 10.1021/acs.jcim.6b00686
J. Chem. Inf. Model. 2017, 57, 115−121



Row ID	Query	Subject	Similarity	Identity
Row0	5ht1b_human	5ht2b_human	51	31
Row1	5ht1b_human	acm1_human	52	33
Row2	5ht1b_human	acm2_human	53	33
Row3	5ht1b_human	acm3_rat	51	34
Row4	5ht1b_human	acm4_human	53	34
Row5	5ht1b_human	adrb1_melga	57	42
Row6	5ht1b_human	adrb2_human	58	38
Row7	5ht1b_human	drd3_human	59	39
Row8	5ht1b_human	hrh1_human	53	36

Figure 1. KNIME workflows to exploit cheminformatics and bioinformatics information on GPCRs (GPCRdb nodes) and protein kinases (KLIFS nodes). In the GPCRdb workflow, KNIME nodes are used to enable the extraction and combination of protein information, sequence, alternative numbering schemes, mutagenesis data, and experimental structures for a selected receptor from GPCRdb. The lower branch of the workflow returns all sequence identities and similarities of the TM domain for the selected receptors and can be used for further structural chemogenomics analyses[44] using, e.g., structural and structure-based sequence alignments of the ligand binding site residues of crystallized aminergic receptors (available in the VM as a PyMOL session). In the KLIFS workflow, KNIME nodes enable the integrated analysis of structural kinase−ligand interactions from all structures for a specific kinase in KLIFS (human MAPK in the example). Kinase−ligand complexes with a specific hydrogen bond interaction pattern between the ligand and residues in the hinge region of the kinase (stacked bar chart) are selected for an all-against-all comparison of their structural kinase−ligand interactions fingerprints (heat map). The ligands from the selected structures are compared and the ligand pair with the lowest chemical similarity and a high interaction fingerprint similarity are retrieved from KLIFS for binding mode comparison. Meta nodes in the workflows in panels A and B are indicated with a star (*). The full workflows are provided in the Supporting Information, Figures S2 and S3.

customizable workflows for virtual screening, off-target prediction, and ligand design, including bioisostere detection based on protein−ligand interaction pharmacophore features (KRIPO[30]) and consideration of ligand-based metabolite prediction (SyGMa[31]). The integrated structural cheminformatics infrastructure enables large-scale structural chemogenomics studies, where protein−ligand binding interaction and bioactivity data are considered across multiple ligands and targets.

3D-e-Chem-VM. KNIME, PostgreSQL,[32] and chemistry-aware open source tools were integrated to become the backbone of a desktop cheminformatics infrastructure (Supporting Information, Figure S1). This system has been augmented by new tools to use structural protein−ligand interaction data from KRIPO,[30] GPCRdb,[27] and KLIFS[28,29] databases and has been made publically available on GitHub (http://3d-e-chem.github.io). The previously reported myChEMBL VM[33] provided a useful template to design the 3D-e-Chem-VM and a local copy of the ChEMBL database[19] can optionally be incorporated into the VM (https://github.com/3D-e-Chem/3D-e-Chem-VM/wiki/Datasets#chembl). The 3D-e-Chem-VM is available in the Vagrant[34] box catalog of HashiCorp called Atlas.[35] The Vagrant box is automatically constructed using Packer,[36] which creates a VirtualBox[37] machine image, installs Lubuntu, and finally executes our Ansible[38] playbooks to install all the additional software and enhancements (Supporting Information, Figure S1). To obtain a copy of the 3D-e-Chem-VM on a local PC, the user installs VirtualBox and Vagrant, then downloads the Vagrant box, and starts the VM by running two Vagrant commands: "vagrant init nlesc/3d-e-chem" then "vagrant up". New

