Supporting Information

STarFish: A **S**tacked Ensemble **Tar**get **Fish**ing Approach and its Application to Natural Products

Nicholas T. Cockroft[†]*, Xiaolin Cheng*[†]**, and James R. Fuchs*[†]***

† Division of Medicinal Chemistry & Pharmacognosy, College of Pharmacy, The Ohio State University, Columbus, Ohio 43210, United States

*E-mail: cheng.1302@osu.edu., *E-mail: fuchs.42@osu.edu.

Table S1. Results of KNN base classifier tuning on a stratified random 90/10 train/test split of the training dataset for each cross-validation fold.

metric	n_neighbors	micro_AUROC	macro_AUROC	Frac_1_in_top10	Frac_all_in_top10	micro_BEDROC	macro_BEDROC	coverage
jaccard	1	0.796388 (0.00176)	0.780529 (0.002284)	0.617222 (0.00322)	0.594705 (0.00355)	0.612085 (0.003563)	0.583117 (0.004359)	744.301234 (6.726699)
jaccard	5	0.923237 (0.001405)	0.909096 (0.002784)	0.868757 (0.00295)	0.844148 (0.002442)	0.853131 (0.002763)	0.826756 (0.005218)	259.071313 (5.575081)
jaccard	10	0.940646 (0.000933)	0.924678 (0.001974)	0.899521 (0.002092)	0.875196 (0.002173)	0.885127 (0.001885)	0.855073 (0.003874)	194.402763 (2.762402)
minkowski	1	0.794396 (0.002119)	0.779087 (0.002493)	0.613079 (0.00411)	0.590713 (0.00434)	0.608348 (0.004196)	0.580374 (0.00478)	752.279818 (8.264903)
minkowski	5	0.92039 (0.001419)	0.906489 (0.002227)	0.86295 (0.002556)	0.838582 (0.002263)	0.847581 (0.002752)	0.821665 (0.004188)	270.140751 (5.016876)
minkowski	10	0.937781 (0.001126)	0.922548 (0.00218)	0.89363 (0.002057)	0.869434 (0.002147)	0.87948 (0.002298)	0.85074 (0.004148)	205.32594 (3.692552)

Table S2. Results of MLP base classifier tuning on a stratified random 90/10 train/test split of the training dataset for each cross-validation fold.

hidden_layer_sizes	micro_AUROC	macro_AUROC	Frac_1_in_top10	Frac_all_in_top10	micro_BEDROC	macro_BEDROC	coverage
(100,)	0.983686 (0.001387)	0.984229 (0.000853)	0.871111 (0.004238)	0.847315 (0.004101)	0.919742 (0.004053)	0.917158 (0.002717)	35.779642 (2.875651)
(1000, 100)	0.985217 (0.001098)	0.982851 (0.000956)	0.85854 (0.003482)	0.834136 (0.003386)	0.914784 (0.003445)	0.911042 (0.002612)	30.289004 (1.233247)
(1000, 1000)	0.980335 (0.001652)	0.978336 (0.001368)	0.862274 (0.003771)	0.838016 (0.004102)	0.901066 (0.005523)	0.896809 (0.004918)	30.565095 (1.180071)
(1000, 1000, 100)	0.982636 (0.001496)	0.979935 (0.001754)	0.846172 (0.003339)	0.822105 (0.003833)	0.900011 (0.00702)	0.895426 (0.00801)	30.411428 (0.740376)
(1000, 1000, 1000)	0.982331 (0.00317)	0.981026 (0.002729)	0.848929 (0.005684)	0.824962 (0.005778)	0.896086 (0.014113)	0.890992 (0.015034)	29.929988 (2.233989)
(1000,)	0.978462 (0.002088)	0.97966 (0.001521)	0.884938 (0.002679)	0.860374 (0.002652)	0.916133 (0.003451)	0.914374 (0.003722)	38.572175 (3.145195)

Table S3. Results of RF base classifier tuning on a stratified random 90/10 train/test split of the training dataset for each cross-validation fold.

max_features	n_estimators	micro_AUROC	macro_AUROC	Frac_1_in_top10	Frac_all_in_top10	micro_BEDROC	macro_BEDROC	coverage
0.333	10	0.889216 (0.001681)	0.876576 (0.001867)	0.80111 (0.003377)	0.776924 (0.003388)	0.787611 (0.003383)	0.764548 (0.003436)	388.936712 (6.855073)
0.333	100	0.930415 (0.001785)	0.914428 (0.002373)	0.872884 (0.003036)	0.848404 (0.003051)	0.863443 (0.003669)	0.833557 (0.004559)	231.333976 (6.515776)
auto	10	0.895672 (0.001981)	0.881752 (0.002302)	0.813003 (0.00384)	0.788617 (0.003898)	0.798712 (0.003663)	0.773639 (0.00409)	364.072808 (7.762124)
auto	100	0.9454 (0.001129)	0.927314 (0.002217)	0.893945 (0.002532)	0.869258 (0.002393)	0.889633 (0.002503)	0.855677 (0.004295)	173.219298 (3.652797)
auto	1000	0.965165 (0.000616)	0.946285 (0.002496)	0.911903 (0.001612)	0.887386 (0.001935)	0.917467 (0.001513)	0.88161 (0.004596)	99.560729 (2.017088)

