Supporting Information

Improving Measures of Chemical Structural Similarity Using Machine Learning on Chemical-Genetic Interactions

Hamid Safizadeh,[†,‡] Scott W. Simpkins,[§,◊] Justin Nelson,[§,¤] Sheena C. Li,[||,Δ] Jeff S. Piotrowski,[Δ,¤] Mami Yoshimura,[Δ] Yoko Yashiroda,[Δ] Hiroyuki Hirano,[Δ] Hiroyuki Osada,[Δ] Minoru Yoshida,[Δ,#] Charles Boone,[||,⊥,Δ] Chad L. Myers[‡,§,]*

[†]Department of Electrical and Computer Engineering, University of Minnesota-Twin Cities, Minneapolis, Minnesota 55455, United States

[‡]Department of Computer Science and Engineering, University of Minnesota-Twin Cities, Minneapolis, Minnesota 55455, United States

[§]Bioinformatics and Computational Biology Graduate Program, University of Minnesota-Twin Cities, Minneapolis, Minnesota 55455, United States

[||]The Donnelly Centre, University of Toronto, Toronto, Ontario M5S 3E1, Canada

[⊥]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 3E1, Canada

[Δ]RIKEN Center for Sustainable Resource Science (CSRS), Wako, Saitama 351-0198, Japan

[#]Department of Biotechnology and Collaborative Research Institute for Innovative Microbiology, The University of Tokyo, Bunkyo City, Tokyo 113-8654, Japan

◊Present address: Octant Inc., Emeryville, California 94608, United States

¤Present address: Yumanity Therapeutics, Boston, Massachusetts 02135, United States

*Correspondence to chadm@umn.edu

Table of Contents

Supporting algorithm. Algorithm for finding the best performing structural similarity measures

	Steps for the systematic benchmarking of structural similarity measures based on chemical-genetic interaction data	Results (RIKEN high-confidence set)
	Selecting the best-performing molecular fingerprints	
1	Generate molecular fingerprints for the compound collections for which we have chemical-genetic interaction profiles.	
2	Establish the binarized gold standard for biological activity (10% of the most similar compound pairs based on chemical-genetic cosine similarity).	
3	Compute precision for each structural similarity measure at several recall thresholds (Each structural similarity measure is defined as one molecular fingerprint paired with one similarity coefficient).	
4	Select up to the top three fingerprints based on the highest precision at each recall threshold (Precision values within the 10% of the maximum precision were considered equivalent).	Recall = 0.002: ASP, DFS, RAD2D Recall = 0.005: ASP, LSTAR, RAD2D Recall = 0.02: ASP, LSTAR, RAD2D Reall = 0.05: ASP, LSTAR, RAD2D Recall = 0.2: LSTAR
	Top-performing molecular fingerprints: ASP, LSTAR, and RAD2D	
	Selecting the best-performing similarity coefficients	
1	Remove the similarity coefficients for which the precision was < 80% of the maximum precision achieved at the majority of recall thresholds for the majority of molecular fingerprints.	Exclude: Dot-product, Euclidean, Russel/Rao, Simpson
2	Select the similarity coefficients that appear in the top three in terms of precision at all recall thresholds (based on the best-performing molecular fingerprint for that similarity coefficient).	Recall = 0.002: Braun-Blanquet, Tullos Recall = 0.005: Braun-Blanquet, Tullos Recall = 0.02: Braun-Blanquet, Dice, Sokal/Sneath, Tanimoto, Tullos Recall = 0.05: Braun-Blanquet, Dice, Sokal/Sneath, Tanimoto, Tullos Recall = 0.2: Braun-Blanquet, Dice, Sokal/Sneath, Tanimoto, Tullos
	Top-performing similarity coefficients: Braun-Blanquet and Tullos	

