Orphan drug legislation with data fusion rules using multiple fingerprints measurements

Moustafa Zein[1], Ahmed Abdo[1], Ammar Adl[1], Aboul Ella Hassanien[1,2], Mohamed F. Tolba[3] and Vaclav Snasel[4]

[1] Scientific Research Group in Egypt (SRGE), http://www.egyptscience.net
[2] Faculty of Computers and Information, Cairo University, Egypt
[3] Faculty of Medicine, Ain Shams University, Cairo, Egypt
[4] Electrical Engineering & Computer Science,VSB-TU Ostrava

Abstract. The orphan drug certification process from the European committee is depending on experts opinions that it is not similar to any other drug, this stage is very complicated and those opinions differ based on the expertise. So, this paper introduces computational model that gives one accurate probability of similarity, using multiple fingerprints measurements to similarity, and fuse these measurements by data fusion rules, that give one probability of similarity helping experts to determine that drug is similar to existing anyone or not.

Keywords Similarity coefficients, fingerprints, Orphan drug, drug legislation, Data fusion

1 Introduction

The development of a new drug from discovery to registration is a challenging, time-consuming and cost-intensive process [9], [10]. The consuming of the time and cost in design a new drug is reasonable if this is drug used in therapy of dangerous or orphan disease. Computational techniques contributed in finding alternatives to medical to medicinal chemistry experiments for studying a lot of chemical trends [9]. The work with rare diseases needs a lot of interest in development more than normal diseases. The drugs that can be process with rare diseases has a name is orphan drugs [15]. An orphan drug is a medicine that has been developed specifically to treat a rare medical condition. These medical conditions are normally referred to as rare diseases and there is much current interest in the development of orphan drugs for the treatment of such diseases [10].

An orphan drug has been got the certification from European committee as orphan drug; no similar drug can be brought to the European market for a period of ten years [10]. One of the most techniques that used in detect structure similarity is Similarity searching. It is techniques for ligand-based virtual screening [7][10]. There are three components that any similarity measure depend on (i) the representation that is used to describe each of the structures. (ii) Weighting scheme. (iii) Similarity coefficient [24, 3, 21]. There is no similarity measure has optimal solution, so we need to use more than one [16, 22].

Using the 2D fingerprint methods to support the assessment of structural similarity in orphan drug [10], Willett, Peter [27] described the use of supervised fusion rules and techniques in cheminformatics to solve the similarity problem. It first found application in similarity fusion, where a single reference structure is searched using different similarity measures; it has been extended to encompass multiple reference structures, this group fusion approach normally being noticeably superior to similarity fusion. Both approaches can benefit from the availability of training data linking similarity scores and probabilities of activity, but unsupervised fusion rules are available that enables effective searches to be carried out even in the absence of such data. Given the general success of data fusion, it is natural to seek a satisfactory theoretical description of the approach that could yield further improvements in screening effectiveness [27].

Fingerprint-based measures of similarity can be used to assess the structural novelty of molecules that are being submitted for consideration as new medicines for rare diseases. The results obtained by Willett, Peter [10] here demonstrate clearly that simple 2D fingerprint representations, provide measures of structural similarity that mimic closely the judgments of experts, using both training-set, molecule-pairs extracted from Drug Bank and test-set molecule-pairs typical of the work of the CHMP. This is so despite the fact that the two sets of molecules are rather different in character.

For every structure there is more than one measure, so in this study we will introduce to give use on accurate similarity that is Data fusion. Data fusion is the name given to a body of techniques that combine multiple sources of data into a single source, with the expectation that the resulting fused source will be more informative than will the individual input sources [27]. Data fusion has been widely used as way of enhancing the effectiveness of similarity-based virtual screening [7][10]. A fusion rule takes as input $n(n \geq 2)$sets of Nsimilarities or ranks and produces as output a single such set, normally as a ranked list from which the top-ranked database structures can be selected for further analysis. Fusion rules can either be unsupervised, meaning that the fusion rule operates directly on the similarity or rank information, or supervised, meaning that an additional training procedure is required [27][7].

The rest of this paper is structured as follows. Section 2 discuss the problem definition. Section 3 presents the proposed approach. Section 4 presents the experimental result. Finally, Section 5 addresses conclusions.