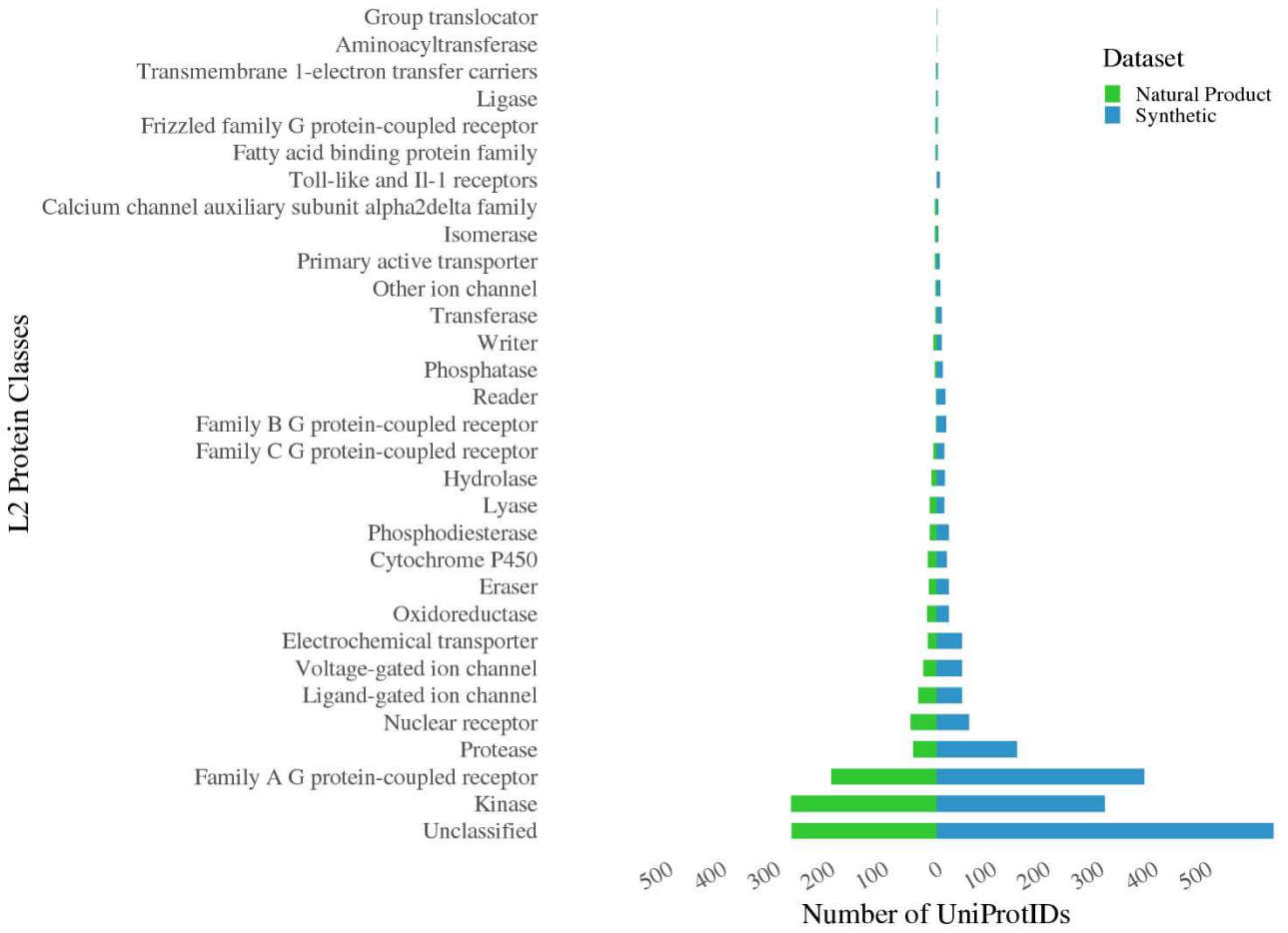

Figure S1. Comparison of the ChEMBL L2 protein classes between the synthetic compound dataset and the natural product dataset.

Table S4. Results of 10-fold cross-validation on the synthetic compound dataset.

Model	micro_AUROC	macro_AUROC	Frac_1_in_top10	Frac_all_in_top10	micro_BEDROC	macro_BEDROC	coverage	Type
KNN	0.942301 (0.00204)	0.926946 (0.002619)	0.903073 (0.003441)	0.878424 (0.004442)	0.887923 (0.003782)	0.858709 (0.004589)	187.432522 (5.965479)	Not Stacked
MLP	0.979047 (0.00154)	0.980881 (0.001518)	0.889626 (0.003988)	0.864936 (0.005078)	0.919088 (0.004065)	0.918307 (0.004824)	37.669495 (2.086298)	Not Stacked
RF	0.966171 (0.001907)	0.947825 (0.002528)	0.915714 (0.002656)	0.890845 (0.004556)	0.919514 (0.003409)	0.884665 (0.00463)	96.05143 (4.288706)	Not Stacked
MLP_RF	0.985001 (0.000851)	0.982912 (0.001303)	0.917994 (0.002563)	0.893223 (0.004476)	0.933013 (0.003042)	0.916546 (0.004095)	29.402247 (1.140787)	Not Stacked
KNN_MLP	0.981905 (0.001228)	0.982763 (0.001351)	0.91647 (0.003034)	0.891882 (0.004705)	0.938412 (0.003113)	0.931853 (0.003854)	34.485059 (1.744688)	Not Stacked
KNN_RF	0.966719 (0.001945)	0.948387 (0.002586)	0.917974 (0.003122)	0.893239 (0.004643)	0.923396 (0.00365)	0.887987 (0.004632)	94.662067 (4.420363)	Not Stacked
KNN_MLP_RF	0.985188 (0.000868)	0.983078 (0.00131)	0.919445 (0.002943)	0.89475 (0.004735)	0.935549 (0.003141)	0.91861 (0.0041)	29.285386 (1.155606)	Not Stacked
KNN	0.989754 (0.001177)	0.989983 (0.001013)	0.926564 (0.002868)	0.902647 (0.005331)	0.950907 (0.003202)	0.950178 (0.002287)	21.891085 (2.285872)	Stacked
MLP	0.97467 (0.001503)	0.985998 (0.000695)	0.849289 (0.004211)	0.827107 (0.005889)	0.890829 (0.002729)	0.945159 (0.002604)	54.948871 (2.853801)	Stacked
RF	0.992117 (0.000812)	0.993382 (0.00081)	0.942336 (0.002908)	0.918346 (0.004259)	0.961141 (0.001832)	0.96196 (0.002096)	17.607684 (1.705223)	Stacked
MLP_RF	0.990756 (0.00075)	0.992894 (0.000855)	0.926472 (0.003249)	0.902666 (0.004413)	0.950749 (0.001785)	0.961727 (0.001978)	21.209976 (1.548678)	Stacked
KNN_MLP	0.99245 (0.000724)	0.992431 (0.000764)	0.935409 (0.002499)	0.911633 (0.003481)	0.958234 (0.001665)	0.958074 (0.002156)	17.41799 (1.516647)	Stacked
KNN_RF	0.994133 (0.000625)	0.99308 (0.000792)	0.945509 (0.002734)	0.921512 (0.004051)	0.967476 (0.001585)	0.961067 (0.002037)	13.722058 (1.306009)	Stacked
KNN_MLP_RF	0.993727 (0.000631)	0.993192 (0.000811)	0.940914 (0.00285)	0.91691 (0.003607)	0.963391 (0.001632)	0.961474 (0.002002)	14.998962 (1.308995)	Stacked

Table S5. Model performance results on the natural product benchmark.

Model	micro_AUROC	macro_AUROC	Frac_1_in_top10	Frac_all_in_top10	micro_BEDROC	macro_BEDROC	coverage	Type
KNN	0.700767	0.72812	0.566135	0.339775	0.430205	0.473909	1286.037056	Not Stacked
MLP	0.809043	0.821639	0.555327	0.342816	0.517641	0.513207	492.21719	Not Stacked
RF	0.805814	0.797995	0.601647	0.383253	0.560575	0.55939	870.233145	Not Stacked
MLP_RF	0.850754	0.845468	0.602162	0.386116	0.584146	0.584559	416.970664	Not Stacked
KNN_MLP	0.820458	0.836301	0.592898	0.380032	0.567536	0.587548	482.108595	Not Stacked
KNN_RF	0.806571	0.798452	0.599074	0.386831	0.565894	0.560926	866.966032	Not Stacked
KNN_MLP_RF	0.851126	0.845666	0.602676	0.388978	0.587345	0.585724	416.318065	Not Stacked
KNN	0.935312	0.889168	0.594442	0.425121	0.711535	0.687939	217.677818	Stacked
MLP	0.81811	0.719749	0.427689	0.278046	0.455236	0.485233	425.947504	Stacked
RF	0.917567	0.892882	0.630468	0.44194	0.692452	0.702486	230.226454	Stacked
MLP_RF	0.898711	0.874591	0.587751	0.405618	0.626237	0.68325	267.904786	Stacked
KNN_MLP	0.915233	0.883064	0.60422	0.421363	0.681662	0.676551	225.667524	Stacked
KNN_RF	0.938025	0.899854	0.637159	0.448918	0.732045	0.710923	190.437983	Stacked
KNN_MLP_RF	0.925955	0.890382	0.62069	0.435677	0.704735	0.697931	208.935152	Stacked

