molecular pharmaceutics

Article

pubs.acs.org/molecularpharmaceutics

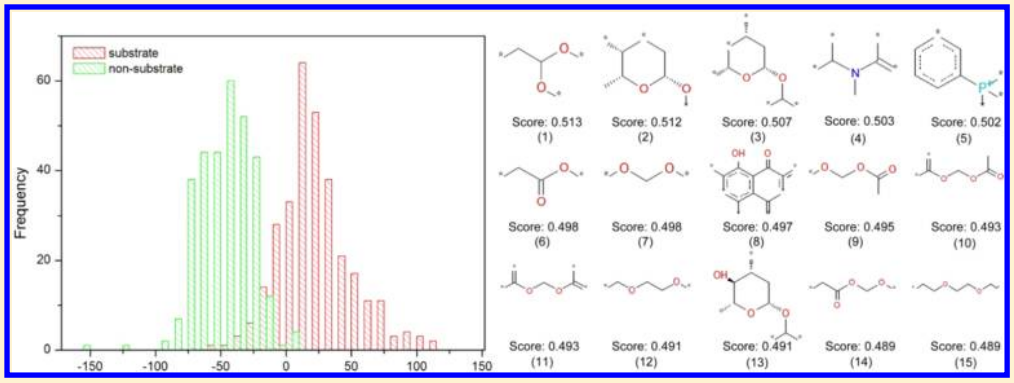
ADMET Evaluation in Drug Discovery. 13. Development of *in Silico* Prediction Models for P-Glycoprotein Substrates

Dan Li,[†] Lei Chen,[‡] Youyong Li,[‡] Sheng Tian,[‡] Huiyong Sun,[‡] and Tingjun Hou*[,†,‡]

[†]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China

[‡]Institute of Functional Nano & Soft Materials (FUNSOM) and Collaborative Innovation Center of Suzhou Nano Science and Technology, Soochow University, Suzhou, Jiangsu 215123, China

Supporting Information



ABSTRACT: P-glycoprotein (P-gp) actively transports a wide variety of chemically diverse compounds out of cells. It is highly associated with the ADMET properties of drugs and drug candidates and, moreover, plays a major role in the multidrug resistance (MDR) phenomenon, which leads to the failure of chemotherapy in cancer treatments. Therefore, the recognition of potential P-gp substrates at the early stages of the drug discovery process is quite important. Here, we compiled an extensive data set containing 423 P-gp substrates and 399 nonsubstrates, which is the largest P-gp substrate/nonsubstrate data set yet published. Comparison of the distributions of eight important physicochemical properties for the substrates and nonsubstrates reveals that molecular weight and molecular solubility are the informative attributes differentiating P-gp substrates from nonsubstrates. Examination of the distributions of eight physicochemical properties for 735 P-gp inhibitors and 423 substrates gives the fact that inhibitors are significantly more hydrophobic than substrates while substrates tend to have more H-bond donors than inhibitors. Then, the classification models based on simple molecular properties, topological descriptors, and molecular fingerprints were developed using the naive Bayesian classification technique. The best naive Bayesian classifier yields a Matthews correlation coefficient of 0.824 and a prediction accuracy of 91.2% for the training set from a 5-fold cross-validation procedure, and a Matthews correlation coefficient of 0.667 and a prediction accuracy of 83.5% for the test set containing 200 molecules. Analysis of the important structural fragments given by the Bayesian classifier shows that the essential H-bond acceptors arranged in distinct spatial patterns and flexibility are quite essential for P-gp substrate-likeness, which affords a deeper understanding on the molecular basis of substrate/P-gp interaction. Finally, the reasons for mispredictions were discussed. It turns out that the presented classifier could be used as a reliable virtual screening tool for identifying potential substrates of P-gp.

KEYWORDS: *P-glycoprotein, substrates, naive Bayesian classification, ADMET, ADME, fingerprint*

INTRODUCTION

ATP-binding cassette (ABC) transporter proteins transport various molecules across the plasma membrane, as well as the intracellular membranes of endoplasmic reticulum, peroxisome, and mitochondria with the energy provided by ATP hydrolysis.[1−3] P-glycoprotein (P-gp) is the best-studied member of the ABC superfamily of transporter proteins and is a product of the ABCB1 gene. P-gp is preferentially expressed in normal tissues important for the absorption (intestinal epithelium), distribution (e.g., endothelial capillaries of the brain comprising the blood−brain barrier), and elimination (hepatocytes, renal proximal tubular cells, and adrenal gland) of drugs and drug candidates.[4] When several drugs are coadministered, the interaction between P-gp and a coadministered drug may change the pharmacokinetics and/or the pharmacodynamics of another drug, thus leading to unwanted drug−drug interactions or even undesired side effects.[5] Furthermore, P-gp is usually overexpressed in cancer cells and can transport many structurally unrelated drugs out of the

Received: July 31, 2013
Revised: January 14, 2014
Accepted: February 5, 2014
Published: February 5, 2014

cancer cells, resulting in the multidrug resistance (MDR) phenomenon that is the major cause of the failure of chemotherapy in cancer treatments.[6]

