JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Drug Effect Prediction by Polypharmacology-Based Interaction Profiling

Zoltán Simon,[†,‡] Ágnes Peragovics,[†,‡] Margit Vigh-Smeller,[†] Gábor Csukly,[§] László Tombor,[§] Zhenhui Yang,[†] Gergely Zahoránszky-Kőhalmi,[†] László Végner,[†] Balázs Jelinek,[†] Péter Hári,[‡] Csaba Hetényi,[†] István Bitter,[§] Pál Czobor,[§] and András Málnási-Csizmadia*[,†]

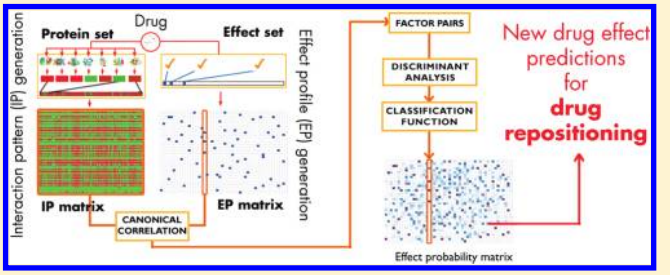
[†]Department of Biochemistry, Institute of Biology, Eötvös Loránd University, Pázmány Péter sétány 1/C, H-1117 Budapest, Hungary

[‡]Delta Informatika, Inc., Szentendrei út 39-53, H-1033 Budapest, Hungary

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

Supporting Information

ABSTRACT: Most drugs exert their effects via multitarget interactions, as hypothesized by polypharmacology. While these multitarget interactions are responsible for the clinical effect profiles of drugs, current methods have failed to uncover the complex relationships between them. Here, we introduce an approach which is able to relate complex drug−protein interaction profiles with effect profiles. Structural data and registered effect profiles of all small-molecule drugs were collected, and interactions to a series of nontarget protein binding sites of each drug were calculated. Statistical analyses confirmed a close relationship between the studied 177 major effect categories and interaction profiles of ca. 1200 FDA-approved small-molecule drugs. On the basis of this relationship, the effect profiles of drugs were revealed in their entirety, and hitherto uncovered effects could be predicted in a systematic manner. Our results show that the prediction power is independent of the composition of the protein set used for interaction profile generation.



INTRODUCTION

One of the most exciting questions in modern science is the prediction of processes or unknown parameters of complex systems. Pharmacology is such a complex system. Predicting effect profiles (EPs) of drugs and drug candidates is a great challenge, which may be improved using their atomic-level structural data. The EP of a drug is a complex feature since a molecule entering the organism usually interacts with multiple targets, as indicated by the theory of polypharmacology.[1−4] Multiple actions may be important for clinical efficacy, especially in the case of complex diseases. For example, psychiatric drugs affecting several well-defined proteins have high efficacy.[5] Thus, single target-based approaches may prove insufficient for identifying the full spectrum of EP of molecules.[4] In addition, considering that our knowledge is limited even for routinely used drugs, the discovery of new effects frequently leads to new indications for existing drugs. Some typical examples include sildenafil, originally an antianginal agent which was repositioned to the treatment of male erectile dysfunction, or topiramate, a former antiepileptic agent, recently approved for treatment of obesity.[6] In a similar manner, the introduction of new compounds frequently reveals unpredicted side effects. Thus, it is becoming increasingly recognized that the prediction of the full EP is essential to revealing the mechanisms of drug actions and side effects.[7] Up until now, heuristic and empirical experiences have played the principal role in identifying various effects of bioactive molecules. Recently developed systematic prediction methods, however, increase the efficiency of drug development and safety control. For example, Keiser et al. related drug targets to each other on the basis of chemical similarity measurements of their ligands[8] and predicted new targets for existing drugs and proved 23 new drug-target associations.[9] Campillos et al. used side effect information to determine the possibility of two drugs sharing the same target,[10] resulting in 13 confirmed interactions out of 20 predictions. Kauvar et al. measured the binding potencies of several compounds against a reference panel of eight (in a later work, 18) proteins that defines the affinity fingerprints of the applied compounds in order to predict the binding properties of the compounds to other proteins not represented in this reference panel.[11,12] Fliri et al., building on the pioneering work of Kauvar et al., found a weak relationship between affinity fingerprints and the side effect data of drug molecules.[13] Bender et al. introduced the "Bayes Affinity Fingerprint" similarity search approach, in which compound similarity is determined by similarities of binding affinity values against a panel of pharmacological target proteins, and proved its superior performance over conventional structural similarity searches.[14,15]

Our working hypothesis was that a feature set must comprise similar complexity to that of clinical effect profiles in order to

Received: May 6, 2011
Published: November 18, 2011

dx.doi.org/10.1021/ci2002022 | J. Chem. Inf. Model. 2012, 52, 134–145

yield systematic information with predictive power for the effect profiles.[4] The task was to extract the relevant information stored in complex feature sets of drug molecules in order to unravel effect profiles in their entirety. To accomplish this, in the present study, an atomic-level strategy is introduced for the prediction of the effect profiles of drugs by systematic mapping of their molecular interactions. For this, the central assumption of polypharmacology is adopted, and it is presumed that similar interaction profiles (IPs) of molecules are related to their similar biological actions. In order to test this assumption empirically, we generated IPs for 1177 FDA-approved drugs by calculating their binding affinities for a set of proteins, and the IPs were correlated with the EPs of all drugs. A correlation between IPs and EPs would hold out the promise for the discovery of novel effects of drugs and the prediction of side effects of drug candidates in the development phase. The aims of the present study are (1) to uncover IP−EP relationships and (2) to derive general rules for effect prediction.

METHODS

Generation of the Interaction Profile (IP) Matrix. IP generation was done as described in our previous work.[16] In short, 1226 FDA-approved drug molecules were extracted from DrugBank database[17] as of June 2009 (Table S1, Supporting Information). A total of 49 entries were removed for various reasons (e.g., structure contains a metal ion, two components under one name, etc.); 149 proteins were collected from RCSB Protein Data Bank[18] (PDB), which met the following requirements:

(1) The structure contained a ligand.
(2) The resolution was better than 2.3 Å.
(3) There was a complete ligand binding site.
(4) If a mutant protein had been selected, the amino acid sequence was not changed in the binding pocket, and fewer than five mutations were in other regions.
(5) Water molecules were not involved in the ligand binding.

Table S2 (Supporting Information) shows the list of the PDB codes of the applied proteins. Docking preparations and calculations were performed using the DOVIS 2.0 software (DOcking-based VIrtual Screening),[19] using the AutoDock4 docking engine,[20] the Lamarckian genetic algorithm and X-SCORE,[21] and AutoDock4 scoring functions. Docking runs were repeated using the AutoDock4 scoring function to assess the impact of different scoring functions on the results, and the same analysis procedure was further applied to them. Explicit hydrogens were added to the drug molecules, and optimization procedures were applied for aromatic rings and for the overall 3D structure before docking using the ChemAxon JChem Base software (version 5.2.0, 2008). All ligands and other molecules were removed during the preparation of the protein PDB file. The docking box was centered on the geometrical center of the original ligand of the protein (as found in the intact PDB file); the box size and grid spacing were set to 22.5 Å and 0.375 Å, respectively. Protein parts outside the box were excluded from the calculations. The applied box size enables each member of the drug set to rotate freely in order to find the conformation with the lowest binding free energy without steric clashing with the box perimeter. No further reductions in box size were applied to smaller ligands. Protein structures were kept rigid during docking according to our initial hypothesis that a uniform, constant discriminative surface is required for creating interaction profiles.