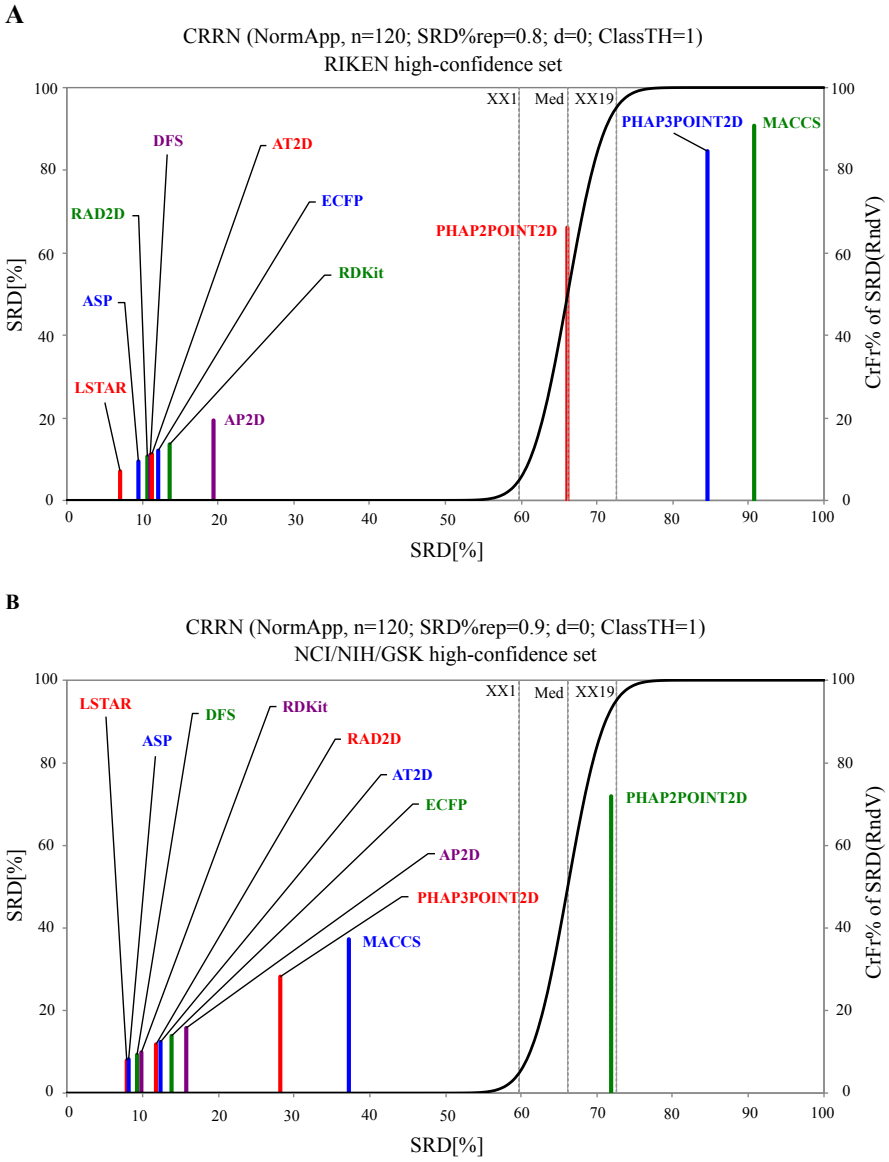


Figure S1. SRD analysis for molecular fingerprints. Since our systematic benchmark compares many structural similarity measures/models, we used the SRD method/model comparison approach[72] as a multi-criteria decision making framework to ensure that we fairly compared the components of our similarity models. We downloaded, from http://aki.ttk.hu/srd, the Excel macro program file for the SRD analysis with ties (i.e., repeated observations)[73]. We generated a matrix of our data that contained the precision at all predefined recall thresholds and the areas under the ROC curves as two sets of criteria for evaluating the performance of our structural similarity

models (120x11 data matrix by transposing and combining all spreadsheet tables from Table S1 or Table S2; rows: 12 similarity coefficients measured at 9 recall thresholds and for the areas under the ROC curves; columns: 11 molecular fingerprints). We used only the fingerprints with a depth of 8 and the symmetric similarity coefficients (Tversky with $\alpha = 0.9$ was removed) for this SRD analysis. We used the maximum of the row values as the reference ranking because we searched for the molecular fingerprint with the highest performance. We validated the resulting sum of absolute ranking differences using the approximated normal distribution of random numbers. The Comparison of Ranks by Random Numbers (CRRN) validation[74] using (A) the RIKEN and (B) the NCI/NIH/GSK high-confidence sets confirmed the superiority of LSTAR and ASP fingerprints over all other molecular fingerprints. The colors were assigned to the bars on a rotational basis by the SRD program file.

