**HitPickV2: a Web Server to predict targets of chemical compounds**

Sabri Hamad 2,# , Gianluca Adornetto 2 # , Jesús Naveja 2,3,4, Aakash Chavan Ravindranath 2, Johannes Raffler 2, Monica Campillos 1,2,\*

1German Center for Diabetes Research, Neuherberg, 85764, Germany,

2Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Neuherberg, 85764, Germany,

3PECEM, Faculty of Medicine, UNAM, Mexico City, 04510. Mexico,

4DIFACQUIM, Faculty of Chemistry, UNAM, Mexico City, 04510, Mexico

\*To whom correspondence should be addressed.

#These authors contribute equally to this work.

5Current address: Feral GmbH, c/o CoLaborator (Bayer), Building S141, Muellerstr. 178 | 13353 | Berlin

**Supplementary Material**

**Supplementary Methods**

**Compound-protein interaction database**

Compound-protein interaction data were obtained from public databases storing quantitative (i.e., ChEMBL 22.1 (Gaulton, et al., 2017), PDSP Ki Database (Roth, et al., 2000), Binding DB (Gilson, et al., 2016)) and qualitative (i.e., DrugBank 5.0.2 (Wishart, et al., 2006), LigandExpo (Feng, et al., 2004), T3DB (Wishart, et al., 2015) and TTD 4.3.02 ) molecular activity of compounds. Compound-protein associations in quantitative datasets were considered positive when the logarithmic potency was 5 or higher (e.g., pIC50, pKi50 :>= 5). Associations with contradictory activity information were removed from the dataset.

Molecule SMILES of compounds from the aforementioned databases were preprocessed and standardized using CDK version 1.5.12 and RDKit version 3.2.4 nodes in KNIME 3.3. Pan-Assay Interference compounds (PAINS) were removed using a published KNIME workflow (Saubern, et al., 2011). SMILES were converted to INCHI keys using RDKit nodes in KNIME 3.3. We remove redundancy of compounds by merging molecules with identical INCHI keys, DrugBank or ChEMBL IDs (if available) and with high structure and functional similarity (defined as Extended Connectivity Fingerprints diameter 4 Tc> 0.7 and target similarity Tc > 0.7).

**Connection between protein Targets and Protein Complexes.**

Information of composition of protein complexes was downloaded from Reactome (Fabregat, et al., 2018) database (file ComplexParticipantsPubMedIdentifiers\_Human.txt from <https://reactome.org/download/>, 19th June 2018). Using this file, we connected the protein targets in HitPickV2 to Reactome protein complexes by grouping the protein targets into non-overlaping small protein complexes. For that, we started with a random HitPickV2 protein target and connected it to the smallest complex it belongs to. We then, excluded other HitPickV2 proteins belonging to this complex. We repeated this procedure until there were not more proteins left. The connection between the protein target and the Reactome complexes is included in an extra column of the output program.

**Murcko scaffolds**

Murcko scaffold’s (Bemis and Murcko, 1996) SMILES are obtained using RDKit for all compounds interacting with the same target in the restricted chemical space. We then binned the scaffolds and report on HitPickV2 website the three most abundant scaffolds along with their occurrence. The number of unique scaffolds is also indicated. Since Murcko scaffolds are not defined for linear compound, compounds without ring systems (acyclic compounds) are reported in HitPickV2 as “acyclic compounds”.

**Supplementary Table 1**

Precision (%) for the ten ranked predicted targets as a function to the three parameters of the restricted space of the query compound: the Tanimoto coefficient (Tc) between a validation compound and the most similar molecule in the training set (in green), occur and Target rank (1st to 10th predicted targets). Cells with a precision higher than 50% are marked in red. The precision for cells marked as \* was determined with a number of compound-target predictions lower than 30. NA: Precision in these cells could not be determined. Note: Tc of 1 does not mean that two molecules are necessarily identical. For predicted compound-targets pairs already annotated in our in-house compound-protein database, we assign the target with 100% target prediction precision.

****

**Supplementary Figure 1**

Comparison of the performance of HitPick V2 in an internal (cross-validation) and in an independent set. Precision is calculated across all ranges of Tc and for the 1st Target rank.

****

**References**

Bemis, G.W. and Murcko, M.A. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry* 1996;39(15):2887-2893.

Fabregat, A.*, et al.* The Reactome Pathway Knowledgebase. *Nucleic acids research* 2018;46(D1):D649-D655.

Feng, Z.*, et al.* Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics (Oxford, England)* 2004;20(13):2153-2155.

Gaulton, A.*, et al.* The ChEMBL database in 2017. *Nucleic acids research* 2017;45(D1):D945-D954.

Gilson, M.K.*, et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research* 2016;44(D1):D1045-1053.

Roth, B.L.*, et al.* The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *Neuroscientist* 2000;6(4):252-262.

Saubern, S., Guha, R. and Baell, J.B. KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries. *Mol Inform* 2011;30(10):847-850.

Wishart, D.*, et al.* T3DB: the toxic exposome database. *Nucleic acids research* 2015;43(Database issue):D928-934.

Wishart, D.S.*, et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* 2006;34(Database issue):D668-672.

Zhu, F.*, et al.* Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic acids research* 2012;40(Database issue):D1128-1136.