ChemMedChem

Supporting Information

Prioritizing Small Sets of Molecules for Synthesis through *in-silico* Tools: A Comparison of Common Ranking Methods

Marko Breznik+, Yunhui Ge+, Joseph P. Bluck+, Hans Briem, David F. Hahn, Clara D. Christ, Jérémie Mortier, David L. Mobley, and Katharina Meier*

Table S1: The PDB IDs of structures used in this study and the RMSD (Å) relative to the previously published benchmarking set. RMSDs were obtained from aligning the backbone CA atoms of chain A (except for thrombin) to the previously published benchmark set.[1]

Target label	PDB ID	RMSD (Å)
cdk2	1H1Q	0.17
cdk8	5HNB	0.15
cmet	4R1Y	0.21
eg5	3L9H	0.31
galectin	5E89	0.24
hif2a	5TBM	0.14
jnk1	2GMX	0.13
mcl1	4HW3	0.00
p38	3FLY	0.19
pfkfb3	6HVI	0.17
ptp1b	2QBS	0.18
shp2	5EHR	0.17
syk	4PV0	0.11
thrombin	2ZFF	0.48
tnks2	4UI5	0.15
tyk2	4GIH	0.18

Confusion matrix analysis shows an overall better performance of MD-based methods compared to docking algorithms and MM/GBSA in ranking compounds.

To further assess the ranking-order, a confusion-matrix analysis was performed. Note that the confusion-matrix analysis in this work is not applied to compounds which are actives or inactives as used in traditional benchmark studies. Instead, the goal of this work is to assess performance of docking algorithms for ligand ranking of a set of ligands that was previously used in MD-based binding free energy calculations. Therefore, all of selected compounds for each target are active binders. Correspondingly, these compounds were defined as potent binders and weak binders in the confusion matrix analysis instead of actives and inactives. More specifically, the top 25% of ranked ligands and the last 25% of ranked ligands were defined as potent and weak binders based on their experimentally measured binding free energies for the purpose of confusion matrix analysis. Then true potent (TP), true weak (TW), false potent (FP) and false weak (FW) rates were computed for each target and the overall dataset, and then used to compare these methods (high TP/TW and low FP/FW rates indicate a better performance).

A true potent/weak (TP/TW) rate of 100% indicates all potent/weak binders picked by the method are real potent/weak binders. In contrast, a false potent/weak (FP/FW) rate of 100% means all potent/weak binders picked by the methods are actually weak/potent binders. So a robust method should return high TP/TW rates and low FP/FW rates.

We first checked overall TP/FP/TW/FW rates by averaging results across all targets for each method. MD-based methods (FEP+, PMX) had highest TP/TW rates and lowest FP/FW rates among all methods (Table S2). Among all docking methods, GoldScore was the best (highest TP/TW rates and lowest FP/FW rates). MM/GBSA calculations had similar performance as GoldScore. We found error bars in Table S2 were large (> 20%), indicating performance fluctuations between different targets.

Table S2: Confusion matrix analysis results of docking algorithms and MD-based methods. Reported are averaged values across all targets (without bootsrapping). Uncertainties are estimated using standard deviations.

Methods	TP (%)	TW (%)	FP (%)	FW (%)
MD-based methods				
PMX	40 ± 21	51 ± 20	60 ± 21	49 ± 20
FEP+	60 ± 13	64 ± 20	40 ± 13	36 ± 20
non-constrained docking				
ChemPLP	31 ± 18	39 ± 21	69 ± 18	61 ± 21
GoldScore	44 ± 26	45 ± 24	56 ± 26	55 ± 24
Glide	23 ± 20	32 ± 20	77 ± 20	68 ± 20
FlexX	26 ± 26	27 ± 24	74 ± 26	73 ± 24
FRED	27 ± 25	35 ± 24	73 ± 25	65 ± 24
AutoDock Vina	29 ± 23	28 ± 26	71 ± 23	72 ± 26
MM/GBSA	40 ± 26	38 ± 26	60 ± 26	62 ± 26
constrained docking				
ChemPLP	37 ± 17	46 ± 21	63 ± 17	54 ± 21
GoldScore	42 ± 19	50 ± 22	58 ± 19	50 ± 22
Glide	35 ± 15	30 ± 25	65 ± 15	70 ± 25
FlexX	30 ± 23	36 ± 23	70 ± 23	64 ± 23
HYBRID	32 ± 27	41 ± 25	68 ± 27	59 ± 25
MM/GBSA	41 ± 22	42 ± 21	59 ± 22	58 ± 21

To assess the performance of these methods for each target, we performed 10000 rounds of bootstrapping in a confusion matrix analysis. In each round of bootstrapping, we randomly selected half of the total compound set for each target and ranked them using the experimental binding free energies. Then we used both the top and last 25% of compounds in each set as potent and weak binders.

A robust method for ligand ranking is expected to return as high as possible TP rates. Figure S1 summarized the mean TP rates from bootstrapping as described above. The uncertainty estimates can be found in Figure S37. In general, MD-based methods (FEP+, PMX) returned higher TP rates than docking algorithms and MM/GBSA. This indicated a better ability for ranking these compounds. We observed some exceptions though. FlexX, GoldSocre and FRED (all without constraints) had higher TP rates than MD-based methods in tnks2, pfkfb3 and hif2a, respectively. MM/GBSA had higher TP rates than MD-based methods in cdk8, ptp1b and thrombin. In constrained docking algorithms, GoldScore, ChemPLP, HYBRID and FlexX outperformed MD methods in galectin, ptp1b, syk and tnks2, respectively. MM/GBSA had higher TP rates than MD-based methods in cdk8 and thrombin. These results suggested the performance of docking algorithms and MM/GBSA varied between targets. In some cases, these methods could outperform MD methods in our confusion matrix analysis. Further investigations for these targets are needed to trace the origin of the challenges in MD simulations but this is beyond the scope of the current work. Despite these exceptions, MD-based methods still yielded higher TP rates than docking algorithms and MM/GBSA in most targets.

We further compared docking algorithms by checking the number of targets where a docking algorithm got the highest TP rate. Among docking algorithms without constraints, GoldScore ranked the first (9 targets). FRED, FlexX, ChemPLP and AutoDock Vina had similar performance (1, 1, 1, 3 targets, respectively). Among docking algorithms with scaffold constraint, ChemPLP, GoldScore and Glide had similar performance and yielded highest TP rate for 4, 5, and 3 targets, respectively. When comparing MM/GBSA and docking algorithms, MM/GBSA yielded the highest TP rate in 5 and 4 targets with non-constrained and constrained docking, respectively.

Similar non-uniform improvement were observed by using scaffold constraints in TP rate as we did in Kendall τ analysis. Even though constrained docking achieved higher population of close-to-reference conformations than non-constrained docking (Figure 1), it does not always lead to a better performance in binding potency predictions as we found here (Figure S1) and in Kendall τ values (Figure 2). For example, HYBRID gets a lower TP rate than FRED in hif2a. The fact that more cases were found in constrained docking that outperform MD methods in TP rates than non-constrained docking indicate improvements in those cases. But still, we should keep in mind that this is not always the case and is system dependent.

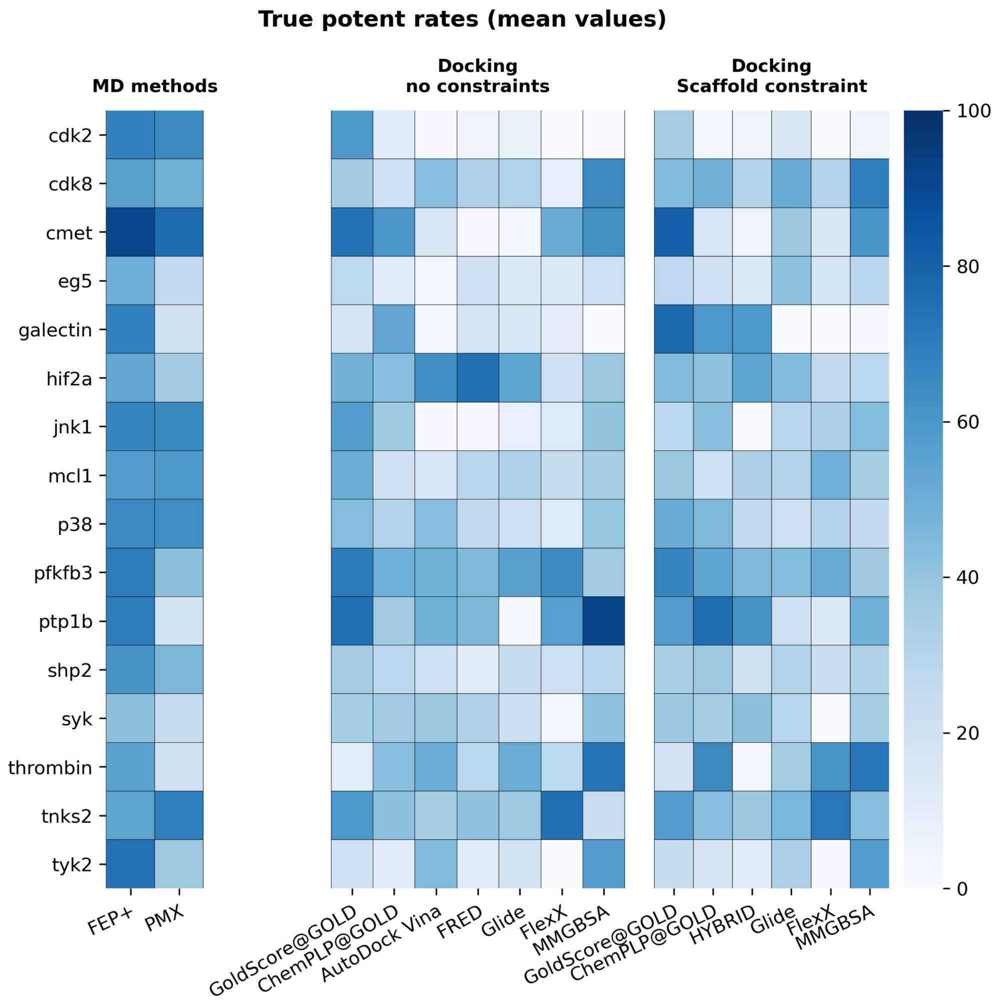

Figure S1: True potent rates (%) for each method across all targets. Mean values of each target after bootstrapping are reported here. MD-based methods have higher true potent rates than docking algorithms although exceptions are also observed.

A robust method for ligand ranking is expected to return as low as possible false potent (FP) rates. False potent rates indicate the percentage of ranked potent binders that are in fact weak binders. It is similar to false positives in virtual screening as in both cases weak binders/inactives are mistakenly ranked more potent/active.

Figure S2 summarizes mean values of FP rates after 10000 bootstrapping trials. Uncertainty estimates can be found in Figure S38. We can see MD-based methods had the lowest FP rates across all targets. Those docking algorithms and MM/GBSA that had good performance in TP rates also outperformed MD-based methods in FP rates. For non-constrained docking, FlexX in tnks2, GoldScore in pfkfb3, FRED in hif2a all had lower FP rates than MD methods. For constrained docking methods, GoldScore in galectin, ChemPLP in ptp1b, HYBRID in syk, FlexX in tnks2 outperformed MD-based methods. MM/GBSA with non-constrained docking in thrombin/ptp1b/cdk8 and with constrained docking in thrombin/cdk8 had lower FP rates than MD methods. Despite these exceptions, MD-based methods had an overall better performance in most targets.

Similar to our analysis in TP rates, we also compared docking algorithms in FP rates by the number of targets where the docking algorithm yielded the lowest FP rate. Among non-constrained docking algorithms, GoldScore had the best performance (9 targets). Other algorithms had similar performance (ChemPLP: 2, FlexX: 1, FRED: 1, AutoDock Vina: 3). Among constrained docking methods, GoldScore (5 targets), ChemPLP (4 targets) and Glide (3 targets) had similar performance and were better than FlexX (2 targets) and HYBRID (2 targets). Compared to docking algorithms, MM/GBSA had the best performance in 5 targets with non-constrained docking and 4 targtes with constrained docking. These results again suggested performance of these docking algorithms and MM/GBSA were highly system dependent.

MD-based methods outperformed docking algorithms and MM/GBSA in both TW and FW rates (Figure S39,S40,S41,S42). There were also cases where docking algorithms and/or MM/GBSA returned higher TW and lower FW rates than MD-based methods. The results are summarized in Table S6 and S7.

Similar to our analysis of TP/FP rates, we compared docking algorithms based on the number of targets where one docking method yielded the highest TW and lowest FW rate. Among non-constrained docking algorithms, GoldScore ranked the 1st for both TW and FW rates (6 targets). ChemPLP (3 targets for both TW and FW rates), AutoDock Vina (3 targets for both TW and FW rates), FRED (2 targets for both TW and FW rates), Glide (1 target for both TW and FW rates) and FlexX (2 targets for both TW and FW rates) had similar performance. Among constrained docking algorithms, HYBRID ranked the 1st

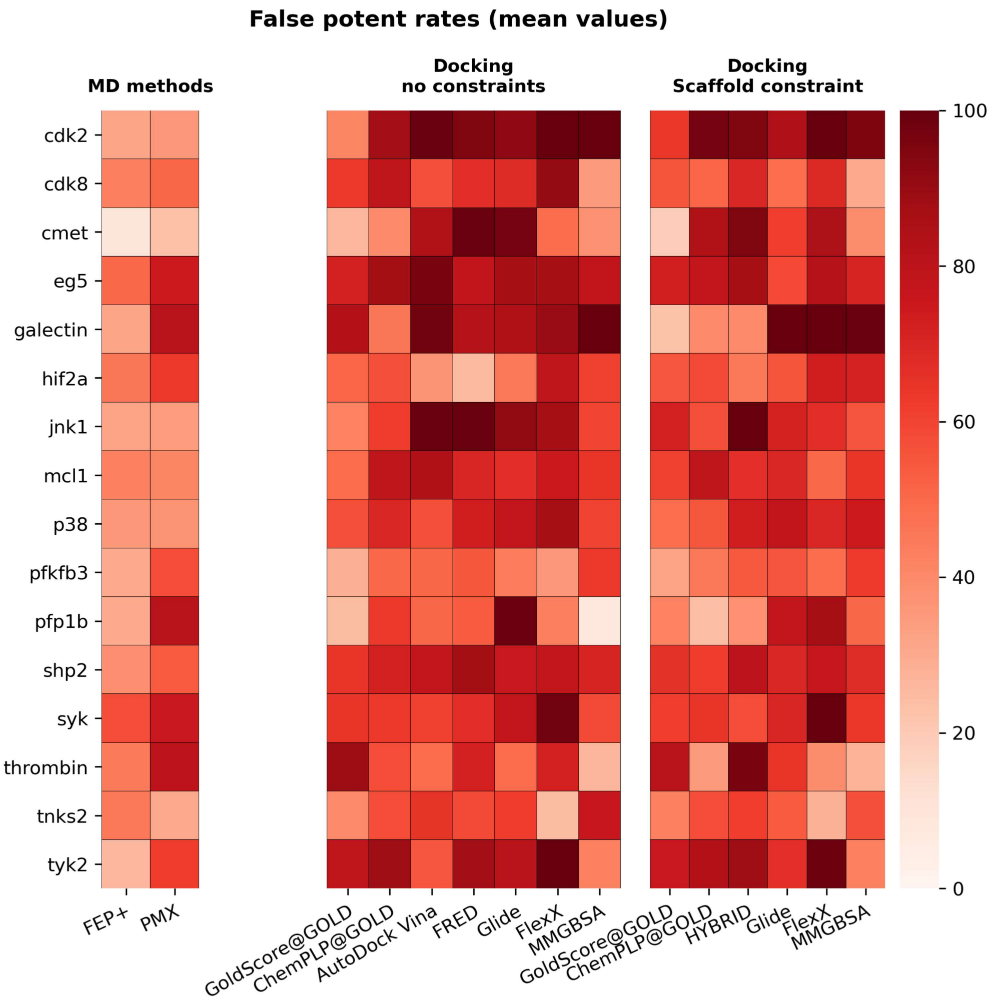

Figure S2: False potent rates (%) for each method across all targets. Mean values of each target after bootstrapping are reported here. MD-based methods have lower false potent rates than docking algorithms although exceptions are also observed.

(5 targets for both TW and FW rates). ChemPLP (4 targets for both TW and FW rates) and GoldScore ranked the 2nd (4 targets for both TW and FW rates) followed by Glide (3 targets for both TW and FW rates). Compared to docking methods, MM/GBSA with non-constrained docking has the best performance in 2 targets for both TW and FW rates and 3 targets for both TW and FW rates with constrained docking.

Table S3: Cases where docking algorithms and MM/GBSA outperform MD-based methods for high level success rates.

non-constrained docking		constrained docking	
AutoDock Vina	thrombin	MM/GBSA	cdk8, thrombin
GoldScore	mcl1		
MM/GBSA	cdk8		

Table S4: Cases where docking algorithms and MM/GBSA outperform MD-based methods for low level success rates.

non-constrained docking		constrained docking	
FRED	hif2a	GoldScore	galectin
GoldScore	mcl1	MM/GBSA	cdk8, thrombin
MM/GBSA	cdk8, thrombin		

References

(1) Hahn, D. F.; Wagner, J. openforcefield/protein-ligand-benchmark: 0.2.0 Addition of new targets. 2021; `https://zenodo.org/record/5679599`.

(2) Schindler, C. E. M. et al. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. *Journal of Chemical Information and Modeling* **2020**, *60*, 5457–5474.

(3) Hahn, D. F.; Gapsys, V. dfhahn/protein-ligand-benchmark-analysis: Release 0.2.0. 2022; `https://zenodo.org/record/6538976`, Version Number: 0.2.0 Type: dataset.