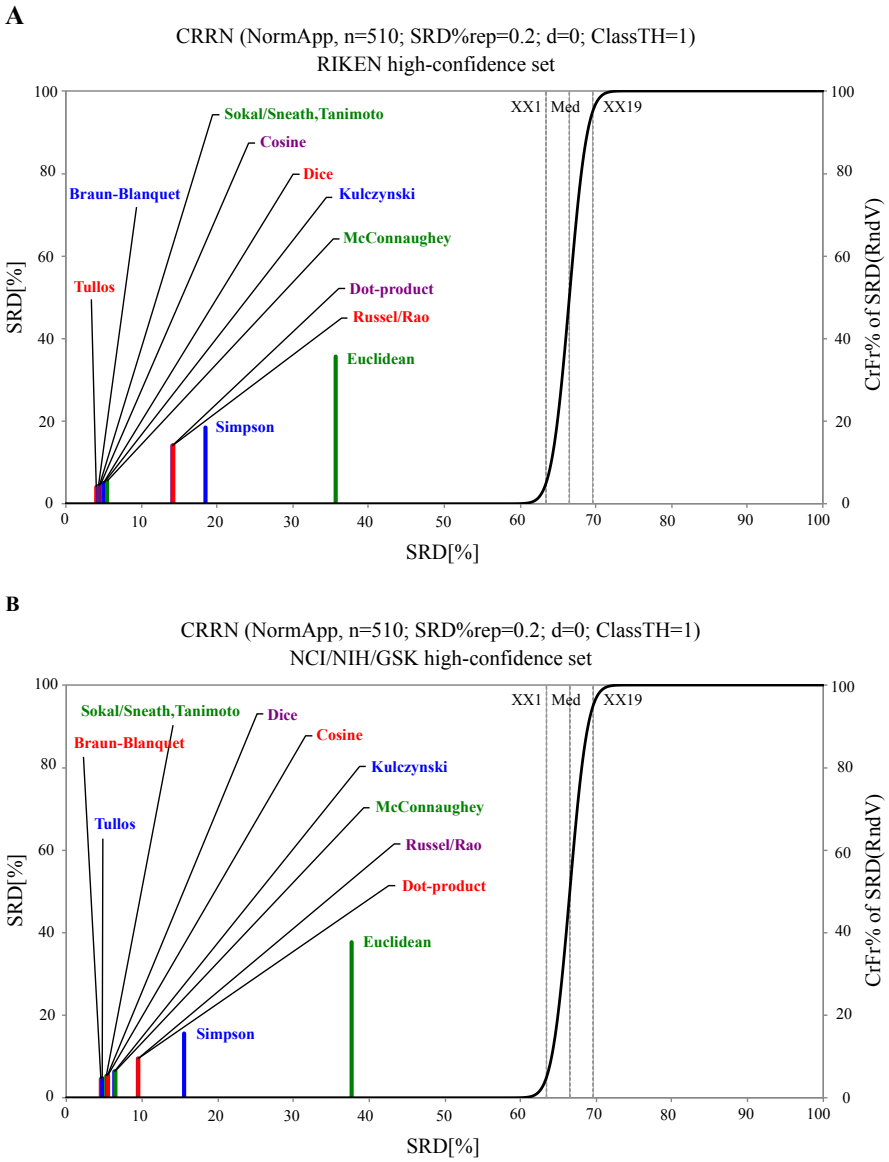


Figure S2. SRD analysis for similarity coefficients. For each compound collection, we generated a matrix of our data (510x12 matrix by combining all spreadsheet tables from Table S1 or Table S2; rows: 11 molecular fingerprints measured at all available depths, at 9 recall thresholds, and for the areas under the ROC curves; columns: 12 symmetric similarity coefficients). Since we included the Dice and Tanimoto similarity coefficients, which are two symmetric instances of the Tversky coefficient, we removed Tversky with $\alpha = 0.9$ from this SRD analysis. We used the maximum of the row values as the reference ranking because we searched for the similarity coefficient with the

