Improving Structure-Based Virtual Screening with Ensemble Docking and Machine Learning: Supporting Information

Joel Ricci-López,[†] Sergio A. Aguila,[*,‡] Michael K. Gilson,[¶] and Carlos A. Brizuela[*,†]

†*Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Ensenada, Baja California, México, C.P. 22860*

‡*Centro de Nanociencias y Nanotecnología, Universidad Nacional Autonoma de México (UNAM), Ensenada, Baja California, México, C.P. 22860*

¶*Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, California 92093, USA.*

E-mail: aguila@cnyn.unam.mx; cbrizuel@cicese.mx

Table **S1**: **Hyperparameter optimization of the two Machine Learning models.**

ML model	Hyperparameter	Grid Values	Default Values	Chosen values			
				CDK2	**FXa**	**EGFR**	**HSP90**
Logistic Regression	C	1e-6, 1e-4, 0.01, 1, 100	0.01	1.0	0.01	1.0	1.0
	penalty	*l1*, *l2*	*l2*	*l2*	*l2*	*l2*	*l2*
Gradient Boosting Trees	n_estimators	200, 300, 500	100	200	200	200	500
	max_depth	3, 5, 10, 20	*None*	20	10	10	5
	learning_rate	0.05, 0.1	*None*	0.05	0.1	0.1	0.05
	gamma	0.01, 0.1, 0.5, 1	*None*	0.01	1	0.01	0.01
	alpha	0.01, 0.1, 0.5, 1	*None*	0.01	0.5	0.01	0.1
	subsample	0.3, 0.5, 1	*None*	0.5	0.5	0.5	0.6
	colsample_bytree	0.3, 0.5, 1	*None*	0.5	1	0.5	0.5

[a] The meaning and the default values of the ML classifiers' hyperparameters can be consulted in the scikit-learn[1] documentation (v0.23.2), and the XGBoost[2] documentation (v1.3.0).

Table **S2**: **Shapiro-Wilk (SW) normality test[3] results per virtual screening method from the 30x4cv analyis.**

Protein	Metric	SW results	csAVG	csGEO	csMIN	DClf	GBT	LR
CDK2	**AUC-ROC**	**W-statistic**	0.99	0.99	0.99	0.99	0.97	0.99
		p-value	0.37	0.34	0.36	0.27	**0.01**	0.40
	NEF	**W-statistic**	0.98	0.99	0.98	0.98	0.98	0.99
		p-value	0.21	0.38	0.14	**0.05**	0.13	0.47
FXa	**AUC-ROC**	**W-statistic**	0.98	0.98	0.98	0.98	0.98	0.99
		p-value	0.10	0.10	**0.08**	**0.03**	0.18	0.50
	NEF	**W-statistic**	0.97	0.98	0.98	0.94	0.98	0.98
		p-value	**0.03**	**0.04**	0.06	**0.00**	**0.05**	**0.04**
EGFR	**AUC-ROC**	**W-statistic**	0.99	0.99	0.99	0.98	0.98	0.99
		p-value	0.73	0.71	0.36	0.18	0.20	0.48
	NEF	**W-statistic**	0.98	0.98	0.99	0.97	0.98	0.99
		p-value	0.13	0.10	0.31	**0.02**	0.14	0.59
HSP90	**AUC-ROC**	**W-statistic**	0.99	0.99	0.99	0.98	0.98	0.99
		p-value	0.35	0.55	0.69	0.17	0.12	0.24
	NEF	**W-statistic**	0.98	0.97	0.98	0.98	0.99	0.98
		p-value	0.07	**0.02**	0.21	**0.05**	0.54	0.17

LR: Logistic Regression; *GBT*: Gradient Boosting Trees; *DClf*: Dummy Classifier; *csAVG*: average; *csGEO*: geometric mean; *csMIN*: minimum.
Bold-font values: Null hypothesis of normality rejected, at $\alpha = 0.05$.

Table **S3**: **Repeated measures ANOVA and Mauchly's tests[4] of sphericity results.**

Protein	Metric	DFn	DFd	F	p	p<.05	ges	W-sphe	p-sphe	p<.05
CDK2	AUC-ROC	5.00	595.00	8203.69	0.00	*	0.98	0.00	0.00	*
	NEF	5.00	595.00	5583.09	0.00	*	0.96	0.03	0.00	*
FXa	AUC-ROC	5.00	595.00	6864.65	0.00	*	0.96	0.00	0.00	*
	NEF	5.00	595.00	3428.12	0.00	*	0.94	0.03	0.00	*
EGFR	AUC-ROC	5.00	595.00	26063.04	0.00	*	0.99	0.00	0.00	*
	NEF	5.00	595.00	17595.87	0.00	*	0.99	0.01	0.00	*
HSP90	AUC-ROC	5.00	595.00	4348.47	0.00	*	0.95	0.00	0.00	*
	NEF	5.00	595.00	3280.88	0.00	*	0.94	0.16	0.00	*

DFn, DFd: Degrees of freedom, n: numerator, d: denominator.
F: F-value. **p**: Repeated measures ANOVA p-value.
ges: Generalized effect size.
W-sphe: Mauchly's statistic. **p-sphe**: Sphericity p-value.

(a)

CDK2 molecular library. Initial and final number of molecules per benchmarking set.

Libraries	Initial num. of molecules	Final num. of molecules
COCRYS	315	292
CSAR	111	110
DEKOIS	2,146	1,825
DUD	1,240	1,239
Total	3,812	**3,466**

CDK2: duplicated molecules among benchmarking sets

Library 1	Library 2	Num. of duplicated mols
COCRYS	COCRYS	2
COCRYS	CSAR	8
COCRYS	DEKOIS	1
COCRYS	DUD	17
CSAR	DUD	1
DEKOIS	DEKOIS	3
DUD	DEKOIS	3
DUD	DUD	503

CDK2: Merged library

Total num. of molecules	Num. of actives	Num. of decoys	Ra value
3,466	415	3,051	0.119

(b)

FXa molecular library. Initial and final number of molecules per benchmarking set.

Libraries	Initial num. of molecules	Final num. of molecules
COCRYS	128	119
DEKOIS	1,240	1,221
DUD	5,891	4,893
Total	7,259	**6,233**

FXa: duplicated molecules among benchmarking sets

Library 1	Library 2	Num. of duplicated mols
COCRYS	COCRYS	3
COCRYS	DEKOIS	1
COCRYS	DUD	5
DEKOIS	DEKOIS	21
DUD	DEKOIS	2
DUD	DUD	2,001

FXa: Merged library

Total num. of molecules	Num. of actives	Num. of decoys	Ra value
6,233	300	5,933	0.048

Figure **S1**: **Molecular libraries.** Tables showing the initial and the final number of molecules inside each molecular library after eliminating duplicates. The number of duplicates among pairs of molecular libraries is also shown. **(a)** CDK2 protein. **(b)** FXa protein.

(c)

EGFR molecular library. Initial and final number of molecules per benchmarking set.

Libraries	Initial num. of molecules	Final num. of molecules
COCRYS	109	105
DEKOIS	1,240	1,229
DUD	16,471	14,176
Total	17,820	**15,510**

EGFR: duplicated molecules among benchmarking sets

Library 1	Library 2	Num. of duplicated mols
COCRYS	COCRYS	4
COCRYS	DEKOIS	0
COCRYS	DUD	0
DEKOIS	DEKOIS	41
DUD	DEKOIS	5
DUD	DUD	3,170

EGFR: Merged library

Total num. of molecules	Num. of actives	Num. of decoys	Ra value
15,510	585	14,925	0.037

(d)

HSP90 molecular library. Initial and final number of molecules per benchmarking set.

Libraries	Initial num. of molecules	Final num. of molecules
COCRYS	200	191
DEKOIS	1,240	1,234
DUD	1,016	877
Total	2,456	**2,302**

HSP90: duplicated molecules among benchmarking sets

Library 1	Library 2	Num. of duplicated mols
COCRYS	COCRYS	0
COCRYS	DEKOIS	3
COCRYS	DUD	6
DEKOIS	DEKOIS	3
DUD	DEKOIS	6
DUD	DUD	214

HSP90: Merged library

Total num. of molecules	Num. of actives	Num. of decoys	Ra value
2,302	256	2,046	0.111

Figure **S1**: **(Continuation) Molecular libraries.** Tables showing the initial and the final number of molecules inside each molecular library after eliminating duplicates. The number of duplicates among pairs of molecular libraries is also shown. **(c)** EGFR protein. **(d)** HSP90 protein.

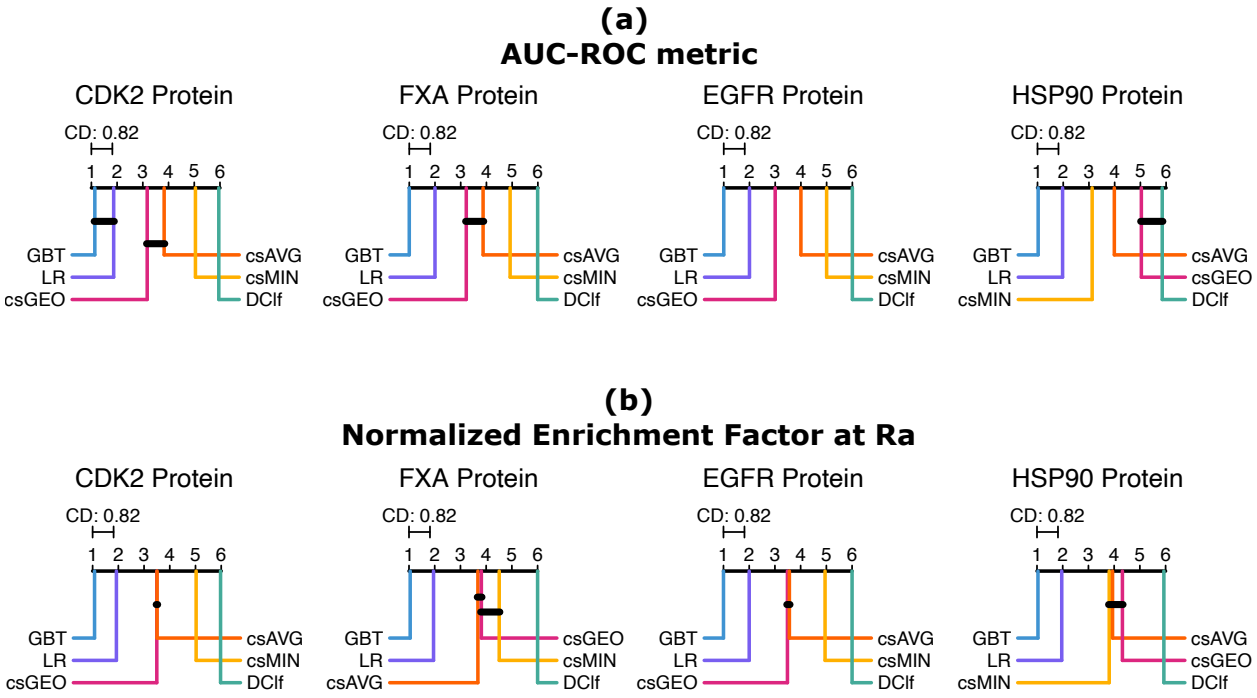

Figure **S2**: **Critical Difference (CD) plots of the pairwise Nemenyi comparison test for the 30×4cv analysis.** SBVS methods with average ranks within the CD (0.82) are not significantly different (at $\alpha = 0.01$).

