Supporting Information: In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening

Jochen Sieg, Florian Flachsenberg, and Matthias Rarey*

Center for Bioinformatics, Research Group for Computational Molecular Design,
University of Hamburg, Bundesstraße 43, 20146 Hamburg, Germany

E-mail: rarey@zbh.uni-hamburg.de

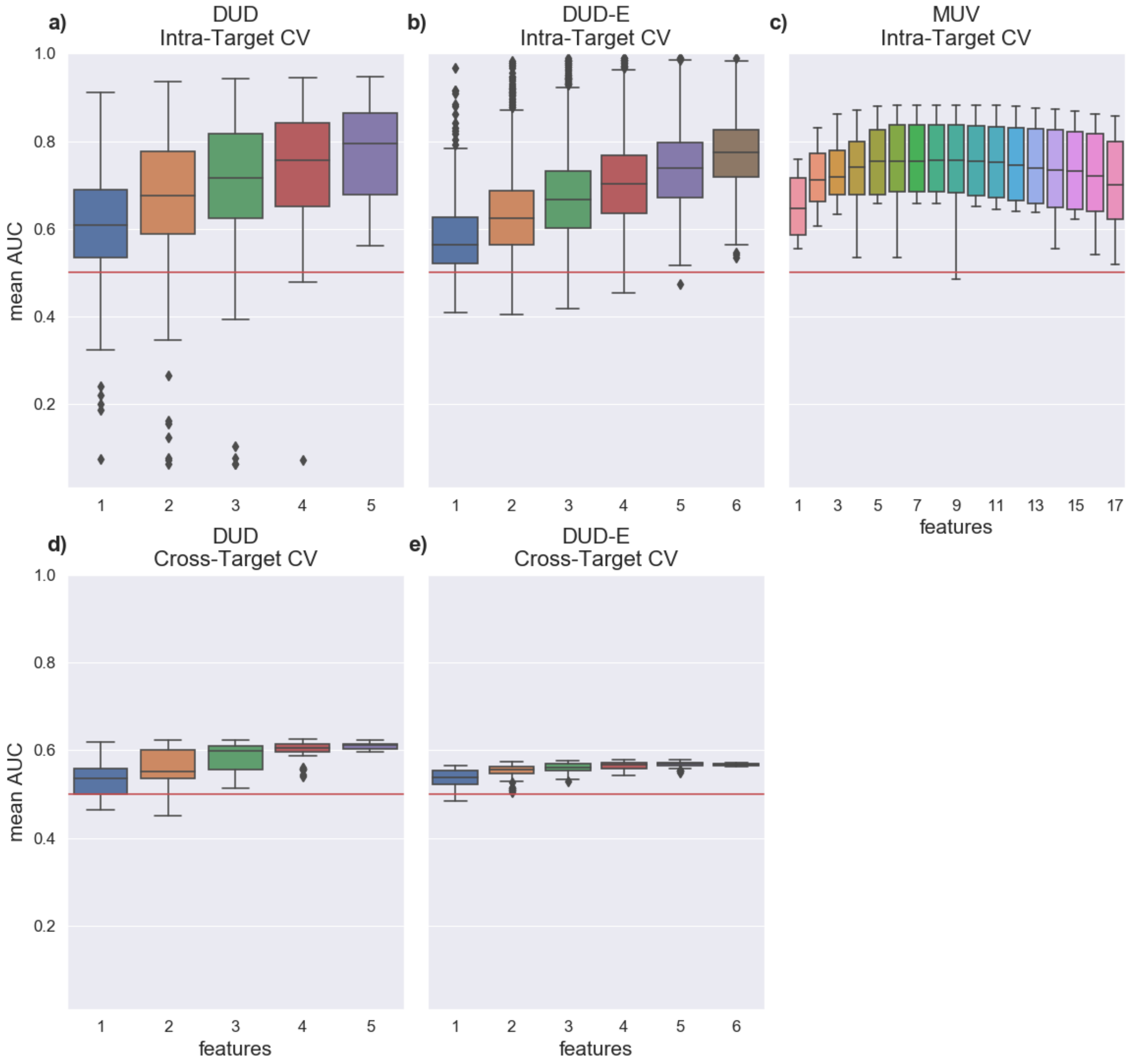

Figure S1: Box plot of the evaluation of unbiased features with logistic regression (LR) of DUD, DUD-E and MUV using AUC. The first row shows results of the intra-target CV for (a) DUD, (b) DUD-E and (c) MUV. In the second row results of cross target CV are depicted for (d) DUD and (e) DUD-E.

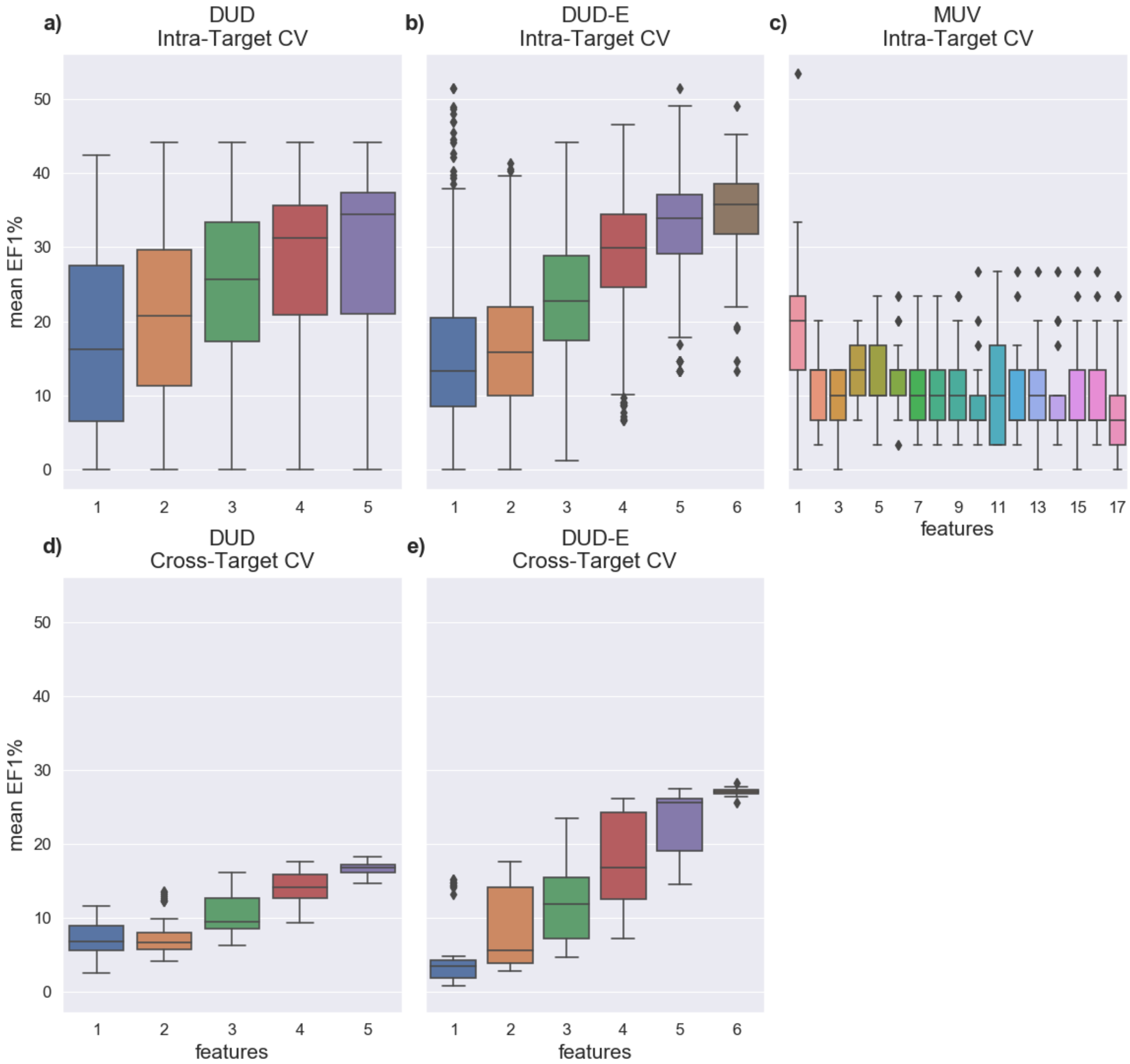

Figure S2: Box plot of the evaluation of unbiased features with random forest (RF) of DUD, DUD-E and MUV. Performance is assessed with the mean enrichment factor on one percent of the test sets (EF1%). The first row shows results of the intra-target CV for (a) DUD, (b) DUD-E and (c) MUV. In the second row results of cross target CV are depicted for (d) DUD and (e) DUD-E.

