A Linear Combination of Pharmacophore Hypotheses as a New Tool in Search of New Active Compounds – An Application for 5-HT$_{1A}$ Receptor Ligands

Dawid Warszycki[1], **Stefan Mordalski**[1], **Kurt Kristiansen**[2], **Rafał Kafel**[1], **Ingebrigt Sylte**[2], **Zdzisław Chilmonczyk**[3], **Andrzej J. Bojarski**[*,1]

1 Medicinal Chemistry Department, Institute of Pharmacology, Polish Academy of Sciences, Kraków, Poland, **2** Medicinal Pharmacology and Toxicology, Department of Medicinal Biology, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway, **3** Department of Cell Biology, National Medicines Institute, Warsaw, Poland

Abstract

This study explores a new approach to pharmacophore screening involving the use of an optimized linear combination of models instead of a single hypothesis. The implementation and evaluation of the developed methodology are performed for a complete known chemical space of 5-HT$_{1A}$R ligands (3616 active compounds with K_i < 100 nM) acquired from the ChEMBL database. Clusters generated from three different methods were the basis for the individual pharmacophore hypotheses, which were assembled into optimal combinations to maximize the different coefficients, namely, MCC, accuracy and recall, to measure the screening performance. Various factors that influence filtering efficiency, including clustering methods, the composition of test sets (random, the most diverse and cluster population-dependent) and hit mode (the compound must fit at least one or two models from a final combination) were investigated. This method outmatched both single hypothesis and random linear combination approaches.

Citation: Warszycki D, Mordalski S, Kristiansen K, Kafel R, Sylte I, et al. (2013) A Linear Combination of Pharmacophore Hypotheses as a New Tool in Search of New Active Compounds – An Application for 5-HT$_{1A}$ Receptor Ligands. PLoS ONE 8(12): e84510. doi:10.1371/journal.pone.0084510

Editor: Andrea Cavalli, University of Bologna & Italian Institute of Technology, Italy

Received August 21, 2013; **Accepted** November 22, 2013; **Published** December 18, 2013

Funding: This study was supported by the project UDA-POIG.01.03.01-12-063/09-00 Prokog, which is co-financed by the European Union from the European Fund of Regional Development (EFRD) and by Norway Grants within the Polish-Norwegian Research Programme, grant no. Pol-Nor/198887/73/2013. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: bojarski@if-pan.krakow.pl

Introduction

A pharmacophore model (also called a pharmacophore hypothesis) is one of the most important concepts in medicinal chemistry. It is defined as the spatial orientation of different features of a molecule (thus, pharmacophore modeling is a ligand-based method) required for the activity towards a biomolecular target [1–3]. Such a model can be used to describe a large number of structurally diverse compounds with only a handful of general features. Pharmacophore filtering is widely used in virtual screening campaigns [4–10] and in other drug development processes [11,12] This filter may be applied as a standalone [7,8] or as one of the subsequent steps in a screening cascade [9,10].

The attempts at pharmacophore modeling the known ligands of 5-HT$_{1A}$R [13–27], a well-recognized therapeutic target [28,29] also intensively studied in our laboratory [30–33], have focused solely on visualizations and explanations in SAR studies [13–27]. Only very recently published pharmacophores of 5-HT$_{1A}$R ligands were intended for use in virtual screening (VS), however only for the off-target activity of α1-adrenoceptor antagonists [34].

It is nearly impossible to define a universal model that covers the entire chemical space of the ligands of a particular target. The use of multiple models at once led to search parameter improvements, yet the arbitrariness of model selection makes it strongly dependent on the researcher's knowledge and experience. To address the downsides of pharmacophore screening, we developed a novel approach involving the use of a carefully selected collection of pharmacophore models instead of a single hypothesis. The primary goal of the research was to develop and to evaluate a screening protocol that utilized a linear combination of pharmacophore models, i.e. a collection of individual hypotheses covering as much as possible the chemical space defined by ligands of a particular target. From the single hypotheses, created from ligand clusters, a group of models with the best combined performance (chosen using a homemade script) was selected and evaluated on various test sets. In addition, the proposed best combinations were compared with a single hypothesis and with randomly composed groups of pharmacophore models of equal length.

Materials and Methods

General

Figure 1 shows a protocol applied for the development of the optimal linear combination of the pharmacophore models. Compounds with proven activity toward 5-HT$_{1A}$R were acquired from the ChEMBL database version 5 [35], clustered, and structures representative of each cluster were used for the construction of a pharmacophore hypothesis. Each model underwent evaluation via test sets composed of active compounds, true decoys and ligands retrieved from the DrugBank [36] that were assumed inactive. The best linear combinations of pharmacophore models were then composed and validated with new compounds (both active and inactives) retrieved from the ChEMBL database v10 [37].

Data sets

The source of the active compounds was the ChEMBL database version 5 (August 2010), containing 5-HT$_{1A}$R ligands retrieved from approximately 520 published papers. Due to a large diversity of activity measures, only the compounds with defined K_i (IC$_{50}$ – assumed as 2×K_i, pK_i or pIC$_{50}$ were converted to K_i) as assayed on human cloned receptors or on rat cloned or native receptors, were taken into account. In the case of multiple data for one ligand, the K_i and human receptors were given preference; a median value was used in the case of many biological results. The ligands were defined as active when their binding constant was lower than or equal to 100 nM; the threshold of inactivity was set at 1000 nM. The resulting sets consisted of 3616 active (more than half with K_i values lower than 10 nM – see Figure 2 for details) and 438 inactive (decoy) compounds (Figure 1).

Clustering

Three methods of clustering were applied: 3D pharmacophore fingerprint-based (P3D), MOLPRINT 2D fingerprint [38]-based (M2D) and the manual – classical method (grouping the compounds by a common core). The P3D and M2D approaches were performed using the Hierarchical Clustering feature in Canvas [39].

For the P3D method, 31 clusters were created using the Kelly criterion [40]. After merging the singleton and doubleton subsets into a special class, 28 clusters containing 8–497 compounds were obtained.

The same approach, applied to MOLPRINT 2D fingerprints, left one cluster considerably larger than the rest, and its recurrent splitting (applied four times) resulted in a total of 36 collections consisting of 6 to 744 compounds each.

The manual clustering generally followed the classification of 5-HT$_{1A}$R ligands described in the literature (9 basic classes) [41–43]. However, more subgroups were then created, e.g. for arylpiperazines [44] (Figure 3). In the case of the alkylamines (714 compounds), indole derivatives were first extracted and, with the exception of the tetrahydropyridoindoles, were divided depending on the distance between two crucial pharmacophore features: an aromatic system and a basic nitrogen atom. The entire procedure developed 28 clusters, each containing 17 to 605 compounds (Figure 3).

