Supporting Information

Learning Drug Functions from Chemical Structures with Convolutional Neural Networks and Random Forests

Jesse G. Meyer[1,2,3]*, Shengchao Liu[4,5], Ian J. Miller[3], Joshua J. Coon[1,2,3,5,6], Anthony Gitter[4,5,7]

[1] Department of Chemistry
[2] Department of Biomolecular Chemistry
[3] National Center for Quantitative Biology of Complex Systems
[4] Department of Computer Sciences
[5] Morgridge Institute for Research
[6] DOE Great Lakes Bioenergy Research Center
[7] Department of Biostatistics and Medical Informatics

University of Wisconsin—Madison, Madison, Wisconsin 53706, United States

*Correspondence to:
Jesse G. Meyer
425 Henry Mall, room 4449
Madison, WI 53706
jessegmeyer@gmail.com

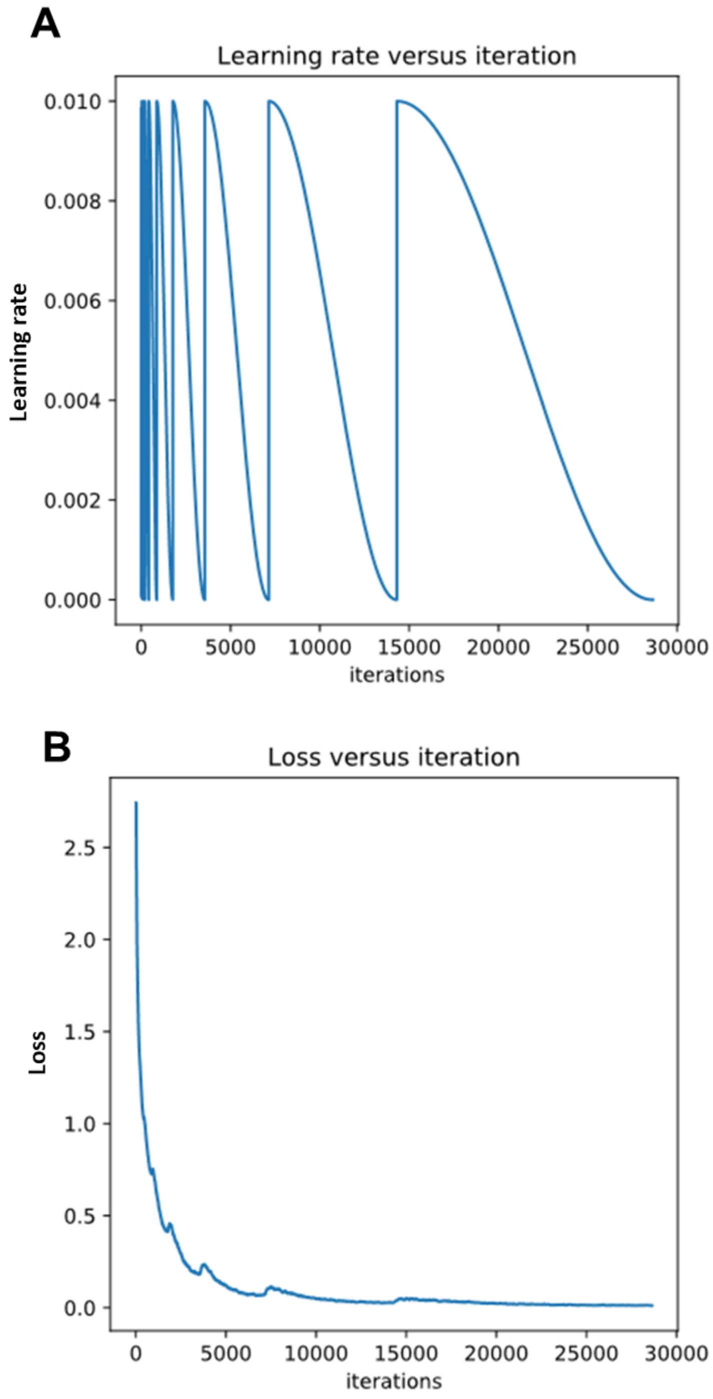

Figure S1: Example of the cosine annealing learning rate strategy used for training the small and large single class models. Seven cycles each with double the time for rate decay were used, resulting in a total of 127 epochs of training. (A) Learning rate versus training batch. (B) Loss versus training batch.