DOI: 10.1021/acs.jcim.6b00686
J. Chem. Inf. Model. 2017, 57, 115−121

functionalities implemented in later 3D-e-Chem-VM releases can be installed using the command "sudo vagrant_upgrade" from a terminal inside the VM. The GPCRdb, KLIFS, KRIPOdb, and SyGMa KNIME nodes included in the 3D-e-Chem-VM are built and tested automatically on the continuous integration platform Travis-CI[39] every time a change is pushed to the Github code repository.[40] The KNIME node development procedure[41] to generate a skeleton, write the code, run tests, and deploy the nodes via the Eclipse User Interface was automated using Tycho[40] based Eclipse plug-ins. The 3D-e-Chem KNIME nodes are tested for KNIME version compatibility (specified in the node config file) and if necessary will be adapted to comply with future KNIME releases. The 3D-e-Chem-VM requires at least 2 GB RAM memory to run, 16 GB of disk space, and the CPU must have virtualization support. The 3D-e-Chem tools and workflows are available for use in any environment as long as the dependencies and prerequisites are correctly installed and configured. The 3D-e-Chem-VM further facilitates the use of the 3D-e-Chem tools and other resources (Supporting Information, Figure S1) by taking care of these dependencies and prerequisites, including the preconfiguration of (i) Python[14] and R[13] packages to facilitate the use of KNIME nodes and workflows, (ii) scripts to set up infrastructures that allow data mining of locally installed databases like the Postgresql[32] and RDKit[3] Postgresql cartridge to exploit a local copy of ChEMBLdb,[19] (iii) additional cheminformatics modeling and visualization software (e.g., PyMOL,[6] Camb,[11] and fpocket[42]), and (iv) OpenPHACTS KNIME functionalities[43] and the new GPCRdb, KLIFS, and KRIPO KNIME nodes to interact with local files and Web servers.

GPCRdb Nodes. GPCRs are the largest group of signal transducing membrane proteins and hence one of the most important target family for drugs that can stimulate, reduce, or block endogenous GPCR activity. GPCR structural chemogenomic analyses require the integration of phylogenetic, sequence, and structure similarity and ligand binding information.[44,45] GPCRdb (http://gpcrdb.org, accessed 25 August 2016) is an online repository of the accumulated knowledge on GPCRs including structure-based annotation of protein sequence alignments of 18 787 sequences of 421 receptor subtypes and of 3096 species, analysis of 142 GPCR crystal structures and GPCR-ligand interactions, and 14 099 mutational data points.[27] For the integration of this data in customizable workflows for systematic structural chemogenomics analyses we have developed seven KNIME nodes that interface with GPCRdb via a web service client generated with Swagger Code Generator.[46] An example workflow utilizing these nodes is shown in Figure 1.

- GPCRDB Protein Families: Extraction of protein family information, including the protein names and classifications of all GPCRs in the four-level hierarchy defined by GPCRdb (class, ligand type, subfamily, subtype).
- GPCRDB Protein Information: Retrieval of source, species, and sequence data from UniProt identifiers or protein family identifier.
- GPCRDB Protein Residues: Retrieval of residues and numbering schemes. This node retrieves all residues of the specified protein with secondary structure annotation, UniProt numbering, and GPCR residue numbering.[47]
- GPCRDB Structures of a Protein: Retrieval of experimental GPCR structures with literature references, PDB codes, and ligands.
- GPCRDB Mutations of a Protein: Retrieval of single point mutations in GPCRs, including the sequence position, mutation, ligand, assay type, mutation effect, protein expression information, and publication reference.
- GPCRDB Structure−Ligand Interactions: Returns the sequence numbers of amino acid residues interacting with ligands in the specified PDB entry. The interaction type is annotated in the output table.
- GPCRDB Protein Similarity: Returns the sequence identity and similarity of a query receptor versus a set of receptors, based on the full sequence or a specified set of residues.

KLIFS Nodes. Protein kinases are important signal pathway regulators and comprise one of the largest protein families that are encoded within the human genome. The KLIFS database (http://klifs.vu-compmedchem.nl, accessed 25 August 2016)[28,29] contains detailed structural kinase−ligand interaction information derived from 3354 structures of catalytic domains of human and mouse protein kinases deposited in the PDB in order to map the structural determinants of kinase−ligand binding and selectivity. To leverage this information for structural chemogenomics analyses we have developed nine KNIME nodes that interface with KLIFS via a web service client generated with Swagger Code Generator.[46] An example workflow of the KLIFS KNIME nodes is shown in Figure 1.

KLIFS Information Nodes.

- Kinase ID Mapper: Maps a user-supplied set of kinase names (names according to Manning et al.[48]), HGNC gene symbols, or UniProt accession codes to a KLIFS kinase ID. The output also contains all related kinase information present within KLIFS (see "Kinase Information Retriever").
- Kinase Information Retriever: Returns a table comprising the KLIFS kinase ID, kinase name, HGNC symbol, kinase group, kinase family, kinase class, species, full name, UniProt accession code, IUPHAR ID, and the amino acid sequence of the pocket based on the KLIFS pocket definition using a consistent alignment of 85 residues.

KLIFS Interactions Nodes.