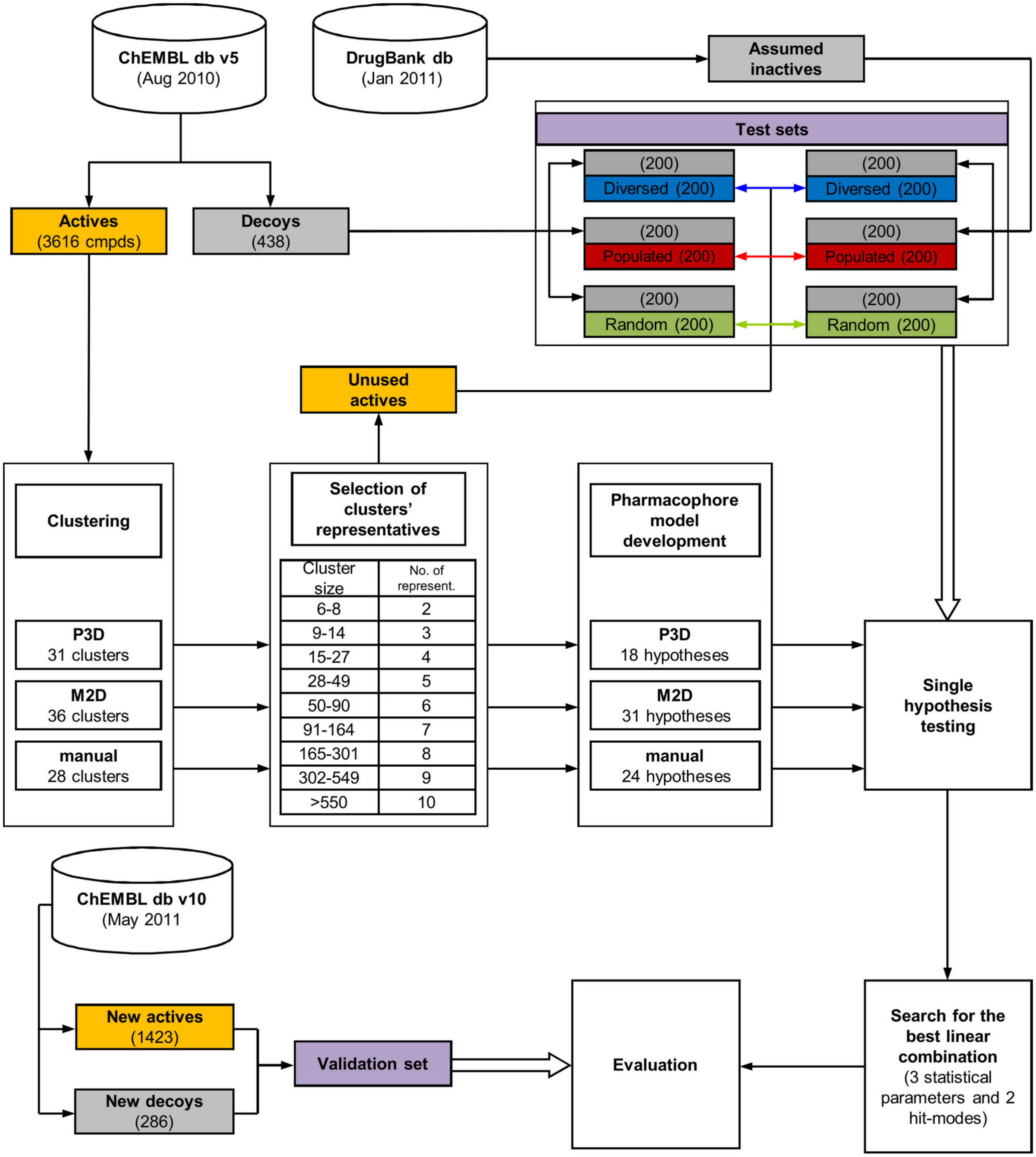
Pharmacophore model development

Cluster representatives, in number proportional to the cluster's size (Figure 1), were selected using the diversity-based selection tool in Canvas (similarity metric – Soergel distance; compounds selection algorithm – sphere exclusion; sphere size – 0.5; initialization – random with random seed). The selected representatives were further used as a basis for the development of a pharmacophore model using Phase [45] under default settings (conformers generated during search, 10 conformers per rotatable bond; not more than 100 conformers per structure; relative energy window between conformations – 10 kcal/mol; RMSD tolerance for match – 2 Å). The best hypothesis for each ligand class must have mapped at least half of the input compounds. Among these hypotheses, the one with the maximum number of features, the highest matching rate and the best selectivity score was selected for further research. For some of the clusters, none of the hypotheses met all the requirements (13 for P3D, 5 for M2D and 4 for manual grouping).

Test sets

All obtained models were first tested on three pairs of 400-compound sets that were composed of an equal number of active and inactive compounds (Figure 1). Of the 5-HT$_{1A}$ ligands not used in the development of the models, the sets of active ligands were selected (i) randomly (marked in green in Figures 1, 4-6 and Figures S1–S6), (ii) to be the most diverse (blue), or (iii) in a way reflecting the abundance of different scaffolds according to manual clustering (red). Out of the 438 compounds from the ChEMBL database with K_i (5-HT$_{1A}$) > 1000 nM, the 200 most diverse compounds (diversity-based selection tool in Canvas) constituted the set of decoys for the above-mentioned ligand groups (Figure 1). Similarly, the active ligands were complemented with compounds with assumed inactivity from the DrugBank database (low potential for binding to 5-HT$_{1A}$ receptor was confirmed by SEA search tool [46]). The most diverse molecules with polarizable nitrogen and no data regarding activity toward the 5-HT$_{1A}$ receptor were selected. All statistical parameters were calculated as an average of the values obtained for the pair of sets that was composed of the same actives (Figure 1).

A validation set was created out of the novel 5-HT$_{1A}$R ligands that appeared in the ChEMBL database version 10 (May 2011). The inactive compounds from this set were a challenge for the method because they were very similar to active compounds. Similarity search using MOLPRINT 2D fingerprint and Tanimoto metric revealed that 24.8% of inactives compounds had similarity coefficient with actives of 0.9 or higher.



Cluster size	No. of represent.
6-8	2
9-14	3
15-27	4
28-49	5
50-90	6
91-164	7
165-301	8
302-549	9
>550	10

Figure 1. The development of an optimal combination of pharmacophore models. The development of an optimal combination of pharmacophore models. Transparent boxes show the logical steps of the workflow; cylinders represent data sources; colored boxes reflect the compound character: gray – inactives, orange – actives or the active's selection method (blue, red or green), which is consequently used in subsequent figures. The population of the compound set is given in brackets. Thick arrows indicate the use of data sets.

doi: 10.1371/journal.pone.0084510.g001

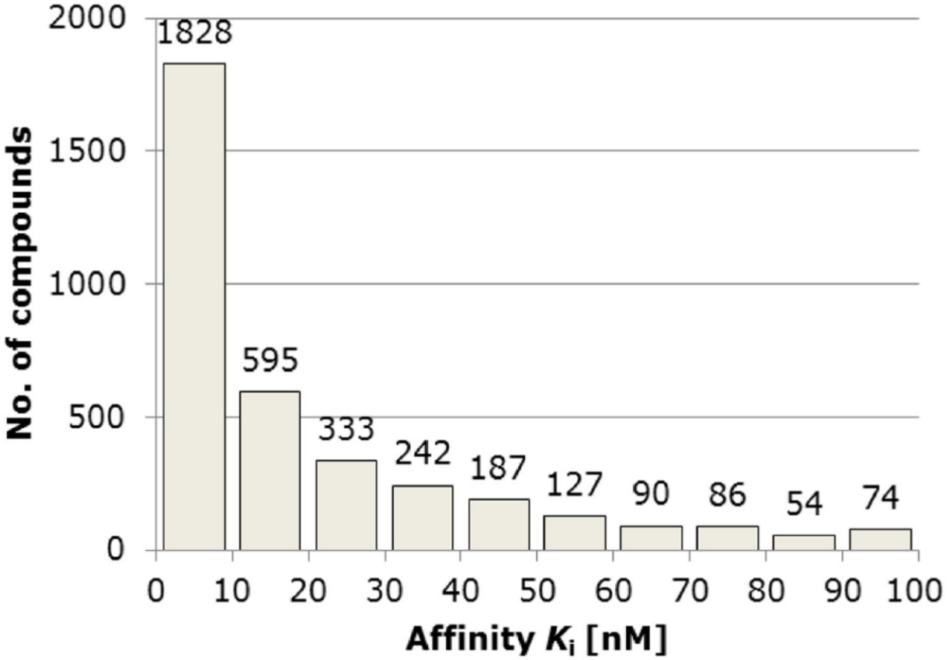


Figure 2. Affinity distribution of 3616 5-HT$_{1A}$R ligands retrieved from the ChEMBL database version 5. Affinity distribution of 3616 5-HT$_{1A}$R ligands retrieved from the ChEMBL database version 5.

