JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Contribution of 2D and 3D Structural Features of Drug Molecules in the Prediction of Drug Profile Matching

Ágnes Peragovics,[†,‡] Zoltán Simon,[‡] Ildikó Brandhuber,[‡] Balázs Jelinek,[†,‡] Péter Hári,[‡] Csaba Hetényi,[$] Pál Czobor,[§] and András Málnási-Csizmadia[†,‡,$,*]

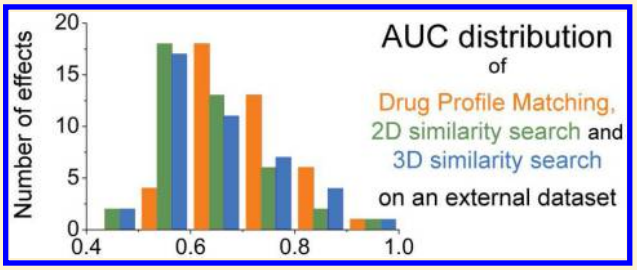
[†]Department of Biochemistry, Institute of Biology, Eötvös Loránd University, Pázmány Péter sétány 1/C, H-1117 Budapest, Hungary

[$]HAS-ELTE Molecular Biophysics Research Group, Pázmány Péter sétány. 1/C, H-1117 Budapest, Hungary

[‡]Drugmotif Ltd., Szent Erzsébet krt. 14, H-2112 Veresegyház, Hungary

[§]Department of Psychiatry and Psychotherapy, Semmelweis University, Balassa utca 6, H-1083 Budapest, Hungary

ABSTRACT: Drug Profile Matching (DPM), a novel virtual affinity fingerprinting method capable of predicting the medical effect profiles of small molecules, was introduced by our group recently. The method exploits the information content of interaction patterns generated by flexible docking to a series of rigidly kept nontarget protein active sites. We presented the ability of DPM to classify molecules excellently, and the question arose, what the contribution of 2D and 3D structural features of the small molecules is to the intriguingly high prediction power of DPM. The present study compared the prediction powers for effect profiles of 1163 FDA-approved drug compounds determined by DPM and ChemAxon 2D and 3D similarity fingerprinting approaches. We found that DPM outperformed the 2D and 3D approaches in almost all therapeutic categories for drug classification except for mechanically rigid structural categories where high accuracy was obtained by all three methods. Moreover, we also tested the predictive power of DPM on external data by reducing the parent data set and demonstrated that DPM can overcome the common screening problems of 2D and 3D similarity methods arising from the presence of structurally diverse molecules in certain effect categories.



INTRODUCTION

Virtual drug design is a key method of pharmaceutical research.[1] From a structural point of view, two main approaches can be followed: target-based design utilizing the 3D structure of the target protein and ligand-based design capitalizing on the concept of molecular similarity,[2] exploiting the information contained in the known active ligands of the investigated effect. This information is often stored in molecular descriptors derived from the molecular structure.

The most simple, nonetheless effective ligand-based approach is the similarity search using 2D fingerprints, i.e., binary strings displaying the presence or absence of molecular fragments. While similarity between two molecules can be characterized by various similarity coefficients,[3] comparative studies have demonstrated that the well-known Tanimoto coefficient remains the most reliable coefficient of choice.[4]

To take into account the spatial features of molecules, 3D fingerprints have been developed and are also widely used as a basis of virtual screening protocols. However, several studies suggest that 2D descriptors are often superior to 3D ones.[5,6] This surprising result is often explained by the nature of the chemical data investigated in drug development processes. In a typical medicinal chemistry program, many compounds are synthesized having the same scaffolds, and only small structural changes are introduced, e.g., with substitution reactions. These minor structural differences favor 2D similarity methods; however, 3D methods can be significantly affected by such small changes.[7] Other issues that raise difficulties concerning 3D comparisons are the identification and alignment of the biologically relevant conformations of the active molecules prior to calculating similarity.

Starting from the 1990s, "affinity fingerprints" have been introduced that are a series of binding affinities or docking scores against a reference panel of proteins that do not include the target protein. The in vitro version of this idea was developed by Kauvar et al. and was used to predict the binding properties of the applied compounds to other proteins not included in the reference panel.[8,9] Bender et al. developed the "Bayes Affinity Fingerprint" similarity search method, in which compound similarity is measured by similarities of experimental binding affinity values against a set of pharmacological target proteins and demonstrated its superiority over a traditional structural similarity search.[10,11] In many studies, the experimental assays were replaced by computational dockings of ligands to the binding pockets of the reference proteins.[12,13] In silico affinity fingerprints proved to be effective in the design of focused libraries,[14,15] and their screening performance was

Received: February 24, 2012
Published: June 14, 2012

   dx.doi.org/10.1021/ci3001056 | J. Chem. Inf. Model. 2012, 52, 1733−1744

comparable to standard 2D methods.[7] Moreover, it was also demonstrated that the hit lists produced by similarity searches on the basis of virtual affinity fingerprints are complementary to those obtained by 2D methods. Because the direct structural information of the active molecules is not exploited with the affinity fingerprints, biological similarity is more likely to be detected beyond structural similarity.

Recently, we reported on Drug Profile Matching (DPM), a novel approach capable of predicting the complete medical effect profiles of small-molecule compounds on the basis of in silico affinity fingerprints.[16−18] It relates complex drug−protein interaction patterns (IPs) with effect profiles (EPs) of drug molecules by multidimensional statistical techniques. This method can be suitable for virtual design as it calculates probability for each drug, showing the likelihood of possessing a given effect. The list of probabilities can be ordered, effect by effect, by decreasing values to get a priority list analogous to that produced by standard 2D and 3D methods (so-called nearest neighbor list).

The present study compares the performance of Drug Profile Matching with conventional 2D and 3D similarity searches for drug classification and also investigates the predictivity of DPM on external data by reducing the parent data set. ChemAxon tools were used to create the 2D and 3D similarity matrices of the studied drug molecules and classification was performed by generating similarity lists. Receiver Operating Characteristic (ROC) analyses were performed on these similarity lists, and the accuracy was determined by calculating the AUC values. DPM predictions were subjected to a rigorous cross-validation procedure in order to prove the reliability of the results.

We found that the DPM method performed better than 2D and 3D approaches for the majority of the effect categories for drug classification, except for mechanically rigid structural categories that resulted in high accuracy with all three methods. Possible explanation for this phenomenon is provided. The predictive power of DPM was also investigated outside the parent data set, and our results suggest that the method outperforms the conventional similarity-based approaches for structurally diverse effect categories.

METHODS

Drug Profile Matching Method. A detailed description of the DPM method was presented in our recent publication.[16] Shortly, 1177 FDA-approved drug molecules extracted from the DrugBank database were docked to 149 nontarget proteins from the RCSB Protein Data Bank (PDB) using DOVIS 2.0 software (DOcking-based VIrtual Screening),[19] AutoDock4 docking engine,[20] Lamarckian genetic algorithm, and X-SCORE scoring function.[21] The docking box was centered to the geometrical center of the original ligand of the protein; box size and grid spacing were set to 22.5 Å and 0.375 Å, respectively. Twenty-five docking runs were performed for each job. The best docking scores for each drug−protein complex were used to create the Interaction Pattern (IP) matrix. Then, a binary matrix called Effect Profile (EP) matrix was formed that displays the presence or absence of the studied 177 effect categories for each drug. Classification of molecules was based on the categories listed in DrugBank and was revised manually. Every effect class contained at least 10 registered drugs to ensure meaningful classification.

Canonical correlation analysis (CCA) was performed between the IP matrix and each effect to generate a highly correlating factor pair. This factor pair served as an input for linear discriminant analysis (LDA) that produced classification functions yielding the probability for each drug-effect pair. These probabilities comprise the Effect Probability matrix.

A major difference from our previous work[16] is that a new hierarchical, two-level effect database was introduced in this study that consists of 62 main therapeutic classes. These main effect groups are further divided into several subgroups on the basis of a common target, mechanism of action, or molecular structure, which also served as a basis of DPM predictions and similarity-based searching. This new effect database structure is in better accordance with clinical practice and therefore gives a more realistic classification of drug molecules. The introduction of second level classification enables us to study the impact of reducing large medical effect classes to smaller but more cohesive subgroups that can improve classification accuracy.

In this work, the studied drug set was reduced to 1163 molecules because 3D alignment could not be performed for 14 molecules. DPM predictions resulting in probabilities for each drug-effect pair were recalculated on this drug set using the new hierarchical effect database. Because only those effects were investigated that contain at least 10 registered drugs, the number of studied therapeutic classes was reduced to 52. A total of 70 corresponding subgroups were used for further analyses, which also comprise at least 10 drugs.