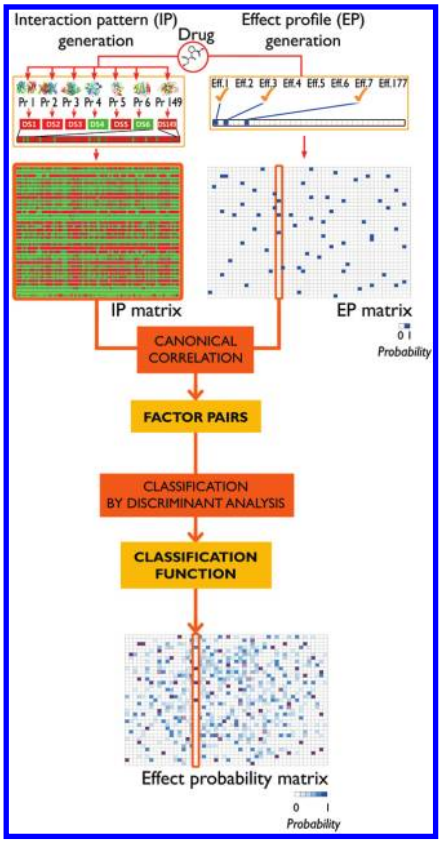


Figure 1. Graphical summary of the Drug Profile Matching method: from the atomic structures to the effect probability matrix. A drug molecule is docked to a set of 149 proteins, and the calculated binding free energies (docking scores, DS1-149) are entered into a row vector, i.e., the interaction profile (IP). IPs of the 1177 studied drugs form the IP matrix. The effect pattern (EP) matrix contains the therapeutic effects of the drugs in a binary coded form (blue and white cells represent the presence and the absence of a given effect from the 177 categories, respectively). Then, a canonical correlation analysis is performed in order to generate highly correlating factor pairs that serve as the input for linear discriminant analysis. This way, classification functions are produced that yield the probability for each drug−effect pair, resulting in the effect probability matrix. Note that the values in this matrix are continuous. See the text and Supporting Information for details of the Drug Profile Matching method.

Twenty-five docking runs were performed for each job on a Hewlett-Packard cluster of 104 CPUs. Each drug molecule was docked to each protein (1177 × 149 = 175 373 dockings, individual docking runs: 175 373 × 25 = 4 384 325). Binding free energies were extracted, and the minima were imported to the IP database (Figure 1). Here, drugs are ordered in rows, and the columns represent the individual proteins. This way, each row forms the interaction profile for the given drug.

Diversity Analysis. We assume that an IP vector with a diverse set of proteins used in the present study might model the interactions formed by a given drug with the human proteome. To check this assumption, the diversity of the protein set was calculated from the similarity values of the binding site geometry descriptors obtained from the PocketPicker software.[22] A total of 95.5% of the values in the protein−protein dissimilarity half-matrix are above the dissimilarity threshold,[22] suggesting a fairly diverse set of proteins.

Protein Set Size Evaluation. An evaluation procedure was applied on different protein set sizes in order to determine

the required number of proteins for efficient classification. Randomly generated protein sets containing 1, 5, 10, 40, 70, 100, and 130 proteins were used to produce the IPs of the drugs. Then, the DPM method was performed effect by effect, as described in the following sections, and the resulting classification accuracy values (AUCs) were extracted. Each protein set was generated three times. The following hyperbolic function was fitted to the mean AUC values at the seven set sizes for each effect:

$$y = \frac{a \times x}{b + x} + c$$

The maximum obtainable AUC equals $a + c$, while parameter b is the number of proteins required to reach 50% of the maximal obtainable AUC.

Generation of Effect Profile (EP) Matrix. As mentioned above, structural and pharmacological information on 1177 FDA-approved small-molecule drugs was extracted from the DrugBank database.[17] Then, a list of 559 effects was formed that contained all effect entries that appeared in the drug information. Effect entries were further refined in order to eliminate initial database inconsistencies. Since effect categories with less than 10 registered drugs contain an insufficient amount of information for meaningful classification, the effect list was reduced to 177 categories. Figure S1 (Supporting Information) shows the distribution of the number of drugs registered to an effect. Then, a binary matrix was formed that shows the presence or absence of the studied 177 effects for each drug. (The appearance of an effect for a drug is marked with a "1" value and *vice versa*.)

Statistical Analyses. *Canonical Correlation Analysis.* In order to match the complex pattern structures of IP and EP matrices, we adopted canonical correlation analysis (CCA). CCA is a "bimultivariate" method that has the advantage of simultaneous handling of two separate sets of variables, which we had in our study (i.e., IP and EP descriptor variables, respectively). In CCA, the relationship between the two sets is studied by creating derived variables ("variates") that are linear composites of the original variables. The principal goal is to simplify complex relationships, while providing some specific insights into the underlying structure of the data. An analogy to factor analysis, a more familiar method, may be helpful in explaining CCA. In factor analysis, variates (factors) are formed from one set of variables to describe the correlation structure in the same set of variables. In CCA, variates in one set are formed to describe the correlation structure in a different set of variables. Therefore, CCA can be viewed as an extension of factor analysis for two separate sets of variables. In particular, the objective of this method is to obtain as high a correlation as possible between the derived variables (here, pairs of variates or "factors" are formed from the two sets) in variable set 1 (i.e., set of IPs in current study) and those in variable set 2 (i.e., set of EPs in current study). In other words, this technique is an optimal linear method for studying interset association: canonical factor pairs from the two sets are extracted jointly to be maximally correlated with a component of the complementary variable set (Figure S3, Supporting Information).

Linear Discriminant Analysis. On the basis of the above-described canonical factor pairs of IPs and EPs, we calculated the probability of each effect for each drug via linear discriminant analysis (LDA; Figure S2, Supporting Information). In particular, LDA is a classical statistical approach to finding an optimal linear transformation for maximizing the between-class variance and minimizing the within-class variance, thereby identifying the best discriminating surfaces or "hyperplanes" in the multidimensional space of feature sets that generate complex pattern classes (such as the interaction profile of drugs at the atomic level, or IPs, in our study). Using the mathematical equation of such discriminating surfaces, classification functions for each effect were determined in order to classify observations into known effect classes based on their IP canonical factors. The performance of the classification function was evaluated by estimating the drug effect probability for each drug with regard to each effect and the rate of correct classification for all drugs with regard to all effects. In order to accomplish this, each observed IP was plugged into the classification function in order to generate the drug−effect probability matrix (Figure 1).

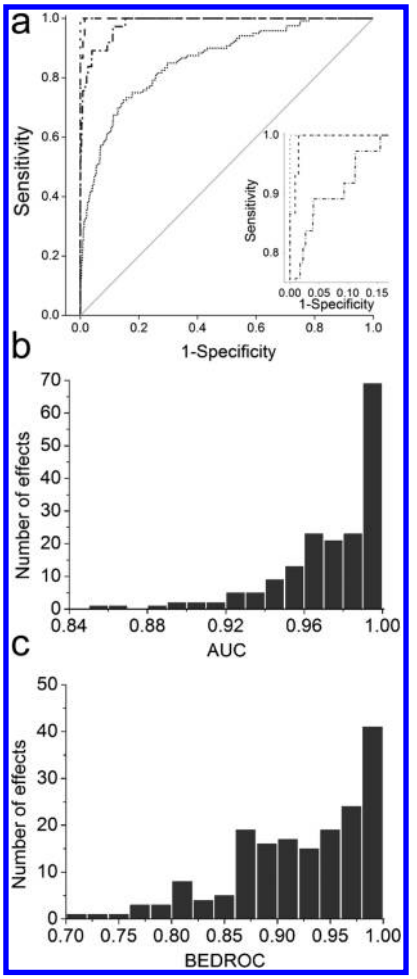


Figure 2. (A) Representative ROC curves. The ROC curve provides a characterization of classification accuracy; here, ROCs of the "tetracycline" (best classification), "ACE inhibitor", "COX inhibitor", and "antineoplastic agent" (our most inefficient classification) effect categories are shown (dotted, dashed, dash-dotted, and short-dotted lines, respectively). The gray diagonal line represents classification based on random guess. The inset shows an enlarged portion of the upper left region of the plot. (B) AUC histogram, showing the distribution of the area under the curve (AUC) values for the studied 177 effects. Results suggest that near-perfect classification was obtained in most cases. (C) Distribution of the BEDROC values for the 177 studied effect categories.

The Statistical Analysis System for Windows (version 9.2; SAS Institute, Cary, NC) was used for the implementation of all statistical analyses, including CCA (CANCORR Procedure) as well as LDA (DISCRIM Procedure).