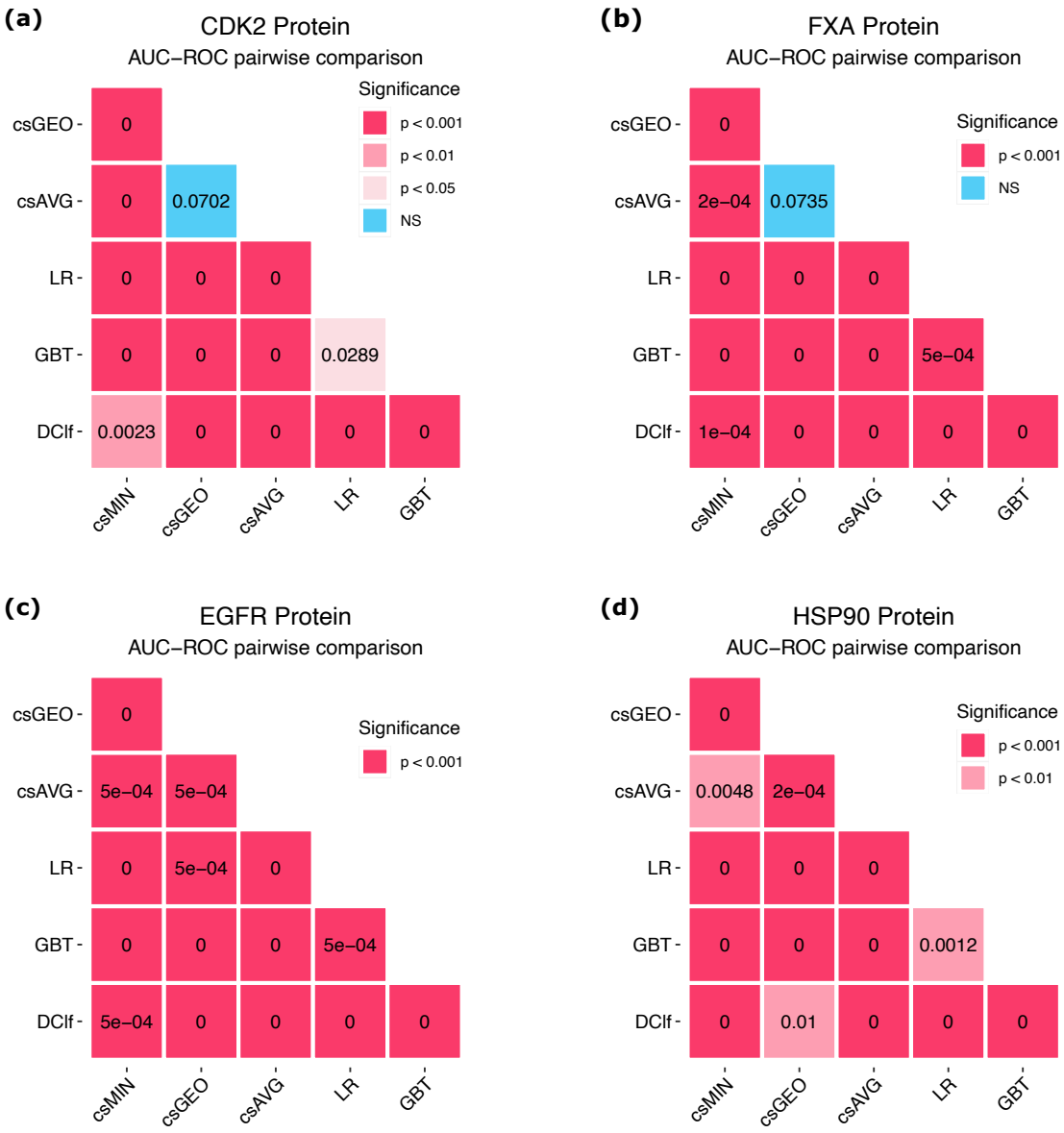

Figure **S3**: **Pairwise comparison between virtual screening methods' AUC-ROC performances using the Nemenyi pos-hoc test.** **(a)** CDK2 protein results. **(b)** FXa protein results. **(c)** EGFR protein results. **(d)** HSP90 protein results. The virtual screening methods compared are: *LR*: Logistic Regression; *GBT*: Gradient Boosting Trees; *DClf*: Dummy Classifier; *csAVG*: average; *csGEO*: geometric mean; *csMIN*: minimum.

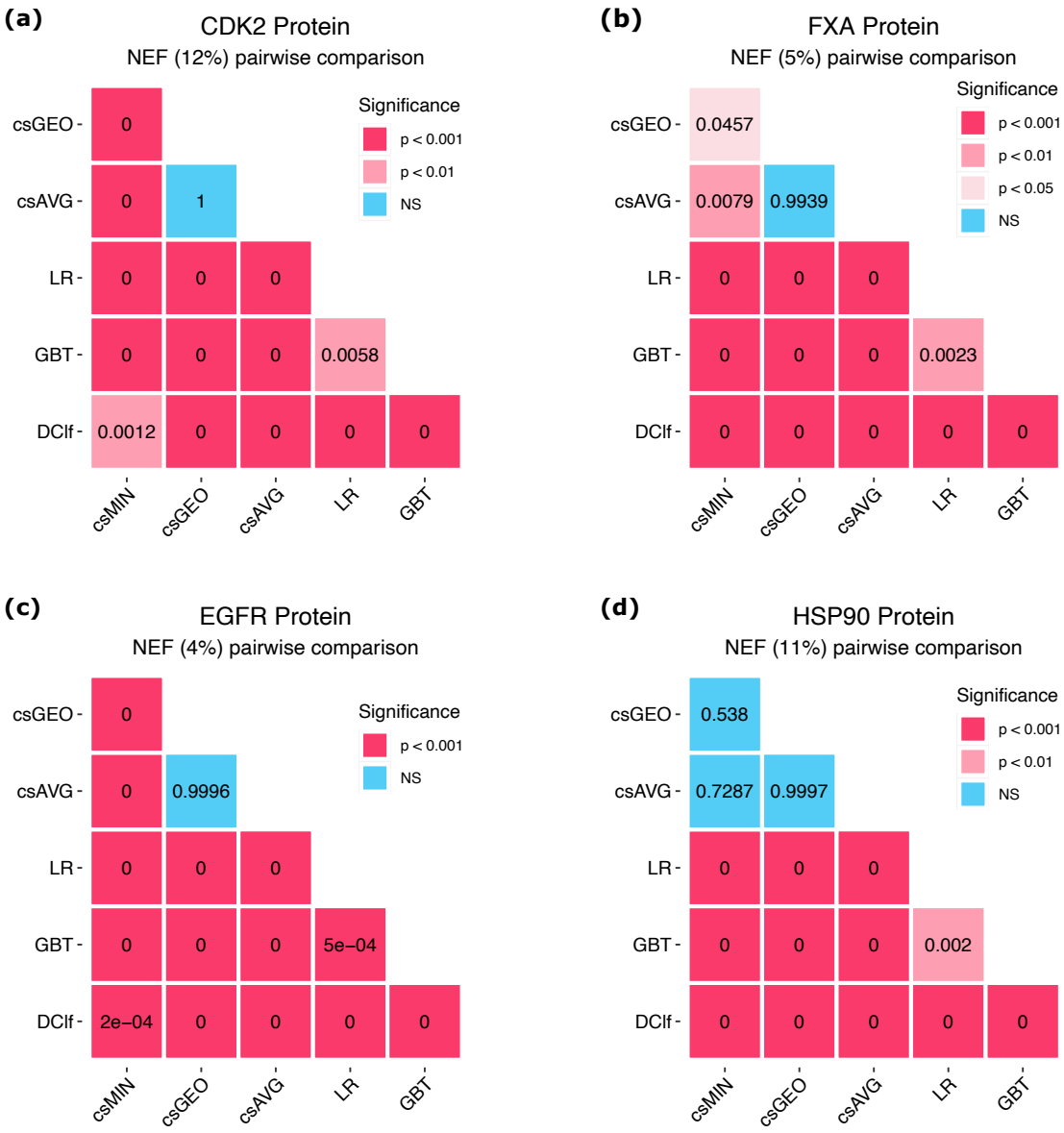

Figure **S4**: **Pairwise comparison between virtual screening methods' Normalized enrichment factor (NEF) performances using the Nemenyi pos-hoc test.** **(a)** CDK2 protein results. **(b)** FXa protein results. **(c)** EGFR protein results. **(d)** HSP90 protein results. The virtual screening methods compared are: *LR*: Logistic Regression; *GBT*: Gradient Boosting Trees; *DClf*: Dummy Classifier; *csAVG*: average; *csGEO*: geometric mean; *csMIN*: minimum.

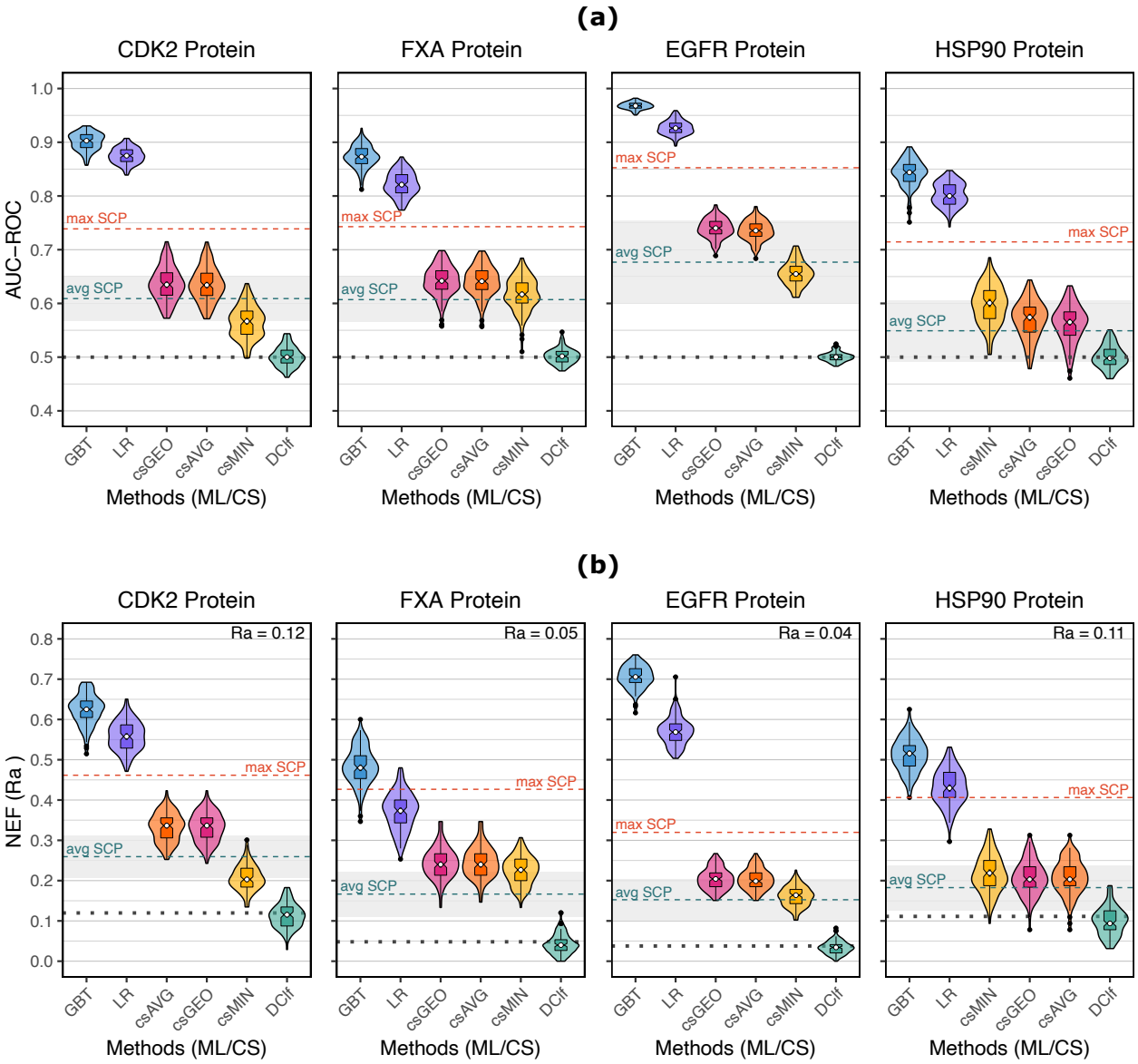

Figure **S5**: **Results from the 30×4cv analysis using the default hyperparameters (Table S1) of the two machine learning algorithms: Logistic Regression and Gradient Boosting trees. (a) AUC-ROC values. (b) NEF values.** Violin boxplots showing the distribution of the performance values of each SBVS method across the 30 repetitions. The white points within the boxes indicate the value of the median, and the notches represent the 95% confidence interval around it. Outliers are shown as black points. The max SCP (single-conformation performance) and the avg SCP dashed lines indicate the maximum and the average performance, respectively, achieved by a single conformation using the raw docking scores. The translucent gray area surrounding the svg SCP value represents one standard deviation from the average SCP. The dotted black lines indicate the expected performance of a random classifier. The results of the consensus strategies and the SCP values are the same as the Figure 2 of the main manuscript because the same validation sets were evaluated. This was done through the use of a random seed parameter of the *RepeatedKFold* object employed to perform the dataset splitting.