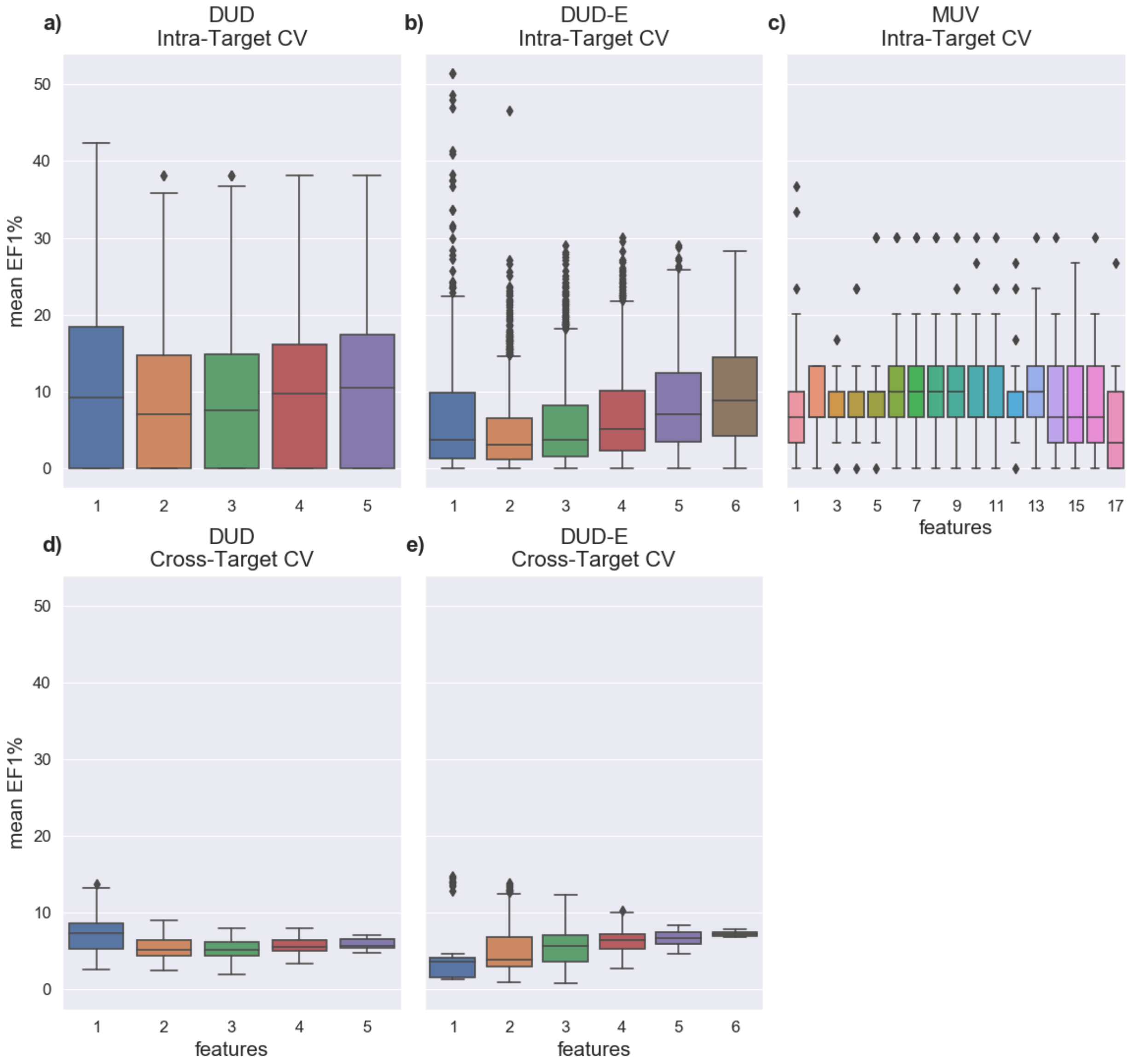

Figure S3: Box plot of the evaluation of unbiased features with logistic regression (LR) of DUD, DUD-E and MUV. Performance is assessed with the mean enrichment factor on one percent of the test sets (EF1%). The first row shows results of the intra-target CV for (a) DUD, (b) DUD-E and (c) MUV. In the second row results of cross target CV are depicted for (d) DUD and (e) DUD-E.