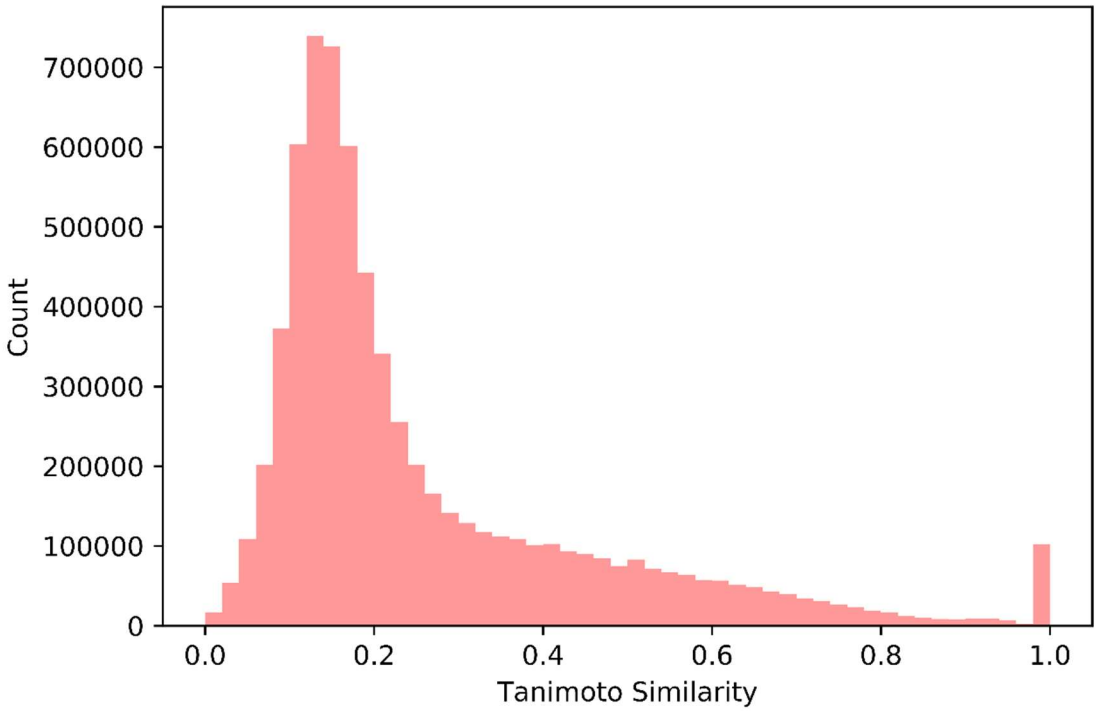

Figure S2. A histogram showing intra-target compound similarities. All pairwise compound similarities were calculated between the training compounds and themselves for each protein target label. The training compound set was obtained from a single cross-validation fold.

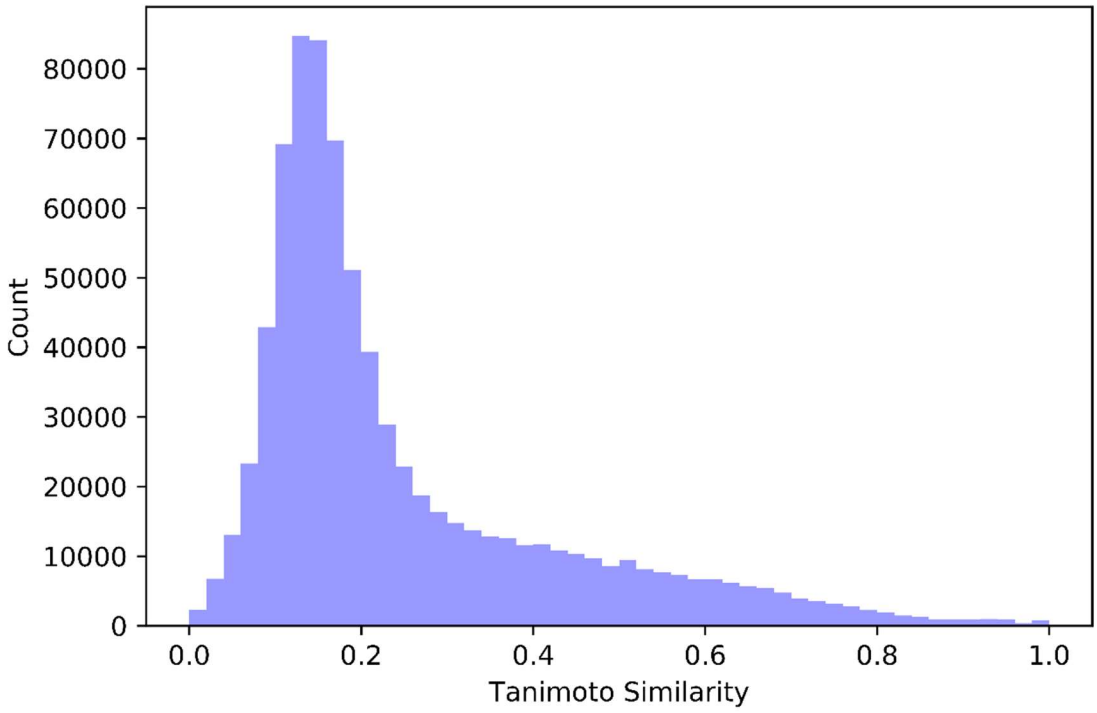

Figure S3. A histogram showing intra-target compound similarities. All pairwise compound similarities were calculated between the training compounds and the test set compounds for each protein target label. The training and test compound sets were obtained from a single cross-validation fold.

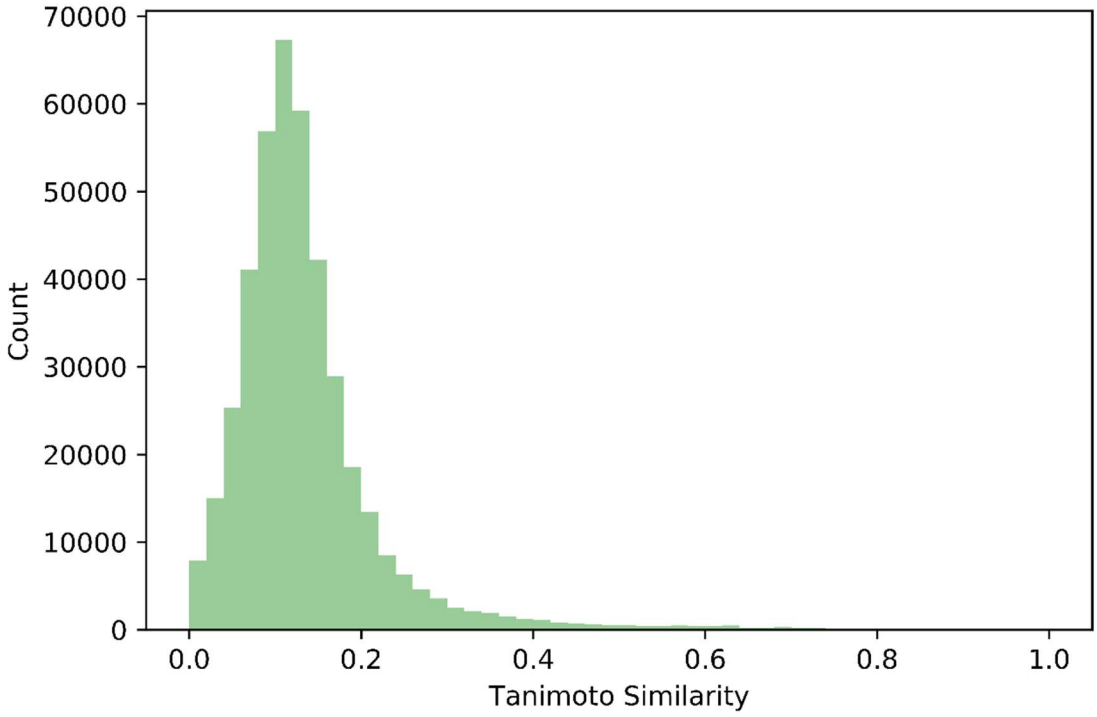

Figure S4. A histogram showing intra-target compound similarities. All pairwise compound similarities were calculated between the training compounds and the natural product benchmark compounds for each protein target label. The training compound set was obtained from a single cross-validation fold.