2 Problem discussion

In the European Union, medicines are authorized for some rare disease only if they are judged to be dissimilar to authorized orphan drugs for that disease. When a company applies to register a new medicine for an indication that has already been granted for an orphan medicine, it is the responsibility of the EMA's Committee for Orphan Medicinal Products (COMP) to decide if the new drug is indeed similar to an existing orphan drug, with an application being successful only when the COMP decides that this is not the case. To date, the evaluations carried out by the COMP have been based largely on human judgments of similarity [10, 17].

Using manual system to measure the similarity of drugs gives a low level of efficiency and need more and more time. From previous related work that using some of finger similarity coefficients to measure the similarity gives limited results that can't be help the chemist to take a final right decision. In addition if change some similarity measurements, the decision of chemist change, so we need to one verified output that represent more than one of similarity measurements that can help to build computational model to predict the probability of similarity to drugs and can be useful for data generalization. There are considerable differences between individual chemists in decision making that lead to do more efforts to reach a final decision about drug authority.

3 Orphan drug legislation with data fusion rules algorithm

3.1 Fingerprints measurements

In this study, we used a similarity measures for comparing chemical structures represented by means of fingerprints are (Dice, Tanimoto, Cosine, Kaczynski, McConnaughey) for ('ECFP6', 'MACCS', 'FCFP6', 'AVALON', 'RDK7', 'RDK6', RDK5', 'HASHAP', 'HASHTT') fingerprints for measuring the similarity of Data set of orphan drug. These fingerprints ranked by Sereina R. in his study [19] [4]. Circular fingerprints were developed more recently [28], and encode circular atom environments up to a certain bond radius from the central atom. There are four type of circular fingerprints extended-connectivity bit string (ECFP), extended connectivity count vector (ECFC), feature-connectivity bit string (FCFP) and feature-connectivity count vector (FCFC) , i.e. ECFP4, ECFC4, FCFP4 and FCFC4, as well as ECFP6 were compared. [28]. All fingerprints were calculated using the RDKit [19]. The RDKit fingerprint, a relative of the well-known Daylight fingerprint [13], is another topological descriptor.

3.2 Data fusion rules

Using Data fusion rules to give one probability of similarity to molecules is a methodology to solve the problem of more than one probability to the same structure similarity. Examples of such arithmetic fusion rules are shown in Figure 3 where, as before, dj denotes the jth database structure and where there are n sets of similarity scores or ranks to be fused (when considering these rules, it should be remembered that a large similarity score, e.g., 0.95 for a Tanimoto coefficient, will correspond to a small rank, i.e., at or near to the top of a ranked list) [27]. Algorithm (1) explain the solution path to the problem and how we can use fusion rules.

```
Define $NOD$ = 2 {No. of drugs.}
Define $NOF$ = 6 {No. of fingerprints measurements.}
Define $NOM$ = 100 {No. of molecules.}
Define $PS(u,v)$ {probability of similarity with authorized drug u and new drug v.}
Loop to $NOF$
$PS(u,v)$ = MACCS fingerprint(authorized drug, new drug)
for $i = 1$ to $N$ do
  for $j = 1$ to $N$ do
    Compute the similarity $si(dj)$ for the $j-th$ database-structure using the $i-th$
    similarity scoring function.
  end for
  for $j = 1$ to $N$ do
    Use function rule $F$ to combine the set of scores $nsi(dj)$ the $j-th$
    database-structure to gives its fused score, $FSj$.
    Rank the database in descending order of the fused scores $FSj$ [10].
  end for
end for
```

Algorithm 1: Fusion rule algorithm

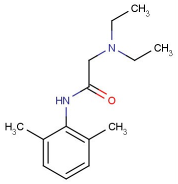
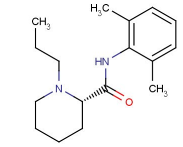
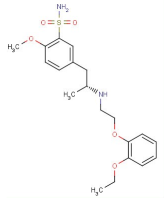
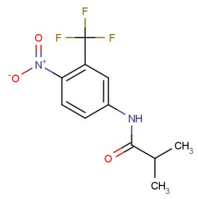
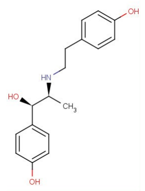
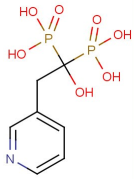
4 Experiments and discussion

The data that used in as a training set form DrugBank [3] with 100 molecule-pairs that compare between molecules structure of authorized drug A and molecules structure of new drug B with chemist opinion about molecules similarity that shown in Table (1). There are a lot of fingerprint algorithms with differences performance and no one is the best [25, 14]. In this study we use multiple actives as query molecules together with some kind of data fusion [19]. The benchmarking platform presented in this paper uses the RDKit [14], an open-source cheminformatics toolkit made available under the permissive Berkeley Software Distribution (*BSD*) license [19].

