Supporting Information:

Virtual Screening with Gnina 1.0

Jocelyn Sunseri and David R. Koes*

Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260

E-mail: dkoes@pitt.edu

Per-Target Confidence Intervals

In addition to evaluating the overall performance of the method (Table S2) it is also insightful to look at per-target performance. There has been some debate about how to statistically assess claims made about machine learning model performance for drug discovery tasks like virtual screening.[1,2] The norm has been, as in much of the other recent machine learning literature that reports model performance on community benchmarks, to avoid statistically evaluating them at all. Our approach is to use bootstrapping to compute 95% confidence intervals of the various metrics and the non-parametric Mann-Whitney U test to evaluate if the bootstrapped distributions of metrics differ significantly. For each target, we perform 1000 rounds of bootstrapping using stratified sampling with replacement to address the class imbalance issue. The resulting distributions are used to calculate the 5%-95% confidence interval and are compared with Mann-Whitney U to get a p-value. Confidence intervals are illustrated in Figures S1 – S10.

Table S1: Analysis of LIT-PCBA assays.

Target	Primary PubChem Assay ID	Confirmatory PubChem Assay ID	Publication	Notes
ADRB2	492947	588463	—	**Cell based assay**
ALDH1	1030	493210	—	
ESR1_ago	743075	743077	Huang et al.[3]	**Cell based assay**
ESR1_ant	743080	743091	Hart et al.[4]	**Cell based assay**
FEN1	588795	488816	Dorjsuren et al.[5]	Paper describes assay, not results.
GBA	2101	2590	Goldin et al.[6]	Both antagonists and agonists
IDH1	602179	624002	Davis et al.[7]	Most compounds selective for R132H, not wild type
KAT2A	504327	588347	—	
MAPK1	995	1742	Marugan et al.[8]	**Cell-based assay**. At least one active inhibits EGFR, not MAPK1
MTORC1	493208	651784, 651789, 651791, 651793	Hoffman et al.[9]	**Cell based assay**. Paper describes assay, not results
OPRK1	1777	2133,2136	Frankowski et al.[10]	**Cell based assay**. Both antagonists and agonists
PKM2	1631	1751	Boxer et al.[11]	Both antagonists and agonists
PPARG	743094	743140	—	**Cell based assay**
TP53	651631	720552	Witt et al.[12]	**Cell based assay**
VDR	602199	602200, 602201, 602202, 602204	Nandhikonda et al.[13]	Actives reported in paper do not match confirmatory screen compounds.

Table S2: Mean AUCs, NEF1% and EF1% values on DUD-E and LIT-PCBA. The best CNN model value for each column is shown in bold. Models whose distributions of per-benchmark metrics are not statistically dissimilar to the model in bold (as computed with a Mann-Whitney U rank test, p-value > 0.05) are shown in italic. Note that RFScore-VS was trained on DUD-E.

Model	DUD-E			LIT-PCBA		
	AUC	NEF1%	EF1%	AUC	NEF1%	EF1%
RFScore-4	0.678	0.0911	5.5	*0.571*	*0.0224*	*1.67*
RFScore-VS	0.944	0.798	49.3	*0.553*	0.0214	1.75
Vina	0.725	0.167	9.93	*0.57*	*0.0247*	*1.71*
Vinardo	0.748	0.204	12.4	*0.58*	0.0235	1.7
General (Pose)	0.712	0.208	12.9	0.509	*0.0251*	*2.01*
General (Affinity)	0.749	0.229	13.9	*0.594*	*0.0445*	*3.7*
Dense (Pose)	*0.767*	**0.34**	**20.9**	0.527	*0.0415*	*3.45*
Dense (Affinity)	**0.793**	*0.303*	*18.8*	**0.616**	*0.0553*	*4.45*
Default (Pose)	0.753	*0.294*	*18.1*	0.518	*0.0292*	*2.41*
Default (Affinity)	*0.781*	*0.29*	*17.7*	*0.61*	**0.0561**	**4.64**

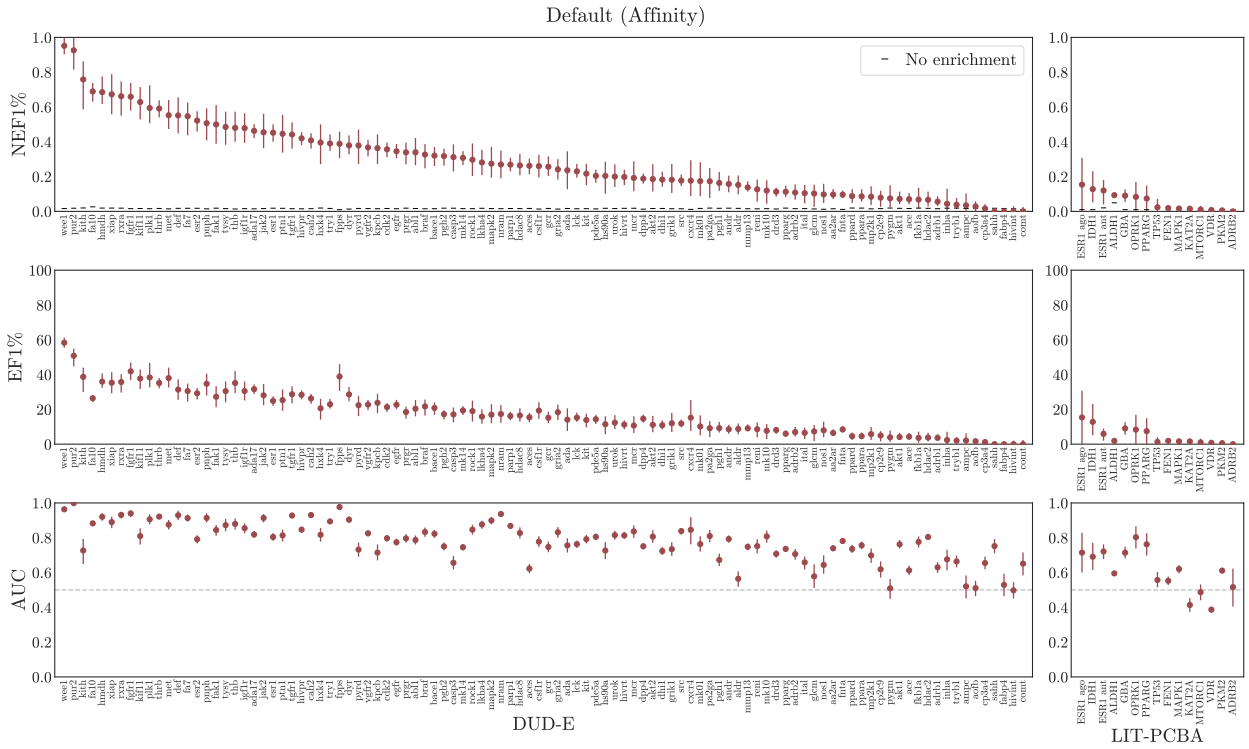


Figure S1: Per-target 5%-95% confidence intervals for various metrics for Default (Affinity).

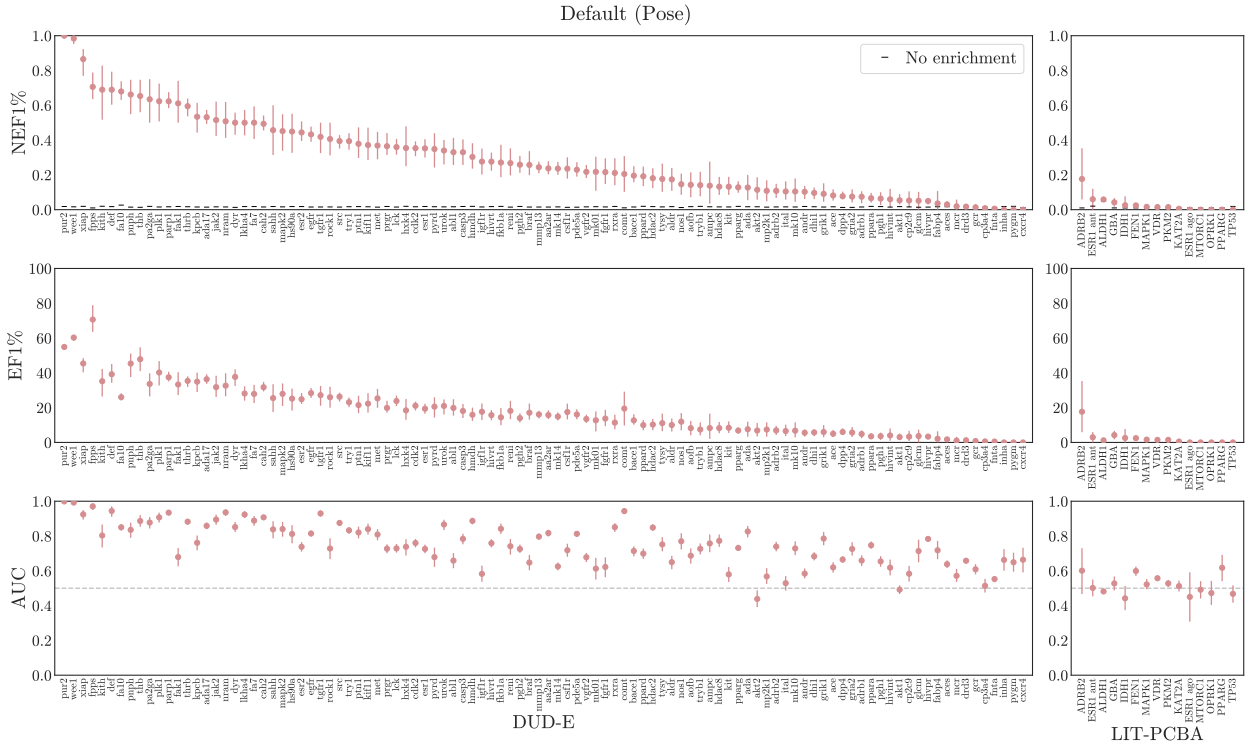


Figure S2: Per-target 5%-95% confidence intervals for various metrics for Default (Pose).

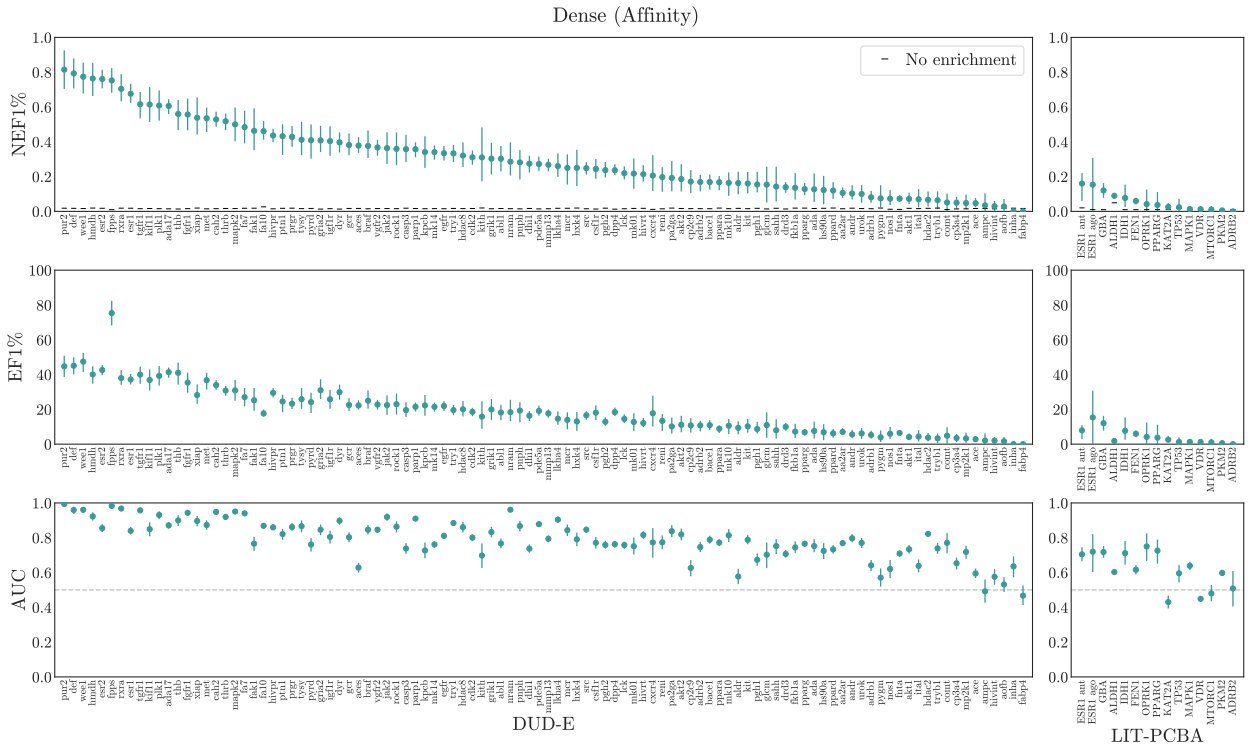


Figure S3: Per-target 5%-95% confidence intervals for various metrics for Dense (Affinity).

