

De novo prediction of drug targets and candidates by chemical similarity-guided network-based inference: Supplementary information

Carlos Vigil-Vásquez and Andreas Schüller

S1 *de novo* network-based inference method

Denoting a set of $N_{\mathcal{F}}$ features as $F = \{f_1, f_2, \dots, f_{N_{\mathcal{F}-1}}, f_{N_{\mathcal{F}}}\}$, a set of $N_{\mathcal{C}}$ novel compounds without known targets as $\mathcal{C} = \{c_1, c_2, \dots, c_{N_{\mathcal{C}-1}}, c_{N_{\mathcal{C}}}\}$, a set of $N_{\mathcal{D}}$ drugs with known targets as $\mathcal{D} = \{d_1, d_2, \dots, d_{N_{\mathcal{D}-1}}, d_{N_{\mathcal{D}}}\}$, and a set of $N_{\mathcal{T}}$ biological targets as $\mathcal{T} = \{t_1, t_2, \dots, t_{N_{\mathcal{T}-1}}, t_{N_{\mathcal{T}}}\}$, a compound-feature-drug-target graph can be represented as the graph $G(V, E)$, where $V = \mathcal{C} \cup \mathcal{F} \cup \mathcal{D} \cup \mathcal{T}$ is the node set and $E = E_{\mathcal{FC}} \cup E_{\mathcal{FD}} \cup E_{\mathcal{DT}}$ is the edge set constructed from the edges between features and novel compounds, between features and drugs and between drugs and targets, respectively.

The initial resources in graph G can be represented as a symmetric adjacency matrix of order $N_{\mathcal{C}} + N_{\mathcal{F}} + N_{\mathcal{D}} + N_{\mathcal{T}}$ given by

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & \mathbf{M}_{\mathcal{CF}} & 0 \\ 0 & 0 & \mathbf{M}_{\mathcal{DF}} & \mathbf{M}_{\mathcal{DT}} \\ \mathbf{M}_{\mathcal{CF}}^T & \mathbf{M}_{\mathcal{DF}}^T & 0 & 0 \\ 0 & \mathbf{M}_{\mathcal{DT}}^T & 0 & 0 \end{bmatrix} \quad (1)$$

where $\mathbf{M}_{\mathcal{CF}}$ is the adjacency matrix for the interactions between compounds and features, $\mathbf{M}_{\mathcal{DF}}$ is the adjacency matrix for the interactions between drugs and features and $\mathbf{M}_{\mathcal{DT}}$ is the adjacency matrix for the interactions between drugs and biological targets.

From \mathbf{R} , one can construct an adjacency matrix for a graph that only contains drugs with known interactions with targets and the features for the selected drugs. This adjacency matrix, that we will denote with \mathbf{R}_0 , is given by

$$\mathbf{R}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{M}_{\mathcal{DF}} & \mathbf{M}_{\mathcal{DT}} \\ 0 & \mathbf{M}_{\mathcal{DF}}^T & 0 & 0 \\ 0 & \mathbf{M}_{\mathcal{DT}}^T & 0 & 0 \end{bmatrix} \quad (2)$$

From \mathbf{R}_0 , the transfer matrix \mathbf{W} of order $N_{\mathcal{C}} + N_{\mathcal{F}} + N_{\mathcal{D}} + N_{\mathcal{T}}$ is defined by

$$\mathbf{W}(i, j) = \begin{cases} \frac{\mathbf{R}_0(i, j)}{\sum_{l=1}^{N_{\mathcal{F}}+N_{\mathcal{C}}+N_{\mathcal{D}}+N_{\mathcal{T}}} \mathbf{R}_0(i, l)}, & \text{if } \sum_{l=1}^{N_{\mathcal{F}}+N_{\mathcal{C}}+N_{\mathcal{D}}+N_{\mathcal{T}}} \mathbf{R}_0(i, l) \neq 0, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

The final resource matrix \mathbf{R}_1 , which correspond to the predicted targets for all the chemical compounds present in the graph G is obtained by

$$\mathbf{R}_1 = \mathbf{R} \times \mathbf{W}^k \quad (4)$$

where \mathbf{R} is the initial resource matrix, \mathbf{W} is the transfer matrix and k is the number of resource spreading processes. Based on previous studies, the number of resource-spreading processes used in this study is equal to 2 ($k = 2$).

S2 Dataset preparation

S2.1 ChEMBL datasets entries selection criteria

From ChEMBL versions 24 (June, 2018) and 28 (February, 2021), all interactions that meet the following criteria are collected:

1. Species identifier equals Homo Sapiens;
2. Assay type equals "protein-ligand binding";
3. K_i , K_d , IC_{50} or EC_{50} between 0 μ M and less than 10 μ M;
4. Molecule must have more than 5 and less than 80 heavy atoms;
5. Activity comment must denote that the ligand is active for annotated target;
6. Activity value must be "minor", "much less", "less than or equal to", "equal" or "equal to equal" (represented as "<", "<<", "<=", "=", and "==", respectively) to the value obtained experimentally and
7. Ligand must have a maximum clinical phase equal to or greater than 1.

S2.2 Tanimoto distributions

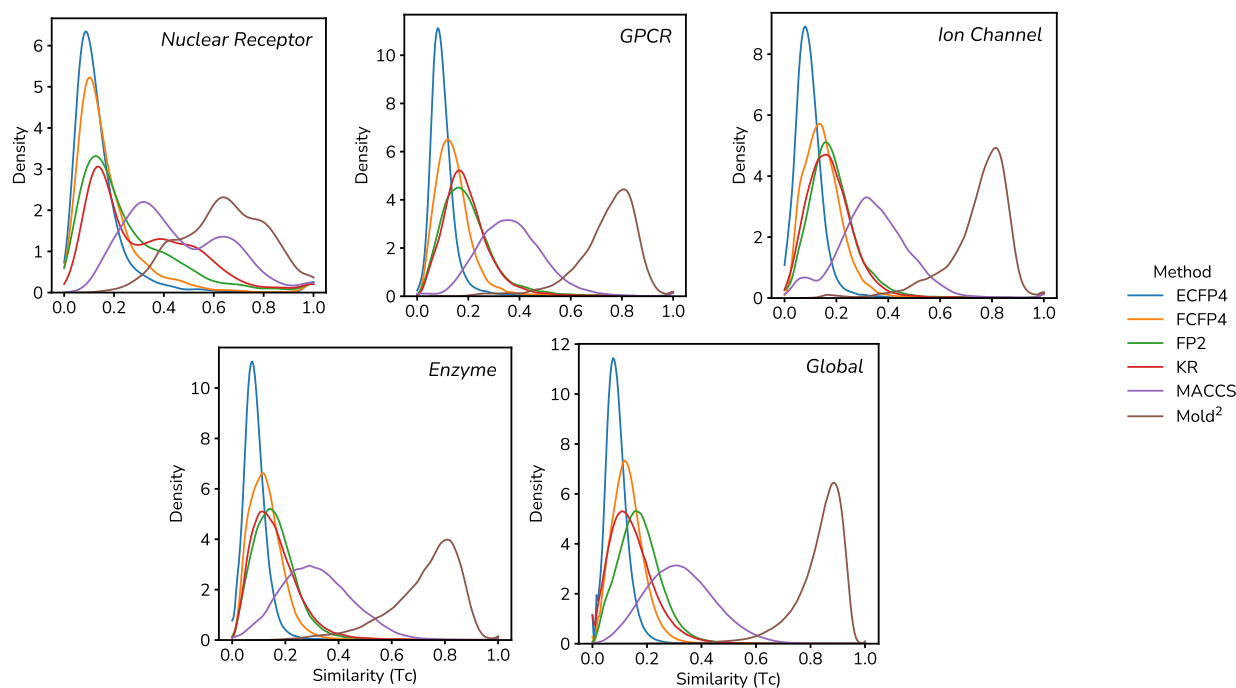


Figure S1: **Tanimoto distribution of the studied datasets and fingerprints.** Kernel density estimate (KDE) of the chemical similarity observed for each molecular descriptor throughout the 5 internal validation datasets

S2.3 Additional information for molecular descriptors

Mold² descriptor was obtained online from <https://www.fda.gov/science-research/bioinformatics-tools/mold2>, accessed on 20 March 2022.

S3 Evaluation metrics description

S3.1 Overall performance

S3.1.1 Area under the Receiver Operating Characteristic curve (AuROC)

To calculate the area under the receiver operating characteristic curve (AuROC), first we calculate the true-positive rate (TPR) and false-positive rate (FPR) for each discrimination thresholds to then calculate the area under the curve constructed from this values. TPR and FPR are given by

$$TPR = \frac{TP}{TP + FN} \quad (5)$$

$$FPR = \frac{FP}{FP + TN} \quad (6)$$

where TP corresponds to the true-positive predictions, TN to the true-negative predictions, FP to the false-positive predictions and FN to the false-negative predictions.

This metric was calculated using the function `roc_auc_score` from the scikit-learn Python package.

S3.1.2 Area under the Precision-Recall curve (AuPRC)

To calculate the area under the precision-recall curve (AuPRC), first we calculate the precision and recall for each discrimination thresholds to then calculate the area under the curve constructed from this values. Precision and recall are given by

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

where TP corresponds to the true-positive predictions, FP to the false-positive predictions and FN to the false-negative predictions.

This metric was calculated using the function `average_precision_score` from the scikit-learn Python package.

S3.2 Early-recognition performance

In virtual screening, only the best predictions obtained from a model are selected for posterior experimental validation. Therefore, understanding the predictive performance of a model for these predictions is essential to (1) make accurate predictions that will translate to biological activity and (2) understand the limitations of the model. The metrics discussed here can be evaluated at a given cut-off rank, considering only the topmost results returned by the predictive method, hence informing of the predictive performance of the model for only the best predictions.