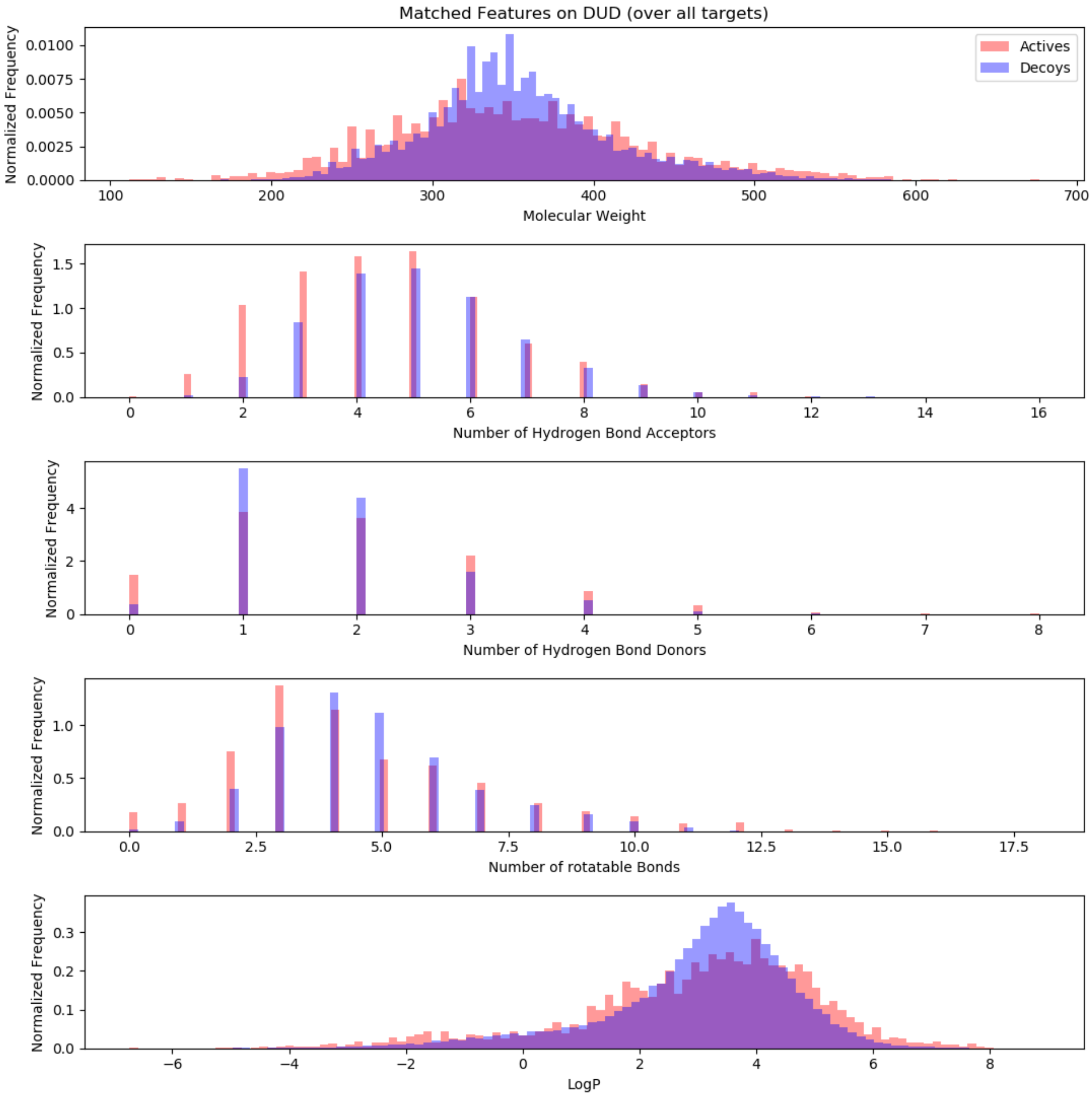

Figure S4: Histograms of all unbiased features of DUD over all targets of the dataset. Actives are marked in red and inactives in blue.

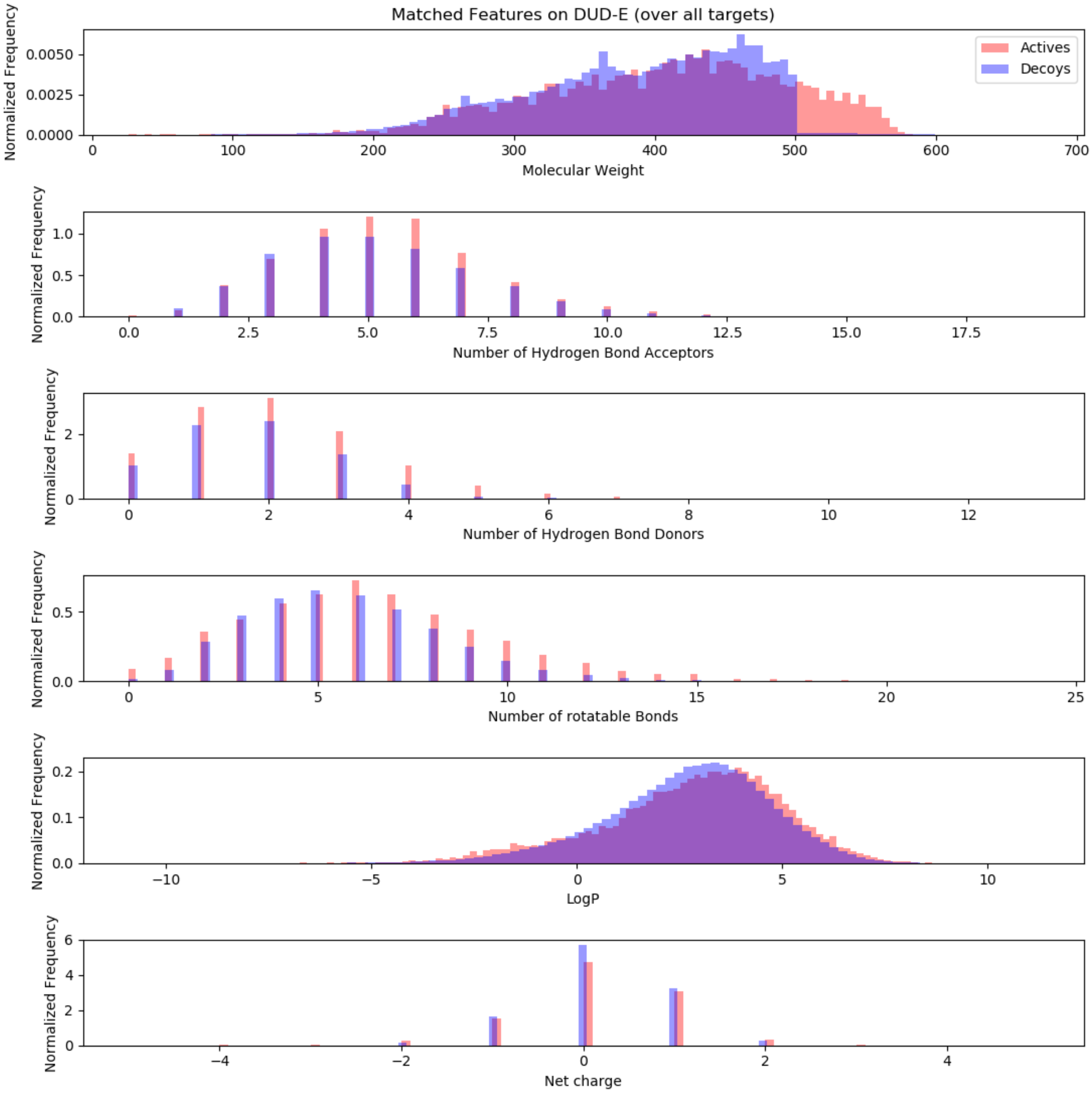

Figure S5: Histograms of all unbiased features of DUD-E over all targets of the dataset. Actives are marked in red and inactives in blue.

