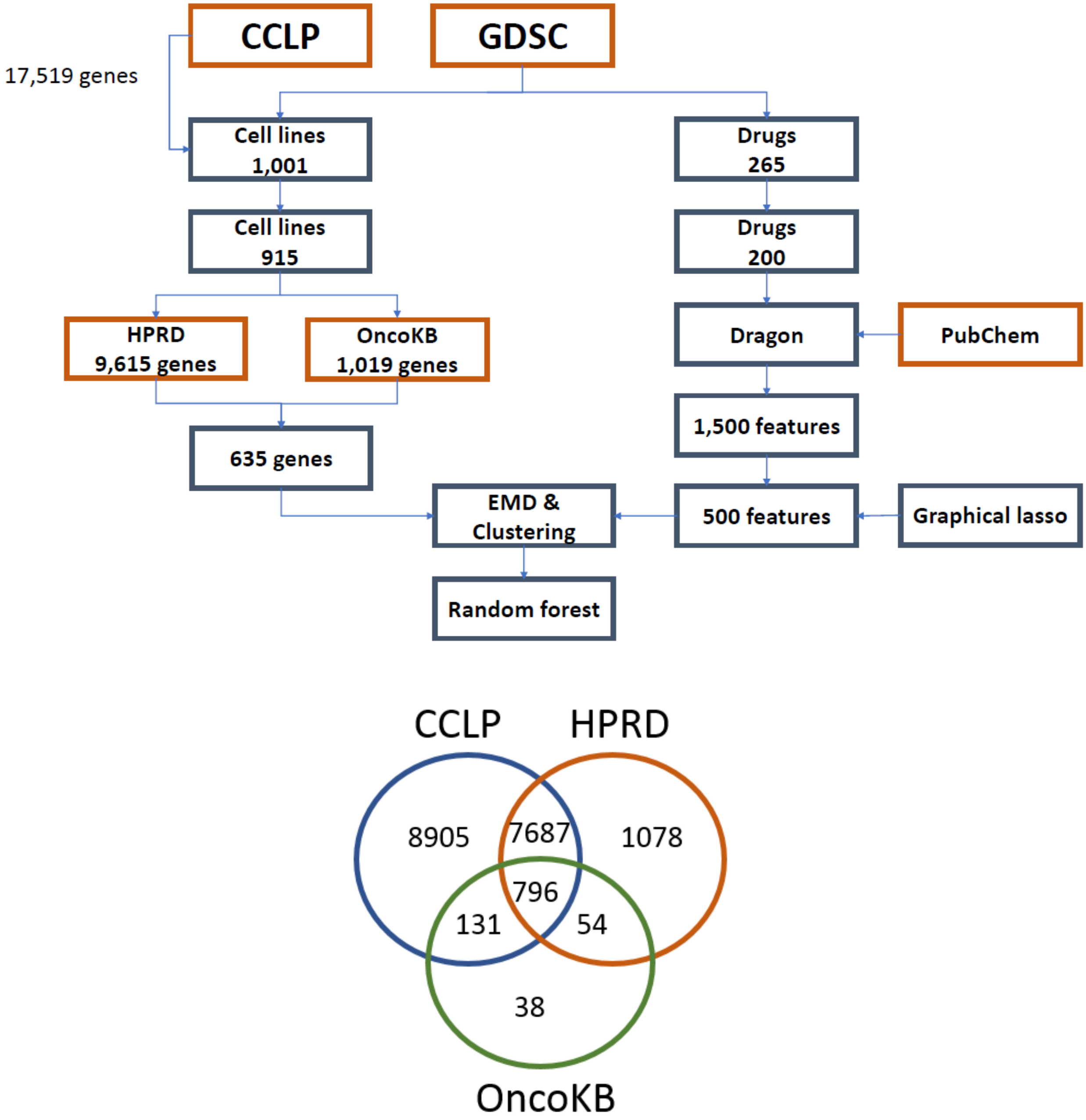

Figure 1. (Top) Overview of data analysis; (Bottom) The number of common genes among the gene sets in CCLP, HPRD, and OncoKB databases.