(4) Bruce Macdonald, H. E. Openforcefield/openff-arsenic. Open Force Field Initiative, 2020.

Table S5: Number of targets where each docking algorithm gets the highest success rate (docking algorithms exclusive).

methods	high level	low level
non-constrained docking		
ChemPLP	2	1
GoldScore	7	8
Glide	0	0
FlexX	1	1
FRED	1	1
AutoDock Vina	5	5
constrained docking		
ChemPLP	5	5
GoldScore	4	7
Glide	2	1
FlexX	1	2
HYBRID	4	1

Table S6: Cases where docking algorithms and MM/GBSA outperform MD-based methods for true weak rates.

non-constrained docking		constrained docking	
FRED	tnks2	MM/GBSA	cdk8, mcl1, pfkfb3
GoldScore	pfkfb3, mcl1	ChemPLP	eg5, ptp1b
ChemPLP	eg5		
Autodock Vina	cdk8		

Table S7: Cases where docking algorithms and MM/GBSA outperform MD-based methods for false weak rates.

non-constrained docking		constrained docking	
FRED	tnks2	ChemPLP	eg5, ptp1b, tnks2
GoldScore	pfkfb3, mcl1	MM/GBSA	pfkfb3, mcl1, cdk8
ChemPLP	eg5	GoldScore	thrombin
Autodock Vina	cdk8		

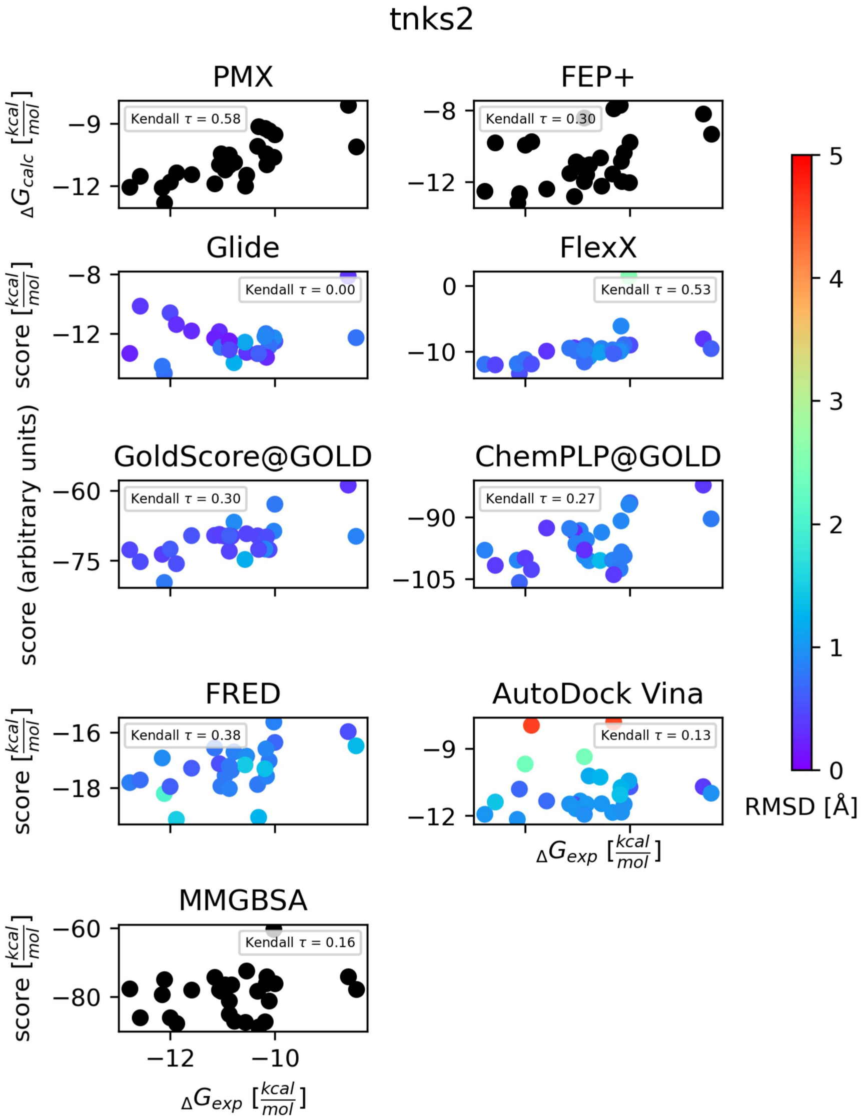

Figure S3: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for tnks2. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). PMX gets the highest Kendall τ value among studied methods. FlexX and FRED get higher Kendall τ values than one MD method (FEP+) in tnks2.

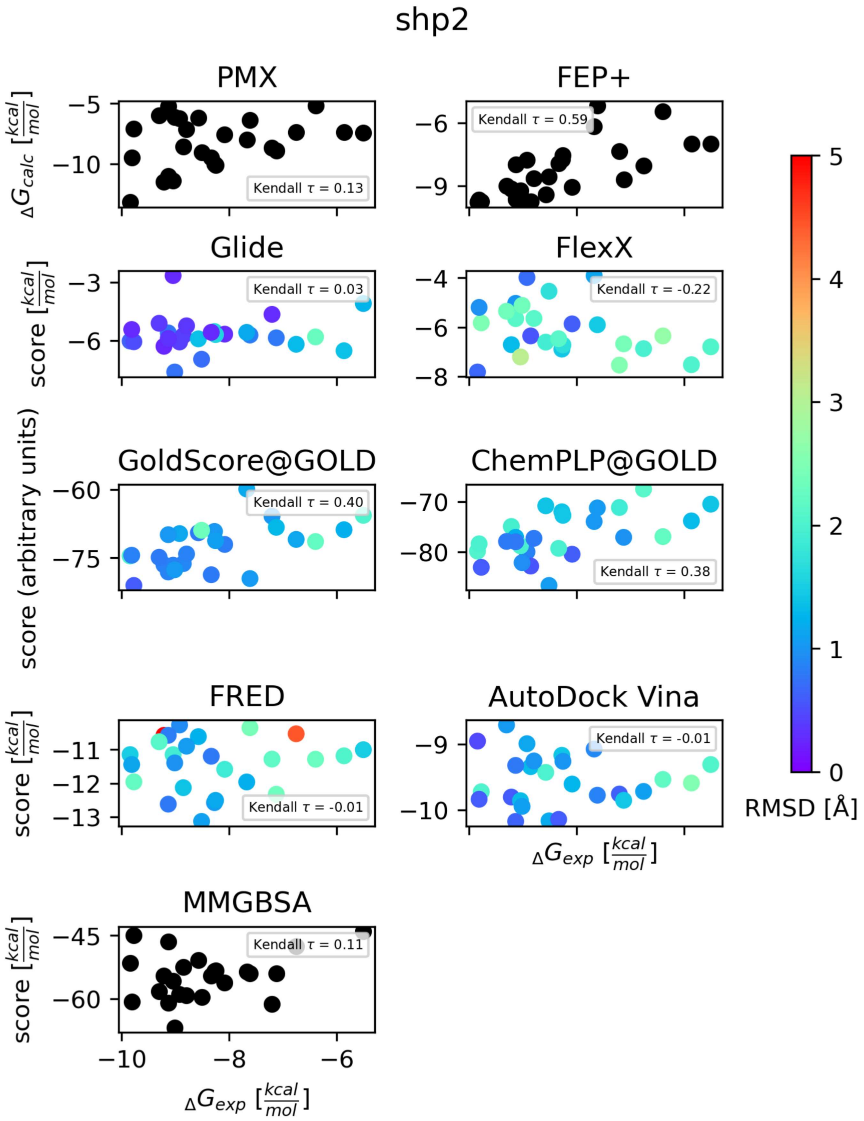

Figure S4: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for shp2. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). FEP+ gets the highest Kendall τ value among studied methods. GoldScore@GOLD and ChemPLP@GOLD get higher Kendall τ values than one MD method (PMX) in shp2.

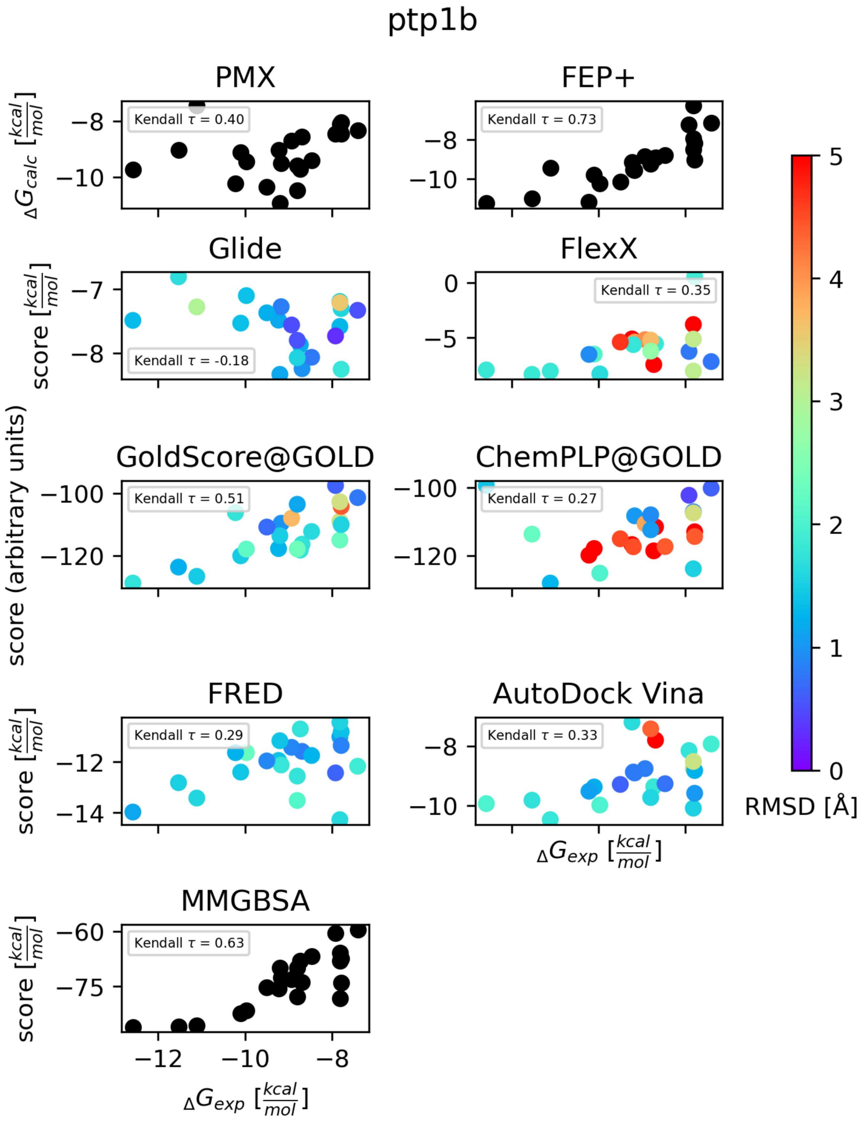

Figure S5: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for ptp1b. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). FEP+ gets the highest Kendall τ value among studied methods. GoldScore@GOLD and MM/GBSA get higher Kendall τ values than one MD method (PMX) in ptp1b.

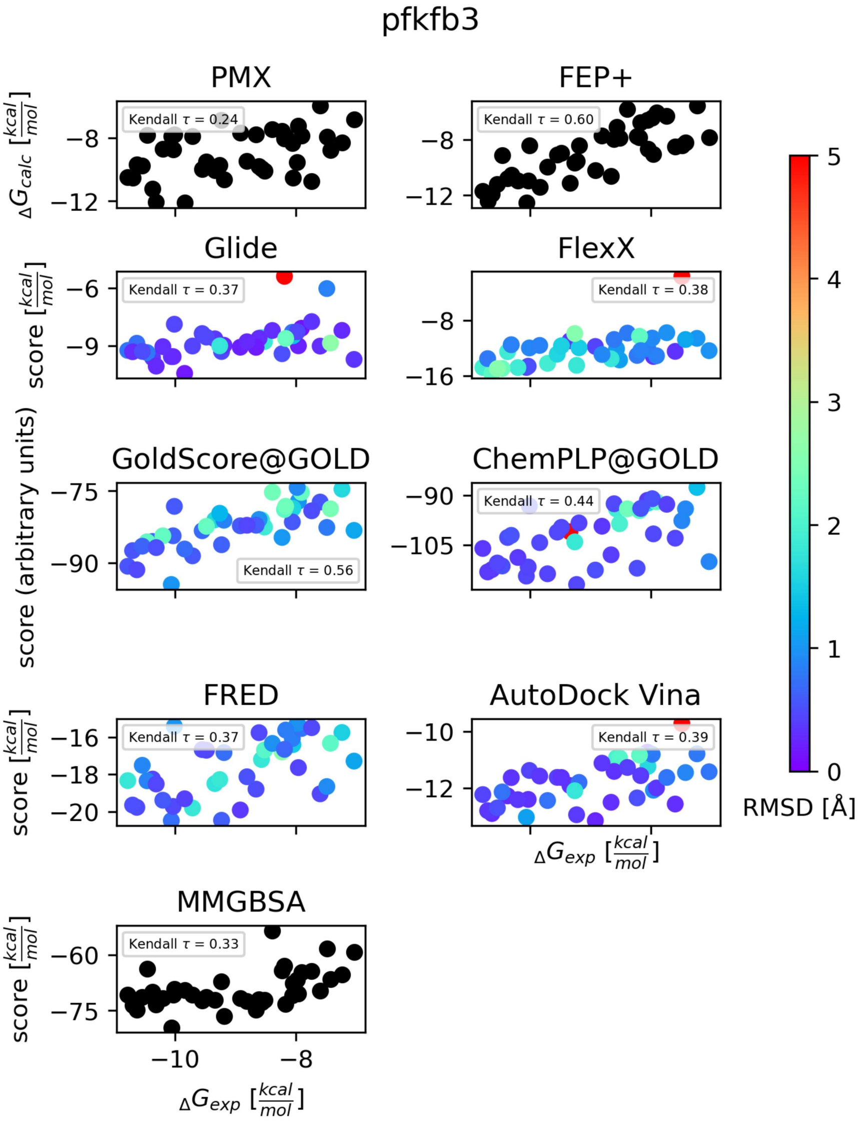

Figure S6: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for pfkfb3. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). FEP+ gets the highest Kendall τ value among studied methods. All Docking algorithms and MM/GBSA get higher Kendall τ values than one MD method (PMX) in pfkfb3.

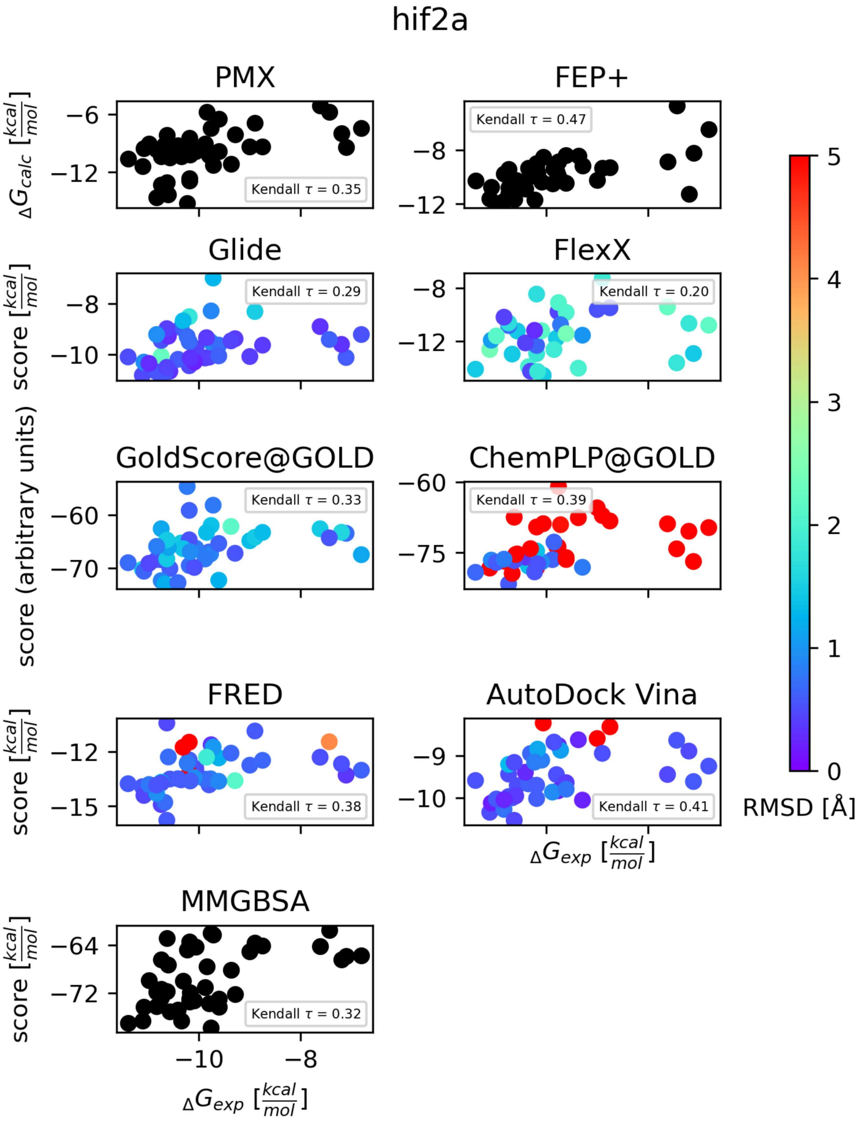

Figure S7: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for hif2a. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). FEP+ gets the highest Kendall τ value among studied methods. ChemPLP@GOLD, FRED and AutoDock Vina get higher Kendall τ values than one MD method (PMX) in hif2a.