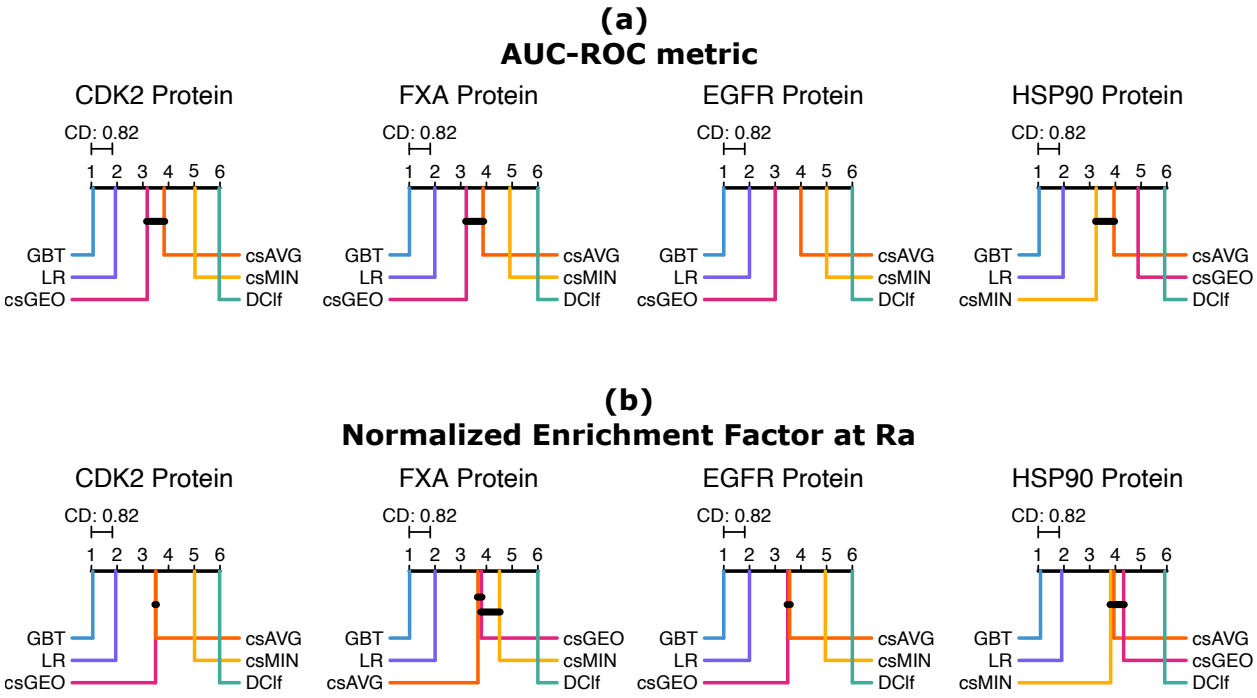

Figure **S6**: **Critical Difference (CD) plots of the pairwise Nemenyi comparison test for the 30×4cv analysis with the ML models using their default hyperparameters (Table S1).** SBVS methods with average ranks within the CD (0.82) are not significantly different (at $\alpha = 0.01$).

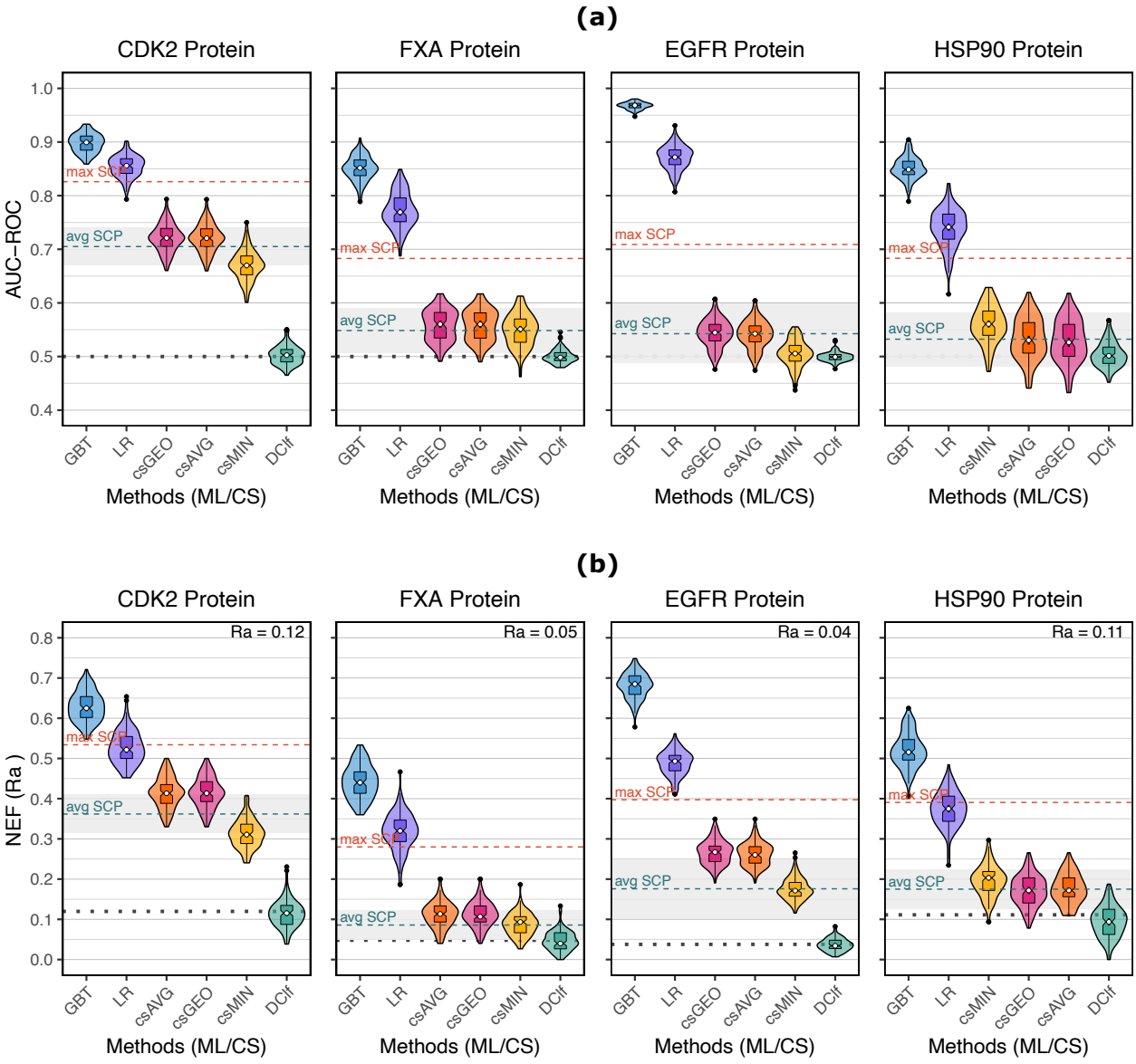

Figure **S7**: **Results from the 30×4cv analysis using Ligand Efficiency scores, which were computed by dividing the raw docking scores by the number of heavy atoms of each molecule.. (a) AUC-ROC values. (b) NEF values.** Violin boxplots showing the distribution of the performance values of each SBVS method across the 30 repetitions. The white points within the boxes indicate the value of the median, and the notches represent the 95% confidence interval around it. Outliers are shown as black points. The max SCP (single-conformation performance) and the avg SCP dashed lines indicate the maximum and the average performance, respectively, achieved by a single conformation using the raw docking scores. The translucent gray area surrounding the svg SCP value represents one standard deviation from the average SCP. The dotted black lines indicate the expected performance of a random classifier. ML algorithms were trained using their default hyperparameters (Table S1).

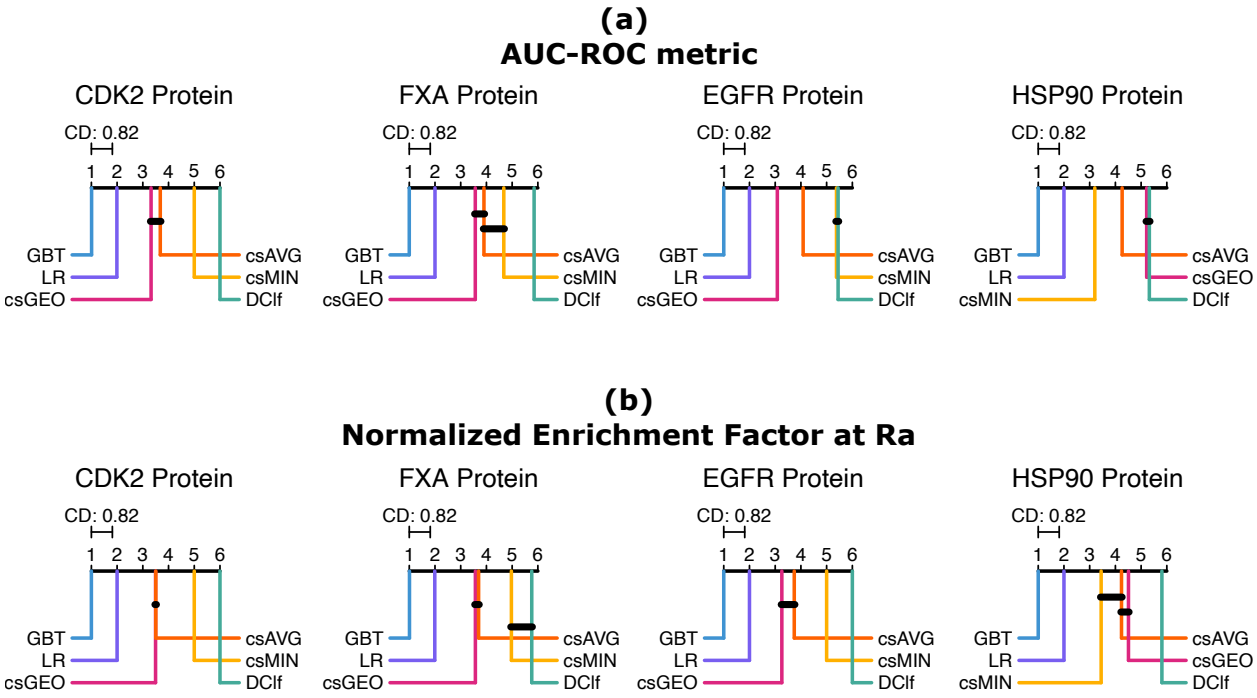

Figure **S8**: **Critical Difference (CD) plots of the pairwise Nemenyi comparison test for the 30×4cv analysis using the Ligand Efficiency scores.** ML algorithms were trained using their default hyperparameters (Table S1). SBVS methods with average ranks within the CD (0.82) are not significantly different (at $\alpha = 0.01$).

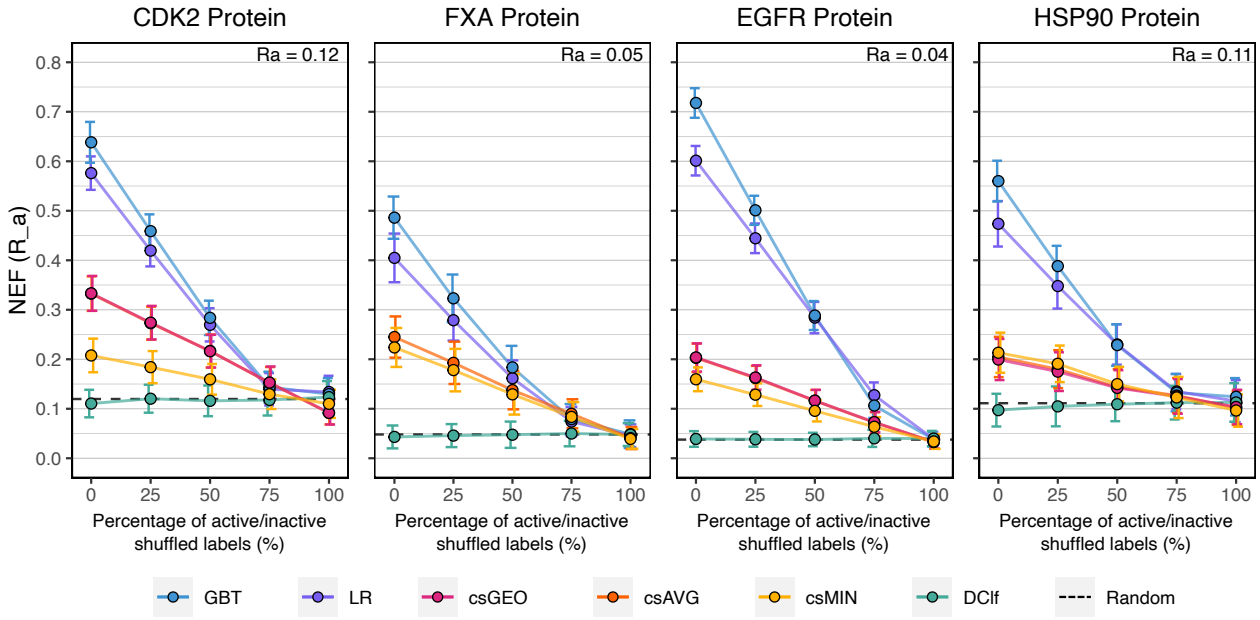

Figure **S9**: **Results of the y-randomization test showing the SBVS methods' average NEF values at different percentages of active/inactive shuffled labels.** Error bars indicate standard deviations. The csAVG and csGEO strategies had practically the same average and standard deviation values.

Table **S4**: **Kruskal-Wallis tests for independent samples comparing different percentages of y-randomization per virtual screening method.**

	Protein	Metric	VS_method	n	KW-statistic	df	p	signif
1	**CDK2**	**NEF**	LR	600	546.15	4	0.00	***
2			GBT	600	548.33	4	0.00	***
3			Dclf	600	11.42	4	0.02	*
4			csAVG	600	530.82	4	0.00	***
5			csGEO	600	530.41	4	0.00	***
6			csMIN	600	346.07	4	0.00	***
7		**ROC-AUC**	LR	600	545.82	4	0.00	***
8			GBT	600	546.82	4	0.00	***
9			Dclf	600	5.23	4	0.26	NS
10			csAVG	600	468.46	4	0.00	***
11			csGEO	600	469.66	4	0.00	***
12			csMIN	600	278.21	4	0.00	***
13	**Fxa**	**NEF**	LR	600	546.48	4	0.00	***
14			GBT	600	549.58	4	0.00	***
15			Dcld	600	6.70	4	0.15	NS
16			csAVG	600	494.44	4	0.00	***
17			csGEO	600	493.25	4	0.00	***
18			csMIN	600	478.56	4	0.00	***
19		**ROC-AUC**	LR	600	549.01	4	0.00	***
20			GBT	600	552.51	4	0.00	***
21			Dclf	600	0.97	4	0.92	NS
22			csAVG	600	421.09	4	0.00	***
23			csGEO	600	423.57	4	0.00	***
24			csMIN	600	359.09	4	0.00	***

Each row shows the Kruskal-Wallis[5] test results of each virtual screening method after comparing the performance of five percentages of y-randomization (0, 25, 50, 75, and 100%).