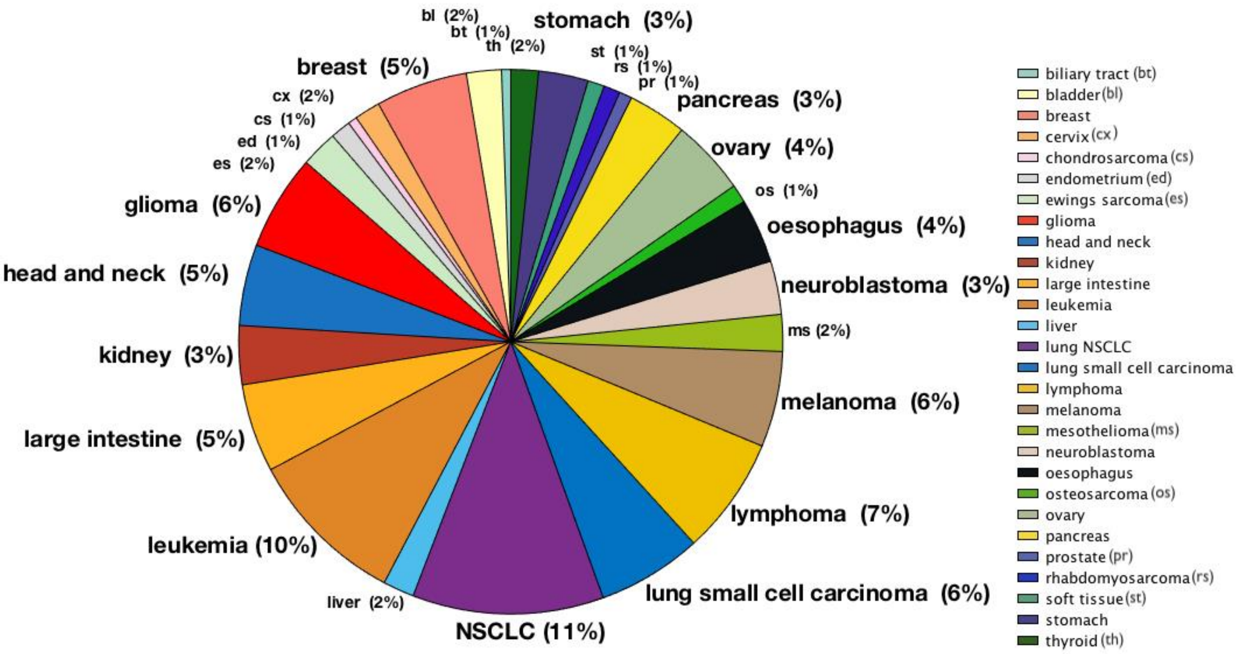

Figure 2. Distribution of the 915 cell lines from the GDSC dataset with respect to their cancer types. The largest cancer types are non-small cell lung cancer (NSCLC) (11%) and leukemia (10%). Lymphoma (7%), glioma (6%), melanoma (6%), and lung small cell carcinoma (6%) are the next major types.