The P-gp (ABCB1) gene encodes 1280 amino acids arranged into a pseudo symmetrical heterodimer, each consisting of two bundles of six transmembrane (TM) helices and two cytosolic nucleotide binding domains (NBDs). Recently, the X-ray structure of the *apo* murine P-gp and two X-ray structures of P-gp in complex with cyclopeptidic inhibitors were resolved.[7] However, the binding mechanisms of P-gp substrates/inhibitors have not yet been fully understood due to the fact that P-gp possesses multiple binding sites for the substrate/inhibitor binding.[7]

A variety of *in vitro* assays, such as monolayer efflux, ATPase activity, and rhodamine-123/calcein-AM fluorescence assays, have been developed to study the binding of substrates to P-gp.[8] Each of these assays has its advantages but also intrinsic disadvantages. For example, the ATPase and calcein-AM assays do not directly measure transport and only give indirect evidence for P-gp substrate identification. Moreover, the experimental assays are expensive and time-consuming. Therefore, *in silico* models for identifying P-gp substrates have been recognized as valuable tools in both virtual screening and rational drug design.[9−12] Earlier attempts have been made to develop a set of relatively simple rules on the basis of molecular descriptors and/or structural features to characterize P-gp substrate specificity. For instance, Ford and Hait observed that hydrophobic compounds, with a molecular weight of 300−2000, are more likely to be P-gp substrates.[13] Gleeson et al. suggested that neutral or basic molecules with MW > 400 and/or log*P* > 4 are preferentially transported by P-gp.[14] Likewise, other similar studies have illustrated that physicochemical properties like hydrogen-bonding ability, molecular weight, and lipophilicity are correlated with P-gp substrate-likeness.[15,16] To give more accurate prediction of P-gp substrates and nonsubstrates, a variety of statistical techniques and machine learning approaches, including multiple linear regression (MLR),[17] partial least-squares discriminant analysis (PLSD),[18] linear discriminant analysis (LDA),[19,20] random forest (RF),[21] support vector machine (SVM),[22−24] and Kohonen self-organizing map (SOM),[25] have been furthermore employed to develop computational models.[9]

It is well-known that a data set with both high quality and quantity is crucial to build up highly reliable theoretical models. Nevertheless, the public data sets for building the prediction models of P-gp substrates used to be quite small. Gombar et al. in 2004 compiled a data set of 98 molecules containing 32 nonsubstrates and 66 substrates identified by *in vitro* monolayer efflux assays.[20] Penzotti et al. in 2002 reported a data set of 195 compounds, which include 108 substrates and 87 nonsubstrates.[26] Xue et al. in 2004 assembled a data set of 116 substrates and 85 nonsubstrates.[24] Wang et al. recently reported a rather large data set of 206 P-gp substrates and 126 nonsubstrates.[23] In our study, in order to develop prediction models with high reliability, we compiled an extensive data set of 822 compounds, which are categorized into 423 substrates and 399 nonsubstrates. To our knowledge, the data set reported here is much larger than those used in previous studies.

Subsequently, based on the comprehensive data set, the distributions of eight important physicochemical properties for 423 P-gp substrates and 399 nonsubstrates were compared, and the performance of each property for distinguishing substrates from nonsubstrates was discussed. Then the distributions of eight physicochemical properties for 423 P-gp substrates and 735 P-gp inhibitors were compared. Finally, the naive Bayesian classifiers based on molecular properties, topological descriptors, and structural fingerprints were developed and validated on an external test set of 200 molecules. Furthermore, we tracked back and discussed the important fragments given by the best Bayesian classifier for substrate predictions.