Table **S4**: **(Continuation) Kruskal-Wallis tests for independent samples comparing different percentages of y-randomization per virtual screening method.**

	Protein	Metric	VS_method	n	KW-statistic	df	p	signif
25	**EGFR**	**NEF**	LR	600	575.04	4	0.00	***
26			GBT	600	574.34	4	0.00	***
27			Dclf	600	1.59	4	0.81	NS
28			csAVG	600	534.35	4	0.00	***
29			csGEO	600	535.36	4	0.00	***
30			csMIN	600	502.66	4	0.00	***
31		**ROC-AUC**	LR	600	573.45	4	0.00	***
32			GBT	600	570.70	4	0.00	***
33			Dclf	600	5.27	4	0.26	NS
34			csAVG	600	564.28	4	0.00	***
35			csGEO	600	565.35	4	0.00	***
36			csMIN	600	522.66	4	0.00	***
37	**HSP90**	**NEF**	LR	600	533.08	4	0.00	***
38			GBT	600	532.68	4	0.00	***
39			Dclf	600	13.33	4	0.01	**
40			csAVG	600	304.57	4	0.00	***
41			csGEO	600	281.23	4	0.00	***
42			csMIN	600	353.34	4	0.00	***
43		**ROC-AUC**	LR	600	535.62	4	0.00	***
44			GBT	600	532.06	4	0.00	***
45			Dclf	600	5.69	4	0.22	NS
46			csAVG	600	297.10	4	0.00	***
47			csGEO	600	264.91	4	0.00	***
48			csMIN	600	380.54	4	0.00	***

Each row shows the Kruskal-Wallis[5] test results of each virtual screening method after comparing the performance of five percentages of y-randomization (0, 25, 50, 75, and 100%).

Table **S5**: **Dunn's post hoc test[6] for pairwise comparison between pair or percentages (pct.) of y-randomization per virtual screening method. Only non-significant differences are shown.**

	Protein	Metric	VS method[a]	pct. 1	pct. 2	n	Dunn's statistic	p	p.adj	p.adj sig.
1	CDK2	NEF	LR	75%	100%	120	-0.41	0.69	0.69	ns
2	CDK2	NEF	GBT	75%	100%	120	-1.44	0.15	0.15	ns
3	CDK2	ROC-AUC	LR	75%	100%	120	0.51	0.61	0.61	ns
4	CDK2	ROC-AUC	GBT	75%	100%	120	0.77	0.44	0.44	ns
5	HSP90	NEF	LR	75%	100%	120	-1.53	0.13	0.13	ns
6	HSP90	NEF	GBT	75%	100%	120	-0.74	0.46	0.46	ns
7	HSP90	ROC-AUC	LR	75%	100%	120	-1.18	0.24	0.24	ns
8	HSP90	ROC-AUC	GBT	75%	100%	120	-0.80	0.42	0.42	ns

Dunn's test was applied among different percentages of y-randomization (0, 25, 50, 75, and 100%) per each virtual screening method, regarding a specific protein and a specific performance metric. Only non-significant differences are shown. Full results can be consulted in the GitHub repository: jRicciL/ML-ensemble-docking

[a] The Dummy Classifier results, all of them non-significant, are omitted.

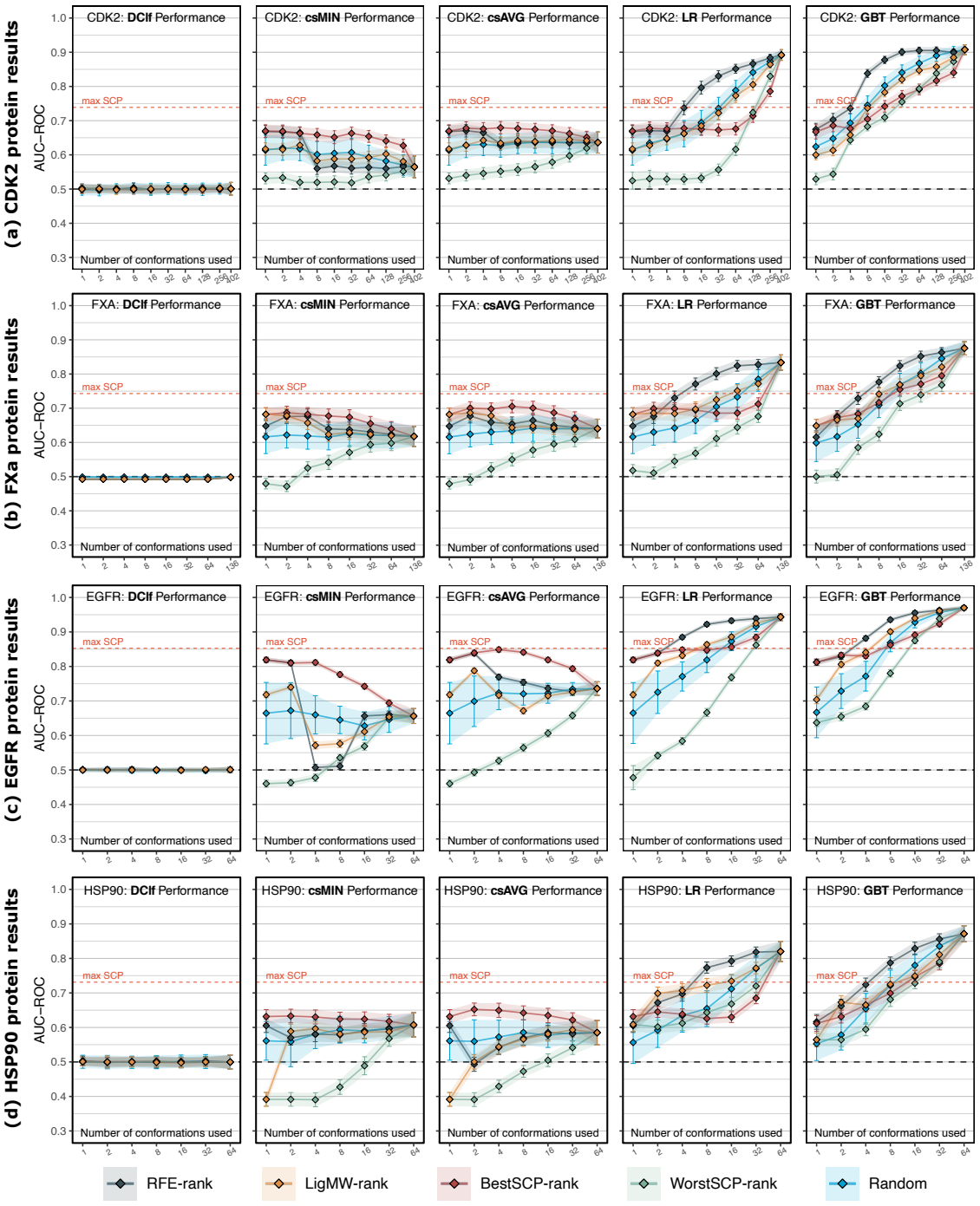

Figure **S10**: **Comparison between the average AUC-ROC performance between machine learning classifiers and traditional consensus strategies using only k protein conformations.** Five different selection criteria are compared. Error bars indicate standard deviations. The max SCP dashed line indicates the maximum performance achieved by a single conformation using the raw docking scores from the $120 \times n$ validation sets generated during the 30×4cv analysis, where all n conformations were considered. **(a)** CDK2 protein results. **(b)** FXa protein results ($k = 128$ values are omitted for clarity). **(c)** EGFR protein results. **(d)** HSP90 protein results. csGEO is ommited for simplicity as its results are similar to the csAVG strategy. Full results can be consulted in the Github repository: jRicciL/ML-ensemble-docking

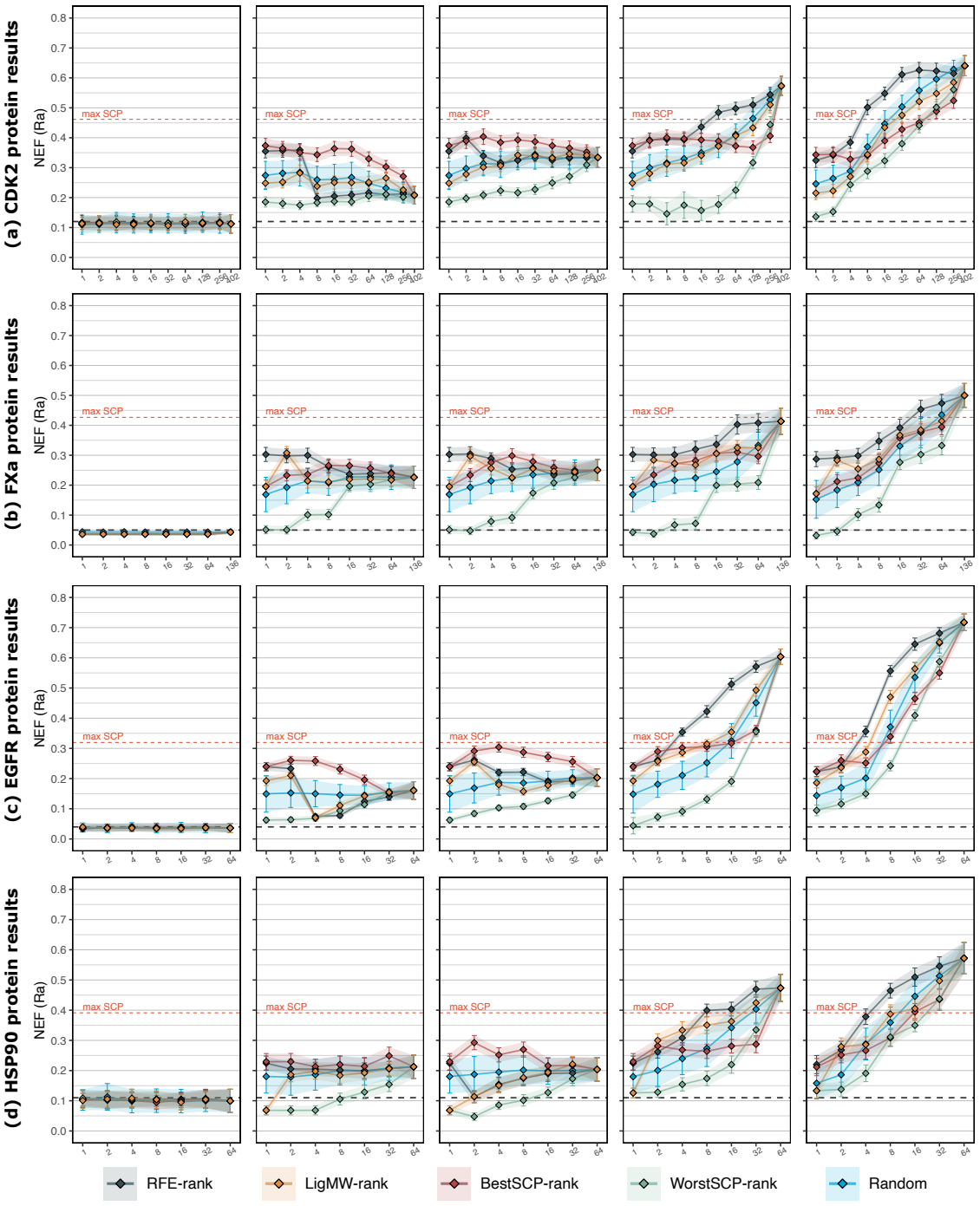

Figure **S11**: **Comparison between the average NEF performance between machine learning classifiers and traditional consensus strategies using only k protein conformations.** Five different selection criteria are compared. Error bars indicate standard deviations. The max SCP dashed line indicates the maximum performance achieved by a single conformation using the raw docking scores from the $120 \times n$ validation sets generated during the 30×4cv analysis, where all n conformations were considered. **(a)** CDK2 protein results. **(b)** FXa protein results ($k = 128$ values are omitted for clarity). **(c)** EGFR protein results. **(d)** HSP90 protein results. csGEO is ommited for simplicity as its results are similar to the csAVG strategy. Full results can be consulted in the Github repository: jRicciL/ML-ensemble-docking