- Interaction Fingerprint Decomposer: Decomposes a protein−ligand interaction fingerprint (IFP)[49] into a human-readable table with annotated interactions for each structure. This node can optionally add the sequence number and the KLIFS residue position[29] for each pocket residue to the table.
- Interaction Fingerprint Retriever: Retrieval of the interaction fingerprint of specific kinase-ligand complexes from KLIFS. The fingerprint has been corrected for gaps/missing residues within the KLIFS pocket thereby enabling all-against-all comparisons.
- Interaction Types Retriever: Retrieves the different interaction types for each bit position of the interaction fingerprint method and can be used in combination with the interaction fingerprint decomposer to identify which kinase−ligand interactions are present in a given set of kinase structures.

KLIFS Ligands Nodes.

- Ligands Overview Retriever: Retrieval of ligand IDs, three-letter PDB-codes, names, molecular structures (SMILES), and InChIKeys for all ligands from (a specific set of) kinase-ligand complexes present within KLIFS.

DOI: 10.1021/acs.jcim.6b00686
J. Chem. Inf. Model. 2017, 57, 115−121

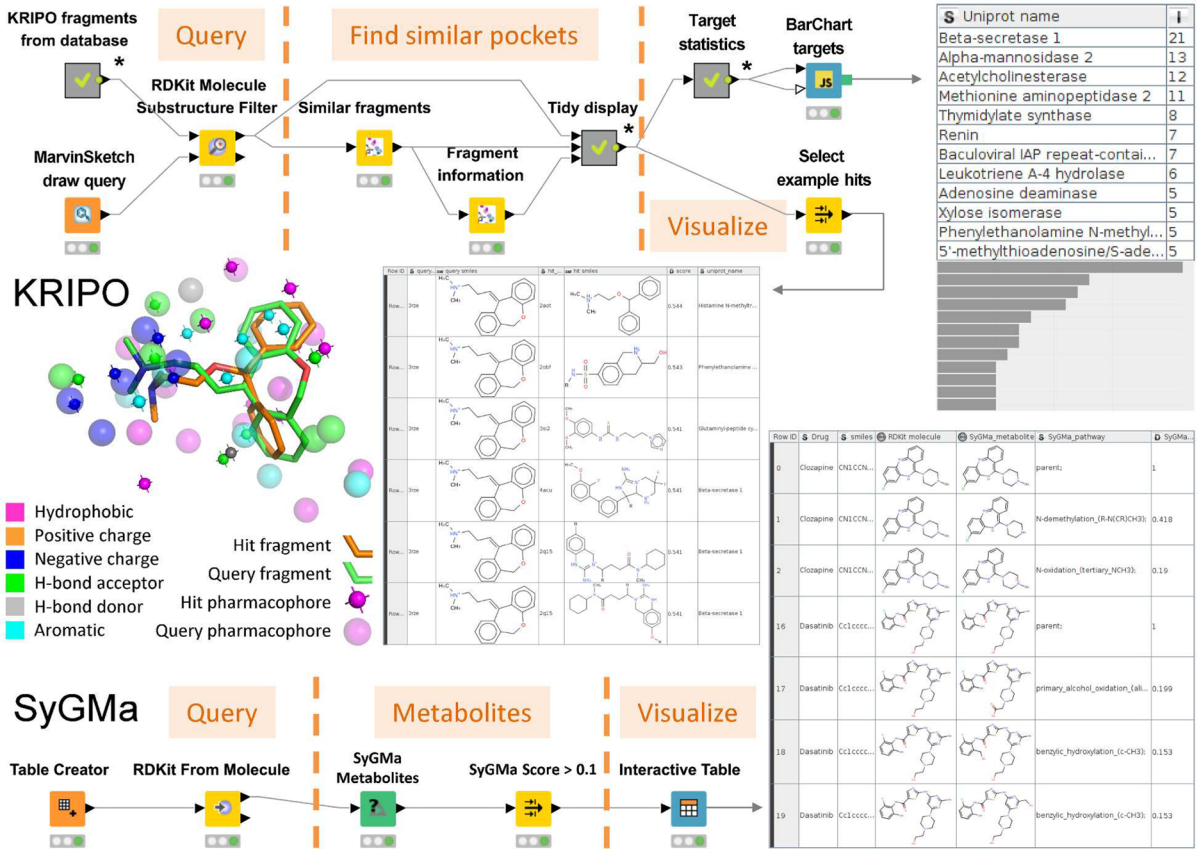


Figure 2. KRIPO binding site similarity based bioisosteric replacement and SyGMa metabolite prediction workflows. Ligands in KRIPOdb that share a chemical (sub)structure with a specified molecule (doxepin in the example) are identified and defined as query fragment(s). Ligand (fragment) binding site hits that share pharmacophore fingerprint similarity with the binding site(s) associated with the query fragment(s) (e.g., the doxepin binding site of the histamine H_1 receptor) are identified and ranked according to Tanimoto similarity score. The occurrence of protein targets in the top hit list is analyzed. The pharmacophore overlay underlying the similarity value of an example hit (histamine methyltransferase, PDB ID: 2aot; available in the VM as a PyMOL session). The full workflow is provided in the Supporting Information (Figure S4). In the SyGMa workflow Smiles strings of clozapine and dasatinib are converted into RDKit molecules for the prediction of metabolites using the SyGMa Metabolites node, filtered based on a SyGMa_score threshold of 0.1. The two tables are subsections of the resulting table, showing the top ranked metabolites of clozapine and dasatinib, consistent with experimental metabolism data.[51,52] Meta nodes are indicated with a star (*).

KLIFS structures nodes.