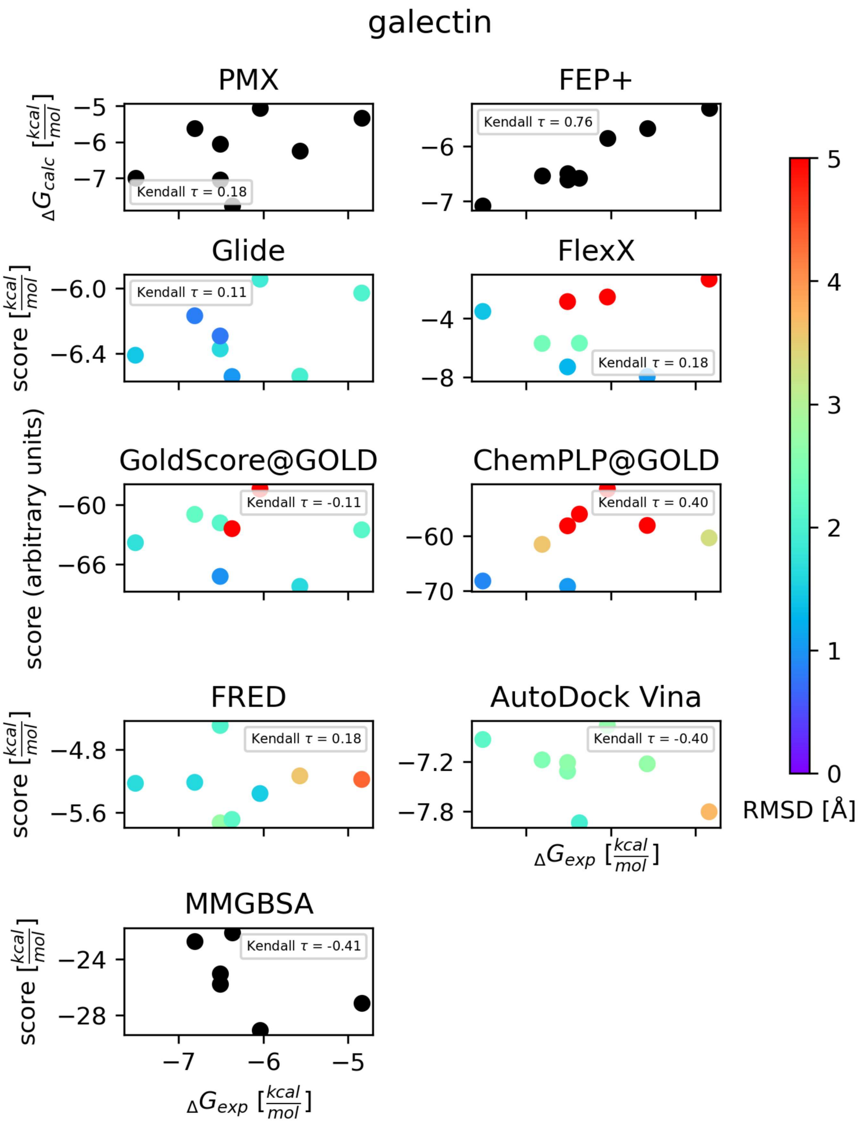

Figure S8: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for galectin. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). FEP+ gets the highest Kendall τ value among studied methods. ChemPLP@GOLD gets a higher Kendall τ value than one MD method (PMX) in galectin.

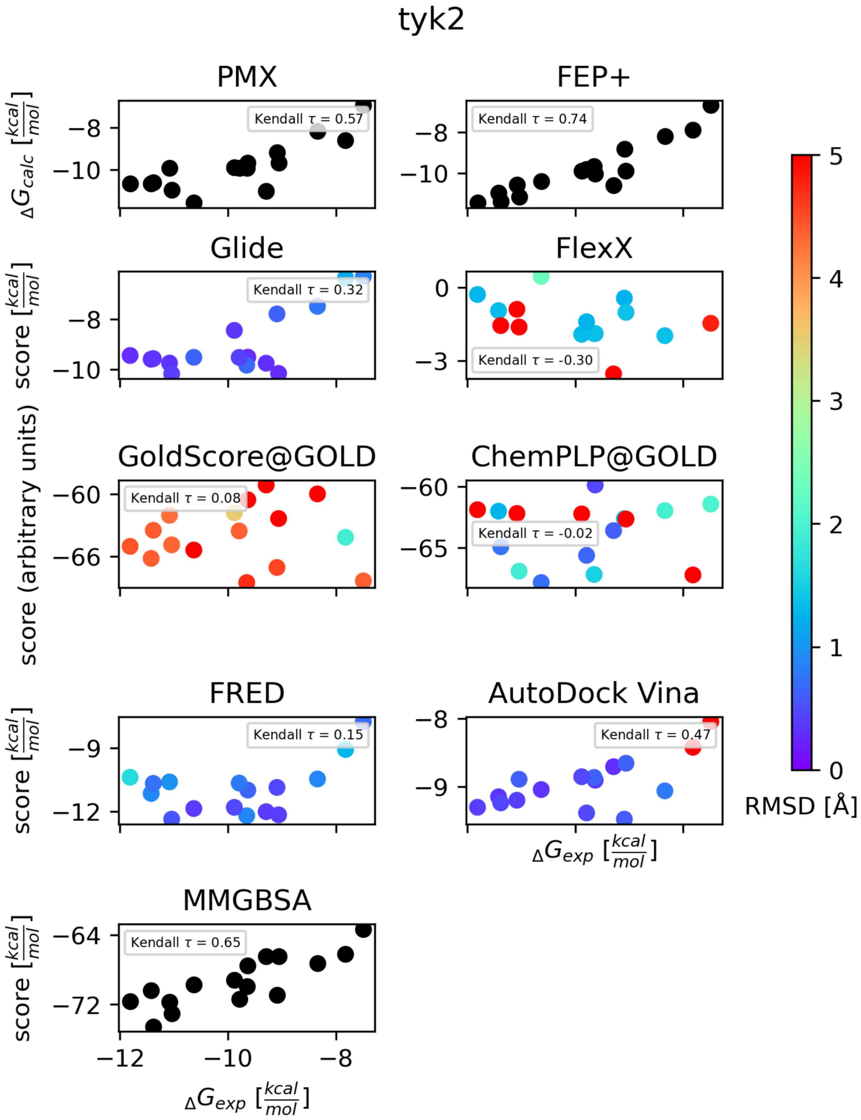

Figure S9: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for tyk2. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). FEP+ gets the highest Kendall τ value among studied methods. MM/GBSA gets a higher Kendall τ than one MD method (PMX) in tyk2.

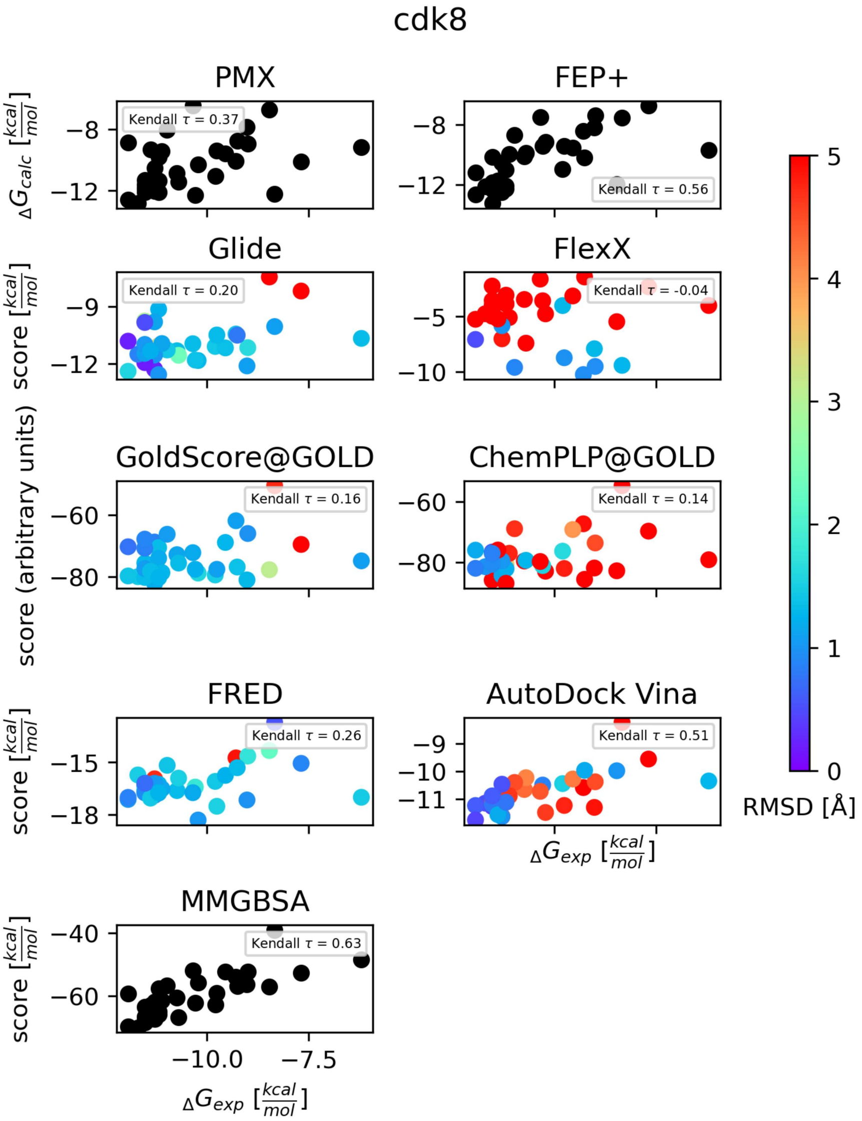

Figure S10: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for cdk8. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MM/GBSA gets the highest Kendall τ value among studied methods. AutoDock Vina gets a higher Kendall τ value than one MD method (PMX) in cdk8.

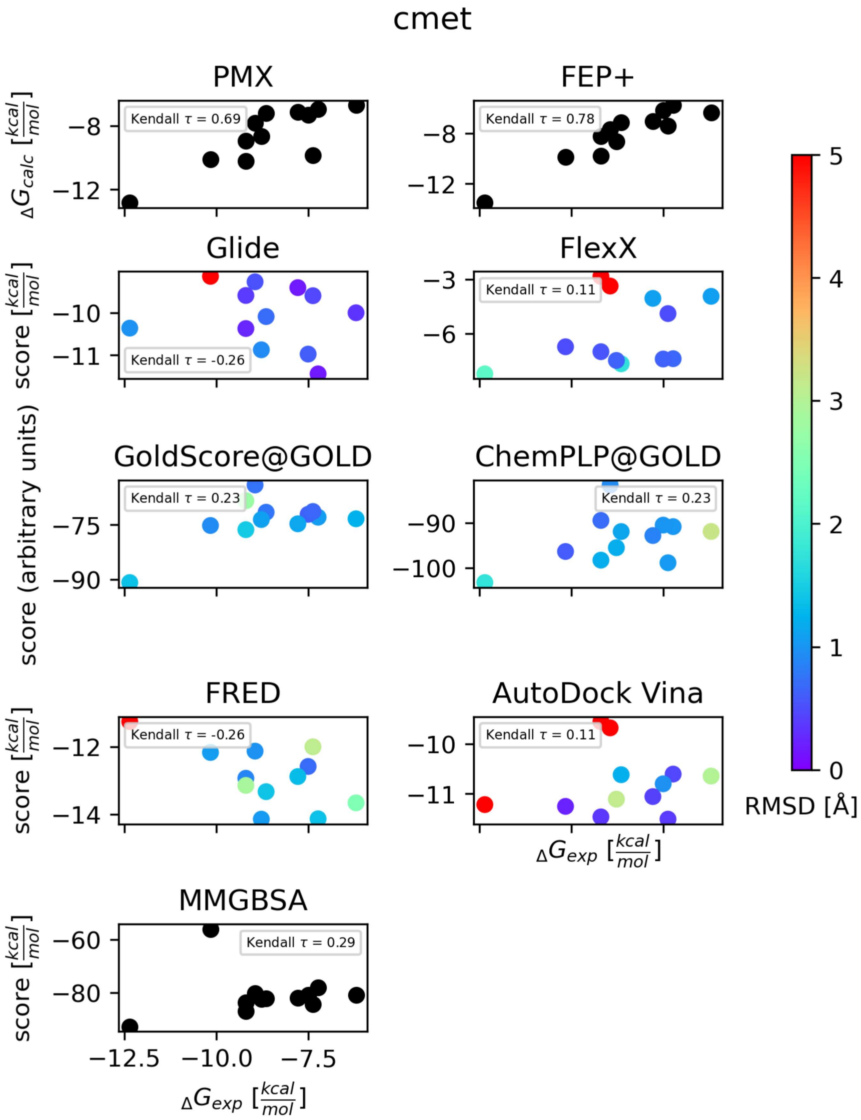

Figure S11: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for cmet. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MD-based methods and MM/GBSA get higher Kendall τ values than docking algorithms and MM/GBSA calculations.

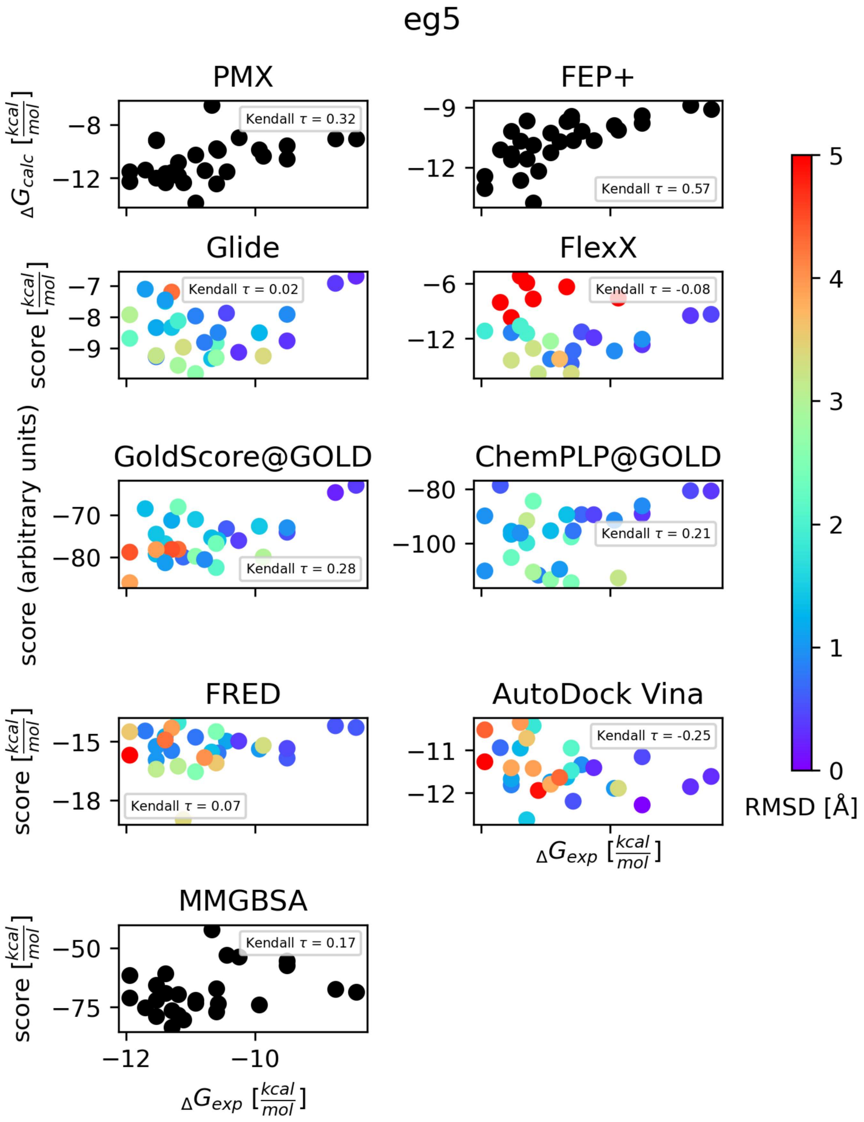

Figure S12: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for eg5. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MD-based methods get higher Kendall τ values than docking algorithms and MM/GBSA calculations.

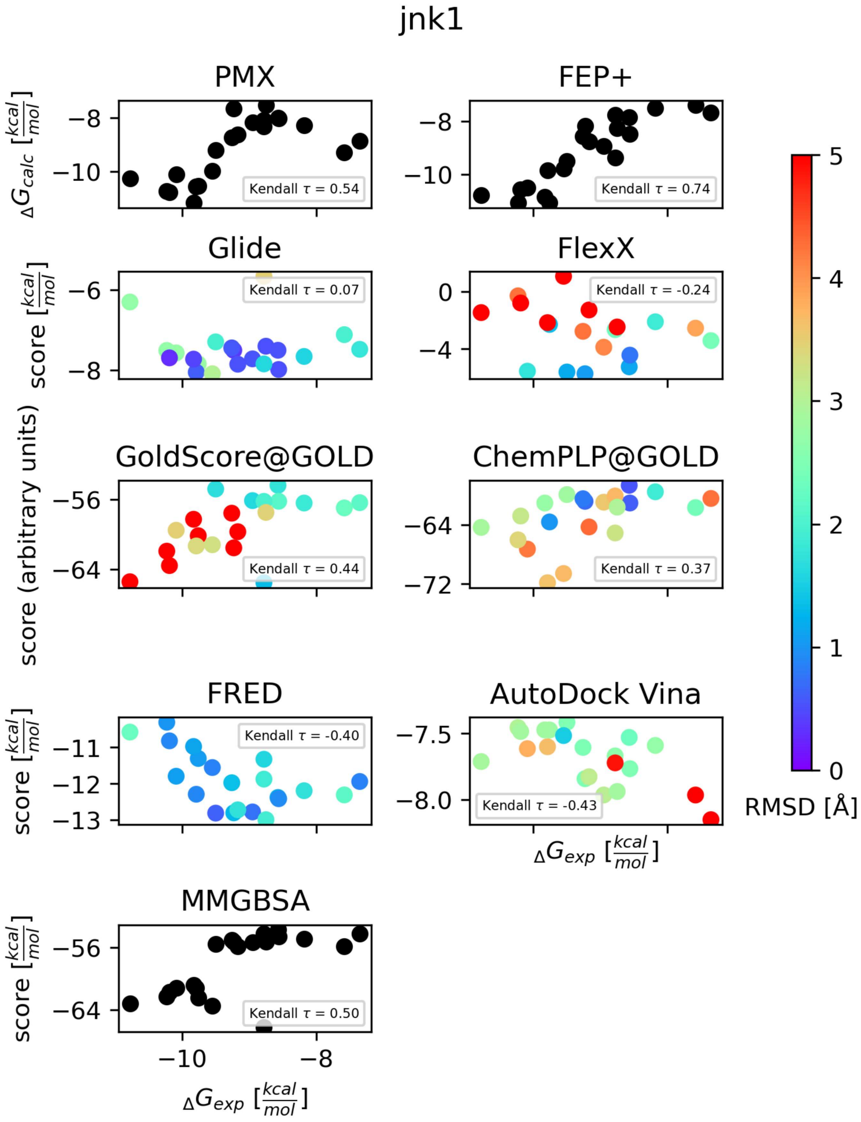

Figure S13: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for jnk1. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MD-based methods and MM/GBSA get higher Kendall τ values than docking algorithms and MM/GBSA calculations.

