CPIExtract: A software package to collect and harmonize small molecule and protein interactions

Andrea Piras[1], Shi Chenghao[2], Michael Sebek[2], Gordana Ispirova[3], and Giulia Menichetti[2,3,4,*]

[1]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133, Milan, Italy

[2]Network Science Institute, Northeastern University, 360 Huntington Ave, 02115, MA, USA

[3]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, 181 Longwood Ave, 02115, MA, USA

[4]Harvard Data Science Initiative, Harvard University, 114 Western Avenue, 02134, MA, USA

*

Abstract

Summary: The binding interactions between small molecules and proteins are the basis of cellular functions. Yet, experimental data available regarding compound-protein interaction is not harmonized into a single entity but rather scattered across multiple institutions, each maintaining databases with different formats. Extracting information from these multiple sources remains challenging due to data heterogeneity. Here, we present CPIExtract (Compound-Protein Interaction Extract), a tool to interactively extract experimental binding interaction data from multiple databases, perform filtering, and harmonize the resulting information, thus providing a gain of compound-protein interaction data. When compared to a single source, DrugBank, we show that it can collect more than 10 times the amount of annotations. The end-user can apply custom filtering to the aggregated output data and save it in any generic tabular file suitable for further downstream tasks such as network medicine analyses for drug repurposing and cross-validation of deep learning models.

*Corresponding author e-mail: giulia.menichetti@channing.harvard.edu

Availability: CPIExtract is an open-source Python package under an MIT license. CPIExtract can be downloaded from https://github.com/menicgiulia/CPIExtract and https://pypi.org/project/cpiextract. The package can run on any standard desktop computer or computing cluster.

1 Introduction

Compound-protein interactions (CPI) dictate the basis of many cellular dynamics, such as signaling, enzymatic reactions, and gene regulations. By understanding these interactions, we can develop drugs to target specific disease phenotypes and study how dietary molecules affect human and microbial health. Often CPI information is generated with high-throughput screening, testing small molecules ($\leq$ 1000 daltons) against large panels of tissues, cells, or proteins to determine their bioactivity. The generated data is gathered and curated into multiple chemical data sources such as BindingDB (Gilson et al., 2016), ChEMBL (Mendez et al., 2019), and PubChem (Kim et al., 2023). Each database is managed by different organizations, each implementing its own format, resulting in significant heterogeneity in the types of information stored. For example, databases with a focus on pharmaceuticals such as DrugBank (Knox et al., 2024) and DrugCentral (Avram et al., 2023) mainly focus on reporting therapeutic targets, whereas databases like STITCH (Szklarczyk et al., 2016) and the Comparative Toxicogenomics Database (Davis et al., 2021), despite compiling extensive biological interaction networks, lack in-depth information on the bioactivities of these interactions. The presence of crowd-sourced databases like Drug Target Commons (Tanoli et al., 2018) and Open Target Platform (Ochoa et al., 2023) further exacerbates this diversity as each user follows individual reporting criteria. Consequently, extracting and integrating CPI data from these sources poses a significant challenge to researchers. The importance of interaction data has driven efforts to unify chemical structure notation (Heller et al., 2015, Pascazio et al., 2023), protein names (The UniProt Consortium, 2017, Seal et al., 2023), and protein-protein interaction data (Szklarczyk et al., 2019, Dimitrakopoulos et al., 2021). Yet, there is currently no solution for consolidating CPI data despite being essential to many applications, including the training of AI pipelines for drug-target discovery (Patten et al., 2022, Chatterjee et al., 2023), developing chemical language models for de-novo drug design (Grisoni, 2023), and serving as seed interactions in network medicine analyses (Nasirian and Menichetti, 2023, Sebek and Menichetti, 2024). The sparsity of CPI data leads researchers to develop predictive tools for annotating understudied compounds and proteins (Tsubaki et al., 2019, Abramson et al., 2024). However, the performance and generalizability of these models are hindered by the extensive effort needed to integrate different data sources. To address these challenges, we present CPIExtract, a Python package enabling users to seamlessly extract experimental inter-

action data from nine compound-protein interaction databases. By automating the filtration, integration, and standardization processes, CPIExtract generates a unified output that can be stored in any standard tabular file format.