doi: 10.1371/journal.pone.0084510.g002

Search for the best linear combination of models

The process of selecting the optimal linear combination was conducted using an in-house script (see Figure S7) because the amount of data and number of combinations (hypotheses from three clustering methods, two hit modes and three different actives sets) rendered manual evaluation nearly impossible. The tool recursively generated all possible combinations of a given length and selected a top-scored combination in terms of the optimized parameter, namely, the Mathews Correlation Coefficient (MCC), accuracy or recall which was obtained using the average of the values received for the pairs of actives vs. the assumed inactives and the actives vs. the decoys.

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

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\mathrm{Re}call = \frac{TP}{TP+FN}$$

Where *TP* stands for the number of true positives (actives labeled as actives), *TN* – true negatives, *FP* – false positives (inactives labeled as actives) and *FN* – false negatives.

MCC takes values from –1 to +1, where +1 represents perfect prediction, 0 represents random prediction and –1 represents an inverse prediction, whereas the accuracy and recall ranged from 0 to 1.

Two modes of compound filtering were evaluated. The "hit-once" mode classified the ligand as active if it was recognized by at least one of the models in the combination and in the "hit-twice" mode if at least two of the hypotheses flagged the ligand as active.

Results

Because the method is designed for VS, various factors influencing filtering performance were investigated. Starting from the active compounds clustered using three different methods, a series of pharmacophore hypotheses were developed (one model per cluster, see sample hypothesis in Figure 7). From the pool of singular models, linear combinations of various lengths were formed (the hypotheses retrieved for different clustering methods were not mixed) and evaluated using diverse test sets. Three coefficients were optimized at two restriction levels (hits must have been recognized by at least one or two models): MCC, accuracy and recall, as the standard measures of screening performance.

Development of the optimal linear combination

The analysis of the approximation to the optimal ensemble of models showed that adding subsequent hypotheses allows for the saturation of the chemical space of the 5-HT$_{1A}$R ligands until the maximum value of the optimized parameter is reached (Figure 8).

The maximization of the MCC parameter led to 6–11 models long combinations for the hit-once and 10–13 of those for the hit-twice mode, depending on the test set/clustering scheme,

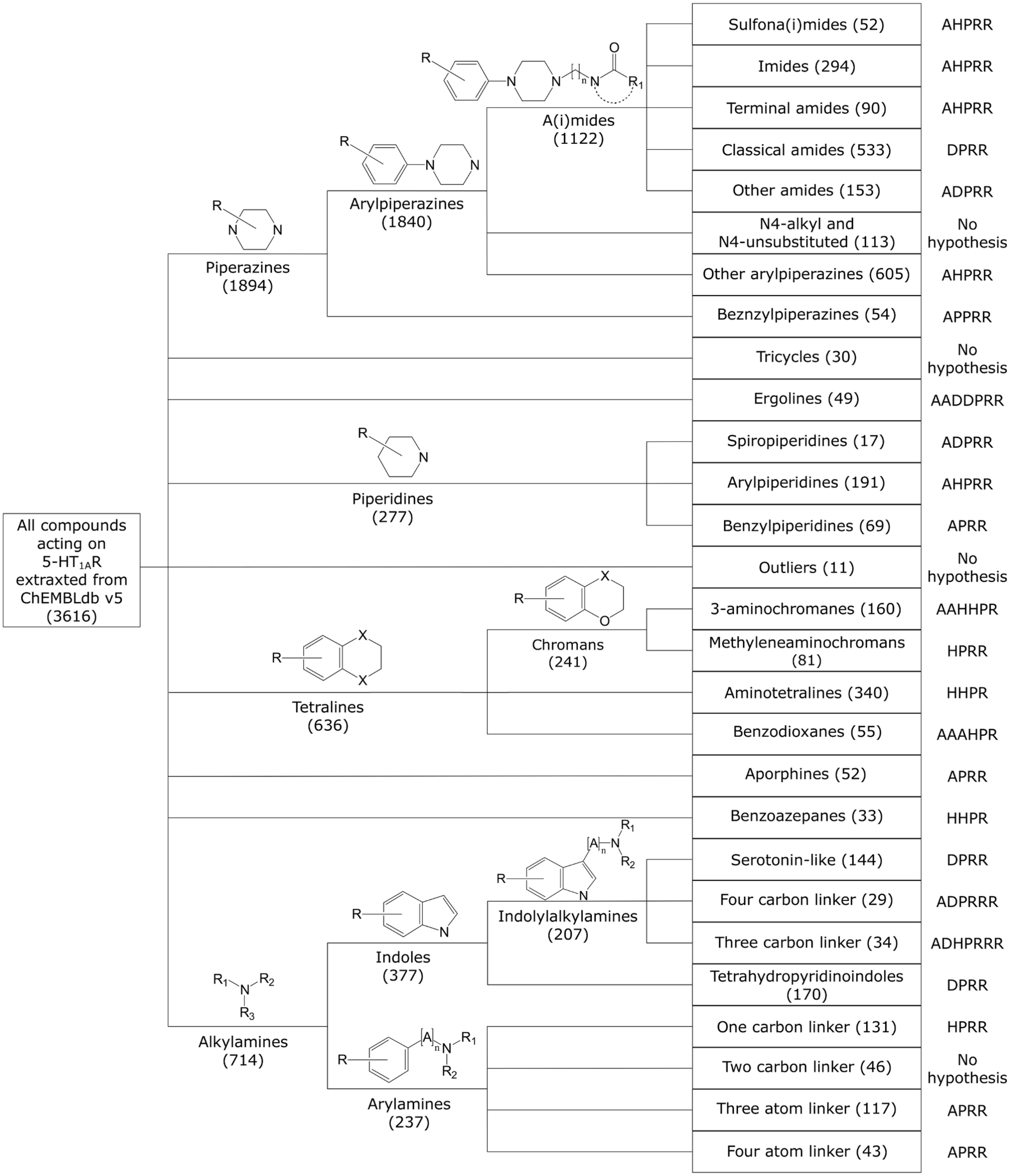


Figure 3. A dendrogram obtained using the manual clustering procedure. A dendrogram obtained using the manual clustering procedure. The number of compounds comprising each cluster is given in brackets. The last column presents a feature composition of the pharmacophore model created for a given cluster. The feature abbreviations used are: hydrogen bond acceptor – **A**, hydrogen bond donor – **D**, hydrophobic group – **H**, positively charged group – **P**, aromatic ring – **R**.

