MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **76** (2016) 359-376

ISSN 0340 - 6253

Application of the Weighted Power-Weakness Ratio (wPWR) as a Fusion Rule in Ligand–Based Virtual Screening

Matteo Cassotti[1,*], Francesca Grisoni[1], Serena Nembri[1], Roberto Todeschini[1]

[1] Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Dept. of Earth and Environmental Sciences, P.za della Scienza 1, 20126, Milano, Italy.

**Corresponding author: Matteo Cassotti, E-mail: cassotti.matteo@gmail.com; Tel.: +39 0264482801*

(Received February 1, 2016)

Abstract

Similarity searching, a technique used for ligand-based virtual screening in drug discovery, exploits the structural information of known active compounds to find new ones with the desired biological activity. The effectiveness of similarity searching can be enhanced by combining independent searches, executed with different reference structures, through the so-called fusion rules. In this pilot study, we applied for the first time the weighted Power-Weakness Ratio (wPWR) as a fusion rule. wPWR was recently proposed as a multi-criteria decision-making and ranking comparison technique that, unlike existing parameter-free fusion rules, (1) is based on pairwise comparisons between compounds, and (2) can use weighting schemes. We compared wPWR with six fusion rules on four datasets using six evaluation metrics. The results indicated that wPWR has a mediocre performance but also the most robust behavior. Moreover, in one case, wPWR had the best accumulation curve on the top 25 compounds. These aspects are important for prospective applications. Weighting the reference structures according to their activity did not show a net beneficial effect on one analyzed dataset and, thus, further investigations are needed.

1 Introduction

The underlying principle of many chemoinformatics techniques asserts that similar chemical structures are likely to exhibit similar physical-chemical properties and biological activities. Built on this basis, ligand-based virtual screening embraces a collection of approaches that aim at finding compounds with a desired biological activity on a specified biological target by exploiting the structural information of known actives [1]. Virtual

screening aims at directing the experimental biological testing on a smaller set of compounds, which have a higher probability of showing the desired activity.

Similarity searching, in particular, directly uses the structural similarity between known actives and a database of compounds and it is probably the most widely used approach to virtual screening [2][3]. Typically, an active compound against the specified target is chosen as a reference structure. The chemical similarity is then evaluated by a chosen measure of similarity [4] (e.g. Jaccard-Tanimoto coefficient) on a set of molecular descriptors, substructures or binary fingerprints [5], which encode structural features. The compounds in the database are then ranked by decreasing similarity, with the top-ranked molecules assumed to having the highest probability of being active.

The effectiveness of similarity searching can be improved by the so-called data fusion (or *consensus* scoring) approaches, in which the output of different independent searches is combined into a single final score, which is used to rank the structures in the database. Data fusion is implemented in two different fashions: (a) in *similarity fusion*, a single reference structure is used and the independent searches are based on different sets of molecular descriptors/fingerprints or similarity measures; (b) in *group fusion*, the set of descriptors/fingerprints and similarity measures is fixed and the independent searches use different reference structures. The individual similarity values (or the induced ranks) are combined together by means of *fusion rules* [6]. Generally, no fusion rule behaves the best in all cases – Chen *et al.* [7] indicate the rRKP rule (see Section 2.2 for details) as providing always the best results; our results confirm rRKP as being *on average* the best rule, but not *always* the best. Moreover, the different searches being combined (whether similarity or group fusion) are given equal importance. On the other hand, a weighting scheme can enable additional criteria to be taken into account, e.g. weighting the reference structures on their potency, ease of synthesis or economical cost. Hence, the focus on new fusion rules for virtual screening with optimal and stable behaviour and embedding a weighting scheme is of relevance.

In this pilot study, we applied the weighted Power-Weakness Ratio (wPWR) as a group fusion rule. The main goal was to have preliminary indications on whether wPWR has a potential for application in virtual screening. wPWR was recently proposed as a weighted multivariate index for multi-criteria decision-making and ranking comparison. Its salient features are: (a) being based on pairwise comparisons between compounds, and (b) the possibility to weigh the considered criteria. We re-adapted wPWR to virtual screening problems and compared the results with most of the parameter-free rules investigated by Chen

et al. [7]. The comparison was carried out using four publicly available datasets and the performance was quantified by means of several benchmarking virtual screening metrics, i.e. RIE, BEDROC, ROC, enrichment factor and AUAC (see Appendix A) [8], and relative scaffold diversity.

2 Theory

The following notation is used: N is the total number of compounds in the database and n is the number of actives. R_a and R_i are the ratio of actives (n/N) and inactives ((N-n)/N), respectively; s_{ik} is the similarity score of the i-th compound in the database with respect to the k-th reference structure and r_{ik} is the corresponding rank derived from the similarity scores (s_i); p is the total number of reference structures used and it is a subset of the actives ($p \leq n$).

2.1 Weighted Power-Weakness Ratio (wPWR)

An extensive description of the original Power-Weakness Ratio (PWR) and its weighted variant (wPWR) can be found in Todeschini *et al.* [9]. Here we briefly introduce the philosophy of wPWR and describe its adaptation to virtual screening. The idea of PWR is to compare a set of objects on a number of different criteria in a pairwise manner. The pairwise comparisons allow compiling a square non-symmetric matrix (**T**), where each element (t_{ij}) indicates the number of criteria for which the i-th object wins (i.e. is better) over the j-th object. In case of a draw, 0.5 is given to both objects. This matrix encodes the "power" of each object, i.e. its ability to win over the others. The elements of the transposed **T** matrix ($\mathbf{T}^{\mathbf{T}}$), indicate the number of criteria for which the i-th object was defeated (i.e. it was worse) by the j-th object. Hence, the $\mathbf{T}^{\mathbf{T}}$ matrix indicates the "weakness" of each object. In the weighted variant of PWR (wPWR), the element t_{ij} of the **T** matrix is the sum of the weights (instead of the count) of the criteria where the i-th object "wins" over the j-th object. Each t_{ij} element is obtained according to Equation 1:

$$t_{ij}^{\mathrm{W}} = \sum_{k=1}^{p} w_k \cdot \delta_{ij,k} \quad \text{where} \quad \delta_{ij,k} = \begin{cases} 1 & \text{if } x_{ik} \rhd x_{jk} \\ 0.5 & \text{if } x_{ik} \triangleq x_{jk} \\ 0 & \text{if } x_{ik} \lhd x_{jk} \end{cases} \quad \text{and} \quad \sum_{k=1}^{p} w_k = 1 \qquad (1)$$

where w_k are the normalized weights pre-assigned to each of the p criteria.