2 CPIExtract Pipeline Workflow

The package allows the user to execute two pipelines, Compound-Proteins-Extraction (Comp2Prot) and the Protein-Compounds-Extraction (Prot2Comp). The Comp2Prot pipeline retrieves proteins interacting with the small molecule passed as input, while Prot2Comp returns compounds interacting with the input protein. Both pipelines comprise three phases: input data extraction, data filtering, and harmonization (Fig. 1A). The package is designed to support multiple user scenarios based on different storage and software availability (SI User scenarios).

2.1 Input data extraction

CPIExtract extracts raw data from the following nine databases: BindingDB (BDB), ChEMBL, Comparative Toxicogenomics Database (CTD), DrugBank (DB), DrugCentral (DC), Drug Target Commons (DTC), Open Targets Platform (OTP), PubChem, and STITCH. Comp2Prot accepts InChI, InChIKey, SMILES, and PubChem ID (CID) as input, while Prot2Comp accepts Entrezgene ID, HGNC ID, Ensembl Peptide ID, Ensembl Gene ID, UniProt ID, ChEMBL ID, and HGNC symbol as input. We use the APIs `pubchempy` for compounds and `Biomart` for proteins to retrieve the identifiers used in each database, unifying the CPI data between databases and removing duplicates.

2.2 Data filtering

CPIExtract focuses exclusively on *Homo sapiens* proteins and protein interactions. Thus, the package removes non-human data and eliminates non-protein interactions, such as RNA interactions. Afterwards, each database is curated depending on the available data: 1) Activity Description: remove interactions with statements indicating inconclusive or undetermined activity (ChEMBL, CTD, DTC, PubChem); 2) Bioactivity Measurements: remove interactions not reporting binding activity (BDB, ChEMBL, DC, DTC, PubChem); 3) Reporting Source: filter to ensure interactions have at least one literature source (ChEMBL, DTC, OTP), or at least one reported experiment (STITCH); and 4) Experimental Conditions: filter interactions by conditions within the human body such as temperature (< 40°C) pH ($5 < pH < 9$) (BDB) and use descriptions of condition validity (ChEMBL).

Since these data are typically only reported when deviating from expected conditions, null values are assumed to represent standard conditions. The package supports optional filters for protein mutations, viral proteins, and compound stereochemistry (SI Optional Filtering Parameters).

2.3 Data harmonization

CPIExtract harmonizes the collected bioactivity data by calculating the pChEMBL value, which is defined as -log_{10}(molar IC50, XC50, EC50, AC50, K_i, K_d, or potency) and provides a numerical value for the binding strength (Bento et al., 2014). BDB, ChEMBL, DC, and PubChem report bioactivity using these standardized metrics, while the user-submitted data in DTC are converted to standardized metrics whenever possible. CPIExtract calculates the average pChEMBL value for unique bioactivities, as databases frequently reference the same sources. The results are provided to the user in a tabular format, which includes the compound and protein identifiers, interaction strength via pChEMBL, and the databases reporting each interaction.

3 Applications

To prove the effectiveness of CPIExtract, we compare it to DrugBank (DB), one of the most important drug-target databases, widely used as training data for machine learning models (Tsubaki et al., 2019) and as a data source for network medicine applications (Sadegh et al., 2021). When focusing on small molecule drugs, DB has 11,340 compounds; however, only 5,420 of these compounds have reported interactions (Fig. 1B). Our package finds interaction data for 8,028 compounds, an additional 2,608 compounds more than DB (23% more coverage). In addition, CPIExtract collects 219,719 interactions compared to the 21,380 interactions in DB, a 10-fold increase in interactions. Lastly, the interactions found by CPIExtract link the compounds to a total of 9,581 proteins, improving the coverage from the 3,044 proteins in DB. The significant increase in CPI data with CPIExtract reveals a fundamental bias in research databases, which often focus on compounds and proteins with known health implications. CPIExtract addresses this bias by integrating diverse datasets, aiding in the mapping of crucial off-target effects.