There are some results in Table (2) of fingerprints (ECFP6, MACCS, RDK7, and HASHAP). It includes the chemist opinion with acceptance and refusal for the similarity between molecules in authorized drug, and the molecules of new drug, in addition to the probability of similarity that calculated by fingerprint algorithms. These results show a lot of difference of similarity measurements [23, 8] from fingerprint to another that will be hard to study for chemist in the decision making process.

We can simulate the results from ECFP6 fingerprint, and compare similarity with the chemist opinion in Figure 1, that shows differences in these similarity like number (4,5) the probability of similarity from ECFP6 fingerprint is larger than chemist yes opinion, and in number (1,3,7) the chemist yes opinion is larger than the probability of similarity that was calculated by the fingerprint. In Figure (2), the differences is smaller than in Figure (1), but there are some undesired results like numbers (4,8,9).

Table 1: Molecules structure representation [3]

Molecule A	Molecule B	Chemist yes opinion	Chemist No opinion
		0.46	0.54
		0.31	0.69
		0.78	0.22

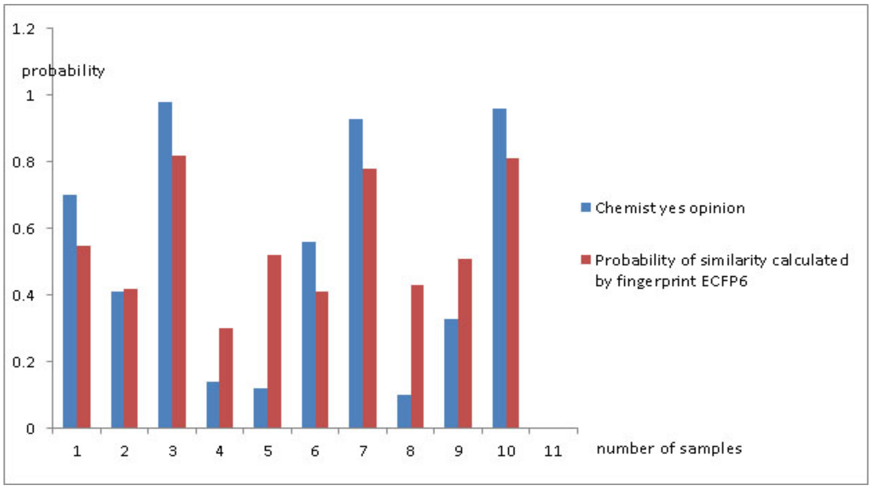


Fig. 1: Correlation between yes and similarity using ECFP6 Fingerprint

Table 2: Probability of similarity for
some fingerprint techniques and similarity measures.

Fingerprint technique	Chemist Yes opinion	Chemist No opinion	Probability of similarity calculated by fingerprint
ECFP6	0.70	0.29	0.55
	0.41	0.58	0.42
	0.98	0.01	0.82
	0.14	0.85	0.30
	0.12	0.87	0.52
	0.56	0.43	0.41
	0.93	0.06	0.78
	0.10	0.89	0.43
	0.33	0.66	0.51
	0.96	0.03	0.81
MACCS	0.70	0.29	0.75
	0.41	0.58	0.44
	0.98	0.01	0.95
	0.14	0.85	0.58
	0.12	0.87	0.40
	0.56	0.43	0.42
	0.93	0.06	0.83
	0.10	0.89	0.34
	0.33	0.66	0.62
	0.96	0.03	0.80
RDK7	0.70	0.29	0.57
	0.41	0.58	0.60
	0.98	0.01	0.81
	0.14	0.85	0.39
	0.12	0.87	0.42
	0.56	0.43	0.42
	0.93	0.06	0.80
	0.10	0.89	0.48
	0.33	0.66	0.59
	0.96	0.03	0.80
HASHAP	0.70	0.29	0.63
	0.41	0.58	0.43
	0.98	0.01	0.79
	0.14	0.85	0.19
	0.12	0.87	0.39
	0.56	0.43	0.48
	0.93	0.06	0.81
	0.10	0.89	0.39
	0.33	0.66	0.67
	0.96	0.03	0.77