METHODS AND MATERIALS

1. Collection of P-gp Substrates and Nonsubstrates.

The data set of 822 nonduplicated molecules was compiled from 517 published papers, including 423 P-gp substrates and 399 nonsubstrates. The experimental data from *in vitro* assays were carefully examined during data collection. Several *in vitro* assays have been used to classify compounds as P-gp substrates, and the most representative approaches are the monolayer efflux, ATPase activity, and rhodamine-123 or calcein-AM fluorescence assays.[27,28] The monolayer efflux assay measures the ratio of basolateral-to-apical (B → A) permeability versus apical-to-basolateral (A → B) permeability of a molecule and is regarded as the standard way to identify P-gp substrates. The ATPase and rhodamine-123/calcein-AM assays do not directly measure transport and give only indirect evidence for P-gp substrate identification. In the Polli study,[27] a molecule is assumed to interact with P-gp (positive) if any of the following three criteria is satisfied: (1) the ratio of B → A permeability versus A → B permeability (BA/AB ratio) given by the monolayer efflux assay is >2.0, (2) GF120918 response providing by the calcein-AM fluorescence assay is >10%, and (3) the ATPase ratio from the ATPase activity assay is >2.0. That is, a molecule is identified as a transported substrate when it is positive in the efflux assay regardless of response in the other assays, a nontransported substrate when it is negative in the efflux assay while positive in ATPase activity and/or calcein-AM fluorescence assays, or a nonsubstrate when it is negative in all three assays. In our study, the BA/AB ratio larger than 1.5 was employed as the most important criterion for P-gp substrate identification according to the suggestion of Schwab.[28] However, all nontransported substrates identified by Polli et al. were classified as nonsubstrates in our study.[27] Moreover, the multidrug resistance (MDR) activities measured in different cell lines were also used as a criterion for P-gp substrate identification. For example, in Ramu's work,[29] the ED_{50} values (the drug concentration effective in inhibiting cell growth by 50%) for each compound were measured in the presence of each drug with/without 200 nM adriamycin (ADR) with P388 murine leukemia cells that are resistant to ADR. A molecule was classified as a P-gp substrate if RF (ED_{50} without ADR/ED_{50} with ADR) is >4.2, and as a P-gp nonsubstrate if RF is ≤2.0 in line with Estrada et al.[30] In Scala's study, a compound was classified as a P-gp substrate when the cytotoxicity to SW620 Ad300 cells was increased ≥4-fold by the addition of cyclosporine A (CsA).[31] In Tang-Wai's study, a minimal degree of 3−4-fold resistance measured in LR73 cells was believed to classify a compound as a P-gp substrate.[32] Certainly, the division of compounds into substrate or nonsubstrate class regarding MDR activity is somewhat arbitrary. It needs to be emphasized that careful curation of experimental data is extremely necessary because the category of a compound reported by different studies may be not consistent. For example, *in vitro* studies indicated that meperidine is a P-gp substrate while *in vivo* brain uptake studies and antinociceptive

studies showed that it is a nonsubstrate.[33] Similar to 5-fluorouracil, in Takara's study 5-fluorouracil is most likely a P-gp substrate because its relative resistance (IC_{50} in Hvr100-6 cells/IC_{50} in HeLa cells) is 4.02,[34] but it was thought to be a nonsubstrate due to its equal sensitivity to K562/adriamycin (ADM) and K562 cells as reported by Naito and his co-workers.[35] Therefore, if the experimental data concerning the P-gp affinity status for the same molecule from variant studies are conflicting, the consensus class given by most studies was adopted in our study.

The structures of most compounds were retrieved from the PKKB database,[36] and the new structures not available in PKKB were built using the Sybyl molecular simulation package.[37] Each molecule was optimized by molecular mechanics (MM) with the MMFF94 force field.[38] All molecules were saved into a MACCS sdf file. The whole data set was split into a 622-compound training set and a 200-compound test set. The 200 compounds in the independent test set were extracted from the whole data set utilizing the *Find Diverse Molecules* protocol in Discovery Studio 2.5 (DS2.5),[39] which makes sure that the compounds have the largest diversity evaluated by the Tanimoto distance based on FCFP_4 fingerprints.[39] The remaining 622 compounds were included in the training set for model establishment. The training set contains 313 substrates and 309 nonsubstrates, and the test set consists of 110 substrates and 90 nonsubstrates. The whole data set and the complete list of the literature can be accessed from the supporting Web site: http://cadd.suda.edu.cn/admet.

2. Calculations of Molecular Descriptors and Fingerprints. In total, 13 simple physicochemical properties widely adopted in ADME predictions[40−42] and 43 topological descriptors afforded by DS2.5 were used for model development. The 13 physicochemical properties are octanol−water partitioning coefficient (Alog*P*),[43] the apparent partition coefficient at pH = 7.4 (log*D*) based on the Csizmadia's method, molecular solubility (log*S*) out from the Tetko's multiple linear regression model,[44] molecular weight (MW), the number of hydrogen bond donors (n_{HBD}), the number of hydrogen bond acceptors (n_{HBA}), the number of rotatable bonds (n_{rot}), the number of rings (n_R), the number of aromatic rings (n_{AR}), the sum of oxygen and nitrogen atoms (n_{O+N}), polar surface area (PSA), fractional polar surface area (FPSA), and surface area (SA). The 43 topological descriptors include Wiener index,[45] Zagreb index,[46] 2 Balaban indices,[47] 6 subgraph counts descriptors,[48] 7 Kappa shape indices,[49] 12 Kier and Hall connectivity indices,[48] and 14 graph-theoretical infocontent descriptors.[46] All these descriptors were calculated using DS2.5.[39]

In addition, a variety of molecular fingerprint sets, including the SciTegic extended-connectivity fingerprint sets (FCFC, ECFC, and LCFC) and Daylight-style path-based fingerprint sets (FPFC, EPFC, and LPFC), were employed as the descriptors in model building to characterize the substructural features of the studied molecules.[50] These fingerprints have been described in detail in the previous studies.[50,51] For each fingerprint class, the fourth character is followed by an underscore and a number to represent the maximum distance in generating the fingerprint. For extended-connectivity fingerprints, the number is the maximum diameter (in bond lengths) of the largest structure represented by the fingerprint or the maximum length of the path, and for path fingerprints, it is the maximum length of the path. In our study, one of the three different maximum distances, 6, 8, and 10, was used to generate the fingerprints. Compared with the fingerprints coming from the predefined functional groups, the extended-connectivity and path-based fingerprints do not need to be preselected or predefined since they are generated directly from the molecules. The fingerprints were generated by DS2.5.[39]