S3.2.1 Precision and Recall at L (P@L and R@L)

These metrics correspond to the precision and recall of the best L predictions obtained from a model. In this study we calculate these metrics over the ligands, i.e., we calculate the precision and recall for the best L prediction for each ligand in the test set and average them to have a single score. Precision at L (P@L) and Recall at L (R@L) are given by

$$P@L = \frac{1}{N_D} \sum_{i=1}^D \frac{X_i(L)}{L} \quad (9)$$

$$R@L = \frac{1}{N_D} \sum_{i=1}^D \frac{X_i(L)}{X_i} \quad (10)$$

where L is the number of top predictions to evaluate, ND corresponds to the number of ligands that participated in the evaluation (equal to 1 in leave-one-out cross-validation), Xi is the number of drug-target interactions removed for a given drug in the data splitting procedure and Xi(L) are the number of true-positives in the top L predictions. In this study we employed a length of the list equal to 10 (L = 20) based in previous studies [1–5].

This metric was calculated using an in-house implementation in Python.

S3.2.2 Boltzmann-Enhanced Discrimination of ROC (AuBEDROC)

BEDROC corresponds to a transformation of the commonly used Receiver Operating Characteristic in order to quantify the early recognition capability of a predictive model. This metric differentially weights the entries based on their position in the predictions list, giving more importance to the first entries in a list, i.e., the ones with higher probabilities or predictive scores. BEDROC is given by

$$BEDROC = \frac{RIE}{\alpha} + \frac{1}{1 - e^{\alpha}} \quad (11)$$

where RIE corresponds to the robust initial enhancement proposed by Sheridan et al. (2001) and corresponds to the early recognition weighting parameter. Based in previous studies [6], a value of 20 was employed for α .

This metric was calculated using an in-house implementation in Python.

S3.3 Binary prediction performance

The last common practice in virtual screening is to assign a score or probability threshold for the predictions obtained from a model and manually select or cherry-pick predictions for experimental validation. In order to evaluate the predictive performance under this paradigm, we calculated the score for each metric at all possible discrimination thresholds and then retrieved the maximum observed value.

S3.3.1 Matthews correlation coefficient (MCC)

Evaluation metric that measures the correlation between the true and predicted binary classification, where values of -1 indicate total disagreement between truth and prediction, 0 indicates a random prediction and +1 indicates total agreement between truth and prediction. MCC is given by

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (12)$$

where TP corresponds to the true-positive predictions, TN to the true-negative predictions, FP to the false-positive predictions and FN to the false-negative predictions.

This metric was calculated using an in-house implementation in Python.

S3.3.2 F₁ score (F₁)

Evaluation metric that measures the classification performance as the harmonic mean between the precision and recall scores. The highest possible value of an F-score is 1.0, indicating perfect precision and recall, and the lowest possible value is 0, if either the precision or the recall is zero. F1 score is given by

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (13)$$

where TP corresponds to the true-positive predictions, FP to the false-positive predictions and FN to the false-negative predictions.

This metric was calculated using an in-house implementation in Python.

S3.3.3 Balanced Accuracy (bACC)

Balanced accuracy (bACC) corresponds to the arithmetic mean of sensitivity (also known as true-positive rate or recall) and specificity (also known as true-negative rate) and its use case is assessing binary or multi-class classification performance in imbalanced datasets, i.e., one of the classes is over-represented in the dataset. bACC is given by

$$bACC = \frac{TPR + TNR}{2} \quad (14)$$

where TPR corresponds to the true-positive rate and TNR to the true-negative rate.

True-negative rate (TNR) is given by

$$TNR = \frac{TN}{TN + FP} \quad (15)$$

where TN corresponds to the true-negative predictions and FP to the false-positive predictions.

This metric was calculated using a in-house implementation in Python.

S4 Predictive performance

S4.1 Leave-one-out cross-validation

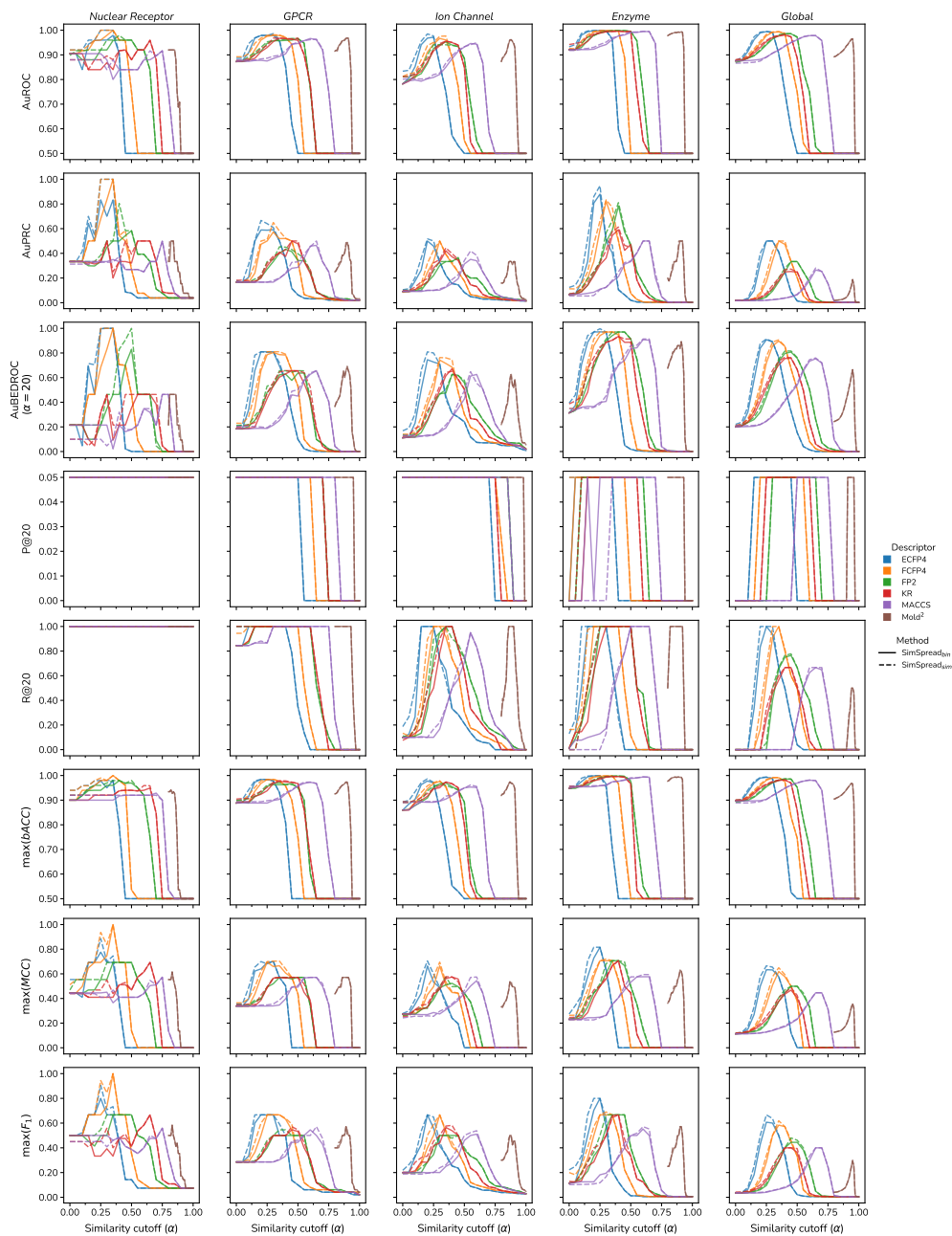


Figure S2: α parameter optimization over LOO CV. Median score for metrics obtained for LOO CV over 5 datasets using values between 0 and 1 with a step size of 0.05. A total of six different molecular descriptors were used for the method optimization: ECFP4 (blue), FCFP4 (orange), FP2 (green), KR (red), MACCS keys (purple) and Mold2 (brown). Solid lines correspond to SimSpread_{bin} while dashed lines correspond to SimSpread_{sim}.