doi: 10.1371/journal.pone.0084510.g003

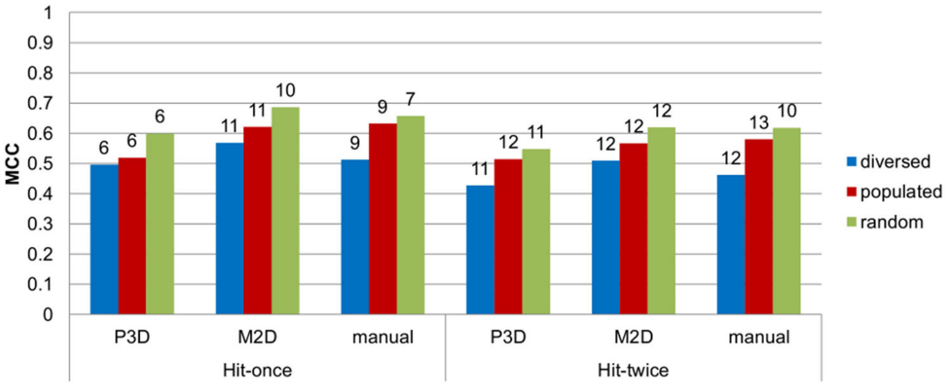


Figure 4. The optimized values of MCC for each possible scheme. The optimized values of MCC for each possible scheme. The length of combination is shown at the top of the bars. The composition of combinations based on the manual clustering approach is shown in Figure 5.

doi: 10.1371/journal.pone.0084510.g004

and the range of the maximum MCC values was from 0.427 to 0.686 (Figure 4). Figure 9 shows details of the MCC-optimized linear combination of 7 models developed on manual clustering, random test set and hit-once mode. The MCC at the highest level indicates misclassifications of only 12% of the active ligands and of one third of the inactive ligands. The experiments proved that the “hit-once” method was slightly better than the “hit-twice” method, and the difference between the best respective combinations was 0.069. In terms of clustering methods, the M2D and manual methods outmatched the approach based on 3D pharmacophore fingerprints.

The analysis of the top-scored combinations revealed frequent occurrences of short hypotheses (formed from four or five features), yet the size of the cluster, the feature count of the pharmacophore model and the pharmacophore efficiency could not be correlated with the performance of the combination. For example, the benzylpiperidines cluster (consisting of only 55 compounds) produced a short, four-feature hypothesis occurring in 17 of 18 optimized combinations. However, the hypothesis representing the largest cluster (other arylpiperazines) was not part of any combination. The results may also suggest that the hypotheses with high feature counts (e.g. ergolines with a seven-element hypothesis) are too strict to participate in optimal combinations; however, there is no statistically significant evidence to support this statement (Figures 5, S5, S6 and S8).

Regarding accuracy optimization, the process established the length of the combination on 6–11 hypotheses for the hit-once and 11–16 for the hit-twice approach. Again, the M2D fingerprint-based clustering method led to the best results (an accuracy of 0.840 for the random actives test set). The hit-once method of optimization was better than the hit-twice; however, this difference did not exceed 0.049. The details of the experiments can be found in Figure S1 .

The optimization of the recall returned compositions of 9 to 21 pharmacophore models for the hit-once method and 12 to 25 for the hit-twice method, and the values of the recall ranged from 0.445 to 0.920 (Figure S2). In that case, the combination curve reached a plateau after climbing to the maximum value. This effect was caused by a lack of FP count in the parameter definition; thus, the misclassifications were ignored. The results for the hit-once method were significantly better than those for the hit-twice selection. In addition, the combination for the hit-twice selection was longer (the greatest difference in length was for the manual/diverse scheme (13 hypotheses)). In this experiment, the hypotheses based on manual clustering dominated because they provided the best combination of the populated actives set and the random actives, as well as the best overall linear combination in terms of recall.

Validation results

An ensemble of top combinations in terms of MCC for each construction scheme was tested using the validation set (Figure 6), again showing the superiority of M2D clustering-based models and the hit-once search approach. The best MCC coefficient was 0.294, which was a low but acceptable value (MCC is normalized in the range of –1 to 1) for such a demanding test set. A validation set consisted of an imbalanced amount of active compounds (1423) in relation to decoys (286).

The validation of the best accuracy-optimizing combinations confirmed the advantage of the M2D clustering method. Manual clustering showed better results than the P3D method but was unable to compete with the M2D method. The highest accuracy obtained for a validation set was 0.710 for the M2D/populated/hit-once scheme (Figure S3).

Combinations reaching the highest recall values achieved up to 0.743 for the validation set for the M2D/diverse/hit-once scheme (Figure S4). Again, in this case, the M2D clustering-based models performed better than the other models. The P3D clustering method also showed the worst results in that system.

Random combination

The aim of this benchmark was to determine whether the performance of the combination of hypotheses was not