The average Tanimoto distance (referred to as Tanimoto diversity) was calculated for each of the studied effect groups to quantify the structural diversity of the registered molecules.

Drug classification was carried out on the basis of the calculated probabilities for each effect via ordering the molecules by decreasing probability. The accuracy of classification was measured by Receiver Operating Characteristic analysis, i.e., plotting the true positive rate (TPR) as a function of the false positive rate (FPR), using a sliding cutoff parameter from 1 to 0 for the probability list. Molecules are classified "positive" if they possess a greater probability for an effect than the actual cutoff value, while they are considered "negative" in the opposite case. TPR is the rate of true positives among the positives, i.e., the portion of positives that were classified correctly, and it is also called sensitivity. FPR is the rate of false positives among the negatives, i.e., a fraction of negatives that were wrongly classified as positive, and is often referred to as 1-specificity. Perfect classification of molecules is obtained when the curve runs through the (0,1) point, i.e., all of the molecules classified as positive are true positives, and no false positives are detected. The Area Under the ROC curve (AUC) was used to evaluate the performance: the closer its value to 1 the better the classification. A random classification would result in a diagonal ROC curve (AUC of 0.5). ROC analysis is a recognized performance criteria in virtual screening;[22,23] however, it has been criticized that AUC values are not sensitive to "early recognition", a specific problem to virtual screening.[24] Active molecules must be ranked early in an ordered list by virtual screening methods because only a limited number of compounds can be tested. Actives appearing at the top of the list contribute with higher score to the Boltzmann enhanced discrimination of ROC (BEDROC) metric, making it suitable for studying early recognition. However, in our recent work,[16] we demonstrated that the DPM method can overcome the "early recognition" problem; therefore, only AUC values were calculated in the current study.

In order to evaluate the robustness of the results and control for overfitting, 10-fold cross-validation was performed. We have partitioned the data into 10 complementary sets (so-called

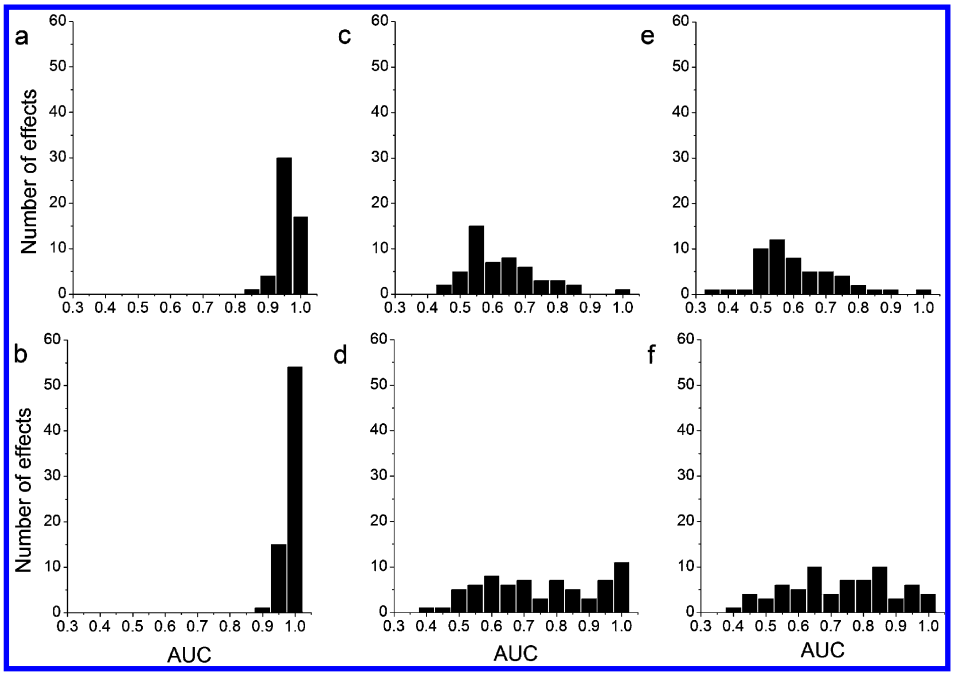


Figure 1. Distribution of AUC values obtained for drug classification. The histograms show the distribution of the AUC values for the studied 52 main effects (first row) and 70 subgroups (second row) obtained by DPM, 2D similarity-based search and 3D similarity-based search. (a) DPM for main effects, (b) DPM for subgroups, (c) 2D similarity search for main effects, (d) 2D similarity search for subgroups, (e) 3D similarity search for main effects, and (f) 3D similarity search for subgroups.

"folds"), and each fold was retained as a test set for validation, while the remaining folds were merged to produce the training set. In each round of the validation, the IP-based classification function was created on the basis of the training set, and effect probabilities were predicted for the test set. This process was performed for each of the 10 folds, and the probability values for each of the originally registered drugs to a given effect were combined. This process was repeated 100 times for each effect.

In order to evaluate the predictivity of DPM on external data, the studied data set was randomly split into two sets of equal size. One part was set as a training set to derive the classification rules, while the other part was used as a test set to estimate class membership of new compounds. According to our previous experiences, DPM can efficiently classify molecules when at least 10 active compounds are provided to derive the IP-based classification function. Therefore, effects having 20 or more registered molecules were used in the following analysis to ensure that a sufficient amount of active molecules is distributed between the two sets. A total of 43 main effect groups and 27 subcategories could fulfill this criterion. For each studied effect, active and inactive molecules were randomly partitioned between the two sets with the only restriction of having 10 active molecules in the training set, while placing the remaining actives in the test set. This way, only the required minimum but adequate amount of information about the active compounds was provided for DPM in the training set. In each round of the analysis, the IP-based classification function was created on the basis of the training set, and effect probabilities were predicted for the test set. Classification accuracy of the test molecules was measured by the ROC analysis described above. The whole process starting from the random splitting of the data was repeated 100 times for each effect, and the AUC values calculated for the test set were averaged.

Similarity-Based Classification Using 2D Fingerprints. Values of 2D chemical fingerprints were generated using ChemAxon's JChem Base software that encodes topological properties of the chemical graph up to six bonds. As a result, a 4096-bit-long binary fingerprint was created for each drug. On the basis of these fingerprints, a similarity matrix was calculated using Tanimoto distance calculation by ChemAxon Similarity plugin.

Then, a molecule registered to a given effect was selected as a query structure, and its similarity list was subjected to ROC analysis, i.e., the determination whether a neighboring molecule belongs to the same effect or not. The accuracy was determined by calculating the AUC value described in the previous section. This procedure was repeated for each drug belonging to the studied effect, and the average AUC value and the standard deviation were calculated. This way, accuracy values were generated for each effect.

In order to produce comparable results to that of the DPM random splitting experiments, the same test sets generated for DPM were subjected to similarity searching using the corresponding 10 active molecules belonging to the former training sets. Each active molecule was set as a reference structure, and the test molecules were sorted by decreasing similarity. AUC values were calculated for each similarity list, and the obtained 10 values were averaged. The inactive molecules of the training set were not considered because conventional similarity searches utilize only the information stored in the active molecules. For each effect, all 100 random splitting sets were used, and the mean of the AUC values were calculated.