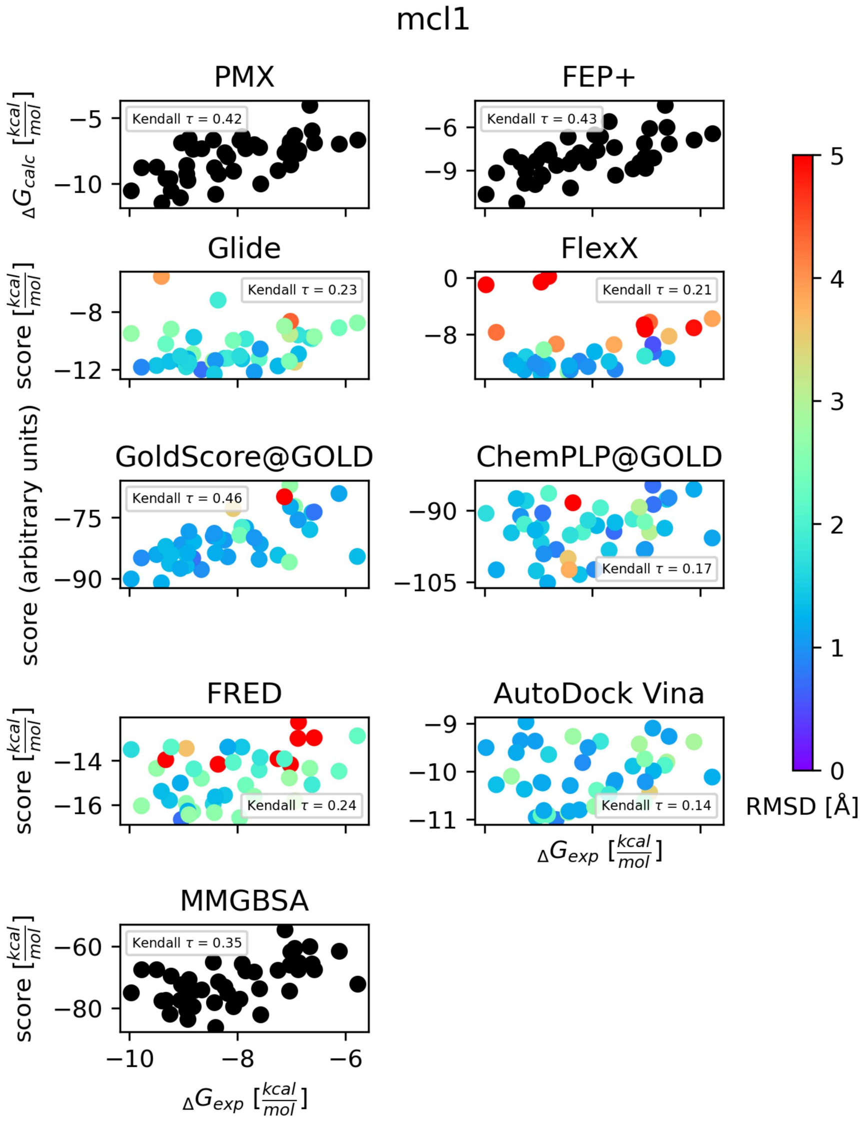

Figure S14: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for mcl1. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). GoldScore@GOLD gets the highest Kendall τ value among studied methods.

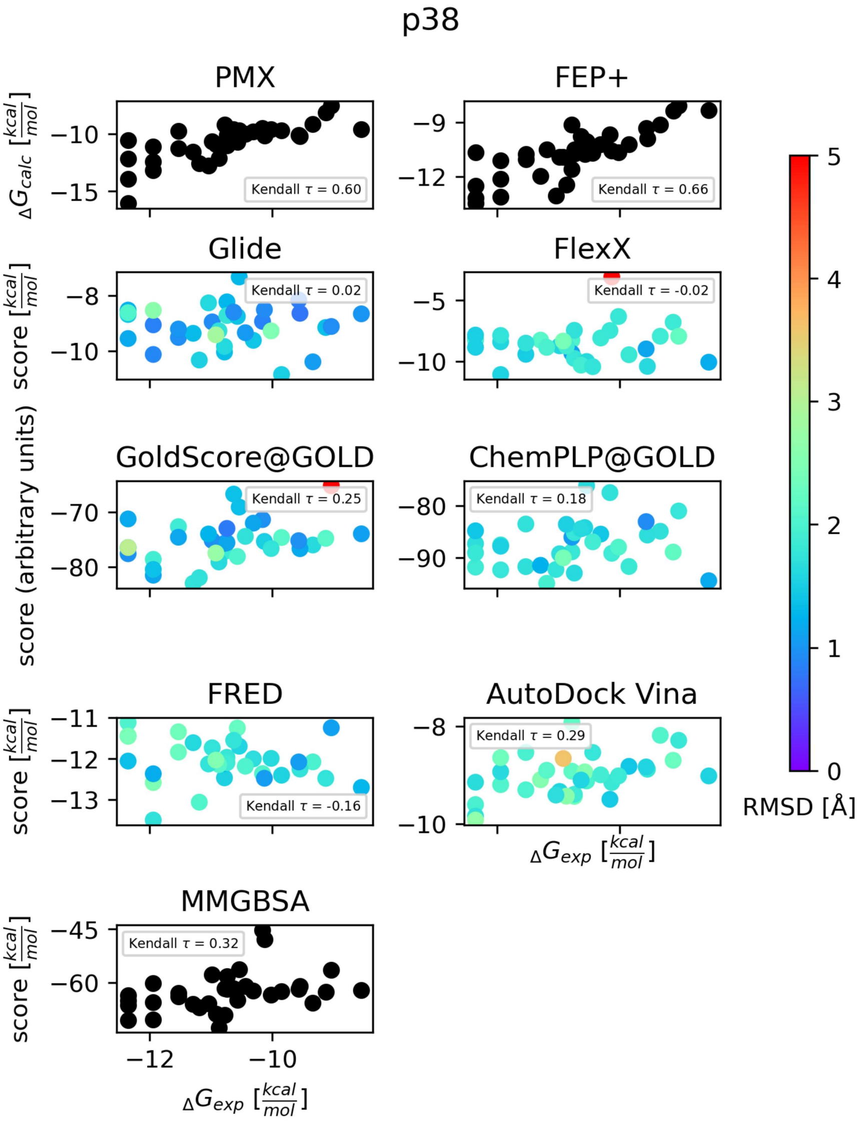

Figure S15: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for p38. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MD-based methods get higher Kendall τ values than docking algorithms and MM/GBSA calculations. MM/GBSA gets a higher Kendall τ value than docking algorithms.

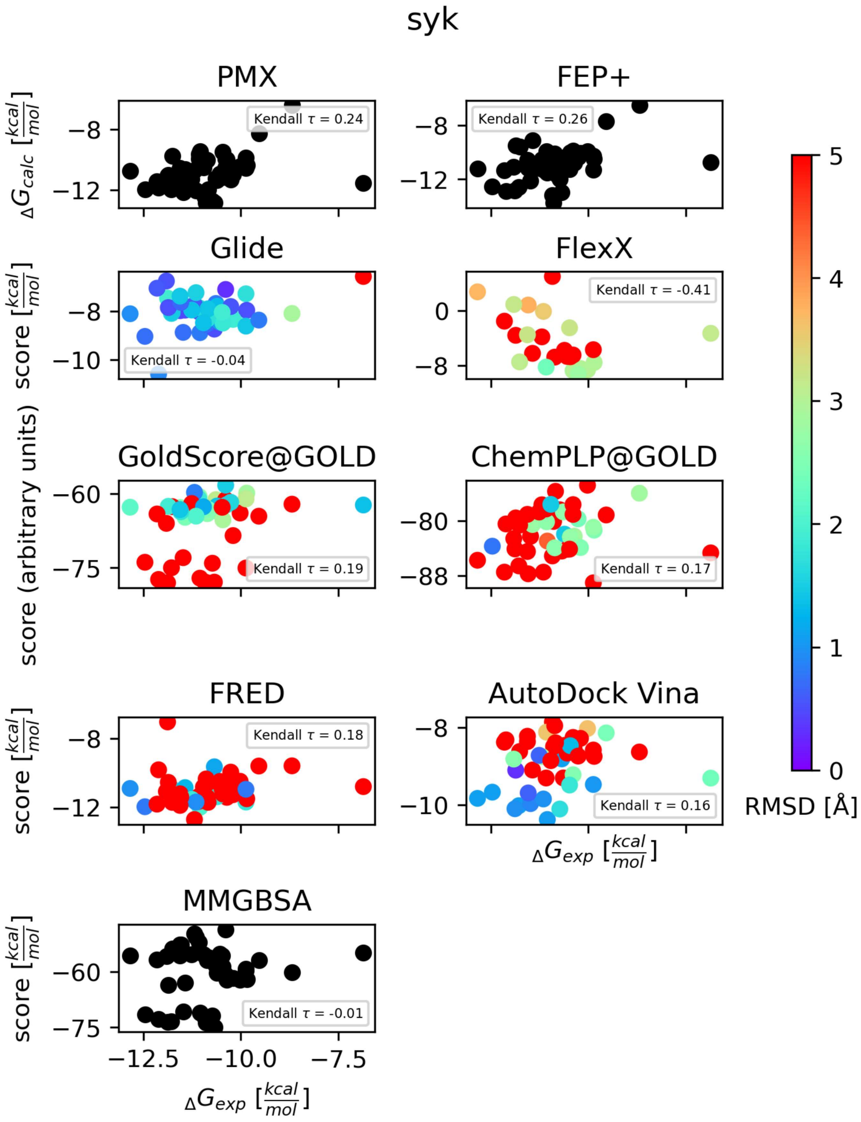

Figure S16: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for syk. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MD-based methods get higher Kendall τ values than docking algorithms and MM/GBSA calculations.

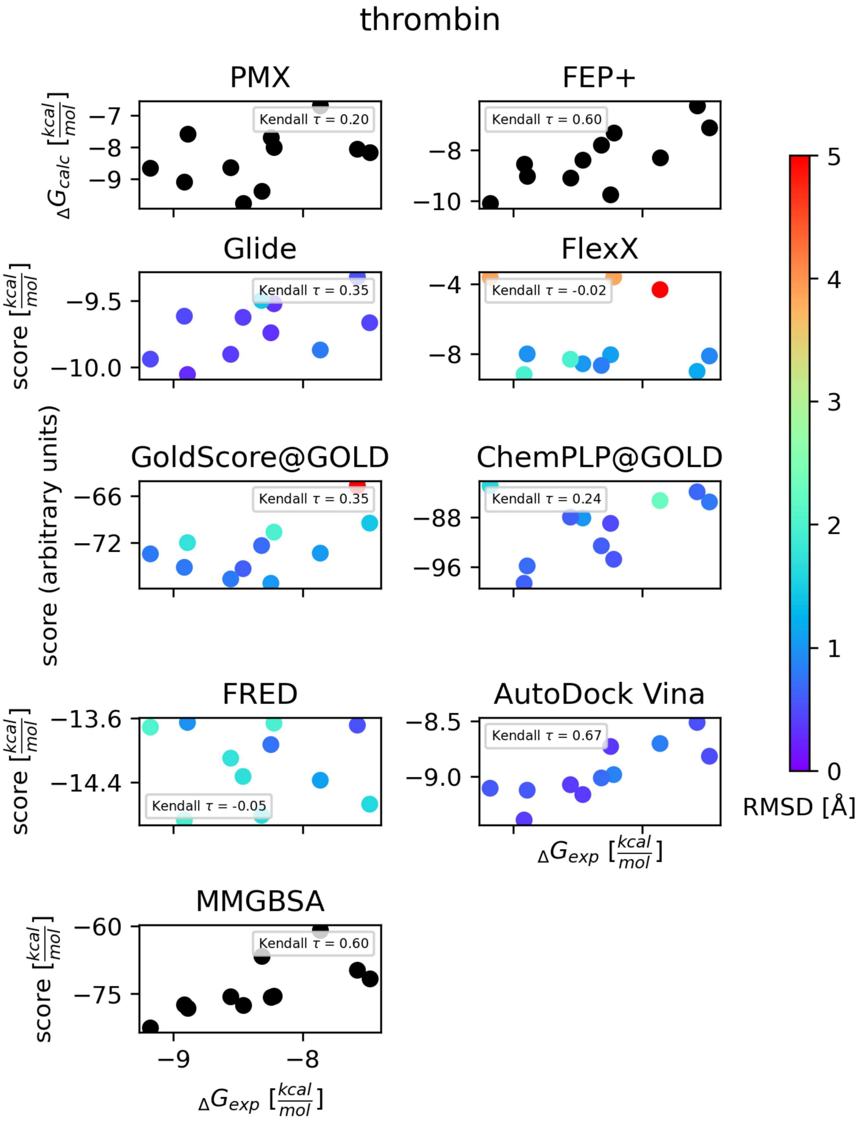

Figure S17: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for thrombin. We showed results of non-constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). AutoDock Vina gets the highest Kendall τ value among studied methods. Glide, GoldScore@GOLD, ChemPLP@GOLD and MM/GBSA get higher Kendall τ values than one MD method (PMX) in thrombin.

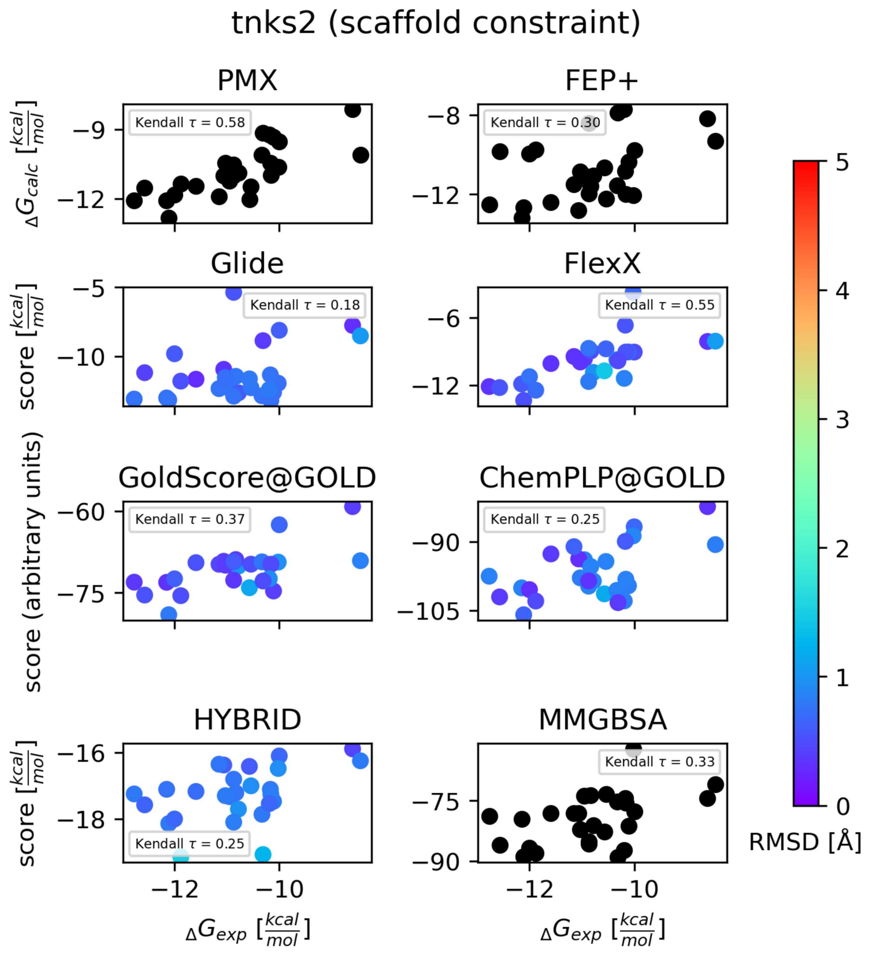

Figure S18: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for tnks2. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). PMX gets the highest Kendall τ value among studied methods. FlexX, GoldScore@GOLD and MM/GBSA get higher Kendall τ values than one MD method (FEP+) in tnks2.

