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

Andrea Piras*[1], Michael Sebek*[2], Shi Chenghao[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

*These authors contributed equally.

[†]Corresponding author e-mail: giulia.menichetti@channing.harvard.edu

## Abstract

**Summary:** The binding interactions between small molecules and proteins are the basis of cellular functions. Yet, experimental data available regarding compound-protein interactions (CPIs) are 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 Python package that automatically retrieves CPI data from nine major repositories, filters non-human and low-quality records, harmonizes chemical and protein identifiers, and computes unified pChEMBL binding values. Compared with MINER, a state-of-the-art CPI extraction algorithm, CPIExtract retrieves 85.5% more compounds, 16-fold more experimentally supported interactions, and over four times more proteins, substantially increasing the availability of strong and weak binders. The resulting harmonized dataset enables custom filtering and export in standard tabular formats for downstream applications such as network medicine, drug repurposing, and training of deep learning models.

**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 (Menichetti et al., 2025). Often CPI information is generated with high-throughput screening, testing small molecules (≤ 1000 Daltons) against large panels of tissues, cells, or proteins to determine their bioactivity. The generated data are 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, no open-source framework currently exists for harmonizing CPI data across repositories. This gap in standardization is increasingly problematic, as CPI datasets form the backbone of modern computational approaches, including AI pipelines for drug–target prediction (Chatterjee et al., 2023), chemical language models for de novo molecular design (Grisoni, 2023), and network-medicine analyses that rely on experimentally supported seed interactions (Nasirian and Menichetti, 2023, Sebek and Menichetti, 2024, Aldana et al., 2025). 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). Yet, these models frequently depend on databases built using heuristics and opaque algorithms that were never rigorously validated, raising concerns about their reliability. To address these challenges, we present CPIExtract, a Python package that enables users to seamlessly extract experimental interaction 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. This framework not only streamlines CPI data retrieval but also emphasizes the importance of a tool that enables transparency and standardization in biochemical research, ensuring more reliable foundations for downstream analyses.

# 2 CPIExtract Pipeline Workflow

The package allows the user to execute two pipelines, Compound-Proteins Extraction (Comp2Prot) and Protein-Compounds Extraction (Prot2Comp). The Comp2Prot pipeline retrieves proteins that interact 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 assess the effectiveness of CPIExtract, we benchmarked it against MINER (Stanford SNAP Group, 2017), the CPI extraction module within the Stanford Network Analysis Platform (SNAP), a widely used graph-mining and network-analysis library (Leskovec and Sosič, 2016). We use the 11,340 small molecules within DrugBank as our comparison dataset between the pipelines. For this dataset, MINER collected 4,433 compounds, 13,625 binding interactions, and 2,096 proteins. In contrast, CPIExtract found 8,221 compounds, 219,719 binding interactions, and 9,581 proteins, representing an 85.5% increase in compounds and a 16-fold increase in interactions over MINER (Fig. 1B). This substantial improvement exposes a bias in existing extraction tools, which rely heavily on a few sources. By integrating diverse repositories, CPIExtract mitigates this bias and supports more comprehensive mapping of off-target effects and understudied interactions.

Since MINER provides interaction data without quantitative bioactivity measurements, we used CPIExtract to annotate MINER's retrieved CPIs with pChEMBL values, enabling the evaluation of interaction strength. Through this integration, we categorized the interactions as non-binding (pChEMBL $\leq 3$), weakly binding ($3 <$ pChEMBL $< 6$), or strongly binding (pChEMBL $\geq 6$) [Chatterjee et al., 2023]. In MINER, 61.6% of interactions lack any strength quantification, with only 0.4% classified as non-binding, 12.3% as weakly binding, and 25.7% as strong binders. CPIExtract dramatically expands these categories, yielding a 32.5-fold increase in weakly binding interactions and a 6.1-fold increase in strong binders (Fig. 1C).

Overall, CPIExtract retrieves 10- to 15-fold more interactions within each pChEMBL bin than MINER, enabling improved contrastive learning workflows (Singh et al., 2023) and providing experimentally validated negatives not obtainable in MINER for training binary binding predictors (Chatterjee et al., 2023). Users may optionally filter interactions by pChEMBL threshold or by the number of independent sources supporting each annotation.

Lastly, we examined the degree $k_i$, defined as the number of unique interactions associated with each compound $comp_i$, and compared its distribution in MINER and CPIExtract. CPIExtract consistently produces degree distributions with substantially heavier tails than MINER (Fig.1D), even under stringent thresholds such as pChEMBL > 8.

This information gain is further quantified by the ratio $k_{i(\mathrm{CPIExtract})}/k_{i(\mathrm{MINER})}$ across pChEMBL cutoffs (Fig. 1E). Notably, the improvement is highly compound-specific: drugs such as chlorpromazine and ponatinib exhibit annotation gains that exceed the median two- to three-fold increase observed up to pChEMBL = 6. We performed analogous analyses to compare the number and type of compound–interaction annotations retrieved per protein (see SI Prot2Comp and MINER) and the number and type of protein targets retrieved per naturally occurring compound (see SI Comp2Prot on Naturally Occurring Compounds). Both analyses revealed a substantial increase in information, with CPIExtract consistently providing far more annotations than MINER. This expanded coverage enables a more complete characterization of competitive binding landscapes and clarifies how multiple molecules may interact with the same protein targets.

## 4 Discussion

CPIExtract provides an extensive, harmonized resource for protein–ligand interactions, addressing long-standing challenges arising from heterogeneous data formats and annotation biases across individual databases. In contrast to tools implemented with limited methodological transparency, CPIExtract offers a clear and systematic workflow that prioritizes experimentally supported evidence rather than inferred associations. By integrating data from nine major repositories, the tool mitigates annotation imbalance and enables users to access a more representative and comprehensive view of compound–protein interactions. 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

*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.