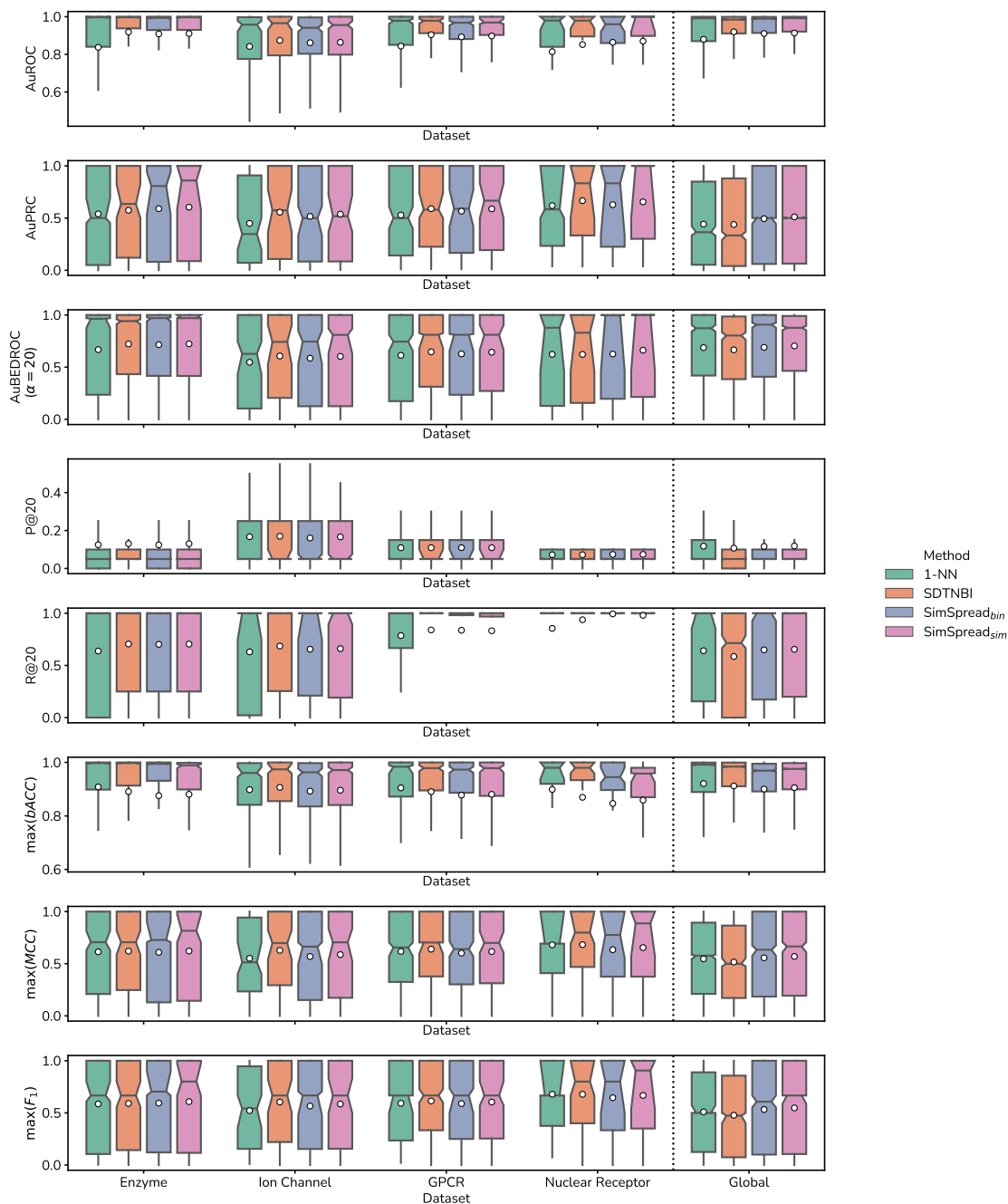


Figure S3: **Predictive performance comparison between the studied methods in LOO CV.** Boxplots for AuROC, AuPRC, AuBEDROC, P@20, R@20, max(MCC), max(F₁) and max(bACC) obtained in the LOO CV over the five validation datasets for 1-NN, NBI, SDTNBI, SimSpread_{bin} and SimSpread_{sim} using ECFP4 molecular descriptor and Tanimoto coefficient as similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the performance of the method for that metric. Boxplot notch represents the bootstrapped 95% CI around the median, whiskers correspond to 1.5x IQR and white dot represents the mean.

Table S1: **Overall average performance for the studied methods in LOO CV.** Mean \pm SD for AuROC and AuPRC observed in LOO CV over the five validation datasets using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the mean performance of the method for that metric.

Dataset	Method	AuROC	AuPRC
NR	NN	0.814 \pm 0.326	0.619 \pm 0.377
	SDTNBI	0.852 \pm 0.284	0.667 \pm 0.373
	SimSpread _{bin}	0.864 \pm 0.209	0.628 \pm 0.401
	SimSpread _{sim}	0.870 \pm 0.210	0.656 \pm 0.397
IC	NN	0.842 \pm 0.231	0.449 \pm 0.392
	SDTNBI	0.874 \pm 0.186	0.556 \pm 0.409
	SimSpread _{bin}	0.862 \pm 0.183	0.516 \pm 0.411
	SimSpread _{sim}	0.864 \pm 0.192	0.537 \pm 0.416
GPCR	NN	0.844 \pm 0.277	0.528 \pm 0.385
	SDTNBI	0.905 \pm 0.177	0.590 \pm 0.376
	SimSpread _{bin}	0.893 \pm 0.175	0.566 \pm 0.397
	SimSpread _{sim}	0.898 \pm 0.174	0.588 \pm 0.400
Enzyme	NN	0.837 \pm 0.299	0.540 \pm 0.435
	SDTNBI	0.919 \pm 0.168	0.574 \pm 0.417
	SimSpread _{bin}	0.909 \pm 0.184	0.590 \pm 0.428
	SimSpread _{sim}	0.911 \pm 0.184	0.604 \pm 0.428
Global	NN	0.880 \pm 0.231	0.442 \pm 0.385
	SDTNBI	0.920 \pm 0.151	0.438 \pm 0.396
	SimSpread _{bin}	0.911 \pm 0.168	0.494 \pm 0.400
	SimSpread _{sim}	0.913 \pm 0.168	0.512 \pm 0.405

Table S2: **Overall median performance for the studied methods in LOO CV.** Q1, median and Q3 for AuROC and AuPRC observed in LOO CV over the five validation datasets using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the mean performance of the method for that metric.

Dataset	Method	AuROC			AuPRC		
		Q1	median	Q3	Q1	median	Q3
NR	NN	0.840	0.980	1.000	0.233	0.583	1.000
	SDTNBI	0.895	0.979	1.000	0.333	0.833	1.000
	SimSpread _{bin}	0.860	0.960	1.000	0.225	0.833	1.000
	SimSpread _{sim}	0.898	1.000	1.000	0.301	1.000	1.000
IC	NN	0.776	0.958	0.998	0.071	0.347	0.909
	SDTNBI	0.795	0.966	1.000	0.108	0.574	1.000
	SimSpread _{bin}	0.803	0.941	0.996	0.083	0.500	1.000
	SimSpread _{sim}	0.798	0.956	1.000	0.084	0.518	1.000
GPCR	NN	0.850	0.978	1.000	0.142	0.500	1.000
	SDTNBI	0.913	0.979	1.000	0.225	0.580	1.000
	SimSpread _{bin}	0.882	0.968	1.000	0.167	0.589	1.000
	SimSpread _{sim}	0.903	0.969	1.000	0.193	0.667	1.000
Enzyme	NN	0.840	0.997	1.000	0.050	0.500	1.000
	SDTNBI	0.938	0.997	1.000	0.120	0.635	1.000
	SimSpread _{bin}	0.929	0.994	1.000	0.081	0.807	1.000
	SimSpread _{sim}	0.930	0.997	1.000	0.088	0.860	1.000
Global	NN	0.870	0.992	1.000	0.053	0.364	0.850
	SDTNBI	0.911	0.985	0.999	0.040	0.333	0.880
	SimSpread _{bin}	0.914	0.989	0.999	0.060	0.500	1.000
	SimSpread _{sim}	0.921	0.992	1.000	0.063	0.500	1.000

Table S3: **Early-recognition average performance for the studied methods in LOO CV.** Mean \pm SD for AuBEDROC, P@20 and R@20 observed in LOO CV over the five validation datasets using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the performance of the method for that metric.

Dataset	Method	AuBEDROC	P@20	R@20
NR	NN	0.623 \pm 0.423	0.072 \pm 0.073	0.855 \pm 0.344
	SDTNBI	0.622 \pm 0.418	0.072 \pm 0.069	0.938 \pm 0.238
	SimSpread _{bin}	0.626 \pm 0.425	0.074 \pm 0.063	0.994 \pm 0.028
	SimSpread _{sim}	0.662 \pm 0.414	0.074 \pm 0.072	0.980 \pm 0.098
IC	NN	0.547 \pm 0.406	0.167 \pm 0.222	0.630 \pm 0.439
	SDTNBI	0.607 \pm 0.398	0.170 \pm 0.199	0.686 \pm 0.408
	SimSpread _{bin}	0.585 \pm 0.408	0.161 \pm 0.211	0.655 \pm 0.415
	SimSpread _{sim}	0.604 \pm 0.411	0.166 \pm 0.217	0.661 \pm 0.424
GPCR	NN	0.613 \pm 0.393	0.109 \pm 0.127	0.786 \pm 0.372
	SDTNBI	0.647 \pm 0.371	0.109 \pm 0.122	0.840 \pm 0.328
	SimSpread _{bin}	0.628 \pm 0.391	0.110 \pm 0.125	0.837 \pm 0.331
	SimSpread _{sim}	0.643 \pm 0.391	0.109 \pm 0.127	0.832 \pm 0.336
Enzyme	NN	0.668 \pm 0.408	0.125 \pm 0.223	0.638 \pm 0.455
	SDTNBI	0.722 \pm 0.369	0.131 \pm 0.218	0.706 \pm 0.423
	SimSpread _{bin}	0.716 \pm 0.390	0.124 \pm 0.215	0.703 \pm 0.426
	SimSpread _{sim}	0.723 \pm 0.392	0.131 \pm 0.225	0.705 \pm 0.425
Global	NN	0.687 \pm 0.362	0.118 \pm 0.166	0.641 \pm 0.421
	SDTNBI	0.666 \pm 0.348	0.107 \pm 0.166	0.586 \pm 0.432
	SimSpread _{bin}	0.690 \pm 0.374	0.116 \pm 0.169	0.650 \pm 0.421
	SimSpread _{sim}	0.704 \pm 0.343	0.119 \pm 0.174	0.655 \pm 0.419

Table S4: **Early-recognition median performance for the studied methods in LOO CV.** Q1, median and Q3 for AuBEDROC, P@20 and R@20 observed in LOO CV over the five validation datasets using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the performance of the method for that metric.

Dataset	Method	AuBEDROC			P@20			R@20		
		Q1	median	Q3	Q1	median	Q3	Q1	median	Q3
NR	NN	0.128	0.878	1.000	0.05	0.05	0.10	1.000	1.000	1.000
	SDTNBI	0.157	0.830	1.000	0.05	0.05	0.10	1.000	1.000	1.000
	SimSpread _{bin}	0.196	0.996	1.000	0.05	0.05	0.10	1.000	1.000	1.000
	SimSpread _{sim}	0.215	1.000	1.000	0.05	0.05	0.10	1.000	1.000	1.000
IC	NN	0.103	0.627	1.000	0.05	0.05	0.25	0.022	1.000	1.000
	SDTNBI	0.205	0.742	1.000	0.05	0.05	0.25	0.254	1.000	1.000
	SimSpread _{bin}	0.125	0.745	1.000	0.05	0.05	0.25	0.211	1.000	1.000
	SimSpread _{sim}	0.126	0.809	1.000	0.05	0.05	0.25	0.192	1.000	1.000
GPCR	NN	0.174	0.745	1.000	0.05	0.05	0.15	0.667	1.000	1.000
	SDTNBI	0.312	0.810	1.000	0.05	0.05	0.15	1.000	1.000	1.000
	SimSpread _{bin}	0.235	0.810	1.000	0.05	0.05	0.15	0.982	1.000	1.000
	SimSpread _{sim}	0.272	0.810	1.000	0.05	0.05	0.15	0.969	1.000	1.000
Enzyme	NN	0.236	0.964	1.000	0.00	0.05	0.10	0.000	1.000	1.000
	SDTNBI	0.433	0.942	1.000	0.05	0.05	0.10	0.250	1.000	1.000
	SimSpread _{bin}	0.417	0.970	1.000	0.00	0.05	0.10	0.250	1.000	1.000
	SimSpread _{sim}	0.416	0.970	1.000	0.00	0.05	0.10	0.250	1.000	1.000
Global	NN	0.419	0.873	1.000	0.05	0.05	0.15	0.156	1.000	1.000
	SDTNBI	0.385	0.803	0.986	0.00	0.05	0.10	0.000	0.714	1.000
	SimSpread _{bin}	0.408	0.908	1.000	0.05	0.05	0.10	0.173	1.000	1.000
	SimSpread _{sim}	0.464	0.877	0.990	0.05	0.05	0.10	0.200	1.000	1.000

Table S5: **Binary prediction average performance for the studied methods in LOO CV.** Mean \pm SD for max(bACC), max(MCC) and max(F₁) observed in LOO CV over the five validation datasets using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the performance of the method for that metric.