Validation. In order to evaluate the robustness of our results, i.e., the extent to which the aforementioned effect classification results would generalize to independent data, the commonly used 10-fold cross-validation was performed (Figure 3A). It partitions the data into 10 complementary sets (also called "folds"). Each fold is retained as a test set for validation, and the remaining folds are used as a training set for the establishment of the classification model. When the standard 10-fold cross-validation approach was adopted in this study, the data set was divided into 10 complementary folds. In each round of validation, one fold was set as a test set, and the remaining folds comprised the training set. CCA and LDA were conducted to derive the IP-based classification function using the training set and computing the drug−effect probability as well as determining (predicting) effect−group membership for the test set. This round was performed for each of the 10 folds, and the cross-validation results for each of the originally registered drugs were then combined to yield a single average estimate for each effect (mean probability value, MPV). The whole process was repeated 100 times.

A more rigorous 3-fold cross-validation was also performed to prove the robustness of the method.

Receiver Operating Characteristic Analysis. The efficacy of the classification functions was assessed by Receiver Operating Characteristic (ROC) analysis, i.e., determining the true positive rate (TPR) and the false positive rate (FPR) for every effect, using the classification function (determined by LDA) and a sliding cutoff parameter running from 1 to 0. Molecules are reclassified at each point, considering compounds as "positive" if they have a greater possibility for an effect than the actual cutoff value and "negative" in the opposite case. Positives can be further divided into true and false positives depending on the binary value originally assigned to the given drug−effect pair; i.e., if a drug had "1" in the effect profile and produced a classification value larger than the cutoff point, it will be considered a "true positive". True and false negatives can be distinguished as well at each step. TPR and FPR are the rate of true positives among the positives and the rate of false positives among the negatives, respectively, and are often referred to as sensitivity and 1−specificity. TPR and FPR values for each cutoff point are plotted on a two-dimensional graph called the ROC curve (Figure 2A). A completely random classification would result in a ROC curve on the diagonal of the graph, meaning that for every true positive hit, a false positive hit also falls into the classification. The better the classification, the closer the curve to the (0,1) point of the graph. Classification accuracy can be characterized by the area under the ROC curve, i.e., the AUC value (ranging from 0 to 1).

Boltzmann-Enhanced Discrimination of ROC. AUC is proved to be a useful metric in many disciplines; however, it does not address the "early recognition" problem specific to virtual screening. Virtual screening methods must rank actives early in an ordered list, since the number of compounds to be tested is generally limited. The Boltzmann enhanced discrimination of ROC (BEDROC) metric uses an exponential weight formula that gives bigger scores to the actives appearing at the top of the list.[23] Similarly to the AUC value, BEDROC also ranges from 0 to 1, and a higher value means better classification in terms of "early recognition".

$$\mathrm{BEDROC} = \frac{\sum_{i=1}^{n} e^{-\alpha r_i/N}}{\frac{n}{N}\left(\frac{1-e^{-\alpha}}{e^{\alpha/N}-1}\right)} \times \frac{R_a \sinh(\alpha/2)}{\cosh(\alpha/2)-\cosh(\alpha/2-\alpha R_a)} + \frac{1}{1-e^{\alpha(1-R_a)}} \quad \text{if } \alpha R_a \ll 1 \text{ and } \alpha \neq 0$$

where r_i is the rank of the ith active in an ordered list, N is the number of total compounds, n is the number of actives, R_a is the ratio of actives (n/N), and α is the tuning parameter. The higher the values of α, the "earlier" the region of the ordered list that is emphasized by higher weighting. $\alpha = 5$ was used in our calculations; this value corresponds to 80% of the score coming from approximately the top 30% of the list.

Top Hit Rate Calculation. The entire set of the 1177 drugs was listed in descending order by the probability value of possessing the given effect, and the *top of the list* was cut at the number of the registered drugs to the studied effect. This top list contains registered and unregistered drugs of the given effect since the unregistered drugs can also gain a high probability value in the Drug Profile Matching method and registered drugs can have a low value.

Classification accuracy can be characterized with the proportion of the registered drugs in the top list. Therefore, the following top hit rate value was calculated for each of the 177 effects:

$$\text{top hit rate} = \frac{\text{number of the registered drugs in the top of the list}}{\text{number of all registered drugs of the given effect}}$$

Here, the number of all registered drugs of the given effect equals the number of drugs in the top list, as discussed above. The distribution of top hit rates can be found in Figure S3 (Supporting Information).

RESULTS AND DISCUSSION

EPs and IPs were generated on the basis of structural and pharmacological information on 1177 FDA-approved small-molecule drugs (Figure 1 and Table S1, Supporting Information). EPs were extracted from the DrugBank database[17] and stored as a row vector for each drug with binary entries, i.e., "1" for the presence and "0" for the absence of a given effect, comprising 177 effect categories. For the IPs, a diverse set of 149 proteins were selected from the Protein Data Bank (Table S2, Supporting Information) on the basis of their suitability for docking studies. Structures of the 1177 × 149 drug−protein complexes were obtained using the popular docking software AutoDock4.[20,24] The corresponding binding affinity values were calculated using X-SCORE and Autodock4 scoring functions[19−21] as described earlier,[16] and the binding affinity values were entered into the IP vectors as recommended by an earlier study.[25] The EP and IP vectors were collected into matrices and used as input databases in the subsequent investigations (Figure 1).

The evaluation of the relationship between IP and EP is a cornerstone of our approach called Drug Profile Matching (DPM) method (Figure 1). In order to match the complex pattern structures, canonical correlations were applied between the IP matrix and each studied effect category, and the basic underlying factor pairs that show maximal correlation between the two data sets were identified. Using these IP and EP factor

Table 1. Prediction and Validation Properties of the Studied 177 Effect Categories[a]