The Perron-Frobenius eigenvector ($\mathbf{e_{PF}}$) associated to the largest eigenvalue of the **T** matrix has large values for objects that win many times. On the other hand, the eigenvector

($\mathbf{e}^*_{\mathbf{PF}}$) associated to the largest eigenvalue of the $\mathbf{T}^{\mathbf{T}}$ matrix has low values for objects that are defeated few times. The wPWR score of the *i*-th object is defined as:

$$\text{wPWR}_i = \frac{\alpha + e_i}{\alpha + e_i^*} \quad \text{where} \quad \alpha = \frac{1}{n + n\log_2 n} \tag{2}$$

where e_i is the *i*-th element of the $\mathbf{e}_{\mathbf{PF}}$ eigenvector obtained from **T** and e_i^* is the *i*-th element of the $\mathbf{e}^*_{\mathbf{PF}}$ eigenvector obtained from $\mathbf{T}^{\mathbf{T}}$. The parameter α was introduced by Todeschini and co-authors to avoid singularities or spikes and to obtain reasonable scales. Objects that win many times and are defeated few times have large (w)wPWR scores.

In the case of virtual screening, the objects being compared are the compounds in the database and the criteria are the *p* similarity vectors (or the corresponding rankings) with respect to the *p* reference structures. When an active compound is taken as a reference structure, it is deleted from the list and then it is re-introduced when another active is chosen as a reference structure. This means that each reference compound has a missing value in the similarity vector when it was the reference structure (i.e. the self-similarity of the reference compound is neglected). When we generate the **T** matrix of pairwise comparisons, the comparison with a missing value is ignored. Consequently the number of valid comparisons is: a) *p* for two compounds never selected as a reference structure; b) *p-1* for comparisons where one compound was selected as a reference and the other one not; c) *p-2* for comparisons between two compounds both selected as a reference. The weights are normalized on the number of valid comparisons.

2.2 Benchmark fusion rules

Six parameter-free combination rules were used for comparison. Five of these were applied to both the similarity scores and the imputed ranks in accordance with Chen *et al.* [7]. The ranks were derived in increasing order so that the highest similarity score corresponds to rank one and accounting for ties. After the application of the fusion rules to the ranks, the scale is inverted (decreasing order) so that the most promising compounds have high scores in agreement with the scores obtained from the application of the combination rules to the similarity vectors. The prefix *s* or *r* is used to distinguish the rules applied to the similarity scores or to the ranks, respectively.

The MIN rule takes the minimum similarity and the lowest rank, accordingly; on the contrary, the MAX rule takes the maximum similarity and the top rank. The SUM and MED rules take the average and the median similarity score or rank, respectively. The EUC rule

computes the Euclidean distance of the similarity scores or ranks. The last rule, RKP, is the summation of the reciprocal ranks. As suggested in Chen *et al.*, this rule is not applied to the similarity scores to avoid the risk of a null denominator. The formulas are collated in Table 1.

Table 1. Parameter-free combination rules compared in this study.

Rule	Formula	Rule	Formula
sMIN	$Score_i = \min\{s_{ik}\}$	sEUC	$Score_i = \sqrt{\sum_{k=1}^{p} s_{ik}^2}$
rMIN	$Score_i = \max\{r_{ik}\}$	rEUC	$Score_i = \sqrt{\sum_{k=1}^{p} r_{ik}^2}$
sMAX	$Score_i = \max\{s_{ik}\}$	sMED	$Score_i = \text{median}\{s_{ik}\}$
rMAX	$Score_i = \min\{r_{ik}\}$	rMED	$Score_i = \text{median}\{r_{ik}\}$
sSUM	$Score_i = \frac{1}{p}\sum_{k=1}^{p} s_{ik}$	rRKP	$Score_i = \sum_{k=1}^{p} \frac{1}{r_{ik}}$
rSUM	$Score_i = \frac{1}{p}\sum_{k=1}^{p} r_{ik}$	wPWR	$Score_i = \frac{\alpha + e_i}{\alpha + e_i^*}$

3 Material and methods

3.1 Data sets

Four publicly available datasets with pharmaceutical relevance were chosen as study cases.

The Beta-glucocerebrosidase (GC) dataset, retrieved from ChEMBL (ID: CHEMBL1613818) [10], comprised 11,377 records with potency expressed as concentration that produces 50% activation/inhibition (AC_{50}). Deficiency of GC triggers Gaucher's disease, a genetic disorder that affects several organs with diverse symptoms. Inhibitors of GC are able to restore the function of mutant proteins by acting as chaperones [11].

The Pyruvate kinase isoenzyme M2 (PKM2) dataset, obtained from ChEMBL (ID: CHEMBL1613996), contained AC_{50} values for 6,331 records. Pyruvate kinase catalyzes the transfer of phosphoryl groups in the glycolytic pathway. The embryonic isoform PKM2 is a selective target for cancer therapy because it is re-expressed in cancer tissues, where it promotes aerobic glycolysis [12].

The Aldose reductase (ALDR) dataset, consisting in 9,159 records flagged as active/inactive, was retrieved from the DUD-E website [13]. Differences between structures are considered at the level of protonation state. Aldose reductase is involved in the reduction of glucose to sorbitol. Inhibitors of ALDR are potential agents in diabetic therapies because accumulation of sorbitol is hypothesized to cause diabetes [14].