Dataset	Method	max(bACC)	max(MCC)	max(F ₁)
NR	NN	0.899 \pm 0.164	0.681 \pm 0.348	0.679 \pm 0.342
	SDTNBI	0.870 \pm 0.255	0.683 \pm 0.361	0.678 \pm 0.362
	SimSpread _{bin}	0.847 \pm 0.262	0.635 \pm 0.410	0.645 \pm 0.387
	SimSpread _{sim}	0.859 \pm 0.244	0.655 \pm 0.413	0.667 \pm 0.387
IC	NN	0.898 \pm 0.137	0.552 \pm 0.349	0.522 \pm 0.375
	SDTNBI	0.907 \pm 0.146	0.628 \pm 0.355	0.604 \pm 0.381
	SimSpread _{bin}	0.893 \pm 0.157	0.569 \pm 0.397	0.566 \pm 0.392
	SimSpread _{sim}	0.896 \pm 0.160	0.587 \pm 0.403	0.585 \pm 0.398
GPCR	NN	0.905 \pm 0.145	0.618 \pm 0.341	0.592 \pm 0.359
	SDTNBI	0.891 \pm 0.225	0.640 \pm 0.336	0.614 \pm 0.358
	SimSpread _{bin}	0.878 \pm 0.233	0.602 \pm 0.380	0.589 \pm 0.382
	SimSpread _{sim}	0.881 \pm 0.233	0.617 \pm 0.383	0.604 \pm 0.386
Enzyme	NN	0.909 \pm 0.159	0.614 \pm 0.394	0.586 \pm 0.416
	SDTNBI	0.891 \pm 0.247	0.621 \pm 0.382	0.590 \pm 0.405
	SimSpread _{bin}	0.876 \pm 0.260	0.609 \pm 0.417	0.594 \pm 0.423
	SimSpread _{sim}	0.881 \pm 0.248	0.622 \pm 0.418	0.608 \pm 0.424
Global	NN	0.921 \pm 0.128	0.547 \pm 0.351	0.510 \pm 0.374
	SDTNBI	0.912 \pm 0.186	0.516 \pm 0.357	0.477 \pm 0.382
	SimSpread _{bin}	0.900 \pm 0.187	0.557 \pm 0.377	0.533 \pm 0.388
	SimSpread _{sim}	0.906 \pm 0.187	0.570 \pm 0.381	0.548 \pm 0.392

Table S6: **Binary prediction median performance for the studied methods in LOO CV.** Q1, median and Q3 for max(bACC), max(MCC) and max(F₁) observed in LOO CV over the five validation datasets using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the performance of the method for that metric.

Dataset	Method	max(bACC)			max(MCC)			max(F ₁)		
		Q1	median	Q3	Q1	median	Q3	Q1	median	Q3
NR	NN	0.920	0.980	1.000	0.410	0.693	1.000	0.375	0.667	1.000
	SDTNBI	0.934	0.980	1.000	0.469	0.799	1.000	0.400	0.800	1.000
	SimSpread _{bin}	0.897	0.945	1.000	0.376	0.776	1.000	0.333	0.800	1.000
	SimSpread _{sim}	0.870	0.958	0.980	0.376	0.887	1.000	0.350	0.906	1.000
IC	NN	0.842	0.961	0.998	0.235	0.514	0.943	0.155	0.542	0.947
	SDTNBI	0.855	0.974	1.000	0.294	0.698	1.000	0.220	0.667	1.000
	SimSpread _{bin}	0.835	0.963	0.997	0.152	0.664	1.000	0.154	0.667	1.000
	SimSpread _{sim}	0.841	0.972	1.000	0.173	0.705	1.000	0.156	0.667	1.000
GPCR	NN	0.872	0.984	1.000	0.325	0.656	1.000	0.235	0.667	1.000
	SDTNBI	0.895	0.979	1.000	0.378	0.703	1.000	0.333	0.667	1.000
	SimSpread _{bin}	0.887	0.973	1.000	0.302	0.640	1.000	0.250	0.667	1.000
	SimSpread _{sim}	0.875	0.979	1.000	0.313	0.699	1.000	0.254	0.667	1.000
Enzyme	NN	0.898	0.998	1.000	0.210	0.707	1.000	0.105	0.667	1.000
	SDTNBI	0.913	0.998	1.000	0.247	0.706	1.000	0.143	0.667	1.000
	SimSpread _{bin}	0.931	0.995	1.000	0.130	0.728	1.000	0.121	0.704	1.000
	SimSpread _{sim}	0.899	0.989	0.998	0.144	0.816	1.000	0.116	0.800	1.000
Global	NN	0.889	0.992	1.000	0.211	0.577	0.894	0.125	0.500	0.889
	SDTNBI	0.911	0.984	1.000	0.171	0.499	0.866	0.074	0.473	0.857
	SimSpread _{bin}	0.892	0.969	0.995	0.185	0.634	1.000	0.100	0.607	1.000
	SimSpread _{sim}	0.899	0.976	0.998	0.194	0.666	1.000	0.105	0.667	1.000

S4.2 10-times 10-fold cross-validation

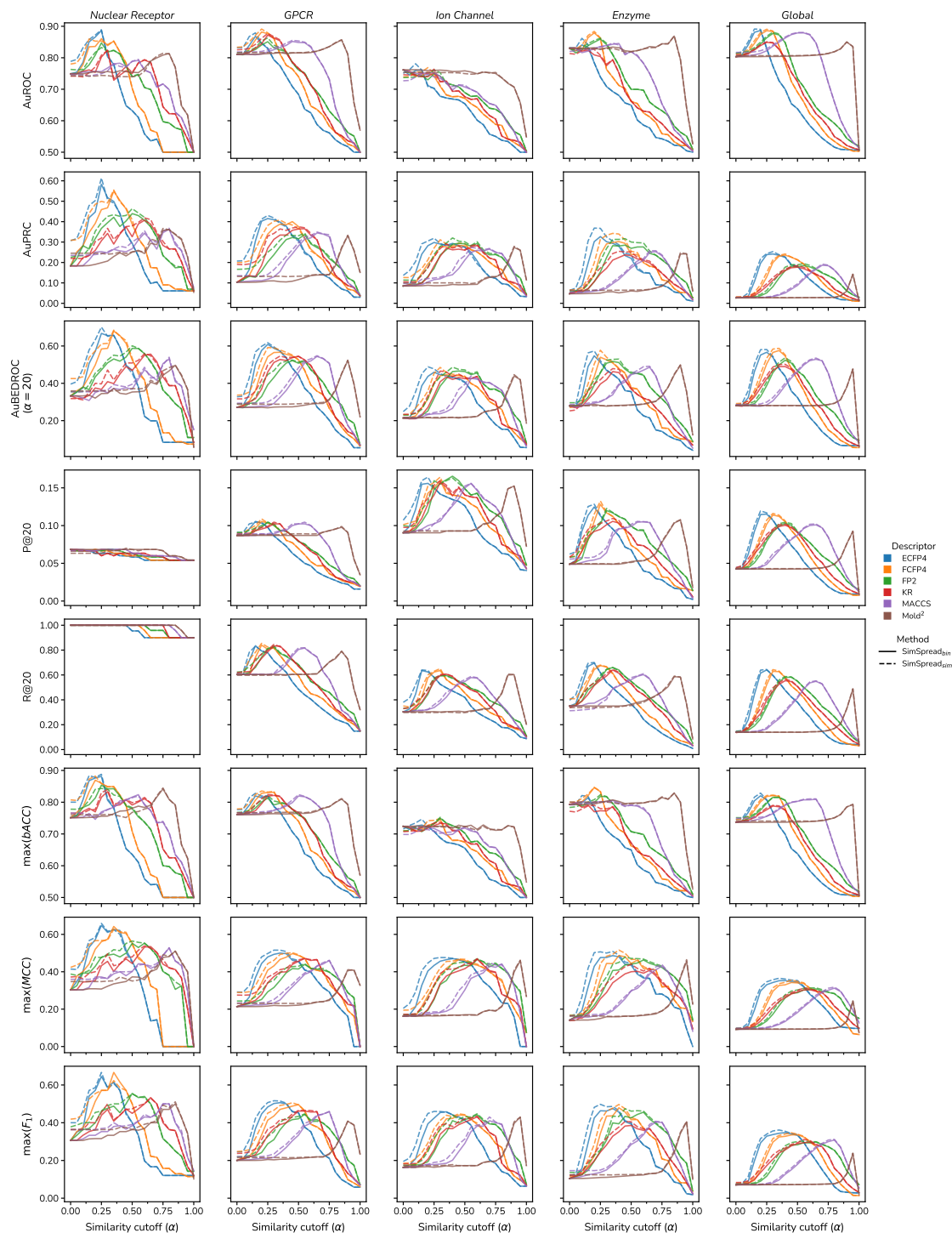


Figure S4: α parameter optimization over 10-times 10-fold CV. Median score for metrics obtained for 10-times 10-fold CV over 5 datasets using values between 0 and 1 with a step size of 0.05. A total of six different molecular descriptors were used for the method optimization: ECFP4 (blue), FCFP4 (orange), FP2 (green), KR (red), MACCS keys (purple) and Mold2 (brown). Solid lines correspond to SimSpread_{bin} while dashed lines correspond to SimSpread_{sim}.