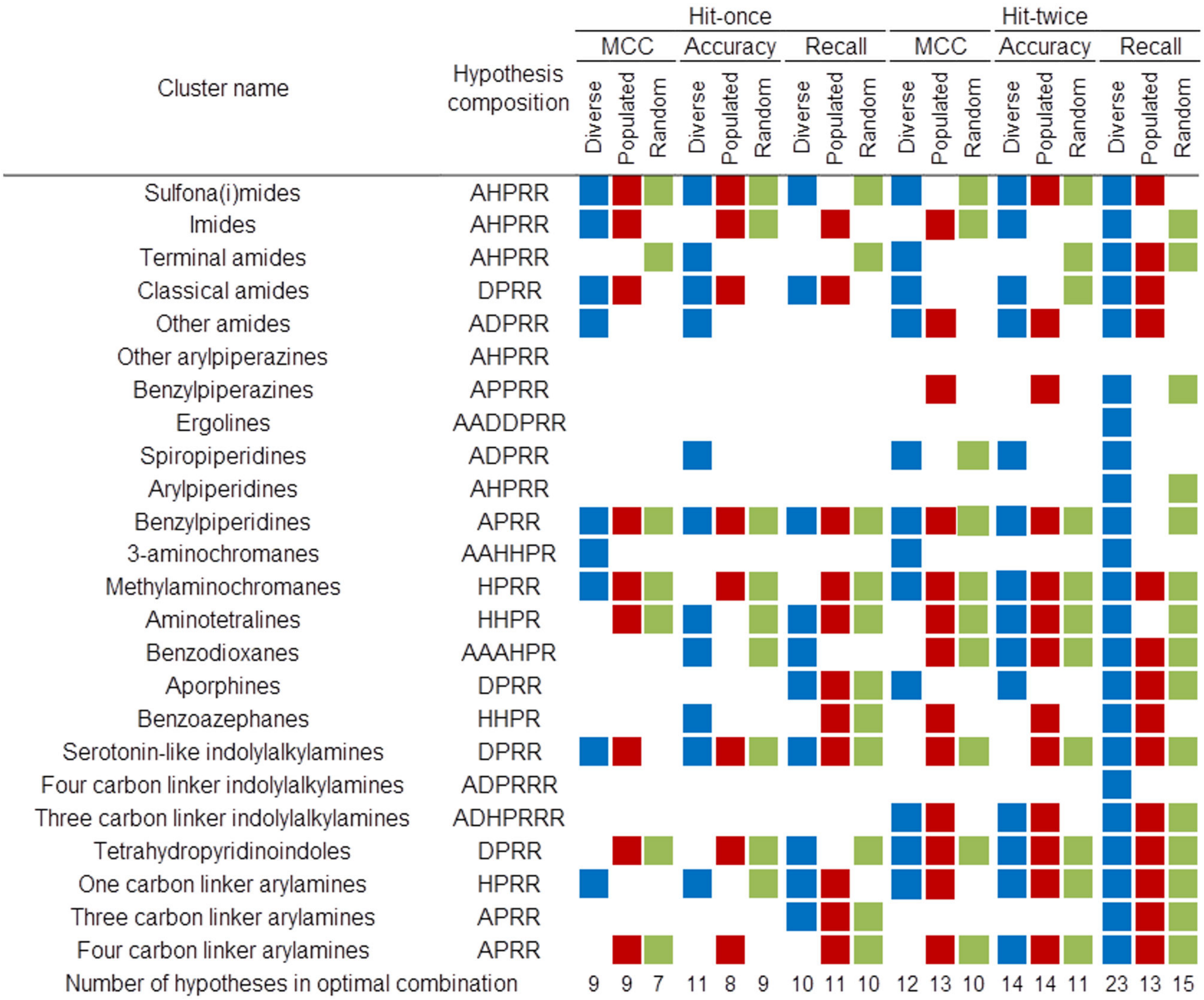


Figure 5. A composition of each top-ranked linear combination obtained using the manual clustering procedure. A composition of each top-ranked linear combination obtained using the manual clustering procedure. Each filled square denotes presence of a hypothesis developed on a particular cluster in the optimal combination for appropriate conditions. Colors code the type of the test set: blue – diverse, red – populated, and green – random. The last row contains the total number of hypotheses forming a respective top-ranked combination. The values of the optimized statistical parameters for manual clustering are shown in Figure 4, and those for accuracy and recall are shown in Figures S1 and S2, respectively. The exemplary linear combination (manual/random/hit-once; 7 hypotheses long) is shown in Figure 9.

doi: 10.1371/journal.pone.0084510.g005

influenced by size alone. For the three schemes that generated the best combinations for the respective statistic parameter (M2D/random/hit-once for MCC and accuracy and M2D/populated/hit-once for recall), ten random collections of hypotheses containing the same number of elements as the optimized hypotheses (10 for MCC and recall and 8 for accuracy) were prepared, and the respective statistics were calculated and averaged. The results (Table 1) clearly showed the superiority of the optimized combination over the random hypotheses ensemble, especially in the case of MCC.

Single hypothesis benchmark

The benchmark against the single pharmacophore hypothesis was essential in comparing its performance with the proposed approach. To cover the full chemical space of the 5-$HT_{1A}R$ ligands, a representative from each cluster was selected (either a centroid or a random pick) to develop a single (universal) hypothesis that was then tested on the validation sets.

The results (Table 2) showed that the single hypothesis performed similarly to the P3D-based combination of hypotheses in terms of MCC. However, for all other parameters

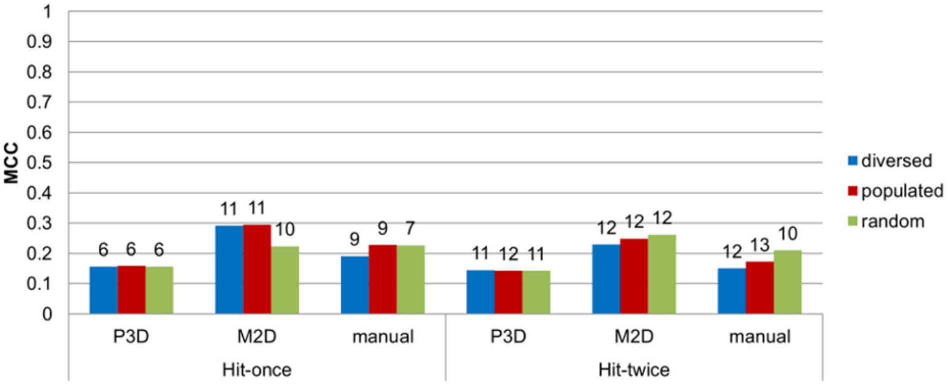


Figure 6. The MCC results for the validation of the top linear combinations. The MCC results for the validation of the top linear combinations. The length of combination is shown on top of the bars.
doi: 10.1371/journal.pone.0084510.g006

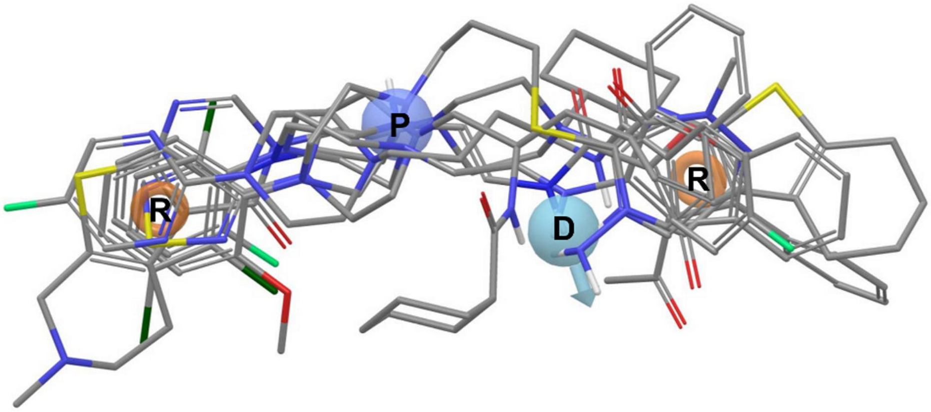


Figure 7. Exemplary pharmacophore hypothesis selected for arylpiperazines with classical amide fragment. Exemplary pharmacophore hypothesis selected for arylpiperazines with classical amide fragment mapping 6 out of 10 cluster representatives. The model fit 462 of the 533 compounds (87%) in the cluster. The feature abbreviations are: hydrogen-bond donor – **D**, positively charged group – **P**, aromatic ring – **R**.
doi: 10.1371/journal.pone.0084510.g007

and combination schemes, the single hypothesis was significantly outmatched.

Discussion

Because the method was designed for VS purposes, the high efficiency of such an experiment was the primary concern. The results showed an increased performance in the linear combination of pharmacophore models in VS compared with the single hypothesis, as measured using the standard parameters of MCC, accuracy and recall. The screening evaluation of the method, however, precipitated observations that require further discussion.

The process of finding the optimal linear combination of pharmacophore models is resource- and time-consuming. The combination of twelve element sets out of twenty-four hypotheses led to nearly three million possible combinations. A subsequent evaluation of all of these combinations required a significant amount of resources and thus was a challenging task even for a powerful workstation. However, once the optimal combination is selected, the screening process is conducted in an amount of time comparable to that required by a single hypothesis approach.

The ensemble of pharmacophore models shows a reasonable performance in declining the compounds assumed to be decoys (up to 198 out of 200 properly classified), yet the more challenging true decoy set remains an issue. None of the proposed combinations found all of the active compounds from the test sets. The reason for this is the presence of clusters that did not produce a pharmacophore hypothesis that was suitable for screening and thus did not support the coverage for the chemical subspace of the active compounds. Thus, the importance of the choice of the clustering method (being the fundament of single hypothesis development) and algorithm