A. Aldana, M. Sebek, G. Ispirova, R. Dorantes-Gilardi, J. Loscalzo, A. L. Barabási, and G. Menichetti. Netmedpy: a python package for large-scale network medicine screening. *Bioinformatics*, 41, 9 2025. ISSN 13674811. doi: 10.1093/BIOINFORMATICS/BTAF338.

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.

J. Leskovec and R. Sosič. Snap: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(1):1, 2016.

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

G. Menichetti, A.-L. Barabási, and J. Loscalzo. Chemical complexity of food and implications for therapeutics. *New England Journal of Medicine*, 392:1836–1845, 5 2025. ISSN 0028-4793. doi: 10.1056/NEJMRA2413243.

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.

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, et al. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proc Natl Acad Sci. U.S.A.*, 120:e2220778120, 2023.

Stanford SNAP Group. Miner: Gigascale multimodal biological network. *SNAPGitHub Repository*, 2017.

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.

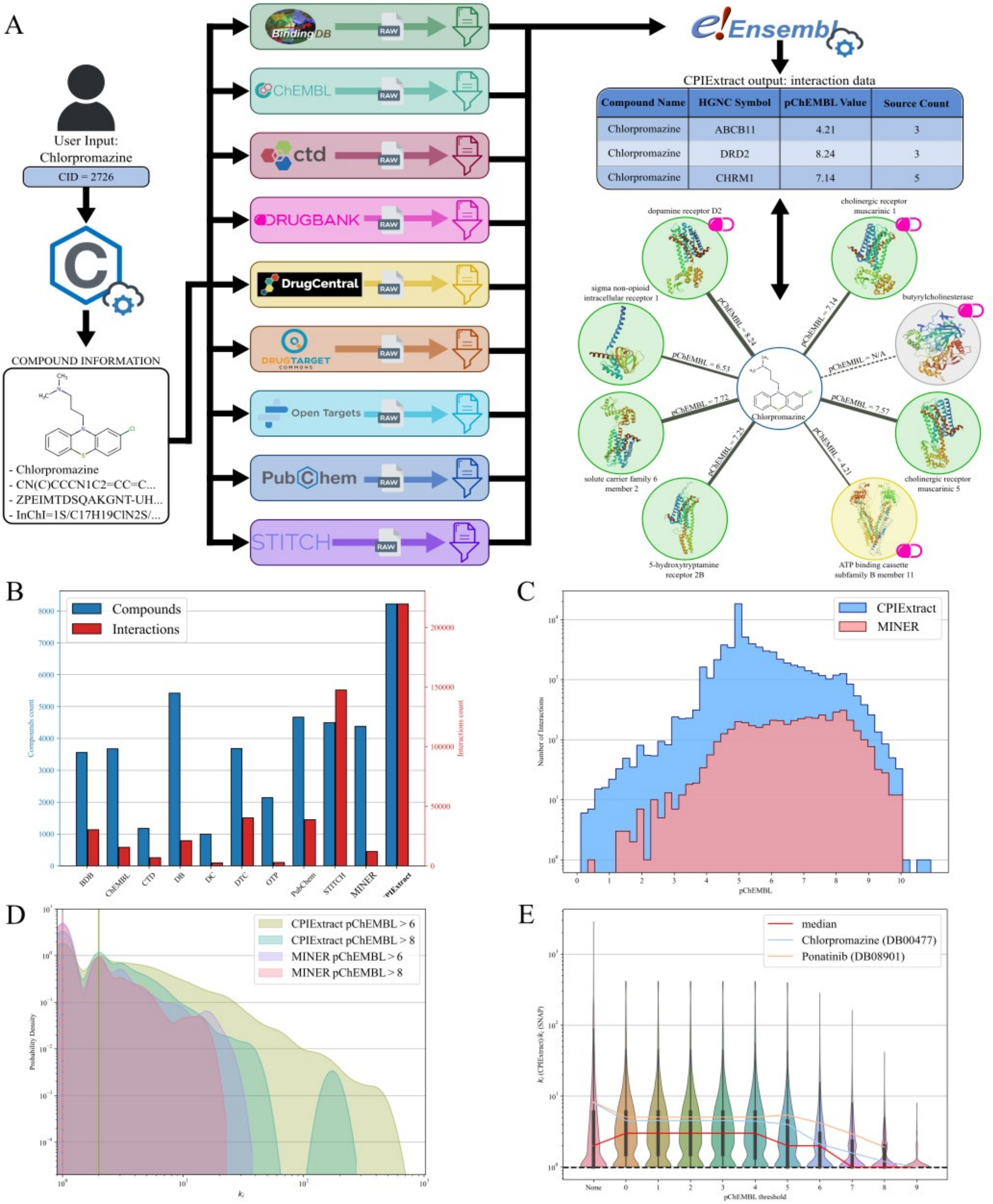

Figure 1: **CPIExtract pipeline and MINER. A)** Schematic of the Comp2Prot pipeline to extract protein interactions for a given compound. The user provides the CID for Chlorpromazine. CPIExtract extracts compound identifiers from PubChem and then collects and filters the raw interaction data for each 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 DrugBank logo marks annotations in DrugBank. **B)** CPI data found in each database for the small molecule drugs from MINER (blue: total number of compounds, red: total number of interactions). **C)** The pChEMBL distribution for CPIExtract and MINER annotations. **D)** Density distribution of $k_i$ for CPIExtract and MINER data at different pChEMBL thresholds. CPIExtract has more interactions per compound than MINER, even with stronger thresholds. **E)** $k_i$ ratio distribution between CPIExtract and MINER. $k_i(CPIExtract)/ k_i(MINER) = 1$ dotted line highlights an equal number of interactions reported.