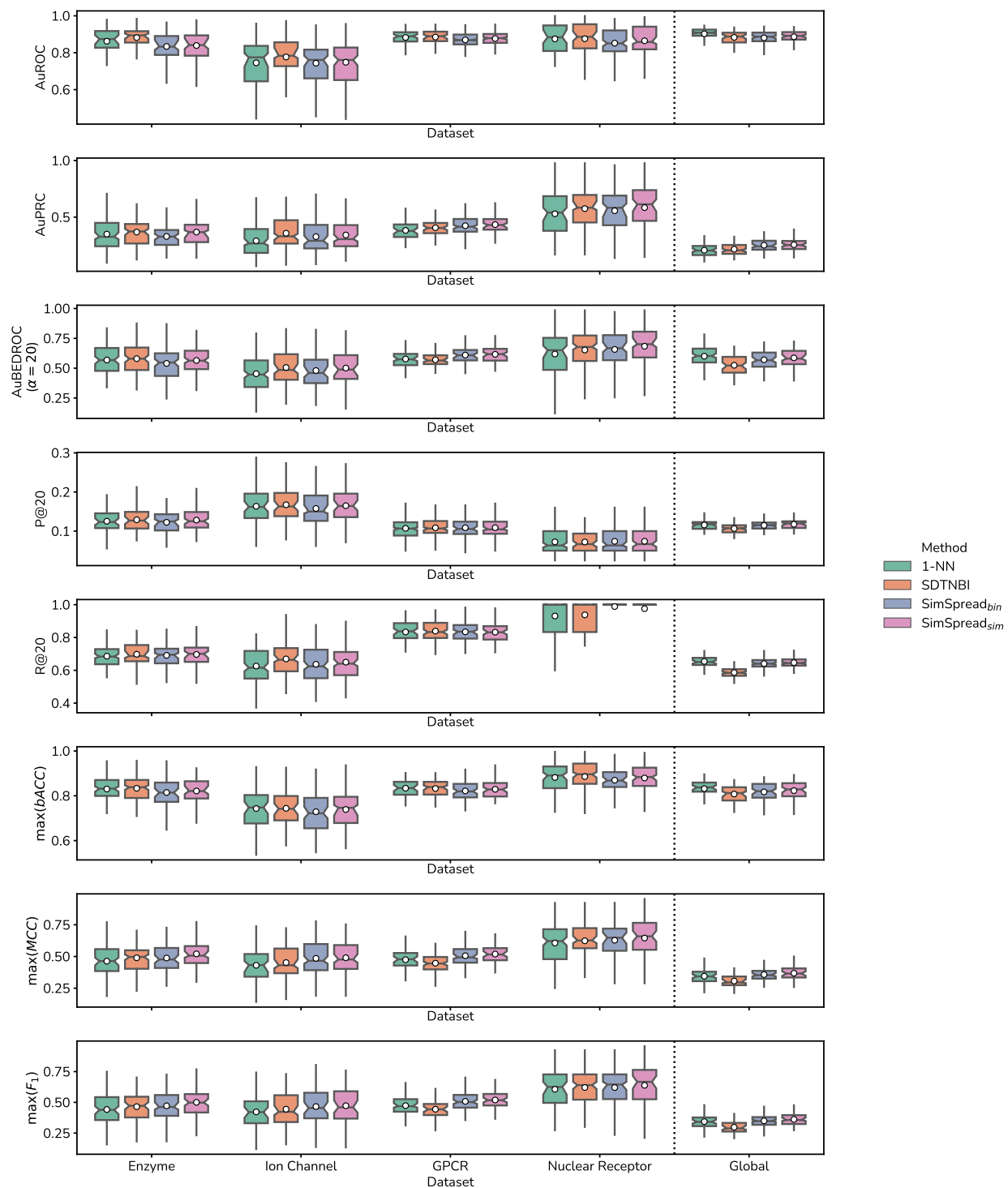


Figure S5: **Predictive performance comparison between the studied methods in 10-times 10-fold CV.** Boxplots for AuROC, AuPRC, AuBEDROC, P@20, R@20, max(MCC), max(F₁) and max(bACC) obtained in the 10-times 10-fold CV over the five validation datasets for 1-NN, NBI, SDTNBI, SimSpread_{bin} and SimSpread_{sim} using ECFP4 molecular descriptor and Tanimoto coefficient as similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the performance of the method for that metric. Boxplot notch represents the bootstrapped 95% CI around the median, whiskers correspond to 1.5x IQR and white dot represents the mean.

Table S7: **Overall average performance for the studied methods in 10-times 10-fold CV.** Mean \pm SD for AuROC and AuPRC observed in 10-times 10-fold CV over the five validation datasets using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the mean performance of the method for that metric.

Dataset	Method	AuROC	AuPRC
NR	NN	0.875 \pm 0.090	0.530 \pm 0.198
	SDTNBI	0.876 \pm 0.092	0.575 \pm 0.181
	SimSpread _{bin}	0.852 \pm 0.082	0.556 \pm 0.188
	SimSpread _{sim}	0.865 \pm 0.084	0.584 \pm 0.201
IC	NN	0.745 \pm 0.120	0.291 \pm 0.135
	SDTNBI	0.777 \pm 0.108	0.358 \pm 0.139
	SimSpread _{bin}	0.743 \pm 0.109	0.327 \pm 0.135
	SimSpread _{sim}	0.748 \pm 0.115	0.342 \pm 0.132
GPCR	NN	0.883 \pm 0.040	0.382 \pm 0.079
	SDTNBI	0.885 \pm 0.037	0.406 \pm 0.073
	SimSpread _{bin}	0.870 \pm 0.037	0.424 \pm 0.078
	SimSpread _{sim}	0.877 \pm 0.038	0.434 \pm 0.080
Enzyme	NN	0.862 \pm 0.066	0.350 \pm 0.126
	SDTNBI	0.882 \pm 0.051	0.368 \pm 0.112
	SimSpread _{bin}	0.834 \pm 0.072	0.329 \pm 0.103
	SimSpread _{sim}	0.839 \pm 0.073	0.368 \pm 0.121
Global	NN	0.903 \pm 0.031	0.208 \pm 0.054
	SDTNBI	0.883 \pm 0.034	0.216 \pm 0.050
	SimSpread _{bin}	0.881 \pm 0.036	0.252 \pm 0.048
	SimSpread _{sim}	0.886 \pm 0.031	0.256 \pm 0.049

Table S8: **Overall median performance for the studied methods in 10-times 10-fold CV.** Q1, median and Q3 for AuROC and AuPRC observed in 10-times 10-fold CV over the five validation datasets using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the mean performance of the method for that metric.

Dataset	Method	AuROC			AuPRC		
		Q1	median	Q3	Q1	median	Q3
NR	NN	0.809	0.884	0.948	0.379	0.540	0.685
	SDTNBI	0.824	0.886	0.954	0.453	0.583	0.696
	SimSpread _{bin}	0.808	0.851	0.921	0.428	0.583	0.690
	SimSpread _{sim}	0.818	0.858	0.941	0.467	0.613	0.739
IC	NN	0.645	0.775	0.837	0.184	0.268	0.395
	SDTNBI	0.727	0.783	0.857	0.266	0.330	0.473
	SimSpread _{bin}	0.661	0.761	0.817	0.223	0.290	0.431
	SimSpread _{sim}	0.652	0.761	0.828	0.243	0.305	0.430
GPCR	NN	0.861	0.889	0.911	0.324	0.378	0.435
	SDTNBI	0.863	0.889	0.914	0.358	0.404	0.449
	SimSpread _{bin}	0.845	0.868	0.898	0.371	0.415	0.484
	SimSpread _{sim}	0.853	0.879	0.902	0.389	0.430	0.483
Enzyme	NN	0.827	0.874	0.918	0.242	0.326	0.449
	SDTNBI	0.855	0.894	0.917	0.267	0.373	0.439
	SimSpread _{bin}	0.788	0.833	0.890	0.254	0.327	0.386
	SimSpread _{sim}	0.784	0.841	0.894	0.279	0.367	0.433
Global	NN	0.892	0.909	0.925	0.164	0.205	0.245
	SDTNBI	0.856	0.888	0.909	0.175	0.206	0.254
	SimSpread _{bin}	0.862	0.887	0.909	0.212	0.242	0.291
	SimSpread _{sim}	0.871	0.891	0.912	0.218	0.253	0.289

Table S9: **Early-recognition average performance for the studied methods in 10-times 10-fold CV.** Mean \pm SD for AuBEDROC, P@20 and R@20 observed in 10-times 10-fold CV over the five validation datasets using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the performance of the method for that metric.

Dataset	Method	AuBEDROC	P@20	R@20
NR	NN	0.619 \pm 0.191	0.072 \pm 0.031	0.931 \pm 0.104
	SDTNBI	0.653 \pm 0.166	0.072 \pm 0.029	0.938 \pm 0.097
	SimSpread _{bin}	0.657 \pm 0.166	0.074 \pm 0.030	0.989 \pm 0.044
	SimSpread _{sim}	0.684 \pm 0.177	0.074 \pm 0.030	0.976 \pm 0.054
IC	NN	0.453 \pm 0.147	0.164 \pm 0.048	0.626 \pm 0.103
	SDTNBI	0.505 \pm 0.144	0.167 \pm 0.043	0.669 \pm 0.096
	SimSpread _{bin}	0.480 \pm 0.144	0.158 \pm 0.047	0.637 \pm 0.105
	SimSpread _{sim}	0.501 \pm 0.146	0.165 \pm 0.047	0.650 \pm 0.103
GPCR	NN	0.577 \pm 0.069	0.107 \pm 0.025	0.834 \pm 0.068
	SDTNBI	0.570 \pm 0.061	0.109 \pm 0.025	0.839 \pm 0.061
	SimSpread _{bin}	0.610 \pm 0.070	0.109 \pm 0.026	0.835 \pm 0.064
	SimSpread _{sim}	0.616 \pm 0.070	0.109 \pm 0.026	0.833 \pm 0.061
Enzyme	NN	0.568 \pm 0.122	0.125 \pm 0.028	0.686 \pm 0.063
	SDTNBI	0.580 \pm 0.123	0.129 \pm 0.028	0.698 \pm 0.070
	SimSpread _{bin}	0.540 \pm 0.130	0.123 \pm 0.028	0.691 \pm 0.066
	SimSpread _{sim}	0.565 \pm 0.111	0.128 \pm 0.029	0.697 \pm 0.065
Global	NN	0.600 \pm 0.079	0.116 \pm 0.011	0.654 \pm 0.033
	SDTNBI	0.524 \pm 0.080	0.107 \pm 0.012	0.586 \pm 0.031
	SimSpread _{bin}	0.571 \pm 0.078	0.115 \pm 0.012	0.641 \pm 0.031
	SimSpread _{sim}	0.587 \pm 0.078	0.118 \pm 0.012	0.647 \pm 0.031

Table S10: **Early-recognition median performance for the studied methods in 10-times 10-fold CV.** Q1, median and Q3 for AuBEDROC, P@20 and R@20 observed in 10-times 10-fold CV over the five validation datasets using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the performance of the method for that metric.