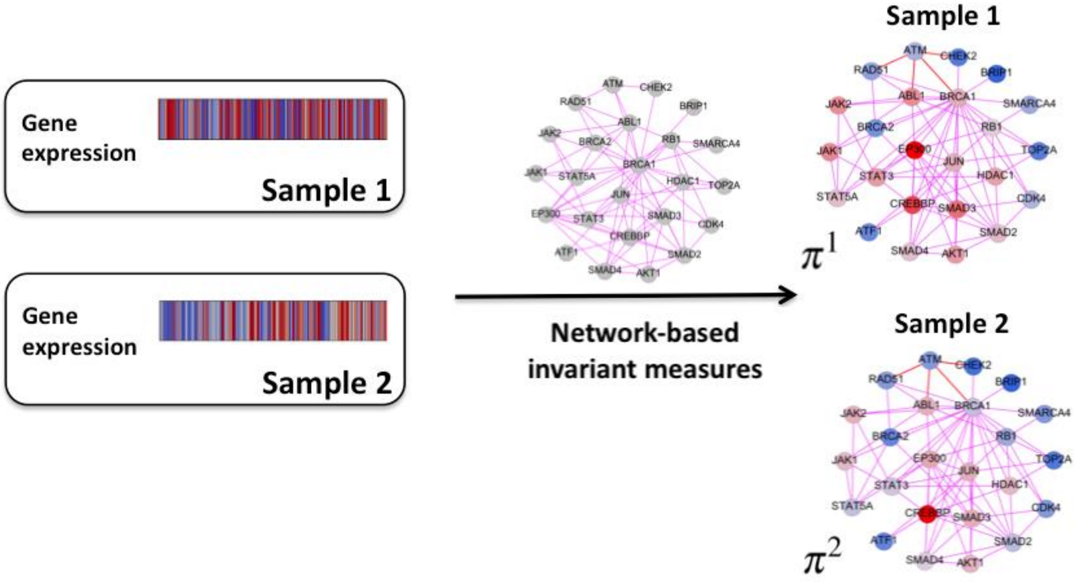

Figure 3. The invariant measures define a weighted network for each sample. The Wasserstein distance (EMD) calculates the most efficient way to move the distribution of invariant measure from one sample to another sample, where the cost is the shortest path in the network. Here, we show a small network for the purpose of illustration.

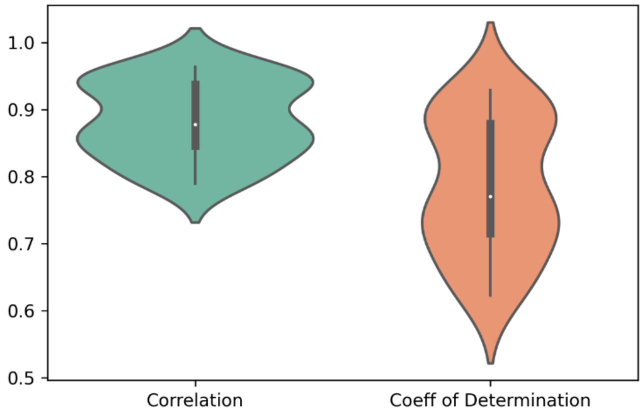

Figure 4. The distribution of correlation (R) and coefficient of determination (R^2) of the predicted and observed log(IC50) values in the 30 paired clusters of cell lines and drugs. The average values of R and R^2 were 0.88 and 0.78, respectively.

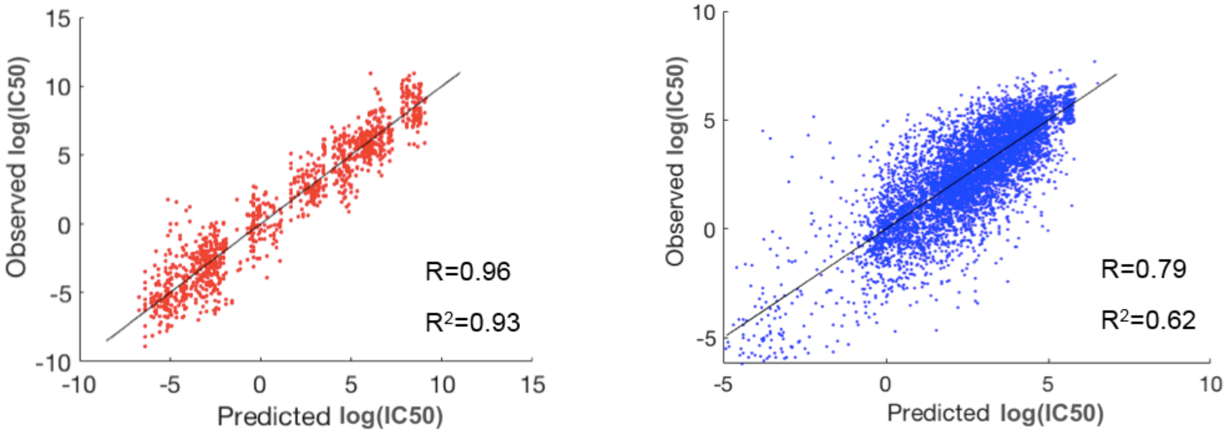

Figure 5. The best (red) and worst (blue) clusters among the 30 paired clusters with respect to the prediction accuracy. The best prediction lies in the pair of cluster 3 in cell lines (mainly glioma and melanoma) and cluster 1 in drugs, and the worst prediction lies in the pair of cluster 6 in cell lines (mainly consisting of breast, head and neck, large intestine, and stomach cancers) and cluster 5 in drugs.