Similarity-Based Classification Using 3D Fingerprints. A 3D similarity matrix was created using ChemAxon's Flexible 3D Alignment tool that aims at maximizing the overlap of the same atom types for two molecules. Alignment is carried out by optimizing a potential function (E_{pot}) involving intra- and

dx.doi.org/10.1021/ci3001056 | J. Chem. Inf. Model. 2012, 52, 1733−1744

Table 1. Results Based on Drug Profile Matching, 2D Similarity Search, and 3D Similarity Search[a]

			Drug Profile Matching								
				10-fold cross-validation		random splitting		2D similarity search		3D similarity search	
effect	*n*	Tanimoto diversity	AUC	mean	std	AUC	std	AUC	std	AUC	std
Adrenal gland hormone	27	0.50	0.98	0.70	0.02	0.87	0.04	0.78	0.24	0.81	0.21
Glucocorticoid*	33	0.65	1.00	0.91	0.02	0.98	0.01	0.99	0.01	1.00[†]	0.00
Anesthetic agent, general	19	0.32	0.98	0.30	0.05			0.54	0.06	0.53	0.02
Anesthetic agent, intravenous	17	0.38	0.99	0.37	0.05			0.62	0.06	0.59	0.02
Anesthetic agent, local	17	0.43	0.99	0.45	0.03			0.73	0.09	0.72	0.08
Anorexigenic agent	14	0.44	0.98	0.44	0.03			0.72	0.21	0.62	0.11
Central anorexigenic agent	12	0.51	0.99	0.49	0.04			0.85[†]	0.17	0.69	0.10
Anti-HIV agent	21	0.47	0.99	0.44	0.04	0.76	0.07	0.78[†]	0.07	0.62	0.06
Antiglaucoma agent	25	0.31	0.96	0.43	0.04	0.72	0.06	0.49	0.04	0.51	0.04
Alpha2 receptor agonist*	10	0.36	0.97	0.39	0.08			0.63	0.10	0.61	0.12
Beta receptor antagonist*	11	0.56	1.00	0.54	0.05			0.94[†]	0.05	0.85	0.09
Glucocorticoid*	33	0.65	1.00	0.91	0.02	0.98	0.01	0.99	0.01	1.00[†]	0.00
Anti-inflammatory agent	84	0.35	0.93	0.63	0.01	0.70	0.06	0.52	0.11	0.50	0.03
Antigout agent	70	0.37	0.96	0.68	0.01	0.78	0.05	0.55	0.13	0.52	0.06
Antirheumatoid arthritis agent	13	0.32	0.95	0.00	0.00			0.52[†]	0.04	0.46	0.04
Glucocorticoid*	33	0.65	1.00	0.91	0.02	0.98	0.01	0.99	0.01	1.00[†]	0.00
NSAID, nonselective COX inhibitor	35	0.37	0.98	0.60	0.02	0.81	0.05	0.63	0.10	0.74[†]	0.06
Antianginal agent	59	0.38	0.93	0.50	0.02	0.65	0.04	0.60[†]	0.04	0.58	0.04
ACE inhibitor*	13	0.62	1.00	0.50	0.05			0.94[†]	0.07	0.86	0.07
Beta1 receptor antagonist*	10	0.64	1.00	0.53	0.06			0.99[†]	0.01	0.86	0.11
Antiarrhythmic agent	30	0.40	0.94	0.31	0.03	0.58	0.05	0.66	0.09	0.71[†]	0.06
Antiasthmatic agent	40	0.37	0.94	0.39	0.02	0.60	0.05	0.54	0.04	0.56[†]	0.03
Beta2 receptor agonist*	16	0.65	1.00	0.50	0.04			0.98[†]	0.01	0.78	0.10
Glucocorticoid*	33	0.65	1.00	0.91	0.02	0.98	0.01	0.99	0.01	1.00[†]	0.00
Muscarinic antagonist*	29	0.49	0.96	0.68	0.01	0.79	0.04	0.82	0.10	0.84	0.05
Antibiotic agent	134	0.45	0.94	0.72	0.01	0.73	0.04	0.72[†]	0.09	0.50	0.08
Antibiotic, quinolone derivative	15	0.72	1.00	0.78	0.03			0.99	0.01	0.97	0.05
Antibiotic, sulfonamide	13	0.52	1.00	0.84	0.02			0.96	0.03	0.93	0.07
Antibiotic, tetracycline	10	0.89	1.00	0.76	0.07			1.00	0.00	0.97	0.04
Beta-lactam antibiotic, cephalosporin	32	0.78	1.00	0.87	0.01	0.96	0.02	1.00[†]	0.00	0.90	0.08
Beta-lactam antibiotic, penicillin	19	0.79	1.00	0.79	0.03			1.00[†]	0.00	0.91	0.05
Antibiotic agent, NOS	12	0.32	0.99	0.14	0.04			0.54	0.05	0.43	0.06
Anticytokine agent	33	0.65	1.00	0.91	0.01	0.98	0.01	0.99	0.01	1.00[†]	0.00
Antidepressant and antimanic agent	36	0.36	0.97	0.42	0.03	0.65	0.05	0.63	0.13	0.70[†]	0.05
Antiemetic agent	34	0.39	0.97	0.38	0.03	0.63	0.05	0.57	0.07	0.73[†]	0.05
Dopamine D2 receptor antagonist	11	0.40	1.00	0.26	0.06			0.68	0.11	0.84	0.03
Histamine H1 receptor antagonist*	40	0.49	0.98	0.64	0.01	0.82	0.03	0.68	0.11	0.86[†]	0.07
Muscarinic antagonist*	29	0.45	0.96	0.68	0.01	0.79	0.04	0.82	0.10	0.84	0.05
Serotonin receptor antagonist*	12	0.35	0.98	0.37	0.06			0.76	0.04	0.76	0.05
Antiemetic agent, NOS	10	0.43	0.99	0.36	0.05			0.59	0.03	0.60	0.09
Antiepileptic agent	45	0.45	0.97	0.54	0.02	0.72	0.05	0.55[†]	0.07	0.50	0.11
GABA agent	24	0.38	1.00	0.60	0.01	0.86	0.05	0.69	0.09	0.66	0.15
Antifungal agent	23	0.38	0.97	0.33	0.03	0.65	0.07	0.61	0.09	0.68[†]	0.12
Antihistamine	44	0.40	0.96	0.59	0.01	0.74	0.04	0.63	0.09	0.77[†]	0.08
Histamine H1 receptor antagonist*	40	0.43	0.98	0.64	0.01	0.82	0.03	0.68	0.11	0.86[†]	0.07
Antihyperlipidemic agent	30	0.45	0.95	0.45	0.03	0.65	0.05	0.68[†]	0.06	0.63	0.06
Antihyperlipidemic agent, prevention	24	0.45	0.96	0.47	0.03	0.71	0.05	0.71[†]	0.07	0.64	0.09
Antihypertensive agent	103	0.37	0.93	0.55	0.01	0.62	0.04	0.59[†]	0.03	0.56	0.04
ACE inhibitor*	13	0.62	1.00	0.50	0.05			0.94[†]	0.07	0.86	0.07
Antihypertensive agent, diuretic	21	0.47	0.99	0.38	0.02	0.80	0.05	0.80[†]	0.10	0.67	0.14
Sodium depletion diuretic	11	0.45	1.00	0.18	0.01			0.79[†]	0.08	0.68	0.07
Sympathetic blocker	34	0.41	0.97	0.47	0.02	0.72	0.05	0.67[†]	0.11	0.61	0.05
Vasodilator*	10	0.23	0.98	0.00	0.00			0.41	0.14	0.45	0.14
Antineoplastic agent	99	0.32	0.88	0.49	0.01	0.60	0.05	0.50	0.05	0.50	0.02
Antineovascularisation agent	11	0.56	0.97	0.58	0.04			0.85[†]	0.04	0.78	0.06
Cell proliferation agent, hormone system associating agent	20	0.35	0.99	0.52	0.02	0.79	0.06	0.54	0.08	0.57	0.03