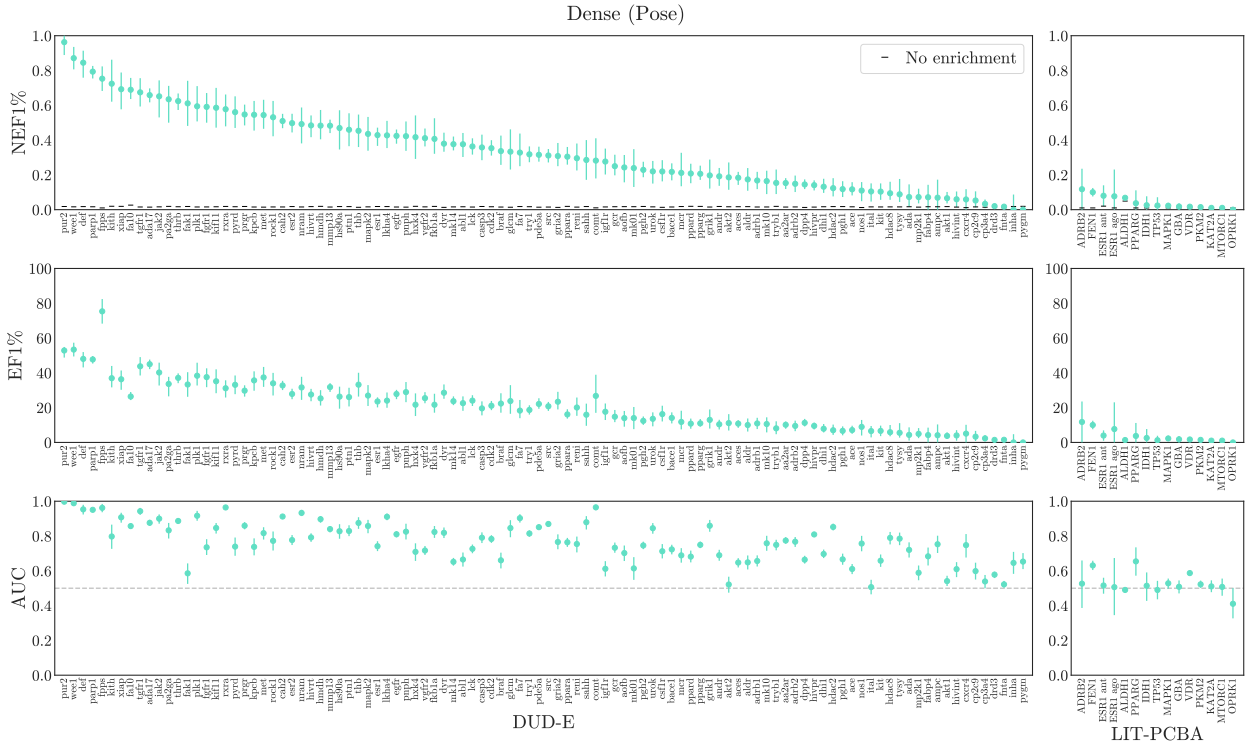


Figure S4: Per-target 5%-95% confidence intervals for various metrics for Dense (Pose).

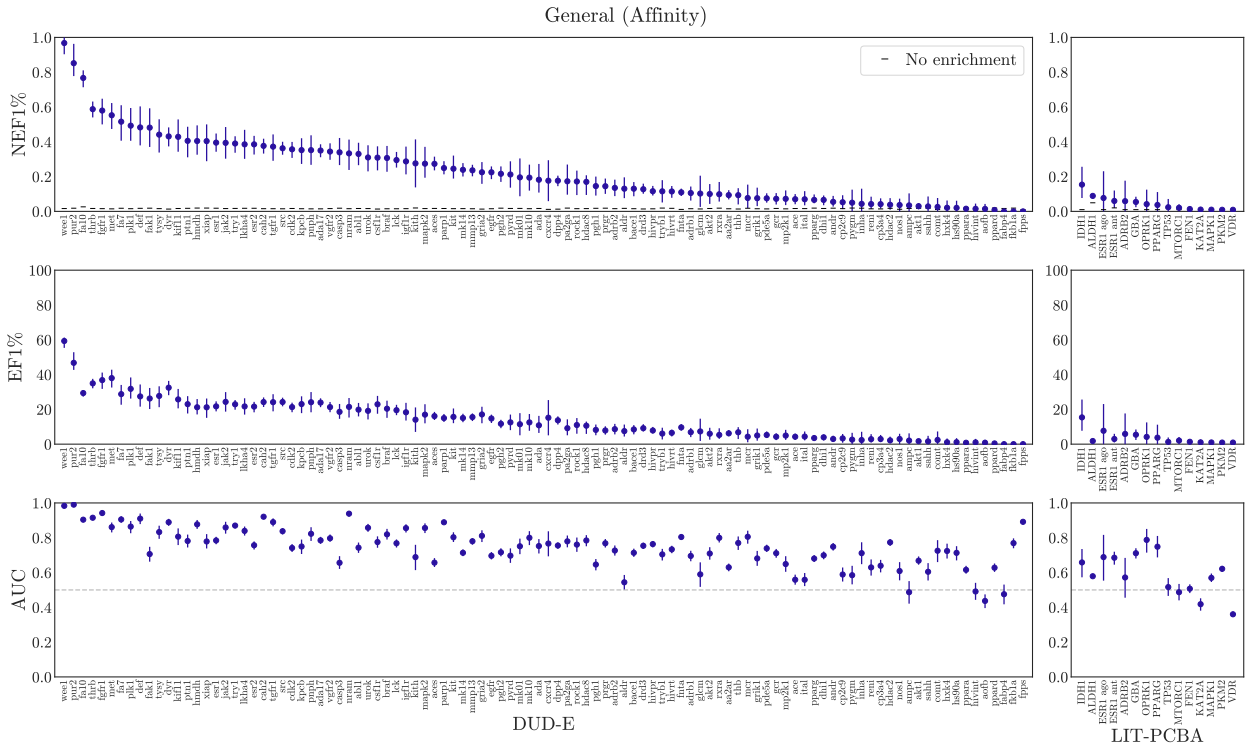


Figure S5: Per-target 5%-95% confidence intervals for various metrics for General (Affinity).

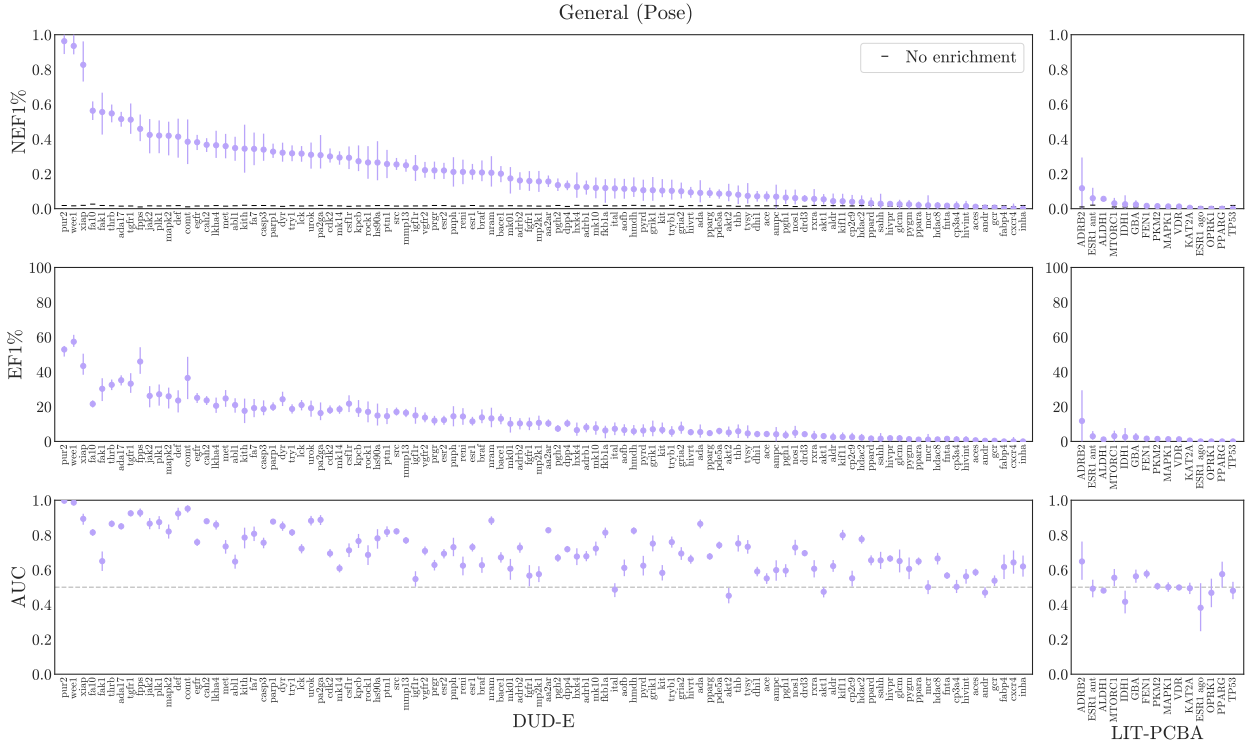


Figure S6: Per-target 5%-95% confidence intervals for various metrics for General (Pose).

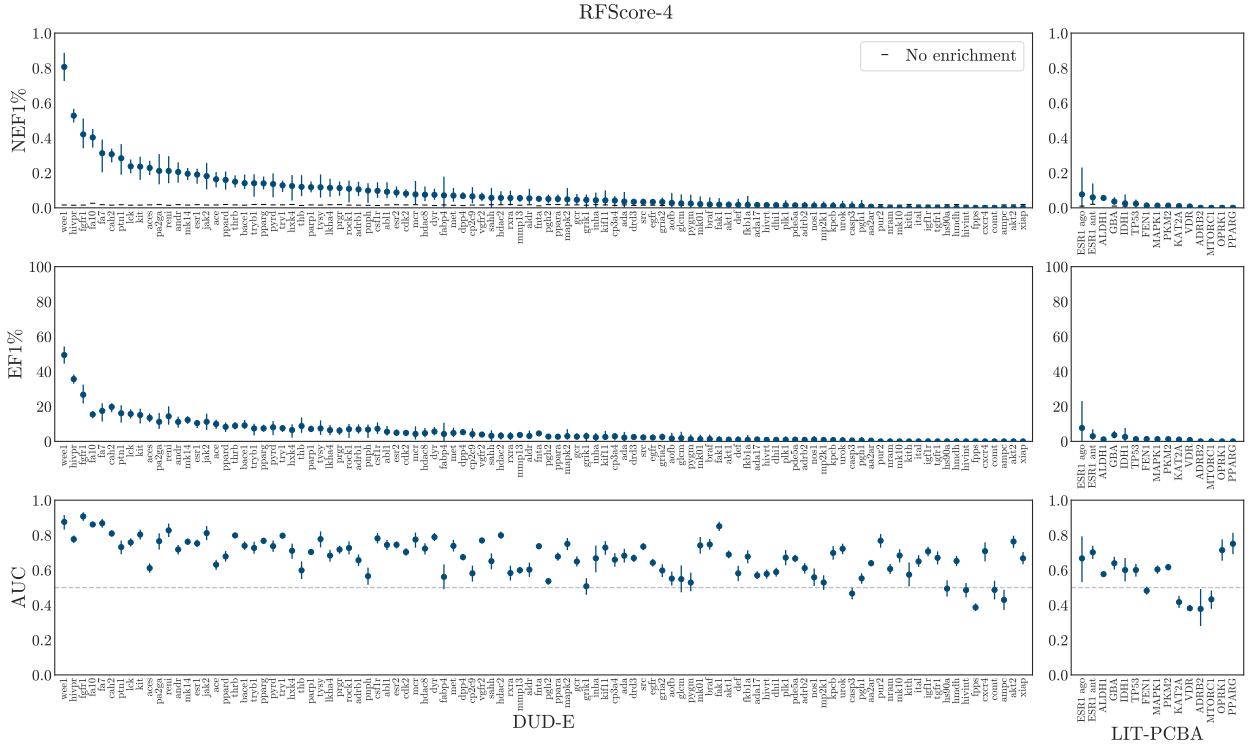


Figure S7: Per-target 5%-95% confidence intervals for various metrics for RFScore-4.