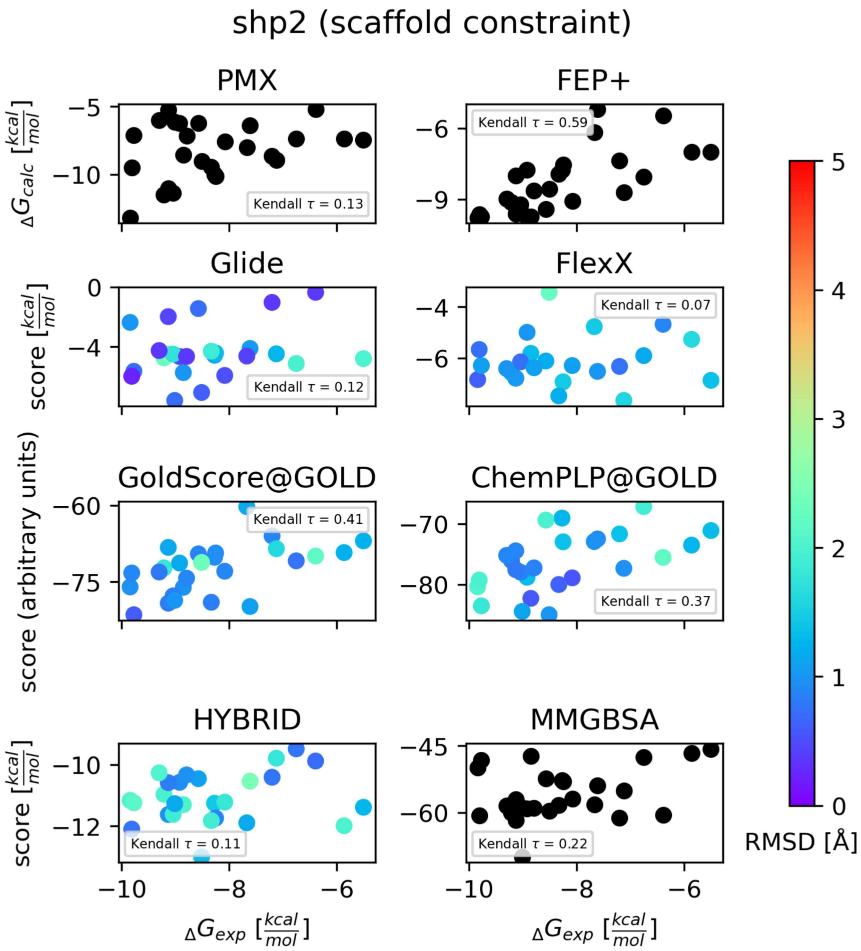

Figure S19: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for shp2. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). FEP+ gets the highest Kendall τ value among studied methods. GoldScore@GOLD, ChemPLP@GOLD and MM/GBSA get higher Kendall τ values than one MD method (PMX) in shp2.

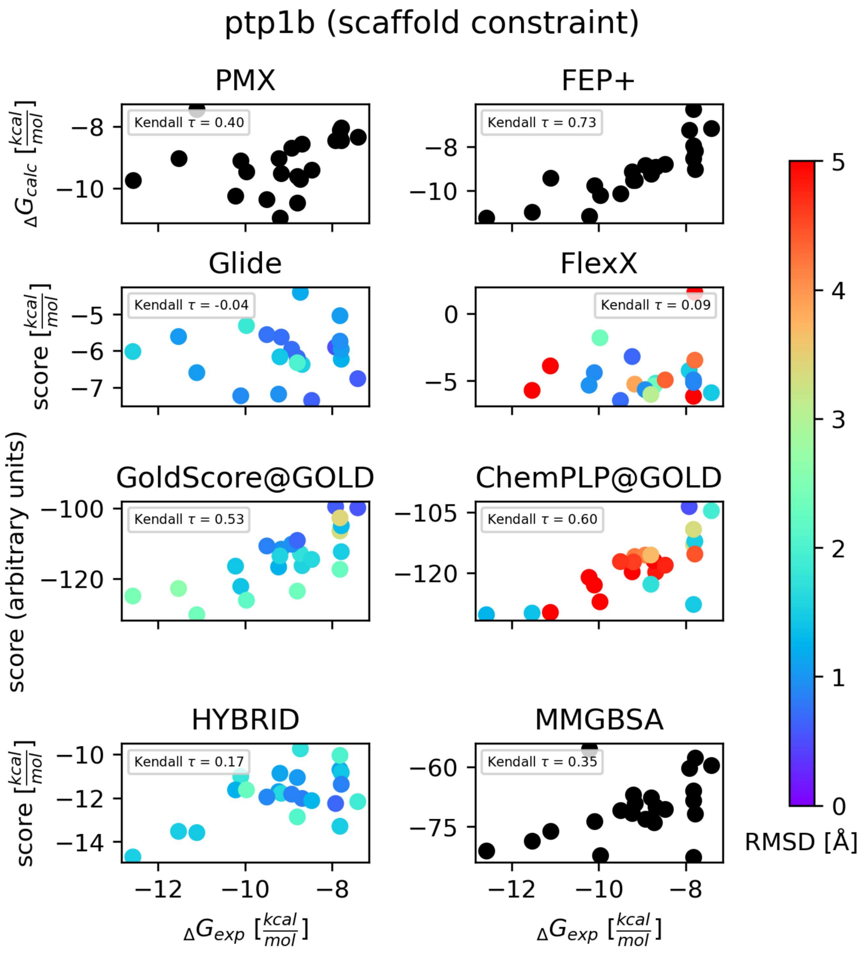

Figure S20: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for ptp1b. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). FEP+ gets the highest Kendall τ value among studied methods. GoldScore@GOLD and ChemPLP@GOLD get higher Kendall τ values than one MD method (PMX) in ptp1b.

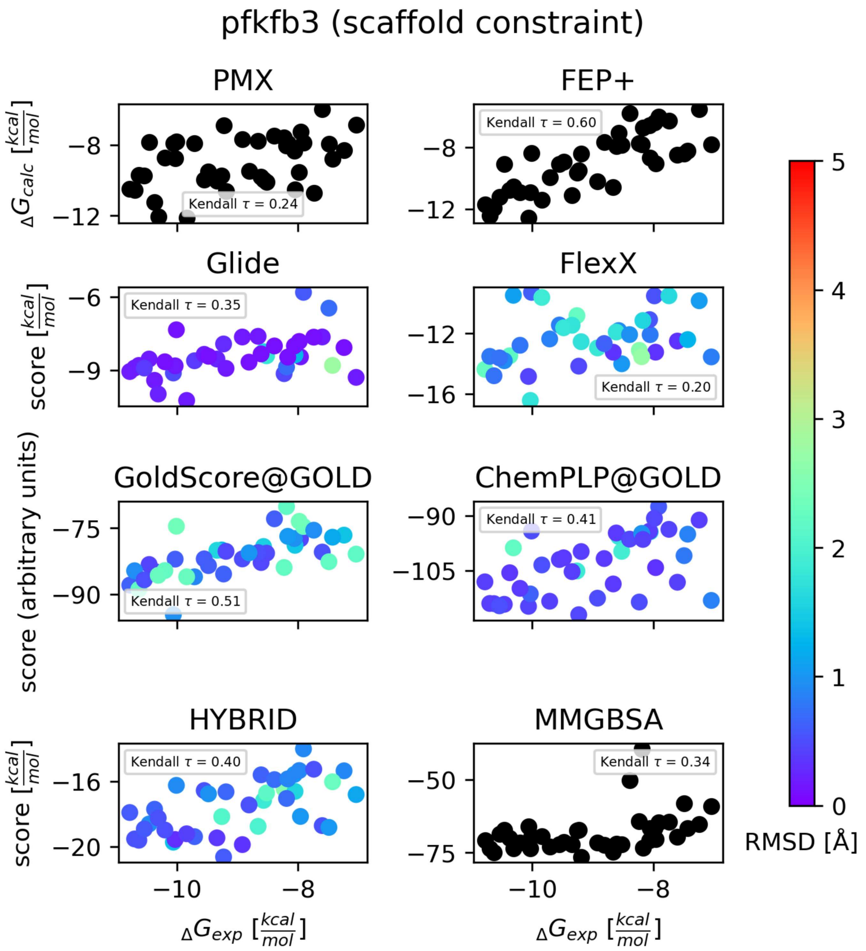

Figure S21: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for pfkfb3. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). FEP+ gets the highest Kendall τ value among studied methods. Glide, GoldScore@GOLD, ChemPLP@GOLD, HYBRID and MM/GBSA get higher Kendall τ values than one MD method (PMX) in pfkfb3.

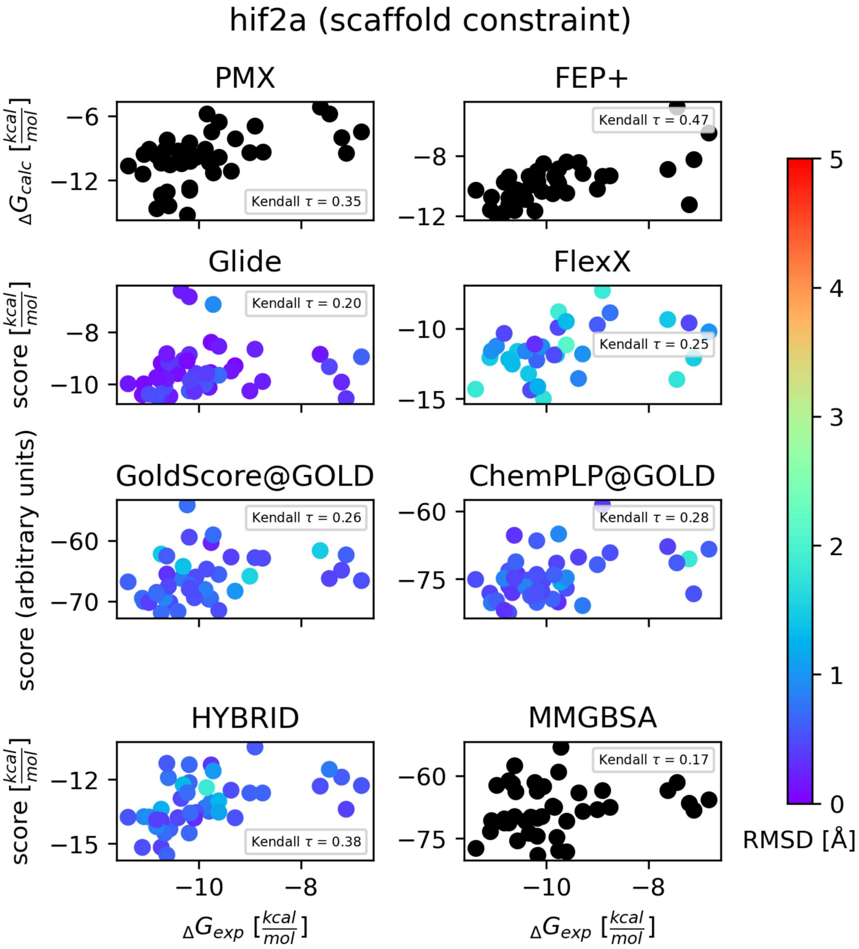

Figure S22: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for hif2a. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). FEP+ gets the highest Kendall τ value among studied methods. HYBRID gets a higher Kendall τ value than one MD method (PMX) in hif2a.

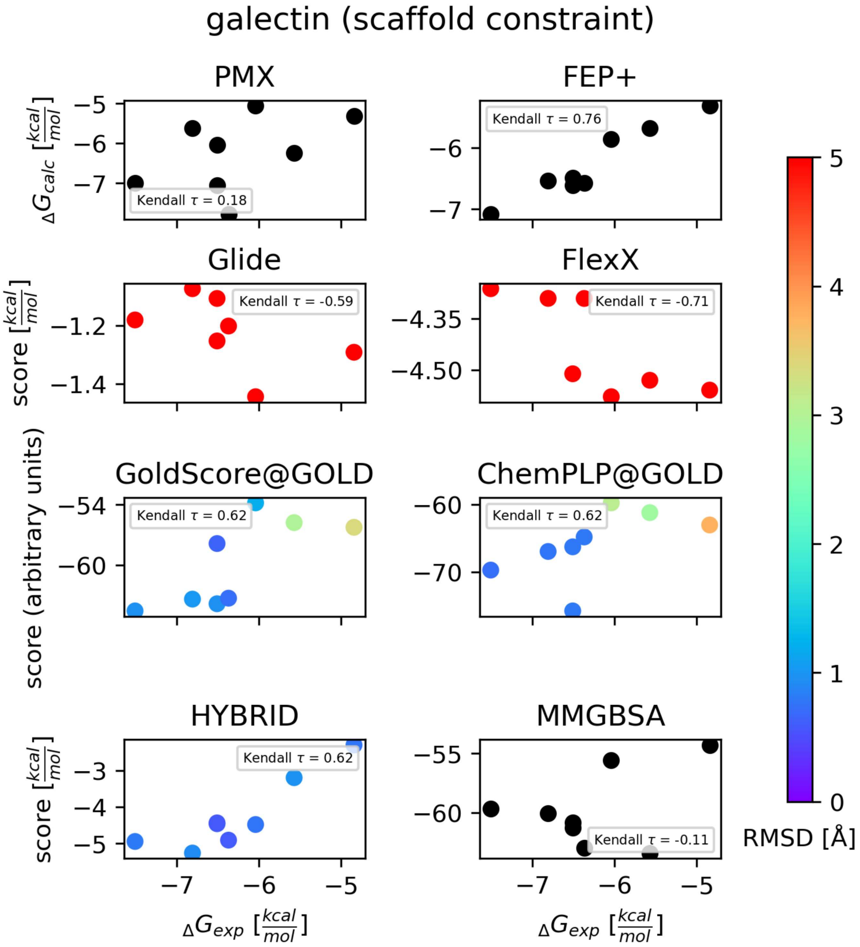

Figure S23: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for galectin. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). FEP+ gets the highest Kendall τ value among studied methods. The GoldScore@GOLD, ChemPLP@GOLD and HYBRID get higher Kendall τ values than one MD method (PMX) in galectin.

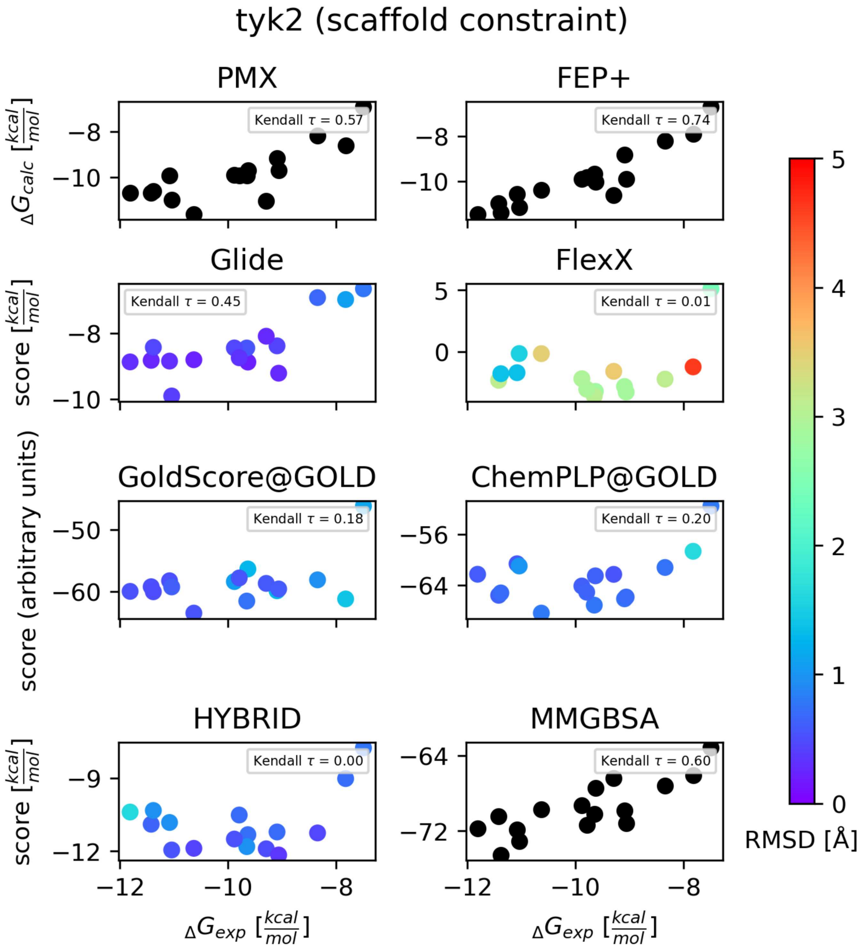

Figure S24: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for tyk2. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). FEP+ gets the highest Kendall τ value among studied methods. MM/GBSA gets a higher Kendall τ value than one MD method (PMX).