dx.doi.org/10.1021/ci3001056 | J. Chem. Inf. Model. 2012, 52, 1733−1744

Table 1. continued

effect	n	Tanimoto diversity	Drug Profile Matching					2D similarity search		3D similarity search	
				10-fold cross-validation		random splitting					
			AUC	mean	std	AUC	std	AUC	std	AUC	std
Cytotoxic agent, alkylating agent	15	0.24	1.00	0.25	0.05			0.45	0.16	0.39	0.10
Cytotoxic agent, topoisomerase inhibitor	10	0.63	0.99	0.77	0.05			0.88	0.08	0.94	0.03
Cytotoxic agent, antimetabolite	17	0.44	0.96	0.49	0.04			0.74†	0.14	0.53	0.12
Antineoplastic agent, NOS	10	0.32	0.99	0.03	0.02			0.52	0.07	0.57	0.08
Antiparasitic agent	14	0.26	0.97	0.02	0.02			0.45	0.09	0.48†	0.03
Antiprotozoal agent	24	0.34	0.97	0.17	0.03	0.55	0.06	0.55	0.07	0.56	0.06
Antimalarial agent	15	0.40	0.99	0.15	0.03			0.70	0.09	0.71	0.04
Antiprotozoal agent, NOS	12	0.30	0.98	0.04	0.02			0.48	0.06	0.46	0.05
Antipsychotic	40	0.47	0.98	0.69	0.01	0.88	0.03	0.78	0.12	0.88†	0.04
Standard antipsychotic, phenothiazine	19	0.69	1.00	0.94	0.02			1.00†	0.00	0.98	0.01
Antitussive and expectorant	23	0.45	0.98	0.27	0.02	0.64	0.06	0.76†	0.12	0.66	0.09
Expectorant, secretomotoric	14	0.57	0.99	0.33	0.03			0.93†	0.03	0.73	0.10
Antiviral agent	20	0.38	0.96	0.31	0.05	0.66	0.08	0.66†	0.13	0.57	0.09
Anxiolytic agent	46	0.40	0.94	0.66	0.02	0.82	0.04	0.59	0.08	0.60	0.11
Barbiturate*	15	0.67	1.00	0.98	0.02			0.99†	0.03	0.77	0.17
Benzodiazepine*	22	0.62	1.00	0.88	0.02	0.97	0.02	0.96	0.04	0.95	0.03
Beta receptor antagonist*	11	0.56	1.00	0.54	0.05			0.94†	0.05	0.85	0.09
Serotonin receptor agonist*	16	0.49	0.99	0.43	0.05			0.83	0.08	0.81	0.06
Calcium and bone metabolism agent	31	0.34	0.98	0.53	0.03	0.81	0.05	0.53†	0.08	0.48	0.07
Hormone and hormone-like synthetic agent	10	0.57	1.00	0.6	0.02			0.80	0.25	0.82	0.19
Calcium and bone metabolism agent, NOS	14	0.54	1.00	0.57	0.03			0.87	0.16	0.77	0.19
Carbohydrate metabolism agent	24	0.39	0.98	0.48	0.04	0.79	0.05	0.65†	0.08	0.57	0.06
Oral antidiabetic agent, sulfonylurea	11	0.60	1.00	0.62	0.10			0.97†	0.03	0.85	0.04
Cholinolytic	39	0.40	0.97	0.60	0.02	0.74	0.04	0.68	0.09	0.72†	0.06
Muscarinic antagonist*	29	0.49	0.96	0.68	0.01	0.79	0.04	0.82	0.10	0.84	0.05
Cholinomimetic	15	0.30	1.00	0.36	0.04			0.53	0.13	0.53	0.07
Indirect cholinomimetic (cholinesterase inhibitor)	10	0.34	1.00	0.23	0.06			0.63	0.15	0.64	0.09
Cutaneous disease agent	262	0.35	0.86	0.60	0.00	0.63	0.04	0.55†	0.05	0.51	0.03
Cutaneous disease agent, antibiotic agent	114	0.45	0.95	0.71	0.01	0.74	0.04	0.71†	0.08	0.51	0.07
Cutaneous disease agent, antifungal agent	23	0.38	0.97	0.33	0.03	0.65	0.06	0.61	0.09	0.68	0.12
Cutaneous disease agent, antihistamine	36	0.39	0.98	0.58	0.02	0.78	0.03	0.66	0.10	0.84†	0.08
Cutaneous disease agent, immunosuppressive agent	41	0.53	0.96	0.74	0.01	0.84	0.03	0.82	0.21	0.79	0.19
Anti-inflammatory agent, NSAID	36	0.37	0.98	0.60	0.02	0.82	0.04	0.64	0.09	0.74†	0.06
Anti-inflammatory agent, glucocorticoid	33	0.65	1.00	0.91	0.01	0.98	0.01	0.99	0.01	1.00†	0.00
Detoxification agent	11	0.20	0.99	0.09	0.05			0.43	0.12	0.34	0.06
Diuretic	37	0.40	0.97	0.40	0.02	0.69	0.06	0.65†	0.09	0.54	0.10
Minor diuretic	12	0.56	1.00	0.49	0.04			0.72	0.11	0.73	0.07
Thiazide or thiazide-like diuretic	13	0.61	1.00	0.60	0.02			0.98†	0.02	0.87	0.12
Gastrointestinal agent, NOS	12	0.24	0.95	0.13	0.03			0.51	0.15	0.42	0.08
Gastrointestinal ulcer agent	22	0.34	0.98	0.44	0.03	0.77	0.06	0.53	0.07	0.55	0.04
Heart failure agent	50	0.38	0.96	0.46	0.02	0.64	0.05	0.58†	0.03	0.55	0.04
ACE inhibitor*	13	0.62	1.00	0.50	0.05			0.94†	0.07	0.86	0.07
Beta receptor antagonist*	11	0.56	1.00	0.54	0.05			0.94†	0.05	0.85	0.09
Heart failure agent, diuretic	10	0.47	1.00	0.31	0.02			0.76	0.11	0.63	0.15
Vasodilator*	10	0.23	0.98	0.00	0.00			0.41	0.14	0.45	0.14
Immunosuppressive agent	40	0.55	0.97	0.76	0.01	0.86	0.03	0.84	0.20	0.81	0.20
Glucocorticoid*	33	0.65	1.00	0.91	0.02	0.98	0.01	0.99	0.01	1.00†	0.00
Inflammatory bowel disease agent	12	0.42	1.00	0.33	0.08			0.66†	0.17	0.52	0.07
Glucocorticoid*	33	0.65	1.00	0.91	0.02	0.98	0.01	0.99	0.01	1.00†	0.00
Neurodegenerative disease agent	93	0.35	0.88	0.49	0.01	0.58	0.03	0.56	0.06	0.64†	0.04
Alzheimer disease agent	15	0.33	0.94	0.15	0.03			0.56	0.06	0.54	0.06
Parkinson disease agent	29	0.37	0.96	0.30	0.04	0.59	0.05	0.62	0.10	0.65	0.05
Sclerosis multiplex agent	23	0.32	0.95	0.19	0.04	0.55	0.06	0.51	0.07	0.55†	0.05
Tardive dyskinesia agent	22	0.55	0.98	0.66	0.02	0.86	0.05	0.84	0.16	0.93†	0.05
Obstipant	23	0.41	0.95	0.45	0.04	0.64	0.06	0.68	0.09	0.73†	0.05

dx.doi.org/10.1021/ci3001056 | J. Chem. Inf. Model. 2012, 52, 1733−1744

Table 1. continued

			Drug Profile Matching								
				10-fold cross-validation		random splitting		2D similarity search		3D similarity search	
effect	*n*	Tanimoto diversity	AUC	mean	std	AUC	std	AUC	std	AUC	std
Opioid	29	0.55	0.97	0.57	0.01	0.88	0.04	0.85	0.07	0.84	0.09
Synthetic opioid	17	0.55	1.00	0.53	0.04			0.88†	0.08	0.85	0.05
Platelet aggregation inhibitor	10	0.37	0.98	0.01	0.01			0.60	0.04	0.56	0.04
Primer headache treatment	42	0.32	0.90	0.38	0.02	0.54	0.05	0.53	0.11	0.59†	0.05
Antimigraine agent, antiattack	12	0.34	1.00	0.19	0.05			0.54	0.06	0.64†	0.09
Antimigraine agent, prevention	27	0.34	0.95	0.43	0.03	0.65	0.05	0.61	0.13	0.65	0.06
Prokinetic agent	19	0.31	0.98	0.35	0.04			0.53	0.10	0.54	0.05
Prokinetic agent, cholinomimetic	15	0.30	1.00	0.36	0.04			0.53	0.13	0.53	0.07
Sedative and/or hypnotic agent	181	0.35	0.90	0.64	0.01	0.64	0.04	0.56	0.06	0.61†	0.04
Barbiturate*	15	0.67	1.00	0.98	0.02			0.99†	0.03	0.77	0.17
Benzodiazepine*	22	0.62	1.00	0.88	0.02	0.97	0.02	0.96	0.04	0.95	0.03
Sedative and/or hypnotic agent, NOS	161	0.36	0.90	0.61	0.01	0.60	0.04	0.58	0.06	0.64	0.05
Serotonin receptor agonist*	16	0.49	0.99	0.43	0.05			0.83	0.08	0.81	0.06
Serotonin agent	34	0.42	0.95	0.45	0.03	0.66	0.04	0.69	0.05	0.75†	0.04
Serotonin receptor agonist*	16	0.49	0.99	0.43	0.05			0.83	0.08	0.81	0.06
Serotonin receptor antagonist*	12	0.45	0.98	0.37	0.06			0.76	0.04	0.76	0.05
Sexual hormone and sexual activity agent	53	0.36	0.93	0.60	0.01	0.73	0.04	0.59	0.15	0.64	0.11
Progestin	11	0.59	1.00	0.62	0.06			0.94	0.12	0.98	0.01
Smooth muscle agent	26	0.34	0.96	0.48	0.03	0.70	0.05	0.56	0.10	0.61†	0.04
Spasmolytic	13	0.35	0.98	0.31	0.04			0.60	0.10	0.67	0.05
Striated muscle agent	32	0.30	0.97	0.35	0.02	0.65	0.06	0.48	0.04	0.45	0.07
Central striated muscle relaxant	26	0.32	0.98	0.38	0.03	0.71	0.06	0.50	0.05	0.49	0.10
Sympatholytic	52	0.42	0.95	0.50	0.02	0.66	0.04	0.67	0.09	0.67	0.06
Beta1 receptor antagonist*	10	0.64	1.00	0.53	0.06			0.99†	0.01	0.86	0.11
Nonselective beta receptor antagonist	11	0.54	1.00	0.80	0.02			0.90	0.09	0.89	0.04
Norepinephrine liberation blocker or stimulant	11	0.35	0.99	0.08	0.03			0.58	0.04	0.62	0.07
Sympathomimetic	61	0.44	0.95	0.58	0.01	0.75	0.04	0.76†	0.14	0.62	0.07
Alpha2 receptor agonist*	10	0.36	0.97	0.39	0.08			0.63	0.10	0.61	0.12
Beta2 receptor agonist*	16	0.65	1.00	0.50	0.04			0.98†	0.01	0.78	0.10
Norepinephrine reuptake inhibitor (cell)	16	0.44	0.99	0.43	0.04			0.78	0.12	0.79	0.05
Vitamin and analogue	22	0.34	0.99	0.46	0.04	0.77	0.05	0.55	0.21	0.52	0.16
Fat soluble vitamin	13	0.51	1.00	0.58	0.05			0.81	0.22	0.78	0.18