		accuracy			10-fold cross-validation probability values			
effect	*n*	AUC	BEDROC	top hit rate	mean	std	mean 75%	std 75%
adrenergic agent	132	0.9186	0.8091	0.6136	0.6157	0.0080	0.7750	0.0095
adrenergic agonist	38	0.9677	0.8963	0.6053	0.5579	0.0203	0.7292	0.0263
adrenergic α agonist	20	0.9904	0.9597	0.7000	0.5163	0.0476	0.6883	0.0634
adrenergic α antagonist	27	0.9806	0.9227	0.6667	0.3704	0.0307	0.4728	0.0390
adrenergic antagonist	61	0.9521	0.8567	0.5738	0.5984	0.0110	0.7643	0.0124
adrenergic β agonist	17	0.9953	0.9791	0.7647	0.7177	0.0364	0.9203	0.0396
adrenergic β antagonist	23	0.9905	0.9612	0.8261	0.7170	0.0188	0.9048	0.0219
adrenergic uptake inhibitor	19	0.9901	0.9559	0.6842	0.4366	0.0435	0.5529	0.0551
alkylating agent	17	0.9781	0.9287	0.6471	0.2225	0.0300	0.2909	0.0392
amphetamine	16	0.9928	0.9675	0.6875	0.6197	0.0185	0.8263	0.0247
analgesic agent	92	0.8966	0.7669	0.5652	0.5449	0.0119	0.7106	0.0153
analgesic agent, opioid	23	0.9900	0.9562	0.6957	0.5423	0.0220	0.6929	0.0281
analgesic agent, non-narcotic	12	0.9839	0.9331	0.6667	0.0614	0.0232	0.0819	0.0309
anesthetic agent	41	0.9661	0.8724	0.4878	0.4456	0.0206	0.5776	0.0262
anesthetic agent, intravenous	12	0.9956	0.9802	0.7500	0.0925	0.0389	0.1232	0.0518
anesthetic agent, local	25	0.9747	0.9242	0.6400	0.5254	0.0300	0.6807	0.0376
angiotensin-converting enzyme inhibitor	15	0.9986	0.9933	0.8000	0.4197	0.0426	0.5228	0.0525
anthelmintic agent	10	0.9959	0.9810	0.8000	0.0666	0.0535	0.0833	0.0669
antiallergic agent	63	0.9452	0.8369	0.5556	0.5591	0.0142	0.7198	0.0175
antianginal agent	21	0.9647	0.8739	0.5714	0.2661	0.0379	0.3492	0.0497
antianxiety agent	50	0.9269	0.8365	0.6400	0.5419	0.0146	0.7127	0.0191
antiarrhythmic agent	62	0.9216	0.8034	0.5161	0.4946	0.0143	0.6292	0.0174
antiasthmatic agent	31	0.9717	0.8887	0.5161	0.3810	0.0305	0.4899	0.0391
antibacterial agent	127	0.9557	0.9146	0.7638	0.7375	0.0091	0.9206	0.0084
antibiotic	132	0.9424	0.8753	0.7045	0.6903	0.0079	0.8919	0.0084
anticholesteremic agent	13	0.9959	0.9806	0.6923	0.3161	0.0472	0.4109	0.0614
anticoagulant	10	0.9921	0.9665	0.8000	0.2408	0.0843	0.3010	0.1053
anticonvulsant	60	0.9614	0.9022	0.6500	0.6175	0.0177	0.8123	0.0223
antidepressant	40	0.9570	0.8891	0.6750	0.5374	0.0237	0.7057	0.0299
antidepressant, second-generation	14	0.9768	0.9124	0.7143	0.2453	0.0451	0.3121	0.0573
antidyskinesia agent	26	0.9775	0.9096	0.5769	0.2861	0.0291	0.3704	0.0376
antiemetic agent	48	0.9354	0.7938	0.4375	0.5080	0.0184	0.6502	0.0225
antifungal agent	30	0.9796	0.9184	0.5667	0.3423	0.0310	0.4443	0.0401
antiglaucoma agent	23	0.9680	0.8831	0.6087	0.3360	0.0314	0.4291	0.0401
anti-HIV agent	24	0.9724	0.9077	0.6667	0.4184	0.0384	0.5578	0.0512
antihypertensive agent	112	0.8983	0.7521	0.5357	0.5209	0.0102	0.6630	0.0118
antihypocalcemic agent	12	0.9974	0.9876	0.6667	0.4648	0.0472	0.6197	0.0630
anti-infective agent	212	0.8524	0.7701	0.6132	0.5753	0.0057	0.7384	0.0067
anti-infective agent, local	11	0.9927	0.9659	0.5455	0.1833	0.0434	0.2240	0.0530
anti-infective agent, urinary	13	0.9884	0.9503	0.6923	0.2364	0.0278	0.3074	0.0362
anti-inflammatory agent	102	0.9103	0.8175	0.6176	0.6108	0.0089	0.7923	0.0106
antimalarial agent	18	0.9935	0.9705	0.7222	0.2411	0.0421	0.3096	0.0541
antimanic agent	12	0.9946	0.9758	0.6667	0.2026	0.0456	0.2701	0.0609
antimetabolite	30	0.9739	0.9047	0.6667	0.4735	0.0327	0.6176	0.0427
antimigraine agent	19	0.9535	0.8680	0.5263	0.2620	0.0259	0.3318	0.0328
antimuscarinic agent	33	0.9757	0.9026	0.5152	0.5840	0.0233	0.7404	0.0269
antineoplastic agent	113	0.8604	0.7048	0.4690	0.4475	0.0124	0.5756	0.0153
antineoplastic agent, alkylating	15	0.9766	0.9303	0.6667	0.2857	0.0413	0.3571	0.0516
antineoplastic agent, antimetabolite	14	0.9956	0.9799	0.7143	0.4700	0.0538	0.5982	0.0685
antineoplastic agent, hormonal	19	0.9865	0.9383	0.4737	0.3609	0.0342	0.4565	0.0431
antiobesity agent	12	0.9938	0.9706	0.5000	0.1789	0.0617	0.2386	0.0822
antioxidant	10	0.9901	0.9552	0.6000	0.0028	0.0041	0.0035	0.0051

dx.doi.org/10.1021/ci2002022 |*J. Chem. Inf. Model.* 2012, 52, 134–145

Table 1. Continued

		accuracy			10-fold cross-validation probability values			
effect	*n*	AUC	BEDROC	top hit rate	mean	std	mean 75%	std 75%
antiparkinson agent	30	0.9712	0.8973	0.6000	0.3574	0.0390	0.4635	0.0502
antiprotozoal agent	19	0.9597	0.8748	0.6316	0.1078	0.0384	0.1365	0.0486
antipruritic agent	41	0.9597	0.8790	0.5854	0.5120	0.0181	0.6680	0.0232
antipsychotic	45	0.9639	0.8747	0.5556	0.5776	0.0136	0.7553	0.0172
antipyretic	25	0.9873	0.9529	0.7600	0.6735	0.0266	0.8854	0.0348
antirheumatic agent	18	0.9874	0.9435	0.5556	0.1805	0.0258	0.2321	0.0332
antispasmodic agent	24	0.9610	0.8727	0.5417	0.4753	0.0237	0.6272	0.0310
antitussive	10	0.9941	0.9718	0.5000	0.4403	0.0556	0.5503	0.0695
antiulcer agent	23	0.9788	0.9276	0.6957	0.4788	0.0387	0.6086	0.0489
antiviral agent	45	0.9613	0.8864	0.5556	0.4837	0.0228	0.6373	0.0298
barbiturate	17	0.9998	0.9990	0.8824	0.9913	0.0133	1.0000	0.0000
benzimidazole	12	0.9905	0.9586	0.7500	0.4304	0.0720	0.5737	0.0959
benzodiazepine	25	0.9988	0.9946	0.9200	0.8712	0.0116	0.9999	0.0002
β-lactame antibiotic	56	0.9942	0.9782	0.8750	0.8337	0.0081	0.9961	0.0017
bone density conservation agent	17	0.9837	0.9425	0.6471	0.3985	0.0196	0.5211	0.0256
bronchodilator agent	29	0.9463	0.8367	0.5172	0.3779	0.0197	0.4976	0.0259
calcium channel agent	30	0.9602	0.8899	0.6000	0.3664	0.0281	0.4771	0.0365
calcium channel blocker	28	0.9659	0.9078	0.5714	0.3726	0.0328	0.4960	0.0437
carbohydrate derivative	20	0.9971	0.9867	0.8500	0.5605	0.0371	0.7473	0.0495
cardiotonic agent	14	0.9929	0.9685	0.7857	0.2297	0.0384	0.2924	0.0489
cardiovascular agent	19	0.9894	0.9536	0.6842	0.2954	0.0398	0.3741	0.0504
catecholamine	11	0.9994	0.9970	0.8182	0.7065	0.0571	0.8635	0.0697
cell wall synthesis inhibitor	58	0.9874	0.9600	0.8448	0.8066	0.0097	0.9913	0.0025
central nervous system agent	23	0.9632	0.8724	0.5217	0.3551	0.0215	0.4537	0.0275
central nervous system stimulant	12	0.9922	0.9632	0.5000	0.3344	0.0456	0.4458	0.0608
cephalosporin	32	0.9988	0.9947	0.9063	0.8546	0.0197	0.9874	0.0049
cholinergic agent	42	0.9664	0.8779	0.5476	0.5410	0.0164	0.6957	0.0204
cholinergic antagonist	37	0.9741	0.9007	0.5676	0.5935	0.0199	0.7698	0.0249
cholinesterase inhibitor	13	0.9960	0.9815	0.7692	0.2975	0.0423	0.3867	0.0550
contraceptive agent	13	0.9995	0.9975	0.9231	0.7958	0.0696	0.9553	0.0551
corticosteroid	31	0.9979	0.9903	0.9032	0.8939	0.0170	1.0000	0.0000
corticosteroid, topical	12	0.9971	0.9858	0.7500	0.7688	0.0564	0.9643	0.0430
cyclooxygenase inhibitor	37	0.9892	0.9569	0.8108	0.6931	0.0198	0.9026	0.0208
depressant	37	0.9302	0.8141	0.5405	0.4686	0.0159	0.6189	0.0209
dermatologic agent	16	0.9816	0.9338	0.6875	0.2510	0.0510	0.3344	0.0679
dihydropyridine	10	0.9991	0.9959	0.9000	0.5485	0.0601	0.6856	0.0751
diuretic	29	0.9508	0.8631	0.6552	0.4321	0.0285	0.5695	0.0375
dopamine agent	75	0.9220	0.7922	0.5467	0.5479	0.0109	0.7061	0.0135
dopamine agonist	11	0.9992	0.9962	0.8182	0.1151	0.0334	0.1407	0.0408
dopamine antagonist	45	0.9694	0.8919	0.5778	0.6150	0.0154	0.7942	0.0171
dopamine uptake inhibitor	13	0.9936	0.9697	0.6154	0.1696	0.0542	0.2205	0.0705
ergoline derivative	10	0.9998	0.9992	0.9000	0.6267	0.0440	0.7832	0.0550
ergosterol synthesis inhibitor	12	0.9968	0.9845	0.7500	0.1997	0.0487	0.2639	0.0644
estrogen	11	0.9996	0.9981	0.9091	0.6657	0.0518	0.8127	0.0628
ethanolamine derivative	33	0.9454	0.8295	0.5455	0.3308	0.0255	0.4344	0.0335
fluoroquinolone	12	1.0000	1.0000	1.0000	0.8334	0.0001	1.0000	0.0000
folic acid antagonist	19	0.9871	0.9491	0.7895	0.5675	0.0263	0.7187	0.0333
GABA agent	65	0.9761	0.9253	0.7692	0.6770	0.0156	0.8894	0.0191
gastrointestinal agent	12	0.9675	0.8793	0.6667	0.0549	0.0296	0.0732	0.0394
glucocorticoid	31	0.9979	0.9906	0.9032	0.9208	0.0107	0.9999	0.0001
glutamate receptor antagonist	18	0.9654	0.8847	0.6111	0.2730	0.0503	0.3510	0.0647
guanidine derivative	22	0.9813	0.9331	0.7273	0.4477	0.0284	0.5793	0.0367
histamine agent	73	0.9401	0.8619	0.6438	0.6528	0.0094	0.8370	0.0106