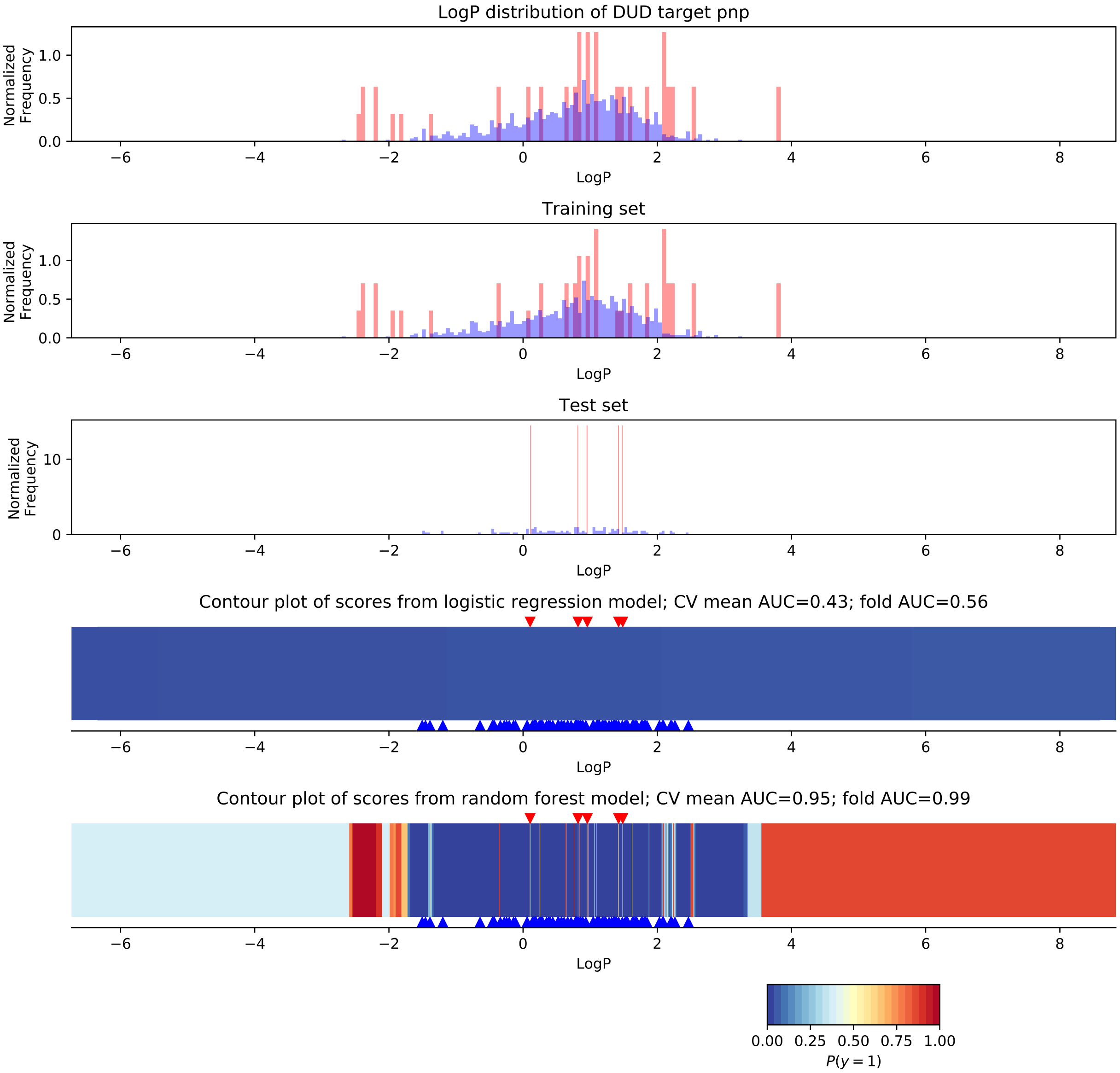

Figure S6: Illustration of the intra-target CV experiment with LogP on the protein target PNP of DUD. In this experiment the highest mean AUC value with random forest was achieved over all DUD targets and single features. Histograms depict the LogP distributions on the target level and of the training and test set of one fold of the CV. Additionally, contour plots of the scores of the random forest model and logistic regression model fitted on the training set is shown. The red triangles on the top of the contour plots mark the LogP value for the actives in the test set, while the blue triangles at the bottom mark the LogP values of the test decoys.

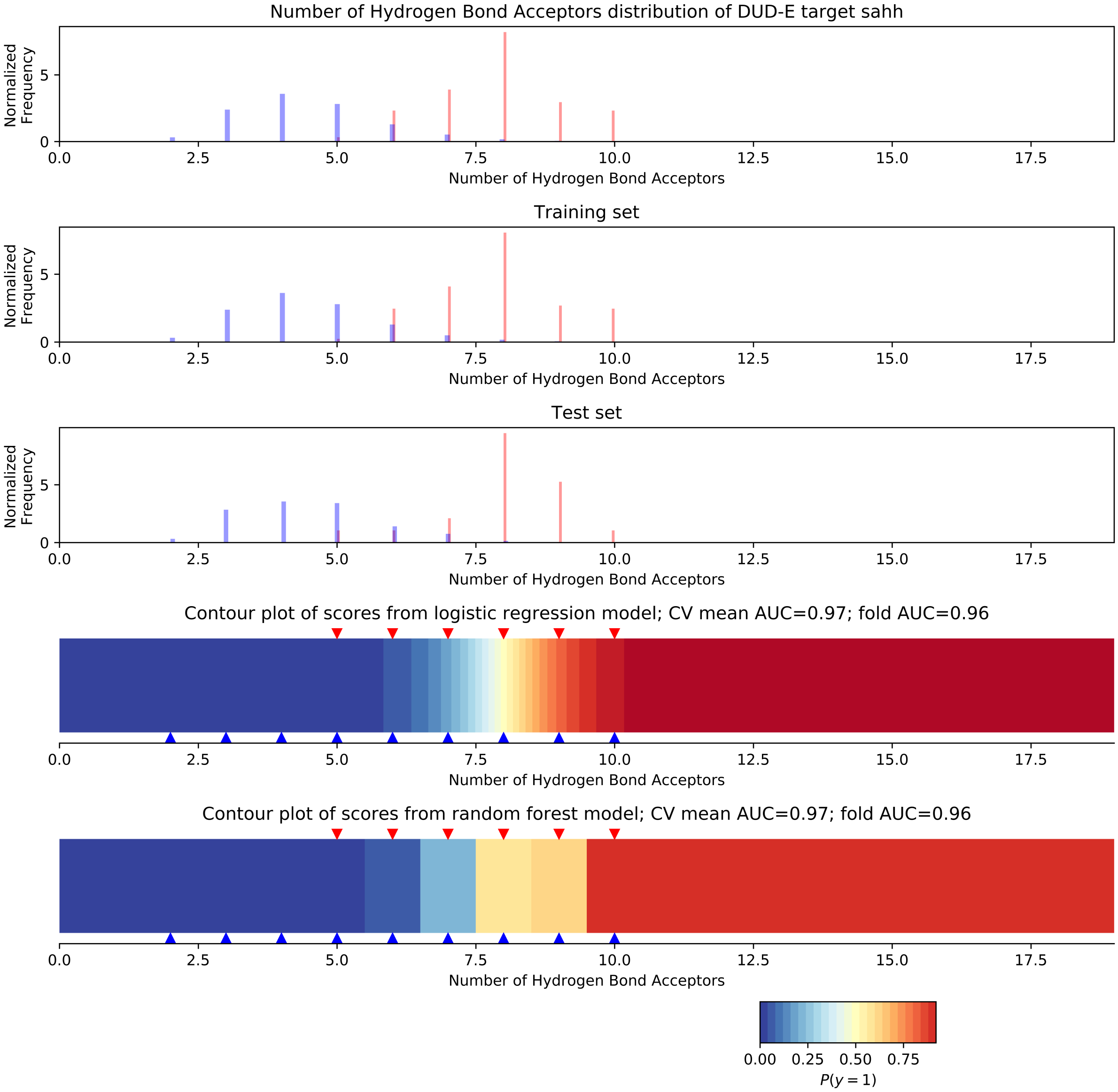

Figure S7: Illustration of the intra-target CV experiment with the number of hydrogen bond acceptors on the protein target SAHH of DUD-E. In this experiment the highest mean AUC value with linear regression was achieved over all DUD-E targets with single features. Histograms depict the acceptor distributions on the target level and of the training and test set of one fold of the CV. Additionally, contour plots of the scores of the random forest model and logistic regression model fitted on the training set is shown. The red triangles on the top of the contour plots mark the number of acceptors for the actives in the test set, while the blue triangles at the bottom mark the number of acceptors of the test decoys.