From Table (3), the first two rules, *MAX* and *MIN*, involve assigning a database structuredj a score that is the maximum, or the minimum, similarity that it has achieved

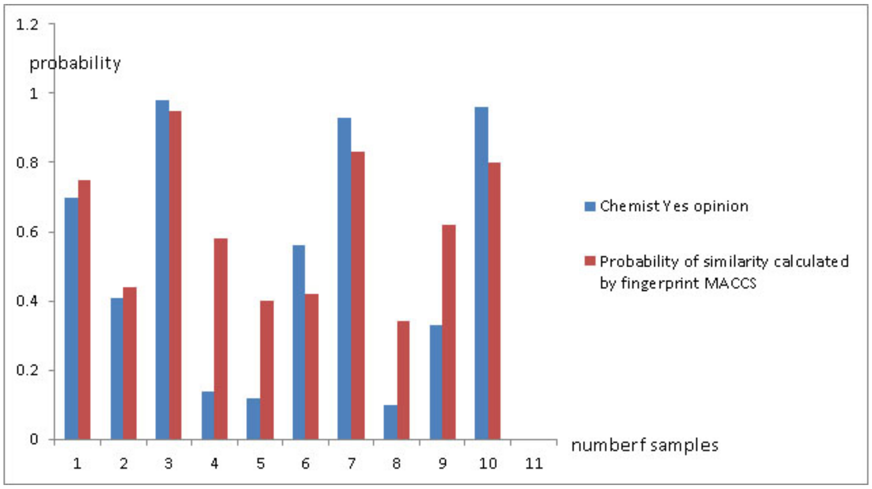


Fig. 2: Correlation between yes and similarity using MACCS Fingerprint

Table 3: Fusion rules

Fusion Rule Name	Equation
MAX	$max\{S_1(d_j), S_2(d_j) \ldots S_i(d_j) \ldots S_n(d_j)\}$
MIN	$min\{S_1(d_j), S_2(d_j) \ldots S_i(d_j) \ldots S_n(d_j)\}$
SUM	$\frac{1}{N}\sum_{i=1}^{N} S_i(d_j)$
MED	$med\{S_1(d_j), S_2(d_j) \ldots S_i(d_j) \ldots S_n(d_j)\}$

in the complete set of n similarity searches. The next two rules compute average scores: SUM is the arithmetic mean of the individual scores (or equivalently the arithmetic sum of them), while MED eschews the mean in favor of the median; the geometric or harmonic means might also be used for this purpose. Virtual screening often uses some cutoff similarity or rank position, with only those molecules above this threshold (e.g., the top-1% of the ranking) being passed on for further processing, this essentially meaning that all database structures below the threshold are assigned a score of zero.

Let p $(p \leq n)$ is the number of such nonzero similarity scores for a database structure; then the rules ANZ and MNZ are obtained by multiplying SUM by either $1/p$ or p, respectively, so that these rules focus on database structures that occur frequently above the threshold. The EUC rule views the set of similarity scores for each database structure as an n-dimensional vector and, then, computes the Euclidean norm for the vector [27].

Now, we have unique probability of similarity for every compound that be shown table3 that describe chemist opinion and data fusion rules output of the probability of similar-

Table 4: Unique probability of similarity using Data fusion rules.

Chemist yes opinion	Chemist no opinion	SUM Fusion Rule	MED Fusion Rule	Max Fusion Rule
0.70	0.29	0.69	0.63	0.75
0.41	0.58	0.48	0.43	0.60
0.98	0.01	0.91	0.81	0.95
0.14	0.85	0.15	0.39	0.58
0.12	0.87	0.17	0.42	0.52
0.56	0.43	0.53	0.42	0.48
0.93	0.06	0.86	0.81	0.83
0.10	0.89	0.19	0.37	0.43
0.33	0.66	0.36	0.62	0.67
0.96	0.03	0.93	0.80	0.80

ity between molecules structure similarity. There are three fusion rules in Table (3) that show one of probability of similarity between molecules structure of authorized and new drugs. We found that *SUM* rule give probability of similarity is closer to chemist yes opinion than *MED* and *MAX* as shown in Table (4).