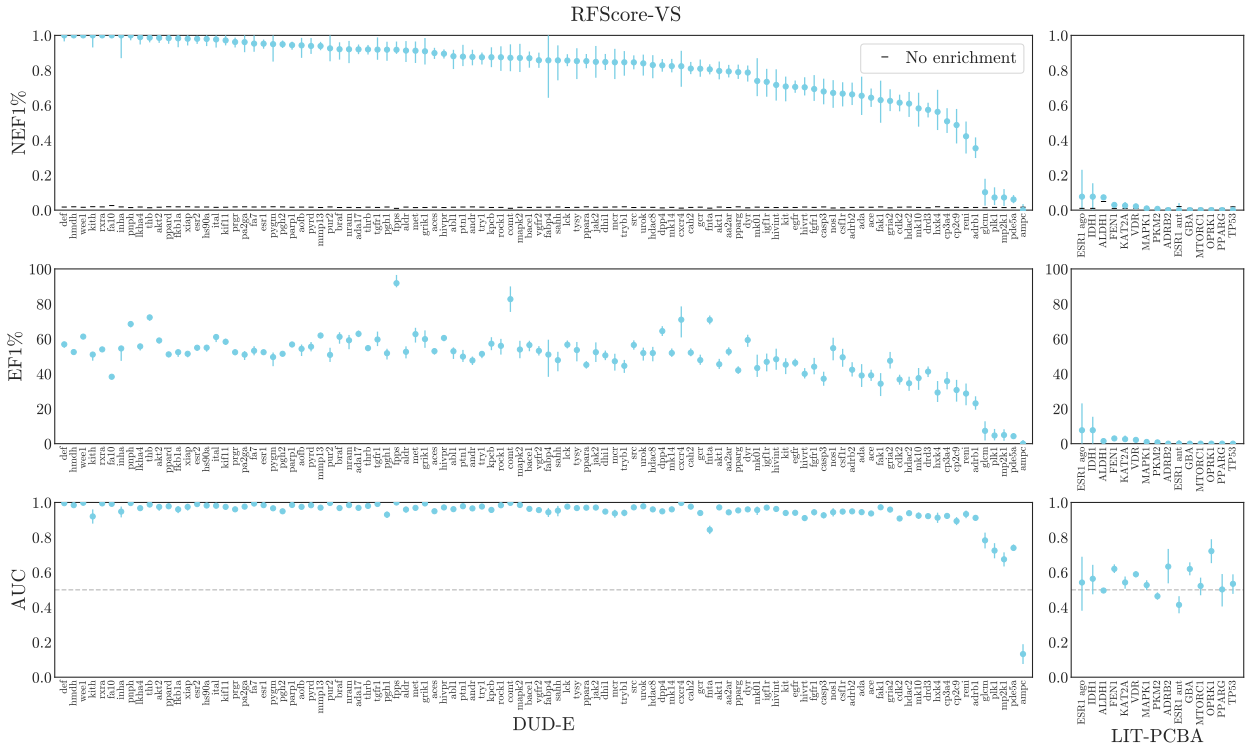


Figure S8: Per-target 5%-95% confidence intervals for various metrics for RFScore-VS.

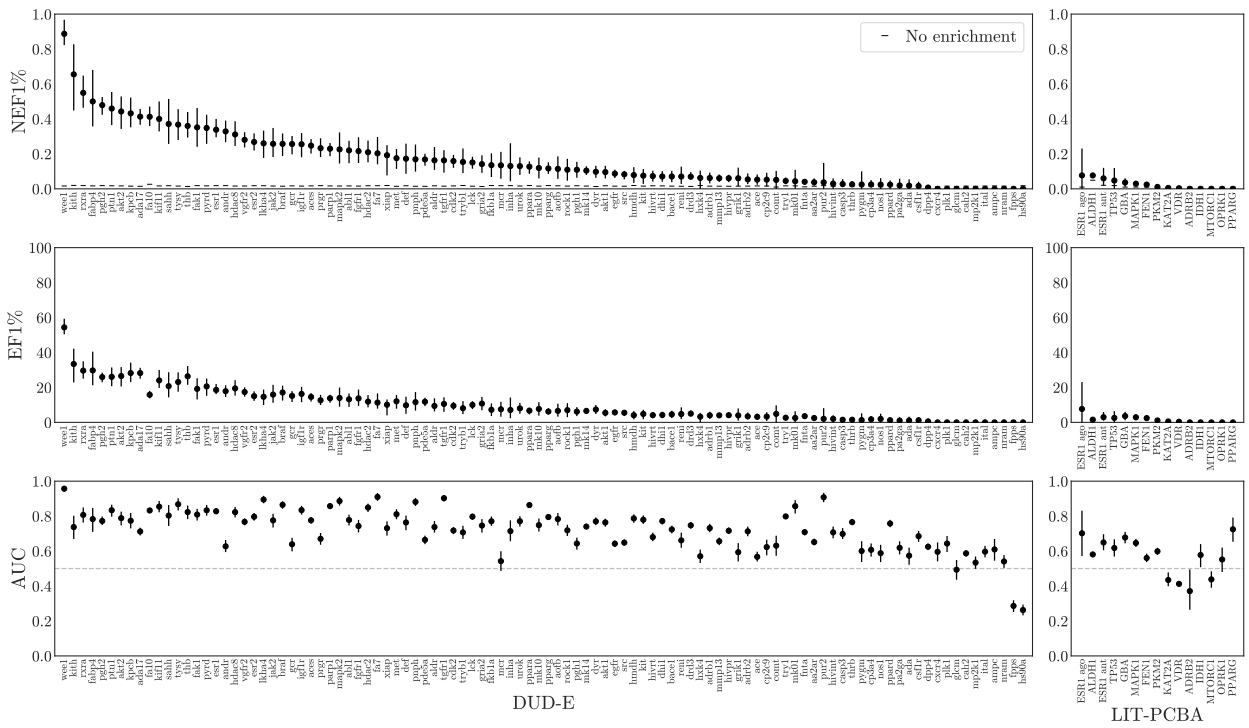


Figure S9: Per-target 5%-95% confidence intervals for various metrics for Vina.

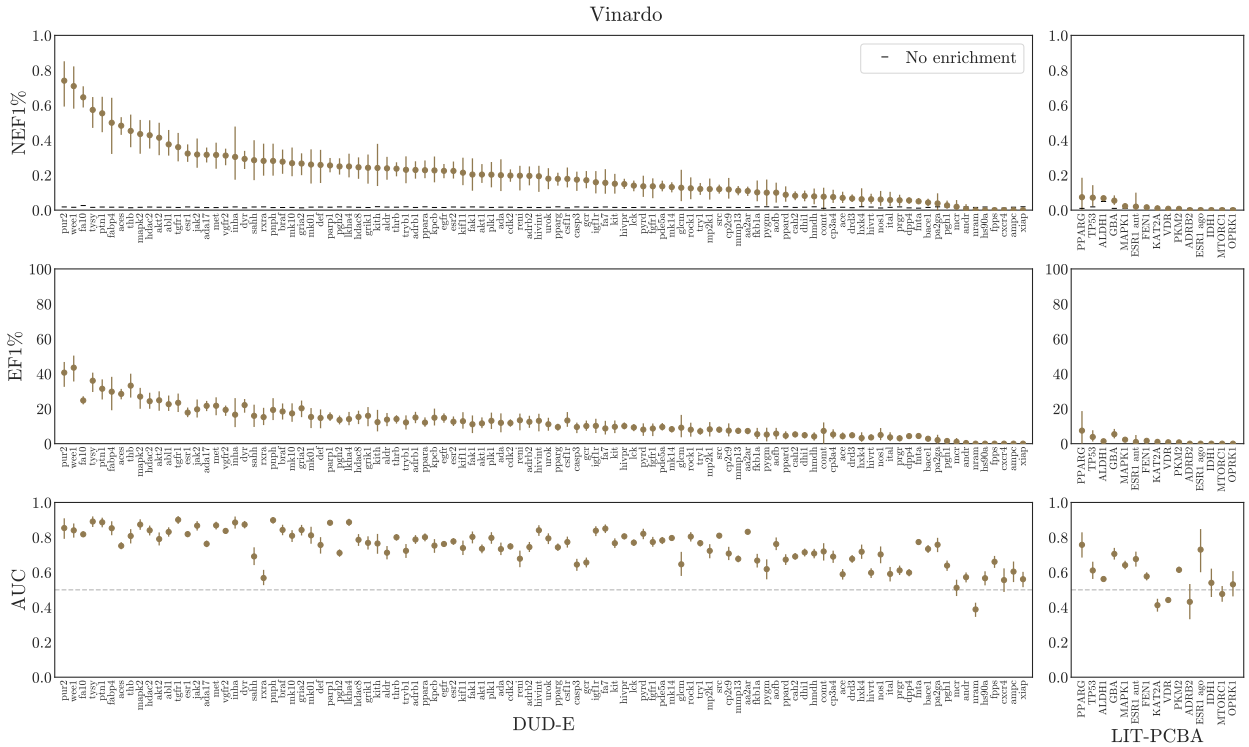


Figure S10: Per-target 5%-95% confidence intervals for various metrics for Vinardo.

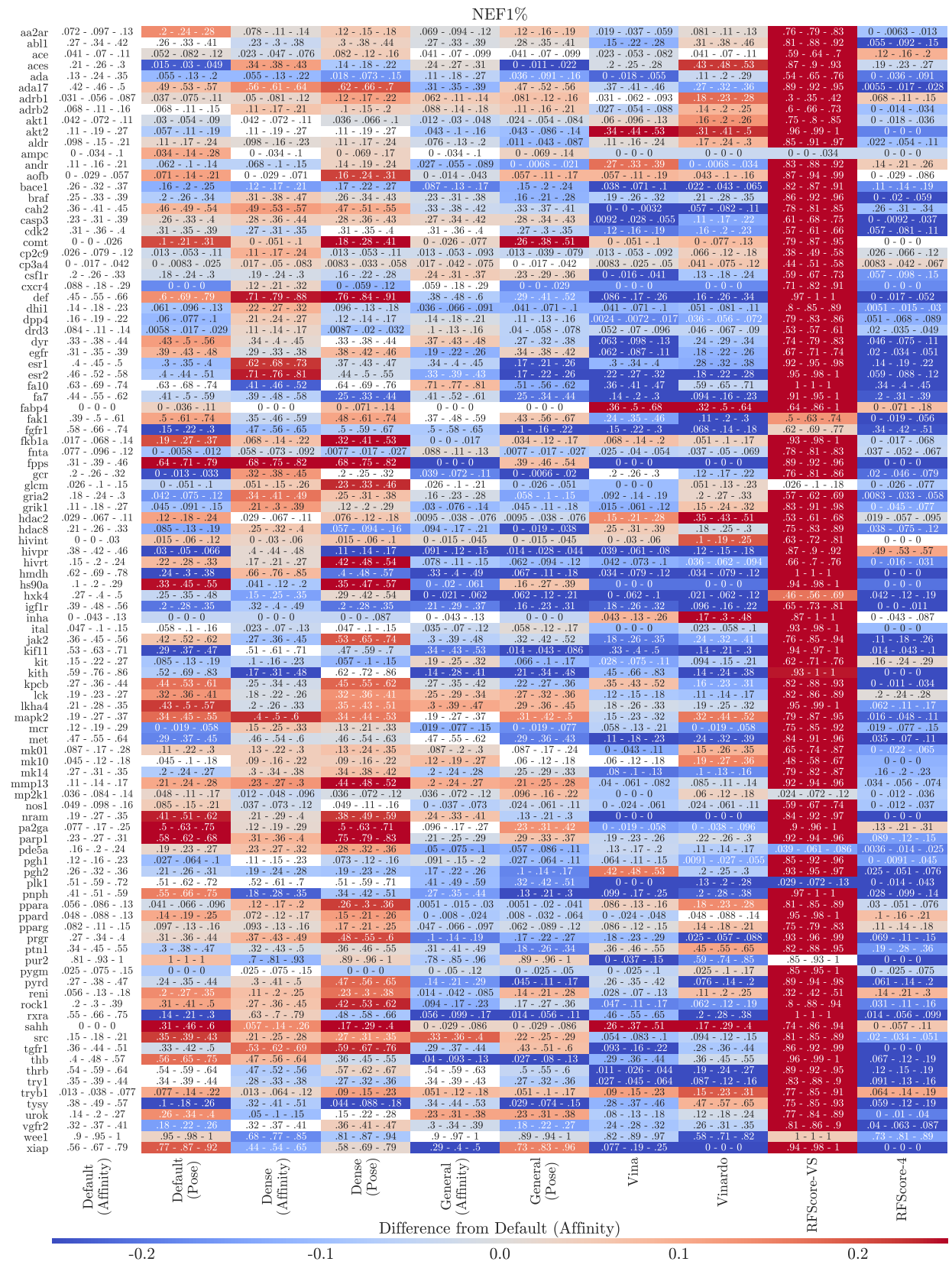


Figure S11: NEF1% performance, with 5%–95% confidence intervals, shaded by difference with Default (Affinity). White boxes are not significant ($p > 0.0005$). Red means the method is better than Default (Affinity) and blue worse.