dx.doi.org/10.1021/ci2002022 |*J. Chem. Inf. Model.* 2012, 52, 134–145

Table 1. Continued

		accuracy			10-fold cross-validation probability values			
effect	*n*	AUC	BEDROC	top hit rate	mean	std	mean 75%	std 75%
histamine antagonist	71	0.9399	0.8613	0.6479	0.6659	0.0089	0.8462	0.0099
histamine H1 antagonist	49	0.9671	0.8999	0.6531	0.6505	0.0159	0.8293	0.0175
histamine H1 antagonist, nonsedating	10	0.9991	0.9954	0.8000	0.2928	0.0462	0.3659	0.0577
hormone replacement agent	11	0.9984	0.9925	0.8182	0.3007	0.0666	0.3674	0.0814
hypnotic and/or sedative	63	0.9456	0.8660	0.6984	0.6450	0.0099	0.8432	0.0127
hypoglycemic agent	22	0.9916	0.9631	0.7273	0.4212	0.0276	0.5428	0.0353
imidazole derivative	35	0.9480	0.8544	0.6000	0.4083	0.0240	0.5280	0.0309
immunosuppressive agent	28	0.9555	0.8653	0.6429	0.3290	0.0343	0.4385	0.0457
indole derivative	20	0.9856	0.9387	0.6500	0.2441	0.0333	0.3250	0.0443
muscarinic agent	36	0.9722	0.8888	0.5000	0.5397	0.0218	0.6983	0.0271
muscle relaxant	60	0.9355	0.8181	0.5833	0.4565	0.0140	0.6041	0.0186
muscle relaxant, central	13	0.9941	0.9729	0.6923	0.0636	0.0344	0.0826	0.0448
muscle relaxant, skeletal	35	0.9653	0.8863	0.6286	0.4685	0.0229	0.6072	0.0297
narcotic	22	0.9882	0.9493	0.6364	0.4914	0.0263	0.6359	0.0340
neuroprotective agent	13	0.9684	0.8827	0.4615	0.1309	0.0251	0.1701	0.0327
neurotransmitter uptake inhibitor	42	0.9495	0.8482	0.5714	0.5570	0.0179	0.7239	0.0226
nitro compound	26	0.9929	0.9703	0.8077	0.6340	0.0289	0.8207	0.0366
nonsteroidal anti-inflammatory agent	69	0.9306	0.8094	0.5652	0.5284	0.0122	0.6929	0.0157
norepinephrine reuptake inhibitor	15	0.9965	0.9841	0.8000	0.5640	0.0486	0.7042	0.0604
nucleic acid synthesis inhibitor	80	0.9097	0.8049	0.6250	0.5199	0.0150	0.6903	0.0199
nucleoside or nucleotide	22	0.9995	0.9978	0.9091	0.8197	0.0288	0.9905	0.0098
nucleoside or nucleotide analogue	13	0.9980	0.9903	0.7692	0.3849	0.0160	0.5004	0.0209
opiate agent	31	0.9865	0.9439	0.6452	0.5554	0.0182	0.7173	0.0236
opiate agonist	27	0.9899	0.9558	0.6296	0.5505	0.0241	0.7077	0.0310
opioid	22	0.9869	0.9486	0.7727	0.6335	0.0077	0.8198	0.0100
parasympatholytic	16	0.9743	0.9148	0.6875	0.6170	0.0395	0.8181	0.0515
parasympathomimetic	10	0.9985	0.9926	0.8000	0.2228	0.0711	0.2785	0.0889
penicillin	20	0.9999	0.9996	0.9500	0.7600	0.0433	0.9415	0.0280
phenothiazine	25	0.9958	0.9815	0.7600	0.8811	0.0194	0.9977	0.0018
phosphodiesterase inhibitor	16	0.9927	0.9682	0.6875	0.2308	0.0508	0.3076	0.0678
piperazine derivative	57	0.9766	0.9198	0.6842	0.6495	0.0148	0.8276	0.0164
piperidine derivative	66	0.9508	0.8533	0.6061	0.6133	0.0131	0.7686	0.0152
platelet aggregation inhibitor	16	0.9721	0.9037	0.4375	0.0688	0.0220	0.0916	0.0293
potassium channel agent	18	0.9903	0.9665	0.8889	0.4876	0.0420	0.6250	0.0537
potassium channel blocker	16	0.9850	0.9555	0.8750	0.5137	0.0538	0.6765	0.0691
progestin	12	0.9996	0.9983	0.8333	0.7847	0.0499	0.9731	0.0403
prostaglandin derivative	11	0.9991	0.9955	0.8182	0.5674	0.0576	0.6911	0.0693
protein synthesis inhibitor	32	0.9661	0.9076	0.7500	0.5700	0.0157	0.7598	0.0209
purine derivative	12	0.9981	0.9913	0.8333	0.6334	0.0598	0.8433	0.0788
pyridine derivative	49	0.9266	0.7866	0.4694	0.3310	0.0190	0.4265	0.0241
pyrimidine derivative	17	0.9807	0.9358	0.5882	0.1828	0.0372	0.2390	0.0486
quaternary amine	35	0.9569	0.8842	0.7143	0.4478	0.0296	0.5798	0.0384
quinoline derivative	14	0.9941	0.9745	0.8571	0.4334	0.0551	0.5513	0.0699
quinolone	15	0.9993	0.9955	0.9333	0.7475	0.0315	0.9287	0.0347
respiratory smooth muscle relaxant	14	0.9678	0.9102	0.6429	0.3013	0.0663	0.3835	0.0843
respiratory system agent	41	0.9048	0.7660	0.4634	0.4101	0.0183	0.5413	0.0241
reverse transcriptase inhibitor	14	0.9794	0.9280	0.7143	0.3283	0.0431	0.4178	0.0549
serotonin agent	63	0.9502	0.8412	0.5556	0.5888	0.0144	0.7396	0.0159
serotonin agonist	13	0.9968	0.9850	0.8462	0.3870	0.0495	0.5028	0.0642
serotonin antagonist	31	0.9699	0.9011	0.6774	0.5427	0.0283	0.6903	0.0352
serotonin reuptake inhibitor	21	0.9835	0.9326	0.6190	0.4327	0.0362	0.5649	0.0471
sodium channel blocker	39	0.9282	0.8119	0.5385	0.3807	0.0198	0.4899	0.0255
sodium chloride symporter inhibitor	13	0.9992	0.9962	0.7692	0.5746	0.0243	0.7470	0.0315