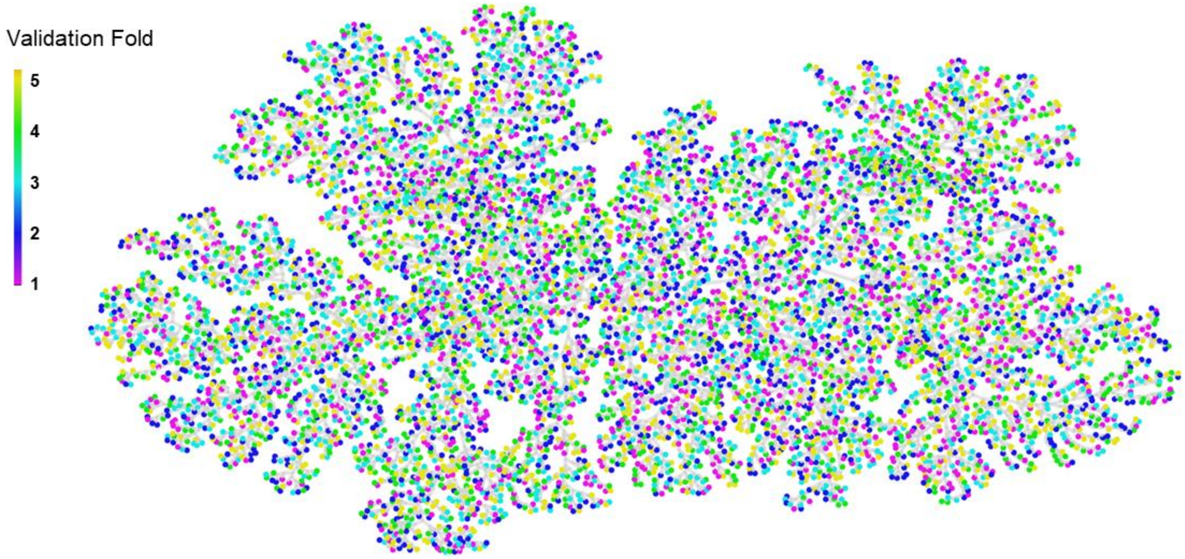

Figure S2: ChemTreeMap showing random chemical similarity of the molecules in each of the 5 validation folds for the large dataset (6,955 molecules).