Next, we consider the strength of the interactions within the DB and CPIExtract data. We use pChEMBL as the evaluation metric, which we divide into three categories: non-binding (pChEMBL≤ 3), weakly binding ($3 <$pChEMBL< 6), and strongly binding (pChEMBL≥ 6) (Chatterjee et al., 2023). Our analysis reveals that 70.9 % of interactions in DB lack bioactivity data. Additionally, DB comprises 0.5 % non-binding, 10.8 % weakly binding, and 17.8 % strongly binding interactions. With CPIExtract, we observed a remarkable 21-fold increase

in weakly binding interactions and a 5.1-fold increase in strong interactions. Overall, CPIExtract retrieved 5 to 10 times more interactions for each pChEMBL bin value compared to DB, facilitating contrastive learning strategies (Singh et al., 2023) and providing valuable examples of experimentally validated negatives to improve the training of binary binding predictors (Chatterjee et al., 2023). Users can filter the retrieved interactions based on the desired pChEMBL value or the number of sources validating the interaction.

Lastly, we define the degree k_i as the number of unique interactions for each compound comp$_i$ and examine its probability distribution in DB and CPIExtract. Our analysis reveals that CPIExtract distributions exhibit larger tails than DB (Fig. 1D), even when applying stringent thresholds of pChEMBL> 8. The information gain is evident when measuring the ratio $k_i(\text{CPIExtract})/k_i(\text{DB})$ at varying pChEMBL thresholds (Fig. 1E). Notably, there is remarkable variability across drugs, as illustrated by the annotation gain for compounds like Chlorpromazine and Ponatinib, which consistently surpass the overall median gain of 2 to 3 times more annotations up to pChEMBL$= 6$.

We conducted a similar analysis on the DB protein list using the Prot2Comp pipeline (see SI Prot2Comp and DB). This analysis uncovered a significant increase in information for protein targets, with CPIExtract providing substantially more annotations. This enhancement facilitates a better understanding of the number of molecules competing for similar targets.

4 Discussion

CPIExtract offers an extensive and integrated resource for protein-ligand binding interactions, addressing challenges across various domains by merging heterogeneous data and mitigating bias and annotation imbalances from single sources. Specifically, CPIExtract facilitates the stratification of binding information by chemical classes, enabling the creation of balanced datasets that can enhance model training performance. In the field of network medicine, CPIExtract's comprehensive interaction lists enhance the prediction of health implications. In food science, the tool addresses the issue of limited bioactivity knowledge across databases, providing a more robust framework for understanding compound interactions. Overall, CPIExtract empowers researchers to thoroughly characterize their compounds or proteins of interest, systematically exploring mechanisms of action with robust experimental evidence.

5 Acknowledgments

We thank Bnaya Gross for testing the package during development and Daria Koshkina for her help in designing the figure.

Author Contributions A.P. and M.S. developed the methodology. A.P., S.C., M.S., and G.I. built the software. A.P., M.S., and G.M. contributed to the writing and editing. G.M. conceptualized, administered, and funded the project.

Funding G.M. is supported by NIH/NHLBI K25HL173665 and AHA 24MERIT1185447.

Competing interests Authors declare no competing interests.

References

J. Abramson, J. Adler, J. Dunger, et al. Accurate structure prediction of biomolecular interactions with alphafold3. *Nature*, 5 2024.

S. Avram, T. B. Wilson, R. Curpan, et al. Drugcentral 2023 extends human clinical data and integrates veterinary drugs. *Nucleic Acids Res*, 51(D1):D1276–D1287, 2023.

A. P. Bento, A. Gaulton, A. Hersey, et al. The chembl bioactivity database: an update. *Nucleic Acids Res*, 42(D1): D1083–D1090, 2014.

A. Chatterjee, R. Walters, Z. Shafi, et al. Improving the generalizability of protein-ligand binding predictions with ai-bind. *Nat Commun.*, 14(1):1989, 2023.

A. P. Davis, C. J. Grondin, R. J. Johnson, et al. Comparative toxicogenomics database (ctd): update 2021. *Nucleic Acids Res*, 49(D1):D1138–D1143, 2021.

G. N. Dimitrakopoulos, M. I. Klapa, and N. K. Moschonas. PICKLE 3.0: enriching the human meta-database with the mouse protein interactome extended via mouse–human orthology. *Bioinformatics*, 37(1):145–146, 2021.