dx.doi.org/10.1021/ci2002022 |*J. Chem. Inf. Model.* 2012, 52, 134–145

Table 1. Continued

		accuracy			10-fold cross-validation probability values			
effect	*n*	AUC	BEDROC	top hit rate	mean	std	mean 75%	std 75%
steroidal	73	0.9976	0.9901	0.9178	0.8811	0.0061	0.9998	0.0001
steroidal anti-inflammatory agent	33	0.9991	0.9962	0.9697	0.9334	0.0158	1.0000	0.0000
stimulant	15	0.9900	0.9542	0.5333	0.2236	0.0493	0.2794	0.0616
sulfonamide	78	0.9535	0.8629	0.6282	0.6179	0.0123	0.7867	0.0145
sulfone	17	0.9736	0.9238	0.6471	0.1822	0.0395	0.2382	0.0517
sulfonylurea	11	0.9999	0.9996	0.9091	0.6053	0.0918	0.7388	0.1118
sympatholytic	23	0.9688	0.8940	0.5652	0.3894	0.0408	0.4971	0.0522
sympathomimetic	33	0.9744	0.9029	0.6364	0.5909	0.0133	0.7799	0.0176
tetracycline	10	1.0000	1.0000	1.0000	0.7350	0.0723	0.9065	0.0794
tetrazole derivative	20	0.9898	0.9606	0.8000	0.6422	0.0402	0.8545	0.0535
thiazide	12	0.9995	0.9976	0.9167	0.6056	0.0214	0.8075	0.0285
thiazole	22	0.9936	0.9730	0.8182	0.5156	0.0324	0.6654	0.0415
tocolytic agent	11	0.9888	0.9498	0.5455	0.3296	0.0572	0.4029	0.0699
triazole derivative	16	0.9596	0.8770	0.5625	0.2738	0.0298	0.3650	0.0397
tricyclic antidepressant	14	0.9979	0.9901	0.7857	0.6472	0.0473	0.8205	0.0591
trifluormethyl derivative	32	0.9607	0.8814	0.6563	0.4206	0.0245	0.5535	0.0316
vasoconstrictor	42	0.9495	0.8677	0.6190	0.5603	0.0154	0.7335	0.0200
vasodilator	77	0.8837	0.7389	0.5195	0.4491	0.0135	0.5780	0.0169
2-hydroxy-3-aminopropoxy derivative	21	0.9955	0.9797	0.8095	0.7216	0.0237	0.9373	0.0297

[a] The first column (*n*) lists the number of registered drugs to the given effect. Accuracy (AUC, BEDROC, and top hit rate) and 10-fold cross-validation results (mean and standard deviation of MPV and mean and standard deviation of the upper 75% MPV, respectively) are presented.

pairs, we calculated the probability of each effect for each drug based on the drug's IP by linear discriminant analyses, yielding a classification function for all effects. As shown in Figure 1, each observed IP was plugged into the classification function in order to generate the drug−effect probability matrix.

To quantitatively assess the potential clinical relevance of the drug−effect probability values, we first examined the Receiver Operating Characteristic (ROC) curves (Figure 2) and then performed an independent cross-validation (Figure 3) of our results. ROC analysis characterizes classification performance in terms of sensitivity and specificity of drug−effect classification (see the Supporting Information for the details). ROC curves allow the fine-tuning of the detection threshold in order to optimize for sensitivity and/or specificity. Classification accuracy was characterized by the AUC and BEDROC values. An AUC close to 1, i.e., a ROC that ascends rapidly, indicates high-accuracy classification, while a random guess classification would result in a diagonal ROC yielding an AUC value of 0.5 (see Figure 2A for selected examples). Figure 2B shows the distribution of the AUCs for the entire effect set. A total of 84% of the effects yielded an AUC value larger than 0.95, indicating that an excellent classification was obtained (see Table 1 for the complete list of the studied effects). From another perspective, an effect ROC curve is based upon a list of drugs ordered by descending probability values, regardless of their FDA effect registration. High classification accuracy is obtained if the registered drugs of the given effect appear on the top of the list. If we cut the list at the number of the registered drugs to the given effect, we found that here, on average, 69% of the registered drugs appear (Figure S1, Supporting Information). If we consider this number, more than two-thirds of the registered drugs are in the top 2.6% of the list, since on average 32 out of 1177 drugs belong to an effect (enrichment: 26.54). In order to assess the early recognition problem and calculate a more rigorous measure for classification accuracy, BEDROC scores were also determined at $\alpha = 5$ (Table 1, Figure 2C). The results were similar to the previously determined AUC values: the antineoplastic agent category resulted in the worst but still considerable classification accuracy values (AUC and BEDROC values were 0.860 and 0.705, respectively). For 116 effects out of 177, BEDROC values are above 0.9, suggesting that the DPM method can overcome the early recognition problem. High correlation ($R^2 = 0.962$) was found between AUC and BEDROC values, so the calculation of BEDROC values did not result in a substantially different conclusion about the performance of our method.

To check the validity of the effect classification results from Drug Profile Matching, an independent 10-fold cross-validation was performed and repeated 100 times (Figure 3A). For each effect, we calculated a mean probability value (MPV), i.e., the mean of the calculated probabilities for each drug registered to the given effect. Finally, the mean of the MPVs of the 100-times repeated 10-fold cross-validation experiments were calculated (Table 1). A high mean MPV indicates the method's robustness that is the resistance of the classification system against the loss of information due to the removal of 10% of the molecule entries, when the classification rules are established during the validation. Figure 3B and C show the means of the MPVs for the studied 177 effects and some selected examples. A total of 48.6% of the studied effects are validated by a mean probability value larger than 0.5. (Using a randomized EP list would result in an average probability value of 0.027.) We observed for certain effects that a small number of the registered compounds were validated with low probability, which may reflect the existence of subgroups within the effect categories (Figure S4, Supporting Information). Therefore, we also present the mean probability

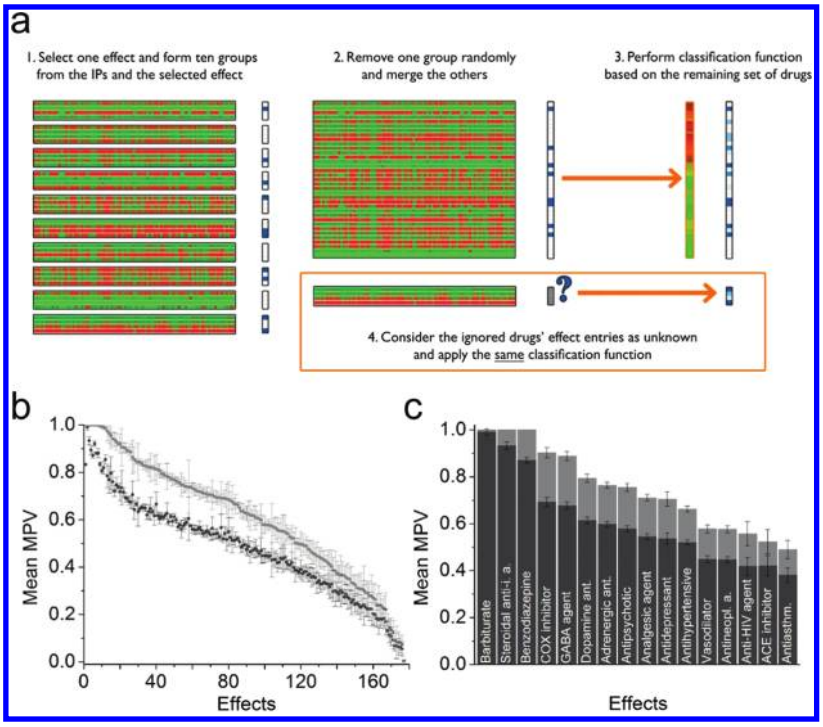


Figure 3. (A) 10-fold cross-validation of a selected effect category. In the first step, the IP matrix and a selected effect are partitioned into 10 groups ("folds"). One group is removed, and the rest are merged in order to produce a classification function on the remaining set of molecules. This function is applied to calculate the classification probabilities of the drugs from the removed group. The same process is repeated for each fold and then performed for each effect. Finally, the whole cross-validation procedure was repeated 100 times. (B) Means of mean probability values for the 177 studied effect categories, obtained from 10-fold cross-validation. Dark dots refer to the mean MPVs of the whole set of drugs registered to the given effect; light gray dots represent the upper 75%, i.e., the subset giving the best 75% calculated probability values. Standard deviations are also plotted. Using a randomized EP list would result in an average probability value of 0.027. (C) Mean MPVs and standard deviations for some selected effect categories. Dark and light gray bars represent the same values as for the previous panel. Abbreviations: anti-i. a., anti-inflammatory agent; ant., antagonist; antineopl. a., antineoplastic agent; antiasthm., antiasthmatic agent.

values for the upper 75% of the drugs (Figure 3B,C). We found that, applying this portion of the drugs, 67% of the effects have a mean probability value above 0.5. We also performed a 3-fold cross-validation which gave similar results: the mean and standard deviation of the MPV values were 0.478 ± 0.031 and 0.419 ± 0.060 for the 10-fold and 3-fold cross-validation, respectively, implying the robustness of the DPM method.