- Structures Overview Retriever: Retrieves a list of all corresponding structures within KLIFS based on a user-supplied set of KLIFS kinase or ligand IDs (e.g., from a specific kinase family). The node returns the structure ID, kinase name, kinase ID, PDB-code, and all other structural annotation data within KLIFS (e.g., pocket sequence, resolution, quality, ligands, DFG conformation, targeted subpockets, waters).[29]
- Structures PDB Mapper: Maps a set of PDB-codes to structure IDs from KLIFS and provides all related structural information from KLIFS.
- Structures Retriever (MOL2): Retrieves from KLIFS a set of structures, (optionally the full complex, the protein, the pocket, or the ligand) in MOL2 format, based on a user-supplied set of Structure IDs. As output the node provides a table of aligned structures based on the KLIFS pocket definition.

KRIPOdb and KRIPO Nodes. The KRIPOdb includes an SQLite database with more than 2.3×10^{11} pairwise ligand binding site similarity scores based on KRIPO pharmacophore fingerprints[30] of 483 083 subpockets associated with the substructures (fragments) of small-molecule ligands identified in the binding sites of all PDB entries released until 29 June 2016. The full similarity matrix is available as a web service (http://3d-e-chem.vu-compmedchem.nl/kripodb/ui/), whereas a similarity matrix calculated between all crystallized GPCRs and the whole PDB above a similarity threshold of 0.45 (calculated as a modified Tanimoto similarity score[50]) is included in the 3D-e-Chem-VM as compact HDF5 file. The KRIPO Python library with a command line interface is provided inside the VM to extract and manipulate fragment structural data in KRIPOdb. We have developed the following two KNIME nodes to efficiently extract and integrate the information in KRIPOdb.

- Similar Fragments: Retrieval of ligand fragments that share a similar subpocket with the query fragment, based on a specified similarity matrix (local HDF5 file or web service URL), similarity threshold, and maximum number of fragment hits.
- Fragment Information: Retrieval of the chemical structures of the fragment, the full ligand, and the associated PDB based on the fragment identifier.

Figure 2 presents an example KRIPO KNIME workflow to identify similar ligand binding sites (for e.g. off-target prediction) and search for bioisosteric replacements based on ligand binding site similarity.

DOI: 10.1021/acs.jcim.6b00686
J. Chem. Inf. Model. 2017, 57, 115−121

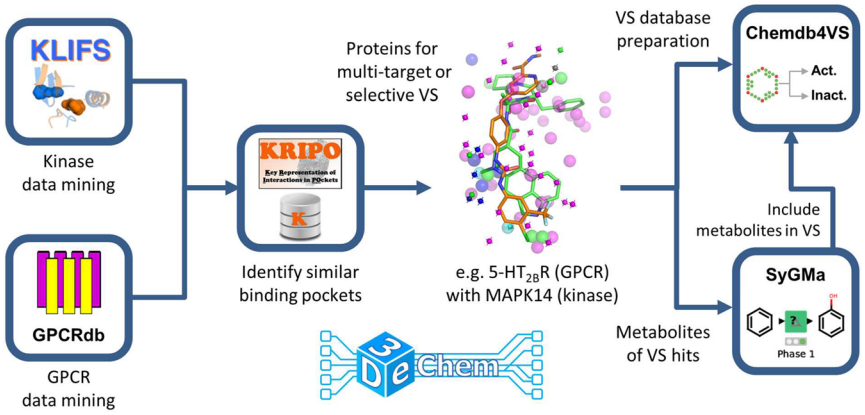


Figure 3. Schematic diagram of possible interactions of the 3D-e-Chem-VM virtual machine elements: KLIFS and GPCRdb web service connector nodes, KRIPOdb, KRIPO, and SyGMa nodes, and the Chemdb4VS workflow (full workflow presented in the Supporting Information, Figure S6) integrated in a GPCR-kinase cross-reactivity prediction workflow.