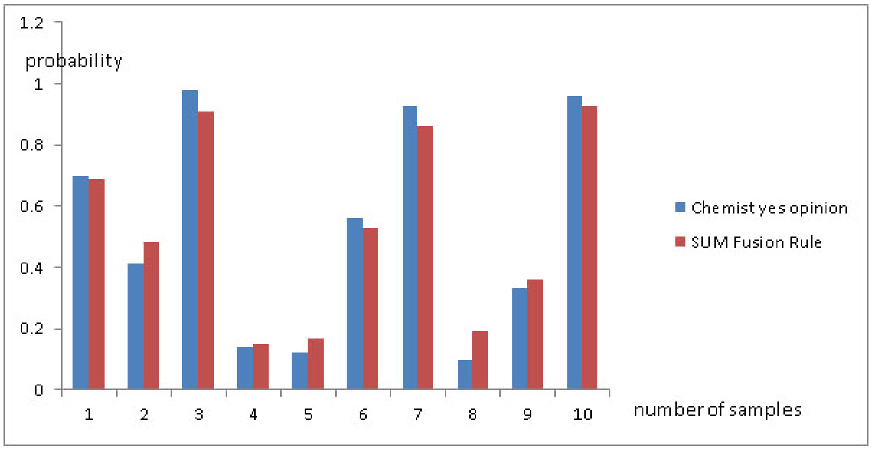


Fig. 3: Correlation between yes and similarity using SUM fusion rule.

Figure (4) and Figure (5) show the correlation between chemist yes opinion and the probability of similarity that calculated by fusion rule, and we can note that the results

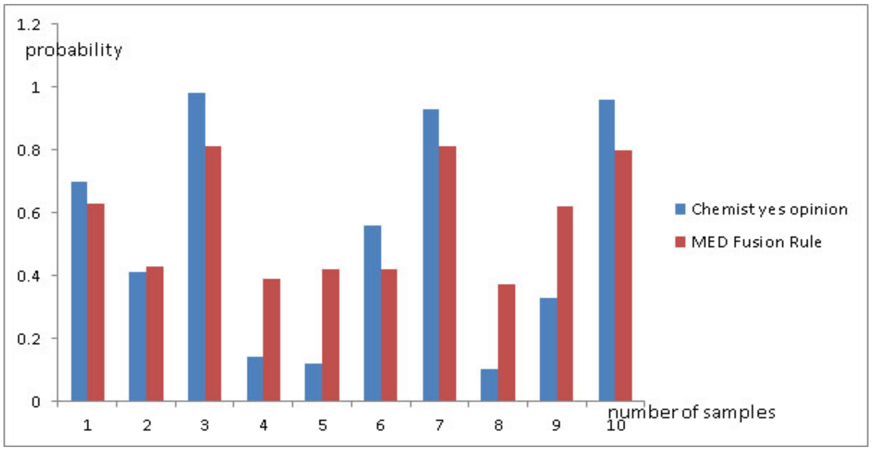


Fig. 4: Correlation between yes and similarity using MED fusion rule.

from *SUM* rule in Figure (4) is better than the result of *MED* rule in Figure (5), so in this paper we use *SUM* rule to detect the similarity of molecule-pairs in two drugs.

5 Conclusions

In this paper, we introduce a solution to the differences in similarity measurements of fingerprint algorithms and how these differences cause to the chemist some difficulty in the decision making process. We used fusion rules to solve and give the chemist an accurate probability of similarity to molecule-pairs structure in authorized drug, and new drug that needs to take a legislation to be shared in market.

Acknowledgment

This work was supported by the Bio-Inspired Methods: research, development and knowledge transfer project, reg. no. CZ.1.07/2.3.00/20.0073 funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic.