3. Naive Bayesian Classifiers. The naive Bayesian categorization technique was used to develop the classifiers to discriminate between P-gp substrates and nonsubstrates.[52] In our study, each compound can be categorized into the substrate (+) or nonsubstrate (−) class, and a vector $\mathbf{f} = \langle f_1, f_2, ..., f_n \rangle$, where $f_1, f_2, ..., f_n$ are the calculated values for the variables (molecular properties, topological descriptors, and fingerprints) $f_1, f_2, ..., f_n$. According to Bayes's theorem, the following equation (eq 1) was derived:

$$p(C|F_1, F_2, ..., F_n) = \frac{p(C)p(F_1, ..., F_n|C)}{p(F_1, ..., F_n)} \quad (1)$$

where C refers to a compound's class, $p(C|f_1, f_2, ..., f_n)$ is the posterior probability of the compound class, $p(C)$ is the prior probability induced from the training set, $p(f_1, ..., f_n|C)$ is the probability that a compound has certain descriptors given that it is a substrate or nonsubstrate, and $p(f_1, ..., f_n)$ is the marginal probability that the descriptors will occur in the data set. The three probabilities on the right side of eq 1 can be learned from the training set, and the mathematical details for naive Bayesian categorization have been described.[51−53] Naive Bayesian categorization only requires a small amount of training data to estimate the parameters (means and variances of the variables), and no tuning parameters are required beyond the selection of the input descriptors from which to learn. Furthermore, it can process large amounts of data, can learn fast, and is tolerant of random noise. The naive Bayesian classifiers were developed using the *Create Bayesian Model protocol* in DS2.5.[39]

4. Validating the Prediction Accuracies of the Bayesian Classifiers. Each classifier was assessed by the following parameters: true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), sensitivity (SE), specificity (SP), global accuracy (GA), and Matthews correlation coefficient (C).

$$SE = \frac{TP}{TP + FN} \quad (2)$$

$$SP = \frac{TN}{TN + FP} \quad (3)$$

$$GA = \frac{TP + TN}{TP + TN + FN + FP} \quad (4)$$

$$C = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (5)$$

The validation of each classifier was carried out by a 5-fold cross-validation test. Additionally, the actual prediction power of each classifier was assessed with an independent external test set containing 200 compounds.

RESULTS AND DISCUSSION

1. Analysis of Property Distributions for P-gp Substrates and Nonsubstrates. Previous studies have shown that a variety of molecular physicochemical properties, such as hydrogen-bonding ability, molecular weight, and