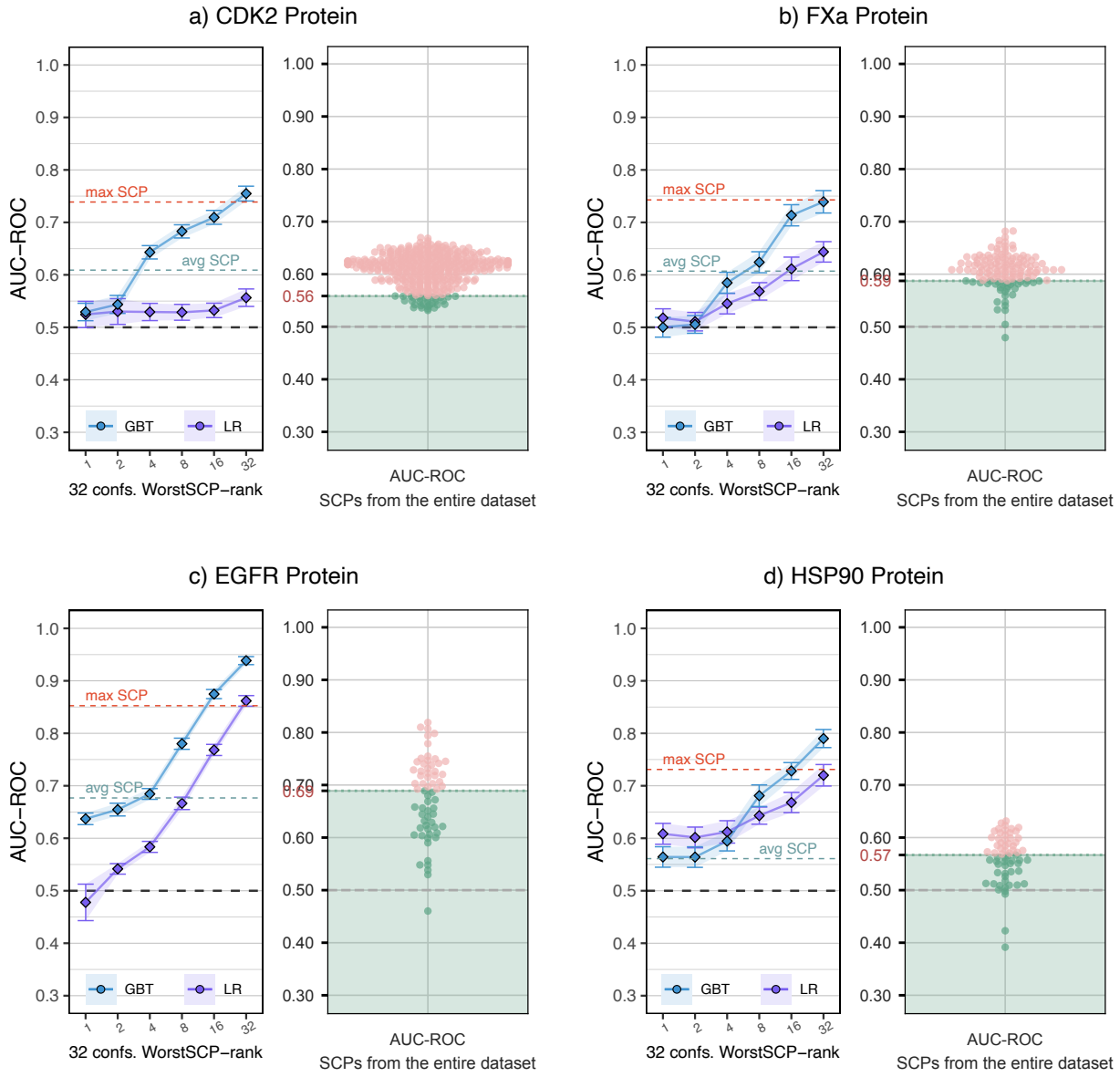

Figure **S12**: **Gradient Boosting Trees (GBT) and Logistic Regression (LR) classifiers trained with the worst 32 conformations as the input features. (a)** CDK2 protein. **(b)** FXa protein. **(c)** EGFR protein. **(d)** HSP90 protein. **Line plot panels:** performance of the GBT (blue) and LR (purple) classifiers evaluated with the 30 repetitions of 4-fold stratified cross-validation (30x4cv) using the *worst* k conformations (WorstSCP-rank). With k ranging from the 1 to 32 *worst* conformations. The *worst* conformations were determined by their single-conformation performance (SCP) in terms of AUC-ROC computed from the entire dataset. The max SCP and avg SCP, are the maximum and the average values, respectively, achieved for a single protein using the docking scores from the $120 \times n$ validation sets generated during the 30×4cv analysis, where all n conformations were considered (see Figure 2 of the main manuscript). **Swarm plot panels:** SCP (AUC-ROC) values obtained by each protein conformation according to its performance using the raw docking scores from the entire dataset. The 32 conformations with the lowest SCP AUC-ROC values (*worst* conformations) are shown in green.

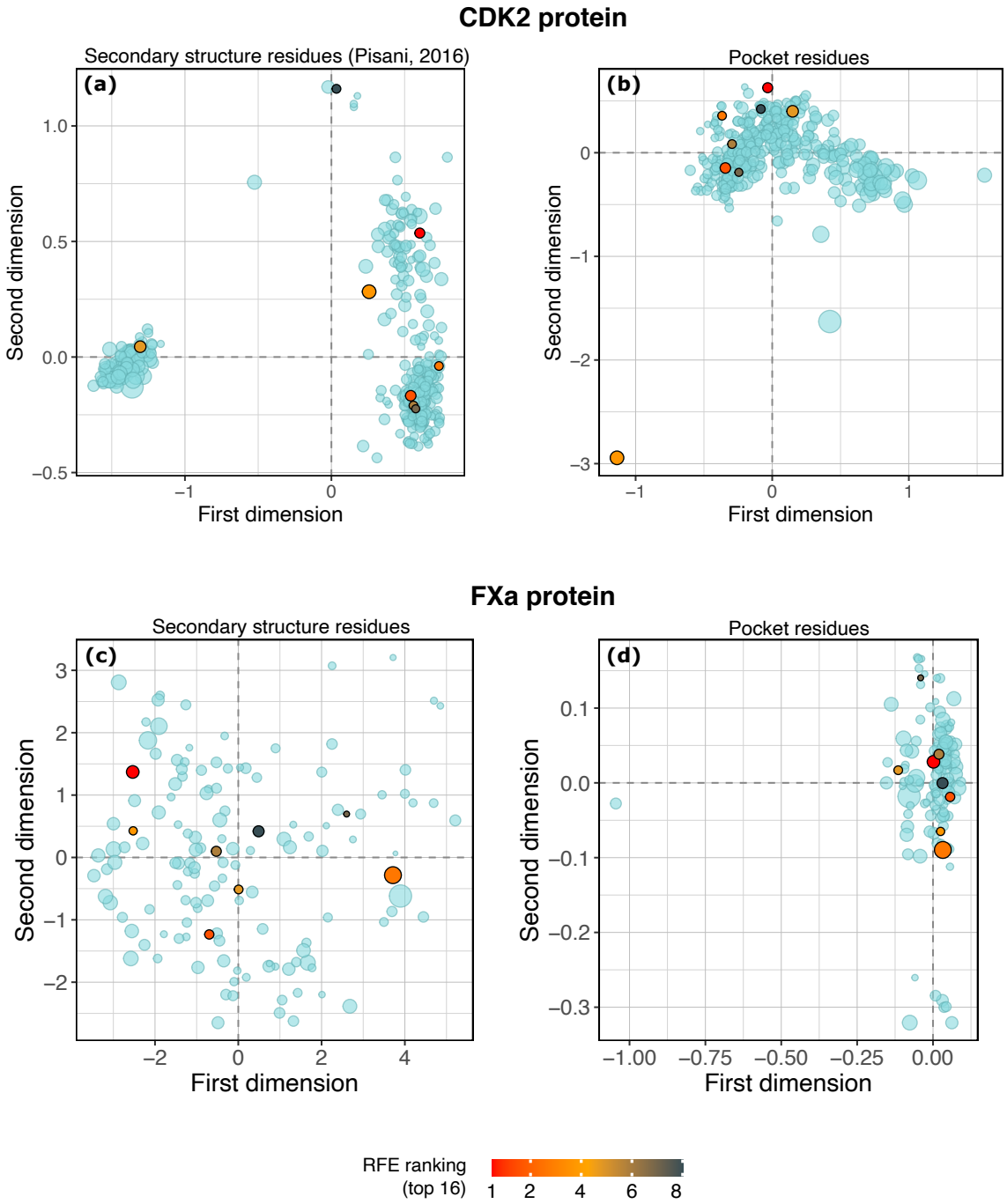

Figure **S13**: **Classical Multidimensional Scaling (cMDS) showing the top 8 (red orange to black points) protein conformations selected by the RFE procedure using GBT as a base estimator.** cMDS were computed from the pairwise RMSD matrix considering $C\alpha$ atoms from protein secondary structure residues, **(a)** and **(c)**, and protein pocket's residues, **(b)** and **(d)**. Particularly, in plot **(a)**, we used the same the secondary structure residues used by Pisani et al.[7]. Each point represents a protein conformation. The point's size is proportional to the protein pocket's volume, computed by POVME3.[8]

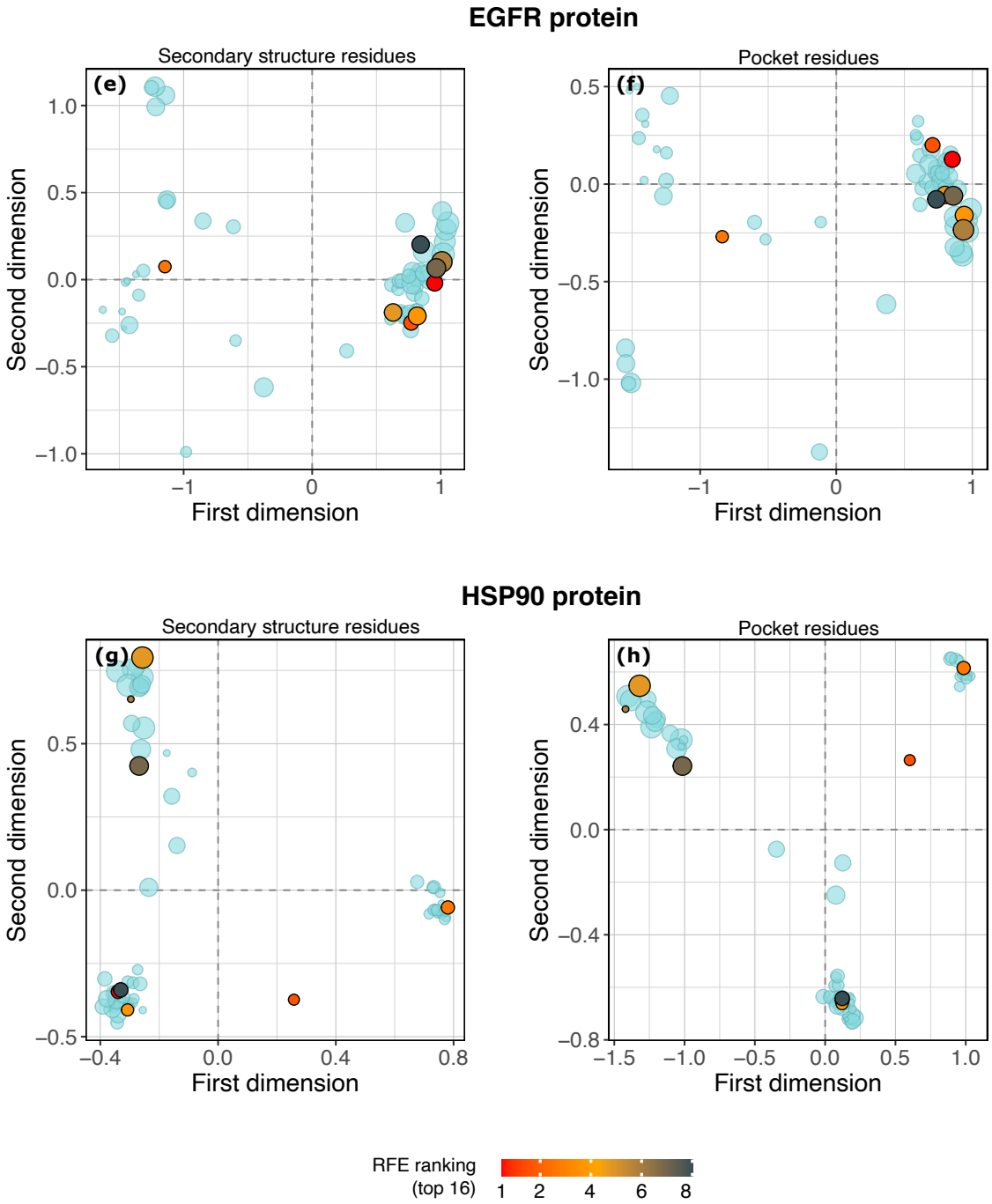

Figure **S13**: **(Continuation) Classical Multidimensional Scaling (cMDS) showing the top 8 (red orange to black points) protein conformations selected by the RFE procedure using GBT as a base estimator.** cMDS were computed from the pairwise RMSD matrix considering $C\alpha$ atoms from protein secondary structure residues, **(e)** and **(g)**, and protein pocket's residues, **(f)** and **(h)**. The point's size is proportional to the protein pocket's volume, computed by POVME3.[8]