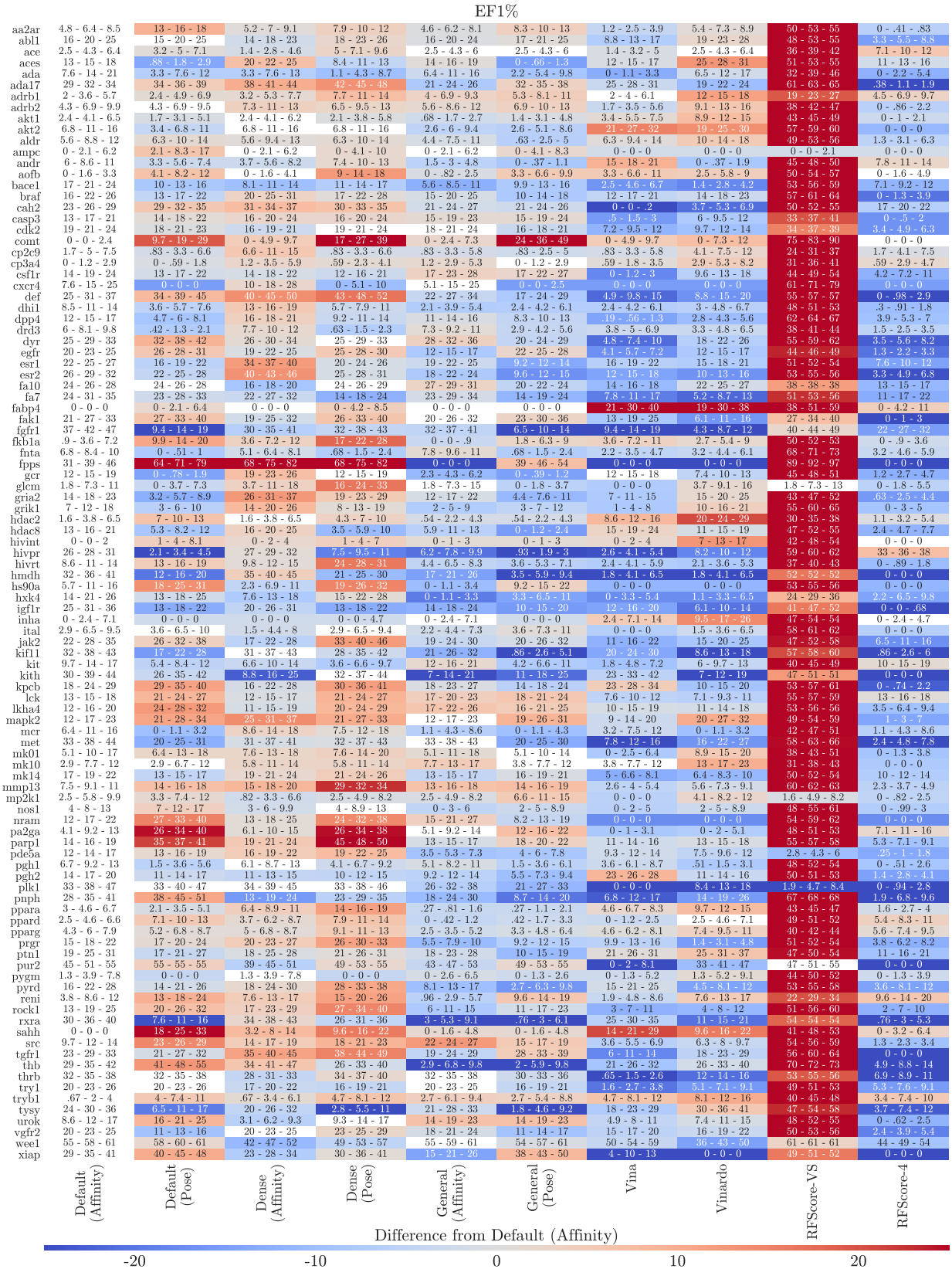


Figure S12: DUD-E EF1% performance, with 5%–95% confidence intervals, shaded by difference with Default (Affinity). White boxes are not significant ($p > 0.0005$). Red means the method is better than Default (Affinity) and blue worse.

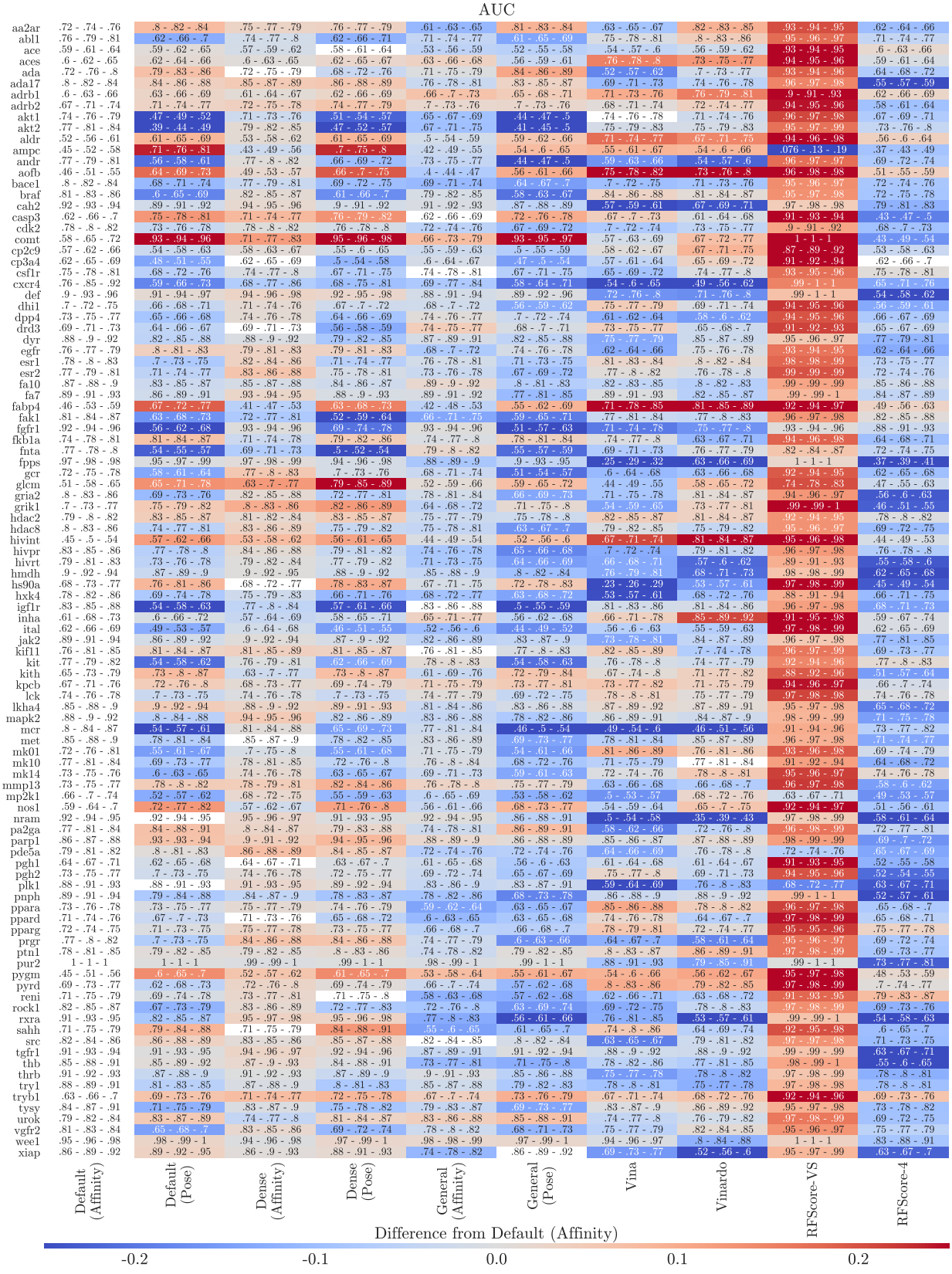


Figure S13: DUD-E AUC performance, with 5%–95% confidence intervals, shaded by difference with Default (Affinity). White boxes are not significant ($p > 0.0005$). Red means the method is better than Default (Affinity) and blue worse.

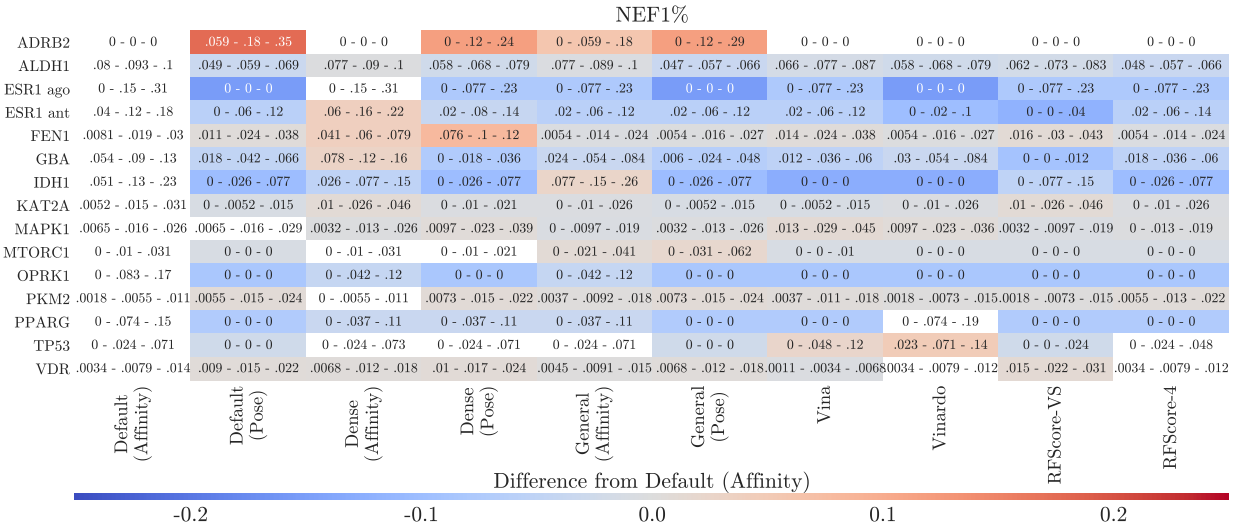


Figure S14: LIT-PCBA NEF1% performance, with 5%–95% confidence intervals, shaded by difference with Default (Affinity). White boxes are not significant ($p > 0.0005$). Red means the method is better than Default (Affinity) and blue worse.

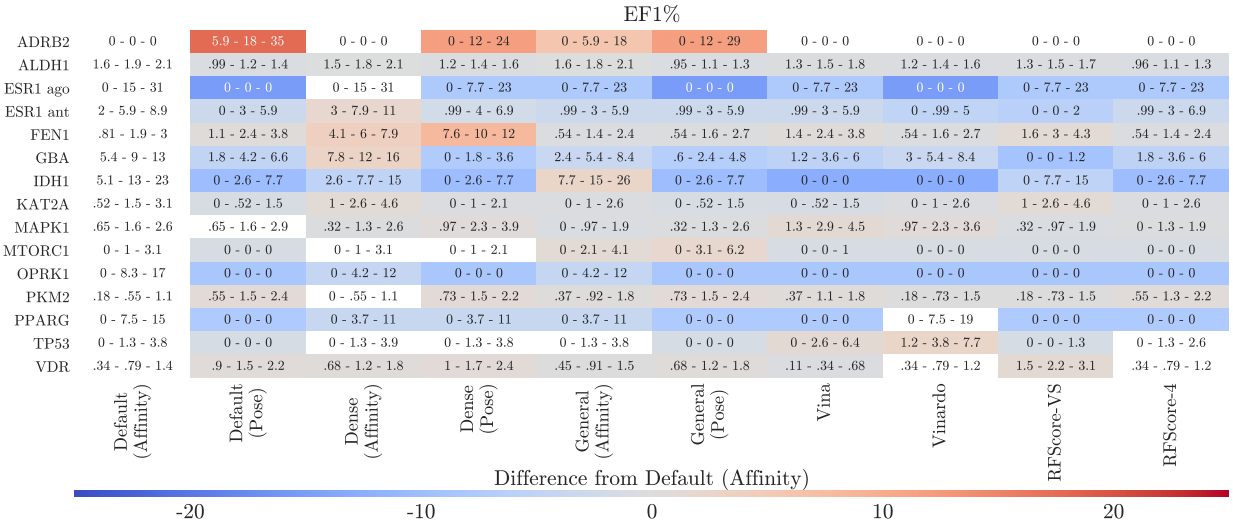


Figure S15: LIT-PCBA EF1% performance, with 5%–95% confidence intervals, shaded by difference with Default (Affinity). White boxes are not significant ($p > 0.0005$). Red means the method is better than Default (Affinity) and blue worse.