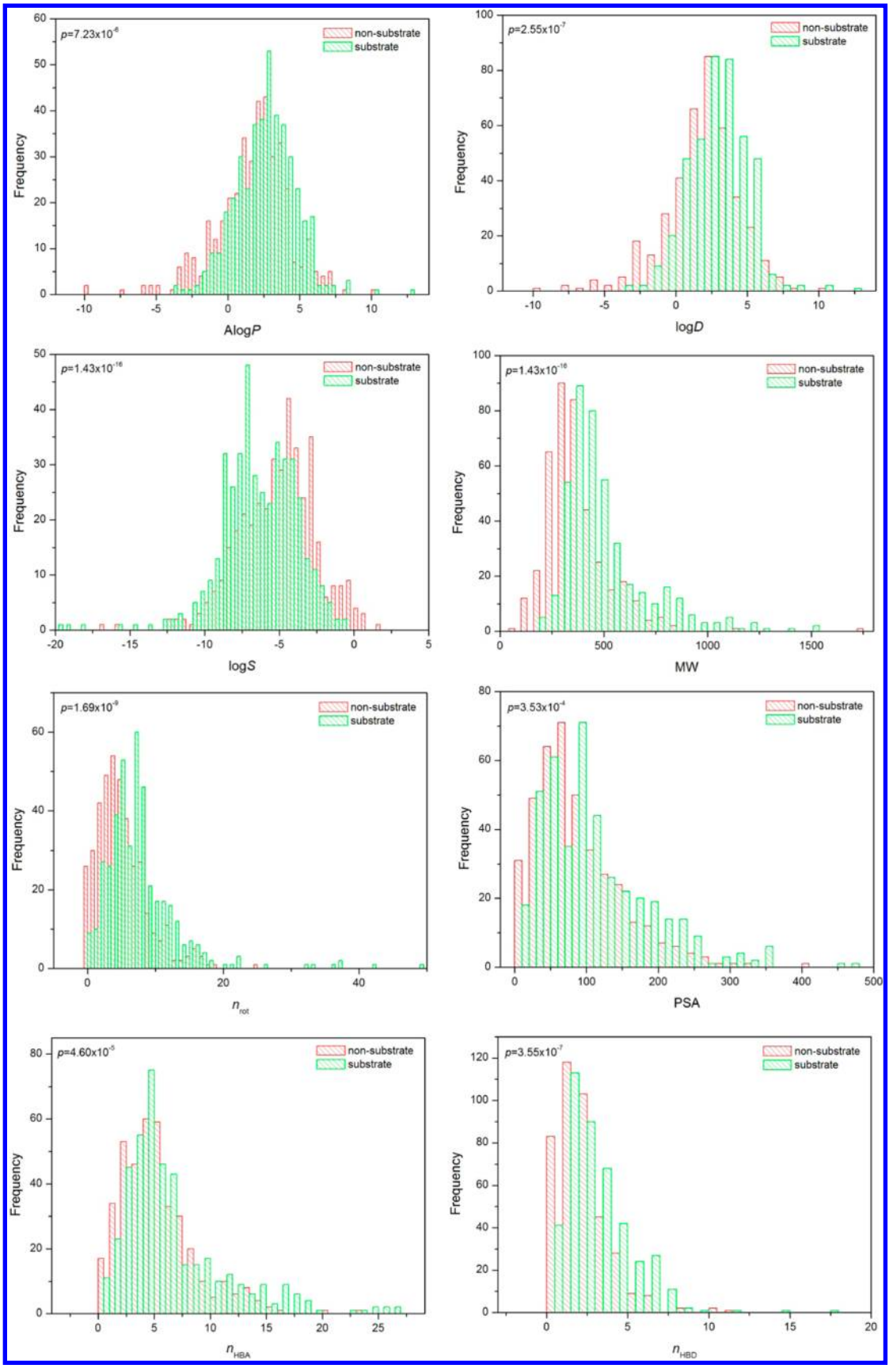


Figure 1. Distributions of eight molecular properties, including AlogP, logD, logS, MW, PSA, n_{rot}, n_{HBD}, and n_{HDA}, for the substrate and nonsubstrate classes.