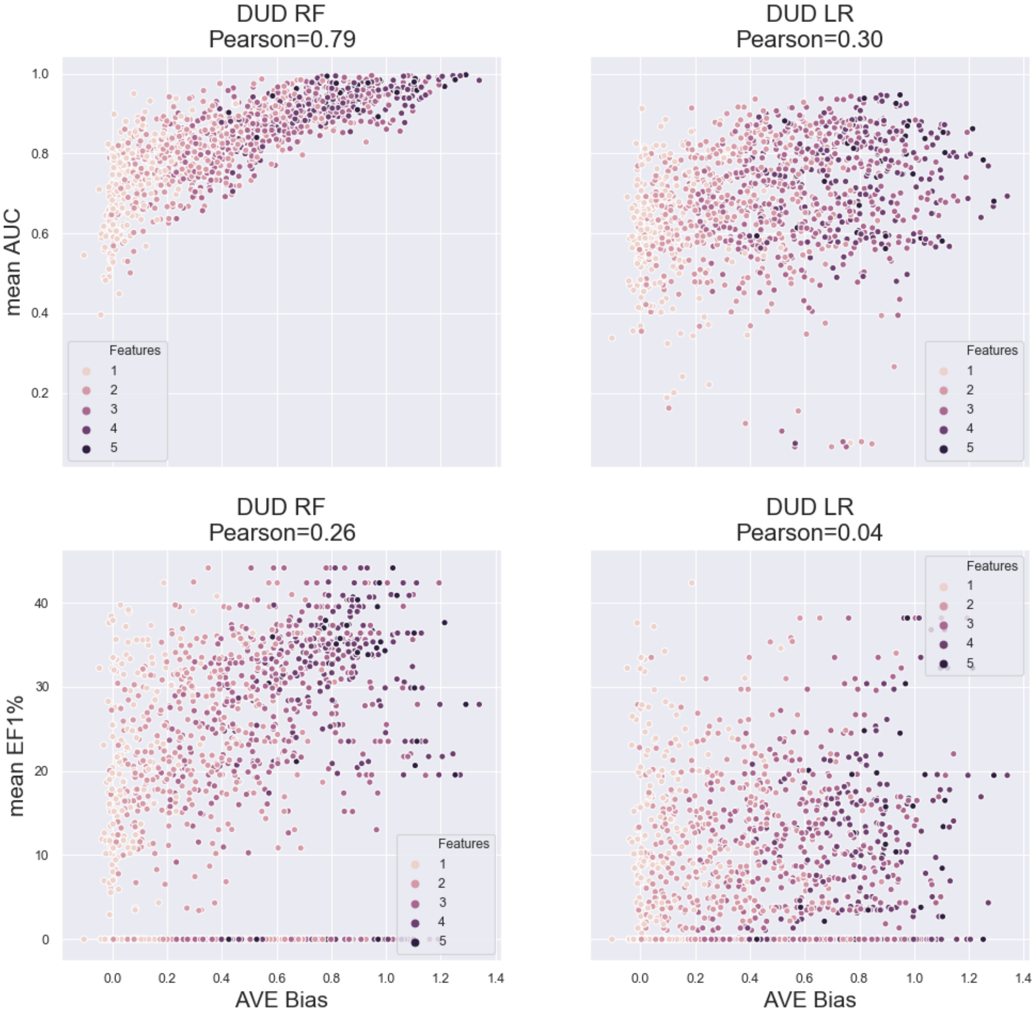

Figure S8: Correlation plots of AVE bias and AUC in the intra-target CV of DUD. One point represents the mean performance in dependence of the AVE bias over a single intra-target CV experiment (this plot sets the results of Figure 2, S1, S2, S3 in relation to AVE bias). For RF evaluated with AUC there is a notable correlation between performance and AVE bias. The coloring of the points illustrates the number of features used in the experiments (light coloring means less features and darker more features). It can be seen that experiments with less features exhibit low or no AVE bias, while experiments with many features have increased AVE bias. Even though there is correlation when using RF with AUC there are also many examples, especially with a single feature, where there is very high performance, but no AVE bias. Therefore, not all bias in DUD is explained by AVE.

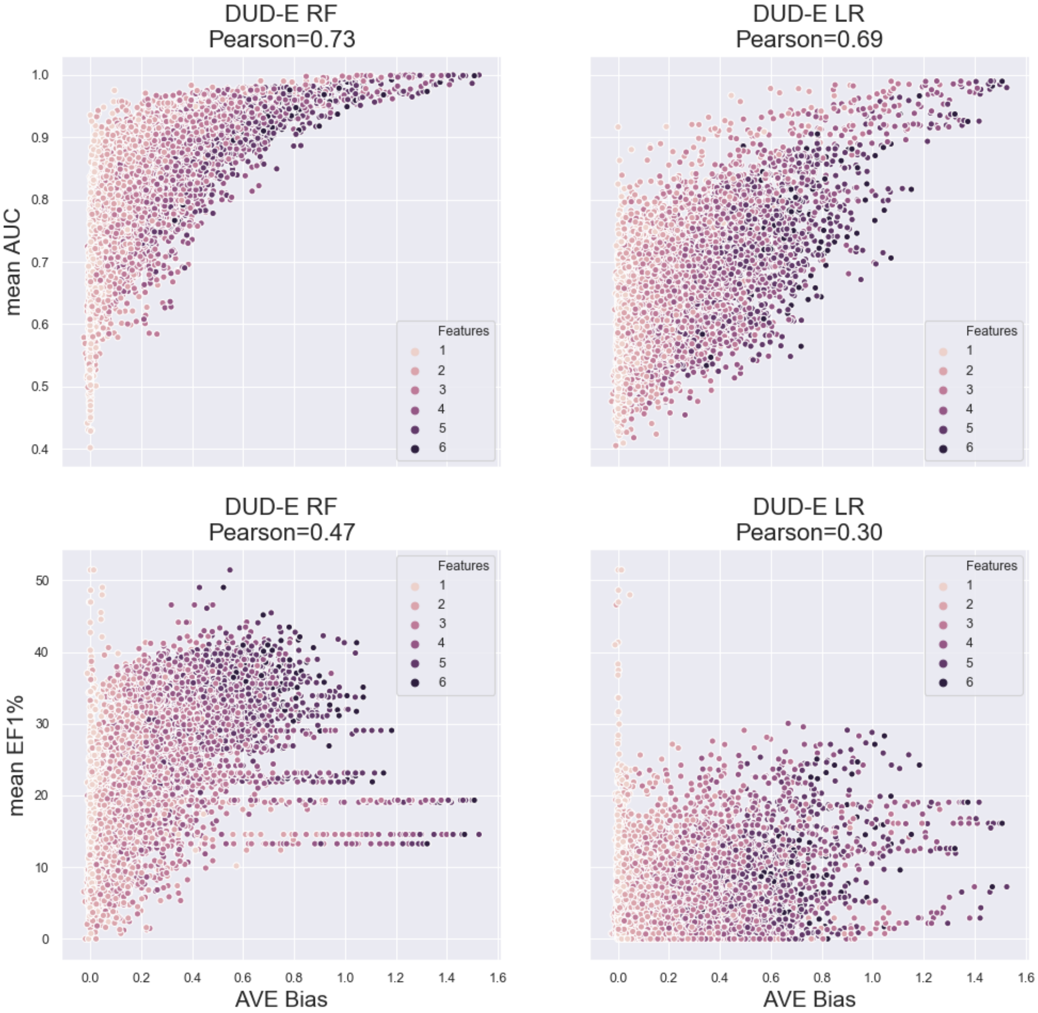

Figure S9: Correlation plots of AVE bias and AUC in the intra-target CV of DUD-E. One point represents the mean performance in dependence of the AVE bias over a single intra-target CV experiment (this plot sets the results of Figure 2, S1, S2, S3 in relation to AVE bias). For RF and LR evaluated with AUC there is a notable correlation between performance and AVE bias. For experiments with EF1% there is a moderate and weak correlation observable. The coloring of the points illustrates the number of features used in the experiments (light coloring means less features and darker more features). It can be seen that experiments with less features exhibit low or no AVE bias, while experiments with many features have increased AVE bias. Even though there is correlation when using RF and LR with AUC there are also many examples, especially with a single feature, where there is very high performance, but no AVE bias. Therefore, not all bias in DUD-E is explained by AVE.