Dataset	Method	AuBEDROC			P@20			R@20		
		Q1	median	Q3	Q1	median	Q3	Q1	median	Q3
NR	NN	0.486	0.651	0.754	0.050	0.063	0.100	0.833	1.000	1.000
	SDTNBI	0.561	0.678	0.774	0.050	0.067	0.094	0.833	1.000	1.000
	SimSpread _{bin}	0.567	0.668	0.777	0.050	0.063	0.100	1.000	1.000	1.000
	SimSpread _{sim}	0.590	0.701	0.806	0.050	0.067	0.100	1.000	1.000	1.000
IC	NN	0.342	0.447	0.565	0.133	0.162	0.196	0.550	0.616	0.718
	SDTNBI	0.403	0.485	0.617	0.138	0.163	0.198	0.594	0.664	0.735
	SimSpread _{bin}	0.375	0.460	0.571	0.126	0.150	0.191	0.552	0.624	0.726
	SimSpread _{sim}	0.410	0.489	0.609	0.136	0.164	0.196	0.570	0.641	0.712
GPCR	NN	0.525	0.574	0.620	0.089	0.107	0.123	0.796	0.838	0.888
	SDTNBI	0.534	0.564	0.608	0.095	0.107	0.126	0.797	0.833	0.890
	SimSpread _{bin}	0.565	0.609	0.653	0.093	0.106	0.124	0.793	0.835	0.876
	SimSpread _{sim}	0.563	0.618	0.662	0.093	0.104	0.124	0.788	0.832	0.869
Enzyme	NN	0.477	0.568	0.669	0.108	0.123	0.146	0.637	0.684	0.729
	SDTNBI	0.485	0.580	0.673	0.107	0.127	0.149	0.655	0.687	0.754
	SimSpread _{bin}	0.435	0.548	0.625	0.102	0.125	0.143	0.643	0.695	0.732
	SimSpread _{sim}	0.492	0.563	0.647	0.109	0.125	0.149	0.651	0.701	0.739
Global	NN	0.548	0.602	0.664	0.106	0.118	0.123	0.633	0.651	0.676
	SDTNBI	0.462	0.522	0.595	0.097	0.107	0.115	0.566	0.585	0.607
	SimSpread _{bin}	0.513	0.567	0.631	0.106	0.115	0.122	0.623	0.640	0.663
	SimSpread _{sim}	0.534	0.580	0.645	0.108	0.120	0.125	0.629	0.644	0.667

Table S11: **Binary prediction average performance for the studied methods in 10-times 10-fold CV.** Mean \pm SD for max(bACC), max(MCC) and max(F₁) observed in 10-times 10-fold CV over the five validation datasets using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the performance of the method for that metric.

Dataset	Method	max(bACC)	max(MCC)	max(F ₁)
NR	NN	0.882 \pm 0.070	0.606 \pm 0.150	0.606 \pm 0.151
	SDTNBI	0.886 \pm 0.068	0.624 \pm 0.135	0.621 \pm 0.138
	SimSpread _{bin}	0.870 \pm 0.055	0.628 \pm 0.151	0.620 \pm 0.157
	SimSpread _{sim}	0.879 \pm 0.054	0.645 \pm 0.157	0.639 \pm 0.167
IC	NN	0.743 \pm 0.085	0.431 \pm 0.134	0.422 \pm 0.137
	SDTNBI	0.744 \pm 0.080	0.452 \pm 0.128	0.444 \pm 0.134
	SimSpread _{bin}	0.729 \pm 0.091	0.485 \pm 0.134	0.466 \pm 0.141
	SimSpread _{sim}	0.739 \pm 0.085	0.490 \pm 0.127	0.472 \pm 0.143
GPCR	NN	0.834 \pm 0.037	0.474 \pm 0.070	0.473 \pm 0.071
	SDTNBI	0.833 \pm 0.035	0.447 \pm 0.064	0.444 \pm 0.062
	SimSpread _{bin}	0.822 \pm 0.043	0.506 \pm 0.071	0.508 \pm 0.071
	SimSpread _{sim}	0.829 \pm 0.036	0.519 \pm 0.069	0.519 \pm 0.070
Enzyme	NN	0.830 \pm 0.056	0.463 \pm 0.116	0.442 \pm 0.118
	SDTNBI	0.833 \pm 0.054	0.489 \pm 0.103	0.465 \pm 0.118
	SimSpread _{bin}	0.815 \pm 0.062	0.488 \pm 0.104	0.472 \pm 0.117
	SimSpread _{sim}	0.822 \pm 0.052	0.521 \pm 0.105	0.500 \pm 0.117
Global	NN	0.832 \pm 0.038	0.345 \pm 0.055	0.343 \pm 0.054
	SDTNBI	0.808 \pm 0.037	0.307 \pm 0.046	0.297 \pm 0.048
	SimSpread _{bin}	0.817 \pm 0.041	0.357 \pm 0.042	0.350 \pm 0.046
	SimSpread _{sim}	0.823 \pm 0.041	0.368 \pm 0.048	0.361 \pm 0.047

Table S12: **Binary prediction median performance for the studied methods in 10-times 10-fold CV.** Q1, median and Q3 for max(bACC), max(MCC) and max(F₁) observed in 10-times 10-fold CV over the five validation datasets using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the performance of the method for that metric.

Dataset	Method	max(bACC)			max(MCC)			max(F ₁)		
		Q1	median	Q3	Q1	median	Q3	Q1	median	Q3
NR	NN	0.834	0.891	0.931	0.479	0.621	0.715	0.496	0.625	0.727
	SDTNBI	0.853	0.896	0.944	0.565	0.626	0.724	0.522	0.640	0.727
	SimSpread _{bin}	0.839	0.868	0.906	0.545	0.648	0.720	0.526	0.645	0.727
	SimSpread _{sim}	0.844	0.884	0.925	0.553	0.659	0.765	0.525	0.667	0.764
IC	NN	0.677	0.748	0.803	0.341	0.433	0.520	0.330	0.420	0.509
	SDTNBI	0.690	0.742	0.799	0.368	0.431	0.563	0.339	0.423	0.558
	SimSpread _{bin}	0.655	0.722	0.791	0.392	0.466	0.598	0.370	0.454	0.578
	SimSpread _{sim}	0.679	0.747	0.796	0.402	0.476	0.590	0.368	0.459	0.590
GPCR	NN	0.805	0.834	0.862	0.429	0.478	0.528	0.424	0.468	0.526
	SDTNBI	0.805	0.839	0.862	0.398	0.445	0.496	0.396	0.443	0.487
	SimSpread _{bin}	0.793	0.820	0.853	0.451	0.493	0.559	0.456	0.502	0.561
	SimSpread _{sim}	0.797	0.828	0.857	0.470	0.516	0.566	0.475	0.516	0.568
Enzyme	NN	0.799	0.832	0.870	0.385	0.464	0.558	0.356	0.439	0.542
	SDTNBI	0.790	0.838	0.871	0.403	0.497	0.549	0.377	0.472	0.546
	SimSpread _{bin}	0.773	0.814	0.859	0.410	0.477	0.567	0.390	0.462	0.560
	SimSpread _{sim}	0.788	0.821	0.865	0.448	0.507	0.582	0.418	0.498	0.567
Global	NN	0.818	0.838	0.859	0.305	0.343	0.380	0.306	0.341	0.377
	SDTNBI	0.779	0.812	0.838	0.274	0.296	0.342	0.263	0.288	0.333
	SimSpread _{bin}	0.791	0.820	0.853	0.325	0.353	0.387	0.317	0.348	0.380
	SimSpread _{sim}	0.798	0.829	0.856	0.334	0.365	0.406	0.324	0.356	0.396