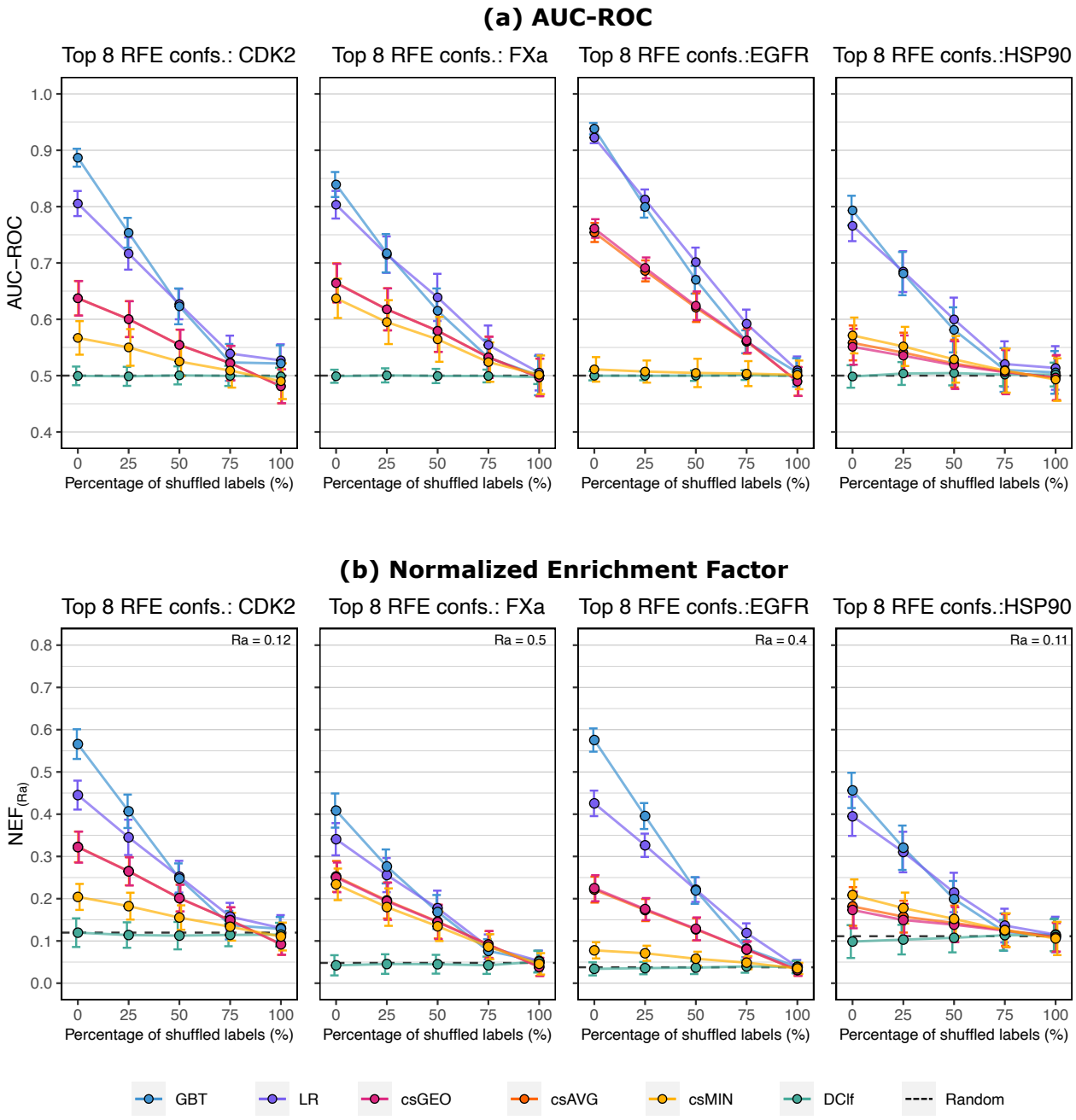

Figure **S14**: **Results of the y-randomization test using the top 8 conformations selected by the RFE procedure.** The VS methods' average performances are shown at different percentages of active/inactive shuffled labels. Error bars indicate standard deviations. **(a)** AUC-ROC values. **(b)** NEF values. The csAVG and csGEO strategies had practically the same average and standard deviation values.

Table **S6**: **CDK2 protein. Top 9 molecules as judged by the GBT classifier after the 30×4cv implementation.** Ranks according to the other methods are also shown.

Library	**Lig. Name**	Activity	rank LR	**rank GBT**	rank csAVG	rank csGEO	rank csMIN	MW	Num. atoms	Num. rots
COCRYS	03Z	Active	135	**1**	113	112	236	374.4	25	5
CSAR	CS13	Active	4	**2**	588	588	1098	316.8	21	3
COCRYS	FRV	Active	36	**3**	103	105	400	384.5	27	6
	04Z	Active	193	**4**	95	93	232	406.5	27	6
DUD	ligand_61	Active	2	**5**	495	515	1374	304.3	23	2
COCRYS	X40	Active	99	**6**	106	106	335	375.4	25	5
	CDK	Active	10	**7**	204	207	419	435.5	29	8
CSAR	CS14	Active	5	**8**	410	413	1186	300.4	21	3
COCRYS	26Z	Active	47	**9**	36	35	96	389.5	26	5

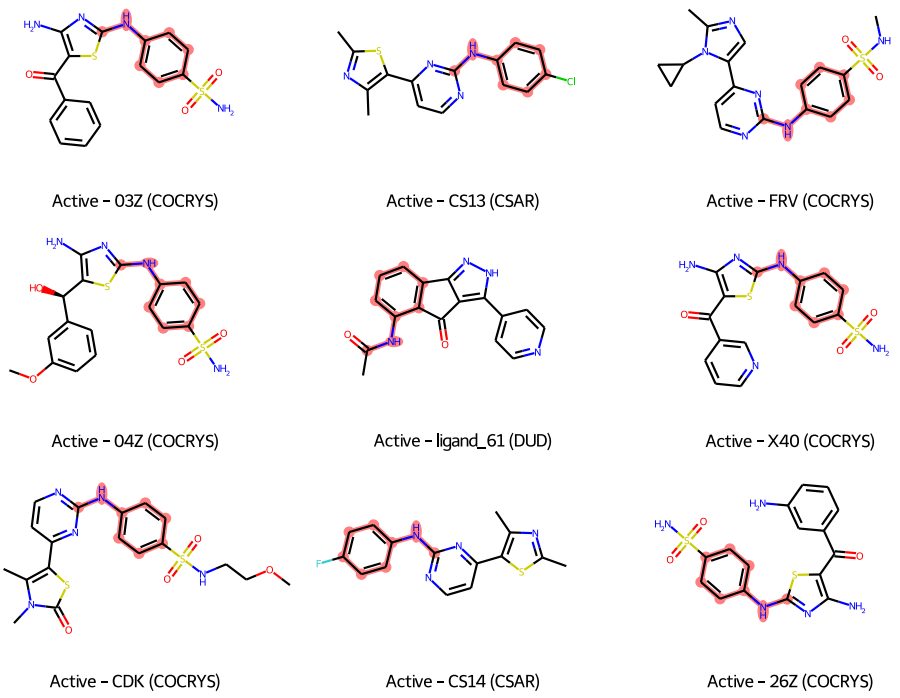

Figure **S15**: **CDK2 protein.** Structures of the top 9 molecules as judged by the **GBT** classifier after the 30×4cv implementation. The maximum common substructure is highlighted in red (*rdkit*: `completeRingsOnly = False`). An extended version can be consulted in the GitHub repository: jRicciL/ML-ensemble-docking

.

Table **S7**: **CDK2 protein. Top 9 molecules as judged by the csMIN classifier after the 30×4cv implementation.** Ranks according to the other methods are also shown.

Library	**Lig. Name**	Activity	rank LR	rank GBT	rank csAVG	rank csGEO	**rank csMIN**	MW	Num. atoms	Num. rots
COCRYS	LQ5	Active	232	247	209	208	**1**	503.5	36	6
DEKOIS	decoy_1191	Inactive	1537	178	3	3	**2**	484.5	36	7
COCRYS	2KD	Active	1376	487	21	22	**3**	503.9	36	7
DEKOIS	decoy_722	Inactive	2649	378	15	15	**4**	543.5	37	9
CSAR	CS112	Inactive	3131	357	82	89	**5**	421.4	31	8
DEKOIS	decoy_1063	Inactive	2273	139	4	4	**6**	434.5	33	6
COCRYS	EZR	Active	101	117	67	67	**7**	454.6	34	4
DEKOIS	decoy_755	Inactive	515	265	42	41	**8.5**	393.4	29	4
	decoy_1175	Inactive	2222	342	15	16	**8.5**	504.6	37	8

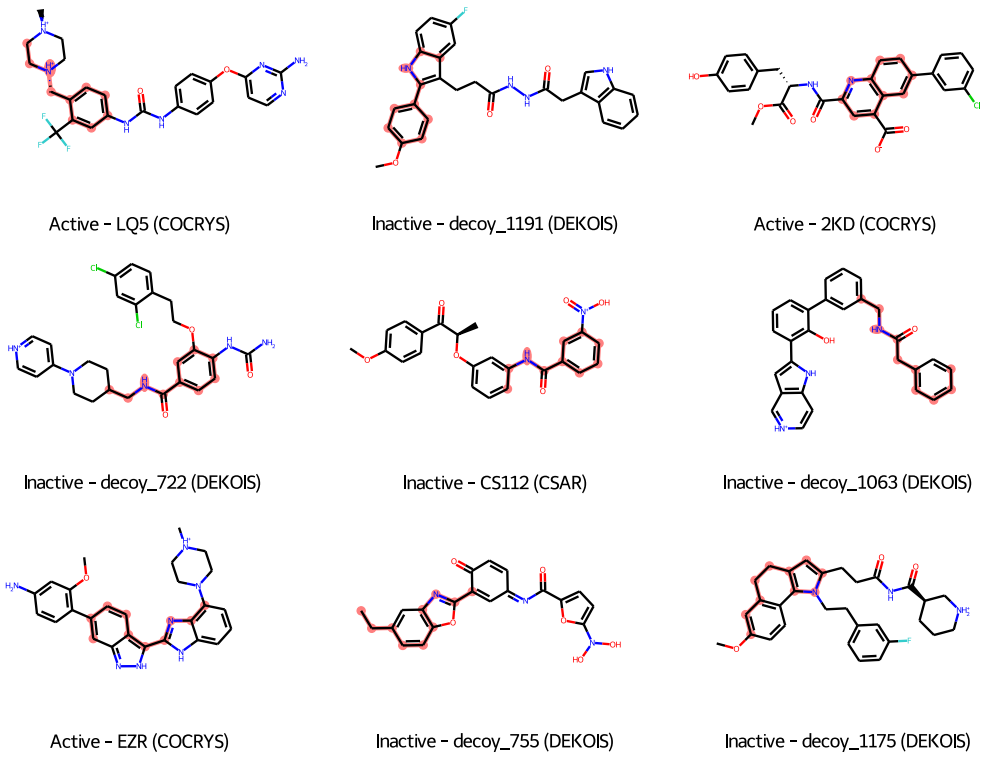

Figure **S16**: **CDK2 protein.** Structures of the top 9 molecules as judged by the **csMIN** classifier after the 30×4cv implementation. The maximum common substructure is highlighted in red (*rdkit*: `completeRingsOnly = False`). An extended version can be consulted in the GitHub repository.

Table **S8**: **FXa protein. Top 9 molecules as judged by the GBT classifier after the 30×4cv implementation.** Ranks according to the other methods are also shown.