The GSK TCAMS dataset derives from a screening of 1,986,056 compounds performed by GlaxoSmithKline (GSK) against the proliferation of *Plasmodium falciparium* strain 3D7 in human erythrocytes [15]. The dataset contains concentrations producing 50% inhibition (XC_{50}) for the 13,533 compounds that showed at least 80% inhibition of parasite growth at a concentration of 2 μM. *P. falciparium* is one of the *Plasmodium* species that cause malaria, an infectious disease that affects both humans and animals.

3.2 Data curation

A common curation procedure was applied to the datasets. Potency values reported as "greater/smaller than" (> or <) were removed (applicable only to the GC, PKM2 and GSK datasets). For replicates (multiple experimental measurements for the same chemical), the mean potency was calculated for the GC, PKM2 and GSK datasets, whereas for ALDR the agreement between class assignations was examined (no disagreement detected). Disconnected structures and structures considered not valid by Dragon [16] were removed. Only the GSK TCAMS dataset was considerably reduced because of the large number of disconnected structures (5234). Ninety-six invalid structures were detected, all in the ALDR dataset. These compounds had pentatomic N-containing rings with incorrectly defined aromatic bonds. All of them were inactive.

Characteristics of the datasets with number of records before and after data curation are reported in Table 2.

Table 2. Characteristics of the analyzed datasets.

Dataset	Response	Initial no. records[a]	Final no. records[b]	Class distribution	Cutoffs [μM]	Source
PKM2	AC_{50}	6331	6111	5965/146[c]	0.5	[10]
GC	AC_{50}	11377	10477	10260/183/34[d]	0.5/0.1	[10]
GSK TCAMS	XC_{50}	13533	8000	7750/250[c]	0.1	[15]
ALDR	Class	9159	9062	8903/159[c]	1.0	[13]

[a] number of records retrieved from database; [b] number of records after data curation; [c] inactive/active; [d] inactive/moderate/potent.

3.3 Data preparation

The potency values of the PKM2, GC and GSK datasets were converted into activity classes using the cutoff values of Table 2. Three classes (inactive, moderate and potent) were defined for the GC dataset in order to investigate the weighting ability of wPWR.

Structural information was encoded by extended connectivity binary fingerprints (ECFP) [17][18], a benchmark in virtual screening. ECFP of 1024 bits were generated setting the *radius* to four. The other parameters used to discriminate fragments were: atom type, connectivity (total), attached hydrogens, charge, aromaticity and bond order.

3.4 Selection of reference structures

For each dataset, 10% of the active compounds were selected as the reference structures. Two sets of reference structures were generated using the *k*-Means method, an iterative distance-based clustering technique that partitions *n* objects into a predefined number of clusters [19]. First, the *n* actives were partitioned into *k* clusters, with *k* equal to the number of reference structures ($k = n*0.1$), using the Hamming distance [20]. Then, the two sets of reference structures were taken: 1) *Set1* collects one reference structure from each cluster in order to maximize their diversity; 2) *Set2* gathers *k* reference structures from the biggest cluster to pursue high similarity among them, as opposed to *Set1*.

3.5 Weighting schemes

Three classes of activity were defined for the GC dataset, namely inactive, moderate and potent. The activity class was used to give different weights to the reference structures in order to understand the effect of the weighting scheme in wPWR. The hypothesis was that the weighting scheme could be used to place on the top of the final ranked list compounds more similar to the reference structures with higher weights. To test this, the weights unbalance between the moderate and potent class was increased as follows: 1:1 (unweighted), 1:2, 1:4 and 1:8, respectively.

3.6 Evaluation of virtual screening

The ability of the fusion rules to place active compounds in the top positions of the final ranked list was compared by means of five metrics: (a) Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC) [8], (b) Enrichment Factor (EF) [21], (c) Area Under the Accumulation Curve (AUAC) [8], (d) area under the Receiver Operating Characteristic (ROC) [22] and (e) Robust Initial Enhancement (RIE) [23]. The interested reader can find a thorough discussion in the work of Truchon and Bayly [8]. The discrete formulas are presented in Appendix A. The enrichment factor was calculated on the top 5% of the final ranked list; BEDROC and RIE were calculated with $\alpha = Log_{10}(N)$.

It is often desired that the top ranked actives are also as different as possible, because this gives medicinal chemists additional possibilities. The ability to retrieve diverse active

structures is called scaffold hopping ability. We investigated this property for the analyzed fusion rules using the relative Murcko scaffold diversity (%), calculated on the actives in the top 100 positions of the final ranked list [24].

3.7 Software

Dataset curation was carried out in KNIME v. 2.9.0 [25] by means of workflows designed by the authors. Dragon 7 [16] was used to check the validity of the SMILES strings and generate the extended connectivity fingerprints; the node RDKit Find Murcko Scaffolds v. 2010/06/12 of KNIME was used to generate the Murcko scaffolds. *K*-Means clustering was carried out with the statistics toolbox of MATLAB v. R2012a [26]. All the calculations (application of fusion rules and evaluation of the virtual screening) were carried out in MATLAB v. R2012a [26] by means of a toolbox written by the authors.

Table 3. Results on the ALDR and GSK TCAMS datasets; no weighting scheme applied to wPWR. The best and worst result within each metric is highlighted in bold face.