M. K. Gilson, T. Liu, M. Baitaluk, et al. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res*, 44(D1):D1045–D1053, 2016.

F. Grisoni. Chemical language models for de novo drug design: Challenges and opportunities. *Curr Opin Struct Biol.*, 79:102527, 2023.

S. R. Heller, A. McNaught, I. Pletnev, et al. InChI, the IUPAC International Chemical Identifier. *J Cheminform.*, 7 (1):23, 2015.

S. Kim, J. Chen, T. Cheng, et al. Pubchem 2023 update. *Nucleic Acids Res*, 51(D1):D1373–D1380, 2023.

C. Knox, M. Wilson, C. M. Klinger, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic Acids Res*, 52(D1):D1265–D1275, 2024.

D. Mendez, A. Gaulton, A. P. Bento, et al. Chembl: towards direct deposition of bioassay data. *Nucleic Acids Res*, 47(D1):D930–D940, 2019.

F. Nasirian and G. Menichetti. Molecular Interaction Networks and Cardiovascular Disease Risk: The Role of Food Bioactive Small Molecules. *Arterioscler Thromb Vasc Biol.*, 43(6):813–823, 2023.

D. Ochoa, A. Hercules, M. Carmona, et al. The next-generation open targets platform: reimagined, redesigned, rebuilt. *Nucleic Acids Res*, 51(D1):D1353–D1359, 2023.

L. Pascazio, S. Rihm, A. Naseri, et al. Chemical Species Ontology for Data Integration and Knowledge Discovery. *J Chem Inf Model.*, 63(21):6569–6586, 2023.

J. Patten, P. T. Keiser, D. Morselli-Gysi, et al. Identification of potent inhibitors of sars-cov-2 infection by combined pharmacological evaluation and cellular network prioritization. *Iscience*, 25(9), 2022.

S. Sadegh, J. Skelton, E. Anastasi, J. Bernett, D. B. Blumenthal, G. Galindez, M. Salgado-Albarrán, O. Lazareva, K. Flanagan, S. Cockell, et al. Network medicine for disease module identification and drug repurposing with the nedrex platform. *Nature Communications*, 12(1):6848, 2021.

R. L. Seal, B. Braschi, K. Gray, et al. Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res*, 51(D1): D1003–D1009, 2023.

M. Sebek and G. Menichetti. Chapter 20 - Network Science and Machine Learning for Precision Nutrition. In D. Heber, Z. Li, and J. Ordovas, editors, *Precision Nutrition*, pages 367–402. Academic Press, 2024.

R. Singh, S. Sledzieski, B. Bryson, L. Cowen, and B. Berger. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proceedings of the National Academy of Sciences of the United States of America*, 120:e2220778120, 6 2023. ISSN 10916490. doi: 10.1073/PNAS.2220778120/SUPPL_FILE/PNAS.2220778120.SAPP.PDF. URL https://www.pnas.org/doi/abs/10.1073/pnas.2220778120.

D. Szklarczyk, A. Santos, C. Von Mering, et al. Stitch 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Res*, 44(D1):D380–D384, 2016.

D. Szklarczyk, A. L. Gable, D. Lyon, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*, 47(D1): D607–D613, 2019.

Z. Tanoli, Z. Alam, M. Vähä-Koskela, et al. Drug target commons 2.0: a community platform for systematic analysis of drug–target interaction profiles. *Database*, 2018:bay083, 2018.

The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*, 45(D1):D158–D169, 2017.