dx.doi.org/10.1021/mp400450m | Mol. Pharmaceutics 2014, 11, 716−726

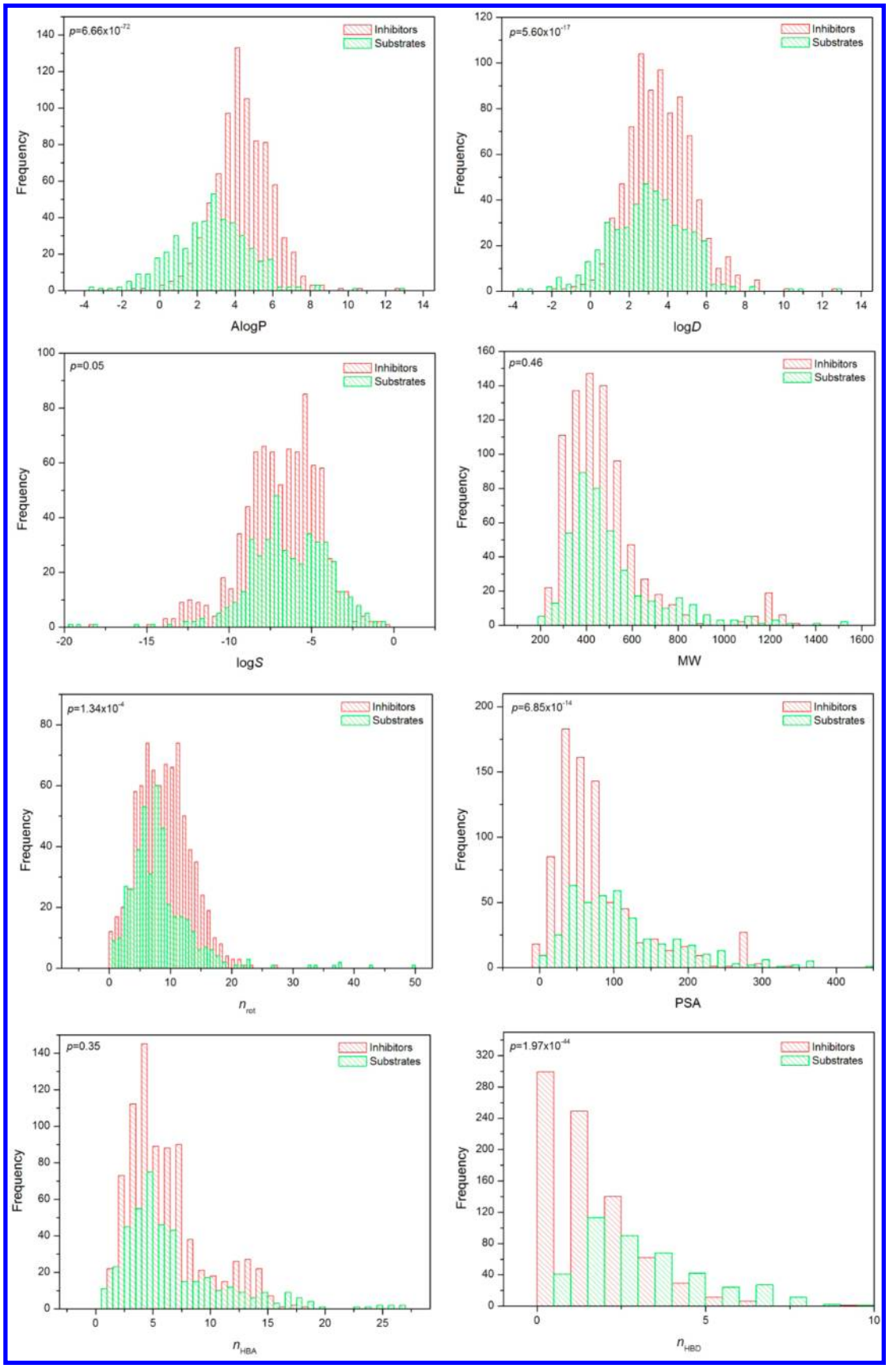


Figure 2. Distributions of eight molecular properties, including Alog*P*, log*D*, log*S*, MW, PSA, n_{rot}, n_{HBD} and n_{HDA}, for the substrates and inhibitors.

hydrophobicity, are closely related to the P-gp substrate-likeness.[13,15,16] Here, we examined the distributions of eight important physicochemical properties (MW, log*S*, log*D*, Alog*P*, PSA, n_{HBA}, n_{HBD}, and n_{rot}) for the P-gp substrates and

dx.doi.org/10.1021/mp400450m | *Mol. Pharmaceutics* 2014, 11, 716−726

nonsubstrates in the data set (Figure 1). For each property, the difference between the means of the two distributions was evaluated by the Student's *t*-test.

In the eight analyzed properties, Alog*P*, log*D*, and log*S* are related to the lipophilicity of a molecule. The mean values of Alog*P* for the 423 substrates and 399 nonsubstrates are 2.55 and 1.79 respectively, and those of log*D* are 2.83 and 1.97 to each. In other words, more hydrophobic molecules have more chance to be P-gp substrates. The mean values of log*S* for the substrates and nonsubstrates are −6.49 and −4.97 individually, implying that the nonsubstrates of P-gp tend to be more soluble. Interestingly, as shown in Figure 1, the distributions of log*S* demonstrate more significant dissimilarity between the compound classes than those of Alog*P* and log*D* whereas the *p*-value associated with the distribution difference of log*S* for the substrates and nonsubstrates is 1.43×10^{-16}. The distributions of log*S* for the substrates and nonsubstrates is obviously more separated than those of Alog*P* or log*D*. Indubitably, log*S* is a more informative attribute than Alog*P* or log*D* for the differentiation between substrates and nonsubstrates, though the two distributions of log*S* are still strongly overlapped. Our results are interesting because it appears that P-gp binding is not directly connected with solubility. However, solubility is also an important indicator of the hydrophobicity of a molecule.[54] Therefore, it is quite possible that log*S* gives a better description for the hydrophobicity of the studied molecules.

As shown in Figure 1, the *p*-value associated with the distinctness in the mean MW values of the two classes is 3.16×10^{-20} at the 95% confidence level, indicating that MW prominently differs between the compound classes. The mean MW values of the substrates and nonsubstrates are 498.4 and 374.9, respectively, demonstrating that P-gp substrates usually have higher molecular weights. However, it appears that the results from the present analyses are not well consistent with Gleeson's observations: the molecules with MW > 400 are more likely to be transported by P-gp.[14] In our data set, 261 (61.8%) substrates have a molecular weight higher than 400, suggesting that many compounds with a molecular weight less than 400 are also substrates. Using the criterion of MW > 400, the substrates and nonsubstrates were detected with 61.7% (261/423) and 68.7% (274/399) accuracy, separately. It is clear that the MW criterion cannot be used as a reliable filter to distinguish P-gp substrates from nonsubstrates. The descriptor n_{rot} can characterize not only the flexibility but also the bulkiness of a molecule as a larger molecule usually has more rotatable bonds. The mean values of n_{rot} for the substrates and nonsubstrates are 7.63 and 5.33 to each, illustrating that the substrates tend to have more flexible bonds. However, n_{rot} shows worse capability than MW to discriminate substrates from nonsubstrates because the *p*-value associated with the difference in the mean n_{rot} values of the substrates versus those of the nonsubstrates is 1.69×10^{-9}.

The other three properties (PSA, n_{HBD}, and n_{HBA}) are primarily used to describe the H-bonding features of a molecule. As shown in Figure 1, the distributions of these three properties do not notably vary between the compound classes. The *p*-values associated with the distribution variations of the two classes for PSA, n_{HBD}, and n_{HBA} are 3.68×10^{-6}, 3.55×10^{-7}, and 4.60×10^{-5}, individually. It appears that our results are not in agreement with previous observations though it has been reported that the H-bonding features are important factors regarding P-gp substrate-likeness.[14,55,56] The inconsistence is not so surprising because the descriptors PSA, n_{HBD}, and n_{HBA} are too general to characterize the spatial distributions of the H-bonding elements. In order to form favorable H-bonding interactions with P-gp, the H-bonding elements in P-gp substrates need to be arranged in distinct spatial patterns.