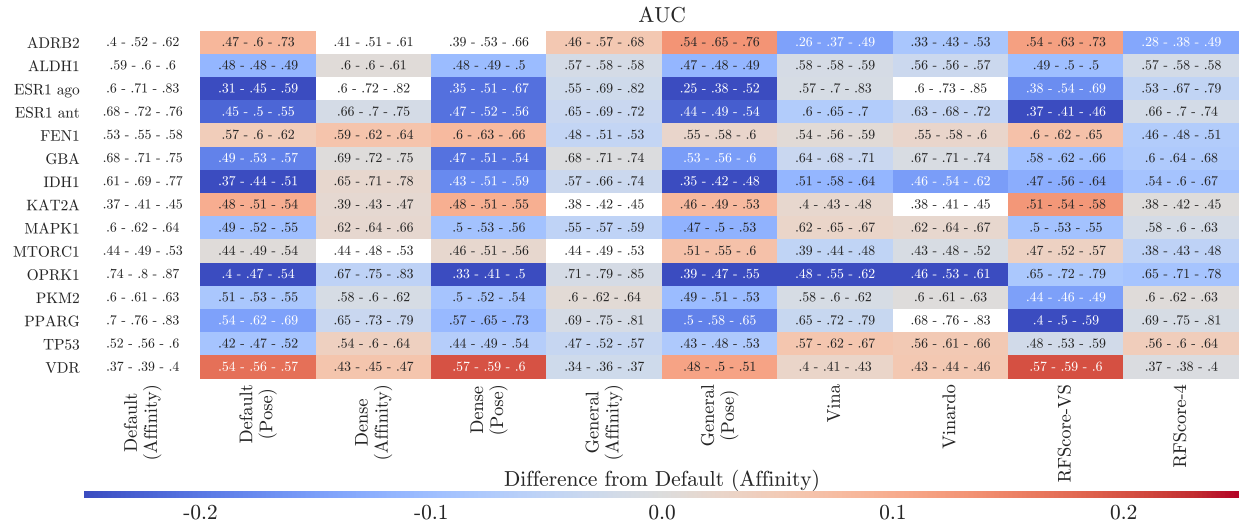
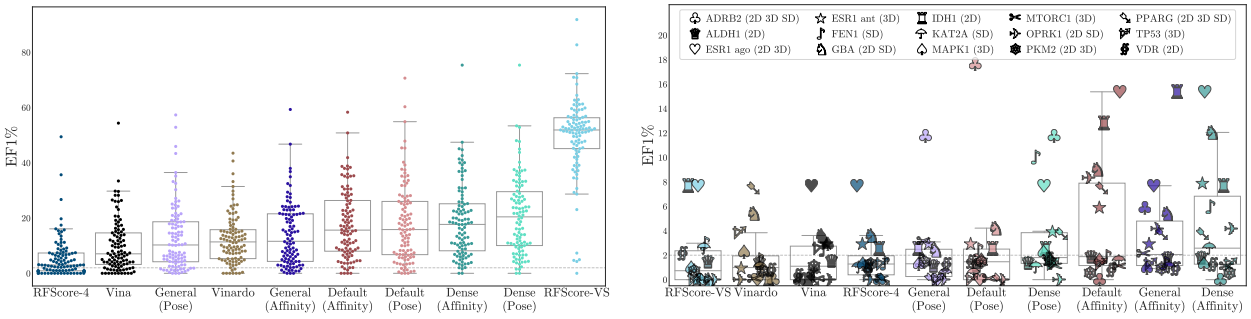


Figure S16: LIT-PCBA AUC performance, with 5%–95% confidence intervals, shaded by difference with Default (Affinity). White boxes are not significant ($p > 0.0005$). Red means the method is better than Default (Affinity) and blue worse.



(a) (b)

Figure S17: Assessment of virtual screening performance using EF1% on (a) DUD-E and (b) LIT-PCBA. The x-axis is sorted in order of increasing median performance. Each data point is the performance of the method on a single target. Note the different y-axis scales: DUD-E targets demonstrate significantly better enrichment. LIT-PCBA targets are assigned a unique symbol; the legend notes which method achieved an EF1%> 2 in Tran-Nguyen et al.[14], where '1D' signifies fingerprints, '2D' signifies ligand shape shape, and '3D' signifies docking with Surflex-Dock. Note that while EF1% $= 1$ is equivalent to random performance, we have annotated EF1% $= 2$ since that was the minimum enrichment required for inclusion in LIT-PCBA.[14]

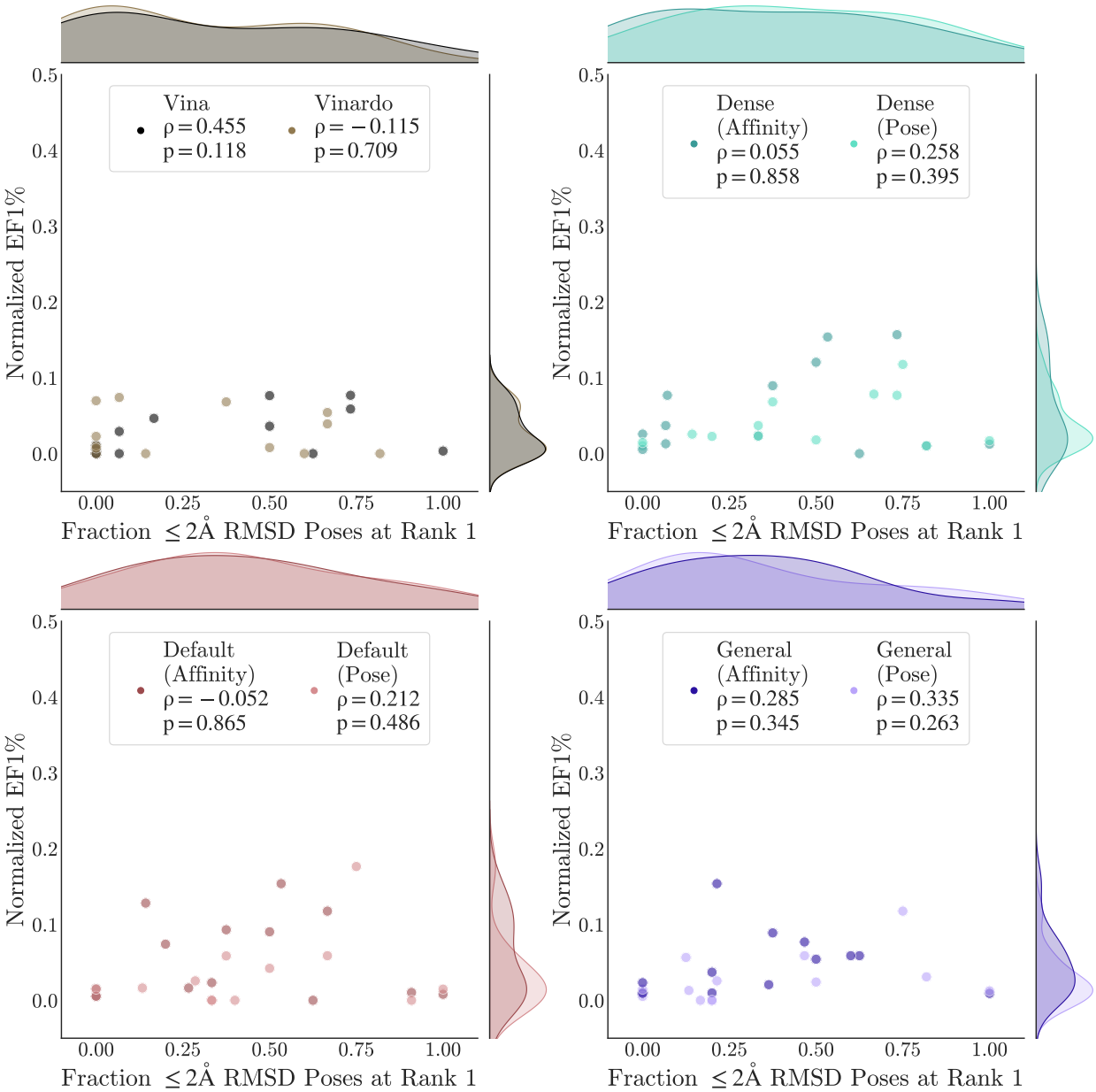


Figure S18: Correlation between normalized EF1% for the virtual screening task and the fraction of compounds with a good pose ranked first in the pose prediction task. The Spearman correlation and its p-value is shown (there are no statistically significant correlations).

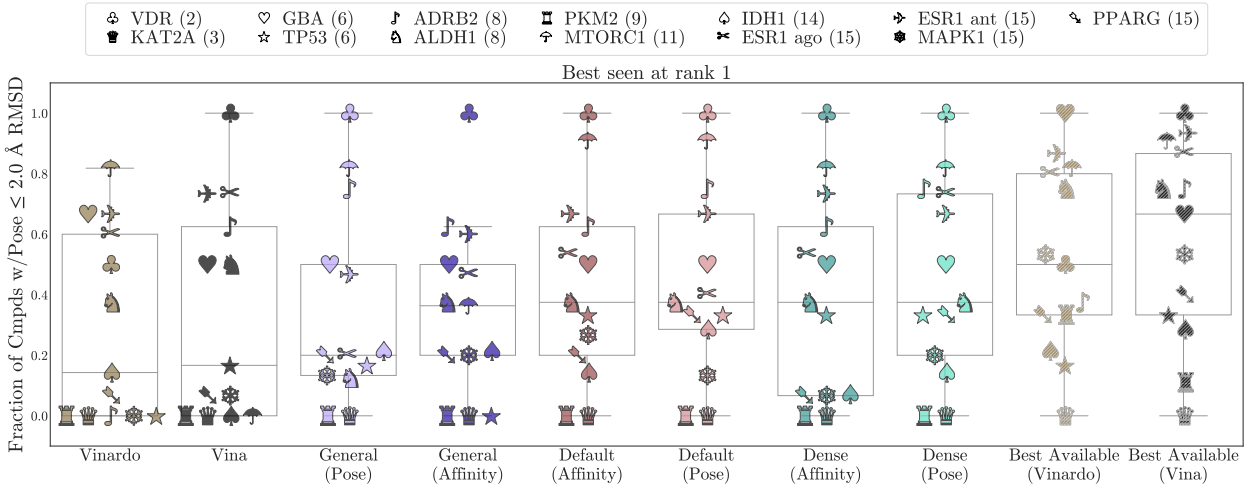


Figure S19: Boxplots showing the fraction of compounds, per-target, that have a pose $\leq$ 2Å RMSD from the provided crystal pose ranked first by each method when cross-docked to non-cognate structures. The two plots with hatched markers show the fraction of compounds for which such a pose was *sampled*. Numbers in the legend indicate the number of reference receptors provided for each target.