Rule	ALDR - *Run1*					ALDR - *Run2*				
	AUAC	ROC	$EF_{5\%}$	RIE	BEDROC	AUAC	ROC	$EF_{5\%}$	RIE	BEDROC
rRKP	**0.93**	**0.94**	**14.72**	**3.38**	**0.86**	**0.87**	**0.88**	12.83	**3.03**	**0.77**
rMIN	0.74	0.74	5.41	2.09	0.53	**0.84**	**0.84**	**11.57**	**2.79**	**0.71**
sMIN	**0.73**	**0.73**	**4.03**	**1.97**	**0.49**	0.85	0.86	11.95	2.87	0.73
rMAX	**0.93**	0.93	**14.72**	3.34	0.85	**0.87**	0.87	12.08	2.94	0.75
sMAX	0.92	0.92	14.09	3.27	0.84	0.86	0.87	12.08	2.91	0.74
rSUM	0.84	0.85	10.06	2.76	0.70	0.86	0.86	12.83	2.96	0.76
sSUM	0.88	0.88	12.45	3.00	0.77	0.86	0.87	**13.21**	3.02	**0.77**
rEUC	0.82	0.83	8.30	2.60	0.66	0.85	0.86	12.45	2.92	0.74
sEUC	0.88	0.89	12.83	3.06	0.78	**0.87**	0.87	**13.21**	**3.03**	**0.77**
rMED	0.84	0.84	10.94	2.79	0.71	0.86	0.86	13.08	2.96	0.75
sMED	0.84	0.85	11.19	2.79	0.71	0.86	0.87	13.08	2.98	0.76
wPWR	0.84	0.84	9.81	2.74	0.70	0.86	0.86	12.83	2.96	0.75
Rule	GSK TCAMS - *Run1*					GSK TCAMS - *Run2*				
	AUAC	ROC	$EF_{5\%}$	RIE	BEDROC	AUAC	ROC	$EF_{5\%}$	RIE	BEDROC
rRKP	**0.68**	**0.69**	**6.08**	**1.95**	**0.51**	0.48	0.48	2.00	1.12	0.28
rMIN	0.44	0.44	0.88	0.85	0.21	0.42	0.42	**0.64**	0.80	0.19
sMIN	0.49	0.49	1.04	1.00	0.25	**0.50**	**0.50**	1.36	1.11	0.28
rMAX	**0.68**	**0.69**	5.44	1.94	**0.51**	**0.50**	**0.50**	**2.40**	**1.16**	**0.29**
sMAX	0.64	0.65	4.16	1.71	0.44	0.46	0.46	1.68	0.98	0.24
rSUM	0.45	0.44	0.56	0.87	0.21	0.40	0.40	0.96	0.77	0.19
sSUM	0.54	0.54	1.76	1.22	0.31	0.40	0.40	1.28	0.78	0.19
rEUC	**0.43**	**0.43**	0.56	0.85	0.21	0.40	0.40	0.88	0.77	0.19
sEUC	0.58	0.58	2.32	1.44	0.37	0.40	0.40	1.20	0.77	0.19
rMED	0.44	**0.43**	0.80	0.86	0.21	**0.39**	**0.38**	**0.64**	0.74	**0.18**
sMED	0.44	0.44	**0.40**	**0.83**	**0.20**	**0.39**	**0.38**	0.88	**0.73**	**0.18**
wPWR	0.45	0.45	0.56	0.87	0.21	0.40	0.40	0.88	0.77	0.19

4 Results and discussion

Two runs were executed on each dataset with two different sets of reference structures as previously described. *Run1* and *Run2* are the results obtained with the reference structures in *Set1* and *Set2*, respectively.

Results are discussed by comparing the methods on calculations without weighting scheme for wPWR (unweighted cases). The effect of the weighting scheme is analyzed in section 4.2.

4.1 Unweighted cases

In the unweighted cases, the moderate and potent classes of the GC dataset were merged and considered as a single active class. The results on the four analyzed datasets are collated in Tables 3 and 4.

The results highlighted that the pairs of metrics AUAC - ROC and RIE - BEDROC are highly correlated with each other ($\rho = 1$). Hence, we arbitrarily chose to consider only ROC, EF and BEDROC from now on. Tables 3 and 4 suggest that virtual screening is more effective on the ALDR dataset as indicated by the higher values of the metrics. By averaging the results of each metric for each run, it appears that *Run1* gave better results in the GSK and PKM2 datasets. The situation is reversed for the ALDR and GC datasets.

A Principal Component Analysis (PCA) [27] was carried out to analyze the results in a holistic way. PCA was calculated on the matrices collecting all the results evaluated by the ROC, EF and BEDROC separately (Figures 1a, 1b, 1c). In order to facilitate the interpretation, two theoretical rules were added: the Best rule always takes the best result among the ones provided by the analyzed combination rules; the Worst rule, on the other hand, always takes the worst result. This approach was already adopted to compare methods because it allows to stretch the results along the Best-Worst direction, thus enabling easy spotting of the methods with better/worse performance [28]. Moreover, the deviation from the Best-Worst direction gives an indication of the sensitivity of the methods to the particular dataset. In other words, methods close to the line connecting Best and Worst are robust. On the contrary, methods far away from this line are more sensitive to the dataset.

Additionally, a PCA was conducted on the matrix were each combination rule was described by its average ROC, EF and BEDROC calculated across all computational runs of all datasets (Figure 1d).

Table 4. Results on the PKM2 and GC datasets; no weighting scheme applied to wPWR. The best and worst result within each metric is highlighted in bold face.