2. Analysis of Property Distributions for P-gp Substrates and Inhibitors. The molecules that can interact with P-gp are roughly classified into three classes: substrates, inhibitors, and modulators.[9] Substrates can be actively transported by P-gp, while inhibitors can compromise the transporting function of P-gp. In principle, P-gp substrates and inhibitors are different and should belong to two different classes of molecules. Therefore, the two classes of molecules can be clearly distinguished from each other via carefully designed criteria. However, this question is not so simple. P-gp inhibitors have been typically assumed to competitively inhibit substrate transport, suggesting that the molecular characteristics of the inhibitors and substrates of P-gp may share similarities in structures or binding patterns.[57] In this study, in the interest of defining features that dictate whether a molecule is inhibitor or can be transported by P-gp, the distributions of eight important molecular properties (MW, log*S*, log*D*, Alog*P*, PSA, n_{HBA}, n_{HBD}, and n_{rot}) for 423 P-gp substrates and 797 P-gp inhibitors were systematically compared. The 797 P-gp inhibitors were extracted from the PKKB database developed by our group.[36,51] The distributions of the molecular properties for the P-gp substrates and inhibitors are shown in Figure 2.

The mean values of Alog*P* for the inhibitors and substrates are 4.54 and 2.55, respectively. As shown in Figure 2, the *p*-value associated with the distribution differences of Alog*P* for the inhibitors and substrates is 6.66×10^{-72}, indicating that these two distributions are considerably different. When Alog*P* > 3.25 was chosen as a criterion to distinguish the P-gp inhibitors from substrates, 660 inhibitors (78.0%) and 269 substrates (63.7%) were correctly classified. Apparently, on average, P-gp inhibitors are principally more hydrophobic than P-gp substrates. Bain et al. also observed that the mean logK_{ow} (log octanol/water partitioning coefficient) of 18 P-gp inhibitors is much higher than that of 11 substrates.[57] Compared with Alog*P*, log*D* and log*S* exhibit worse capability for the differentiation between P-gp substrates and inhibitors because of the higher *p*-values of 5.60×10^{-17} and 0.05 (in Figure 2).

The mean MW values of the P-gp inhibitors and substrates are quite similar (489.65 and 498.38, respectively). At the 95% confidence level, the *p*-value of the difference of MW for the P-gp inhibitors and substrates is 0.46. It means that MW does not have any ability to distinguish substrates from inhibitors. Similar to MW, the distributions of n_{rot} do not show obvious difference for the inhibitors and substrates. Bain et al. observed that the mean MW of 11 substrates is significantly larger than that of 18 inhibitors, which is not consistent with our results. Regarding the much more extensive data set used in our analysis, our results apparently more fulfill the statistical significance, and are accordingly more reliable.

The mean values of n_{HBD} for the substrates and inhibitors are 2.59 and 1.17, respectively, those of n_{HBA} for the substrates and inhibitors are 6.08 and 5.85 to each, and those of PSA for the substrates and inhibitors are 110.07 and 79.62. Therefore, P-gp substrates tend to have more H-bonding elements than P-gp inhibitors. For PSA, n_{HBD}, and n_{HDA}, the *p*-values associated with the distribution differences for the inhibitors and substrates are 6.85×10^{-14}, 1.97×10^{-44}, and 0.35, individually

Table 1. The Performance of the Bayesian Classifiers for the Training and Test Sets

descriptors	TP	FN	TN	FP	SE	SP	GA	C	TP	FN	TN	FP	SE	SP	GA	C
MP[a]	238	75	184	125	0.760	0.595	0.678	0.361	83	27	52	38	0.755	0.578	0.675	0.338
MP+ECFC_6	271	42	281	28	0.866	0.909	0.887	0.776	79	31	71	19	0.718	0.789	0.750	0.505
MP+ECFC_8	281	32	282	27	0.898	0.913	0.905	0.810	96	14	68	22	0.873	0.756	0.820	0.636
MP+ECFC_10	280	33	287	22	0.895	0.929	0.912	0.824	99	11	68	22	0.900	0.756	0.835	0.667
MP+EPFC_6	253	60	251	58	0.808	0.812	0.810	0.621	78	32	63	27	0.709	0.700	0.705	0.408
MP+EPFC_8	250	63	274	35	0.799	0.887	0.842	0.688	74	36	67	23	0.673	0.744	0.705	0.415
MP+EPFC_10	255	58	278	31	0.815	0.900	0.857	0.717	76	34	69	21	0.691	0.767	0.725	0.455
MP+FCFC_6	267	46	275	34	0.853	0.890	0.871	0.743	86	24	69	21	0.782	0.767	0.775	0.547
MP+FCFC_8	280	33	277	32	0.895	0.896	0.895	0.791	92	18	68	22	0.836	0.756	0.800	0.595
MP+FCFC_10	280	33	284	25	0.895	0.919	0.907	0.814	98	12	64	26	0.891	0.711	0.810	0.617
MP+FPFC_6	249	64	253	56	0.796	0.819	0.807	0.614	78	32	66	24	0.709	0.733	0.720	0.440
MP+FPFC_8	246	67	277	32	0.786	0.896	0.841	0.686	76	34	68	22	0.691	0.756	0.720	0.444
MP+FPFC_10	249	64	287	22	0.796	0.929	0.862	0.730	73	37	71	19	0.664	0.789	0.720	0.452
MP+LCFC_6	270	43	287	22	0.863	0.929	0.895	0.793	95	15	64	26	0.864	0.711	0.795	0.585
MP+LCFC_8	278	35	286	23	0.888	0.926	0.907	0.814	94	16	67	23	0.855	0.744	0.805	0.605
MP+LCFC_10	279	34	289	20	0.891	0.935	0.913	0.827	93	17	68	22	0.845	0.756	0.805	0.605
MP+LPFC_6	282	31	272	37	0.901	0.880	0.891	0.781	83	27	64	26	0.755	0.711	0.735	0.465
MP+LPFC_8	274	39	282	27	0.875	0.913	0.894	0.788	80	30	65	25	0.727	0.722	0.725	0.448
MP+LPFC_10	272	41	288	21	0.869	0.932	0.900	0.802	75	35	74	16	0.682	0.822	0.745	0.504