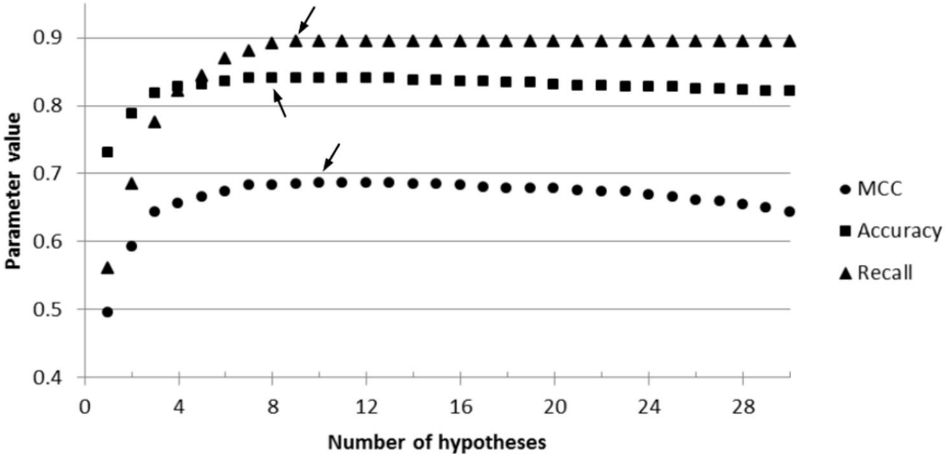


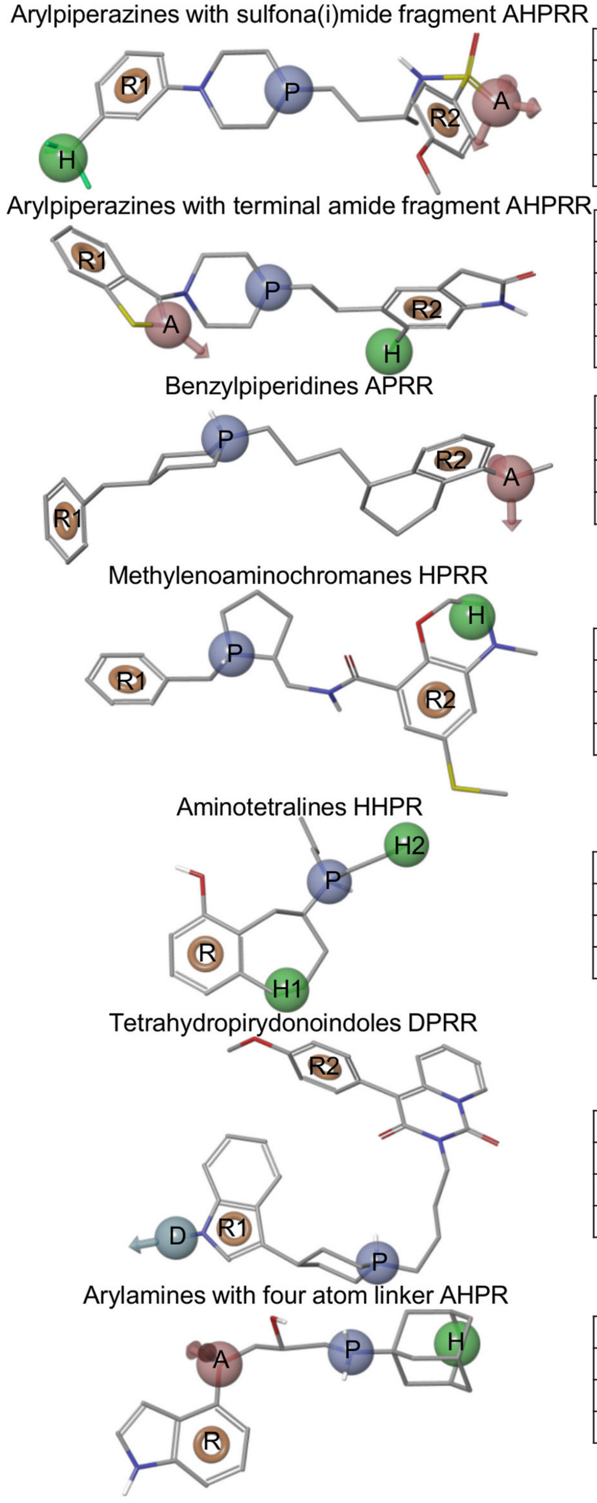
Figure 8. An optimization curve for the investigated parameters of a top-ranked linear combination of MCC. An optimization curve for the investigated parameters of a top-ranked linear combination of MCC (M2D/random/hit-once); arrows indicate the maximum value: MCC reached a rate of 0.686 for 10 hypotheses (also see Figure 6); the optimization of accuracy and recall had the highest values for a combination of 8 and 9 hypotheses, respectively (also see Figures S1 and S2).
doi: 10.1371/journal.pone.0084510.g008

used should be adjusted to the goals of the screening. The P3D method provided the best filter for decoy structures, but its performance in finding active compounds was significantly weaker, thus lowering the measured VS parameters. However, M2D showed an increased rate of locating active compounds at the cost of decoy recognition. M2D is the best method given all optimized parameters. The performance of manual clustering appeared to be a balance between the aforementioned algorithms (the ratios of TP and TN were acceptable), and the parameter-measured performance of manual clustering was not drastically lower than that for M2D. The manual division of compounds can to some extent compete with automatic approaches; however, the time consumption and human factor impacting the final outcome provide disincentives to the wide use of manual clustering. Nevertheless, this splitting method provided structural information unavailable from different approaches, and moreover, reported the classification of entire chemical space of the 5-HT$_{1A}$R ligands stored in ChEMBL.

The approach requiring the selection of one ligand using at least two hypotheses appeared to be too strict. The results proved that different hypotheses primarily do not overlap each other, leading to an increased number of false negatives in the VS experiments and thus significantly reduced screening parameter values.

Conclusions

The results showed improved performance of the proposed method in virtual screening experiments. All investigated VS parameters outmatched both single hypothesis and random linear combination approaches. The experiments also proved that the automatic method of hierarchical clustering (based on the MOLPRINT 2D fingerprint) is a good option for screening. The computational cost of optimization increased, but the outcome compensated for that increase. Given the proposed method's success, it will be incorporated into our screening workflow [9] and applied for the next extended set of targets. Further improvement of the script interface will be undertaken, thus making it usable for other research groups.



Arylpiperazines with sulfona(i)mide fragment AHPRR

	H	P	R1	R2
A	14.52	6.86	12.25	3.94
H		7.92	3.28	13.84
P			5.50	6.63
R1				11.95

Arylpiperazines with terminal amide fragment AHPRR

	H	P	R1	R2
A	9.17	4.43	3.57	9.07
H		5.35	11.79	3.13
P			6.62	5.28
R1				11.75

Benzylpiperidines APRR

	P	R1	R2
A	9.92	15.04	2.80
P		6.45	7.71
R1			13.38

Methylenoaminochromanes HPRR

	P	R1	R2
H	8.60	12.32	3.16
P		3.74	7.37
R1			10.85

Aminotetralines HHPR

	H2	P	R
H1	6.94	4.19	3.18
H2		3.87	8.30
P			5.19

Tetrahydropirydonoindoles DPRR

	P	R1	R2
D	7.49	2.17	8.70
P		5.43	7.78
R1			7.20

Arylamines with four atom linker AHPR

	H	P	R
A	8.27	4.90	2.80
H		4.01	9.21
P			6.07

Figure 9. The best linear combination of pharmacophore models obtained for manual clustering and MCC optimization. The best linear combination of pharmacophore models obtained for manual clustering and MCC optimization (manual/random/hit-once; see also Figures 4 and 6). For each hypothesis the best fitting compound is presented, along with a matrix of distances (in angstroms) between features and a name of cluster it was developed on. The feature abbreviations used are: hydrogen bond acceptor – **A**, hydrogen bond donor – **D**, hydrophobic group – **H**, positively charged group – **P**, aromatic ring – **R**.

doi: 10.1371/journal.pone.0084510.g009

Table 1. A comparison between the optimized parameter values and those obtained for randomly selected combinations consisting of the same number of single hypotheses as optimized combinations.

Parameter	Optimized	Random	SD	Gain[a]
MCC	0.686	0.504	0.086	36.10%
Accuracy	0.840	0.726	0.044	15.66%
Recall	0.920	0.773	0.047	19.02%

[a] Percent increase of the value of the optimized statistical parameters compared with random combinations

The random values from ten different random linear combinations are averaged.

doi: 10.1371/journal.pone.0084510.t001

Table 2. A comparison between a single hypothesis and a linear combination of pharmacophore models.