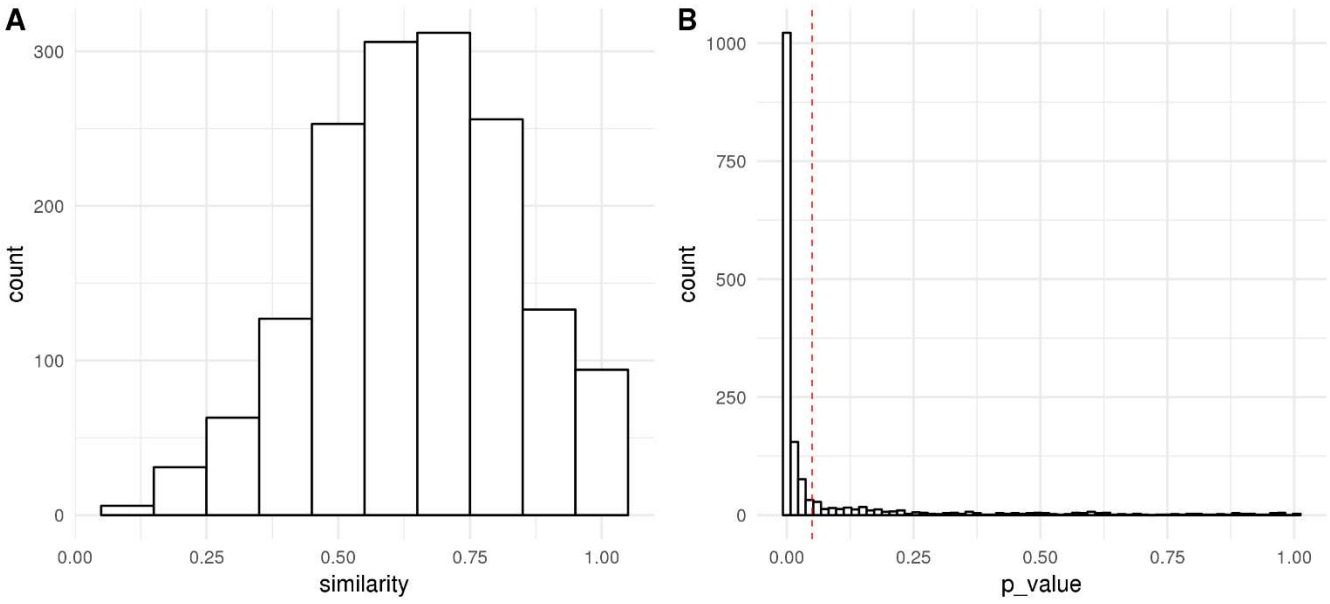

Figure S5. Comparison of protein functional similarity measured by semantic similarity of molecular function gene ontology (GO) ID annotations for each protein UniProt ID. Sets of protein target labels, as UniProt IDs, were obtained from the coefficients of the logistic regression models that were trained to predict each protein target label in the KNN stacked model. (A) Distribution of the average semantic similarities of each protein target label set. (B) Distribution of *p*-values for the average similarity values of each group of protein target labels. The dashed red line is placed at the p-value 0.05. 80% of the protein target label sets had a *p*-value < 0.05. Significance of group similarity for each query group of labels was assessed by a permutation test.

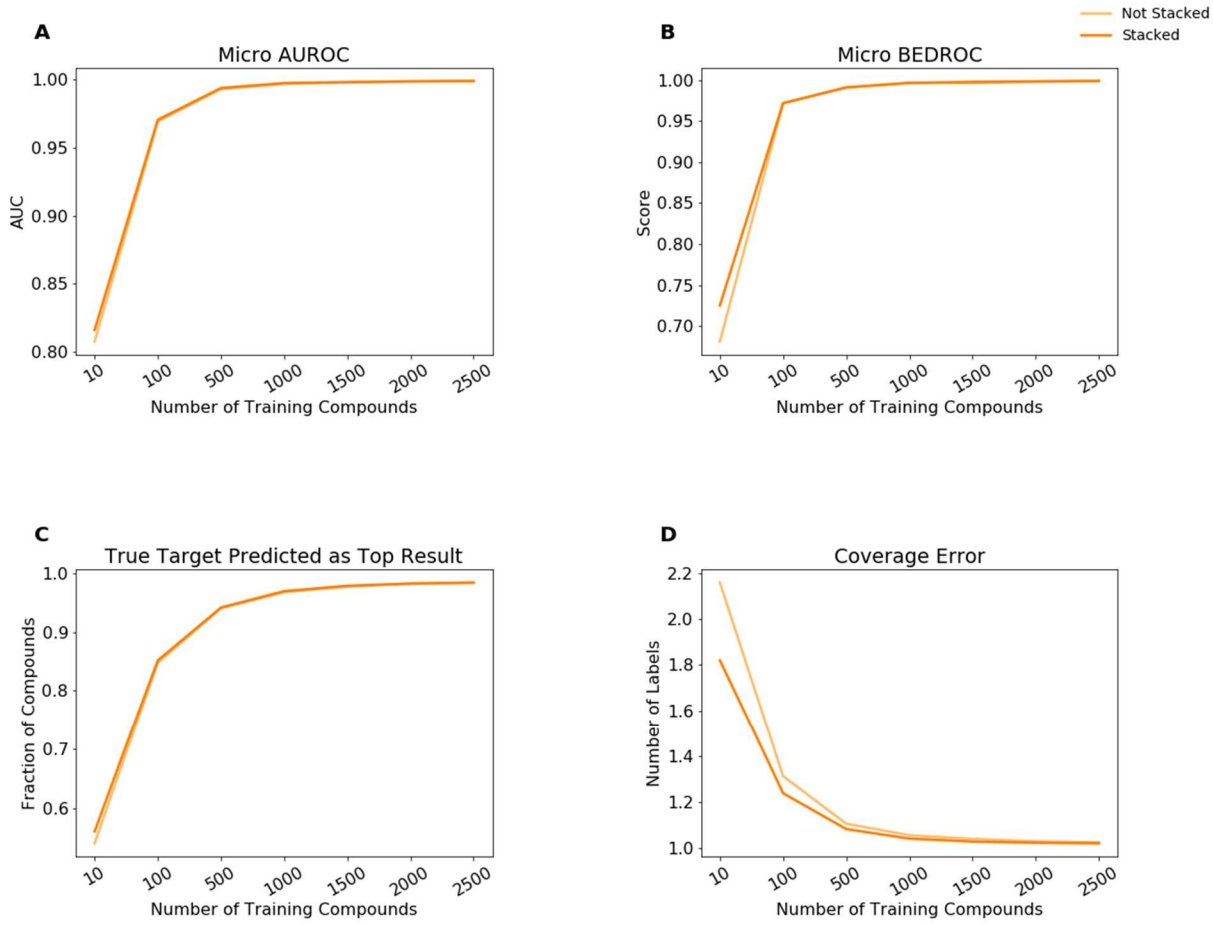

Figure S6. Model performance for stratified 10-fold cross-validation on datasets containing various numbers of compound training records for each protein target label for the KNN classifier. For a single model, "Not Stacked" indicates that the probability predictions of the listed model were used directly. If more than one model is listed, the mean probabilities for each label were used. "Stacked" indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by **(A)** micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, **(B)** micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic

(BEDROC), **(C)** the fraction of compounds which yielded a true target as the top prediction, and **(D)** coverage error are shown.