highest performance. The CRRN validation for (A) the RIKEN and (B) the NCI/NIH/GSK high-confidence sets indicated the relative superiority of the Tullos and Braun-Blanquet coefficients over all other similarity coefficients. Although several similarity coefficients were competing for high performance based on the CRRN validations for our compound collections, the Dot-product, Euclidean, Russel/Rao, and Simpson similarity coefficients were distinctly the worst coefficients, whereas the Tullos and Braun-Blanquet similarity coefficients were the best ones. While the Tullos coefficient achieved comparable performance to the Braun-Blanquet coefficient, the simplicity of the Braun-Blanquet coefficient makes this coefficient preferable in most practical scenarios. We also repeated this SRD analysis for individual recall thresholds (51x12 data matrices), which again confirmed the superiority of the Braun-Blanquet and Tullos similarity coefficients over all other coefficients at our predefined recall thresholds (Tables S1 and S2 – SRD analysis spreadsheets). The colors were assigned to the bars on a rotational basis by the SRD program file.

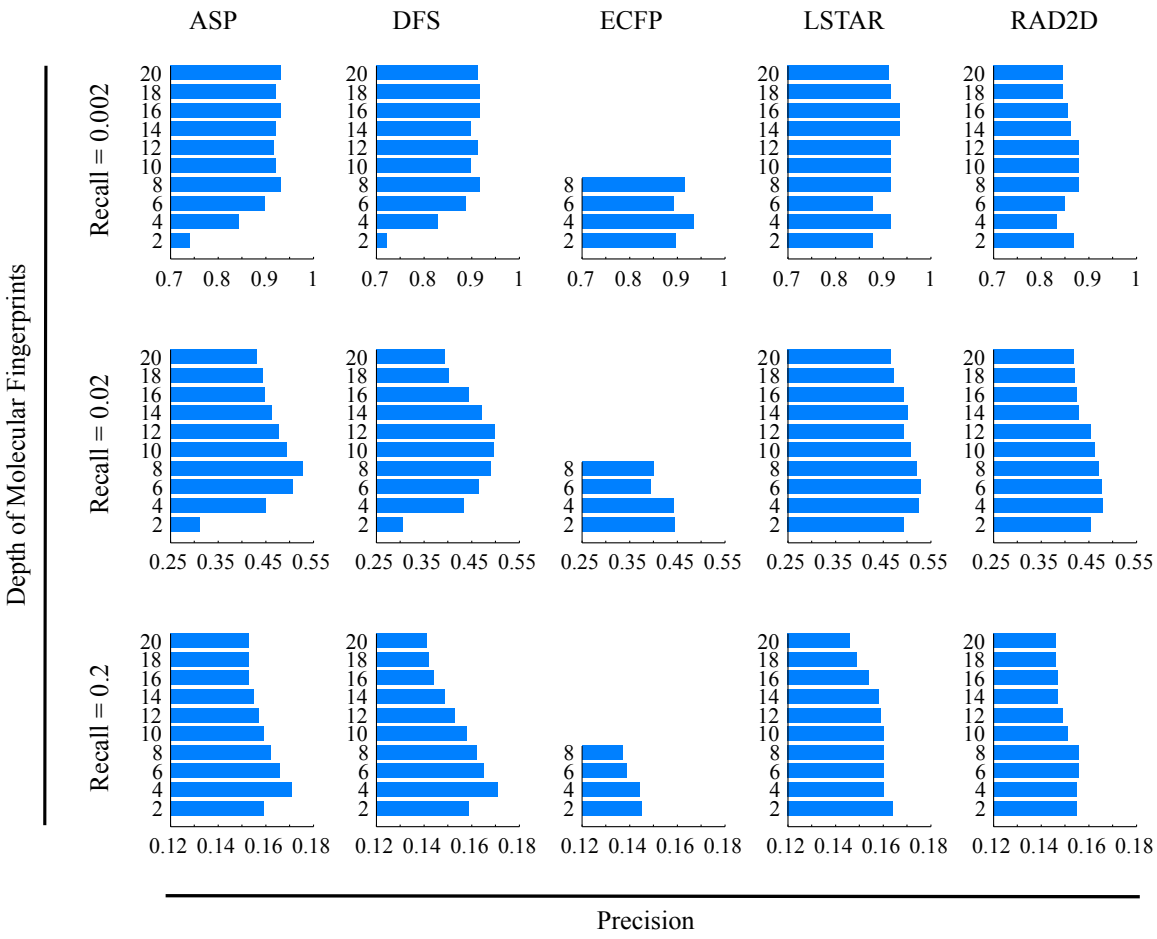
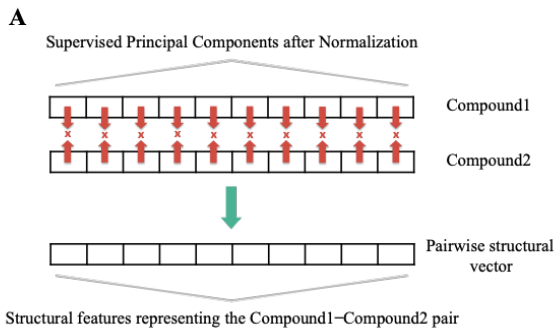