Library	**Lig. Name**	Activity	rank LR	**rank GBT**	rank csAVG	rank csGEO	rank csMIN	MW	Num. atoms	Num. rots
DUD	ligand_90	Active	4	**1**	2831	2858	1541	426.5	32	6
	ligand_118	Active	21	**2**	2034	2007	2988	432.9	31	6
COCRYS	IVK	Active	3	**3**	83	102	26	537.9	37	6
DUD	decoy_5325	Inactive	1	**4**	7	7	6	539.5	38	5
	ligand_91	Active	39	**5**	2248	2221	3140	412.5	31	6
	ligand_103	Active	49	**6**	3013	2979	3650	477.4	31	6
	decoy_1192	Inactive	5	**7**	70	71	92	499.6	36	5
	ligand_26	Active	2	**8**	14	17	18	533.5	37	6
COCRYS	RR8	Active	7	**9**	29	30	22	533.0	36	6

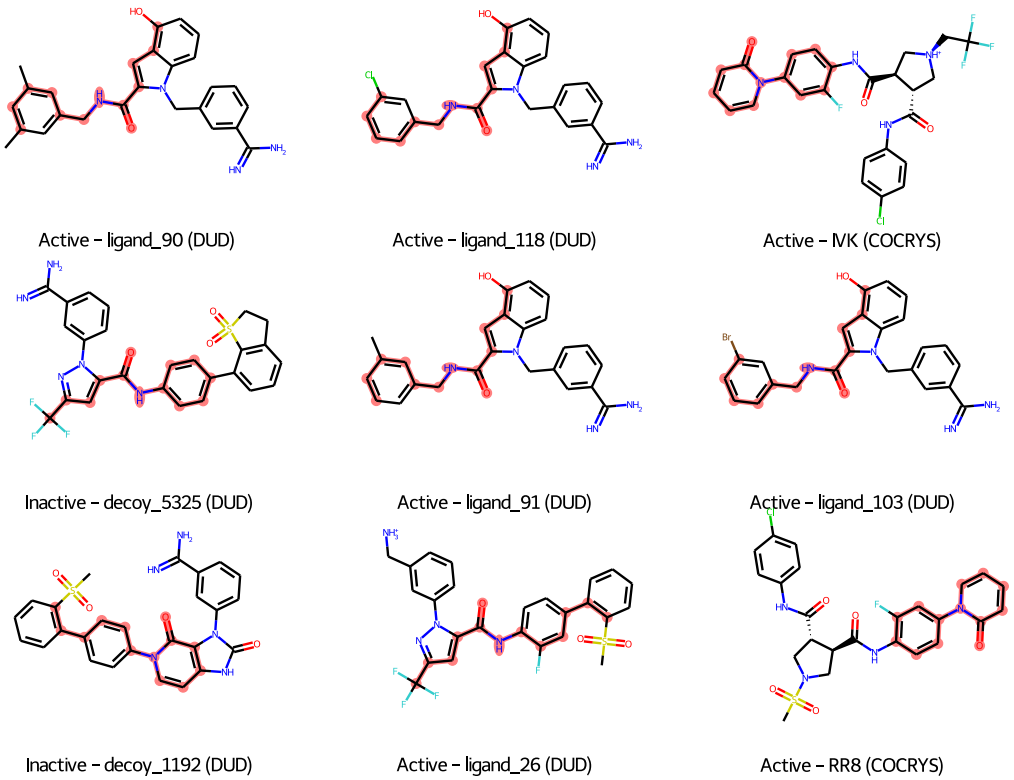

Figure **S17**: **FXa protein.** Structures of the top 9 molecules as judged by the **GBT** classifier after the 30×4cv implementation. The maximum common substructure is highlighted in red (*rdkit*: `completeRingsOnly = False`). An extended version can be consulted in the GitHub repository.

Table **S9**: **FXa protein. Top 9 molecules as judged by the csMIN classifier after the 30×4cv implementation.** Ranks according to the other methods are also shown.

Library	**Lig. Name**	Activity	rank LR	rank GBT	rank csAVG	rank csGEO	**rank csMIN**	MW	Num. atoms	Num. rots
COCRYS	LGJ	Active	34	56	2	2	**1**	586.5	41	6
	LGL	Active	45	28	1	1	**2**	600.6	44	6
	LGK	Active	76	31	15	19	**3**	594.6	42	5
	LGM	Active	53	13	24	26	**4**	546.7	41	6
	FFG	Active	191	129	116	117	**5**	619.6	45	7
DUD	decoy_5325	Inactive	1	4	7	7	**6**	539.5	38	5
	decoy_4894	Inactive	15	35	8	8	**7**	498.6	36	5
	decoy_4854	Inactive	57	50	27	26	**8**	490.6	35	8
DEKOIS	ligand_4	Active	40	37	5	5	**9**	475.6	34	6

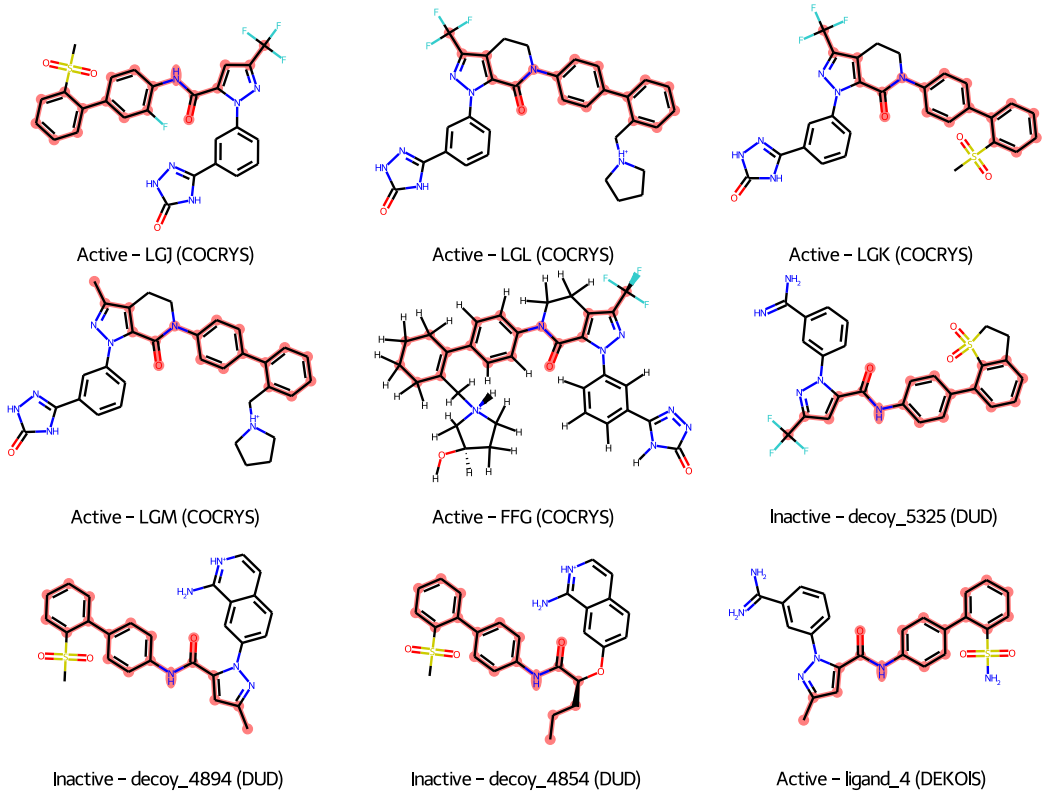

Figure **S18**: **FXa protein.** Structures of the top 9 molecules as judged by the **csMIN** classifier after the 30×4cv implementation. The maximum common substructure is highlighted in red (*rdkit*: `completeRingsOnly = False`). An extended version can be consulted in the GitHub repository.

Table **S10**: **EGFR protein. Top 9 molecules as judged by the GBT classifier after the 30×4cv implementation.** Ranks according to the other methods are also shown.

Library	**Lig. Name**	Activity	rank LR	**rank GBT**	rank csAVG	rank csGEO	rank csMIN	MW	Num. atoms	Num. rots
DUD	ligand_24	Active	133	**1**	9389	9203	9334	255.7	18	2
	ligand_226	Active	124	**2**	7767	7507	8786	330.2	20	3
	ligand_247	Active	225	**3**	8692	8514	10269	335.6	19	2
	ligand_71	Active	232	**4**	8187	7953	10432	331.2	20	3
	ligand_18	Active	225	**5**	9448	9264	9125	235.3	18	2
	ligand_27	Active	323	**6**	10200	9998	9901	300.2	18	2
	ligand_140	Active	279	**7**	6589	6429	8893	266.3	20	3
	ligand_185	Active	111	**8**	7103	6905	6434	329.2	20	3
	ligand_162	Active	107	**9**	10086	9832	11624	331.2	20	3

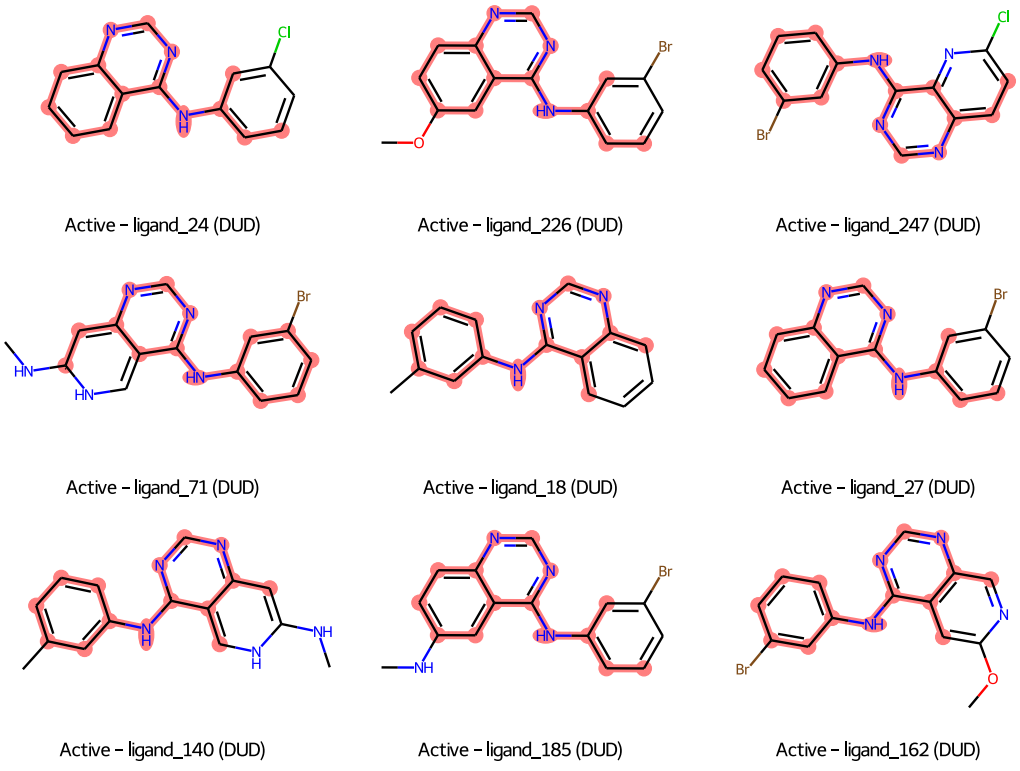

Figure **S19**: **EGFR protein.** Structures of the top 9 molecules as judged by the **GBT** classifier after the 30×4cv implementation. The maximum common substructure is highlighted in red (*rdkit*: `completeRingsOnly = False`). An extended version can be consulted in the GitHub repository.

Table **S11**: **EGFR protein. Top 9 molecules as judged by the csMIN classifier after the 30×4cv implementation.** Ranks according to the other methods are also shown.

Library	**Lig. Name**	Activity	rank LR	rank GBT	rank csAVG	rank csGEO	**rank csMIN**	MW	Num. atoms	Num. rots
COCRYS	ITI	Active	12475	478	38	40	**1**	588.7	43	7
	W19	Active	333	426	8	6	**2**	512.0	35	6
	FMM	Active	1272	365	10	17	**3.5**	582.1	40	11
DEKOIS	decoy_85	Inactive	1722	243	1	2	**3.5**	493.4	36	8
	decoy_136	Inactive	515	701	249	305	**5**	530.7	38	12
COCRYS	N7Q	Active	1511	215	3	3	**6**	563.1	40	10
	N7B	Active	2084	250	6	8	**7**	551.7	41	9
	03P	Active	1517	313	7	9	**8**	548.0	38	9
	W32	Active	460	247	13	14	**9**	586.0	39	7

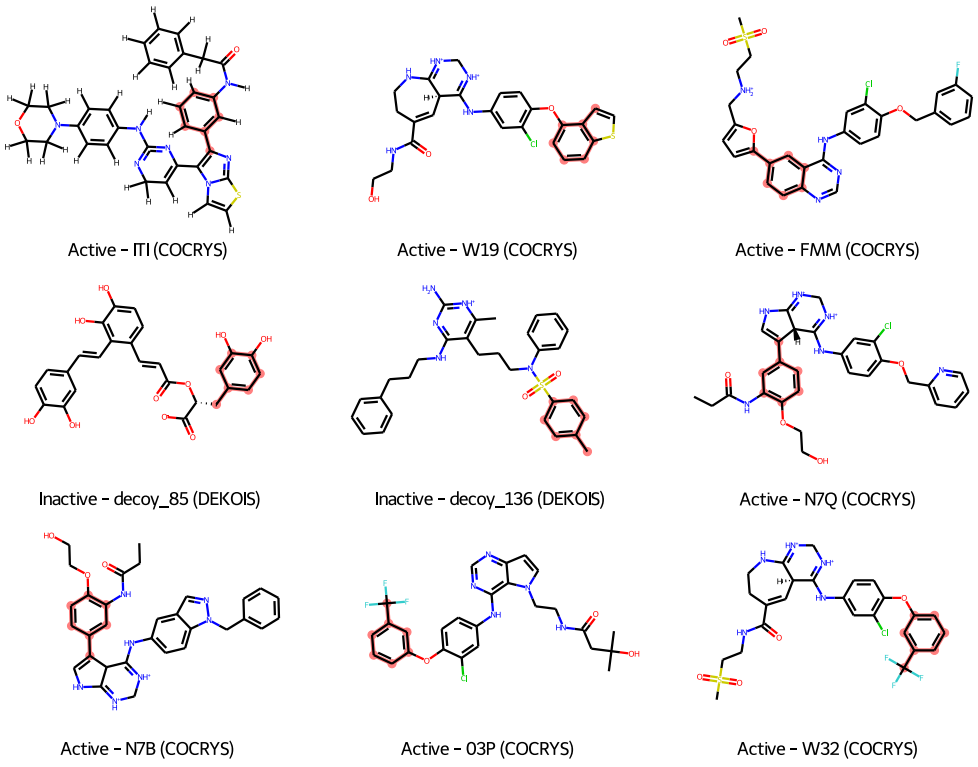

Figure **S20**: **EGFR protein.** Structures of the top 9 molecules as judged by the **csMIN** classifier after the 30×4cv implementation. The maximum common substructure is highlighted in red (*rdkit*: `completeRingsOnly = False`). An extended version can be consulted in the GitHub repository.