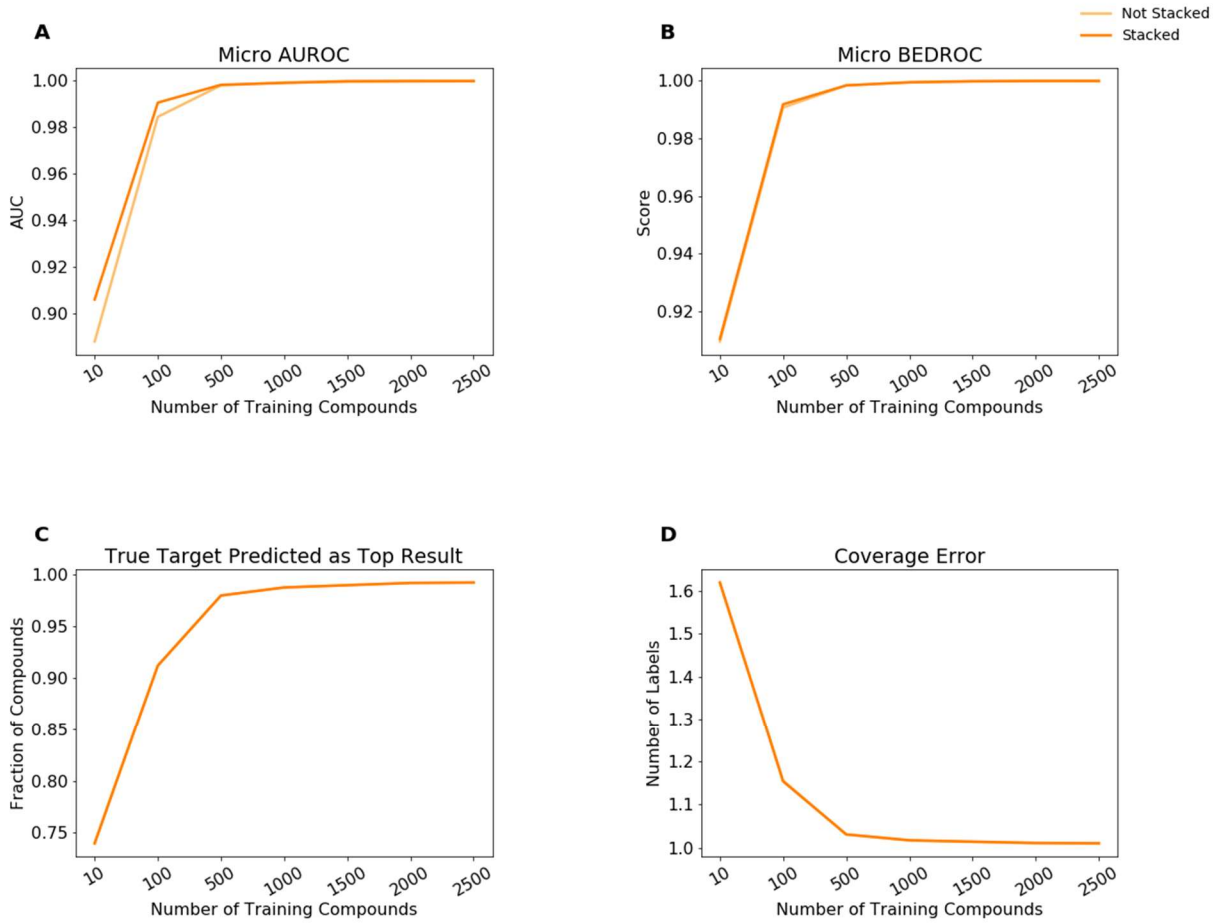

Figure S7. Model performance for stratified 10-fold cross-validation on datasets containing various numbers of compound training records for each protein target label for the MLP classifier. For a single model, "Not Stacked" indicates that the probability predictions of the listed model were used directly. If more than one model is listed, the mean probabilities for each label were used. "Stacked" indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by **(A)** micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, **(B)** micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic

(BEDROC), **(C)** the fraction of compounds which yielded a true target as the top prediction, and **(D)** coverage error are shown.

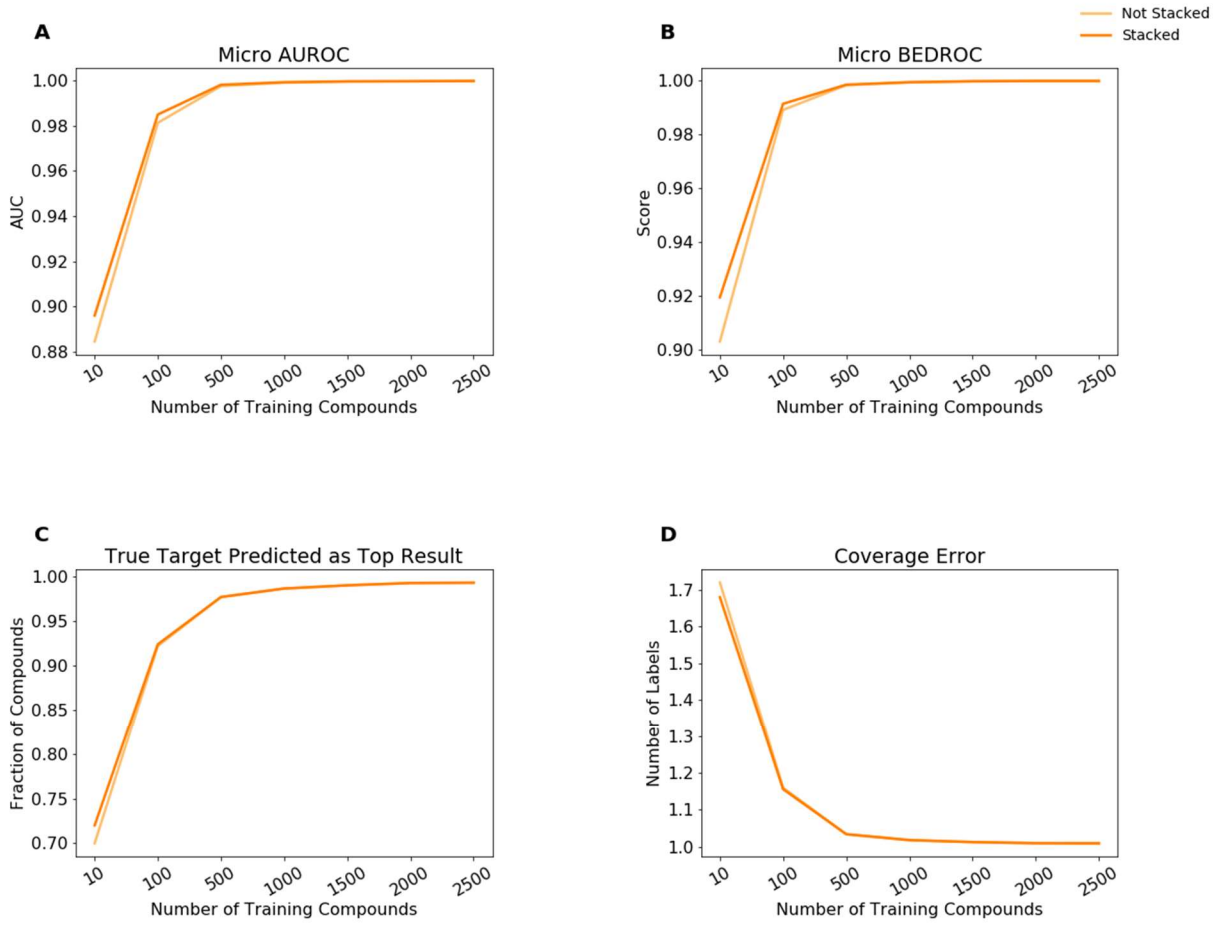

Figure S8. Model performance for stratified 10-fold cross-validation on datasets containing various numbers of compound training records for each protein target label for the RF classifier. For a single model, "Not Stacked" indicates that the probability predictions of the listed model were used directly. If more than one model is listed, the mean probabilities for each label were used. "Stacked" indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by **(A)** micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, **(B)** micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic

(BEDROC), **(C)** the fraction of compounds which yielded a true target as the top prediction, and **(D)** coverage error are shown.