M. Tsubaki, K. Tomii, and J. Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2):309–318, 2019.

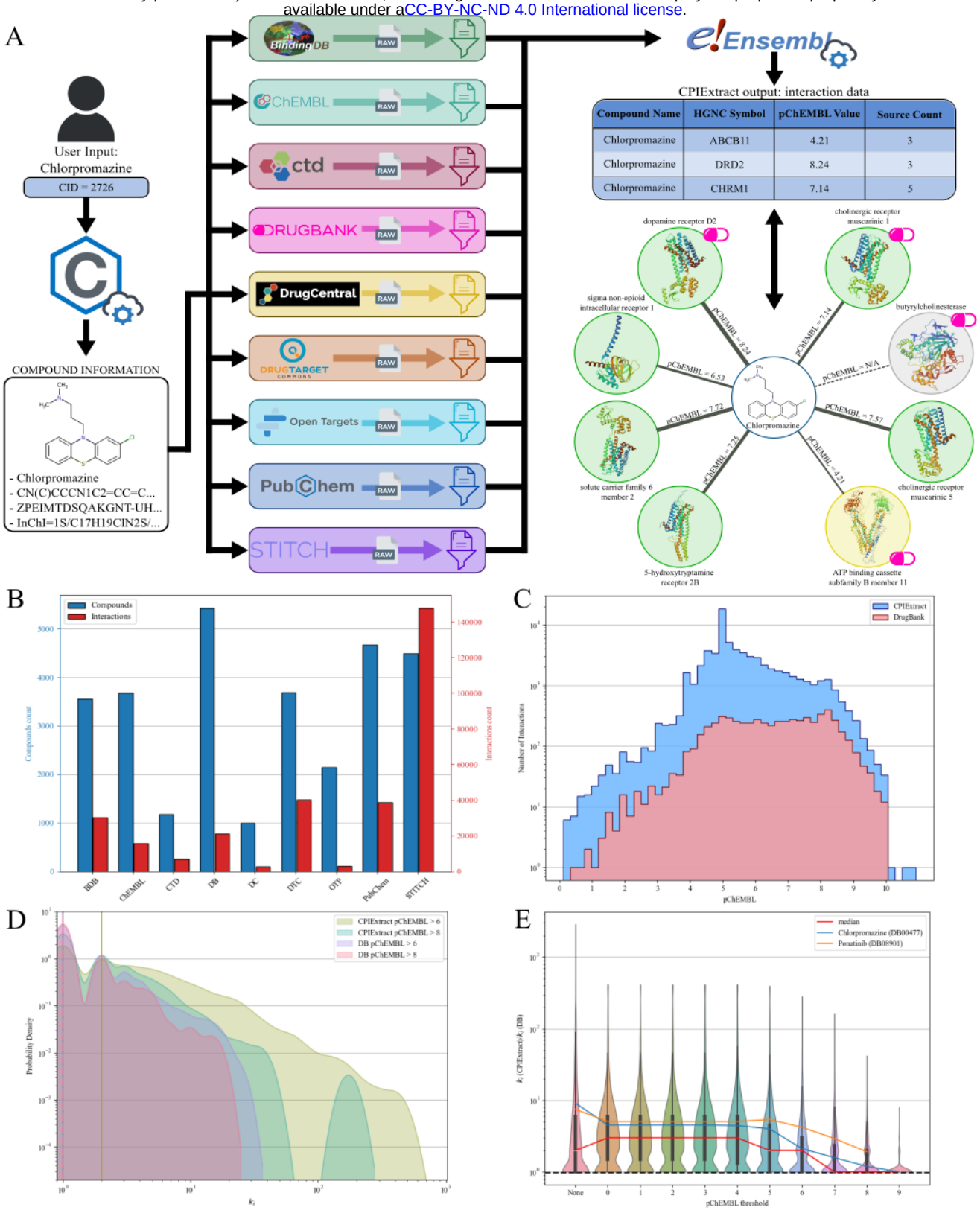

Figure 1: **CPIExtract pipeline and DrugBank. A)** Schematic of the Comp2Prot pipeline to extract protein interactions for a given compound. The user provides the CID for Chlorpromazine. CPIExtract extracts compound identifiers from PubChem and then collects and filters the raw interaction data for each database. The protein information is standardized by Biomart and stored in a tabular output. The interaction network shows binding proteins colored by interaction strength (green: binding, yellow: weak binding, gray: no data). The DB logo marks annotations in DB. **B)** CPI data found in each database for the small molecule drugs from DB (blue: total number of compounds, red: total number of interactions). **C)** The pChEMBL distribution for CPIExtract and DB annotations. **D)** Density distribution of k_i for CPIExtract and DB data at different pChEMBL thresholds. CPIExtract has more interactions per compound than DB, even with stronger thresholds. **E)** k_i ratio distribution between CPIExtract and DB. $k_i(\text{CPIExtract})/k_i(\text{DB}) = 1$ dotted line highlights an equal number of interactions reported.