[a]Each main effect group is marked with bolded, and corresponding subgroups are listed under it. An asterisk (∗) indicates that a given subgroup belongs to more main effects. The number of active molecules (n) and their average Tanimoto distance is displayed for each effect. Because DPM is able to handle all active molecules in a single search, only a single AUC value can be calculated to describe its performance for drug classification. Overfitting is countered by an independent 10-fold cross-validation that is also presented (mean and standard deviation of MPV). The predictive power of DPM on external data was tested in random splitting experiments and was repeated 100 times for each effect. AUC values and standard deviation were calculated. Only effects having more than 20 registered molecules were used in this analysis. In the case of 2D and 3D similarity searches, every molecule registered to a given effect was used as a query structure. The individual AUC values were averaged, and standard deviations were calculated. Two sample *t* tests were applied to test the statistical significance of the difference between the structure-based results. A dagger (†) marks an AUC value, if it is significantly higher than the AUC value produced by the other similarity search at a 95% confidence level.

intermolecular geometrical constraints. The formula of the used E_{pot} is presented and described in detail at the following web link of ChemAxon (http://www.chemaxon.com/library/flexible-3d-alignment-and-its-application-in-virtual-screening/). The applied E_{pot} was optimized during alignment in several steps. For each molecule, three atoms maximally separated in the molecular graph were selected. Rigid, quaternion-based alignment of the two molecules was performed between the two atom pairs. Four starting conformations of randomly selected rotatable dihedrals were used for each paired alignment. In the next step, optimization was done with fixed dihedrals, and then the internal rotations were released. Then for each pair of atoms, the Gaussian constraints were replaced by simpler quadratic terms to find an optimal orientation for the local neighborhood of the overlapping atom. Individual alignments with the smallest E_{pot} values were kept to result in the overall alignment.

In order to get chemically relevant overlapping, extended atom types that discriminate atomic number, hybridization state, and aromaticity were used. The 3D structures of the studied molecules in SDF format were used as starting conformations. The conformation of the molecules was kept flexible in order to maximize their overlap. Flexible ring size was set to eight, meaning that smaller rings were kept rigid unless their nonbarred bonds exceeded the applied limit of four. Molecules were aligned to each other one-by-one with normal accuracy option in all cases, and alignment scores were calculated that comprise the 3D similarity matrix. Then the procedures described in the previous section were applied.

dx.doi.org/10.1021/ci3001056 | J. Chem. Inf. Model. 2012, 52, 1733−1744

The Statistical Analysis System for Windows (version 9.2; SAS Institute, Cary, NC) was used for the implementation of all analyses.

RESULTS AND DISCUSSION

Figure 1 shows the distribution of the AUC values for the 52 main effect groups and the 70 subgroups using the studied three methods. Drug classification performed with our novel multidimensional approach, the Drug Profile Matching method, yielded AUC values greater than 0.86 and 0.90 for the studied main effects and subgroups, respectively (Figure 1 a,b). These values are remarkably higher than the performance of the other approaches, so DPM is clearly superior to the other popular methodologies investigated for this task. The results obtained by 2D and 3D similarity-based searches were found to be very similar both for the main groups and for the subcategories without any significant difference (Figure 1 c–f). Therefore these methods cannot be ranked on the basis of their overall performance on our data. However, for many effect categories either of them proved to perform significantly better; thus, in these cases, the method of choice could be selected.

Detailed results of the drug classification are displayed in Table 1 that shows the AUC values for each effect and each method. The classification performance of DPM can be described by a single AUC value because it is able to deal with all active molecules in a single search, while traditional 2D and 3D similarity-based searches utilize only one query structure. Therefore, every molecule registered to a given effect was applied as a reference structure in these cases. The obtained AUC values were averaged, and standard deviation was calculated. Figure 2 shows ROC curves for a selected effect that served for the calculation of the AUC values.

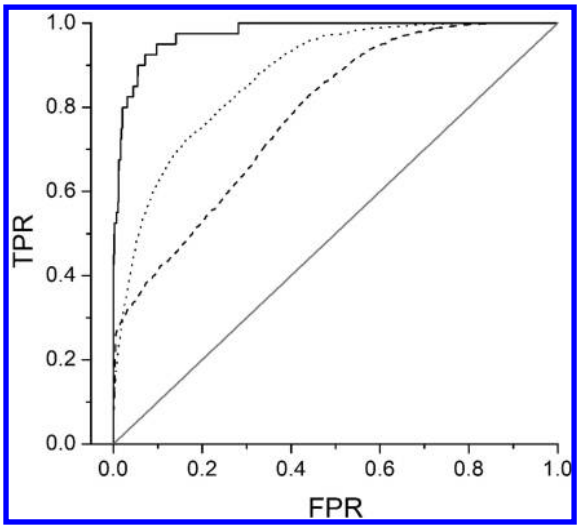


Figure 2. ROC curves for antipsychotic main effect category. The ROC curve is used for the characterization of classification accuracy. Here, ROCs of the antipsychotic main effect are shown produced by DPM, 2D similarity-based search, and 3D similarity-based search (solid, dashed, and dotted lines, respectively). The presented ROC curves for the traditional similarity searches (i.e., 2D and 3D similarity search, dashed and dotted lines) are the average of the individual curves that were obtained by using every registered molecule as a query structure. The gray diagonal line represents classification based on random guess.

Because an AUC value of 0.5 means random classification, the accuracy of conventional methods were considered insufficient in those cases where none of them could reach AUC > 0.6, which in our opinion indicates that meaningful enrichment could not be obtained. Accordingly, 20 main effect categories from the studied 52 could not be handled by the structure-based methods, such as anti-inflammatory agent, antihypertensive agent, or antineoplastic agent. These represent broad categories comprising diverse drug molecules, and 2D and 3D similarity searches typically fail on such groups. In the case of the subgroups, 12 effects from the studied 70 could not be screened by 2D and 3D similarity searches with sufficient accuracy. For example, Alzheimer's disease agent and sclerosis multiplex agent, subgroups of neurodegenerative disease agent main category, could not be treated by the conventional methods. These results also emphasize problematic therapeutic areas in the drug discovery field, where molecular similarity-based concepts obviously have their limitations.