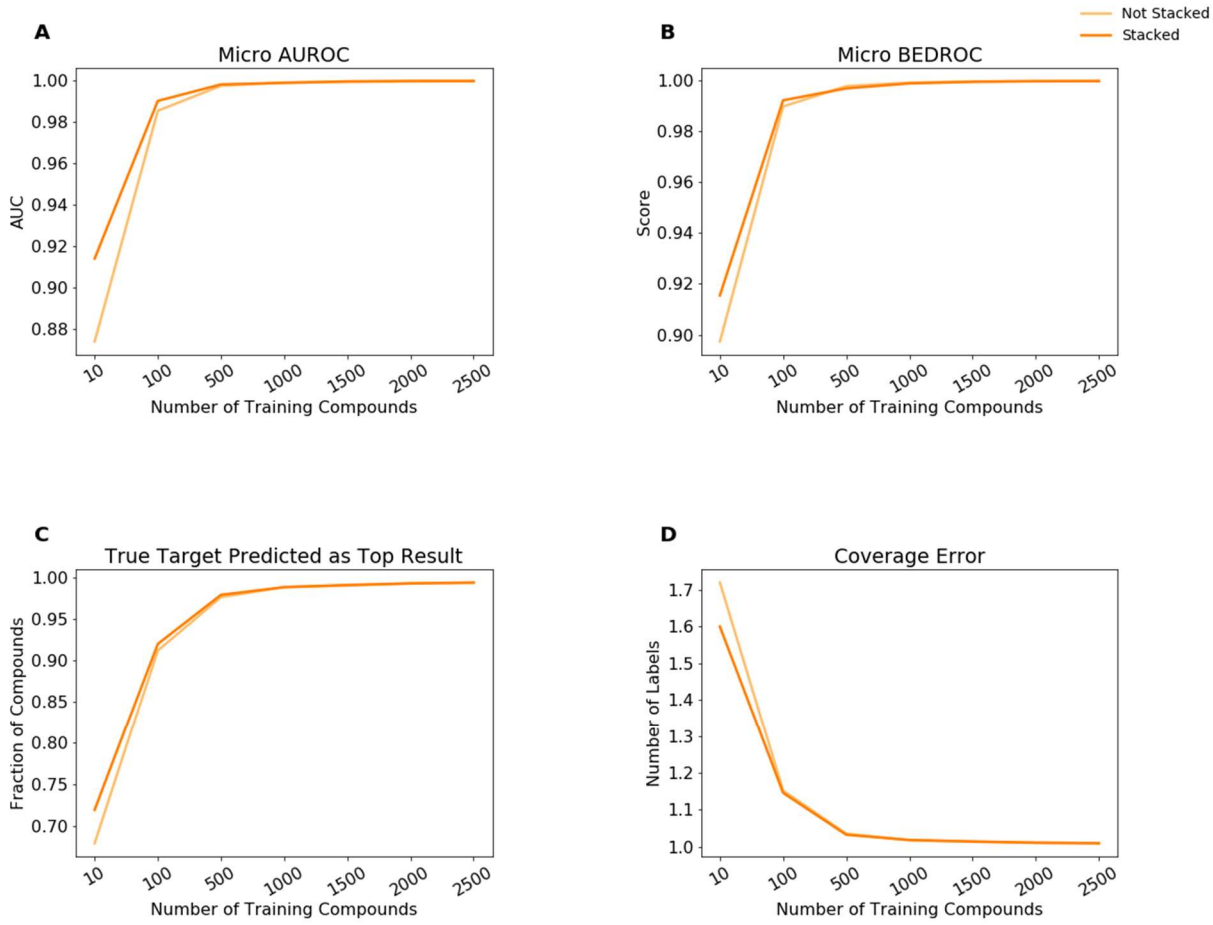

Figure S9. Model performance for stratified 10-fold cross-validation on datasets containing various numbers of compound training records for each protein target label for the KNN_MLP classifier. For a single model, "Not Stacked" indicates that the probability predictions of the listed model were used directly. If more than one model is listed, the mean probabilities for each label were used. "Stacked" indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by **(A)** micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, **(B)** micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic

(BEDROC), **(C)** the fraction of compounds which yielded a true target as the top prediction, and **(D)** coverage error are shown.

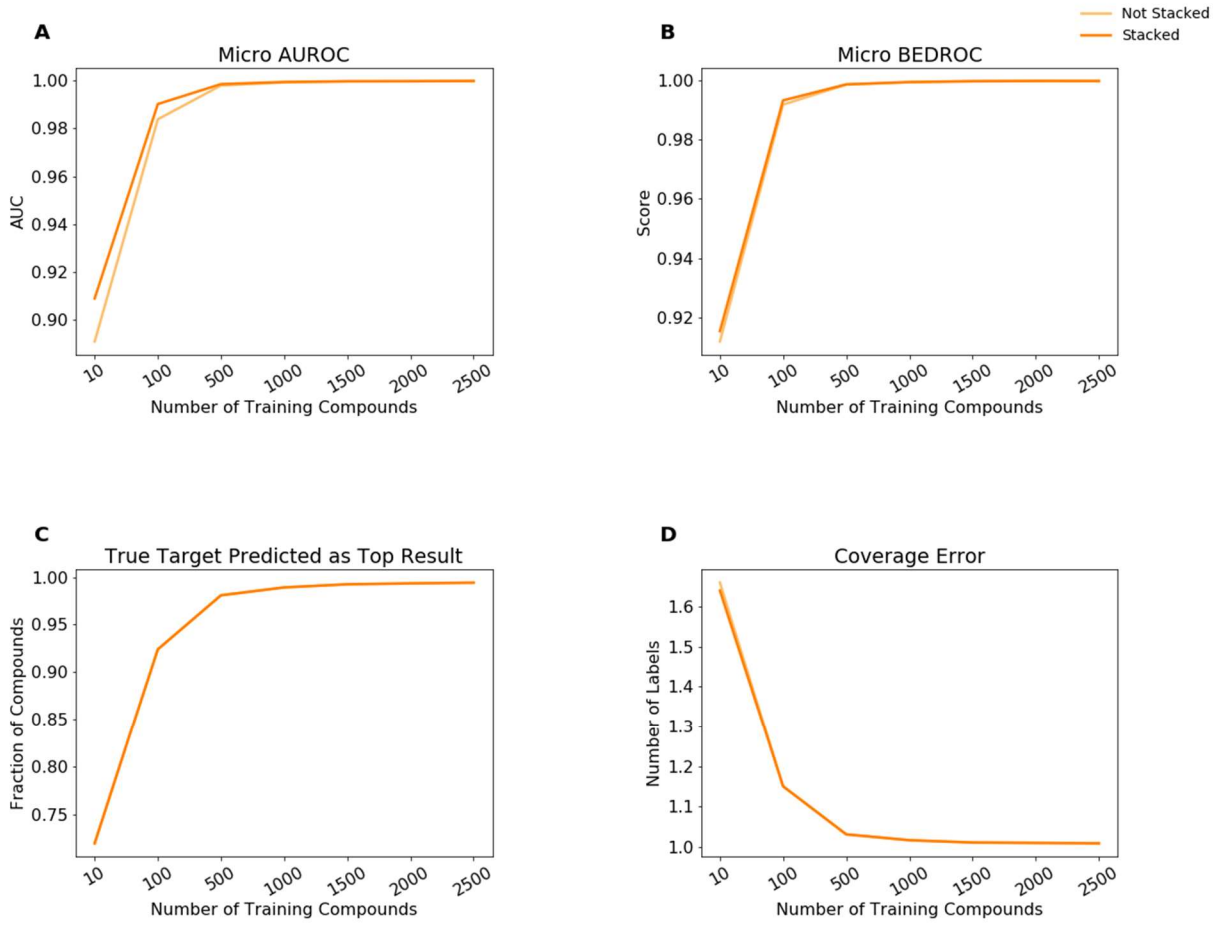

Figure S10. Model performance for stratified 10-fold cross-validation on datasets containing various numbers of compound training records for each protein target label for the MLP_RF classifier. For a single model, "Not Stacked" indicates that the probability predictions of the listed model were used directly. If more than one model is listed, the mean probabilities for each label were used. "Stacked" indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by **(A)** micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, **(B)** micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic

(BEDROC), **(C)** the fraction of compounds which yielded a true target as the top prediction, and **(D)** coverage error are shown.