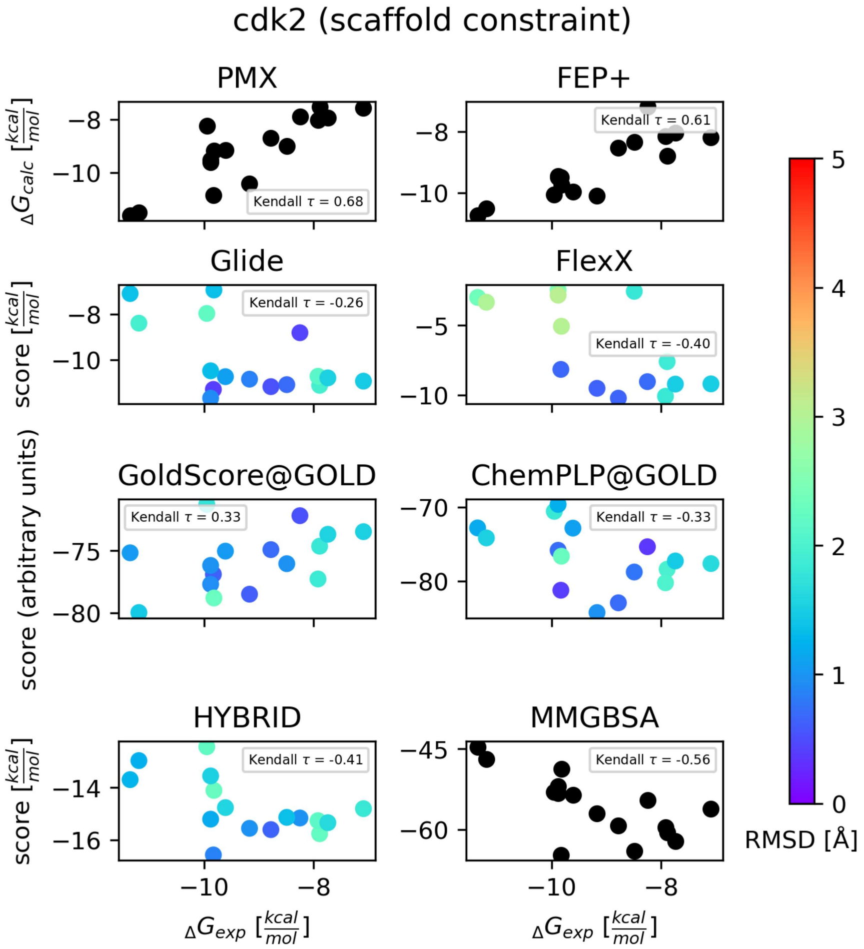

Figure S25: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for cdk2. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MD-based methods get higher Kendall τ values than docking algorithms and MM/GBSA calculations.

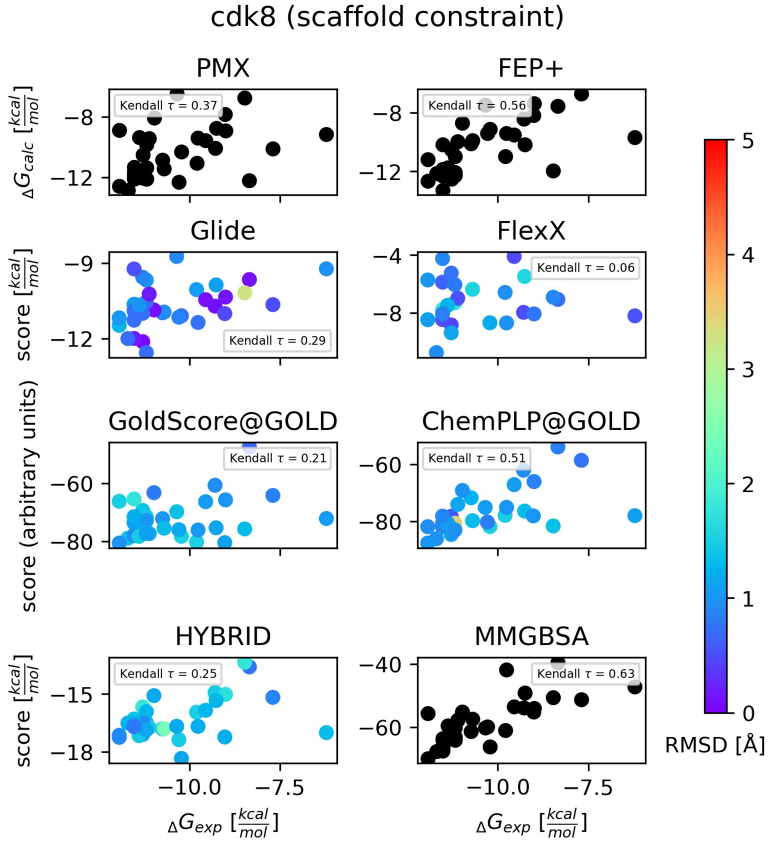

Figure S26: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for cdk8. Results for constrained docking algorithms are shown here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MM/GBSA has the highest Kendall τ value among all methods. The Docking algorithm (ChemPLP@GOLD) gets a higher Kendall τ value than one MD method (PMX) in cdk8. FEP+ and PMX results were retrieved from Ref. 2 and 3, respectively. The relative free energy differences were converted into free energy differences with Arsenic.[4]

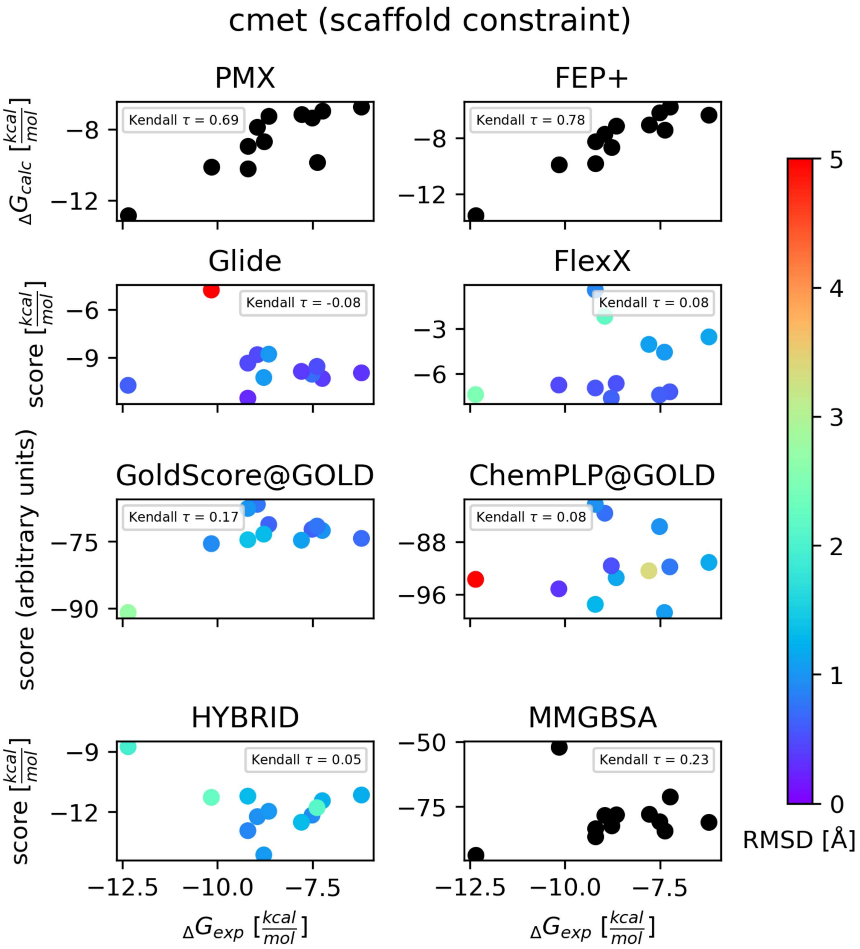

Figure S27: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for cmet. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MD-based methods get higher Kendall τ values than docking algorithms and MM/GBSA calculations. MM/GBSA gets a high Kendall τ than all docking algorithms.

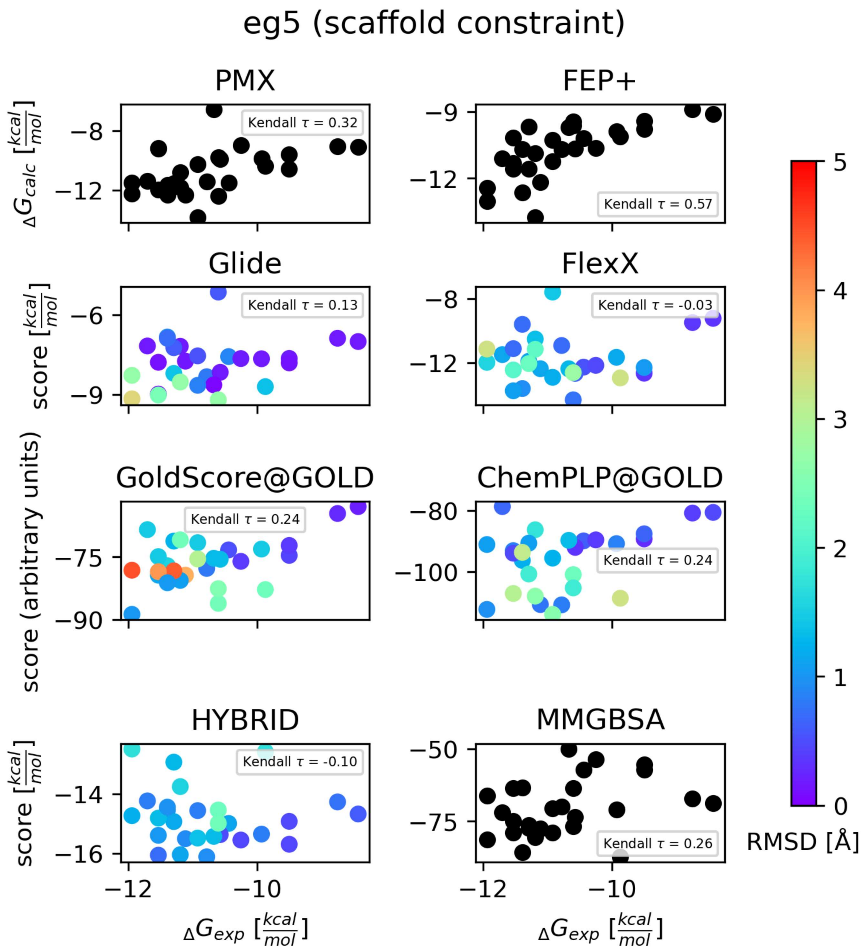

Figure S28: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for eg5. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MD-based methods get higher Kendall τ values than docking algorithms and MM/GBSA calculations. MM/GBSA gets a higher Kendall τ value than all docking algorithms.

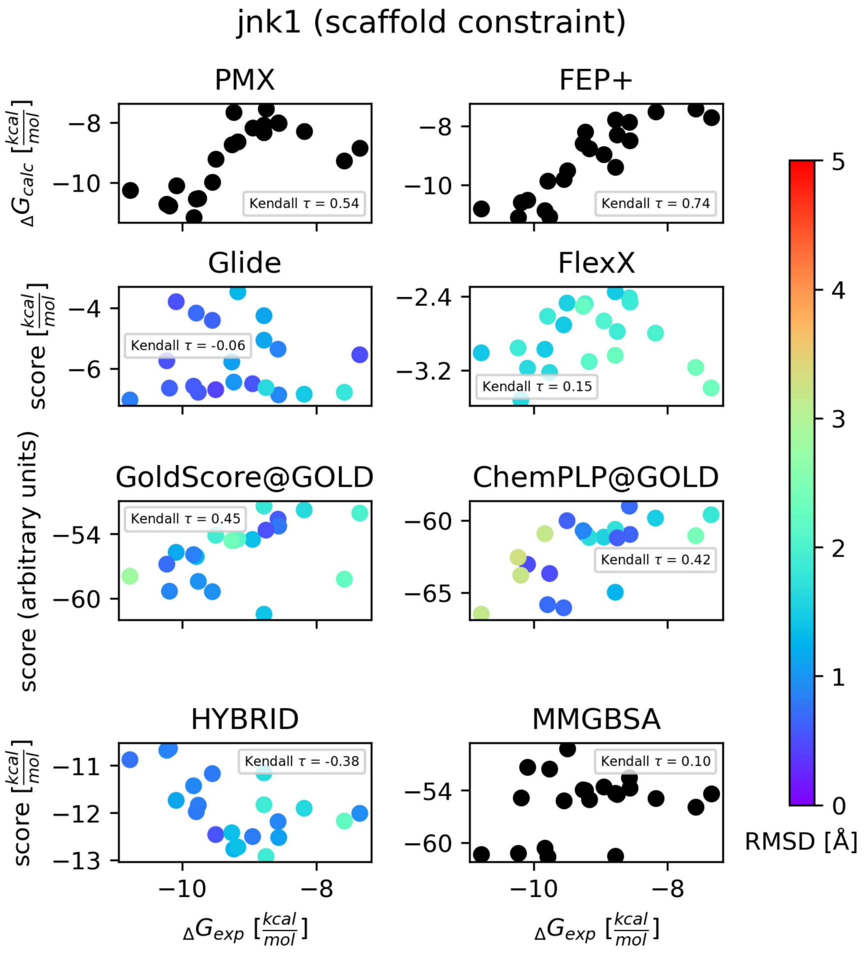

Figure S29: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for jnk1. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MD-based methods get higher Kendall τ values than docking algorithms.

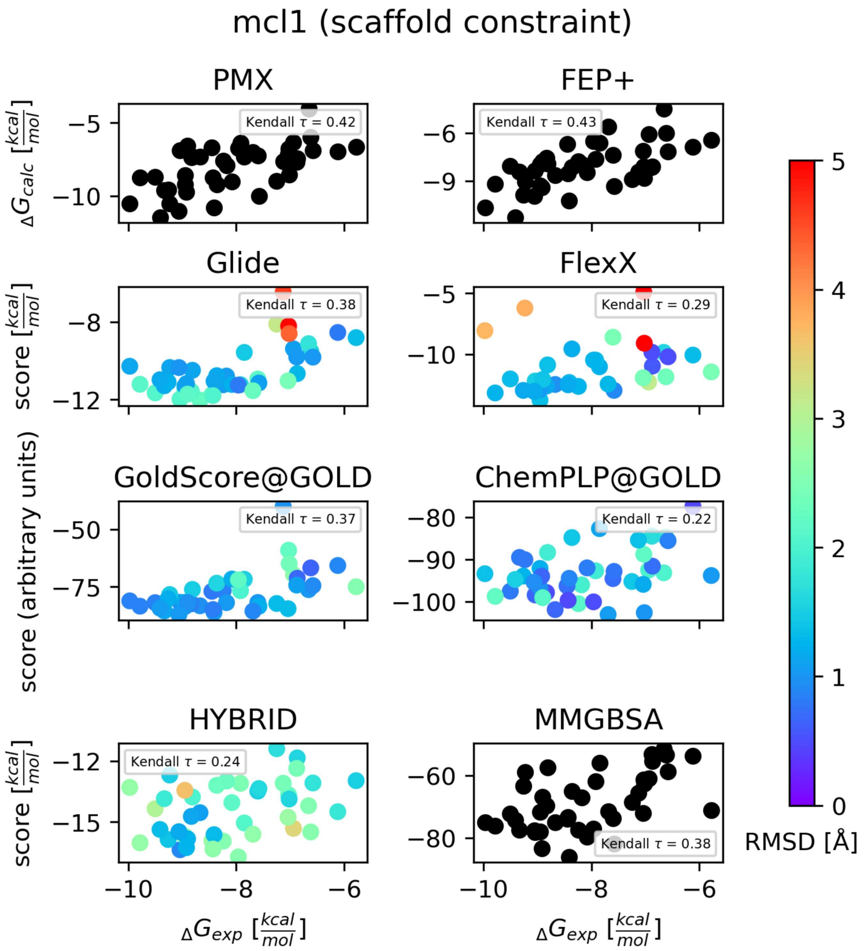

Figure S30: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for mcl1. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MD-based methods get higher Kendall τ values than docking algorithms and MM/GBSA calculations.

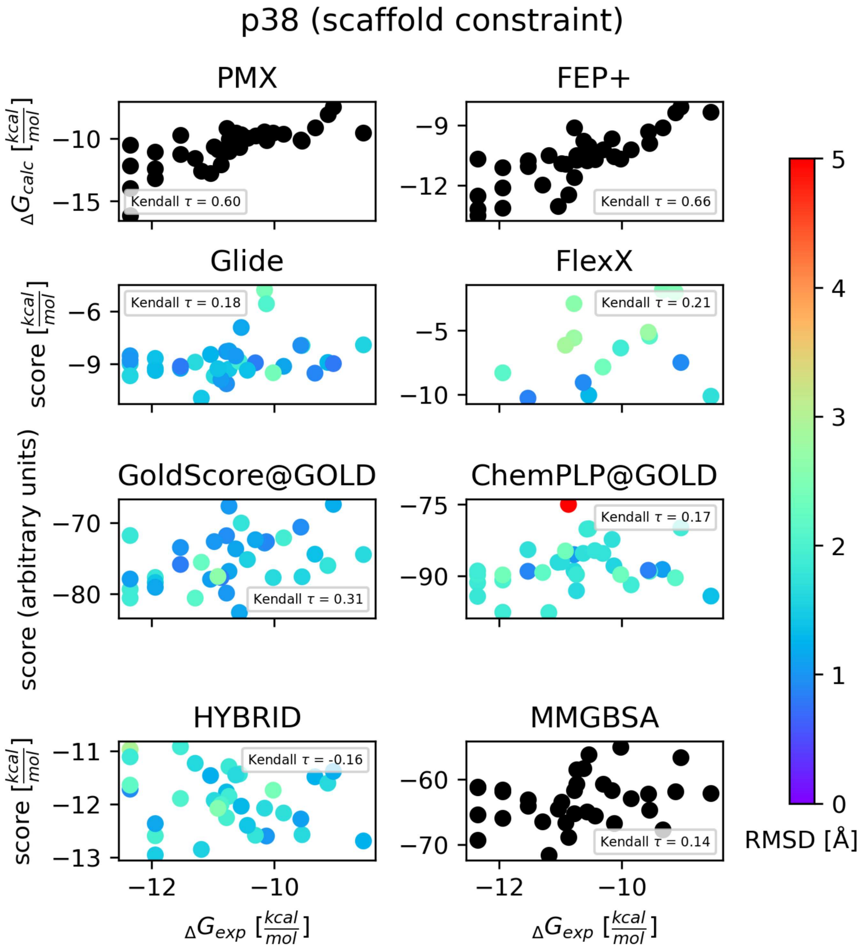

Figure S31: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for p38. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MD-based methods get higher Kendall τ values than docking algorithms and MM/GBSA calculations.