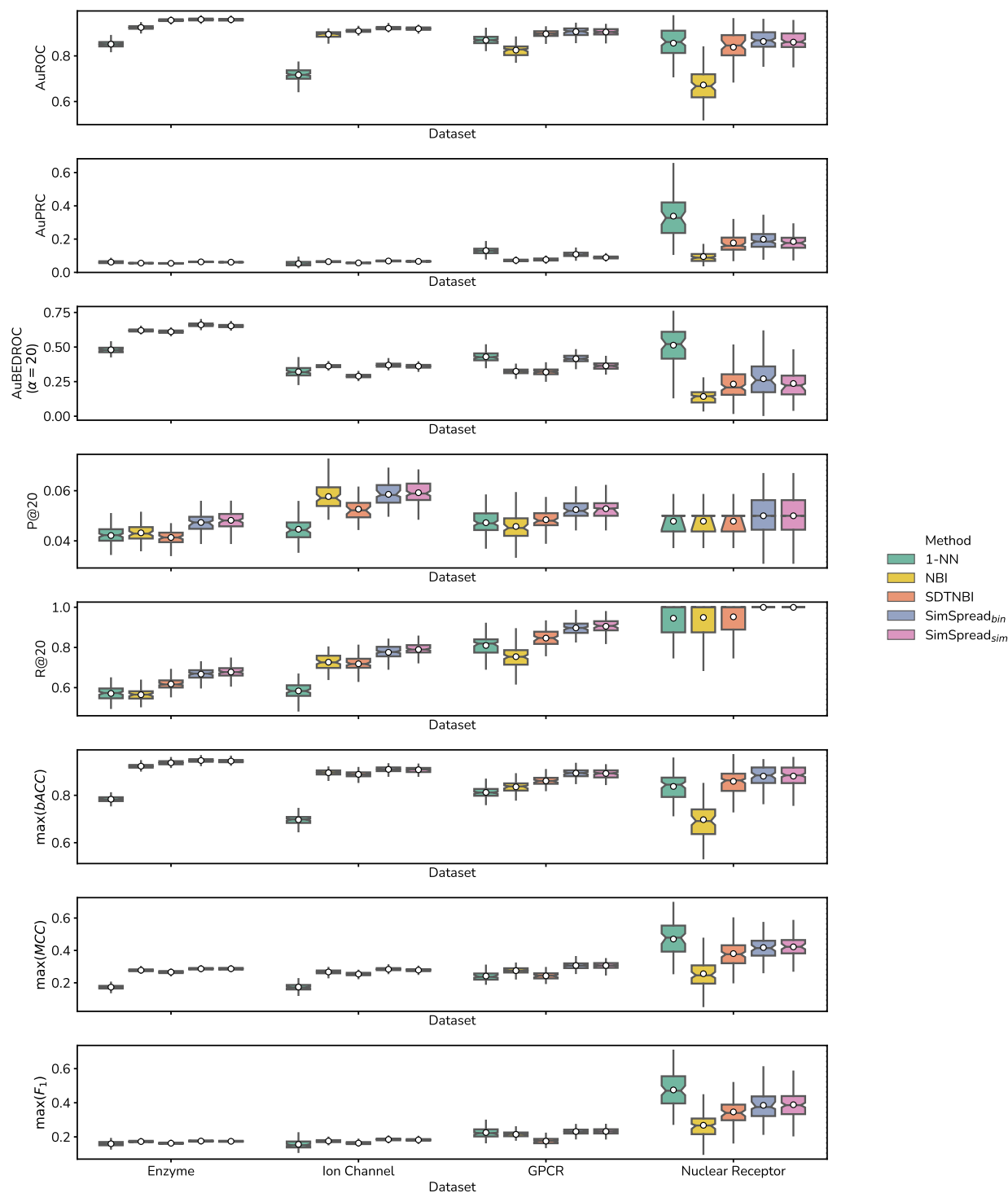


Figure S6: **Predictive performance comparison between the studied methods in 10-times 10-fold CV using DTI holdout.** Boxplots for AuROC, AuPRC, AuBEDROC, P@20, R@20, max(MCC), max(F₁) and max(bACC) obtained in the 10-times 10-fold CV using DTI holdout over the five validation datasets for 1-NN, NBI, SDTNBI, SimSpread_{bin} and SimSpread_{sim} using ECFP4 molecular descriptor and Tanimoto coefficient as similarity measure. For SimSpread variants, the similarity cutoff is equal to the optimal cutoff that maximizes the performance of the method for that metric. Boxplot notch represents the bootstrapped 95% CI around the median, whiskers correspond to 1.5x IQR and white dot represents the mean.

S4.3 Time-split cross-validation

Table S13: **Overall average performance for the studied methods in time-split CV.** Mean \pm SD for AuROC and AuPRC observed in time-split CV using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff α is equal to 0.2.

Method	AuROC	AuPRC
1-NN	0.807 \pm 0.261	0.303 \pm 0.352
SDTNBI	0.765 \pm 0.266	0.252 \pm 0.335
SimSpread _{bin}	0.774 \pm 0.264	0.279 \pm 0.349
SimSpread _{sim}	0.778 \pm 0.265	0.310 \pm 0.367

Table S14: **Overall median performance for the studied methods in time-split CV.** Q1, median and Q3 for AuROC and AuPRC observed in time-split CV using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff α is equal to 0.2.

Method	AuROC			AuPRC		
	Q1	median	Q3	Q1	median	Q3
1-NN	0.699	0.942	0.997	0.014	0.137	0.500
SDTNBI	0.623	0.873	0.990	0.009	0.067	0.417
SimSpread _{bin}	0.500	0.912	0.994	0.009	0.090	0.500
SimSpread _{sim}	0.500	0.932	0.995	0.01	0.111	0.539

Table S15: **Early-recognition average performance for the studied methods in time-split CV.** Mean \pm SD for AuBEDROC, P@20 and R@20 observed in time-split CV using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff α is equal to 0.2.

Method	AuBEDROC	P@20	R@20
1-NN	0.496 \pm 0.409	0.065 \pm 0.092	0.526 \pm 0.464
SDTNBI	0.405 \pm 0.399	0.055 \pm 0.079	0.442 \pm 0.457
SimSpread _{bin}	0.451 \pm 0.402	0.061 \pm 0.085	0.497 \pm 0.464
SimSpread _{sim}	0.476 \pm 0.410	0.063 \pm 0.086	0.517 \pm 0.463

Table S16: **Early-recognition median performance for the studied methods in time-split CV.** Q1, median and Q3 for AuBEDROC, P@20 and R@20 observed in time-split CV using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff α is equal to 0.2.

Method	AuBEDROC			P@20			R@20		
	Q1	median	Q3	Q1	median	Q3	Q1	median	Q3
1-NN	0.016	0.527	0.951	0.000	0.050	0.100	0.000	0.667	1.000
SDTNBI	0.002	0.305	0.840	0.000	0.050	0.050	0.000	0.333	1.000
SimSpread _{bin}	0.004	0.434	0.885	0.000	0.050	0.100	0.000	0.500	1.000
SimSpread _{sim}	0.006	0.477	0.909	0.000	0.050	0.100	0.000	0.500	1.000

Table S17: **Binary prediction average performance for the studied methods in time-split CV.** Mean \pm SD for max(bACC), max(MCC) and max(F₁) observed in time-split CV using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff α is equal to 0.2.

Method	max(bACC)	max(MCC)	max(F ₁)
1-NN	0.885 \pm 0.139	0.418 \pm 0.341	0.370 \pm 0.361
SDTNBI	0.861 \pm 0.145	0.356 \pm 0.329	0.309 \pm 0.343
SimSpread _{bin}	0.829 \pm 0.201	0.376 \pm 0.350	0.339 \pm 0.357
SimSpread _{sim}	0.831 \pm 0.201	0.405 \pm 0.365	0.372 \pm 0.373

Table S18: **Binary prediction median performance for the studied methods in time-split CV.** Q1, median and Q3 for max(bACC), max(MCC) and max(F₁) observed in time-split CV using the ECFP4 molecular descriptor and Tanimoto coefficient as chemical similarity measure. For SimSpread variants, the similarity cutoff α is equal to 0.2.

Method	max(bACC)			max(MCC)			max(F ₁)		
	Q1	median	Q3	Q1	median	Q3	Q1	median	Q3
1-NN	0.811	0.949	0.997	0.099	0.347	0.706	0.030	0.250	0.667
SDTNBI	0.755	0.905	0.994	0.073	0.242	0.576	0.020	0.143	0.500
SimSpread _{bin}	0.643	0.942	0.996	0.048	0.273	0.705	0.017	0.190	0.667
SimSpread _{sim}	0.646	0.947	0.997	0.049	0.313	0.706	0.019	0.222	0.667

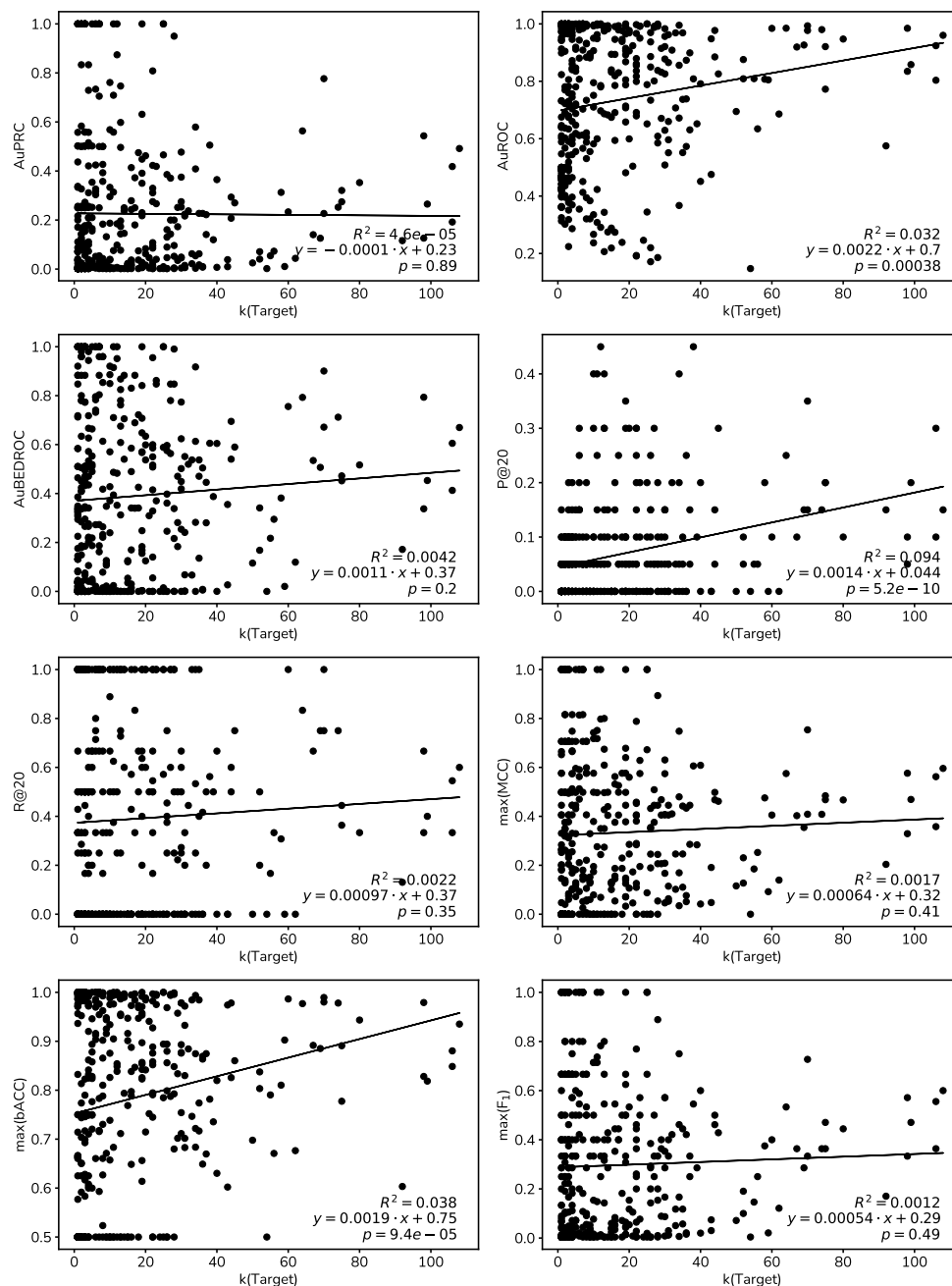


Figure S7: **Relationship between target degree and predictive performance.** Per target performance for SimSpread_{sim} over the ChEMBL CC&D time-split validation using the ECFP4 molecular descriptor. R^2 and p corresponds to the coefficient of determination and p-value for a hypothesis test that the slope is nonzero (Wald Test with t-distribution) obtained from the linear least-squares regression between target degree and predictive performance over the 8 evaluation metrics used in the study.