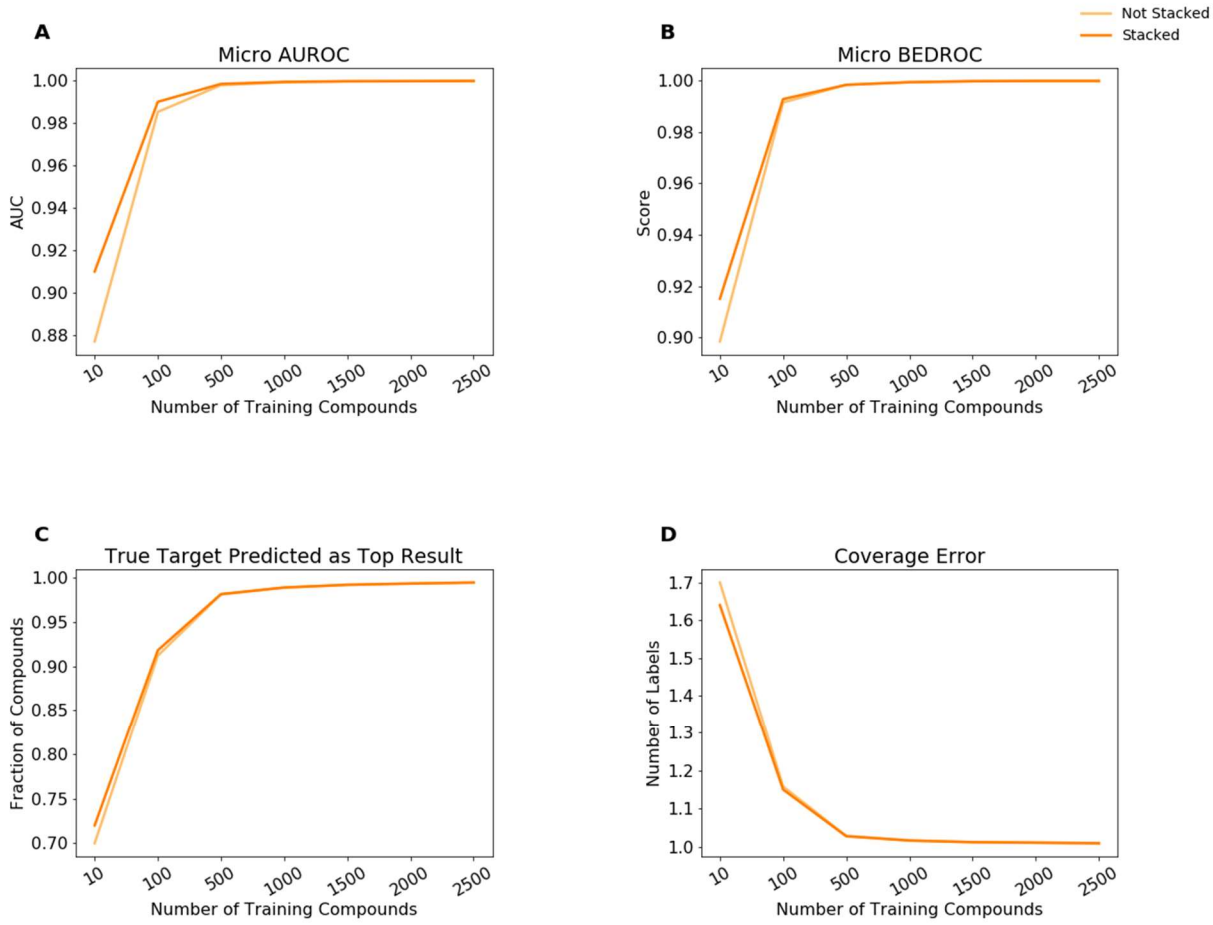

Figure S11. Model performance for stratified 10-fold cross-validation on datasets containing various numbers of compound training records for each protein target label for the KNN_MLP_RF classifier. For a single model, "Not Stacked" indicates that the probability predictions of the listed model were used directly. If more than one model is listed, the mean probabilities for each label were used. "Stacked" indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by **(A)** micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, **(B)** micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic

(BEDROC), **(C)** the fraction of compounds which yielded a true target as the top prediction, and **(D)** coverage error are shown.

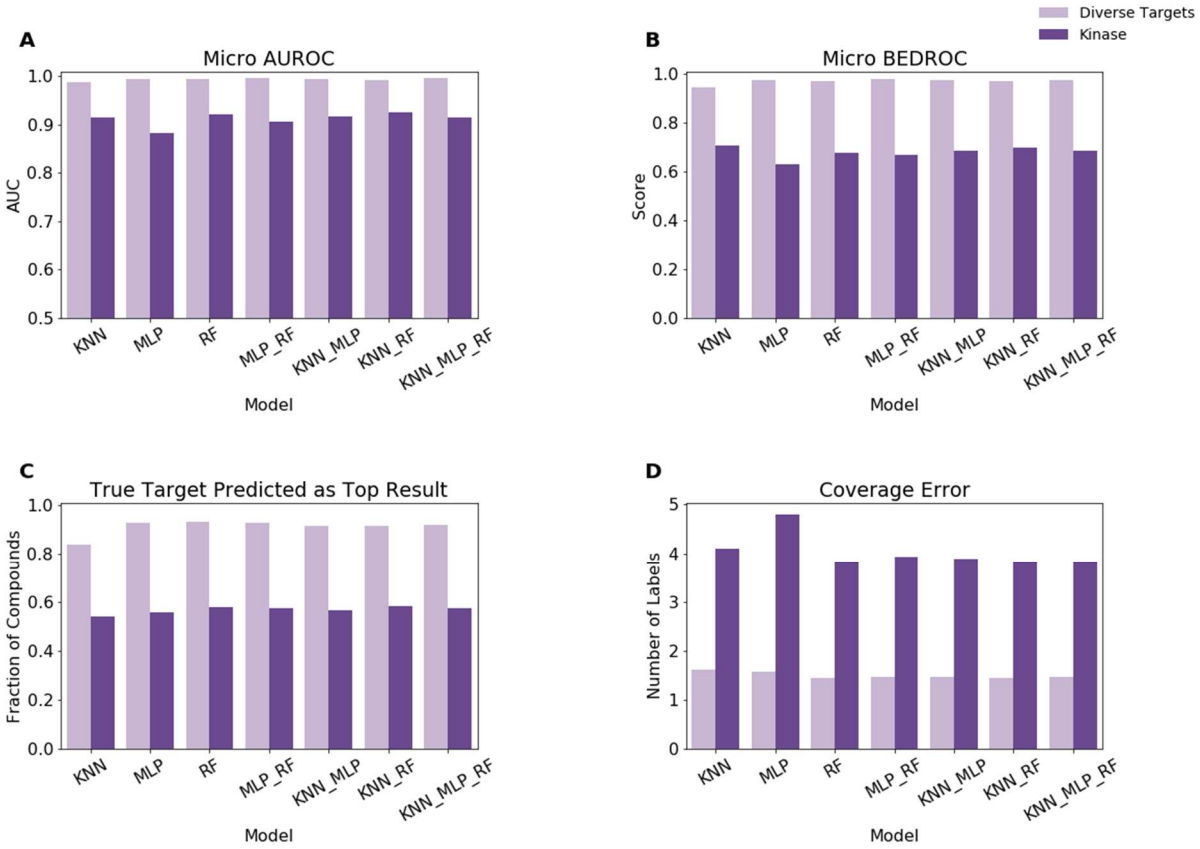

Figure S12. Model performance for stratified 10-fold cross-validation on the diverse target and kinase datasets for stacked classifiers. Model performance as measured by **(A)** micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, **(B)** micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), **(C)** the fraction of compounds which yielded a true target as the top prediction, and **(D)** coverage error are shown.

