Supporting Information for

Predicted Biological Activity of Purchasable Chemical Space

John J. Irwin+%*, Garrett Gaskins+%^#$, Teague Sterling%, Michael M. Mysinger% and Michael J. Keiser%^#$

% Department of Pharmaceutical Chemistry, University of California San Francisco, Byers Hall, 1700 4th St, San Francisco CA 94158-2330

^ Institute for Neurodegenerative Diseases, University of California San Francisco, 675 Nelson Rising Ln, San Francisco CA 94158

# Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, Byers Hall, 1700 4th St, San Francisco CA 94158

$ Institute for Computational Health Sciences, University of California San Francisco, 550 16th St, San Francisco CA 94158

Keywords: chemical tools, ligand discovery, library design

*Corresponding author: jji@cgl.ucsf.edu 415/937-1461

+ These authors contributed equally to this work.

Figure S1. Performance metrics for SEA+TC on ChEMBL cross-validation sets filtered for >5 ligand annotations per target. All curves are derived from independent 5-fold cross-validation runs. Overall performance is measured by either the AUROC (A, C) or the AUPRC (B, D). **A)** ROC curves for ChEMBL cross-validation sets at a variety of threshold values. Each curve is the result of stepping the decision threshold across MaxTC values, while holding the SEA p-value decision threshold constant. Inset shows a zoomed-in version of ROC curves, with the FPR (x-axis) in log units to emphasize low-FPR behavior. **B)** Corresponding PRCs for cross-validation runs described in (A). Pink and blue circles indicate the recommended upper and lower bounds for MaxTC thresholding, respectively (MaxTC = 0.80; 0.40). **C)** Complementary ROC curves to section (A); each curve is the result of stepping across all SEA p-values, while holding the MaxTC decision threshold constant. **D)** Corresponding PRCs for cross-validation runs described in (C). Pink and blue circles indicate the recommended upper and lower bounds for the SEA p-value decision threshold, respectively (pSEA = 80; 40).

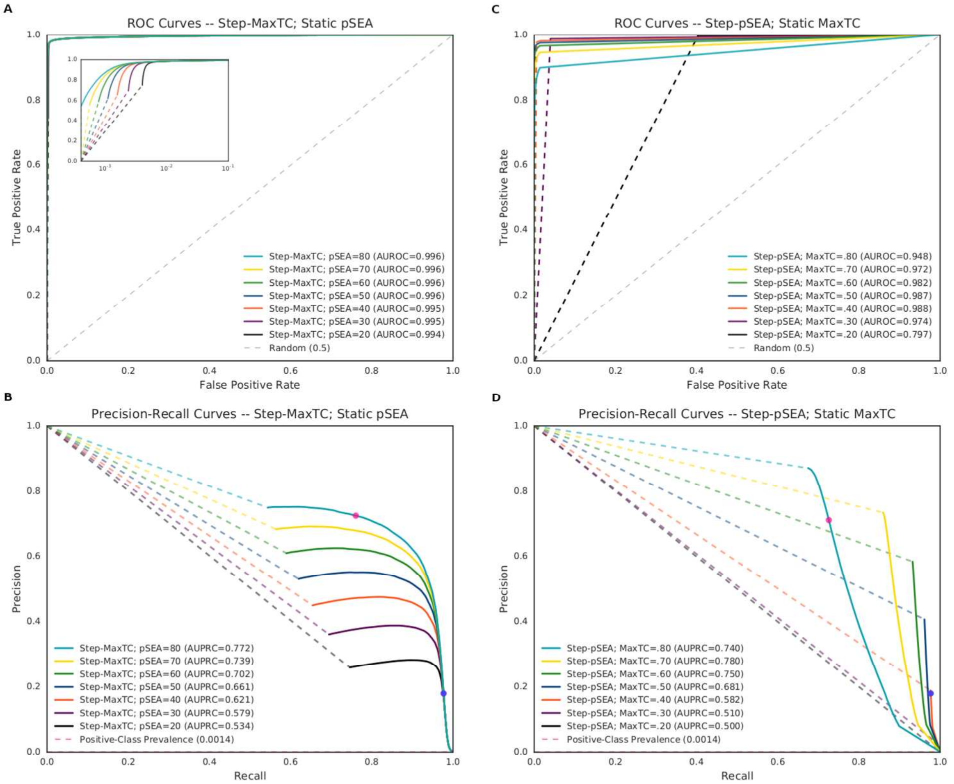


Figure S2. Performance metrics for SEA+TC on ChEMBL cross-validation sets filtered for >50 ligand associations per target. All analyses replicate those undertaken in **Supplementary Figure 1**.

