*Bioinformatics*, YYYY, 0–0

doi: 10.1093/bioinformatics/xxxxx

Advance Access Publication Date: DD Month YYYY

Manuscript Category

|  |
| --- |
| Structural Bioinformatics  **HitPickV2: a Web Server to predict targets of chemical compounds**  Sabri Hamad 2,# , Gianluca Adornetto 2,5 # , 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  Associate Editor: XXXXXXX  Received on XXXXX; revised on XXXXX; accepted on XXXXX  Abstract  **Motivation:** The identification of protein targets of novel compounds is essential to understand compounds’ mechanisms of action leading to biological effects. Experimental methods to determine these protein targets are usually slow, costly and time consuming. Computational tools have recently emerged as cheaper and faster alternatives that allow the prediction of protein targets for a large number of compounds.  **Results:** Here, we present HitPickV2, a novel ligand-based approach for the prediction of human druggable protein targets of multiple compounds. For each query compound, HitPickV2 predicts up to 10 targets out of 2739 human druggable proteins. To that aim, HitPickV2 identifies the closest, structurally similar compounds in a restrictedspace within a vast chemical-protein interaction area, until 10 distinct protein targets are found. Then, HitPickV2 scores these 10 targets based on three parameters: the Tanimoto coefficient (Tc) between the query and the most similar compound interacting with the target in such space, a target rank that considers Tc and Laplacian-modified naïve Bayesian target models scores and a novel parameter introduced in HitPickV2, the number of compounds interacting with each target (occur) in this space. We present the performance results of HitPickV2 in cross-validation as well as in an external dataset.  **Availability:** HitPickV2 is available in [www.hitpickv2.com](http://mips.helmholtz-muenchen.de/HitPickV2/).  **Contact:** mcampillos@gmail.com  **Supplementary information:** Supplementary data are available at *Bioinformatics* online. |

# Introduction

The identification of protein targets of novel compounds is essential to understand compounds’ mechanisms of action leading to biological effects. To determine protein targets of compounds, experimental and computational approaches are followed. The former methods are usually slow, costly and time consuming whereas the latter are cheaper and faster alternatives. Ligand-based approaches (e.g. Similarity Ensemble Approach (SEA) (Keiser, et al., 2007), PASS (Poroikov, et al., 2007), those using multiple-category Bayesian models (Nidhi, et al., 2006) (Wale and Karypis, 2009), SwissTargetPrediction (Gfeller, et al., 2014) and HitPick (Liu, et al., 2013)), are widely used computational approaches that predict targets of compounds considering that similar ligands bind to common proteins (Johnson, et al., 1990). The fast increase in the number of ligand-protein interactions deposited in public databases is expanding the druggable target repertoire predicted by ligand-based approaches, improving their suitability for the systematic molecular analysis of a larger number of compounds.

To facilitate these systems pharmacology analyses, we have developed HitPickV2 webserver, which implements an advanced version of the HitPick target prediction method (Liu, et al., 2013). HitPick V2 covers 2739 human druggable proteins (1350 more than in the former HitPick version (Liu, et al., 2013)) and allows the prediction of protein targets (up to 10 per query compound) for multiple ligands, contributing to the expansion our understanding of the biological and phenotypic activity of small molecules.

# Methods

**2.1 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). The final database consists of 431,352 unique compounds and 891,629 compound-target associations.

**2.2 HitPickV2 novel ligand-based approach**

We first generated Morgan fingerprint with feature invariants similar to Functional-Class Fingerprints (FCFP) for all compounds in our database (2.1) using the RDKit nodes implemented in KNIME 3.3. We then created Laplacian-modified naive Bayesian models (Nidhi, et al., 2006) for the 2739 druggable proteins interacting with at least 3 known ligands using KNIME 3.3 Bayesian model builder. For each protein model, we distributed compounds into an active or inactive group based on whether the compounds interact with the protein or not, respectively.