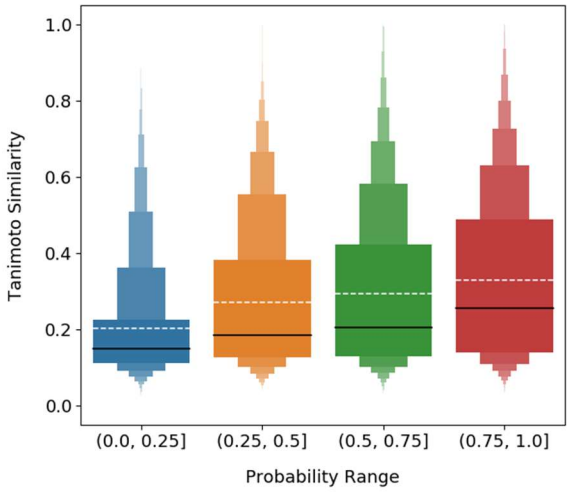

Figure S13. Letter-value plot showing the aggregated pairwise similarity distributions for synthetic test compounds and synthetic training compounds for known positive protein target labels in a cross-validation fold. Similarity distributions were aggregated based on the predicted probability from the KNN_RF stacked classifier for the known protein targets of each synthetic test compound. The solid black line represents the median and the white dashed line the mean. Letter-value plots are similar to box plots, but provide more information about the tails of a distribution. Each box represents a portion of a distribution according to its width shown. The widest box is identical to the interquartile range in a box plot and represents 50% of the data. The next widest boxes, as more than one box now has identical width, comprise 25% of the data. Those boxes are present directly above and below the interquartile range. For each successive box width reduction, the amount of data represented is halved.

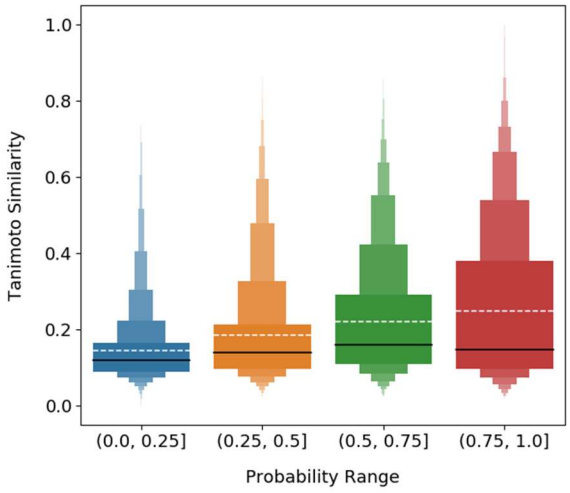

Figure S14. Letter-value plot showing the aggregated pairwise similarity distributions for benchmark natural product compounds and synthetic training compounds for known positive protein target labels. Similarity distributions were aggregated based on the predicted probability from the KNN base classifier for the known protein targets of each natural product. The solid black line represents the median and the white dashed line the mean. Letter-value plots are similar to box plots, but provide more information about the tails of a distribution. Each box represents a portion of a distribution according to its width shown. The widest box is identical to the interquartile range in a box plot and represents 50% of the data. The next widest boxes, as more than one box now has identical width, comprise 25% of the data. Those boxes are present directly above and below the interquartile range. For each successive box width reduction, the amount of data represented is halved.

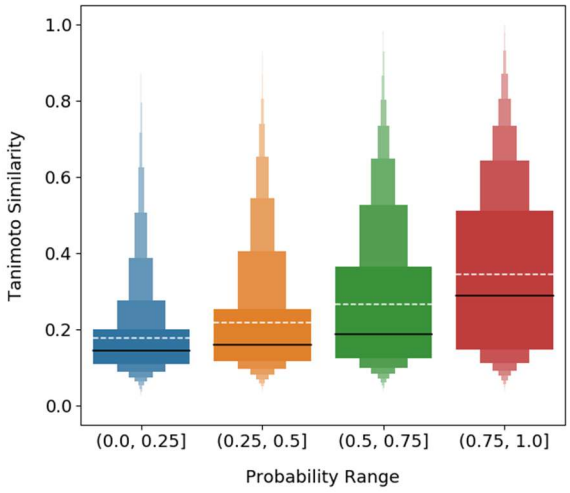

Figure S15. Letter-value plot showing the aggregated pairwise similarity distributions for synthetic test compounds and synthetic training compounds for known positive protein target labels in a cross-validation fold. Similarity distributions were aggregated based on the predicted probability from the KNN base classifier for the known protein targets of each synthetic test compound. The solid black line represents the median and the white dashed line the mean. Letter-value plots are similar to box plots, but provide more information about the tails of a distribution. Each box represents a portion of a distribution according to its width shown. The widest box is identical to the interquartile range in a box plot and represents 50% of the data. The next widest boxes, as more than one box now has identical width, comprise 25% of the data. Those boxes are present directly above and below the interquartile range. For each successive box width reduction, the amount of data represented is halved.

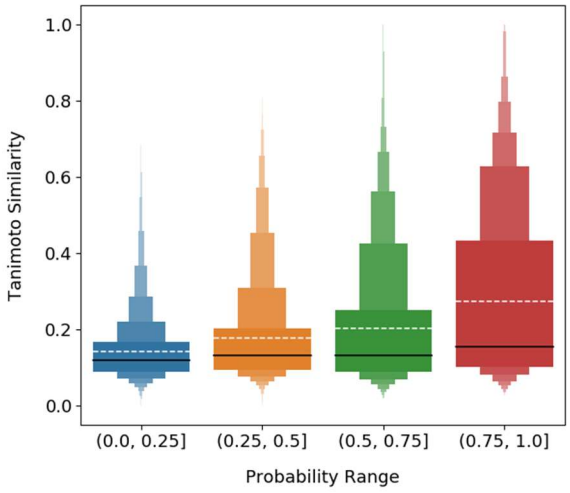

Figure S16. Letter-value plot showing the aggregated pairwise similarity distributions for benchmark natural product compounds and synthetic training compounds for known positive protein target labels. Similarity distributions were aggregated based on the predicted probability from the RF base classifier for the known protein targets of each natural product. The solid black line represents the median and the white dashed line the mean. Letter-value plots are similar to box plots, but provide more information about the tails of a distribution. Each box represents a portion of a distribution according to its width shown. The widest box is identical to the interquartile range in a box plot and represents 50% of the data. The next widest boxes, as more than one box now has identical width, comprise 25% of the data. Those boxes are present directly above and below the interquartile range. For each successive box width reduction, the amount of data represented is halved.

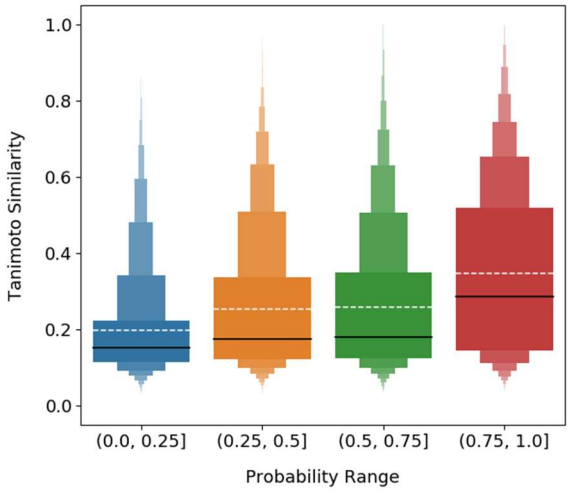

Figure S17. Letter-value plot showing the aggregated pairwise similarity distributions for synthetic test compounds and synthetic training compounds for known positive protein target labels in a cross-validation fold. Similarity distributions were aggregated based on the predicted probability from the RF base classifier for the known protein targets of each synthetic test compound. The solid black line represents the median and the white dashed line the mean. Letter-value plots are similar to box plots, but provide more information about the tails of a distribution. Each box represents a portion of a distribution according to its width shown. The widest box is identical to the interquartile range in a box plot and represents 50% of the data. The next widest boxes, as more than one box now has identical width, comprise 25% of the data. Those boxes are present directly above and below the interquartile range. For each successive box width reduction, the amount of data represented is halved.