Table **S12**: **HSP90 protein. Top 9 molecules as judged by the GBT classifier after the 30×4cv implementation.** Ranks according to the other methods are also shown.

Library	**Lig. Name**	Activity	rank LR	**rank GBT**	rank csAVG	rank csGEO	rank csMIN	MW	Num. atoms	Num. rots
DUD	ligand_33	Active	2	**1**	604	552	1063	432.3	27	5
	ligand_35	Active	5	**2.5**	423	363	1147	387.8	27	5
	ligand_31	Active	12	**2.5**	598	554	1127	373.8	26	4
COCRYS	9ZC	Active	29	**4**	259	217	531	382.4	28	4
DUD	ligand_37	Active	11	**5**	366	310	1227	401.9	28	5
	ligand_25	Active	23	**6**	762	716	600	330.8	23	3
COCRYS	2GG	Active	24	**7**	422	385	832	388.8	27	5
DUD	ligand_5	Active	39	**8**	382	345	537	352.4	26	3
COCRYS	06T	Active	550	**9**	1567	1946	146	494.7	36	2

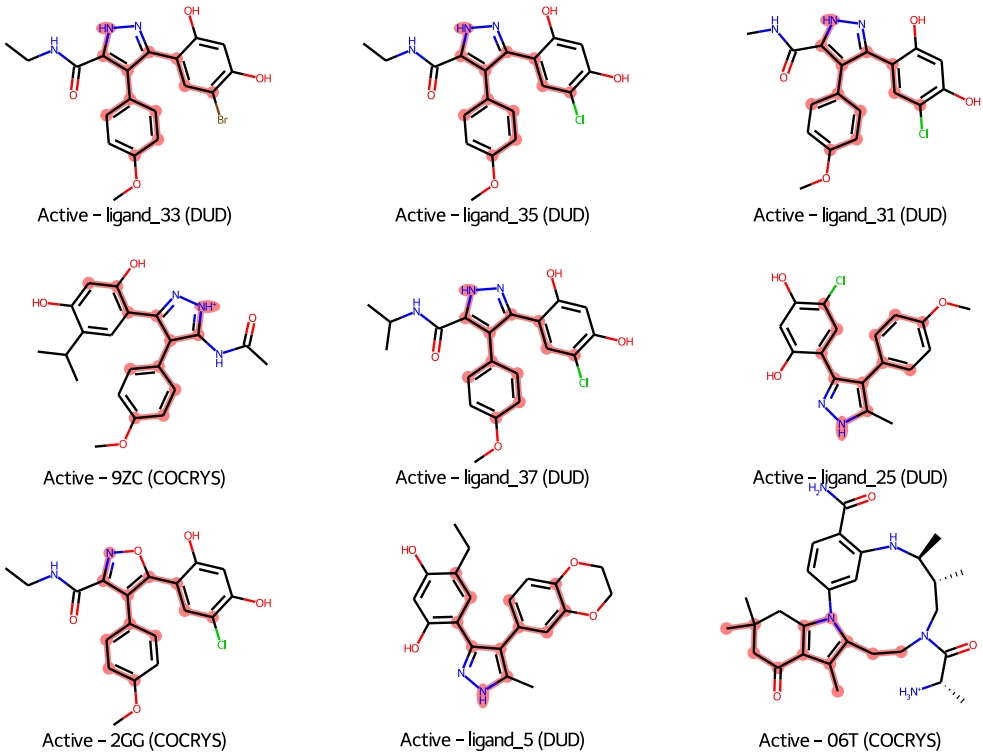

Figure **S21**: **HSP90 protein.** Structures of the top 9 molecules as judged by the **GBT** classifier after the 30×4cv implementation. The maximum common substructure is highlighted in red (*rdkit*: `completeRingsOnly = False`). An extended version can be consulted in the GitHub repository.

Table **S13**: **HSP90 protein. Top 9 molecules as judged by the csMIN classifier after the 30×4cv implementation.** Ranks according to the other methods are also shown.

Library	**Lig. Name**	Activity	rank LR	rank GBT	rank csAVG	rank csGEO	**rank csMIN**	MW	Num. atoms	Num. rots
COCRYS	YKE	Active	66	49	1	1	**1**	453.5	35	3
	73Y	Active	214	56	6	7	**2**	532.4	36	3
	72K	Active	1686	181	75	103	**3**	491.6	37	5
	73S	Active	7	32	3	4	**4**	453.9	33	3
	YKJ	Active	315	40	13	51	**5**	418.5	32	3
	YKI	Active	9	26	7	13	**6**	443.5	34	3
	YKC	Active	1440	205	74	141	**7**	403.4	31	3
	YKB	Active	112	33	22	68	**8**	397.4	30	5
DEKOIS	decoy_1044	Inactive	2233	466	101	143	**9**	496.6	37	9

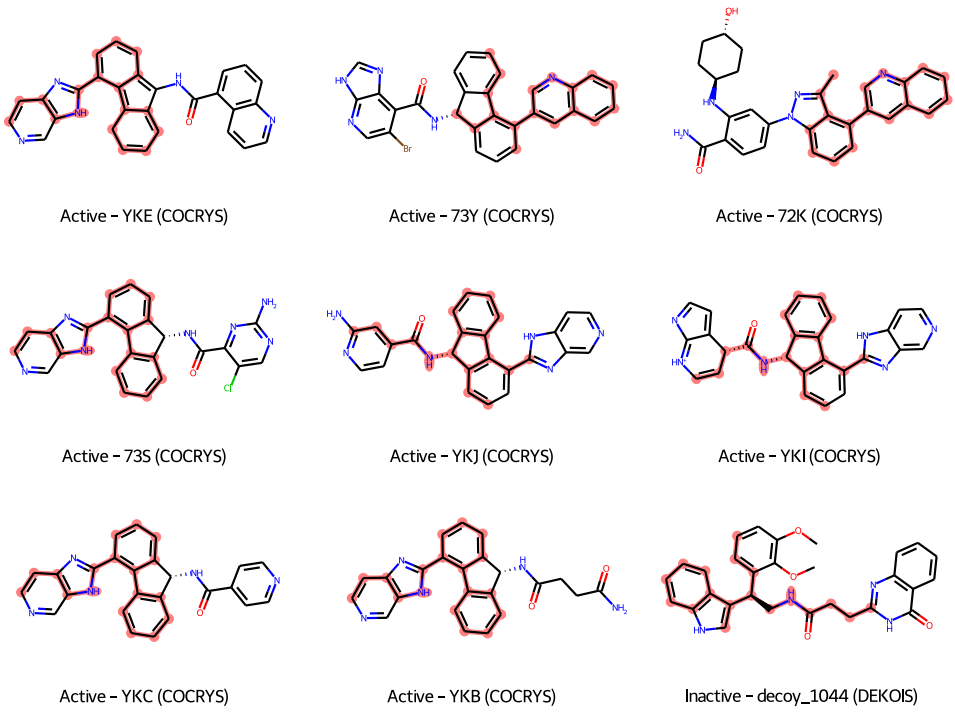

Figure **S22**: **HSP90 protein.** Structures of the top 9 molecules as judged by the **csMIN** classifier after the 30×4cv implementation. The maximum common substructure is highlighted in red (*rdkit*: `completeRingsOnly = False`). An extended version can be consulted in the GitHub repository.

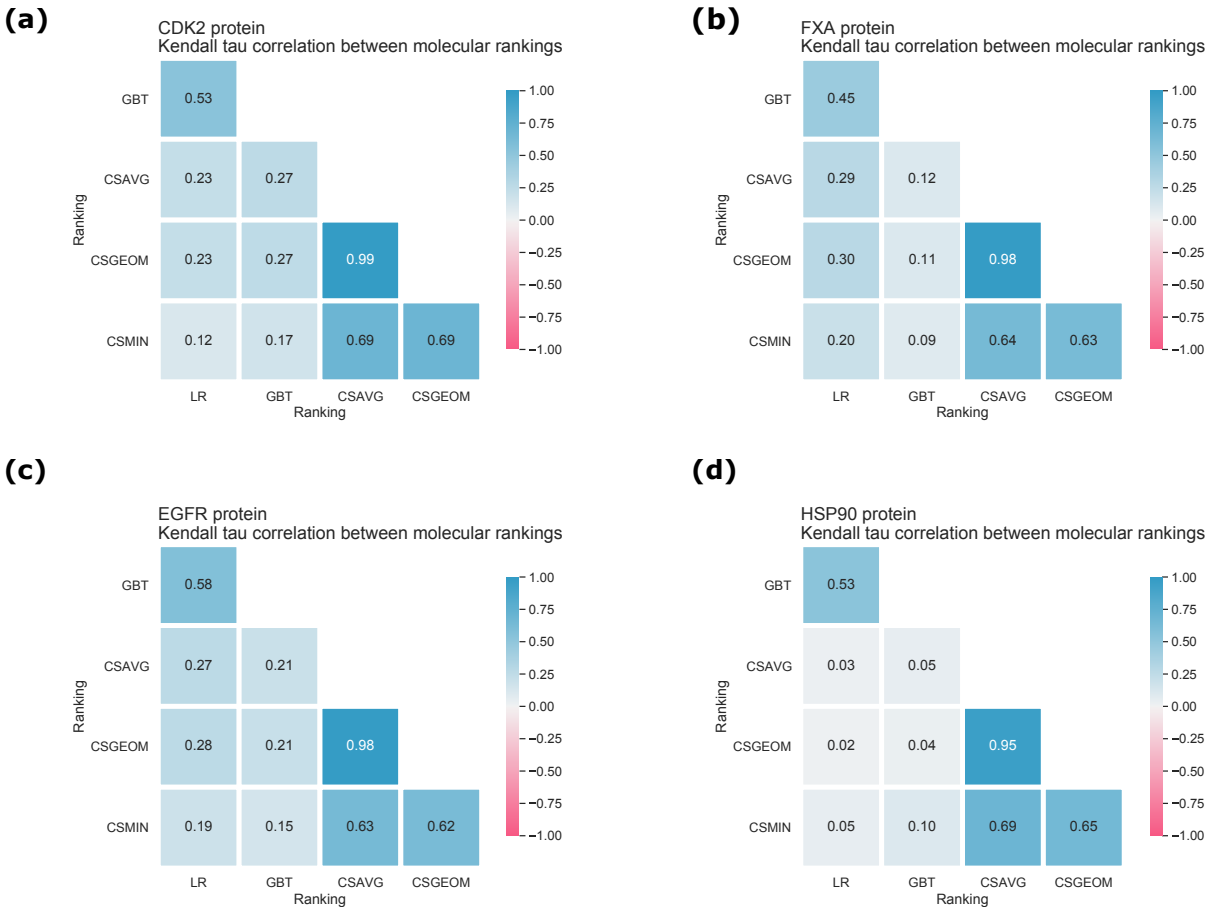

Figure **S23**: **Pairwise Kendall's tau correlations between molecular rankings obtained by the machine learning classifiers and the consensus strategies.** **(a)** CDK2 protein results. **(b)** FXa protein results. **(c)** EGFR protein results. **(d)** HSP90 protein results. The virtual screening methods compared are: *LR*: Logistic Regression; *GBT*: Gradient Boosting Trees; *csAVG*: average; *csGEO*: geometric mean; *csMIN*: minimum.

References

(1) Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **2011**,

(2) Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

(3) Shapiro, S. S.; Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52*, 591–611.

(4) Mauchly, J. W. Significance Test for Sphericity of a Normal n-Variate Distribution. *The Annals of Mathematical Statistics* **1940**, *11*, 204–209.

(5) Kruskal, W. H.; Wallins, W. W. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association* **1952**, *47*, 583–621.

(6) Dunn, O. Multiple comparisons using rank sums. *Technometrics* **1964**, *6*, 241–256.

(7) Pisani, P.; Caporuscio, F.; Carlino, L.; Rastelli, G. Molecular dynamics simulations and classical multidimensional scaling unveil new metastable states in the conformational landscape of CDK2. *PLoS ONE* **2016**, *11*, 1–22.

(8) Wagner, J. R.; Sørensen, J.; Hensley, N.; Wong, C.; Zhu, C.; Perison, T.; Amaro, R. E. POVME 3.0: Software for Mapping Binding Pocket Flexibility. *Journal of Chemical Theory and Computation* **2017**, *13*, 4584–4592.