Rule	PKM2 - *Run1*					PKM2 - *Run2*				
	AUAC	ROC	EF_5 %	RIE	BEDROC	AUAC	ROC	EF_5 %	RIE	BEDROC
rRKP	**0.54**	**0.54**	1.37	1.16	**0.30**	0.49	0.49	0.96	0.98	0.25
rMIN	**0.50**	**0.50**	0.82	1.01	0.25	**0.48**	**0.48**	0.96	0.97	0.24
sMIN	**0.50**	**0.50**	**0.41**	**0.97**	**0.24**	**0.51**	**0.51**	1.10	**1.07**	**0.27**
rMAX	0.53	0.53	**1.78**	**1.17**	**0.30**	0.49	0.49	**0.82**	0.99	0.25
sMAX	**0.54**	**0.54**	1.10	1.15	0.29	**0.48**	**0.48**	0.96	**0.92**	**0.23**
rSUM	0.51	0.51	0.96	1.01	0.25	0.49	0.49	1.10	0.96	0.24
sSUM	0.52	0.52	1.23	1.06	0.27	**0.48**	**0.48**	0.96	0.93	**0.23**
rEUC	0.51	0.51	0.96	1.03	0.26	0.49	0.49	**0.82**	0.98	0.25
sEUC	0.51	0.51	1.37	1.03	0.26	**0.48**	**0.48**	1.10	0.93	**0.23**
rMED	**0.50**	**0.50**	0.82	1.00	0.25	0.49	0.49	**1.23**	0.98	0.25
sMED	0.51	0.51	0.96	1.04	0.26	0.49	0.49	1.10	0.97	0.24
wPWR	0.51	0.51	0.96	1.01	0.25	0.49	0.49	1.10	0.96	0.24
Rule	GC - *Run1*					GC - *Run2*				
	AUAC	ROC	EF_5 %	RIE	BEDROC	AUAC	ROC	EF_5 %	RIE	BEDROC
rRKP	0.48	0.48	0.83	0.93	0.22	0.50	0.50	0.65	0.99	0.24
rMIN	0.49	0.49	0.74	**0.99**	**0.24**	0.52	0.52	1.48	1.10	0.27
sMIN	0.49	0.49	0.83	**0.99**	**0.24**	**0.55**	**0.55**	**2.12**	**1.24**	**0.30**
rMAX	0.49	0.49	0.83	0.94	0.22	**0.49**	**0.49**	0.74	0.96	0.23
sMAX	**0.46**	**0.46**	0.74	**0.89**	**0.21**	**0.49**	**0.49**	**0.28**	**0.94**	**0.22**
rSUM	0.48	0.48	0.65	0.95	0.23	0.51	0.51	1.29	1.05	0.25
sSUM	0.48	0.48	0.83	0.94	0.23	0.51	0.51	0.92	1.01	0.24
rEUC	0.48	0.48	0.74	0.98	0.23	0.51	0.51	1.57	1.06	0.26
sEUC	0.47	0.47	**0.92**	0.91	0.22	0.50	0.50	0.55	0.96	0.23
rMED	0.49	0.49	0.55	0.98	**0.24**	0.51	0.51	1.11	1.06	0.26
sMED	**0.50**	**0.50**	**0.46**	0.98	0.23	0.51	0.52	1.01	1.03	0.25
wPWR	0.48	0.48	0.65	0.95	0.23	0.51	0.51	1.29	1.05	0.25

The score plots agree in indicating rRKP and rMAX as the best performing rules, followed by sMAX, sEUC and sSUM, then all the remainders, with rMIN being usually the last one. This overall ranking is clearly confirmed by the first component of the PCA on the average ROC, EF and BEDROC values (Figure 1d). These considerations confirm the results of Chen *et al.*, who indicated rRKP to be the best performing rule [7]. The explanation given by the authors was that the reciprocal rank closely approximates the probability that a structure in that position shows the same activity as the reference structure. wPWR is always located around the central area, indicating that it has mediocre performance. On the other hand, wPWR is always very close to the Best-Worst line, which suggests that its behavior is robust and not sensitive to the dataset. sMIN and sMAX, instead, seem very sensitive, as they lie far from the Best-Worst line in most cases. This can easily be verified by sorting the methods using the values reported in Tables 3 and 4.

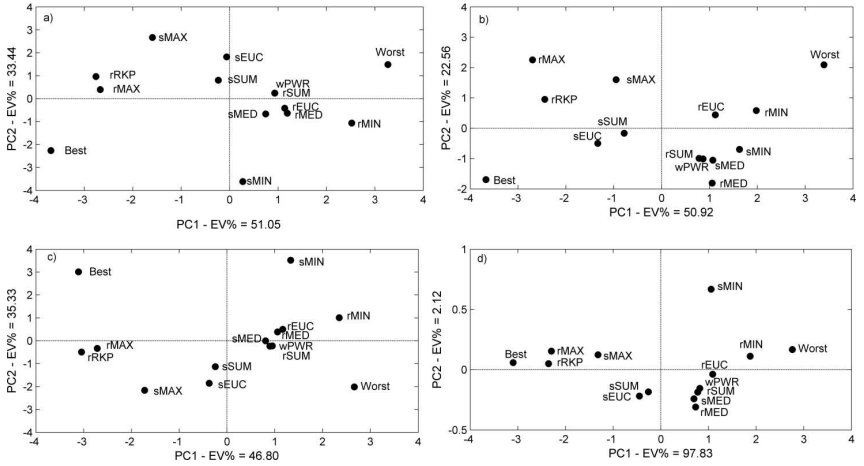

Figure 1. Principal Component Analysis on all the results evaluated by different metrics: (a) ROC; (b) EF; (c) BEDROC; (d) mean values of ROC, EF, BEDROC together.

In order to confirm the considerations drawn from the PCA, the performance of each rule was summarized by the mean and standard deviation calculated separately for each metric on all computational runs of Tables 3 and 4, i.e. average $EF_{5\%}$ and standard deviation of $EF_{5\%}$ on all runs, average RIE and so on. Moreover, the results of Tables 3 and 4 were used to rank the combination rules on each run based on each metric separately. From these rankings, the average, minimum and maximum rank and corresponding range were computed. For the sake of simplicity, we report only the information calculated for $EF_{5\%}$ (Table 5).

Table 5 confirms rRKP and rMAX to be the best performing rules, followed by sMAX, sEUC and sSUM, although none of the differences is significant at $\alpha = 0.05$. The considerations regarding wPWR are corroborated, i.e. it has a mediocre performance and low sensitivity to the particular dataset. The latter aspect is highlighted by a low standard deviation and the lowest range of the ranks, together with rSUM. However, these two rules are the ones that reach the highest minimum rank, i.e. they have difficulties giving very good results.

Table 5. Average statistics from the EF values on all datasets.