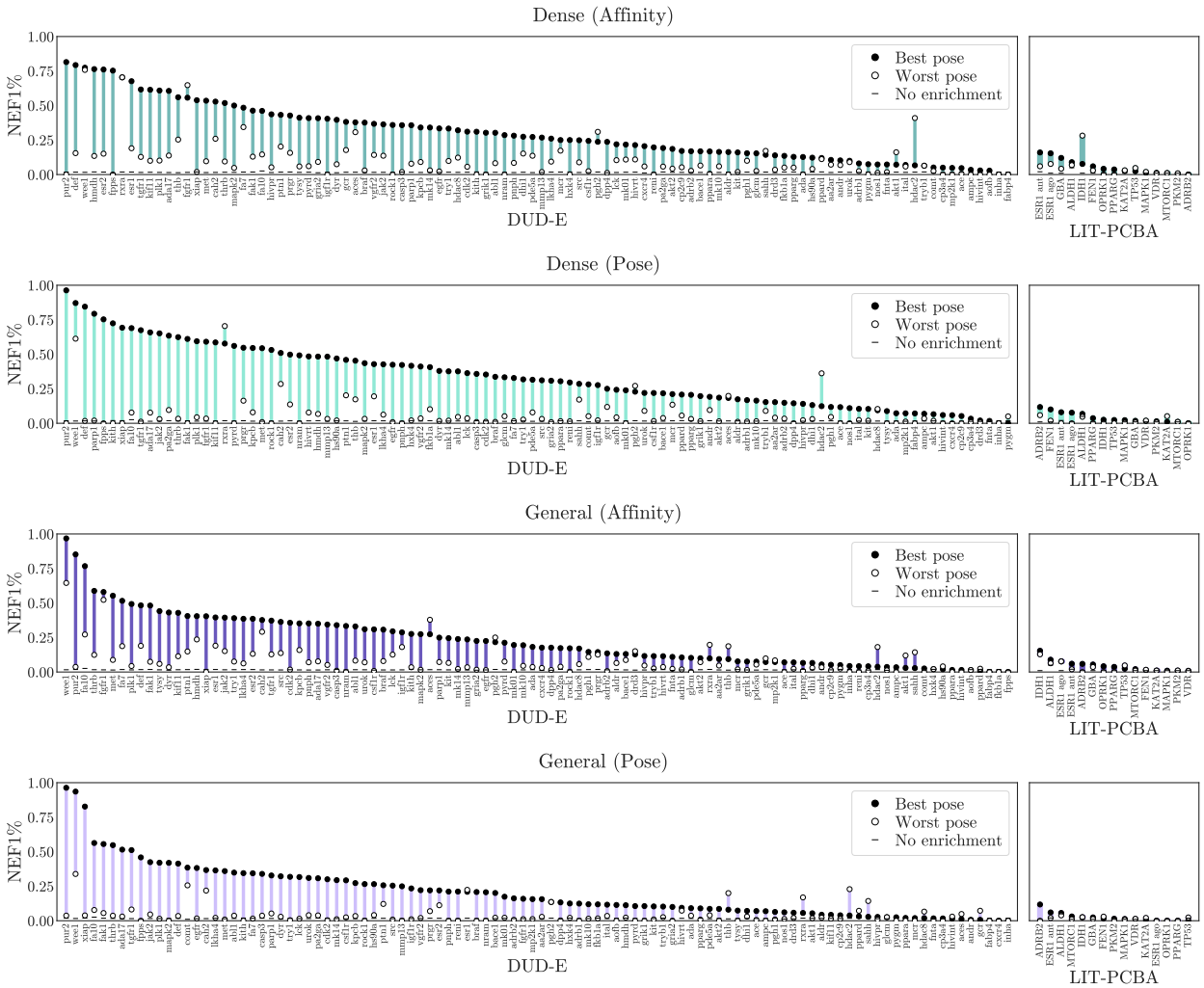


Figure S20: Pose sensitivity assessment of CNN models. For each target the difference in NEF1% when basing the score on the top-ranked versus the bottom-ranked pose is shown.

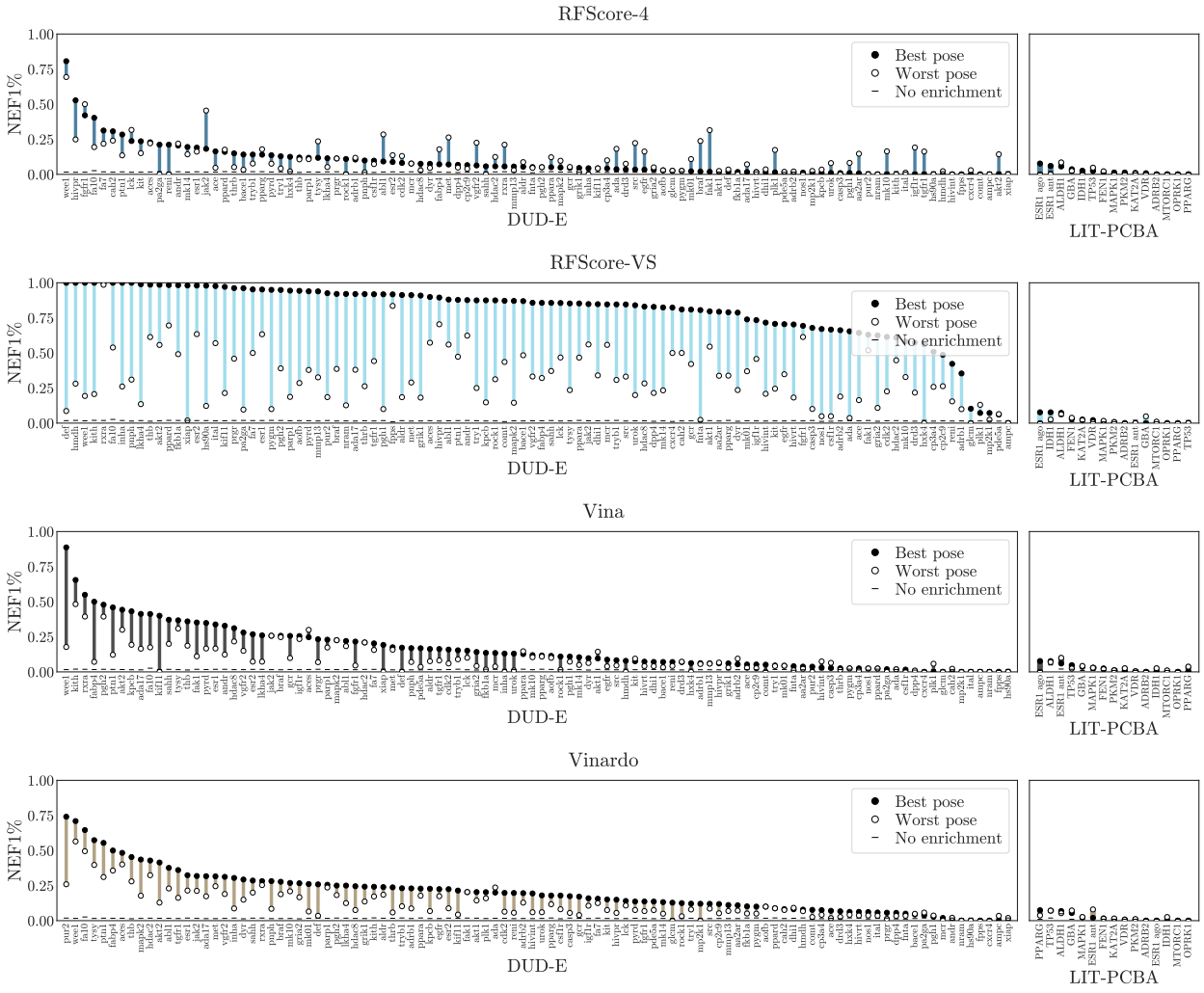


Figure S21: Pose sensitivity assessment of non CNN models. For each target the difference in NEF1% when basing the score on the top-ranked versus the bottom-ranked pose is shown. Note that RFScore-VS was trained on the DUD-E.

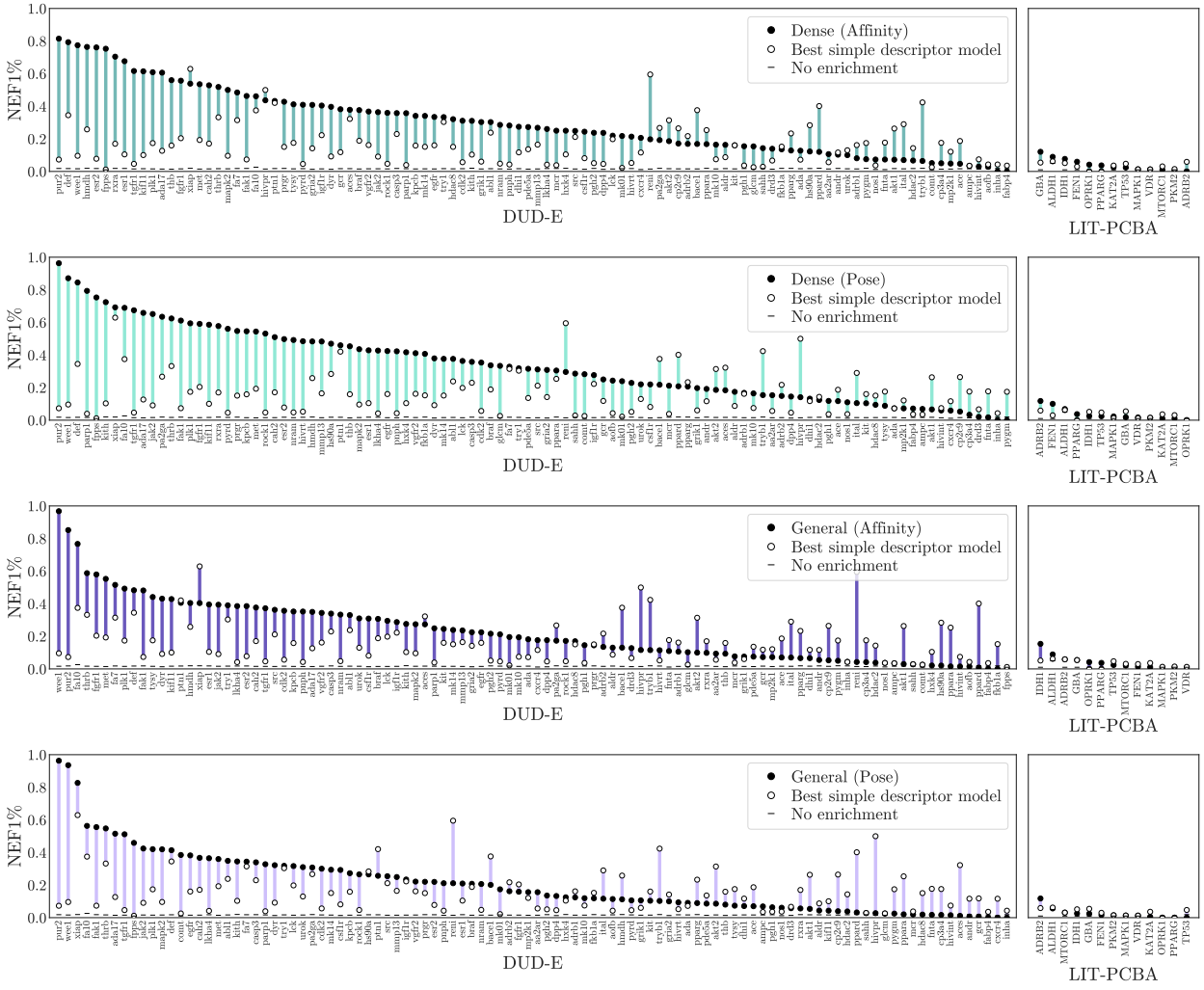


Figure S22: Comparison of CNN models to simple descriptor models. The best achievable performance of a model trained on simple molecular descriptors to predict binding affinities is shown. The x-axes are sorted independently by the NEF1% of the best scoring pose.