In the case of DPM, the limit of meaningful accuracy was not only achieved for all of the studied effects but was also highly exceeded. However, results obtained with multidimensional statistical techniques, such as the classification of molecules performed by DPM, should be evaluated carefully. Overfitting can often take place, which produces misleading results such as artificial enrichment of active molecules in this case. In order to counter this effect, an independent 10-fold cross-validation was performed and repeated 100 times for every effect. The process started by partitioning the IP matrix and a selected effect into 10 complementary folds. One fold was removed, and the remaining 9 folds were merged to produce a classification function. This function was applied to calculate the classification probabilities of the drugs from the removed group. This process was repeated for each fold, and the mean probability value (MPV), which is the mean of the calculated probabilities for each drug registered to the studied effect, was calculated. Then, the whole cross-validation procedure was repeated 100 times to eliminate the impact of the distribution of the molecules on the results. Finally, the MPV values of the 100-times repeated 10-fold cross-validation experiments were averaged to give a single estimate of the robustness of the results (Table 1). This procedure was performed for every main effects and subcategories investigated in this study. Figure 3 shows the mean MPV values with standard deviations for the main and subgroups. Nineteen main effects and 36 subcategories were validated with a mean MPV value greater than 0.5, indicating that the method can classify the majority of the registered molecules into their respective effect classes without any prior knowledge about their activity (a randomized effect would result in an average probability value of 0.029). This effect list contains structure-based categories such as anticytokine agent and opioid (main effects) or barbiturate and glucocorticoid (subgroups), which obviously produce high AUC values with the 2D and 3D structure-based searches as well. However, wide main effects containing structurally diverse molecules such as anti-inflammatory agent and antihypertensive agent can also be found among them that raised difficulties for the conventional methods. In the case of the subgroups, target based effects, e.g., beta receptor antagonist or muscarinic antagonist gained high mean MPVs besides the structural categories. Medium mean MPV values (0.5 > mean MPV > 0.2) were obtained for 28 main effect categories and 23 subgroups, which suggests medium reproducibility (e.g., antiglaucoma agent and its corresponding subgroup alpha2

dx.doi.org/10.1021/ci3001056 | J. Chem. Inf. Model. 2012, 52, 1733–1744

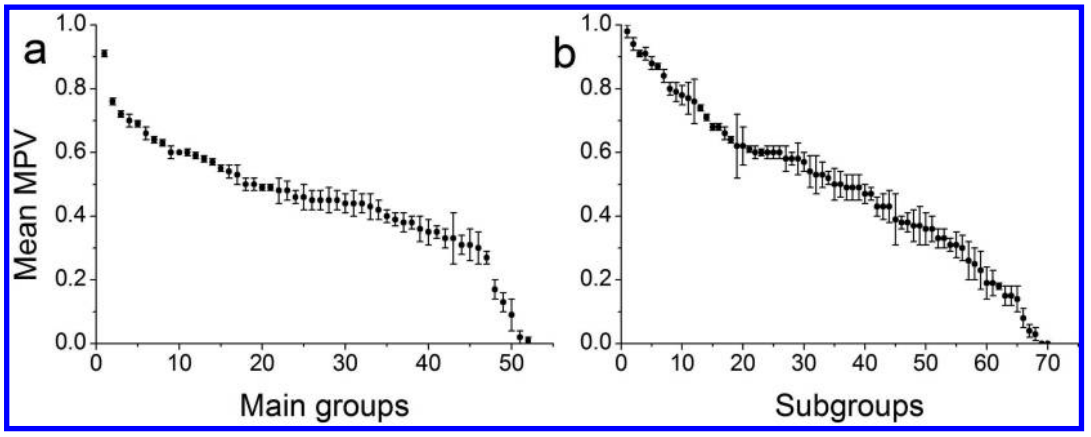


Figure 3. Means of mean probability values (MPVs) with standard deviations obtained from 10-fold cross-validation. To check the validity of DPM for drug classification, an independent 10-fold cross-validation was performed and repeated 100 times. The presented mean MPV values determine the resistance of the system against the 10% information loss. High mean MPV indicates the robustness of the method for a studied effect. (a) Mean MPV for the 52 main groups. (b) Mean MPV for the 70 subgroups.

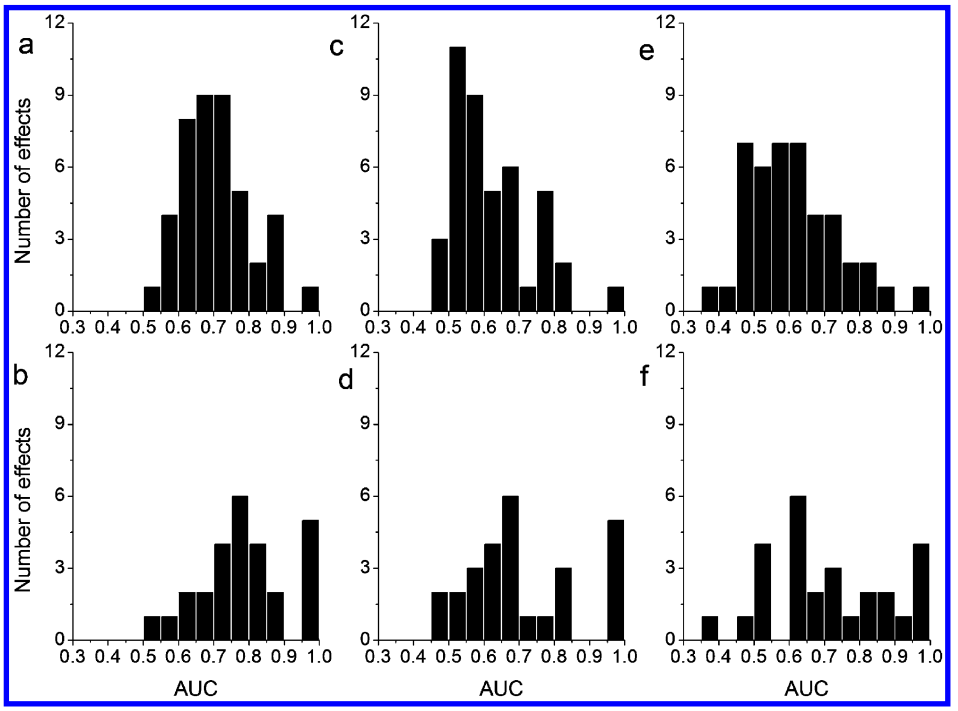


Figure 4. Distribution of AUC values obtained for the random splitting experiments. The histograms show the distribution of the AUC values for the studied 43 main effects (first row) and 27 subgroups (second row) obtained by DPM, 2D similarity-based search, and 3D similarity-based search. (a) DPM for main effects, (b) DPM for subgroups, (c) 2D similarity search for main effects, (d) 2D similarity search for subgroups, (e) 3D similarity search for main effects, and (f) 3D similarity search for subgroups.

receptor agonist with mean MPV values of 0.43 and 0.39, respectively). In these cases, subsets of the registered molecules are often validated with high probability, suggesting that these effects are not entirely cohesive on the basis of their IPs. However, even without the redistribution of the registered drugs, we consider DPM a reliable choice for these effects because there is still a clear convergence of the active molecules, and its results are superior to 2D and 3D analyses. Finally, 5 main effect categories and 11 subgroups produced low mean MPV values (mean MPV < 0.2) in this study, e.g., antiprotozoal agent and its respective subgroup antimalarial agent (their mean MPVs are 0.17 and 0.15, respectively). This result suggests that in these cases the DPM method cannot be applied in its current form but has to be further optimized. For most of these effects, conventional 2D and 3D similarity searches also could not achieve the limit of adequate accuracy.

On the basis of this validation study, we can conclude that the obtained high AUC values for the DPM method are not based on chance, and that a clear relationship can be observed between binding patterns and bioactivity profiles of the studied drug molecules. This result is in agreement with our previous study,[16] where a close relationship was identified between the studied 177 effect categories and the interaction patterns of the small-molecule drugs.