Clustering approach	Selection method	Hypothesis composition	Actives		Decoys		MCC		Accuracy		Recall	
			TP	FN	TN	FP	universal	optimized	universal	optimized	universal	optimized
P3D	centroid	APRR	459	964	250	36	0.162	0.158	0.415	0.474	0.323	0.424
	random	AAPR	456	967	241	45	0.134	0.158	0.408	0.474	0.32	0.424
M2D	centroid	AHPR	639	784	227	59	0.184	0.294	0.507	0.71	0.449	0.743
	random	APRR	442	981	248	38	0.148	0.294	0.404	0.71	0.311	0.743
manual	centroid	APRR	399	1024	251	35	0.136	0.227	0.38	0.548	0.28	0.665
	random	APRR	390	1033	252	34	0.134	0.227	0.376	0.548	0.274	0.665

doi: 10.1371/journal.pone.0084510.t002

Supporting Information

Figure S1. The optimized values of accuracy for each possible scheme. Length of combination is shown on top of the bars.
(TIF)

Figure S2. The optimized values of recall for each possible scheme. Length of combination is shown on top of the bars.
(TIF)

Figure S3. The accuracy results of the validation of top linear combinations. Length of combination is shown on top of the bars.
(TIF)

Figure S4. The recall results of the validation of top linear combinations. Length of combination is shown on top of the bars.
(TIF)

Figure S5. A composition of each top ranked linear combination, obtained for P3D clustering procedure. The length row contains the total number of hypotheses forming a respective top ranked combination. Values of optimized statistical parameters for manual clustering are shown in 4 and for accuracy and recall in Figures S1 and S2, respectively.
(TIF)

Figure S6. A composition of each top ranked linear combination, obtained for M2D clustering procedure. The length row contains the total number of hypotheses forming a respective top ranked combination. Values of optimized statistical parameters for manual clustering are shown in Figure 4 and for accuracy and recall in Figures S1 and S2, respectively.
(TIF)

Figure S7. Pseudocode of in-house script (about 300 lines) used for the search for the best linear combination.
(PDF)

Figure S8. Venn diagrams containing numbers of elements common for the same optimizing schema (clustering approach/actives test set/hit mode) for different screening parameters. Number in overlapping area indicates elements present in two or more linear combinations. Number of hypotheses not used in any of the top-ranked combinations is beyond the circles. Numbers in circles sum up to the size of the given top combination (Figures 5, S5 and S6).
(TIF)

Author Contributions

Conceived and designed the experiments: DW AJB. Performed the experiments: DW KK AJB. Analyzed the data: DW SM AJB. Contributed reagents/materials/analysis tools: SM RK. Wrote the manuscript: DW SM ZC IS AJB.

References

1. Güner OF (2002) History and evolution of the pharmacophore concept in computer-aided drug design. Curr Top Med Chem 2: 1321–1332. doi:10.2174/1568026023392940. PubMed: 12470283.
2. Van Drie JH (2003) Pharmacophore discovery--lessons learned. Curr Pharm Des 9: 1649–1664. doi:10.2174/1381612033454568. PubMed: 12871063.
3. IUPAC Glossary of Terms used in Medicinal Chemistry. Available: http:// www.chem.qmul.ac.uk/iupac/medchem. (Accessed March 27, 2012)
4. Ren J-X, Li L-L, Zheng R-L, Xie H-Z, Cao Z-X et al. (2011) Discovery of novel Pim-1 kinase inhibitors by a hierarchical multistage virtual screening approach based on SVM model, pharmacophore, and molecular docking. J Chem Inf Model 51: 1364–1375. doi:10.1021/ci100464b. PubMed: 21618971.
5. Brunskole Svegelj M, Turk S, Brus B, Lanisnik Rizner T, Stojan J et al. (2011) Novel inhibitors of trihydroxynaphthalene reductase with antifungal activity identified by ligand-based and structure-based virtual screening. J Chem Inf Model 51: 1716–1724. doi:10.1021/ci2001499. PubMed: 21667970.
6. Svensson F, Karlén A, Sköld C (2012) Virtual screening data fusion using both structure- and ligand-based methods. J Chem Inf Model 52: 225–232. doi:10.1021/ci2004835. PubMed: 22148635.
7. Chiu T-L, Amin E a (2012) Development of a Comprehensive, Validated Pharmacophore Hypothesis for Anthrax Toxin Lethal Factor (LF) Inhibitors Using Genetic Algorithms, Pareto Scoring, and Structural Biology. J Chem Inf Model 52: 1886-1897. doi:10.1021/ci300121p. PubMed: 22697455.
8. Manepalli S, Geffert LM, Surratt CK, Madura JD (2011) Discovery of novel selective serotonin reuptake inhibitors through development of a protein-based pharmacophore. J Chem Inf Model 51: 2417–2426. doi: 10.1021/ci200280m. PubMed: 21834587.
9. Kurczab R, Nowak M, Chilmonczyk Z, Sylte I, Bojarski AJ (2010) The development and validation of a novel virtual screening cascade protocol to identify potential serotonin 5-HT(7)R antagonists. Bioorg Med Chem Lett 20: 2465–2468
10. Zajdel P, Kurczab R, Grychowska K, Satała G, Pawłowski M et al. (2012) The multiobjective based design, synthesis and evaluation of the arylsulfonamide/amide derivatives of aryloxyethyl- and arylthioethyl-piperidines and pyrrolidines as a novel class of potent 5-HT(7) receptor antagonists. Eur J Med Chem 56: 348–360. doi:10.1016/j.ejmech. 2012.07.043. PubMed: 22926225.
11. Durdagi S, Duff HJ, Noskov SY (2011) Combined receptor and ligand-based approach to the universal pharmacophore model development for studies of drug blockade to the hERG1 pore domain. J Chem Inf Model 51: 463–474. doi:10.1021/ci100409y. PubMed: 21241063.
12. Sanders MPa, Verhoeven S, de Graaf C, Roumen L, Vroling B et al. (2011) Snooker: a structure-based pharmacophore generation tool applied to class A GPCRs. J Chem Inf Model 51: 2277–2292. doi: 10.1021/ci200088d. PubMed: 21866955.
13. Hibert MF, Gittos MW, Middlemiss DN, Mir a K, Fozard JR (1988) Graphics computer-aided receptor mapping as a predictive tool for drug design: development of potent, selective, and stereospecific ligands for the 5-HT1A receptor. J Med Chem 31: 1087–1093
14. Mellin C, Vallgårda J, Nelson DL, Björk L, Yu H et al. (1991) A 3-D model for 5-HT1A-receptor agonists based on stereoselective methyl-substituted and conformationally restricted analogues of 8-hydroxy-2-(dipropylamino)tetralin. J Med Chem 34: 497–510. doi:10.1021/jm00106a004. PubMed: 1995871.

