1 MATERIALS

1.1 Dataset

The dataset for the experimental evaluation is very critical, and therefore we present the dataset construction in detail here. We constructed a set of bioassays from ChEMBL\footnote{https://www.ebi.ac.uk/chembl/, v.22_1, accessed on 12/08/2016} in accordance with the protocols in Section 1.1.1 and Section 1.1.2 in order to 1) have a sufficiently large number of bioassays to study; and 2) have a sufficiently large number of active and selective compounds in each bioassay to reliably learn models.

1.1.1 Initial Bioassay Selection. We first selected a set of bioassays which are enriched with selective compounds, and meanwhile, the compound selectivity in these bioassays can be largely defined by other selected bioassays. This set of bioassays provides a closed space from which a subset of bioassays will be further constructed (Section 1.1.2) for the experiments. We constructed this initial set of bioassays as follows:

1. Identify all “binding” bioassays with one “single protein” target;
2. From such single-target binding bioassays, find all the bioassays that use IC$_{50}$ to measure compound activities, and keep the compounds in such bioassays that have exact IC$_{50}$ values (i.e., discard from each bioassay the compounds with IC$_{50}$ ranges, for example, IC$_{50}$ ≥ 0.0001µM; also discard compounds whose measurement cannot be converted to IC$_{50}$ values);
3. Combine bioassays of a same target into one bioassay;
4. Clean the combined bioassays as follows:
   a. If a compound appears multiple times with a same IC$_{50}$ value in one bioassay, keep the compound with the unique IC$_{50}$ in the bioassay;
   b. If a compound appears multiple times with different IC$_{50}$ values in one bioassay, remove the compound and all its activities from the bioassay. This is to avoid the complication to resolve conflicts of inconsistent activity values;
   c. If a compound has an invalid IC$_{50}$ value (e.g., negative or zero IC$_{50}$), remove the compound from the bioassay.
5. Select the cleaned bioassays that have at least 20 active compounds.

After the above process, we identified 1,033 bioassays in total. Among these 1,033 bioassays, 594 bioassays have selective compounds that are defined within these 1,033 bioassays. Among these 594 bioassays, 553 bioassays have selective compounds that are defined within these 594 bioassays. Among these 553 bioassays, 227 bioassays have more than 10 selective compounds, and these selective compounds are involved in 529 out of the 553 bioassays. This set of 529 bioassays, denoted as $B_{c0s}$, represents the initial closed set of selectivity-enriched bioassays.

1.1.2 Initial Bioassay Pruning. We defined selectivity for the compounds in each bioassay in $B_{c0s}$ with respect to the other bioassays in $B_{c0}$. These 529 $B_{c0s}$ bioassays are further pruned according to the following protocol in order to have reasonable number of compounds for dCPPP learning:

1. If a bioassay has less than 100 compounds, keep the bioassay as it is;
2. If a bioassay has more than 100 compounds, identify all its selective compounds and x-selective compounds:
   a. If such identified compounds are more than 100, keep all such compounds and discard all the other compounds;
   b. If such identified compounds are less than 100, randomly select active compounds in this bioassay until the total number of selected compounds reaches 100.

The above pruning process retains all the selectivity related information in the $B_{c0s}$ bioassays. All the remaining bioassays and their compounds are used as the final dataset in our experiments. This set of pruned bioassays is denoted as $B_{cs}$. 