Rule	Avg. $EF_{5\%}$[a]	std. dev. $EF_{5\%}$[b]	Avg. Rank[c]	Min. Rank[d]	Max. Rank[e]	Range Rank[f]
rRKP	4.929	5.754	4.4	1.0	10.0	9.0
rMIN	2.812	3.881	8.8	3.0	12.0	9.0
sMIN	2.854	3.842	6.7	1.0	12.0	11.0
rMAX	4.850	5.537	4.9	1.0	11.5	10.5
sMAX	4.384	5.521	6.4	3.0	12.0	9.0
rSUM	3.550	4.935	7.1	4.0	10.0	6.0
sSUM	4.080	5.412	5.1	1.5	8.5	7.0
rEUC	3.285	4.525	8.1	2.0	11.5	9.5
sEUC	4.187	5.475	4.3	1.0	11.0	10.0
rMED	3.647	5.199	7.3	1.0	11.5	10.5
sMED	3.636	5.278	7.6	3.5	12.0	8.5
wPWR	3.509	4.894	7.4	4.0	10.0	6.0

[a] average $EF_{5\%}$ calculated on all runs; [b] standard deviation of $EF_{5\%}$ calculated on all runs; [c] average rank obtained by ranking the rules in each run on $EF_{5\%}$; [d] minimum rank obtained by ranking the rules in each run on $EF_{5\%}$; [e] maximum rank obtained by ranking the rules in each run on $EF_{5\%}$; [f] difference between minimum and maximum rank by ranking the rules in each run on $EF_{5\%}$.

We further analyzed the combination rules for their "early recognition" ability, i.e. their capability to place the actives in the very top positions of the final ranked list. Early recognition is important because in practice only a small percentage of all the compounds, selected from the top of the final ranked list, will be experimentally tested. To this end, we generated accumulation curves considering the top 1000 and top 100 ranked compounds in the list. We report the curves for the two runs were wPWR achieved its best rank (GC *Run2*) and worst rank (GC *Run1*) based on the ROC and AUAC metrics (Figure 2). Figure 2a and 2b indicate wPWR to have an overall intermediate behavior. A closer look at the top 100 compounds (Figure 2b) shows that sMAX, sMIN and sEUC are the rules that give the best "early recognition". sSUM and sMED do not retrieve any active compound in the top 100, but their overall AUAC and ROC values are greater than those of wPWR. This reflects the weakness of AUAC and ROC, i.e. their lack of consideration of the "early recognition" problem. Figure 2c and 2d report the accumulation curves for GC *Run2*. According to the overall AUAC and ROC, this is the case where wPWR performed the best compared to the other methods, even though its tied rank is 6.5, indicating average performance. Nevertheless, the accumulation curve of the top 100 compounds (Figure 2d) clearly shows that wPWR has the best curve up to the 25th compound in the list and it is overtaken only from position 38. From this point, the sMIN and rMIN rules have much nicer curves than all the other rules. Interestingly, in this run the rule rRKP, which overall achieves the best performance, does not retrieve any active in the top 100 compounds: the first active is ranked 188.

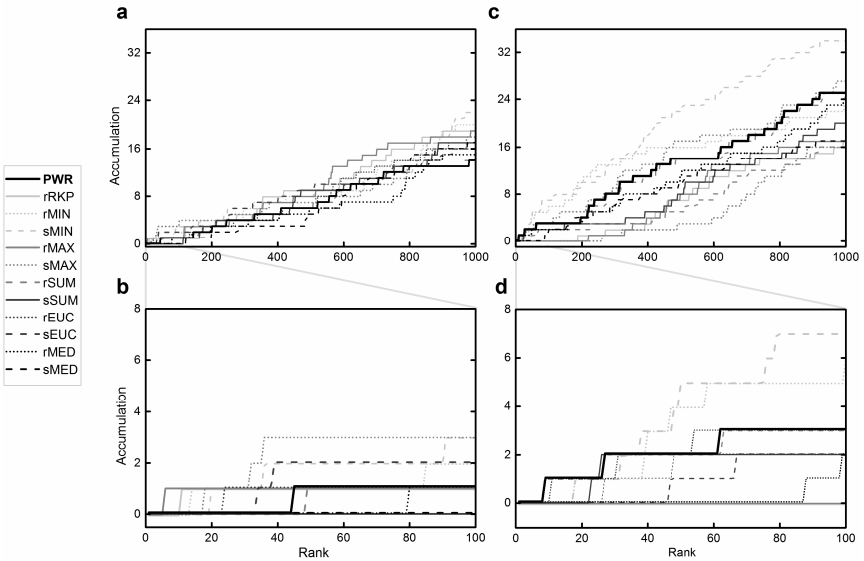

Figure 2. Accumulation curves: (a) GC *Run1*, top 1000 compounds; (b) GC *Run1* top 100 compounds; (c) GC *Run2*, top 1000 compounds; (d) GC *Run2* top 100 compounds.

To evaluate the scaffold hopping ability of the investigated rules, we calculated the relative scaffold diversity (%) of the actives in the top 100 list using the Murcko scaffolds. To visualize the results, we ranked the rules within each computational run and then produced a heat map from the ranks (Figure 3). rRKP, rMAX and sMAX appear to achieve the overall highest scaffold diversity (darker cells). Again wPWR shows mediocre performance and a robust behaviour across the different runs, as indicated by the similar colours. The behaviour of other rules, e.g. rMIN and sMIN, swings much more, as indicated by the fact that on the GC *Run2* they provided the best and second best scaffold diversity, but other times they gave the worst scaffold diversity (ALDR *Run1* and *Run2*, *PKM2 Run1* and *Run2*).

A last remark should be made regarding the computational time of wPWR, which is significantly larger than that of the other rules. The step that takes most of the time is the compilation of the matrix of pairwise comparison (**T**), followed by the calculation of the eigenvectors. Both steps present also potential memory problems if 10^5-10^6 compounds are present in the database. In this regard, Chen *et al.* concluded that best results are obtained when only a small percentage of such big databases are submitted to the fusion rules [7]. This

would come at hand for wPWR because it would allow to generate smaller **T** matrices, thus saving time and memory.