If we examine the mean probabilities of different effect categories, the highest values belong to effects based on a high degree of structural similarity among their registered compounds, as expected. For example, barbiturates, benzodiazepines, and steroidal anti-inflammatory agents result in mean probability values of 0.991, 0.871, and 0.933, respectively. (These categories produce high AUC and BEDROC values as well, see Table 1.) However, effect categories based on common target proteins still show rather high mean probability values (e.g., 0.693 and 0.615 for cyclooxygenase (COX) inhibitors and dopamine antagonists, respectively), despite the fact that these compounds share a low level of chemical similarity. In these cases, the protein set used in the DPM method can be considered as a surrogate creating panel for proteins that are not included in the studied set, a similar phenomenon described in ref 11. Finally, clinical effect categories encompassing an extensive set of drugs with different mechanisms of action also could be characterized by fairly high mean probability values (e.g., 0.578, 0.537, and 0.521 for antipsychotics, antidepressants, and antihypertensive agents, respectively; Figure 3C, Table 1). Many of these categories raise difficulties in conventional prediction approaches. However, they are of crucial practical importance; therefore, these results point to the strength of the DPM method.

We also examined the effect of protein set size on the classification accuracy. Protein sets containing 1, 5, 10, 40, 70, 100, and 130 randomly selected proteins were separated from the complete protein set, and the DPM procedure was applied to them, resulting in a series of effect AUC values for each protein set size. Three independent runs were carried out at each data point. The means and the standard deviations of the resulted AUC values are displayed in Table S3 (except for protein set sizes 1 and 5; Supporting Information). The low values of the standard deviations suggest that the composition of the sets does not affect the AUC values significantly. On the other hand, the increasing number of the applied proteins saturates the AUC values, i.e., the classification accuracy. On the basis of a hyperbolic fitting to the means of the AUC values of an effect at different protein sizes, the maximal obtainable AUC (i.e., the maximum value of the extrapolated hyperbola) and the number of proteins required to reach 90% of this level of AUC can be calculated (Table S4, Supporting Information). The theoretical limit of the AUC is 1.0; therefore it should be noted that the maximal obtainable AUC is linked to a hypothetic protein set of the same diversity as our basic 149-element set. Figure 4A and B display two representative curve fits, while Figure 4C shows the distribution of the number of required

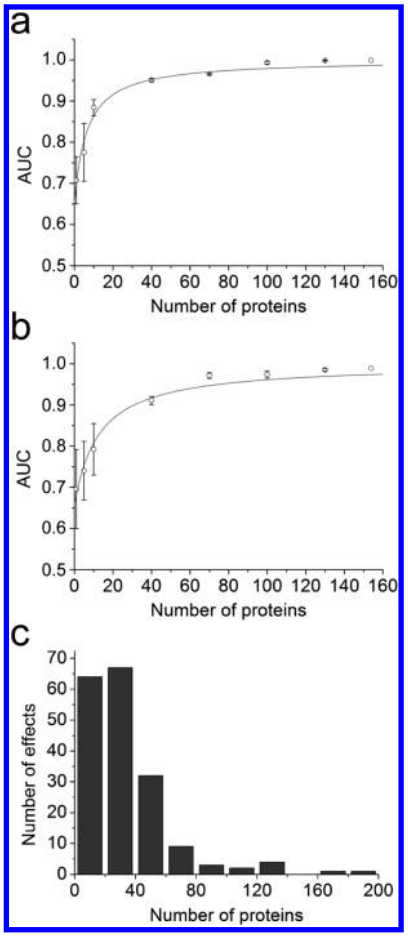


Figure 4. (A and B) Protein set size evaluation results for the angiotensin-converting enzyme inhibitory and the cyclooxygenase inhibitory effect categories, respectively. Hyperboles are fitted to the data points representing the mean and the standard deviation of the AUC values based on a set of 1, 5, 10, 40, 70, 100, and 130 proteins, each performed three times. AUC values obtained using the complete protein set (149 proteins) are also shown in the figure for both effects. The theoretical maximum of AUC is 1.0. (C) Distribution of the number of proteins required to reach 90% of the maximal obtainable accuracy for each effect. More than 90% of the studied 177 effects can be sufficiently classified on the basis of the protein set used for IP generation.

proteins, yielding 90% of the calculated maximal obtainable AUC for each effect category. For 176 of the studied 177 effects, the classification functions based on the complete protein set are sufficient to reach 90% of the maximal obtainable AUC. The remaining one effect, antihypertensive agents contain diverse subcategories with different mechanisms of action. For most of the structural categories, IPs based on 12 proteins are sufficient for effective classification. Target-focused and therapeutic categories also yielded generally low protein size parameters (e.g., 16 and 23 for angiotensin-converting enzyme inhibitors and antidepressants, respectively). These values are comparable with the optimized reference panel of 18 proteins for creating target surrogates described by Kauvar et al.[11] However, antianginal agents and nonsteroidal anti-inflammatory agents require 65 and 100 proteins, respectively. Therefore, we conclude that the relevant effect categories can be appropriately classified with the original protein set used for IP generation.

In order to exclude any artifacts that might originate from the scoring method, we also studied the effect of the applied scoring function on the results. The same analysis procedure presented above was repeated using the AutoDock4 native scoring function and Glide docking with GlideScore scoring, and in all three cases the prediction power did not change significantly (data not shown).

Using the resulting classification functions, probability values were assigned for each drug−effect pair in our data set. For many drugs, a number of unregistered effects were detected with high probability. These "false positive" hits can be indicative of hidden effects which potentially could be used for new drug effect predictions. However, the overall high AUC values make it difficult to judge the actual performance of the DPM system for different therapeutic categories. Therefore, in order to evaluate the predictive power of the DPM method for a given effect category, one must consider the AUC/BEDROC values and the MPVs as well. Effects that produce outstanding values for all of these categories can be accepted as highly predictable categories, e.g., adrenergic β antagonists and antibiotics (AUC/BEDROC/MPV values of 0.991/0.961//0.717, 0.942/0.875/0.690, respectively), as well as the structure-based classes. High AUC/BEDROC values and medium MPVs suggest medium reproducibility (e.g., cholinergic and muscarinic agents with AUC/BEDROC/MPV values of 0.966/0.878/0.541 and 0.972/0.889/0.540, respectively), while low MPVs refer to poorly reproducible effect categories with low predictive power. Two typical examples are platelet aggregation inhibitors and gastrointestinal agents (their MPVs are 0.069 and 0.055, respectively). The mechanisms of actions within these groups are too different for effective prediction; redistribution of the registered drugs can result in better classification in the future (see Conclusions).

In sum, the Drug Profile Matching method is a robust and highly accurate approach that calculates the EPs of drugs solely on the basis of their complex binding properties. The obtained AUC and mean probability values pinpoint the strong relationship between EPs and IPs.

CONCLUSIONS

Polypharmacology is a newly emerging approach which reflects the high complexity of the mechanism of actions of drugs. This aspect of pharmacology has not been fully exploited in drug development. Consequently, the entire effect profiles of drugs and drug candidates have remained unrevealed. We hypothesized that complex molecular feature sets of drugs correlate with the known part of EPs and may therefore provide predictive power to reveal the entire EPs of drugs.