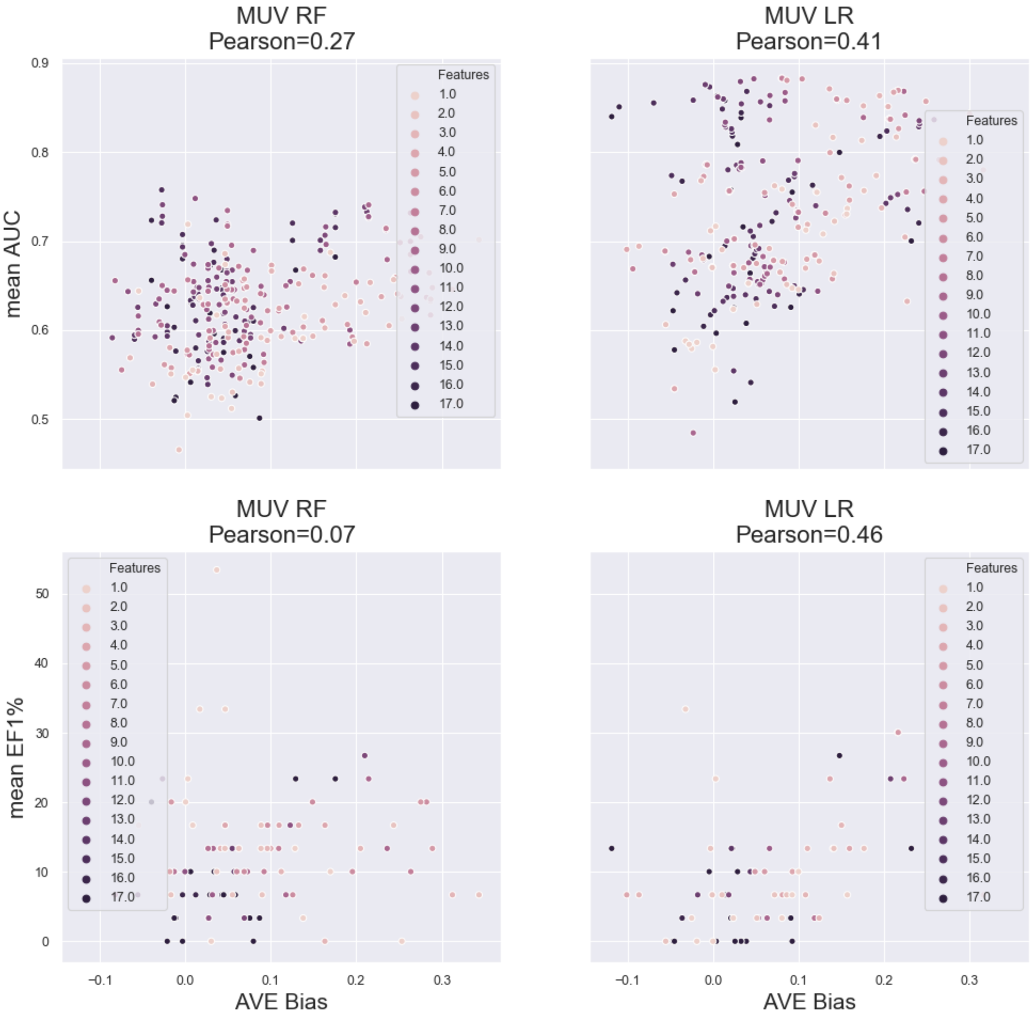

Figure S10: Correlation plots of AVE bias and AUC in the intra-target CV of MUV. One point represents the mean performance in dependence of the AVE bias over a single intra-target CV experiment (this plot sets the results of Figure 2, S1, S2, S3 in relation to AVE bias). There is a weak to moderate correlation between LR results in both AUC and EF1%, while for RF no notable correlation is observed. The coloring of the points illustrates the number of features used in the experiments (light coloring means less features and darker more features). There is no clear trend observable in the coloring.

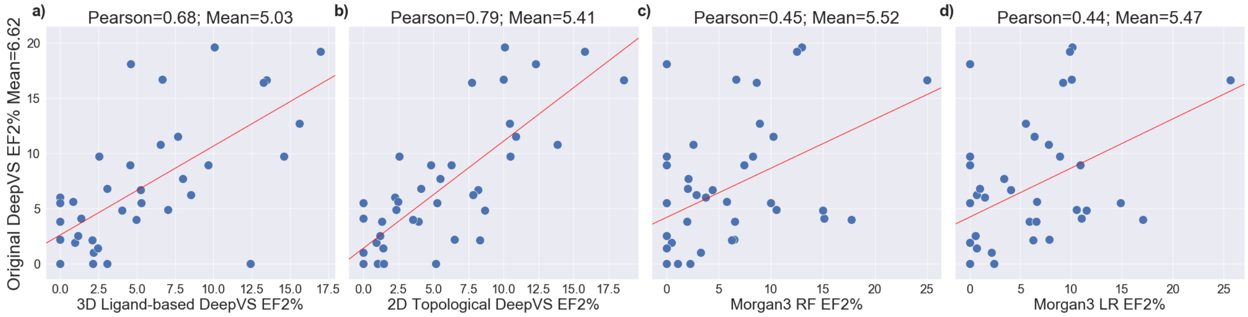

Figure S11: Correlation plots of enrichment factor at 2% (EF2%) values of the structure-based original DeepVS (values taken from *Pereira et al.* 2016) and the four other approaches. **a)** shows the performance of our 3D ligand-based reimplementation. The second plot (**b)**) shows the correlation of the reimplementation using the topological distance on the molecular graph instead of 3D distances. Finally, the performance of RF and LR with Morgan3 fingerprints is plotted against the original DeepVS in **c)** and **d)**, respectively.

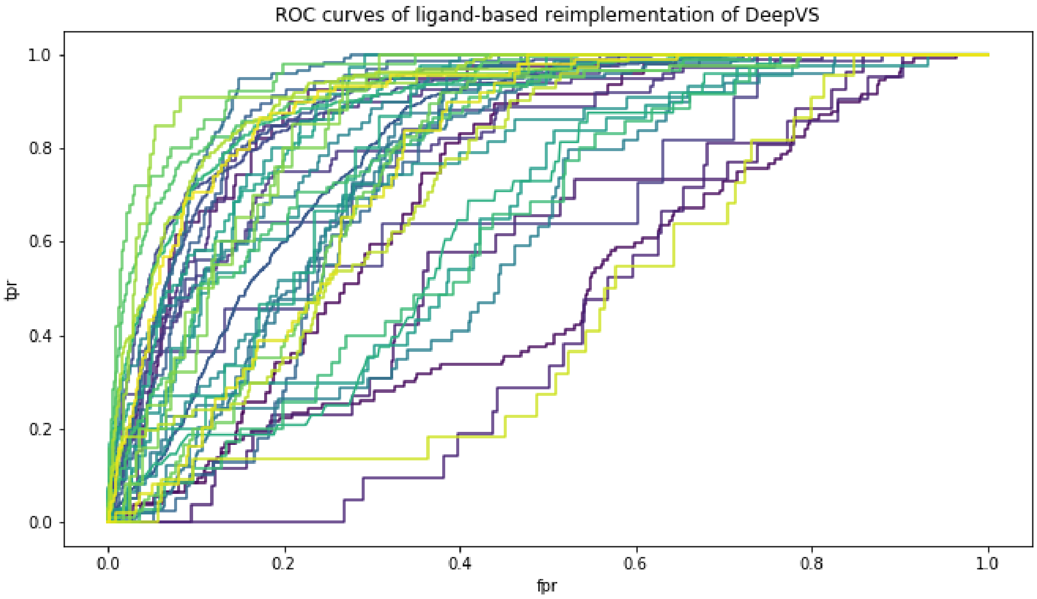

Figure S12: ROC curves of the 40 folds of the LOO-CV with our ligand-based 3D reimplementation of DeepVS.