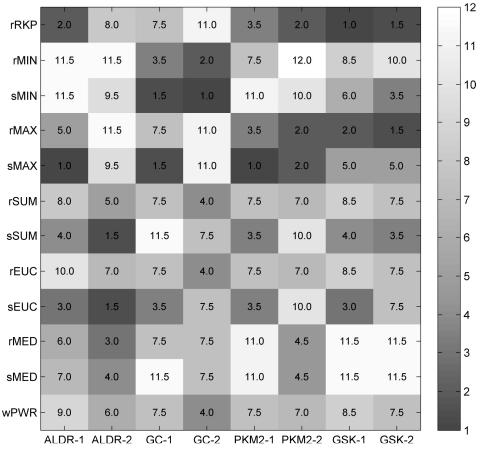

Figure 3. Heat map showing the rank of the fusion rules regarding scaffold diversity within each computational run.

4.2 Effect of weighting on wPWR

The usefulness of the weighting scheme in wPWR was investigated by defining three classes of activity (inactive, moderate and potent) from the potency values of the GC dataset. In general, one can work with only two classes of activity (inactive/active), but use additional criteria (e.g. economical cost, ease of synthesis, patents, etc.) to weigh the reference structures accordingly.

To analyze the results we generated accumulation curves for the top 1000 and top 100 compounds (Figure 4). Accumulation curves were plotted by considering both moderate and potent compounds (Figure 4a and 4b) and only potent ones (Figure 4c). Figure 4a seems to indicate that there is no net effect provided by the weighting scheme on this dataset. Figure 4b, which addresses the “early recognition” ability by zooming on the top 100 compounds, shows that the curves of weights 1:1 and 1:2 seem to have a better start and are taken over by the 1:4 curve at position 89. The 1:8 curve is always below the others, which may indicate a too unbalanced weighting scheme. Figure 4c shows the accumulation curve only for the potent compounds. It is clear that the curves are not good, since the first potent is found at

position 190, 355, 420, 432 with the weighting scheme 1:8, 1:1, 1:4 and 1:2, respectively. A partial justification of the poor "early recognition" can be the very low number of potent compounds, 34 out of 10477 (Table 2). Again, no net effect of the weighting scheme emerges; the most apparent point is though that the 1:8 scheme places the first potent more than 150 positions ahead than the other schemes. The accumulation curve for the top 100 is not shown for obvious reasons.

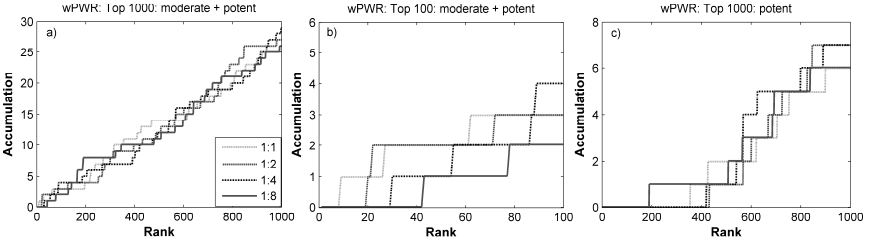

Figure 4. Accumulation curves for wPWR with different weighting schemes: (a) moderate and potent top 1000 compounds; (b) moderate and potent top 100 compounds; (c) potent top 1000 compounds.

5 Conclusions

In this pilot study, we applied the weighted Power-Weakness Ratio (wPWR) algorithm to four publicly available datasets to have indications of its potential as a combination rule for group fusion in virtual screening. We compared wPWR with six parameter-free rules by means of five metrics for virtual screening.

The results highlighted that rRKP is in general the best performing rule, thus confirming the conclusions of Chen *et al.*, even though it does not always outperform the other rules. wPWR achieved mediocre results on the analyzed datasets. However, one characteristic of wPWR that emerged in a number of parts of the analysis (PCA, ranks on $EF_{5\%}$ and on scaffold diversity) is its robustness, i.e. the ability to provide results that vary little with the dataset. On the contrary, the other rules seem more sensitive to the dataset and can behave either well or bad. The robustness of wPWR is an interesting feature for prospective virtual screening where the activity of the compounds in the database is not known and experimental testing is needed. Indeed, the use of wPWR would most likely not lead to the best results, but would also guarantee that the worst ones are avoided, which could instead happen also with a more sensitive well behaving rule.

The investigation of the weighting scheme suggested no net effect and difficulty in retrieving potent compounds (which were only 34 out of 10477). The investigation on the weighting scheme was carried out only on one dataset, hence more studies are needed to have better indications in regards to the usefulness of the weights.

Appendix A

The discrete formulas of the five metrics used to evaluate the performance of the virtual screening are reported here. A thorough analysis and description can be found in Truchon and Bayly [8].

The following notation is used: N is the total number of compounds in the database and n is the number of actives. R_a and R_i are the ratio of actives (n/N) and inactives ($(N-n)/N$), respectively; r_i is the rank of the i-th compound in the final list, sorted according to the output of the combination rules.

The Area Under the Accumulation Curve (AUAC) was calculated according to Equation A1.

$$AUAC = 1 - \frac{1}{nN}\sum_{1=1}^{n} r_i + \frac{1}{2N} \tag{A1}$$

where the last term is a correction for the discrete formula.

The area under the Receiver Operating Characteristic (ROC) was calculated according to its relationship with AUAC derived by Truchon and Bayly (Equation A2):

$$ROC = \frac{AUAC}{R_i} - \frac{R_a}{2R_i} \tag{A2}$$

The Enrichment Factor (EF) was computed as:

$$EF = \frac{\sum_{i=1}^{n} \delta_i}{\chi n} \quad \text{where } \delta_i = \begin{cases} 1 \; if \; r_i \le \chi n \\ 0 \; if \; r_i > \chi n \end{cases} \tag{A3}$$

where χ is the top percentage of ranked list that is evaluated and δ_i equals one if the i-th active is included in the considered top positions and zero otherwise.