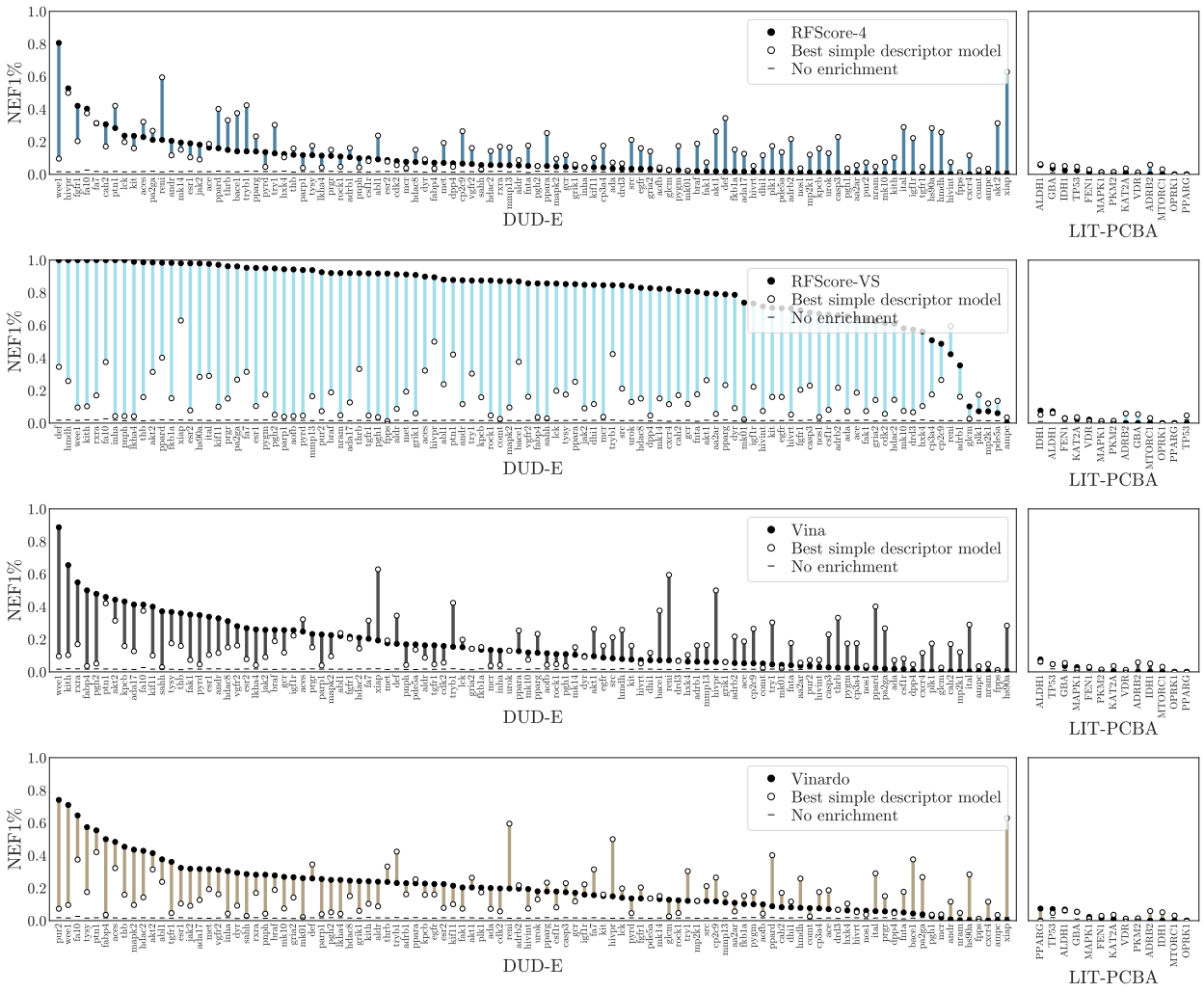


Figure S23: Comparison of CNN models to simple descriptor models. The best achievable performance of a model trained on simple molecular descriptors to predict binding affinities is shown. The x-axes are sorted independently by the NEF1% of the best scoring pose.

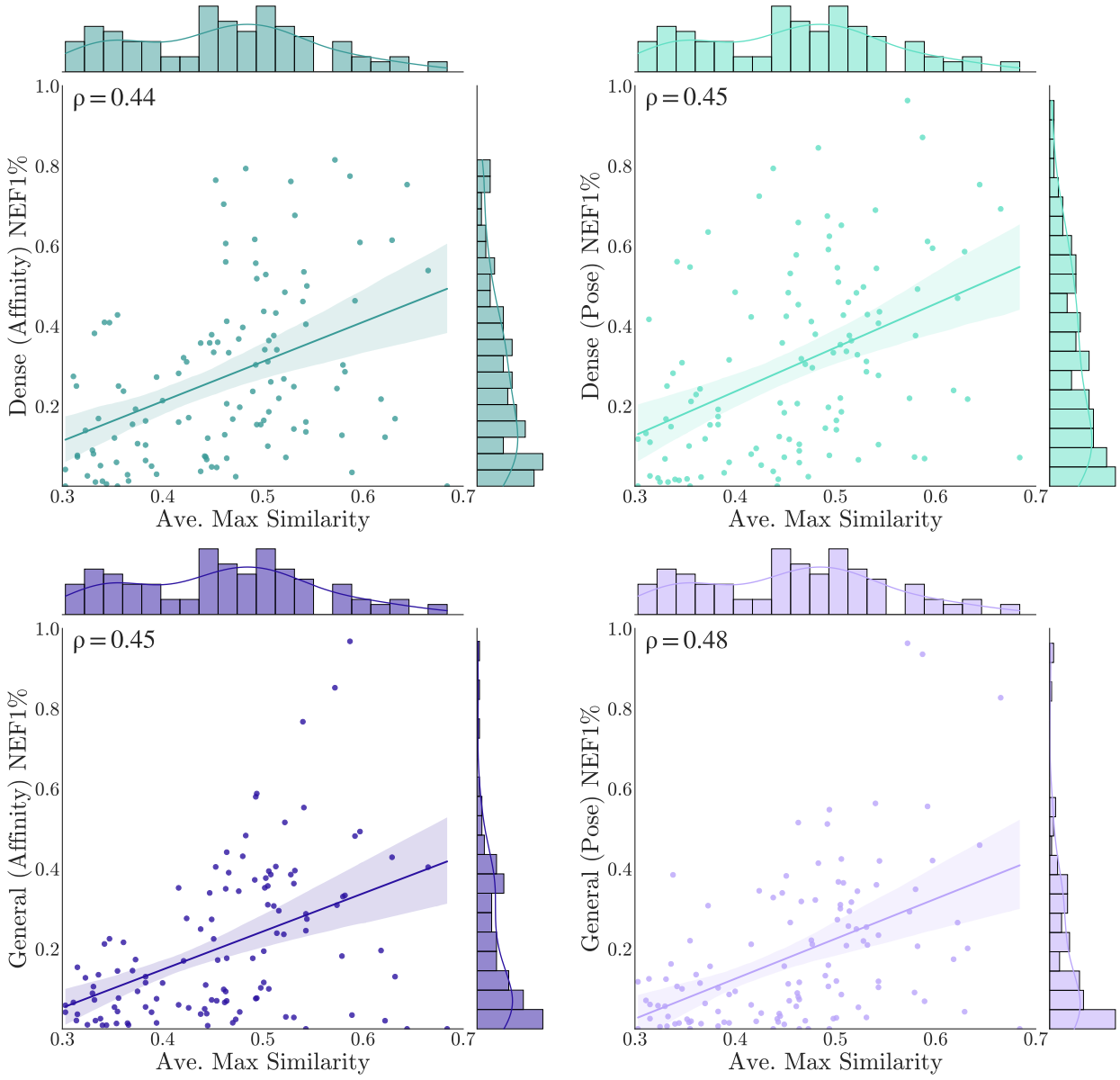


Figure S24: Correlation between similarity with training set and early enrichment performance for the CNN models. For each benchmark, the average of the maximum similarity between active compounds and the PDBbind General set is computed using the Tanimoto coefficient of ECFP4 fingerprints. Only actives are considered since the training set does not include any inactive compounds.

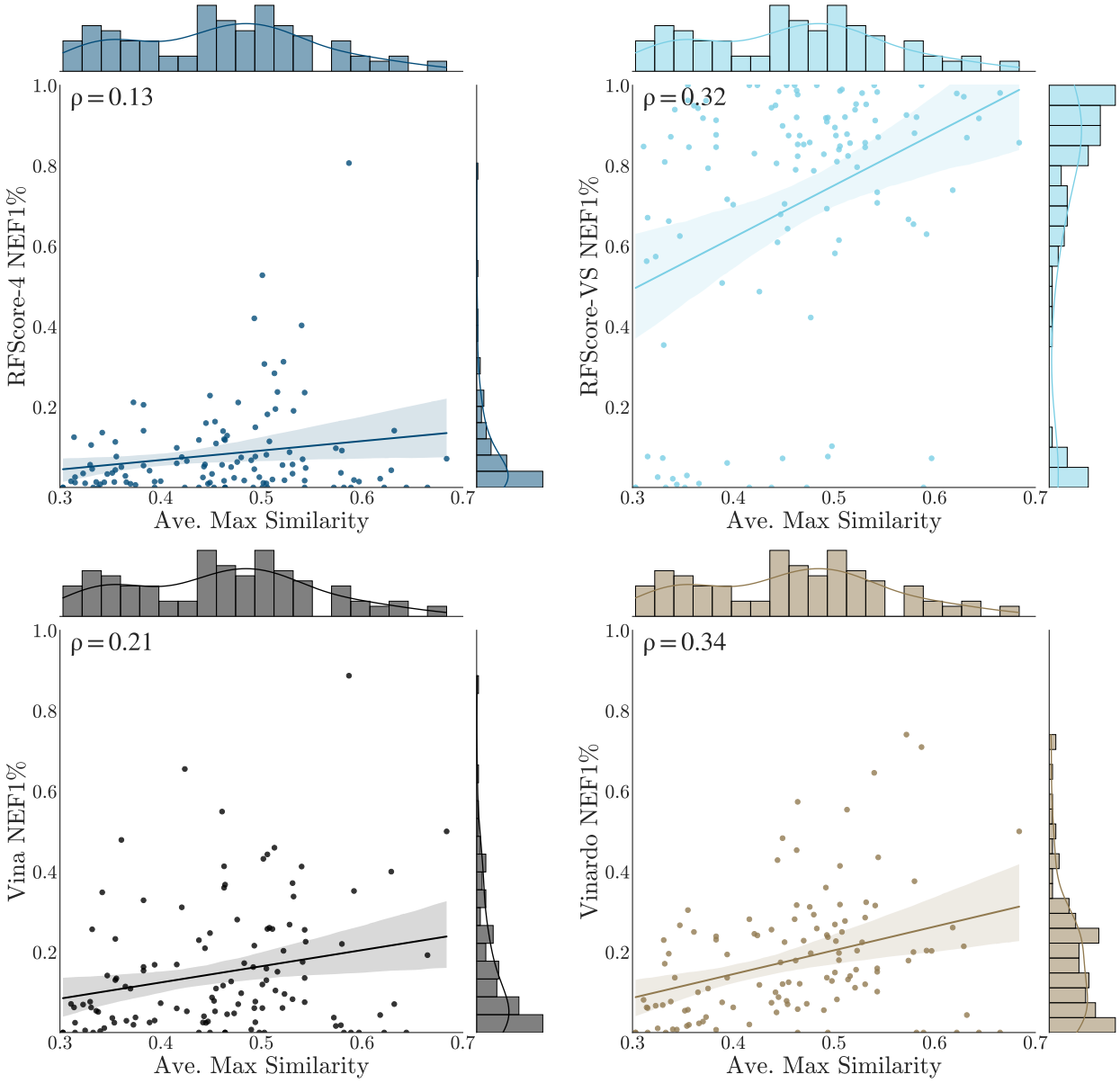


Figure S25: Correlation between similarity with training set and early enrichment performance for the non-CNN models. For each benchmark, the average of the maximum similarity between active compounds and the PDBbind General set is computed using the Tanimoto coefficient of ECFP4 fingerprints. Only actives are considered since the training set does not include any inactive compounds. Note that RFScore-VS used DUD-E for training.

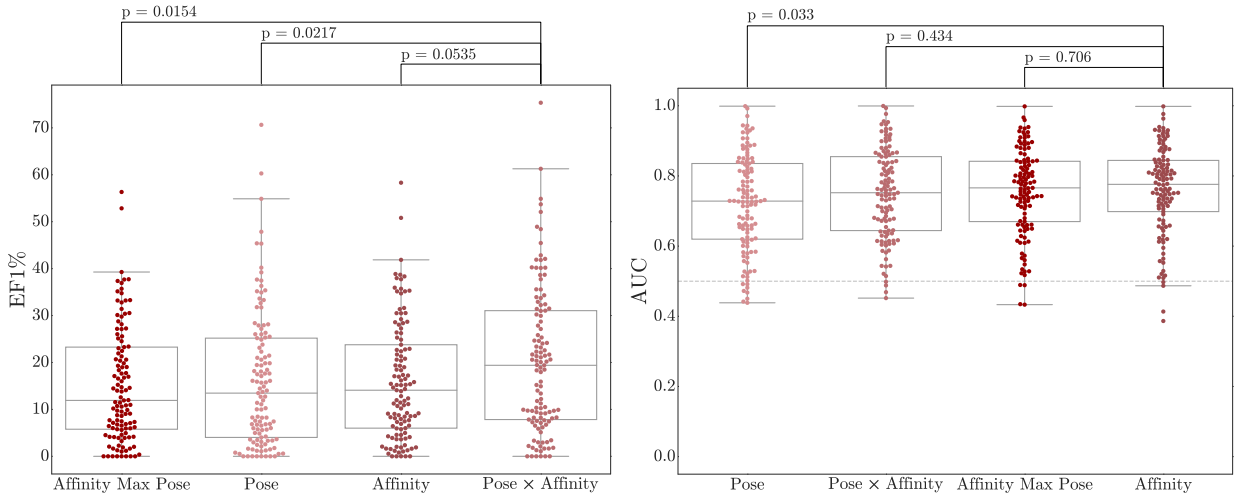


Figure S26: Performance of pose/affinity score combinations measured with (a) EF1% and (b) AUC. Pose and affinity scores are combined either by taking the predicted affinity of the pose with the best pose score, or by multiplying the affinity and pose scores. Performance of the Default ensemble on both DUD-E and LIT-PCBA targets is shown. P-values are computed using the two-sided Mann-Whitney U rank test. The x-axes are independently sorted by the median value.