SyGMa Node. For the assessment or prediction of a complete pharmacological profile, the metabolites of a drug molecule need to be taken into account. SyGMa is a rule-based method for **sy**stematic **g**eneration of potential **meta**bolites.[31] We have developed a SyGMa KNIME node thin wrapper around the SyGMa[31] Python library that enables straightforward generation of the structures of possible metabolites of a specified molecule. The *SyGMa Metabolites* node generates putative metabolites based on the 2D coordinates of molecules in RDKit format, and the definition of the number of phase 1 and phase 2 metabolism cycles in the node dialogue. The SyGMa_metabolite output column contains the resulting metabolite structures, including the parent, ordered by decreasing probability score. The generated 2D chemical structures are aligned to atomic coordinates of the parent, which facilitates visual inspection of the metabolic modifications. The SyGMa_pathway column lists the metabolic reaction rules that were applied to result in the given metabolite structure. The SyGMa_score column lists the probability score, which can be used to filter the results. Figure 2 shows a simple workflow to predict the metabolites for the GPCR antagonist clozapine and kinase inhibitor dasatinib.

3D-e-Chem Workflow Application Example 1: Kinase Interaction Pattern Analysis. In the KLIFS workflow (Figure 1) information on all 14 human MAPK kinases with crystal structure data is retrieved from KLIFS (478 monomers from 312 unique PDB structures). Subsequently, for each MAPK kinase−ligand complex the interaction fingerprints (IFPs), describing the interactions between the residues in the binding site of the enzyme and the ligand, are downloaded. From these IFPs the H-bond donor and acceptor interaction frequency with the hinge region of the kinases are summarized in a stacked bar chart. The IFPs are then filtered to obtain only those kinase−ligand complexes in which the ligand has an H-bond donor for residue hinge.46 (gatekeeper + 1) and an H-bond acceptor for residue hinge.48 (gatekeeper + 3). In 98 of the 478 monomers (58 unique PDB structures), this interaction pattern with the hinge region is observed. The interaction pattern similarity for these monomers is calculated using the Tanimoto coefficient (Tc) on the IFPs as visualized in a heat map, showing that overall IFP similarity is relatively low despite their shared hinge interaction pattern. Finally, this group of monomers is used to identify structures with a high IFP similarity but low structural similarity of the ligands. To this end, the molecular structures of the ligands are obtained and compared to each other using the ECFP-4[53] fingerprint and the Tanimoto coefficient. Subsequently, the IFP and ligand similarity matrices are combined to select the structure pair with a high IFP similarity[54] (Tc $\geq$ 0.75) and the lowest chemical similarity (PDB IDs 3pze and 4qp4, ECFP-4 similarity: 0.07, IFP similarity: 0.76). The 3D ligand binding modes are downloaded from KLIFS and shown in the 3D-viewer MarvinSpace. This workflow can, among others, be used for scaffold hopping purposes by identifying ligands with a high IFP similarity, but a relatively low chemical similarity. For example, the structures with PDB IDs 3gc8 (MAPK11) and 3fl4 (MAPK14) contain ligands that are chemically different (ECFP-4 similarity: 0.2) but share similar binding modes (IFP similarity: 0.76), identifying the pyrazolopyrimidine (3fl4) to dihydroquinazolinone (3gc8) scaffold hop as an interesting design strategy to obtain kinase inhibitors with similar structural interaction patterns.[55]