15. Agarwal A, Pearson PP, Taylor EW, Li HB, Dahlgren T et al. (1993) Three-dimensional quantitative structure-activity relationships of 5-HT receptor binding data for tetrahydropyridinylindole derivatives: a comparison of the Hansch and CoMFA methods. J Med Chem 36: 4006–4014. doi:10.1021/jm00077a003. PubMed: 8258822.
16. Orús L, Pérez-Silanes S, Oficialdegui A-M, Martínez-Esparza J, Del Castillo J-C et al. (2002) Synthesis and molecular modeling of new 1-aryl-3-[4-arylpiperazin-1-yl]-1-propane derivatives with high affinity at the serotonin transporter and at 5-HT(1A) receptors. J Med Chem 45: 4128–4139. doi:10.1021/jm0111200. PubMed: 12213056.
17. Sleight AJ, Peroutka SJ (1991) Identification of 5-hydroxytryptamine1A receptor agents using a composite pharmacophore analysis and chemical database screening. N-S arch pharmacol 343: 109–116
18. Chidester CG, Lin CH, Lahti RA, Haadsma-Svensson SR, Smith MW (1993) Comparison of 5-HT1A and dopamine D2 pharmacophores. X-ray structures and affinities of conformationally constrained ligands. J Med Chem 36: 1301–1315. doi:10.1021/jm00062a001. PubMed: 8496900.
19. Mokrosz MJ, Duszynska B, Bojarski AJ, Mokrosz JL (1995) Structure-activity relationship studies of CNS agents--XVII. Spiro[piperidine-4’, 1-(1,2,3,4-tetrahydro-beta-carboline)] as a probe defining the extended topographic model of 5-HT1A receptors. Bioorgan Med Chem 3: 533–538.
20. Langlois M, Brémont B, Rousselle D, Gaudy F (1993) Structural analysis by the comparative molecular field analysis method of the affinity of beta-adrenoreceptor blocking agents for 5-HT1A and 5-HT1B receptors. Eur J Pharmacol 244: 77–87. doi: 10.1016/0922-4106(93)90061-D. PubMed: 8093601.
21. Van Steen BJ, van Wijngaarden I, Tulp MT, Soudijn W (1994) Structure-affinity relationship studies on 5-HT1A receptor ligands. 2. Heterobicyclic phenylpiperazines with N4-aralkyl substituents. J Med Chem 37: 2761–2773. doi:10.1021/jm00043a015. PubMed: 8064803.
22. Bojarski AJ (2006) Pharmacophore models for metabotropic 5-HT receptor ligands. Curr Top Med Chem 6: 2005–2026. doi: 10.2174/156802606778522186. PubMed: 17017971.
23. Franchini S, Prandi A, Sorbi C, Tait A, Baraldi A et al. (2010) Discovery of a new series of 5-HT1A receptor agonists. Bioorg Med Chem Lett 20: 2017–2020. doi:10.1016/j.bmcl.2010.01.030. PubMed: 20185311.
24. Lepailleur A, Bureau R, Paillet-Loilier M, Fabis F, Saettel N et al. (2005) Molecular modeling studies focused on 5-HT7 versus 5-HT1A selectivity. Discovery of novel phenylpyrrole derivatives with high affinity for 5-HT7 receptors. J Chem Inf Model 45: 1075–1081. doi: 10.1021/ci050045p. PubMed: 16045303.
25. Chilmonczyk Z, Szelejewska-Wozniakowska A, Cybulski J, Cybulski M, Koziol AE et al. (1997) Conformational flexibility of serotonin1A receptor ligands from crystallographic data. Updated model of the receptor pharmacophore. Arch Pharm 330: 146–160. doi:10.1002/ardp.19973300507.
26. Weber KC, Salum LB, Honório KM, Andricopulo AD, da Silva ABF (2010) Pharmacophore-based 3D QSAR studies on a series of high affinity 5-HT1A receptor ligands. Eur J Med Chem 45: 1508–1514. doi: 10.1016/j.ejmech.2009.12.059. PubMed: 20133028.
27. Sanders MP a, Barbosa AJM, Zarzycka B, Nicolaes G a F, Klomp JPG, et al. (2012) Comparative analysis of pharmacophore screening tools. J Chem Inf Model 52: 1607–1620
28. Hoyer D, Hannon JP, Martin GR (2002) Molecular, pharmacological and functional diversity of 5-HT receptors. Pharmacol Biochem Behav 71: 533–554. doi:10.1016/S0091-3057(01)00746-8. PubMed: 11888546.
29. Lanfumey L, Hamon M (2004) 5-HT 1 Receptors. Current Drug Targets - CNS Neurol Disord: 1–10.
30. Paluchowska MH, Mokrosz MJ, Bojarski A, Wesołowska A, Borycz J et al. (1999) On the bioactive conformation of NAN-190 (1) and MP3022 (2), 5-HT(1A) receptor antagonists. J Med Chem 42: 4952–4960. doi: 10.1021/jm991045h. PubMed: 10585205.
31. Bojarski AJ, Paluchowska MH, Duszyńska B, Kłodzińska A, Tatarczyńska E et al. (2005) 1-Aryl-4-(4-succinimidobutyl)piperazines and their conformationally constrained analogues: synthesis, binding to serotonin (5-HT1A, 5-HT2A, 5-HT7), alpha1-adrenergic, and dopaminergic D2 receptors, and in vivo 5-HT1A functional characteristics. Bioorg Med Chem 13: 2293–2303. doi:10.1016/j.bmc.2004.12.041. PubMed: 15727878.
32. Paluchowska MH, Bojarski AJ, Charakchieva-Minol S, Wesołowska A (2002) Active conformation of some arylpiperazine postsynaptic 5-HT(1A) receptor antagonists. Eur J Med Chem 37: 273–283. doi: 10.1016/S0223-5234(01)01312-5. PubMed: 11960662.
33. Nowak M, Kołaczkowski M, Pawłowski M, Bojarski AJ (2006) Homology modeling of the serotonin 5-HT1A receptor using automated docking of bioactive compounds with defined geometry. J Med Chem 49: 205–214. doi:10.1021/jm050826h. PubMed: 16392805.
34. Ngo T, Nicholas TJ, Chen J, Finch AM, Griffith R (2013) 5-HT1A receptor pharmacophores to screen for off-target activity of α1-adrenoceptor antagonists. J Comput Aided Mol Des 27: 305–319. doi: 10.1007/s10822-013-9647-5. PubMed: 23625023.
35. ChEMBL_05, ChEMBL-EBI. http://www.ebi.ac.uk/chembldb/ index.php (accessed August 30 , 2010)
36. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S et al. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36: D901–D906. PubMed: 18048412.
37. ChEMBL_10, ChEMBL-EBI. http://www.ebi.ac.uk/chembldb/ index.php (accessed May 28, 2011) .
38. Sastry M, Lowrie JF, Dixon SL, Sherman W (2010) Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. J Chem Inf Model 50: 771–784. doi:10.1021/ci100062n. PubMed: 20450209.
39. Canvas, version 1.4, Schrödinger, LLC, New York, NY, 2011
40. Kelly J (1956) A new interpretation of information rate. IEEE T INFORM Theory 2: 185–189. doi:10.1109/TIT.1956.1056803.
41. Caliendo G, Santagada V, Perissutti E, Fiorino F (2005) Derivatives as 5HT1A receptor ligands. Past and Present - Curr Med Chem 12: 1721–1753.
42. Olivier B, Soudijn W, van Wijngaarden I (1999) The 5-HT1A receptor and its ligands: structure and function. Prog Drug Res 52: 103–165. PubMed: 10396127.
43. Oh SJ, Ha HJ, Chi DY, Lee HK (2001) Serotonin receptor and transporter ligands - current status. Curr Med Chem 8: 999–1034. doi: 10.2174/0929867013372599. PubMed: 11472239.
44. López-Rodríguez ML, Ayala D, Benhamú B, Morcillo MJ, Viso a (2002) Arylpiperazine derivatives acting at 5-HT(1A) receptors. Curr Med Chem 9: 443–469. doi:10.2174/0929867023371030. PubMed: 11945120.
45. Phase, version 3.3, Schrödinger, LLC, New York, NY, 2011
46. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007) Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25: 197-206. doi:10.1038/nbt1284. PubMed: 17287757.