Figure S3. Impact of the describing depth of molecular fingerprints on the NCI/NIH/GSK high-confidence set. We measured the precision of our prediction models at 10 molecular depths, ranging from 2 to 20, for five different molecular fingerprints. Similarities were calculated with the Braun-Blanquet similarity coefficient, and the precision at three different recall thresholds for each molecular depth is shown.



B

	Bootstrap 1	Bootstrap 2	Bootstrap 3	Bootstrap 4	Mean
Pair 1	X	X		X	3 values
Pair 2	X	X	X	X	4 values
Pair 3			X		1 value
Pair 4	X	X			2 values
Pair 5			X	X	2 values

X: Both compounds in the pair are in the test set and consequently have a predicted chemical-genetic similarity in the bootstrap.

Figure S4. Pairwise structural vectors and bootstrapping used by our machine learning pipeline. (A) We generated the pairwise structural features/vector for a compound pair by the element-wise multiplication of the normalized, low-dimensional structural vectors describing the compounds. (B) Bootstrap aggregating (bagging) example. For an example of five compound pairs and four bootstraps, we illustrate the bootstrap aggregating procedure: We averaged the predicted chemical-genetic similarities of a test pair (represented by X values in a row) over all bootstraps to generate the final prediction for the test pair. For example, compound pair 1 was a test pair in bootstraps 1, 2, and 4; therefore, we averaged these three predictions to form the final prediction for pair 1 (In bootstrap 3, compound pair 1 was either a training pair or an invalid pair for which one compound belonged to the training set and the other to the test set).