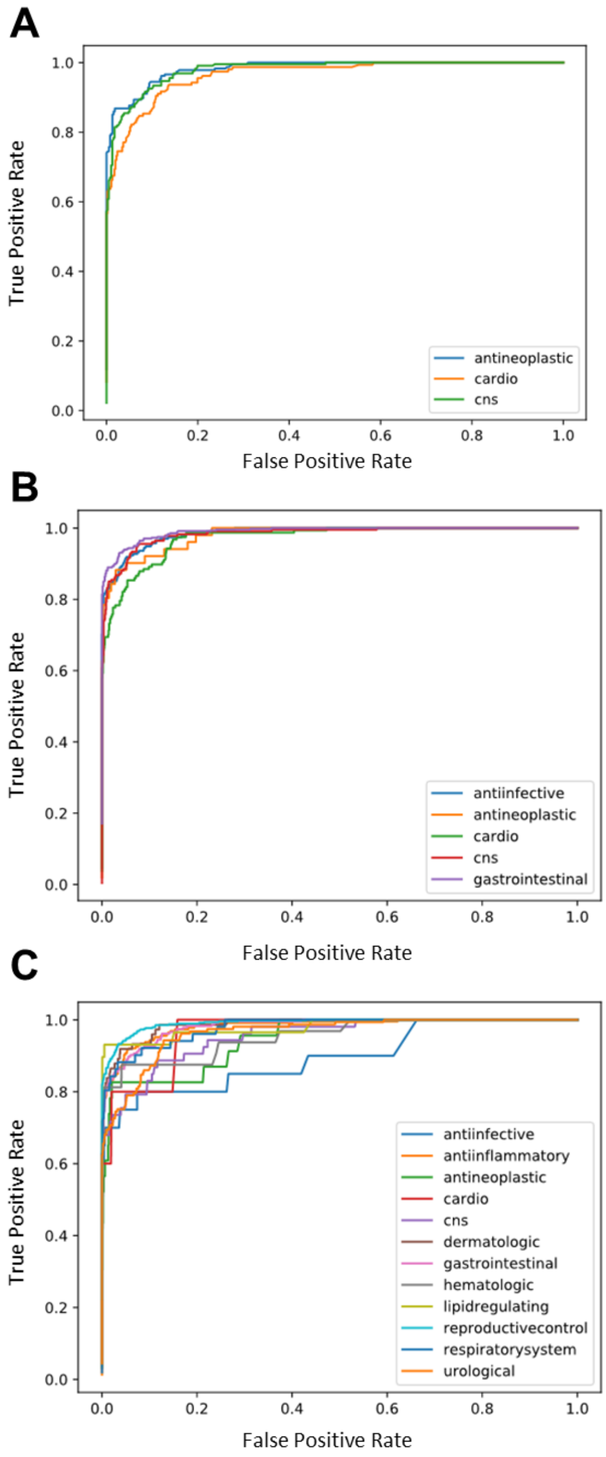

Figure S3: Receiver Operator Characteristic Plots from the Random Forest Models trained with Morgan Molecular Fingerprints on the fifth validation set for the (A) 3, (B) 5, and (C) 12 therapeutic classes.