DPM produced significantly higher AUC values for almost all effect categories then the conventional searches, except for rigid structural groups (e.g., glucocorticoid, benzodiazepine) for which all three methods yielded high classification accuracy.

dx.doi.org/10.1021/ci3001056 | J. Chem. Inf. Model. 2012, 52, 1733−1744

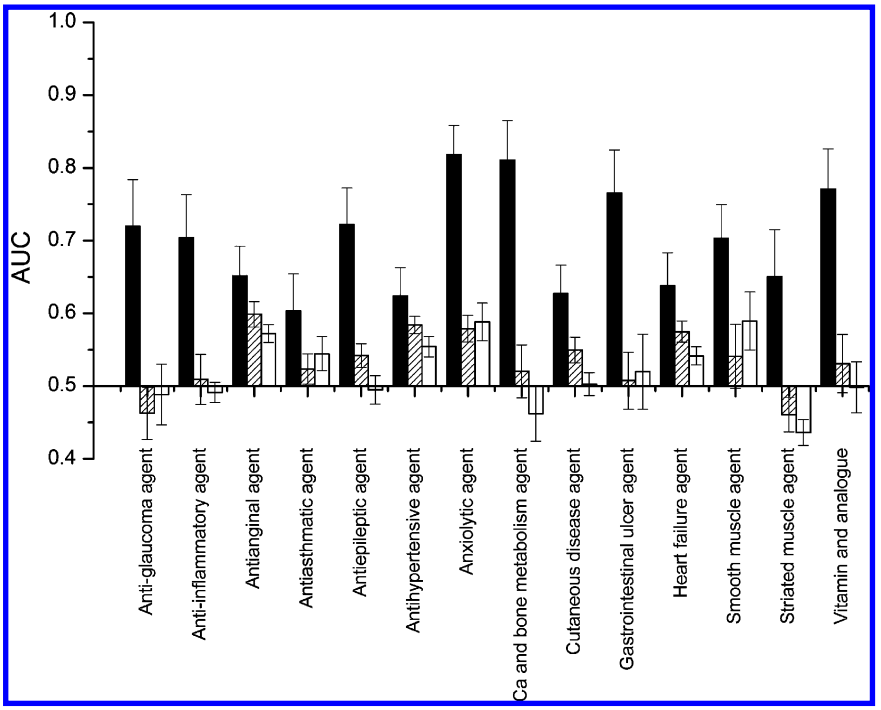


Figure 5. AUC values obtained by random splitting of those main effect categories for which conventional methods showed insufficient performance. AUC values of 14 main effect categories produced by DPM, 2D similarity search, and 3D similarity search (black, striped, and white bars, respectively). An AUC value of 0.5 represents random classification. Traditional similarity-based searches could not reach the limit of meaningful enrichment (AUC > 0.6) for the presented effects; however, DPM produced fair AUC values in all cases.

Thus, these groups can be sufficiently characterized by their 2D and 3D structural features. In the case of more diverse and flexible categories, the use of DPM clearly adds valuable information to the descriptions obtained by traditional structure-based methods. In our opinion, this result indicates that conformational selection of rare conformers might play a role in the docking-based DPM method. In DPM, molecules are characterized via their interaction to nontarget protein sites; thus, biologically more relevant rare conformers can also evolve in the binding pocket.

On the basis of this analysis and on our former study,[16] the applicability of DPM for drug classification was proved. A wider application area would open for DPM if the method could be used in virtual screening experiments. In virtual screening, results are based on extrapolation and on the estimation of class memberships of unknown molecules. To test the predictive power of DPM on external data, the studied data set was randomly split into two parts, and half of the molecules were used to establish a rule for classification while the other half was used for testing. The training set contained intentionally only 10 active molecules of a selected effect, while its remaining fraction was randomly filled with inactive molecules. This way, DPM could utilize only the required minimum amount of information about the active compounds. The second part of the data, subsequently used as a test set, was comprised of the residual active and inactive molecules. Because only effects containing 20 or more registered molecules were investigated in these analyses, the test set contained at least 10 active molecules. A classification function was established using the training set, and then it was applied to the test set to predict class membership probabilities. Test molecules were ordered by decreasing probability, and an AUC value was calculated to describe the accuracy. To eliminate the impact of molecule distribution between the two sets, this process was repeated 100 times for each effect, and an average AUC value was calculated.

To ensure comparability with the conventional similarity searches, the same test sets generated for DPM were used for 2D and 3D similarity searching, and the corresponding 10 active molecules from the former DPM training set were used one-by-one as a reference structure. Because traditional similarity searches utilize only the structure of the active molecules of a selected effect, inactive compounds of the training set were not exploited. Test molecules were ordered by decreasing similarity to the query structure, and AUC values were calculated. The 10 AUC values were averaged to produce a single estimate for accuracy. The whole process was repeated with all of the test sets, and active molecules from the former training sets for each effect.

Figure 4 shows the obtained AUC values for the random splitting experiments. The performance of 2D and 3D similarity searches slightly decreased in this new analysis compared to their drug classification results, but the differences are not significant. Therefore, we conclude that the accuracy of the conventional approaches is stable and predictable on our data. On the other hand, reduced performance was observed for DPM, but the comparison of its accuracy results (displayed in Table 1) to conventional similarity searches revealed that even with this reduced performance, DPM is superior to the other two methods for many effects. Detailed examination of the results enables us to get a more realistic view about its feasibility domain.

Similarly to the drug classification presented previously, the accuracy of the searches were considered insufficient with an AUC < 0.6. In spite of the weaker accuracy of DPM, only five main effects and two subcategories could not reach this limit such as neurodegenerative disease agent main effect and interestingly its two subcategories: sclerosis multiplex agent and

dx.doi.org/10.1021/ci3001056 | J. Chem. Inf. Model. 2012, 52, 1733−1744

Parkinson disease agent. Thus the division of this medical effect to smaller groups did not have a remarkable positive effect on the performance. As expected, structure-based groups showed the smallest reduction in accuracy: AUC values, for example, of opioid main group and glucocorticoid subcategory dropped from 0.97 and 1.00 to 0.88 and 0.98, respectively. Target-based categories such as antihistamine main effect and its corresponding subgroup histamine H1 receptor antagonist still showed appropriate AUC values (0.74 and 0.82, respectively). Although considerably reduced in performance, main effects containing molecules with different mechanisms of action such as sympathomimetic and antiepileptic agent also produced fair AUC values of 0.75 and 0.72, respectively. In spite of having AUCs above the limit of meaningful enrichment, several effects failed to produce sufficient accuracy in the random splitting experiments (e.g., antiemetic agent and antihypertensive agent dropped to 0.63 and 0.62, respectively). For these effects, DPM can be useful for drug classification, but its predictive power is weak. Therefore, its screening capabilities outside the parent data set may be limited.

Detailed comparison of the random splitting results of the three methods revealed those effects that could not be handled by the conventional similarity searches, but DPM was found to be a satisfying solution for them. For 14 of the studied 43 main effects, standard 2D and 3D similarity searches could not reach the limit of meaningful enrichment (AUC < 0.6), but DPM produced fair AUC values for them. Figure 5 displays the AUC values for these effects. The highest AUC differences were observed for anxiolytic agent, calcium and bone metabolism agent, and gastrointestinal ulcer agent (AUCs of 0.82, 0.81, and 0.77 produced by DPM, respectively). Anti-inflammatory agent and antiepileptic agent also raised difficulties for the standard 2D and 3D similarity methods, but could be handled effectively by DPM (AUCs of 0.70 and 0.72, respectively). The Tanimoto diversity values of these 14 effects are between 0.30 and 0.45, indicating a high degree of structural dissimilarity among the registered molecules (the Tanimoto diversity for the studied main groups ranges from 0.30 to 0.65, see Table 1). This result clearly shows the strength of the DPM methodology: It produces sufficient classification performance even for wide effect categories containing structurally diverse molecules, where structure-based similarity searches typically fail.

In the case of the subcategories, there were only 3 of the studied 27 effects for which conventional similarity searches failed to reach the AUC > 0.6 limit, but DPM produced high AUC values (antigout agent, cell proliferation agent, hormone system associating agent, and central striated muscle relaxant). However, Figure 5 shows that DPM produced AUC values comparable to that of the traditional methods, and its performance was considerably higher for several effects (e.g., for NSAID, Nonselective COX inhibitor, the AUC values are as follows: 0.81, 0.63, and 0.74 obtained by DPM and 2D and 3D similarity searches, respectively). Nevertheless, for many effects the more complex DPM approach could not improve the accuracy significantly.

On the basis of this study, we can conclude that DPM does have a predictive power outside the parent data set that could be exploited in real-world virtual screening experiments in the future. Its main strength is its ability to screen wide effect categories containing structurally diverse molecules. A former study also emphasized the ability of affinity fingerprints to identify molecules with similar bioactivity independent of their structural scaffolds.[7] Moreover, their use in virtual screening was found to be superior to a conventional topological similarity approach for 11 studied activity classes.[10]

For several subgroups, the accuracy of all the three applied methodologies was high in both experiments. The molecules registered to these effects usually share a high degree of structural similarity (e.g., glucocorticoid). Therefore, we conclude that in these cases the prediction power of DPM is primarily based on the molecular structure and our more complex multidimensional approach based on the IPs of the drugs does not improve the screening results obtained by the traditional structure-based searches. However, DPM is able to detect biological similarity even in the case of limited structural similarity because no direct topological similarity information on drug molecules is involved in the method.