3D-e-Chem Workflow Application Example 2: GPCR-Kinase Cross-Reactivity Prediction. A workflow combining different 3D-e-Chem functionalities was created to illustrate their integration and applicability for structural chemogenomics studies across different protein families. The full GPCR-kinase cross-reactivity prediction workflow for off-target identification, ligand repurposing, or the discovery of ligands with a desired GPCR-kinase polypharmacological profile is shown in Supporting Information Figure S5. In this workflow the GPCRdb and KLIFS nodes are used to fetch all experimentally determined structures of ligand-protein complexes in the two drug target families. The KRIPO nodes are subsequently used to assess the structure-based pharmacophore similarity between all GPCR and kinase binding sites, yielding 1428 similar GPCR-kinase pairs (modified Tanimoto coefficient[50] >0.5). The analysis for example identified the similar ergotamine bound serotonin 5-HT_{2B} receptor (PDB: 4ib4) and Sorafenib bound MAPK14 (PDB: 3heg, IC_{50} = 57 nM) binding site pair (modified Tc = 0.55), which is consistent with the recent experimental identification of Sorafenib as a high affinity 5-HT_{2B} ligand (K_i = 56 nM).[56]

Combination of the KRIPO pharmacophore similarity assessment and a systematic ChEMBL database[19] search indicated for example that the 5-HT_{2B} receptor also shares a similar binding site and experimentally evaluated ligands with several other kinases, including CDK8, ABL1, DDR1, FGFR1, KIT, HCK, VGFR2, and B-raf. The MAPK14 kinase furthermore shares high binding site similarity and experimentally validated

DOI: 10.1021/acs.jcim.6b00686
J. Chem. Inf. Model. 2017, 57, 115−121

ligands with the adenosine A_{2A}[57,58] and smoothened (SMOR)[59] G protein-coupled receptors, amongst others. The computationally predicted kinase-GPCR pairs offer opportunities for the rational identification and design of ligands with well-defined polypharmacological profiles.[60] The kinase-GPCR cross-reactivity workflow can for example be complemented by the Chemdb4VS workflow for the evaluation and optimization of virtual screening strategies to identify selective or multitarget ligands (Figure 3). In addition, the SyGMa metabolite predictor node can be used to enumerate potential metabolites of ligands identified for drug repurposing or of hits identified in virtual screening (Figure 3).

The 3D-e-Chem-VM provides preconfigured starting points that can be easily adapted to construct flexible structural chemogenomics analysis and drug design workflows using the 3D-e-Chem structural cheminformatics research tools.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00686.

Figures presenting the full versions of the GPCRdb, KLIFS, KRIPO, SyGMa, Chemdb4VS, and GPCR-kinase cross-reactivity prediction example KNIME workflows (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: ross.mcguire@bioaxisresearch.com (R.McG.)
*E-mail: S.Verhoeven@esciencecenter.nl (S.V.)
*E-mail: c.de.graaf@vu.nl (C.d.G.).

ORCID

Ross McGuire: 0000-0003-3404-6456
Stefan Verhoeven: 0000-0002-5821-2060
Márton Vass: 0000-0003-1486-0063
Albert J. Kooistra: 0000-0001-5514-6021
Chris de Graaf: 0000-0002-1226-2150

Author Contributions

#R.McG, S.V., and M.V. contributed equally.

Funding

Netherlands eScience Center/NWO (3D-e-Chem, grant 027.014.201). M.V., R.L., G.V., I.J.P.d.E., A.J.K., and C.d.G. participate in the COST Action CM1207 (GLISTEN). M.V., I.J.P.d.E, R.L., and C.d.G. participate in the GPCR Consortium (gpcrconsortium.org).

Notes

The authors declare no competing financial interest.

Downloads and documentation of the 3D-e-Chem VM, GPCRdb, KLIFS, KRIPO, SyGMa, and Chemdb4VS KNIME nodes and workflows, as well as other 3D-e-Chem tools and databases are accessible from http://3d-e-chem.github.io.

■ ACKNOWLEDGMENTS

Vignir Isberg, Christian Munk, and David Gloriam from University of Copenhagen for useful discussions on the developments of the GPCRdb KNIME nodes.