The Robust Initial Enhancement (RIE) was calculated according to the following equation:

$$RIE = \frac{\frac{1}{n}\sum_{i=1}^{n} e^{-\alpha r_i/N}}{\frac{1}{N}\left(\frac{1-e^{-\alpha}}{e^{\alpha/N}-1}\right)} \tag{A4}$$

where α is a smoothing factor and $1/\alpha$ has a meaning similar to χ in EF. In this study α was defined as:

$$\alpha = Log_{10}(N) \tag{A5}$$

The Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC) was calculated as a scaled RIE, as follows:

$$BEDROC = \frac{RIE - RIE_{\min}}{RIE_{\max} - RIE_{\min}} \tag{A6}$$

where RIE_{min} and RIE_{max} are:

$$RIE_{\min} = \frac{1 - e^{\alpha R_a}}{R_a\left(1 - e^{\alpha}\right)} \tag{A7}$$

$$RIE_{\max} = \frac{1 - e^{-\alpha R_a}}{R_a\left(1 - e^{-\alpha}\right)} \tag{A8}$$

where α is the smoothing factor.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] A. R. Leach, V. J. Gillet, *An introduction to Chemoinformatics*, Springer, Dordrecht, 2007.

[2] B. Chen, C. Mueller, P. Willett, Evaluation of a Bayesian inference network for ligand-based virtual screening, *J. Cheminf.* **1** (2009) 1–5.

[3] P. Willett, Similarity searching in chemical structure databases, in: J. Gasteiger (Ed.), *Handbook of Chemoinformatics*, Wiley-VCH, Weinheim, 2003, pp. 904–915.

[4] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, P. Willett, Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets, *J. Chem. Inf. Model.* **52** (2012) 2884–2901.

[5] R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley-VCH, Weinheim, 2009.

[6] P. Willett, Enhancing the effectiveness of ligand-based virtual screening using data fusion, *QSAR Comb. Sci.* **25** (2006) 1143–1152.

[7] B. Chen, C. Mueller, P. Willett, Combination rules for group fusion in similarity-based virtual screening, *Mol. Inform.* **29** (2010) 533–541.

[8] J. F. Truchon, C. I. Bayly, Evaluating virtual screening methods: Good and bad metrics for the 'early recognition' problem, *J. Chem. Inf. Model.* **47** (2007) 488–508.

[9] R. Todeschini, F. Grisoni, S. Nembri, Weighted power–weakness ratio for multi-criteria decision making, *Chemom. Intell. Lab. Syst.* **146** (2015) 329–336.

[10] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, J. P. Overington, The ChEMBL bioactivity database: an update, *Nucleic Acids Res.* **42** (2014) D1083–D1090.

[11] D. J. Urban, W. Zheng, O. Goker-Alpan, A. Jadhav, M. E. LaMarca, J. Inglese, E. Sidransky, C. P. Austin, optimization and validation of two miniaturized glucocerebrosidase enzyme assays for high-throughput screening, *Comb. Chem. High Throughput Screen.* **11** (2008) 817–824.

[12] C. J. David, M. Chen, M. Assanah, P. Canoll, J. L. Manley, HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer, *Nature* **463** (2010) 364–368.

[13] M. M. Mysinger, M. Carchia, J. J. Irwin, B. K. Shoichet, Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking, *J. Med. Chem.* **55** (2012) 6582–6594.

[14] C. Nishimura-Yabe, Aldose reductase in the polyol pathway: a potential target for the therapeutic intervention of diabetic complications, *Nihon Yakurigaku Zasshi Folia Pharmacol. Jpn.* **111** (1998) 137–145.

[15] F. J. Gamo, L. M. Sanz, J. Vidal, C. de Cozar, E. Alvarez, J. L. Lavandera, D. E. Vanderwall, D. V. S. Green, V. Kumar, S. Hasan, J. R. Brown, C. E. Peishoff, L. R. Cardon, J. F. Garcia-Bustos, Thousands of chemical starting points for antimalarial lead identification, *Nature* **465** (2010) 305–310.

[16] *Dragon (Software for Molecular Descriptor Calculation). Version 7-Beta*. 2015, Talete srl, Milano.

[17] H. L. Morgan, the generation of a unique machine description for chemical structures – A technique developed at chemical abstracts service, *J. Chem. Doc.* **5** (1965) 107–113.

[18] D. Rogers, M. Hahn, Extended-connectivity fingerprints, *J. Chem. Inf. Model.* **50** (2010) 742–754.

[19] J. MacQueen, Some methods for classification and analysis of multivariate observations, *Proc. Fifth Berkeley Symp. Math. Stat. Prob.* **1** (1967) 281–297.

[20] R. W. Hamming, Error detecting and error correcting codes, *Bell Syst. Tech. J.* **29** (1950) 147–160.

[21] T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, J. L. Banks, Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening, *J. Med. Chem.* **47** (2004) 1750–1759.

[22] N. Triballeau, F. Acher, I. Brabet, J.-P. Pin, H.-O. Bertrand, Virtual screening workflow development guided by the 'receiver operating characteristic' curve approach. application to high-throughput docking on metabotropic glutamate receptor subtype 4, *J. Med. Chem.* **48** (2005) 2534–2547.

[23] R. P. Sheridan, S. B. Singh, E. M. Fluder, S. K. Kearsley, Protocols for bridging the peptide to nonpeptide gap in topological similarity searches, *J. Chem. Inf. Comput. Sci.* **41** (2001) 1395–1406.

[24] G. W. Bemis, M. A. Murcko, The properties of known drugs. 1. Molecular frameworks, *J. Med. Chem.* **39** (1996) 2887–2893.

[25] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, KNIME: The Konstanz information miner, in: D. Baier, F. Critchley, R. Decker, E. Diday, M. Greenacre, C.N. Lauro, J. Meulman, P. Monari, S. Nishisato, N. Ohsumi, O. Opitz, G. Ritter, M. Schader (Eds.), *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*, 2007, pp. 319–326.

[26] *MATLAB version R2012a*. MathWorks Inc, Natick.

[27] I. Jolliffe, Principal component analysis, in: B.S. Everitt, D.C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*, Wiley-VCH, New York, 2005.

[28] R. Todeschini, D. Ballabio, M. Cassotti, V. Consonni, N3 and BNN: Two new similarity based classification methods in comparison with other classifiers, *J. Chem. Inf. Model.* **55** (2015) 2365–2374.