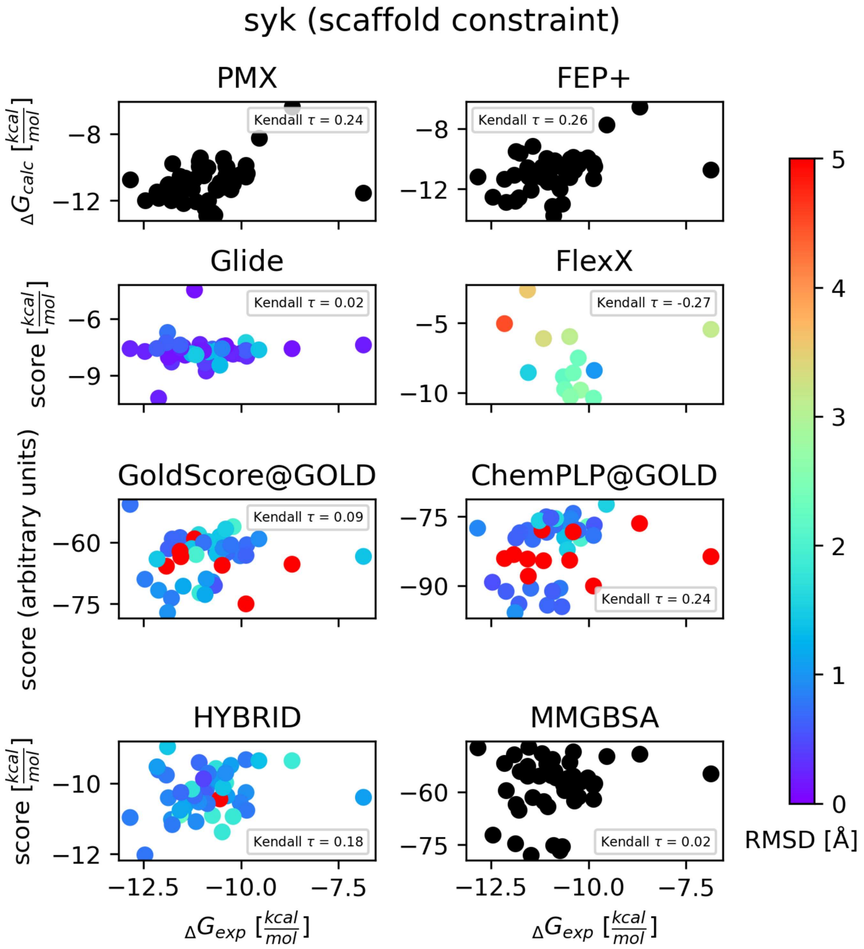

Figure S32: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for syk. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MD-based methods get higher Kendall τ values than docking algorithms and MM/GBSA calculations.

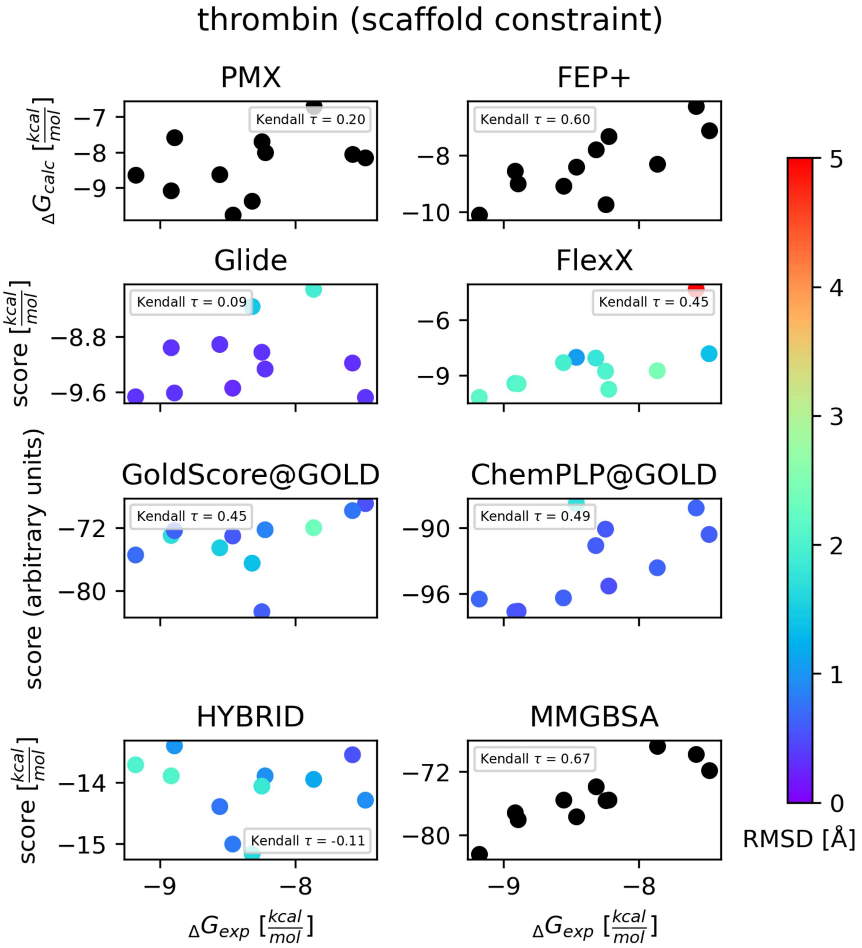

Figure S33: Scatter plot of predicted binding free energies/docking scores vs. experimental binding free energies for thrombin. We showed results of constrained docking algorithms here. The color map shows RMSD (Å) of each compound (we did not compute RMSD values for MD-based methods (PMX, FEP+) and MM/GBSA calculations since it is not applicable to these methods). MM/GBSA gets the highest Kendall τ value among studied methods. FlexX, GoldScore@GOLD and ChemPLP@GOLD get higher Kendall τ values than one MD method (PMX) in thrombin.

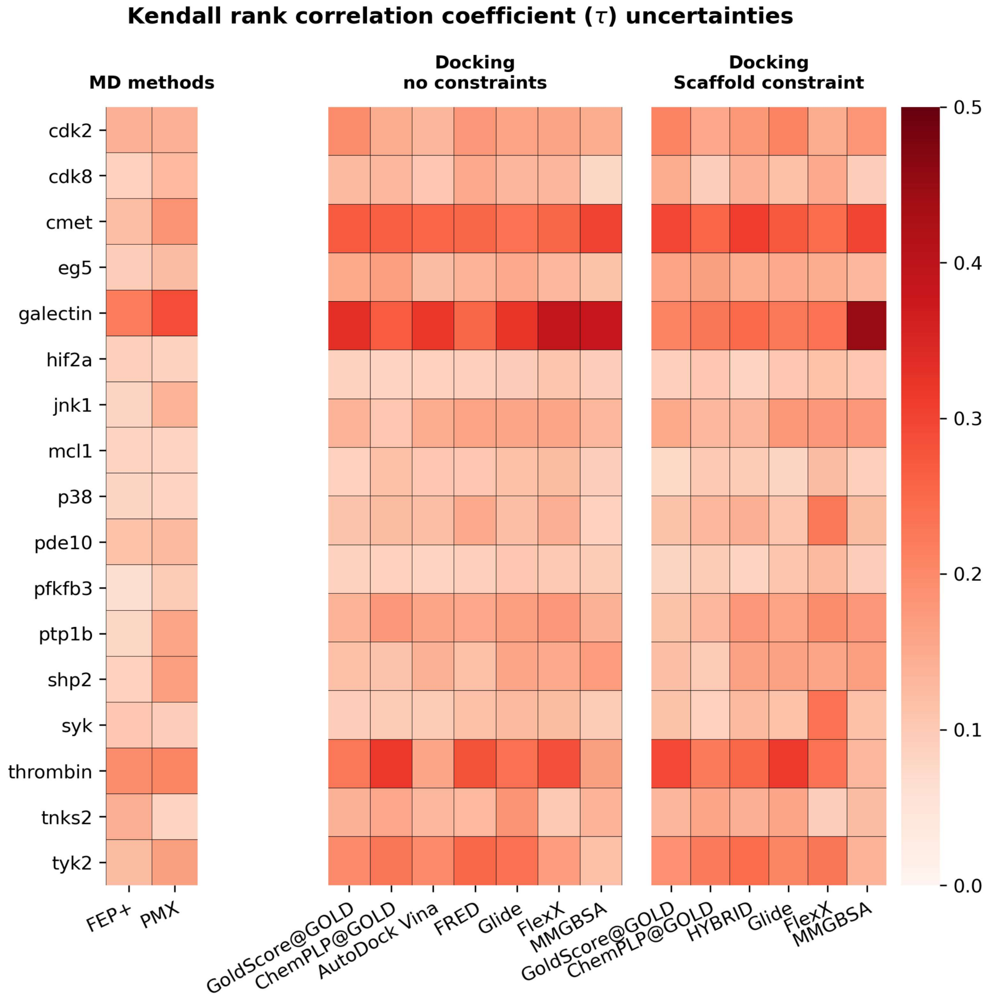

Figure S34: Uncertainties of Kendall rank correlation coefficient (τ) for all docking algorithms across all targets. Standard deviations after 10000 rounds bootstrapping are reported here.

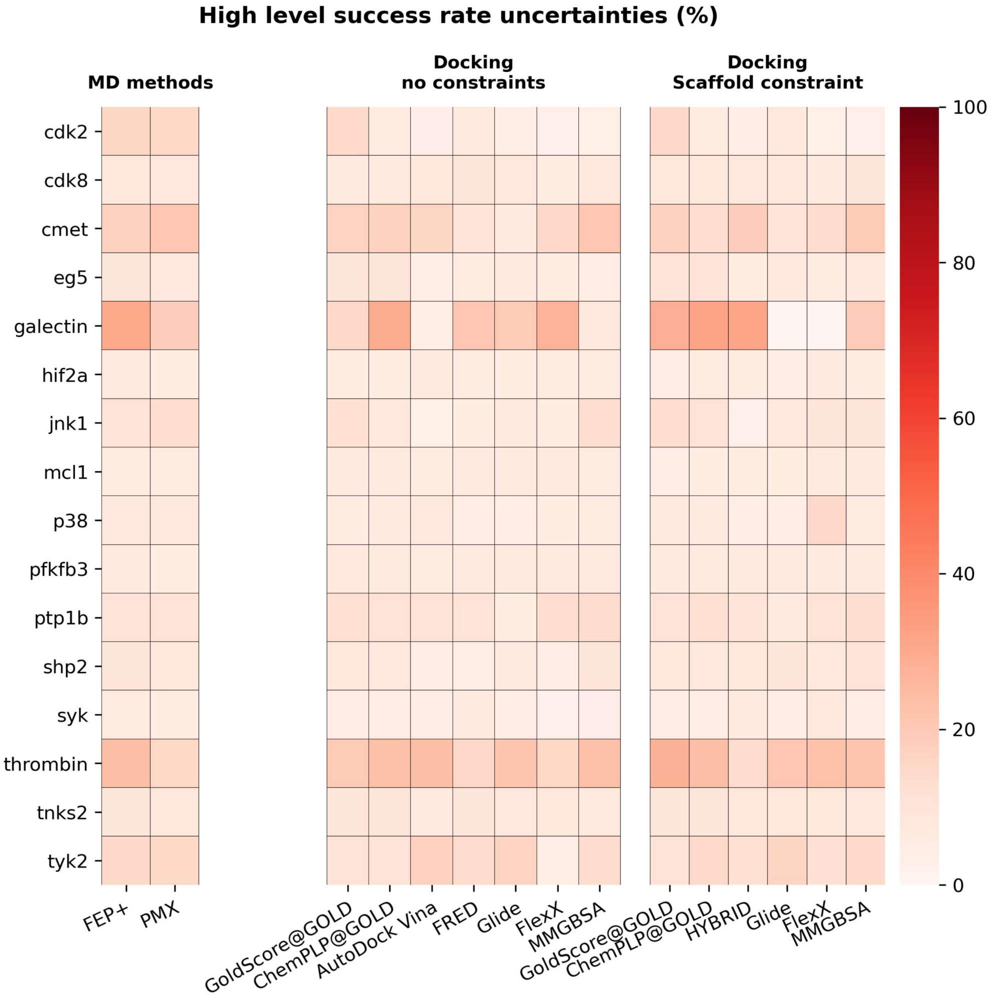

Figure S35: Uncertainties of high level success rate for all docking algorithms across all targets. Standard deviations after 10000 rounds bootstrapping are reported here.

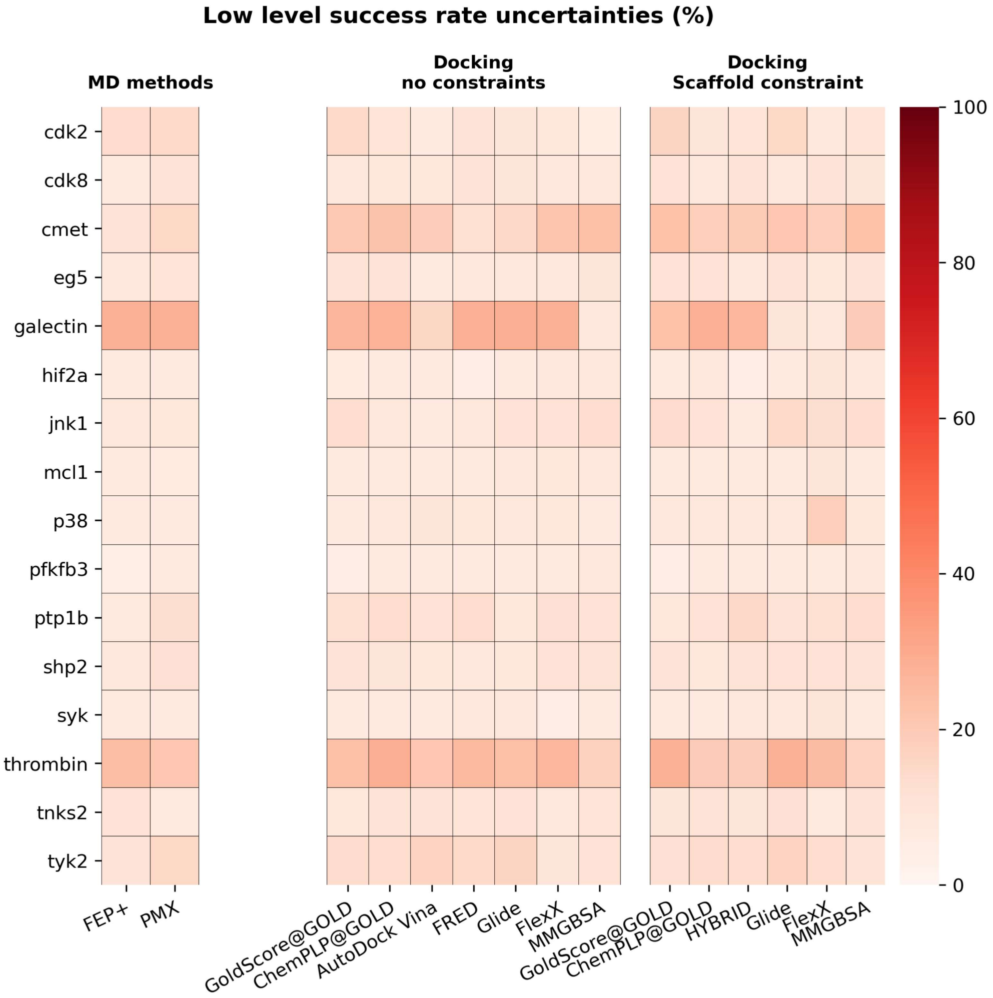

Figure S36: Uncertainties of low level success rate for all docking algorithms across all targets. Standard deviations after 10000 rounds bootstrapping are reported here.

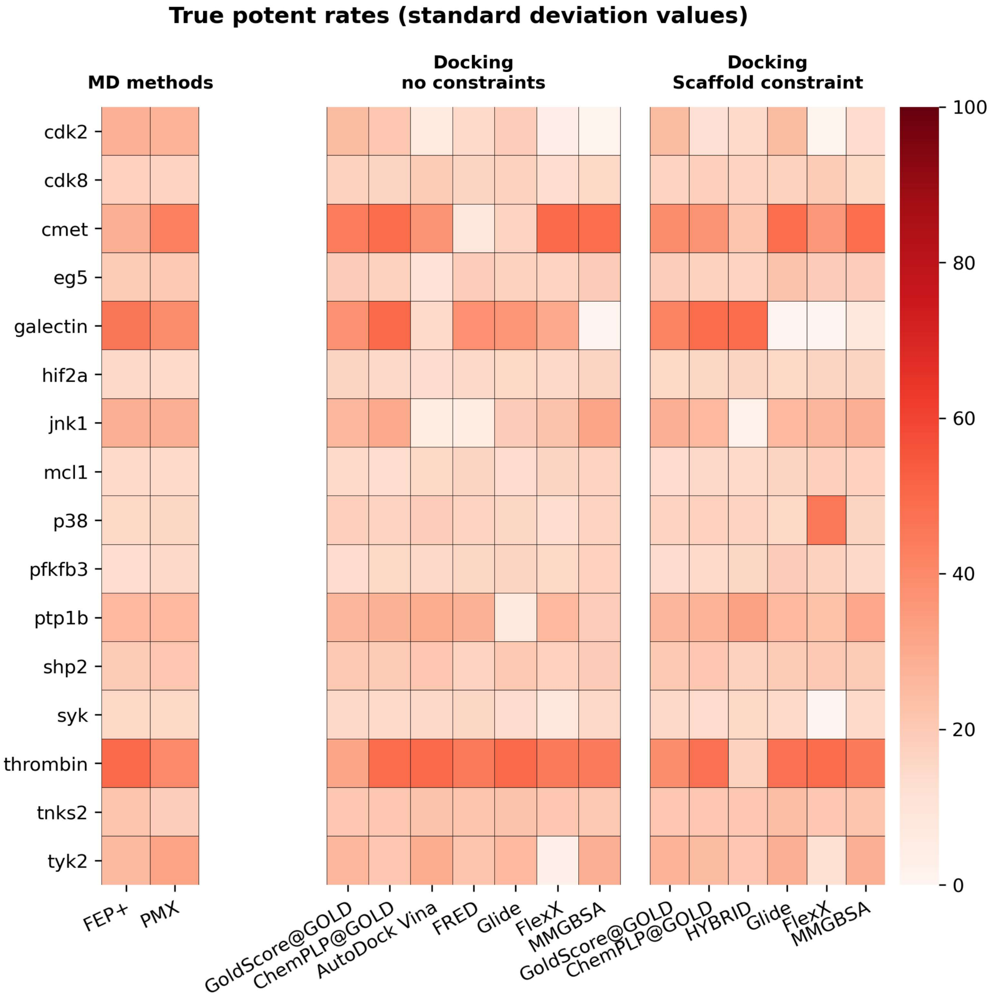

Figure S37: Uncertainties of true potent rates (%) for each method across all targets. Standard deviations after 10000 rounds bootstrapping are reported here.