■ REFERENCES

(1) Hu, Y.; Bajorath, J. Learning from 'big data': compounds and targets. *Drug Discovery Today* **2014**, *19*, 357−60.
(2) Lusher, S. J.; McGuire, R.; van Schaik, R. C.; Nicholson, C. D.; de Vlieg, J. Data-driven medicinal chemistry in the era of big data. *Drug Discovery Today* **2014**, *19*, 859−68.
(3) RDKit. http://www.rdkit.org.
(4) Steinbeck, C. C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.
(5) Jmol. http://jmol.sourceforge.net/.
(6) Pymol. https://www.pymol.org/.
(7) ChemAxon. https://www.chemaxon.com/.
(8) Indigo. http://lifescience.opensource.epam.com/indigo/.
(9) O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open babel: an open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
(10) Beisken, S.; Meinl, T.; Wiswedel, B.; de Figueiredo, L. F.; Berthold, M.; Steinbeck, C. KNIME-CDK: Workflow-driven cheminformatics. *BMC Bioinf.* **2013**, *14*, 257.
(11) Murrell, D. S.; Cortes-Ciriano, I.; van Westen, G. J.; Stott, I. P.; Bender, A.; Malliavin, T. E.; Glen, R. C. Chemically Aware Model Builder (camb): an R package for property and bioactivity modelling of small molecules. *J. Cheminf.* **2015**, *7*, 45.
(12) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. Datawarrior: An Open-Source Program for Chemistry Aware Data Visualization and Analysis. *J. Chem. Inf. Model.* **2015**, *55*, 460−473.
(13) R Core Team. R: *A language and environment for statistical computing*; R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
(14) Python. http://www.python.org.
(15) Java. https://www.oracle.com/java/index.html.
(16) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*; Springer Berlin Heidelberg, 2007; pp 319−326.
(17) Mazanetz, M. P.; Marmon, R. J.; Reisser, C. B.; Morao, I. Drug Discovery Applications for KNIME: An Open Source Data Mining Platform. *Curr. Top. Med. Chem.* **2012**, *12*, 1965−1979.
(18) KNIME Cheminformatics Extensions. https://tech.knime.org/cheminformatics-extensions.
(19) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083−1090.
(20) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2016**, *44*, D1202−1213.
(21) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.
(22) Berman, H. M.; W, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(23) Papadatos, G.; van Westen, G. J.; Croset, S.; Santos, R.; Trubian, S.; Overington, J. P. A document classifier for medicinal chemistry publications trained on the ChEMBL corpus. *J. Cheminf.* **2014**, *6*, 40.
(24) Williams, A. J.; Harland, L.; Groth, P.; Pettifer, S.; Chichester, C.; Willighagen, E. L.; Evelo, C. T.; Blomberg, N.; Ecker, G.; Goble, C.; Mons, B. Open PHACTS: semantic interoperability for drug discovery. *Drug Discovery Today* **2012**, *17*, 1188−1198.
(25) Stierand, K.; Harder, T.; Marek, T.; Hilbig, M.; Lemmen, C.; Rarey, M. The Internet as Scientific Knowledge Base: Navigating the Chem-Bio Space. *Mol. Inf.* **2012**, *31*, 543−546.
(26) Carrascosa, M. C.; Massaguer, O. L.; Mestres, J. PharmaTrek: A Semantic Web Explorer for Open Innovation in Multitarget Drug Discovery. *Mol. Inf.* **2012**, *31*, 537−541.
(27) Isberg, V.; Mordalski, S.; Munk, C.; Rataj, K.; Harpsøe, K.; Hauser, A. S.; Vroling, B.; Bojarski, A. J.; Vriend, G.; Gloriam, D. E.