References

1. (2013) Cheminformatics and machine learning software. http://www.rdkit.org, accessed: 2013
2. (2013) Daylight theory manual. http://www.daylight.com/dayhtml/doctheory/index.pdf
3. (2013) Drug bank. http://www.drugbank.ca/
4. Bender A (2010) How similar are those molecules after all, use two descriptors and you will have three different answers. Expert opinion on drug discovery 5(12):1141-1151.
5. Bender A, Glen RC (2004) Molecular similarity: a key technique in molecular informatics. Organic & biomolecular chemistry 2(22):3204-3218.

6. Bender A, Jenkins JL, Scheiber J, Sukuru SCK, Glick M, Davies JW (2009) How similar are similarity searching methods? a principal component analysis of molec-ular descriptor space. Journal of chemical information and modeling 49(1):108-119.
7. Bonnet P (2012) Is chemical synthetic accessibility computationally predictable for drug and lead-like molecules? a comparative assessment between medicinal and computational chemists. European journal of medicinal chemistry 54:679-689.
8. Cross S, Baroni M, Carosati E, Benedetti P, Clementi S (2010) Flap: Grid molecular interaction fields in virtual screening. validation using the dud data set. Journal of chemical information and modeling 50(8):1442-1450.
9. Dutt R, Madan A (2012) Predicting biological activity: Computational approach using novel distance based molecular descriptors. Computers in biology and medicine 42(10):1026-1041.
10. Franco P, Porta N, Holliday JD, Willett P (2014) The use of 2dngerprint meth-ods to support the assessment of structural similarity in orphan drug legislation. Journal of Cheminformatics 6(1):5.
11. Gedeck P, Rohde B, Bartels C (2006) Qsar-how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. Journal of chemical information and modeling 46(5):1924-1936.
12. Geppert H, Vogt M, Bajorath J (2010) Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application ar-eas, and performance evaluation. Journal of chemical information and modeling 50(2):205-216.
13. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuenhauer A (2004) Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. Organic & biomolecular chemistry 2(22):3256-3266.
14. Jain AN, Nicholls A (2008) Recommendations for evaluation of computational methods. Journal of computer-aided molecular design 22(3-4):133-139.
15. Manley PW, Stie N, Cowan-Jacob SW, Kaufman S, Mestan J, Wartmann M, Wiesmann M, Woodman R, Gallagher N (2010) Structural resemblances and comparisons of the relative pharmacological properties of imatinib and nilotinib. Bioorganic medicinal chemistry 18(19):6977-6986.
16. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon JF, Cornell WD (2007) Comparison of topological, shape, and docking methods in virtual screening. Journal of chemical information and modeling 47(4):1504-1519.
17. Melnikova I (2012) Rare diseases and orphan drugs. Nature Reviews Drug Discovery 11(4):267-268.
18. Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D (2011) The cost of drug development: a systematic review. Health Policy 100(1):4-17.
19. Riniker S, Landrum GA (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. Journal of cheminformatics 5:26.
20. Ripphausen P, Nisius B, Bajorath J (2011) State-of-the-art in ligand-based virtual screening. Drug discovery today 16(9):372-376.
21. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. Journal of chemical information and modeling 50(5):742-754
22. Stumpfe D, Bajorath J (2011) Similarity searching. Wiley Interdisciplinary Reviews: Computational Molecular Science 1(2):260-282.
23. Swann SL, Brown SP, Muchmore SW, Patel H, Merta P, Locklear J, Hajduk PJ (2011) A unified, probabilistic framework for structure-and ligand-based virtual screening. Journal of medicinal chemistry 54(5):1223-1232.
24. Todeschini R, Consonni V, Xiang H, Holliday J, Buscema M, Willett P (2012) Similarity coeficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. Journal of chemical information and modeling 52(11):2884-2901

25. Truchon JF, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the early recognition problem. Journal of chemical information and modeling 47(2):488-508.
26. Willett P (2009) Similarity methods in chemoinformatics. Annual review of information science and technology 43(1):1-117.
27. Willett P (2013) Combination of similarity rankings using data fusion. Journal of chemical information and modeling 53(1):1-10.
28. Bender A, Jenkins JL, Scheiber J, Sukuru SCK, Glick M, Davies JW: (2009) How similar are similarity searching methods? A principal component analysis of molecular descriptor space. J Chem Inf Model 49:108-119.