In the present study, we collected the structural data and registered effect profiles of all small-molecule drugs. Interactions with a series of nontarget protein sites of each drug were calculated, and an IP matrix was constructed. Statistical analyses unveiled a strong correlation between the EPs and IPs, and this relationship was confirmed by independent validation. These findings allowed us to develop a robust and systematic effect prediction method, named Drug Profile Matching.

To our knowledge, no attempt has been made previously to relate large-scale, *in silico* generated affinity fingerprints and pharmacological effects, not only target binding affinity. According to our starting hypothesis, a reference panel of proteins must discriminate between a wide range of compounds in order to be

an effective surface for affinity profiling. We show here that this discrimination has a strong predictive power for clinical effect profiles. However, two critical points about the applied methodology could arise, i.e., the usage of *in silico* calculated scoring values instead of experimentally determined binding constants and the low overall correlation between docking scores and binding constants. First, *in vitro* gained binding affinity values suffer from some serious uncertainty due to the possibility of nonspecific binding of the compound on the receptor and neglecting the information originated from the weak interactions in the immeasurable range. In contrast, in the presented DPM method, these limitations obviously do not exist. Second, the widely discussed problem of reliability of the calculated scores can be overcome by using and comparing different docking/scoring methods as suggested in the recent literature.[26−28] We found that the predictive power of DPM is not influenced significantly by the applied scoring functions. Furthermore, docking scores in DPM are used as descriptor elements of the interaction potency of a compound and not as calculated affinity values that are compared to actual binding affinities. The uniform treatment of the compounds on the discriminator surface is more important in the DPM method than the individual docking scores that are generally unable to determine the measured ligand binding affinity for the given protein.[28] Due to the necessity of uniform treatment, conformational changes in the proteins during ligand binding were banned in this study in order to apply the same discriminator surface (active sites of the proteins) for each drug.

Unlike other similarity-based approaches,[8,9] no direct topological similarity information on drug molecules is involved in the DPM method; therefore, our approach is able to detect EP similarities even in the case of limited structural similarity between compounds. Briem and Kuntz described in an early work that two-dimensional structural similarity methods resulted in better bioactivity prediction power compared to a docking-based interaction fingerprint due to the fact that rigid conformers of ligands were docked to a very limited number of proteins (8).[29] In contrast, we found that the 2D and 3D structural information of drugs used in previous approaches yielded limited EP-prediction power compared to the IP-based DPM method (data not shown). IPs represent binding potencies of drug molecules to protein surfaces, including weak interactions. Binding potency is an essential feature of drugs because, in organisms, drugs may act on series of strong and weak binding partners which play important roles in the mechanism of actions and could be considered as a key factor in polypharmacology.

The DPM method can be improved at many points. As we presented, several inherently diverse effect categories are weakly predictable, but this issue can be expectedly solved by creating more cohesive subgroups based on the individual cross-validation probability values of the drugs (see Figure S4, Supporting Information), e.g., pharmacological effects like "antihypertensive agent" could be handled by the sum of several target-based subgroups. DPM could be further improved by introducing ADME properties into the effect profile matrix. Moreover, different discriminator surfaces can be used for specific therapeutic categories: protein sets that possess a larger discriminative effect on a specific effect group than the protein set used here for general EP prediction.[12] Furthermore, an artificial discriminator surface could be designed and tested in DPM in order to determine the minimum level of complexity of these surfaces required for effective predictions. In a future investigation, it might be an interesting question whether introducing water molecules in the docking procedure increases the predictive power of DPM. Finally, nonlinear discrimination functions might also improve the IP-based effect profile prediction.

Besides network biology, our results can be interpreted from the viewpoint of pharmacochemistry as well. In this regard, the IP of a drug is a representative of a complex chemical feature, the 3D pharmacophore of the small-molecule compound. The observed high level of correlation between IPs and EPs can be originated in the common pharmacophore required to yield the physiological effect through a given mechanism of action. This theory does not contradict the network point of view; on the contrary, it emphasizes the importance of complex feature sets that are required for effect prediction.

The Drug Profile Matching method may be applicable in a number of ways due to the ability to relate complex interaction profiles of molecules with their clinical and pharmacological profiles. First and foremost, it offers an opportunity for systematic and rapid screening of approved drugs in order to discover new therapeutic indications and safety risks. Moreover, it can be a valuable aid in the prediction of the pharmacological effect profiles of drug candidate molecules with high probability, thereby offering a novel approach for lead molecule design and optimization as well. As shown above, the good predictive power of the method holds out the promise for its use with marketed drugs or as a preclinical screen, bringing substantial improvement in the efficacy of future drug development and expediting the development process from drug discovery to marketing.

■ ASSOCIATED CONTENT

Supporting Information. Figure S1 shows the distribution of the top hit rate values among the studied effects. Figure S2 depicts the distribution of the number of registered drugs to the studied 177 effects. Figure S3 summarizes the method of the effect prediction. Figure S4 shows representative probability value curves, the bases of mean probability value calculation. Tables S1 and S2 list the applied small-molecule drugs and the proteins, respectively. Tables S3 and S4 contain information on the protein set size evaluation. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

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

■ ACKNOWLEDGMENT

This work has been supported by the National Development Agency and the European Union (European Regional Development Fund), under the aegis of New Hungary Development Plan (GOP-1.1.1-08/1-2009-0021) and the National Technology Programme (NTP TECH_08_A1/2-2008-0106). The work has also been supported by the European Union and the European Social Fund under the grant agreement no. TÁMOP 4.2.1./B-09/KMR-2010-0003. C.H. is thankful for a János Bolyai Research Scholarship provided by the Hungarian Academy of Sciences.

■ ABBREVIATIONS:

ADME, adsorption−distribution−metabolism−elimination; AUC, area under the curve; BEDROC, Boltzmann-enhanced discrimination of ROC; CCA, canonical correlation analysis; EP, effect profile; FDA, Food and Drug Administration; FPR, false positive rate; IP, interaction profile; LDA, linear discriminant analysis; PDB, Protein Data Bank; ROC, receiver operating characteristic; TPR, true positive rate

■ REFERENCES

(1) Hopkins, A. L. Network pharmacology. *Nat. Biotechnol.* **2007**, *25*, 1110–1111.

(2) Hopkins, A. L.; Mason, J. S.; Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **2006**, *16*, 127–136.

(3) Metz, J. T.; Hajduk, P. J. Rational approaches to targeted polypharmacology: creating and navigating protein-ligand interaction networks. *Curr. Opin. Chem. Biol. 14*, 498-504.

(4) Pujol, A.; Mosca, R.; Farres, J.; Aloy, P. Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol. Sci.* **2010**, *31*, 115–123.

(5) Roth, B. L.; Sheffler, D. J.; Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discovery* **2004**, *3*, 353–359.

(6) Ashburn, T. T.; Thor, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discovery* **2004**, *3*, 673–683.

(7) Merino, A.; Bronowska, A. K.; Jackson, D. B.; Cahill, D. J. Drug profiling: knowing where it hits. *Drug Discovery Today* **2010**, *15*, 749–756.

(8) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.

(9) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175–181.

(10) Campillos, M.; Kuhn, M.; Gavin, A. C.; Jensen, L. J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321*, 263–266.

(11) 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.

(12) Kauvar, L. M.; Villar, H. O.; Sportsman, J. R.; Higgins, D. L.; Schmidt, D. E. Protein affinity map of chemical space. *J. Chromatogr., B.: Biomed. Sci. Appl* **1998**, *715*, 93–102.

(13) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nat. Chem. Biol.* **2005**, *1*, 389–397.

(14) 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.

(15) 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.

(16) 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.

(17) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–906.

(18) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(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 semi-empirical 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) Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* **2007**, *1*, 7.

(23) 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.

(24) Park, H.; Lee, J.; Lee, S. Critical assessment of the automated AutoDock as a new docking tool for virtual screening. *Proteins* **2006**, *65*, 549–554.

(25) 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.

(26) Cole, J. C.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins* **2005**, *60*, 325–332.

(27) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153* (Suppl 1), S7–26.

(28) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(29) Briem, H.; Kuntz, I. D. Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.* **1996**, *39*, 3401–3408.