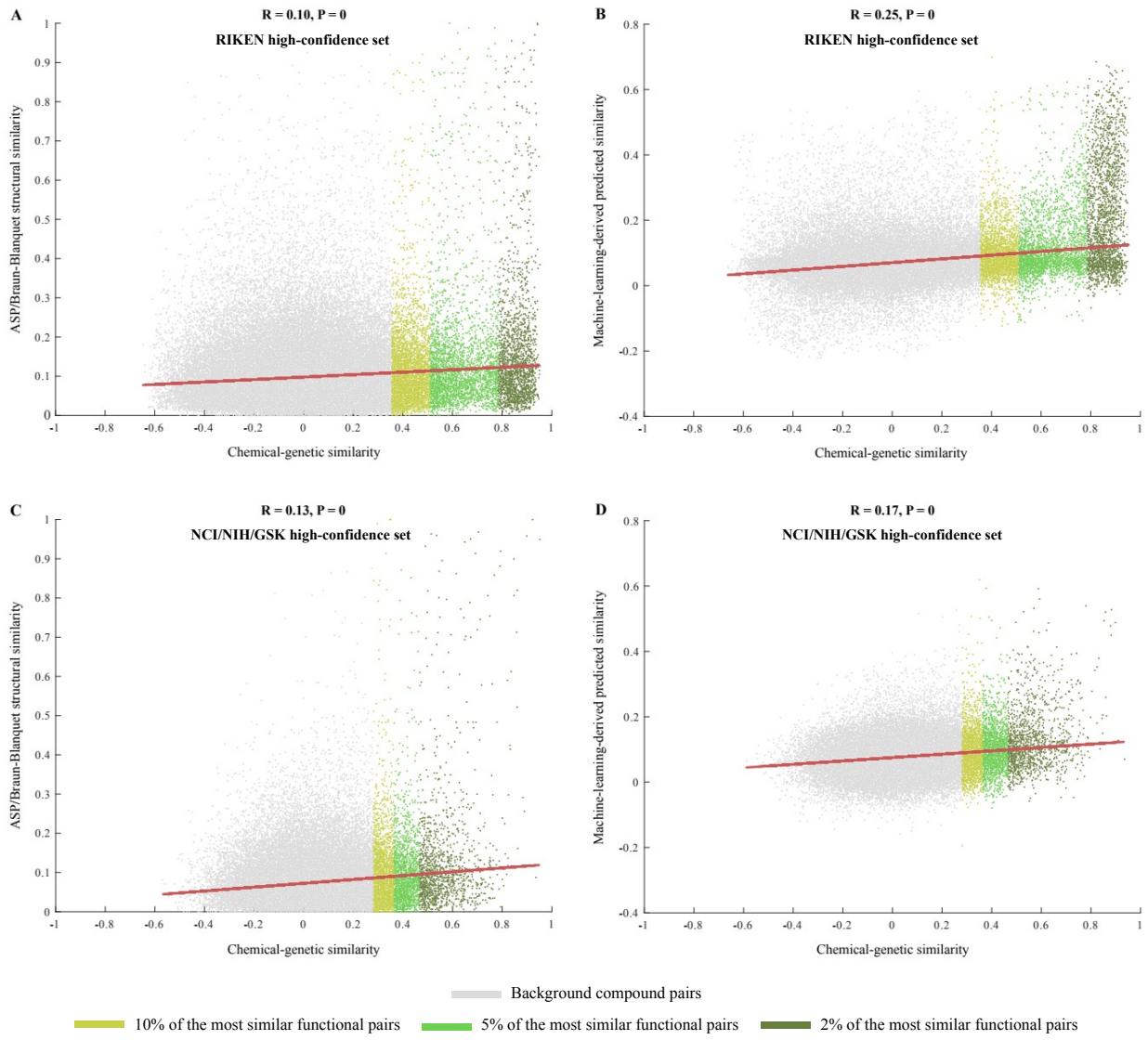


Figure S5. Correlation analysis of the predicted structural similarities with the chemical-genetic similarities. Correlation of (A) the ASP/Braun-Blanquet-derived structural similarity and (B) the machine-learning-derived predicted similarity with chemical-genetic similarity for the RIKEN high-confidence set. Correlation of (C) the ASP/Braun-Blanquet-derived structural similarity and (D) the machine-learning-derived predicted similarity with chemical-genetic similarity for the NCI/NIH/GSK high-confidence set. We computed all the machine-learning-derived predicted similarities from bootstrap aggregating, where we calculated the prediction for a compound pair

as the mean of the model output over all the bootstraps for which the compound pair was in the test set (Figure S4B). R and P represent the Pearson correlation coefficient and the corresponding p-value, respectively.