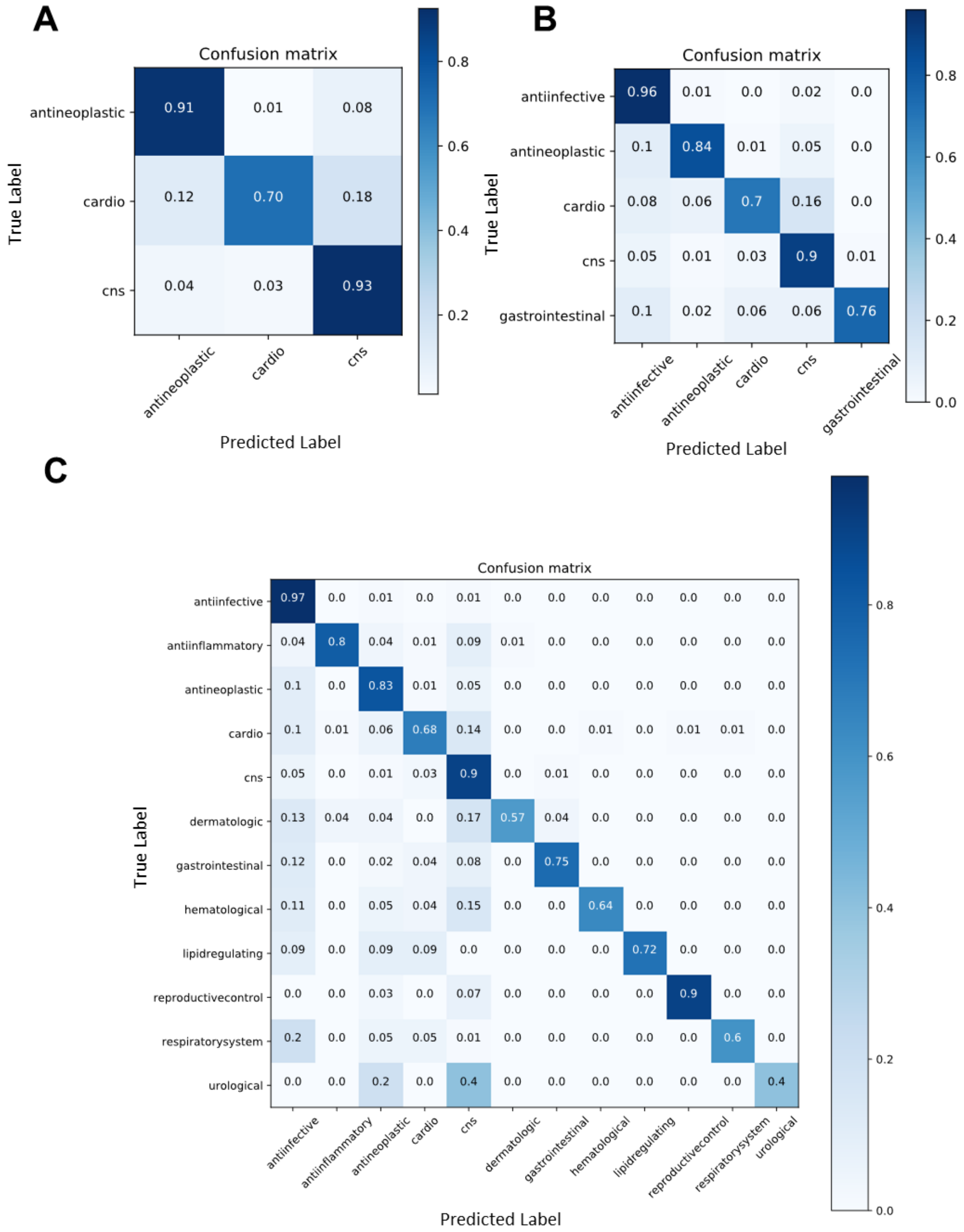

Figure S4: Confusion matrices from MFP+RF classification performance over sets of drug molecules belonging to (A) 3, (B) 5, and (C) 12 therapeutic classes. Each matrix shows the predictions from the fifth validation set using models trained on the large dataset.