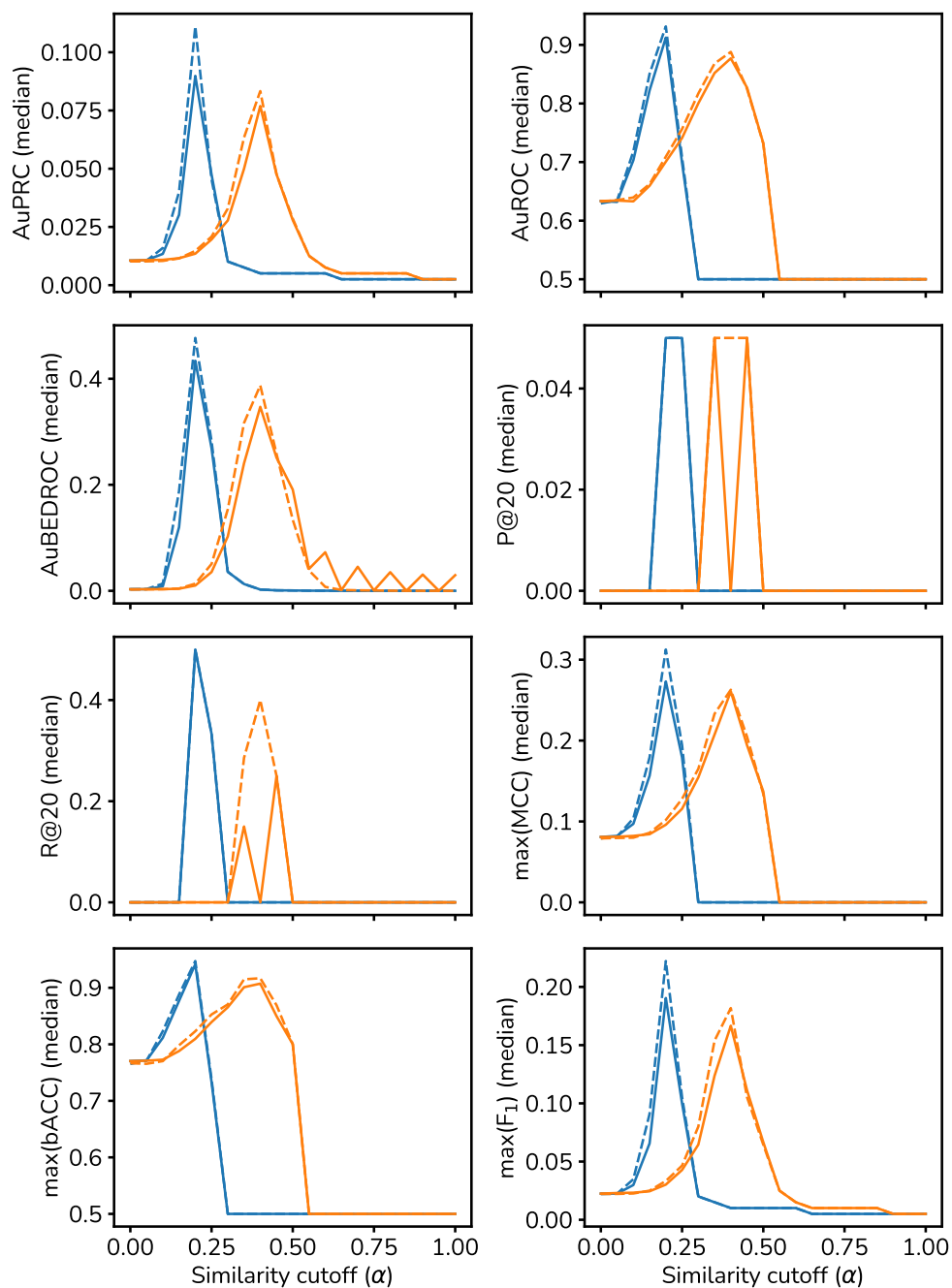


Figure S8: **Relationship between similarity measure and predictive performance for time-split validation.** SimSpread performance obtained for the ChEMBL CC&D time-split validation using the ECFP4 molecular descriptor using 2 similarity measures: Tanimoto coefficient (in blue) and Tversky index with $\alpha = 0.0$ and $\beta = 1.0$ (in orange). Solid line corresponds to SimSpread_{bin} and dashed lines correspond to SimSpread_{sim}.

S5 Computation time

Table S19: **SimSpread_{sim} computation time for 10-times 10-fold cross-validation**

Dataset	Mean time per 10-fold CV (seconds)	Mean time per fold (seconds)
Nuclear Receptor	7.01	0.701
Ion Channel	9.727	0.973
GPCR	8.671	0.867
Enzyme	21.838	2.184
Global	140.137	14.137

Note: Computation times where detemined using the `@elapsed` macro in Julia Language version 1.7.3. Benchmark was run on the following hardware using GPU acceleration through `CUDA.jl` package: Intel(R) Xeon(R) CPU E5-2697 v4 @ 2.30GHz, 536 GB RAM & NVIDIA GeForce GTX 1080-Ti 12Gb VRAM.

S6 Reimplementation of benchmark methods

Table S20: Reimplementation of SDTNBI using the Global dataset with Klethra-Roth descriptor in LOO cross-validation

Implementation	AuROC	P@20	R@20
In-house	0.911 ± 0.150	0.097 ± 0.165	0.500 ± 0.351
Wu, et al (2016)	0.910 ± 0.150	0.099 ± 0.167	0.493 ± 0.438

Note: Values correspond to mean \pm SD over the predictive performance of each fold.

Table S21: Reimplementation of 1-NN using the ChEMBL24 \rightarrow 28 dataset with ECFP4 descriptor under time-split cross-validation

Implementation	AuROC	AuPRC
Julia (network-based)	0.81056	0.13229
C++	0.81056	0.13230

Note: Julia (network-based): Our network-based k-NN algorithm implemented in-house in Julia, which was employed throughout this study. C++: Reference k-NN algorithm implemented in-house in C++ based on the works of Gfeller et al. (2013) [3].

S7 Examples

Table S22: Target prediction for Nintedanib and Pirfenidone over ChEMBL28 CC&D network

Compound	Score	Rank	Target ChEMBL ID	Target Name
Nintedanib	0.006781	1	CHEMBL279	Vascular endothelial growth factor receptor 2
	0.005965	2	CHEMBL3650	Fibroblast growth factor receptor 1
	0.005891	3	CHEMBL1955	Vascular endothelial growth factor receptor 3
	0.005730	4	CHEMBL2007	Platelet-derived growth factor receptor alpha
	0.005473	5	CHEMBL1913	Platelet-derived growth factor receptor beta
	0.005406	6	CHEMBL2742	Fibroblast growth factor receptor 3
	0.005404	7	CHEMBL3973	Fibroblast growth factor receptor 4
	0.005393	8	CHEMBL4142	Fibroblast growth factor receptor 2
	0.005345	9	CHEMBL1868	Vascular endothelial growth factor receptor 1
	0.003402	10	CHEMBL240	Potassium voltage-gated channel subfamily H member 2
	0.003377	11	CHEMBL238	Sodium-dependent dopamine transporter
	0.002833	12	CHEMBL1974	Receptor-type tyrosine-protein kinase FLT3
	0.002408	13	CHEMBL258	Tyrosine-protein kinase Lck
	0.002180	14	CHEMBL267	Proto-oncogene tyrosine-protein kinase Src
	0.002103	15	CHEMBL3905	Tyrosine-protein kinase Lyn
Pirfenidone	0.027070	1	CHEMBL4523354	Nicotinate phosphoribosyltransferase
	0.007884	2	CHEMBL4523441	Deoxyhypusine hydroxylase
	0.006112	3	CHEMBL4596	C-C chemokine receptor type 8
	0.004195	4	CHEMBL214	5-hydroxytryptamine receptor 1A
	0.003968	5	CHEMBL238	Sodium-dependent dopamine transporter
	0.003903	6	CHEMBL228	Sodium-dependent serotonin transporter
	0.003880	7	CHEMBL4685	Indoleamine 2,3-dioxygenase 1
	0.003740	8	CHEMBL3242	Carbonic anhydrase 12
	0.003641	9	CHEMBL3155	5-hydroxytryptamine receptor 7
	0.003485	10	CHEMBL2009	Glutamate receptor 1
	0.003128	11	CHEMBL205	Carbonic anhydrase 2
	0.002962	12	CHEMBL222	Sodium-dependent noradrenaline transporter
	0.002913	13	CHEMBL220	Acetylcholinesterase
	0.002761	14	CHEMBL4295747	Serine hydroxymethyltransferase, mitochondrial
	0.002761	14	CHEMBL4523129	Synaptojanin-2
	0.002761	14	CHEMBL4523136	Synaptojanin-1

Note: Prediction for Nintedanib and Pirfenidone where obtained using SimSpread_{sim} with the ECFP4 molecular descriptor and a similarity cutoff α equal to 0.2. Query network corresponds to ChEMBL28 CC&D dataset from which the time-split datasets were constructed.

References

1. Mold2 — FDA. Available online: <https://www.fda.gov/science-research/bioinformatics-tools/mold2> (accessed on 27 July 2022)
2. Hong, et al (2008). Mold2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *Journal of Chemical Information and Modeling*, 48(7), 1337–1344. <https://doi.org/10.1021/ci800038f>
3. Gfeller et al (2013). Shaping the interaction landscape of bioactive molecules. *Bioinformatics*, 29(23), 3073-3079. <https://doi.org/10.1093/bioinformatics/btt540>