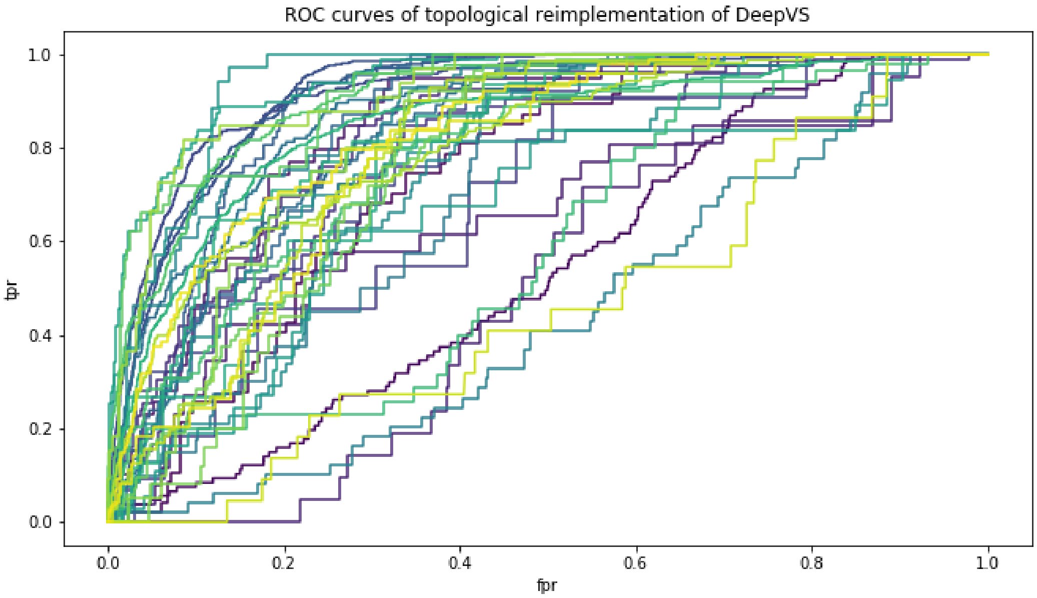

Figure S13: ROC curves of the 40 folds of the LOO-CV with our ligand-based 2D (topological) reimplementation of DeepVS.

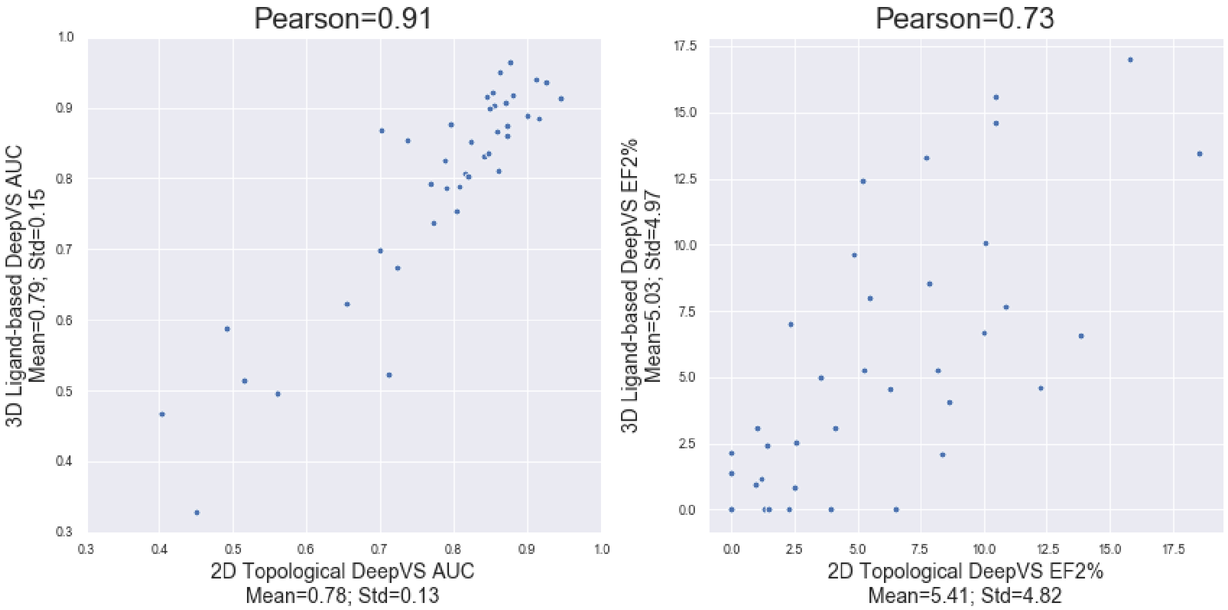

Figure S14: Comparison of AUC and EF2% values of our ligand-based 3D reimplementation and our 2D (topological) reimplementation of DeepVS. Correlation plots over the 40 cross validation folds.

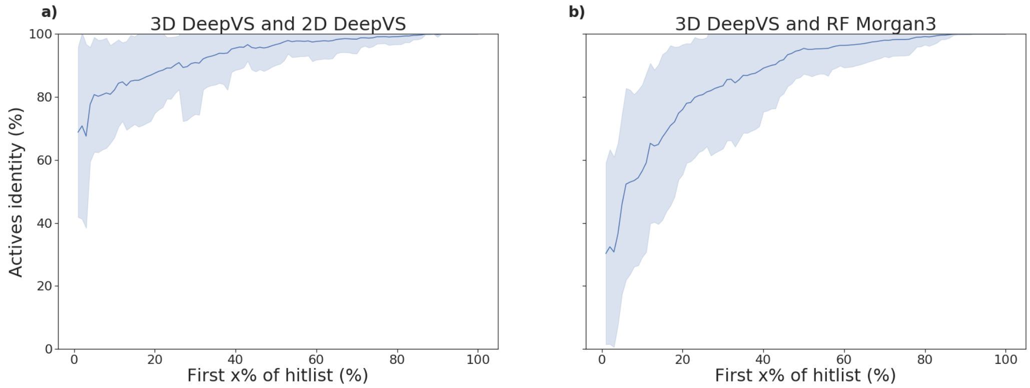

Figure S15: Distribution of overlapping actives between the (a) 3D DeepVS and 2D DeepVS as well as the (b) 3D DeepVS and RF with Morgan3 fingerprints. The mean values over the 40 folds of the LOO-CV on DUD are shown. On the y-axis the overlap of actives between the hitlists is depicted in dependence of the first $x\%$ of the ranked molecules, which is shown on the x-axis. Consequently, one point in the plot describes that in the first $x\%$ of the hitlists $y\%$ of the enriched actives are identical between methods. The light colored band shows the standard deviation.