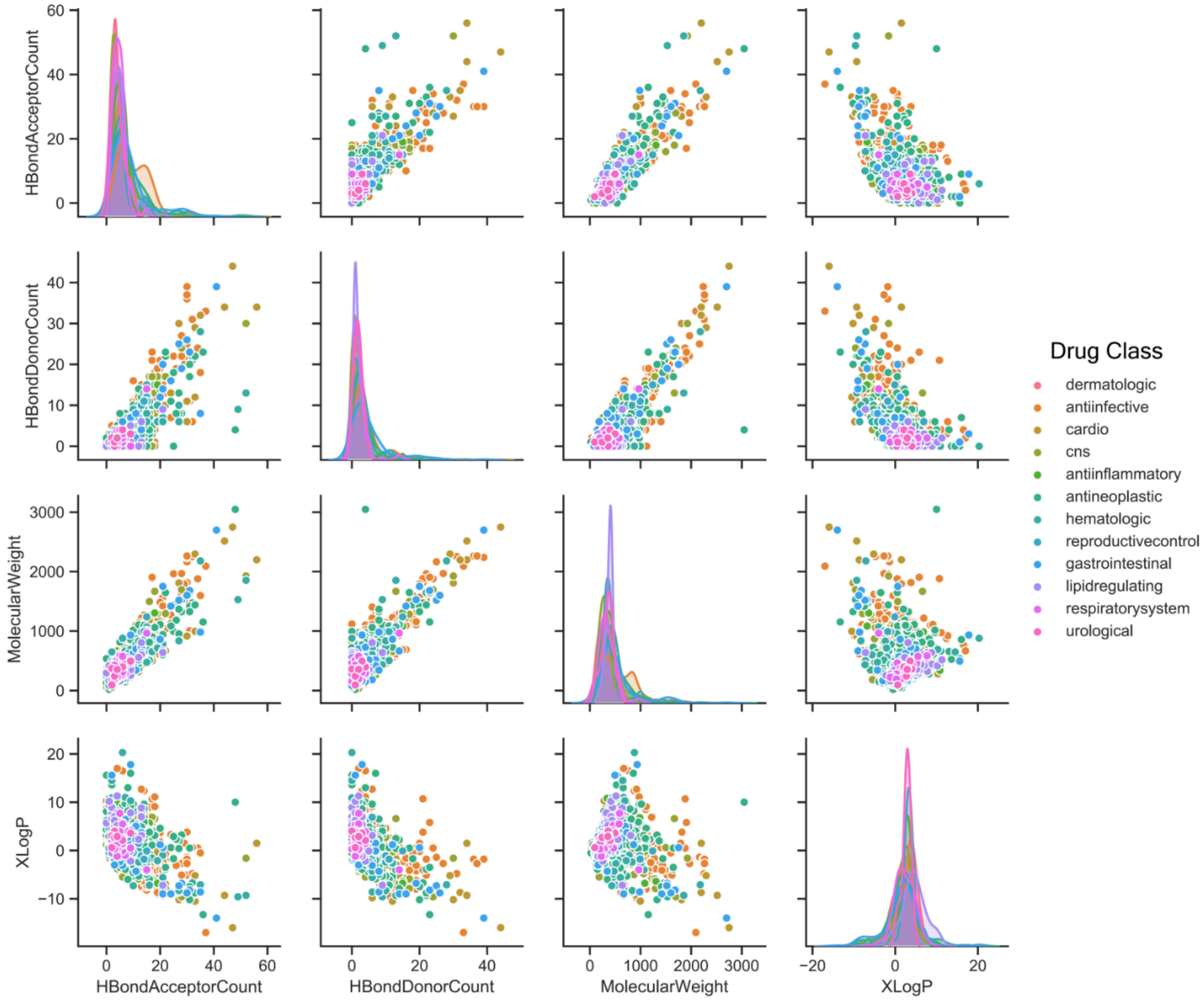

Figure S5: Pairplots showing the distributions and correlations among the molecule properties for each drug classification.

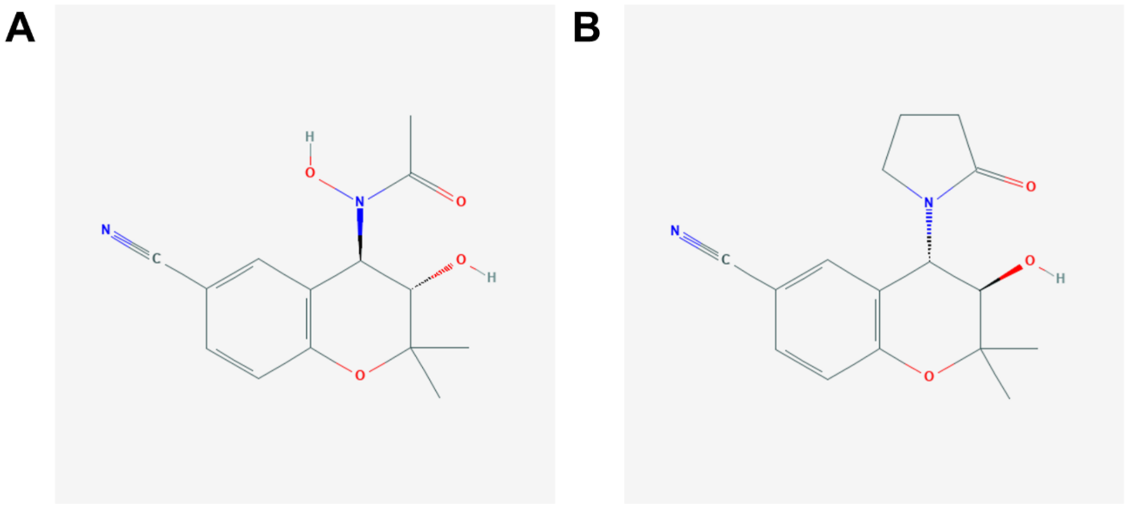

Figure S6: Structures of (A) the molecule with PubChem ID 121878 and (B) cromakalim. Molecule A was misclassified as a respiratory system drug, but shares significant chemical similarity with the known bronchodilator molecule B.