To predict targets of a query compound, HitPickV2 identifies the closest, structurally similar compounds in the chemical-protein interaction space (2.1) using k-nearest neighbours (k-NN) chemical similarity search (Schuffenhauer, et al., 2003) and selects a restricted space comprising compounds interacting with 10 distinct protein targets (Figure 1A). For each protein target, we calculated three parameters: “Tc”, “Target rank”, and “Target occurrence (occur)”. “Tc” refers to the Tanimoto coefficient between the query compound and the most similar compound interacting with each target within that space. We ranked the 10 targets by Tc and afterwards by scores of Laplacian-modified naive Bayesian target models (Nidhi, et al., 2006) for protein targets with the same Tc (when the the most similar compound interacts with more than one target). “Occur” parameter refers to the number of compounds that interact with each target in the restricted space. Finally, we scored the interaction between the query compound and each candidate target by assigning the precision values corresponding to the “Tc”, “Target rank” and “occur” parameters (Supplementary Table 1).

**Figure 1**. A) Illustration of HitPick V2 ligand-based approach. C: Compound; T: Target B) 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.

## 3. Performance

We evaluated HitPickV2 precision in relation to three parameters of the restricted chemical space in cross-validation of the internal dataset (divided in training and validation set in 85%:15% ratio; the validation set comprises 110,603 compound-target pairs). We observed higher precision for increasing values of “Tc” and “occur” and decreasing “Target rank” values (Supplementary Table 1). Notably, the incorporation of the “occur” parameter enables HitPickV2 to predict targets with high precision (>50%) for some compounds with very low chemical similarity to ligands with known targets (0.3>Tc>0.4). We used the resulting performance table (Supplementary Table 1) relating precision values to the three parameters of the restricted chemical space to score the HitPickV2 compound-target predictions.

We then tested HitPickV2 performance in an external data of 359 compound-target interaction pairs comprising 55 compounds (Klaeger, et al., 2017) not present in our internal dataset. For that, we calculated the precision within two “occur” intervals (occur<=10, loccur>10) for the 1st Target ranked across all ranges of Tc in cross-validation. Similar to the observed performance of HitPickV2 in our internal database, we obtained higher precision for increasing values of Target occurrence (occur) (Figure 1 B), reinforcing the influence of the novel parameter “occur” on HitPickV2 performance (Figure 1 A).

**4. Implementation**

HitpickV2 is freely accessible for non-commercial users in a webserver ([www.hitpickv2.com](http://mips.helmholtz-muenchen.de/HitPickV2/)). As input, HitPickV2 requires a list of query compounds pasted on the input window or uploaded in a file, represented as SMILES strings. The output is a table with the Query Compound and its 2D structure, Predicted targets as gene symbol and Precision. Additional information fields can be displayed including: Most Similar Compound and its 2D structure; Tc between the query compound and the closest compound in the k-NN chemical space annotated to the predicted target and Target occurrence. The output table can be downloaded as a text or pdf file.

**Funding**

This work was supported by German Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes Research (DZD e.V.). J. J.-N is thankful to PECEM and UNAM exchange program for the organization of the fellowship.

*Conflict of Interest:* none declared.

References

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.

Gfeller, D.*, et al.* SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic acids research* 2014;42(Web Server issue):W32-38.

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.

Johnson, M.A., Maggiora, G.M. and American Chemical Society. Meeting. Concepts and applications of molecular similarity. New York: Wiley; 1990.

Keiser, M.J.*, et al.* Relating protein pharmacology by ligand chemistry. *Nature biotechnology* 2007;25(2):197-206.

Liu, X.*, et al.* HitPick: a web server for hit identification and target prediction of chemical screenings. *Bioinformatics (Oxford, England)* 2013;29(15):1910-1912.

Nidhi*, et al.* Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *Journal of chemical information and modeling* 2006;46(3):1124-1133.

Poroikov, V.*, et al.* PASS: identification of probable targets and mechanisms of toxicity. *SAR and QSAR in environmental research* 2007;18(1-2):101-110.

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.

Schuffenhauer, A.*, et al.* Similarity metrics for ligands reflecting the similarity of the target proteins. *Journal of chemical information and computer sciences* 2003;43(2):391-405.

Wale, N. and Karypis, G. Target fishing for chemical compounds using target-ligand activity data and ranking based methods. *Journal of chemical information and modeling* 2009;49(10):2190-2201.

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.