DOI: 10.1021/acs.jcim.6b00686
J. Chem. Inf. Model. 2017, 57, 115−121

GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **2016**, *44*, D356−D364.
(28) van Linden, O. P.; Kooistra, A. J.; Leurs, R.; de Esch, I. J.; de Graaf, C. KLIFS: a knowledge-based structural database to navigate kinase−ligand interaction space. *J. Med. Chem.* **2014**, *57*, 249−277.
(29) Kooistra, A. J.; Kanev, G. K.; van Linden, O. P.; Leurs, R.; de Esch, I. J.; de Graaf, C. KLIFS: a structural kinase-ligand interaction database. *Nucleic Acids Res.* **2016**, *44*, D365−371.
(30) Wood, D. J.; de Vlieg, J.; Wagener, M.; Ritschel, T. Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. *J. Chem. Inf. Model.* **2012**, *52*, 2031−2043.
(31) Ridder, L.; Wagener, M. SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites. *ChemMedChem* **2008**, *3*, 821−32.
(32) Postgresql. https://www.postgresql.org/.
(33) Ochoa, R.; Davies, M.; Papadatos, G.; Atkinson, F.; Overington, J. P. myChEMBL: a virtual machine implementation of open data and cheminformatics tools. *Bioinformatics* **2014**, *30*, 298−300.
(34) https://www.vagrantup.com/.
(35) https://atlas.hashicorp.com/boxes/search.
(36) https://www.packer.io/.
(37) https://www.virtualbox.org/.
(38) http://www.ansible.com.
(39) Travis-CI. https://travis-ci.org/.
(40) http://www.eclipse.org/tycho/.
(41) KNIME Developer Guide. https://tech.knime.org/developer-guide.
(42) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinf.* **2009**, *10*, 168.
(43) OPS-KNIME. https://github.com/openphacts/OPS-Knime.
(44) Kooistra, A. J.; Kuhne, S.; de Esch, I. J.; Leurs, R.; de Graaf, C. A structural chemogenomics analysis of aminergic GPCRs: lessons for histamine receptor ligand design. *Br. J. Pharmacol.* **2013**, *170*, 101−26.
(45) Vass, M.; Kooistra, A. J.; Ritschel, T.; Leurs, R.; de Esch, I. J.; de Graaf, C. Molecular interaction fingerprint approaches for GPCR drug discovery. *Curr. Opin. Pharmacol.* **2016**, *30*, 59−68.
(46) http://swagger.io/swagger-codegen.
(47) Isberg, V.; de Graaf, C.; Bortolato, A.; Cherezov, V.; Katritch, V.; Marshall, F. H.; Mordalski, S.; Pin, J. P.; Stevens, R. C.; Vriend, G.; Gloriam, D. E. Generic GPCR residue numbers - aligning topology maps while minding the gaps. *Trends Pharmacol. Sci.* **2015**, *36*, 22−31.
(48) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298*, 1912−1934.
(49) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195−207.
(50) Fligner, M. A.; Verducci, J. S.; Blower, P. E. A modification of the Jaccard−Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics* **2002**, *44*, 110−119.
(51) Nijmeijer, S.; Vischer, H. F.; Rudebeck, A. F.; Fleurbaaij, F.; Falck, D.; Leurs, R.; Niessen, W. M.; Kool, J. Development of a profiling strategy for metabolic mixtures by combining chromatography and mass spectrometry with cell-based GPCR signaling. *J. Biomol. Screening* **2012**, *17*, 1329−38.
(52) Wang, L.; Christopher, L. J.; Cui, D.; Li, W.; Iyer, R.; Humphreys, W. G.; Zhang, D. Identification of the human enzymes involved in the oxidative metabolism of dasatinib: an effective approach for determining metabolite formation kinetics. *Drug Metab. Dispos.* **2008**, *36*, 1828−39.
(53) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−54.
(54) Kooistra, A. J.; Vischer, H. F.; McNaught-Flores, D.; Leurs, R.; de Esch, I. J.; de Graaf, C. Function-specific virtual screening for GPCR ligands using a combined scoring method. *Sci. Rep.* **2016**, *6*, 28288.
(55) Astolfi, A.; Iraci, N.; Manfroni, G.; Barreca, M. L.; Cecchetti, V. A Comprehensive Structural Overview of p38alpha MAPK in Complex with Type I Inhibitors. *ChemMedChem* **2015**, *10*, 957−69.
(56) Lin, X.; Huang, X. P.; Chen, G.; Whaley, R.; Peng, S.; Wang, Y.; Zhang, G.; Wang, S. X.; Wang, S.; Roth, B. L.; Huang, N. Life beyond kinases: structure-based discovery of sorafenib as nanomolar antagonist of 5-HT receptors. *J. Med. Chem.* **2012**, *55*, 5749−59.
(57) DRUGMATRIX: Adenosine A2A radioligand binding assay (ligand: AB-MECA) CHEMBL1909214.
(58) Dombroski, M. A.; Letavic, M. A.; McClure, K. F.; Barberia, J. T.; Carty, T. J.; Cortina, S. R.; Csiki, C.; Dipesa, A. J.; Elliott, N. C.; Gabel, C. A.; Jordan, C. K.; Labasi, J. M.; Martin, W. H.; Peese, K. M.; Stock, I. A.; Svensson, L.; Sweeney, F. J.; Yu, C. H. Benzimidazolone p38 inhibitors. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 919−23.
(59) Yang, B.; Hird, A. W.; Russell, D. J.; Fauber, B. P.; Dakin, L. A.; Zheng, X.; Su, Q.; Godin, R.; Brassil, P.; Devereaux, E.; Janetka, J. W. Discovery of novel hedgehog antagonists from cell-based screening: Isosteric modification of p38 bisamides as potent inhibitors of SMO. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 4907−11.
(60) Peters, J. U. Polypharmacology - foe or friend? *J. Med. Chem.* **2013**, *56*, 8955−71.

DOI: 10.1021/acs.jcim.6b00686
J. Chem. Inf. Model. 2017, 57, 115−121