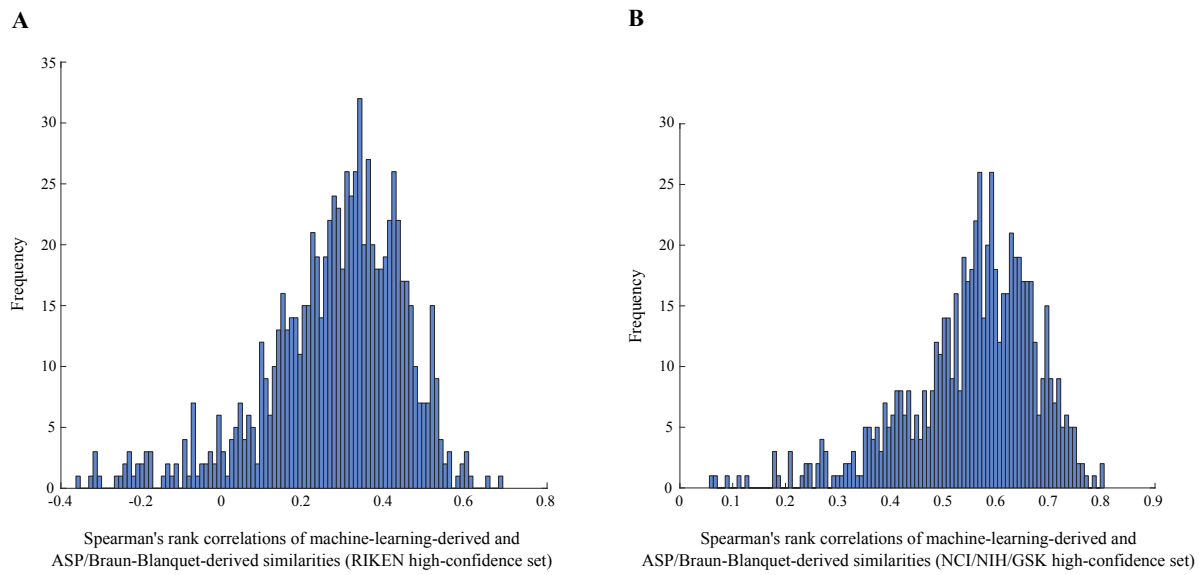


Figure S6. Spearman's rank correlation distribution between the machine-learning-derived and the ASP/Braun-Blanquet-derived structural similarity predictions. We measured the Spearman's rank correlation distribution for (A) the RIKEN and (B) the NCI/NIH/GSK high-confidence sets. Each value represents the Spearman's rank correlation for a compound. To compute the correlation value for a compound, we measured the Spearman's rank correlation between two lists of predicted similarities, where one list was generated by our machine learning model and the other list by the Braun-Blanquet similarity coefficient, both using ASP fingerprints.