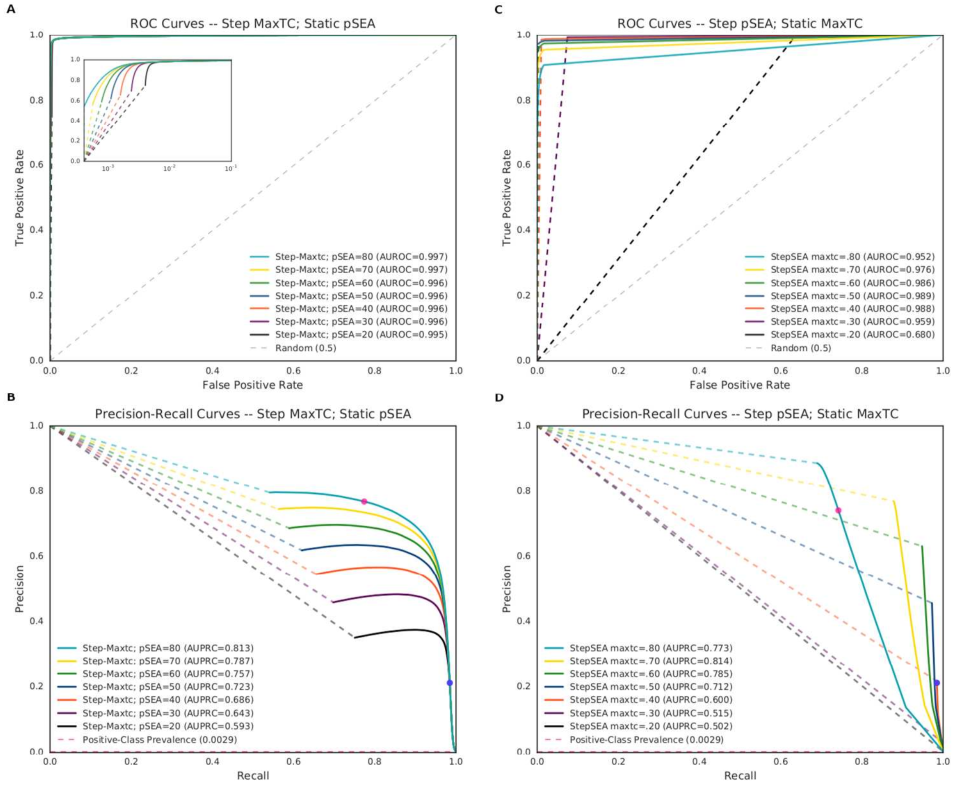
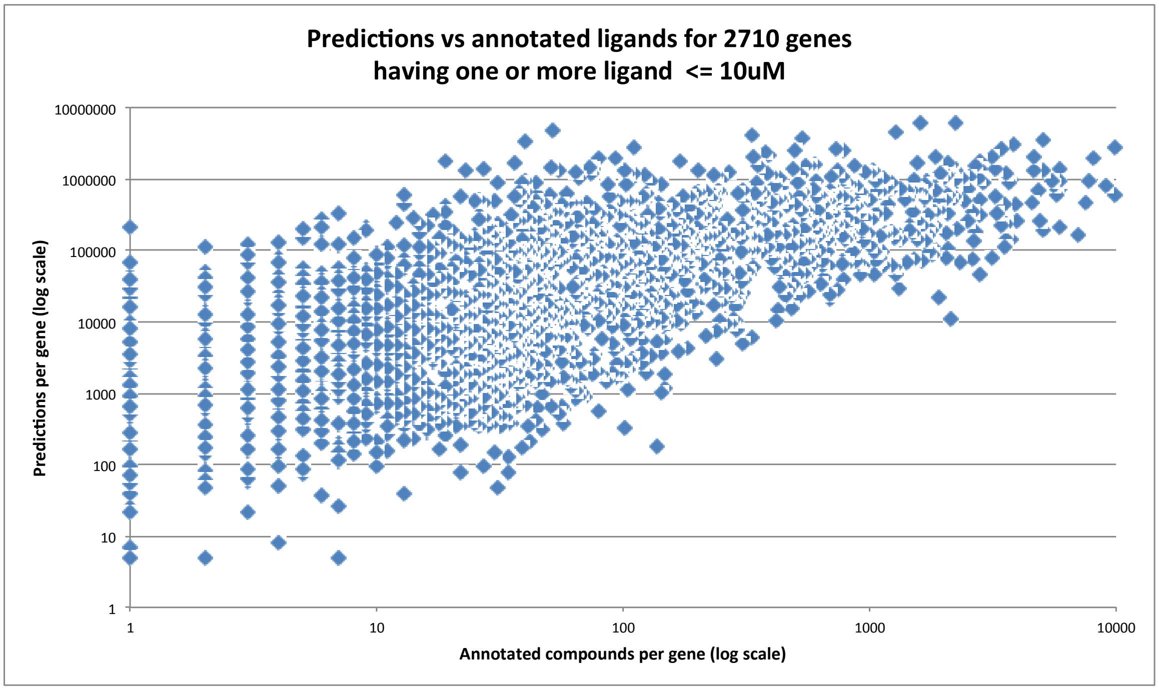


Figure S3


Predictions vs annotated ligands for 2710 genes
having one or more ligand <= 10uM
Predictions per gene (log scale)
10000000
1000000
100000
10000
1000
100
10
1
1
10
100
1000
10000
Annotated compounds per gene (log scale)