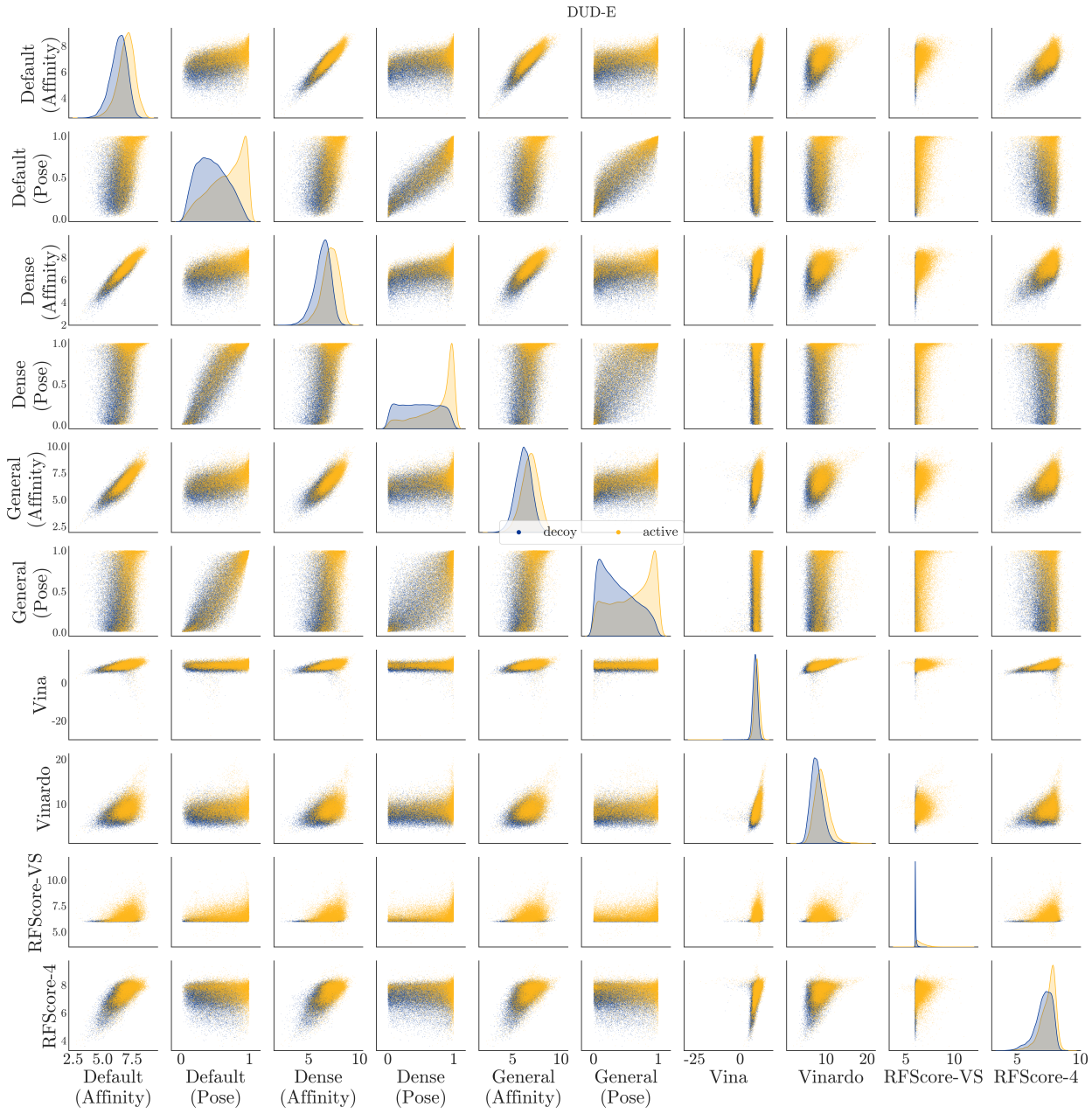


Figure S27: Score distributions and correlations for the different methods tested for DUD-E, separated by class. Note that RFScore-VS was trained on DUD-E. For visualization purposes, decoys were randomly downsampled to the same size as the actives.

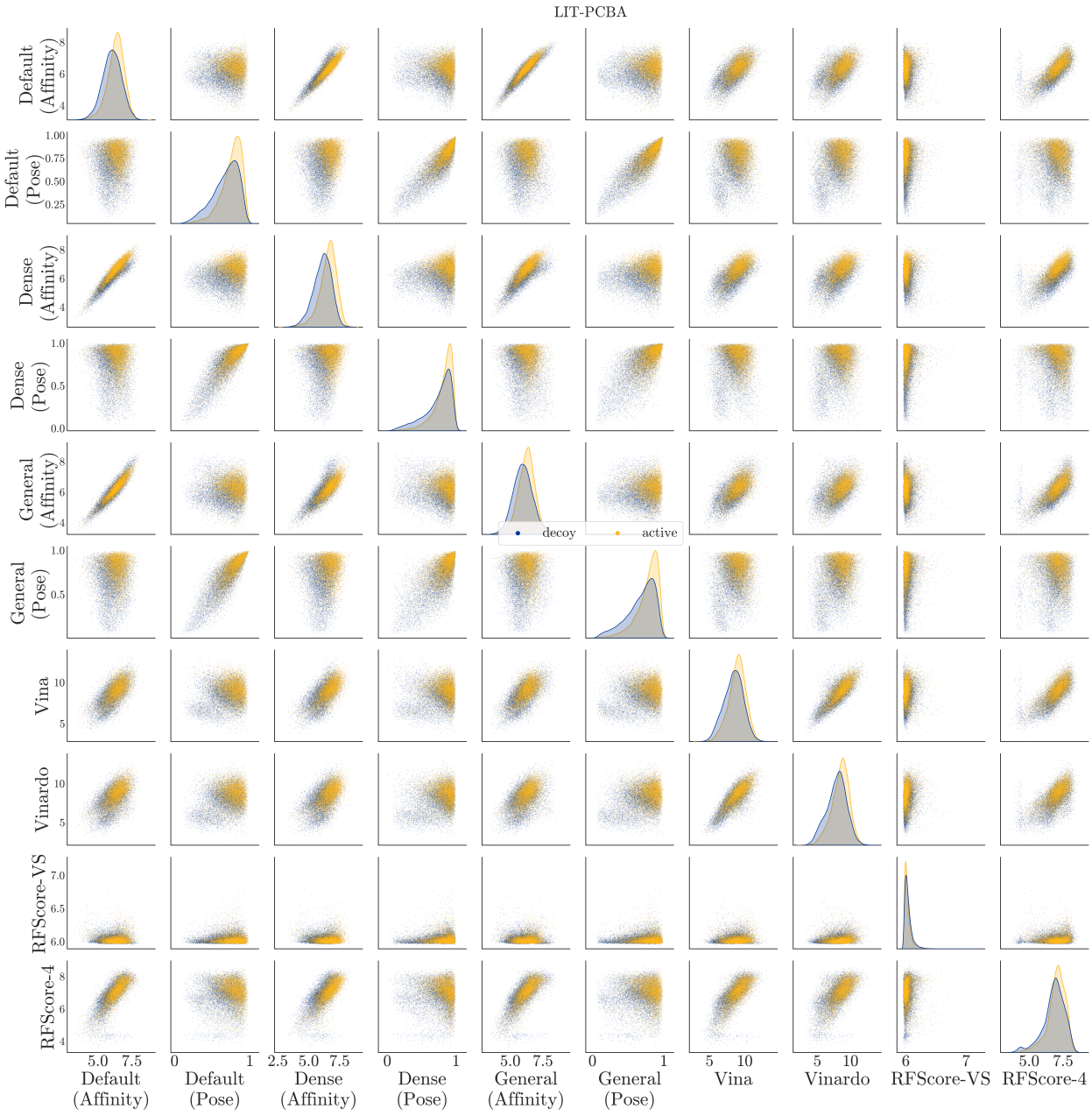


Figure S28: Score distributions and correlations for the different methods tested for LIT-PCBA, separated by class. For visualization purposes, decoys were randomly downsampled to the same size as the actives.

References

(1) Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* **1998**, *10*, 1895–1923.

(2) Nicholls, A. Confidence limits, error bars and method comparison in molecular modeling. Part 1: the calculation of confidence intervals. *Journal of computer-aided molecular design* **2014**, *28*, 887–918.

(3) Huang, R.; Sakamuru, S.; Martin, M. T.; Reif, D. M.; Judson, R. S.; Houck, K. A.; Casey, W.; Hsieh, J.-H.; Shockley, K. R.; Ceger, P., et al. Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. *Scientific reports* **2014**, *4*, 1–9.

(4) Hart, R. J.; Lawres, L.; Fritzen, E.; Mamoun, C. B.; Aly, A. S. Plasmodium yoelii vitamin B 5 pantothenate transporter candidate is essential for parasite transmission to the mosquito. *Scientific reports* **2014**, *4*, 1–7.

(5) Dorjsuren, D.; Kim, D.; Maloney, D. J.; Wilson III, D. M.; Simeonov, A. Complementary non-radioactive assays for investigation of human flap endonuclease 1 activity. *Nucleic acids research* **2011**, *39*, e11–e11.

(6) Goldin, E.; Zheng, W.; Motabar, O.; Southall, N.; Choi, J. H.; Marugan, J.; Austin, C. P.; Sidransky, E. High throughput screening for small molecule therapy for Gaucher disease using patient tissue as the source of mutant glucocerebrosidase. *PloS one* **2012**, *7*, e29861.

(7) Davis, M.; Pragani, R.; Popovici-Muller, J.; Gross, S.; Thorne, N.; Salituro, F.; Fantin, V.; Straley, K.; Su, M.; Dang, L., et al. ML309: A potent inhibitor of R132H mutant IDH1 capable of reducing 2-hydroxyglutarate production in U87 MG glioblastoma cells. *Probe Reports from the NIH Molecular Libraries Program [Internet]* **2013**,

(8) Marugan, J.; Dehdashti, S.; Zheng, W.; Southall, N.; Inglese, J.; Austin, C. HTS for Identification of Inhibitors against the ERK Signaling Pathway using a Homogenous

Cell-based Assay. *Probe Reports from the NIH Molecular Libraries Program [Internet]* **2010**,

(9) Hoffman, G. R.; Moerke, N. J.; Hsia, M.; Shamu, C. E.; Blenis, J. A high-throughput, cell-based screening method for siRNA and small molecule inhibitors of mTORC1 signaling using the In Cell Western technique. *Assay and drug development technologies* **2010**, *8*, 186–199.

(10) Frankowski, K. J.; Hedrick, M. P.; Gosalia, P.; Li, K.; Shi, S.; Whipple, D.; Ghosh, P.; Prisinzano, T. E.; Schoenen, F. J.; Su, Y., et al. Discovery of small molecule kappa opioid receptor agonist and antagonist chemotypes through a HTS and hit refinement strategy. *ACS chemical neuroscience* **2012**, *3*, 221–236.

(11) Boxer, M. B.; Jiang, J.-k.; Vander Heiden, M. G.; Shen, M.; Skoumbourdis, A. P.; Southall, N.; Veith, H.; Leister, W.; Austin, C. P.; Park, H. W., et al. Evaluation of substituted N, N'-diarylsulfonamides as activators of the tumor cell specific M2 isoform of pyruvate kinase. *Journal of medicinal chemistry* **2010**, *53*, 1048–1055.

(12) Witt, K. L.; Hsieh, J.-H.; Smith-Roe, S. L.; Xia, M.; Huang, R.; Zhao, J.; Auerbach, S. S.; Hur, J.; Tice, R. R. Assessment of the DNA damaging potential of environmental chemicals using a quantitative high-throughput screening approach to measure p53 activation. *Environmental and molecular mutagenesis* **2017**, *58*, 494–507.

(13) Nandhikonda, P.; Lynt, W. Z.; McCallum, M. M.; Ara, T.; Baranowski, A. M.; Yuan, N. Y.; Pearson, D.; Bikle, D. D.; Guy, R. K.; Arnold, L. A. Discovery of the first irreversible small molecule inhibitors of the interaction between the vitamin D receptor and coactivators. *Journal of medicinal chemistry* **2012**, *55*, 4640–4651.

(14) Tran-Nguyen, V.-K.; Jacquemard, C.; Rognan, D. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *Journal of Chemical Information and Modeling* **2020**,