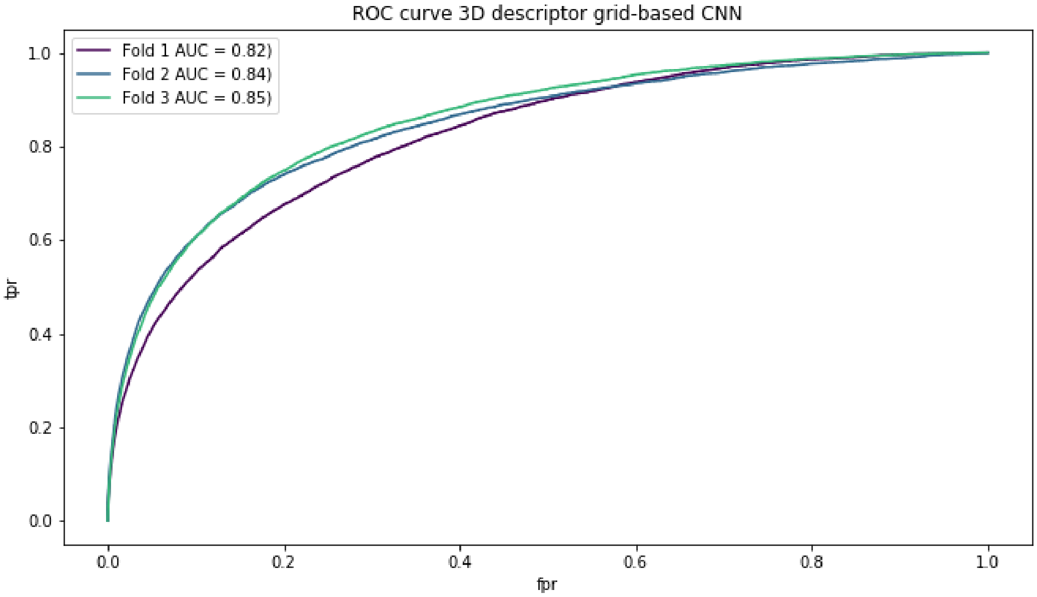

Figure S16: ROC curves for the three folds of the cCCV with our ligand-based 3D reimplementation of the grid-based CNN.

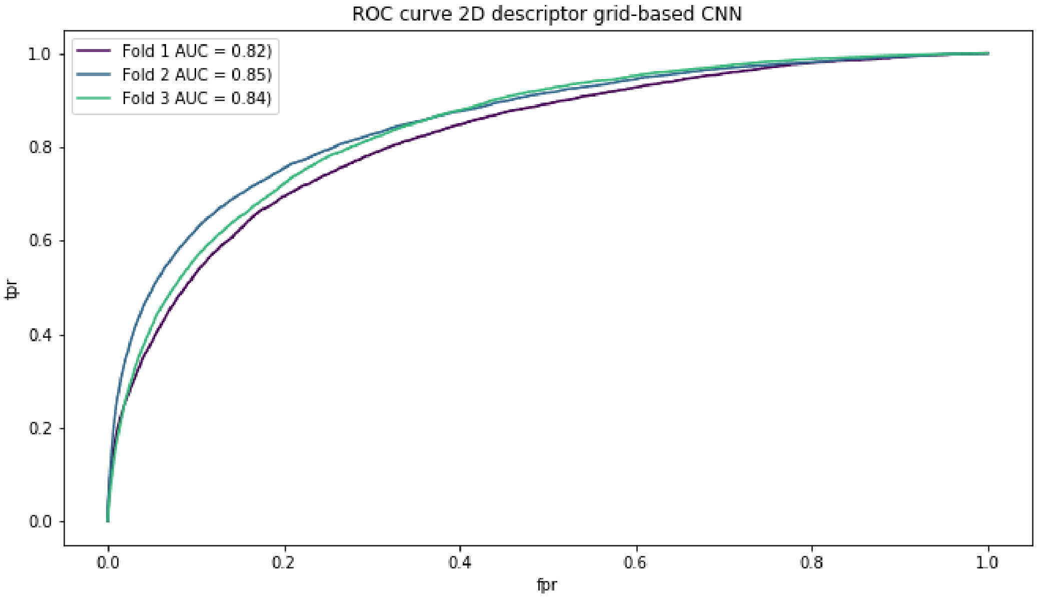

Figure S17: ROC curves for the three folds of the cCCV with our ligand-based 2D reimplementation of the grid-based CNN.

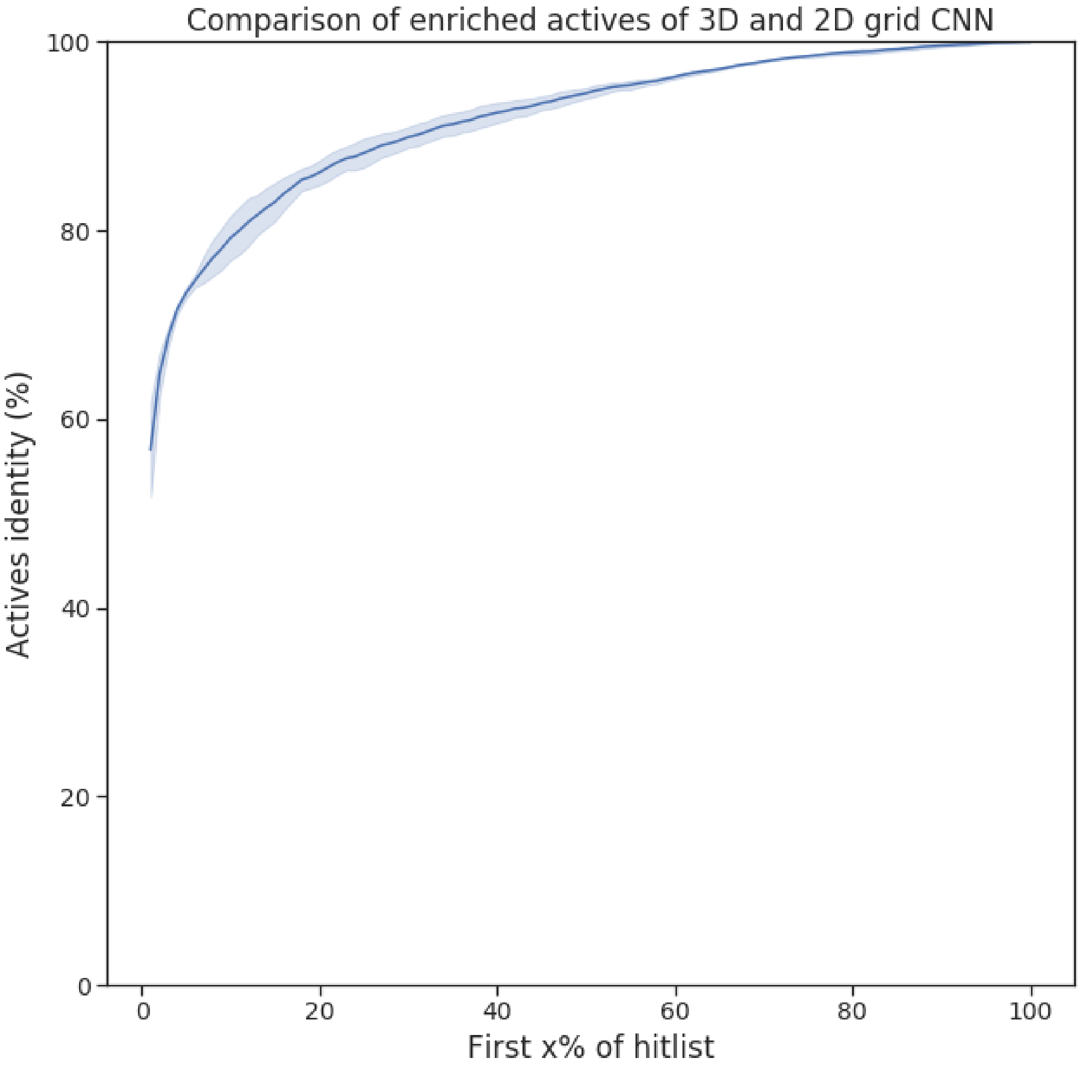

Figure S18: Distribution of overlapping actives between the 3D and 2D version of the grid-based CNN. The mean values over the 3 folds of the cCCV on DUD-E are shown. On the y-axis the overlap of actives between the hitlists is depicted in dependence of the first x% of the ranked molecules, which is shown on the x-axis. Consequently, one point in the plot describes that in the first x% of the hitlists y% of the enriched actives are identical between methods. The light colored band shows the standard deviation.