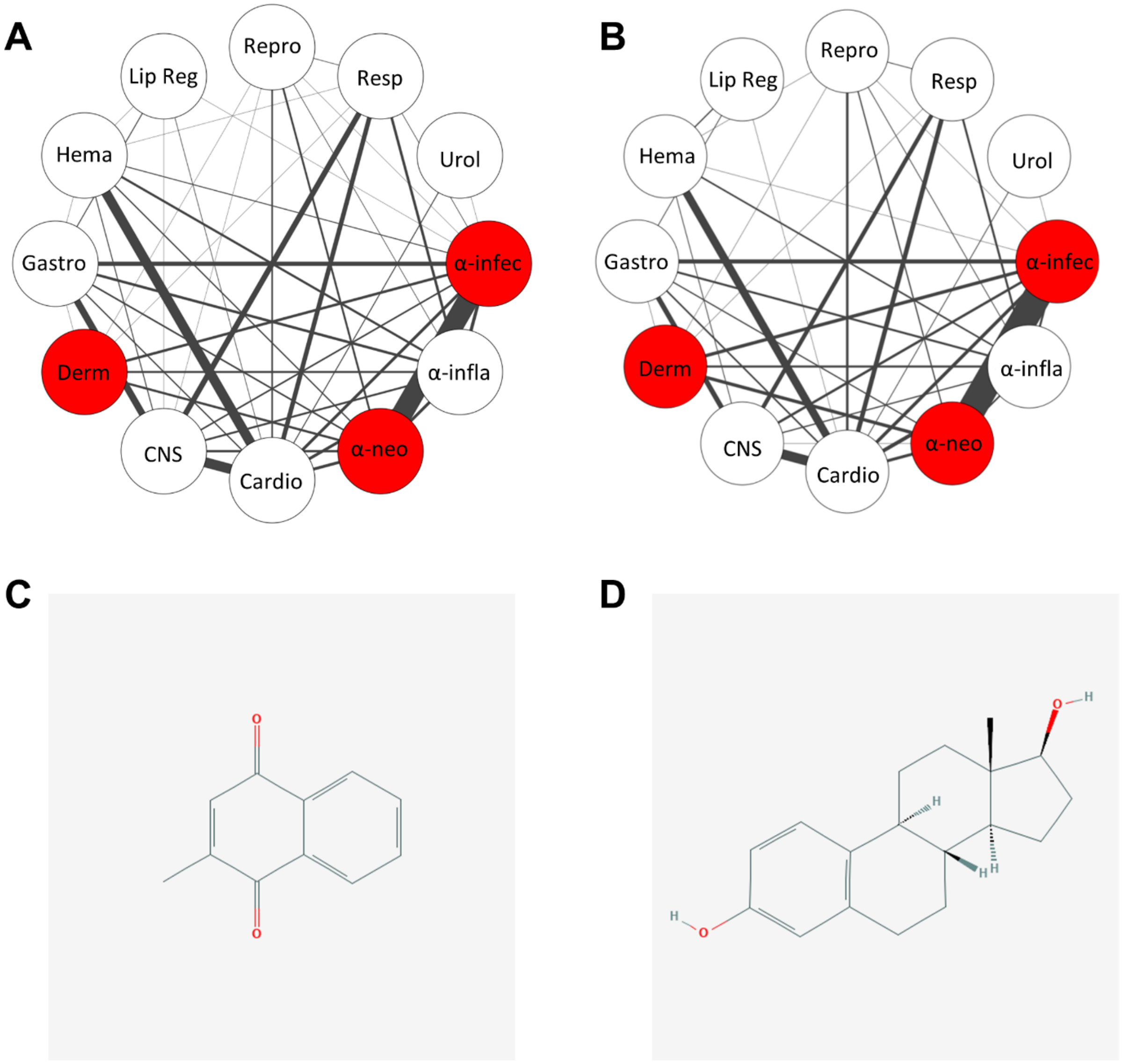

Figure S7: Network analysis of pairwise class relationships from multi-class predictions helps discover new drug repurposing opportunities. Molecules in validation fold #5 were used to make networks of relationships between (A) true class labels and (B) predicted class labels. (C) Chemical structure of menadione, the hematologic compound predicted to also function as reproductive control agent. (D) Chemical structure of estrogen, a common reproductive control drug.