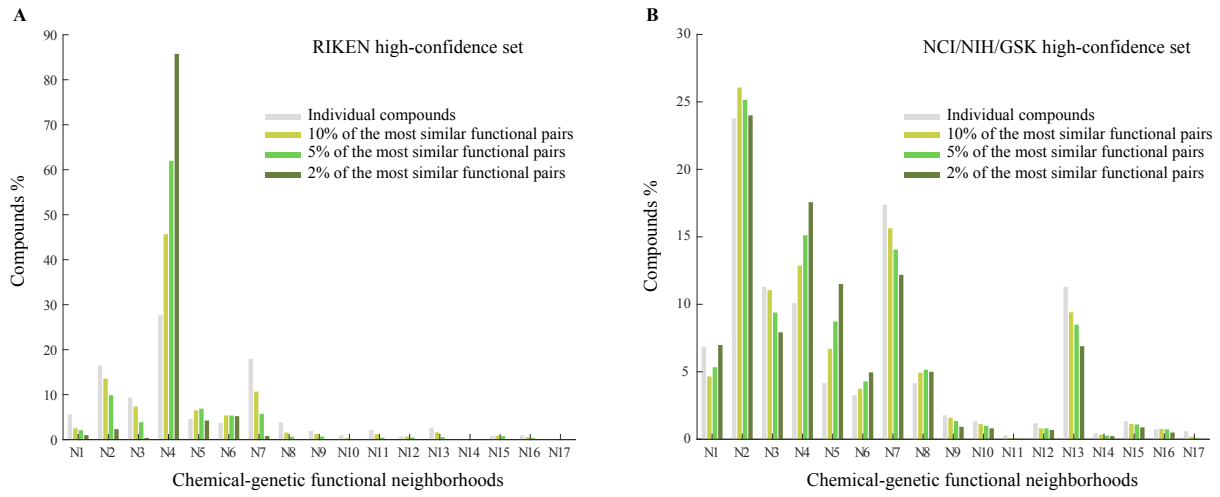


Figure S7. Distribution of the predicted biological functions across 17 broad, previously defined functional neighborhoods[32,34] for (A) the RIKEN and (B) the NCI/NIH/GSK high-confidence sets. The gray bar heights represent the distribution of all the studied compounds in these sets, whereas the green bar heights represent the distribution of a subset of these compounds.

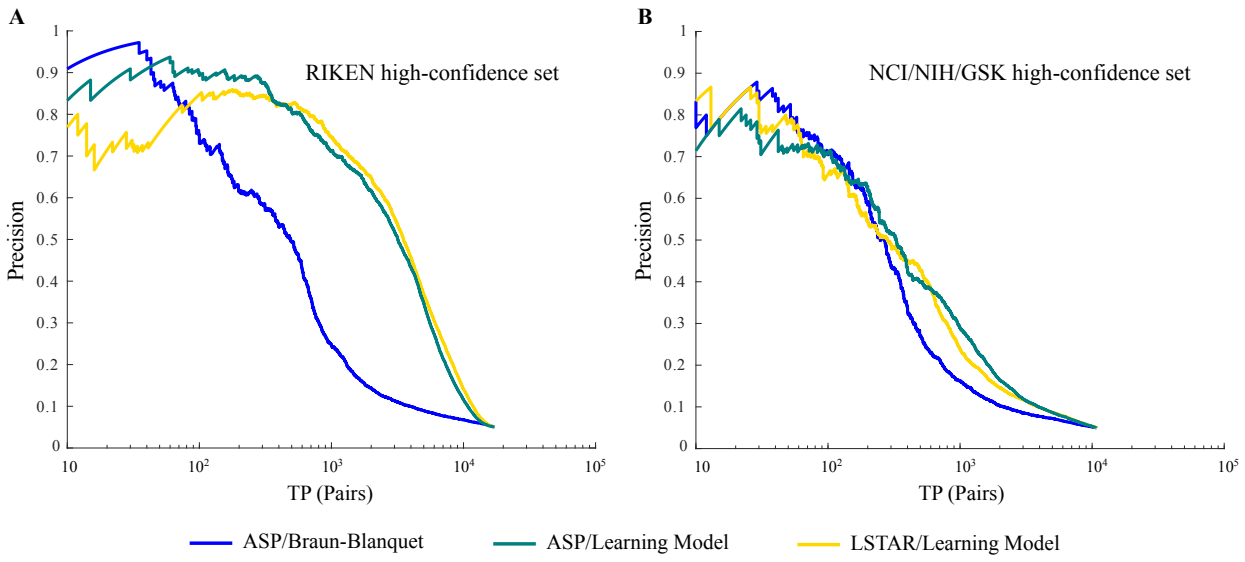


Figure S8. Prediction performance of the machine learning models for the 5% cutoff (as a more stringent cutoff than 10%) on the functional similarity gold standard. (A) Model performance for the RIKEN high-confidence set. The blue PR curve represents the prediction performance gained by the best-performing structural similarity measure (ASP/Braun-Blanquet), whereas the teal and gold PR curves represent the performance of the machine learning models using ASP and LSTAR fingerprints, respectively. A prediction is considered a true positive if the compound pair is within the top 5% of functionally similar compound pairs using chemical-genetic interaction profiles. (B) Model performance for the NCI/NIH/GSK high-confidence set.