The use of the two-level hierarchical effect database in this study enabled us to investigate the impact of reducing large effect classes to smaller but more cohesive subgroups on the basis of molecular structure, mechanism of action, or common target. In most cases, partitioning of wide effect classes resulted in improved classification accuracy as was expected. Highest increment was obtained by dividing a medical effect into structure-based subgroups (e.g., anxiolytic agent main effect into barbiturates and benzodiazepines). Nevertheless, subgroups based on common targets also produced high AUC values such as beta receptor antagonist and serotonin receptor agonist. For certain effects, distribution of molecules into subgroups does not have a remarkable positive effect on the performance, e.g., neurodegenerative disease agent is divided into four subcategories, but notable improvement is only observed for one group (tardive dyskinesia agent).

Besides the presented overall high performance of the DPM method, it also has the advantage of handling the information of all the active molecules of a selected effect in one search. In contrast, a simple virtual screening protocol, as was also performed here, usually computes the similarity between a known target structure and each of the molecules in a compound library and sorts the molecules in decreasing order of the similarities. However, some novel methods have been introduced recently to eliminate this problem such as group fusion that allows the reference molecule to vary and combines the computed similarity values for each molecule to create a new fused score.[25,26] Nevertheless, the high standard deviations of the AUC values obtained in this work suggest that there is great fluctuation in the individual AUC values, and fused searches are unlikely to produce better results than the presented ones here applying the DPM method. Moreover, the DPM method also utilizes information provided by the inactive molecules, whereas conventional methods of molecular similarity are based only on the active compounds.

The applicability of 3D descriptor-based approaches is debated in the literature. Without conformer mapping, 3D descriptors tend to be less efficient than 2D descriptors (while needing larger computational resources). Even after careful handling of the possible conformations, 2D descriptors can still outperform the 3D approaches.[27,28] In this study, the standard structure-based searches resulted in similar screening performance, although slight superiority of 2D search was observed for a considerable number of subgroups. However, the efficacy of 3D similarity search is definitely comparable here to the 2D method. This can be a consequence of the use of the therapeutically more relevant effect database that consists of broad effect categories containing structurally diverse molecules. Also, the screening power of 2D methods can

significantly decrease on such data sets because they often favor molecules with common core structures. In spite of the improved efficacy of 3D similarity search on our data set, the results also show that compound conformational mapping in the presence of highly discriminative surfaces like a series of protein binding sites yields more information than mapping in water or vacuum, which is another advantage of the interaction pattern-based methodology. We assume that rare conformers can also evolve in the binding pockets which may be biologically more relevant, and this can be the reason behind the strikingly high accuracy of the DPM method.

CONCLUSIONS

In this paper, the applicability of Drug Profile Matching for drug classification and its predictive power for virtual screening was investigated, and the results of a comparison study of standard 2D and 3D similarity-based screening approaches and the interaction pattern-based methodology were summarized. We conclude that in the case of drug classification Drug Profile Matching performs better than 2D and 3D descriptor-based approaches for a majority of the studied effect categories. Its predictive power on external data was also analyzed in detail, and we revealed that the highest gain is obtained for complex, wide effect categories. A possible explanation for the high predictive power of DPM might be its ability to select rare conformers via mapping molecular interactions.

A clear relationship was observed between interaction patterns and bioactivity profiles of the drug molecules. Our results also demonstrate that DPM is able to systematically unravel the complete effect profiles of small-molecule drugs, including their target profiles, thus enabling the potential use of DPM in multitarget drug design.[29]

AUTHOR INFORMATION

Corresponding Author

*Phone: +36 1 372 2500, ext. 8780. Fax: +36 1 381 2172. E-mail: malna@elte.hu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The European Union and European Social Fund have provided financial support under the Grant Agreement No. TÁMOP 4.2.1./B-09/KMR-2010-0003. This work is also supported by the Hungarian Academy of Sciences (HAS-ELTE research group ID: 01055). The National Development Agency has also provided financial support under the aegis of the National Technology Programme (NTP_TECH_08_A1/2-2008-0106). C.H. is thankful for a János Bolyai Research Scholarship provided by the Hungarian Academy of Sciences.

ABBREVIATIONS:

AUC: Area under the curve; BEDROC: Boltzmann-enhanced discrimination of ROC; CCA: canonical correlation analysis; DPM: Drug Profile Matching; EP: effect profile; FDA: Food and Drug Administration; FPR: false positive rate; IP: interaction pattern; LDA: linear discriminant analysis; PDB: Protein Data Bank; ROC: Receiver operating characteristic; TPR: true positive rate

REFERENCES

(1) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* **2010**, *53*, 8461−8467.
(2) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225−233.
(3) Willett, P. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(4) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.
(5) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. *J. Med. Chem.* **2006**, *49*, 6802−6810.
(6) Venkatraman, V.; Perez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *J. Chem. Inf. Model.* **2010**, *50*, 2079−2093.
(7) Briem, H.; Lessel, U. F. In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes. *Perspect. Drug Discovery Des.* **2000**, *20*, 231−244.
(8) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**, *2*, 107−118.
(9) Kauvar, L. M.; Villar, H. O.; Sportsman, J. R.; Higgins, D. L.; Schmidt, D. E., Jr. Protein affinity map of chemical space. *J. Chromatogr., B: Biomed. Sci. Appl.* **1998**, *715*, 93−102.
(10) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. "Bayes affinity fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: When are multitarget drugs a feasible concept? *J. Chem. Inf. Model.* **2006**, *46*, 2445−2456.
(11) Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem.* **2007**, *2*, 861−873.
(12) Briem, H.; Kuntz, I. D. Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.* **1996**, *39*, 3401−3408.
(13) Lessel, U. F.; Briem, H. Flexsim-X: A method for the detection of molecules with similar biological activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 246−253.
(14) Fukunishi, Y.; Hojo, S.; Nakamura, H. An efficient in silico screening method based on the protein−compound affinity matrix and its application to the design of a focused library for cytochrome P450 (CYP) ligands. *J. Chem. Inf. Model.* **2006**, *46*, 2610−2622.
(15) Fukunishi, Y.; Mikami, Y.; Takedomi, K.; Yamanouchi, M.; Shima, H.; Nakamura, H. Classification of chemical compounds by protein-compound docking for use in designing a focused library. *J. Med. Chem.* **2006**, *49*, 523−533.
(16) Simon, Z.; Peragovics, A.; Vigh-Smeller, M.; Csukly, G.; Tombor, L.; Yang, Z.; Zahoranszky-Kohalmi, G.; Vegner, L.; Jelinek, B.; Hari, P.; Hetenyi, C.; Bitter, I.; Czobor, P.; Malnasi-Csizmadia, A. Drug effect prediction by polypharmacology-based interaction profiling. *J. Chem. Inf. Model.* **2012**, *52*, 134−145.
(17) Simon, Z.; Vigh-Smeller, M.; Peragovics, A.; Csukly, G.; Zahoranszky-Kohalmi, G.; Rauscher, A. A.; Jelinek, B.; Hari, P.; Bitter, I.; Malnasi-Csizmadia, A.; Czobor, P. Relating the shape of protein binding sites to binding affinity profiles: is there an association? *BMC Struct. Biol.* **2010**, *10*, 32.
(18) Hetenyi, C.; Maran, U.; Karelson, M. A comprehensive docking study on the selectivity of binding of aromatic compounds to proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1576−1583.
(19) Jiang, X.; Kumar, K.; Hu, X.; Wallqvist, A.; Reifman, J. DOVIS 2.0: an efficient and easy to use parallel virtual screening tool based on AutoDock 4.0. *Chem. Cent. J.* **2008**, *2*, 18.

(20) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* **2007**, *28*, 1145−1152.
(21) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided. Mol. Des.* **2002**, *16*, 11−26.
(22) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. 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.* **2005**, *48*, 2534−2547.
(23) Chang, M. W.; Ayeni, C.; Breuer, S.; Torbett, B. E. Virtual screening for HIV protease inhibitors: A comparison of AutoDock 4 and Vina. *PLoS One.* **2010**, *5*, e11955.
(24) Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.
(25) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.
(26) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 746−753.
(27) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(28) Roy, K.; Roy, P. P. Comparative QSAR studies of CYP1A2 inhibitor flavonoids using 2D and 3D descriptors. *Chem. Biol. Drug Des.* **2008**, *72*, 370−382.
(29) Koutsoukas, A.; Simms, B.; Kirchmair, J.; Bond, P. J.; Whitmore, A. V.; Zimmer, S.; Young, M. P.; Jenkins, J. L.; Glick, M.; Glen, R. C.; Bender, A. From in silico target prediction to multi-target drug design: current databases, methods and applications. *J. Proteomics.* **2011**, *74*, 2554−2574.