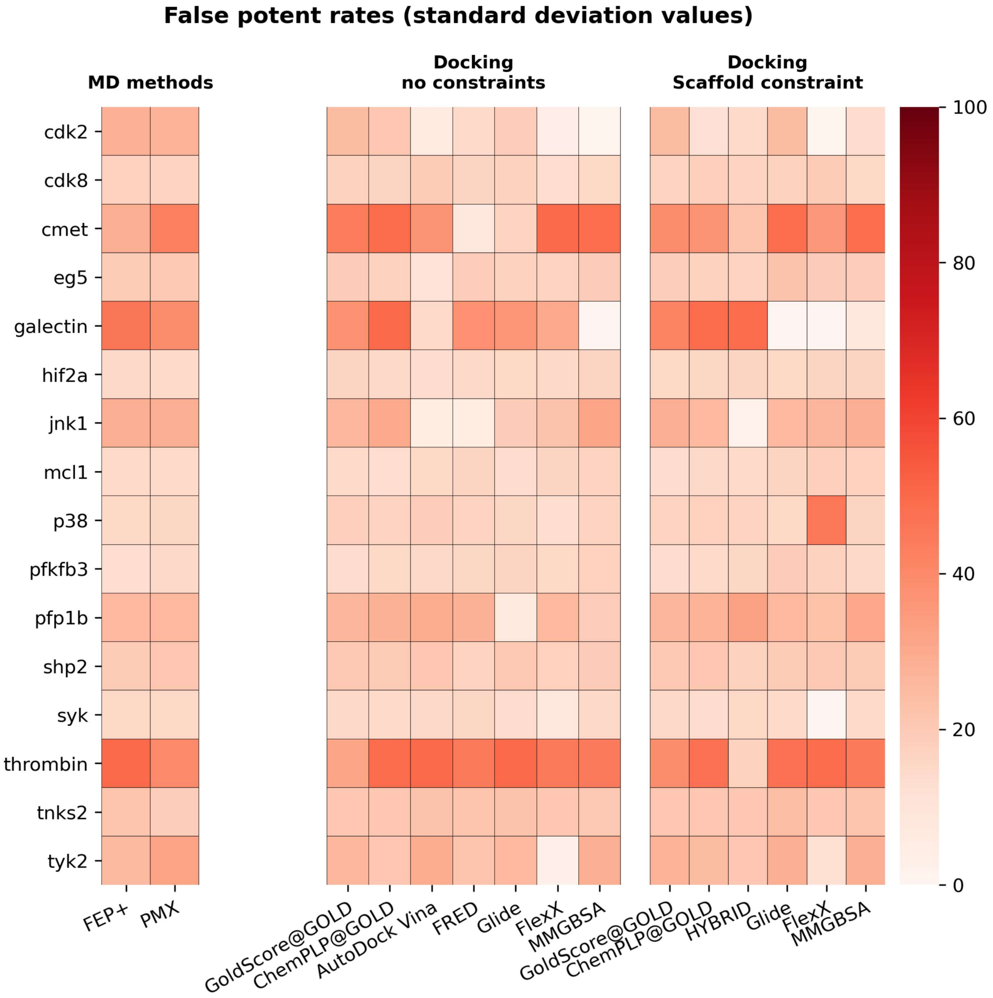

Figure S38: Uncertainties of false potent rates (%) for each method across all targets. Standard deviations after 10000 rounds bootstrapping are reported here.

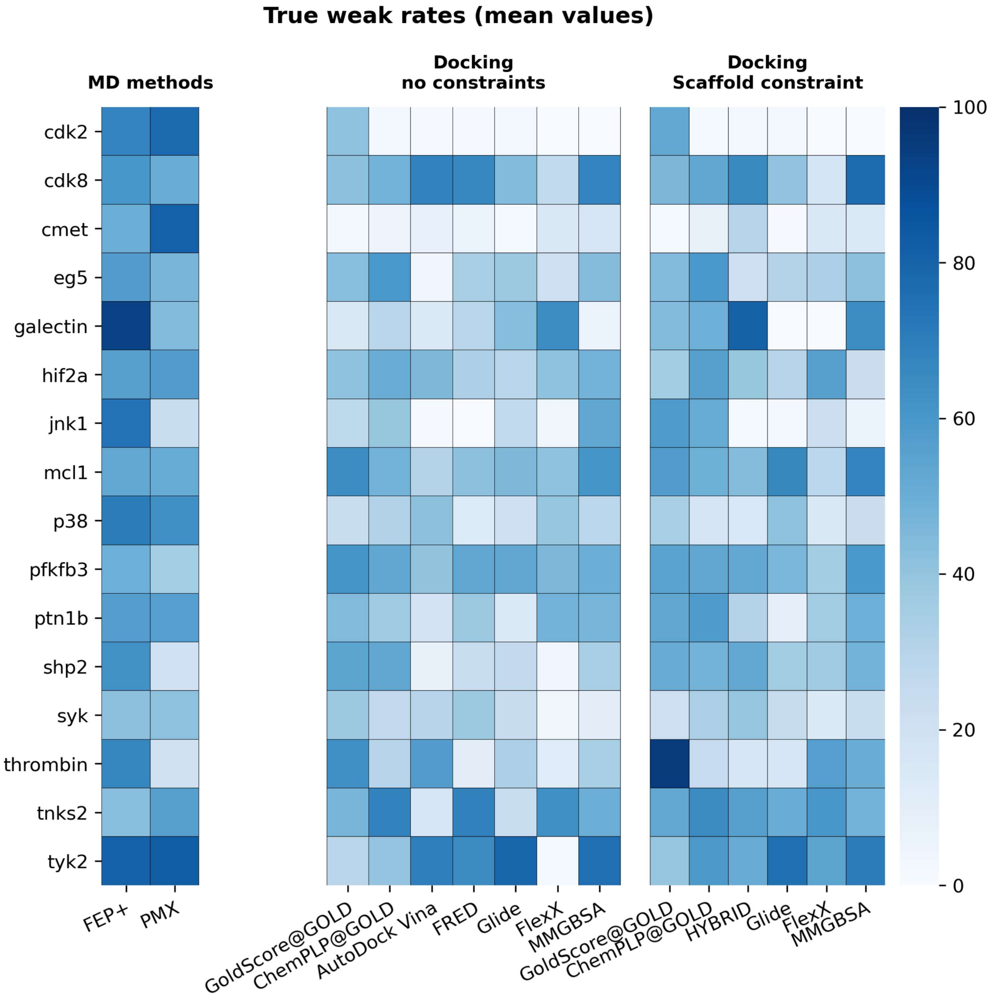

Figure S39: True weak rates (%) for each method across all targets. Mean values of each target after bootstrapping are reported here. MD-based methods have higher true weak rates than docking algorithms although exceptions are also observed.

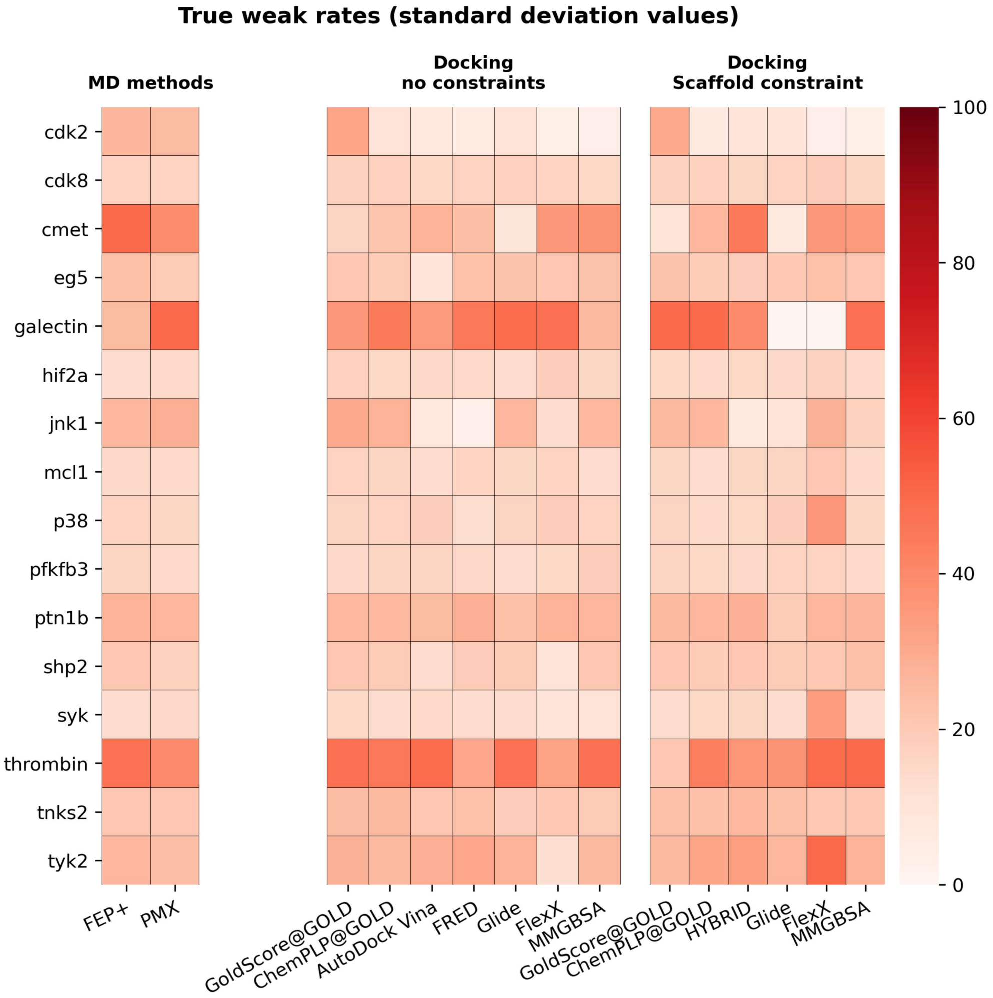

Figure S40: Uncertainties of true weak rates (%) for each method across all targets. Standard deviations after 10000 rounds bootstrapping are reported here.

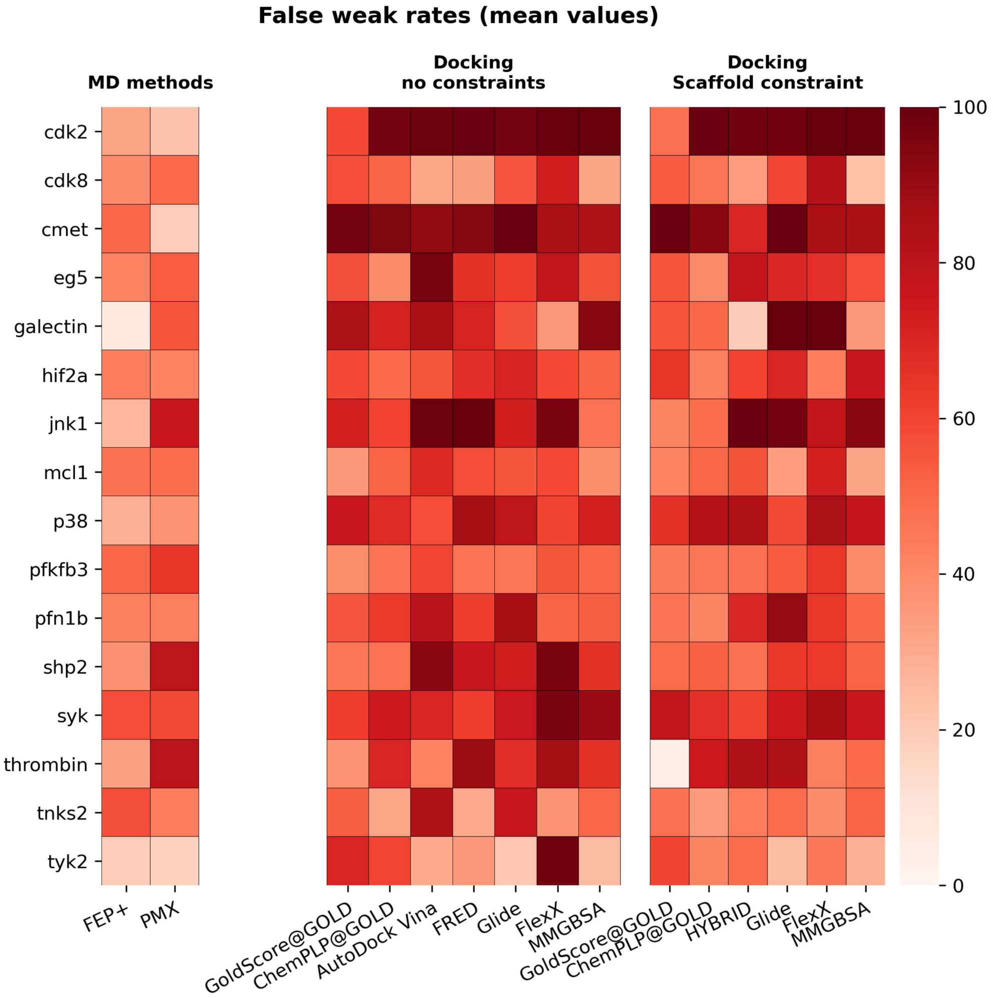

Figure S41: False weak rates (%) for each method across all targets. Mean values of each target after bootstrapping are reported here. MD-based methods have lower false weak rates than docking algorithms although exceptions are also observed.

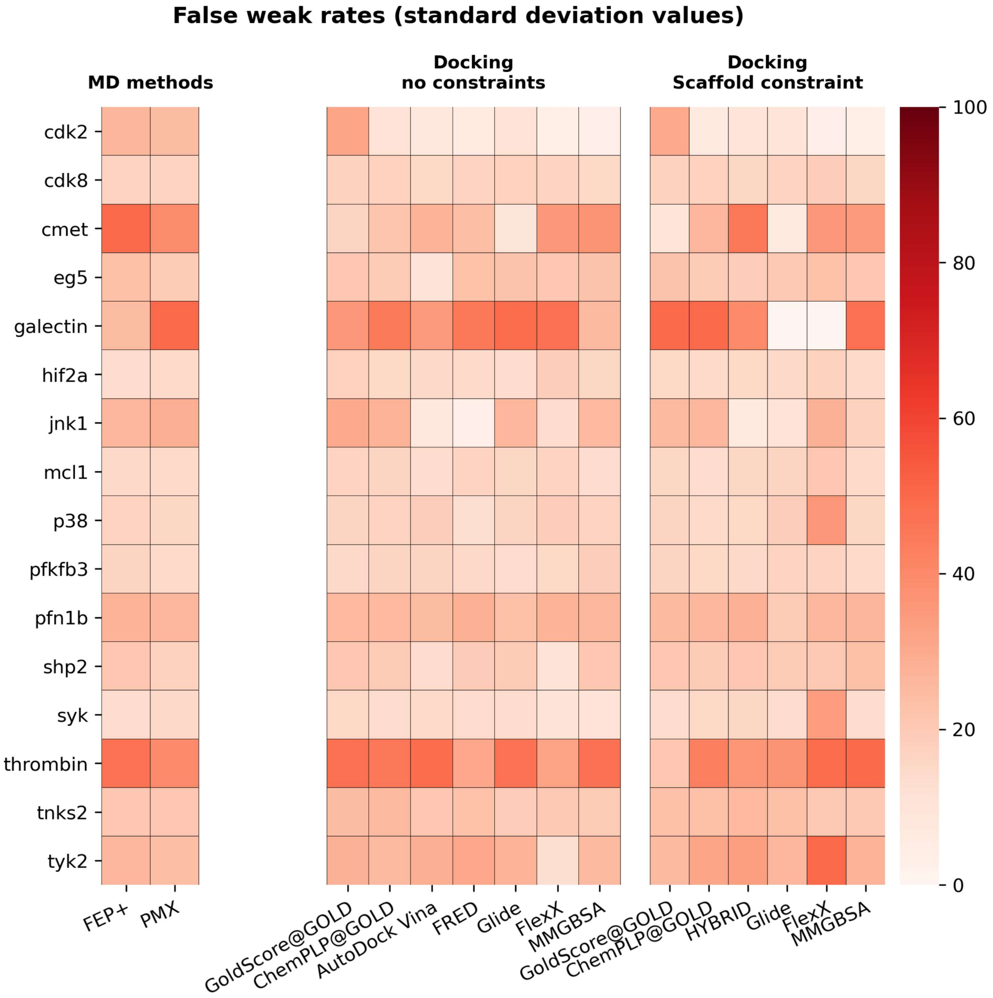

Figure S42: Uncertainties of false weak rates (%) for each method across all targets. Standard deviations after 10000 rounds bootstrapping are reported here.