[a]MP represents molecular properties and topological descriptors.

(Figure 2). Apparently, the descriptor n_{HBD} shows much better capability to discriminate the P-gp inhibitors from substrates than PSA and n_{HBA}. When $n_{HBD} \geq 2.0$ was chosen as a criterion to distinguish P-gp substrates from inhibitors, 548 inhibitors (68.8%) and 266 substrates (62.9%) were accurately categorized. Our results give the fact that the substrates tend to have more H-bond donors than the inhibitors.

3. Naive Bayesian Classifiers Based on Molecular Properties and Fingerprints. According to the previous analysis, it is obvious that a single molecular property cannot be used as a reliable filter for predicting P-gp substrates. More complicated property-based rules have been developed to characterize P-gp substrate specificity.[13−15] Among these rules, the "rule-of-four" proposed by Didziapetris et al. may be the most representative one.[15] This rule states that a molecule is more likely to be a P-gp substrate if it has more than eight nitrogen and oxygen atoms, a molecular weight larger than 400, and an acid pK_a greater than 4.0. In contrast, a molecule is more likely to be a P-gp nonsubstrate if it has fewer than four nitrogen or oxygen atoms, a molecule weight less than 400, and a base pK_a less than 8.0. Here, we validated the prediction accuracy of Didziapetris's rule-of-four on our data set. Unfortunately, only 145 out of 423 substrates were correctly predicted; in other words, most true substrates were falsely predicted as nonsubstrates.[15] Therefore, more complicated and reliable prediction models rather than simple property-based rules are quite necessary. In this study, the naive Bayesian classification technique was employed to develop the classifiers, and its effectiveness for binary classification has been extensively validated in our previous studies.[51,58,59] A Bayesian classifier was first generated only on the basis of molecular properties and topological descriptors, and then a group of classifiers were developed by adding different sets of fingerprints. The statistical results of the naive Bayesian classifiers for the training set from the 5-fold cross-validations are summarized in Table 1. It is evident that the addition of fingerprints can significantly improve the classification accuracy but the performance of different fingerprint sets is quite different. According to the GA values, the classifier based on the fingerprint set of LCFC_10 (SE = 89.1%, SP = 93.5%, C = 0.827, and GA = 91.3%) achieves the best prediction accuracy for the training set, and it is slightly better than that based on the fingerprint set of ECFC_10 (SE = 89.5%, SP = 92.9%, C = 0.824, and GA = 91.2%).

The most rigorous way of testing predictive performance of a model is to predict an independent external data set. All the Bayesian classifiers were then validated by the predictions on the external test set of 200 molecules, and the validation results are summarized in Table 1. In accordance with the GA values of the external test set, the Bayesian classifier based on the fingerprint set of ECFC_10 achieves the best actual prediction capability, and it gives a sensitivity of 89.9%, a specificity of 74.7%, a Matthews correlation coefficient of 0.667, and a global prediction accuracy of 83.5%, indicating that the Bayesian classifier is statistically reliable.

One advantage of a Bayesian classifier is that it can quantify the confidence level of the prediction. The likelihood of a compound (Bayesian score) belonging to a certain class is expressed as the cumulative conditional probability of each feature (molecular property, topological descriptor, or fingerprint) present in this compound. More positive Bayesian score indicates that the compound is more likely to be a P-gp substrate. The distributions of the Bayesian scores for the substrates and nonsubstrates given by the best classifier based on ECFC_10 are plotted in Figure 3a (training set) and Figure 3b (test set). For the training set, the Bayesian scores of the substrates distribute between −51.37 and 114.91 with a mean of 24.51, while those of the nonsubstrates between −154.86 and 9.39 with a mean of −48.43. For the best Bayesian classifier, the most optimal threshold for distinguishing the substrates from nonsubstrates is −10.79. However, the Bayesian scores of the substrates and nonsubstrates in the training set have strong overlap roughly between −25 and 0 (Figure 3a). So the region between −25 and 0 is defined as the "uncertain zone". When the Bayesian score of a molecule predicted by the classifier is located in the uncertain zone, the prediction for this molecule may be not reliable.

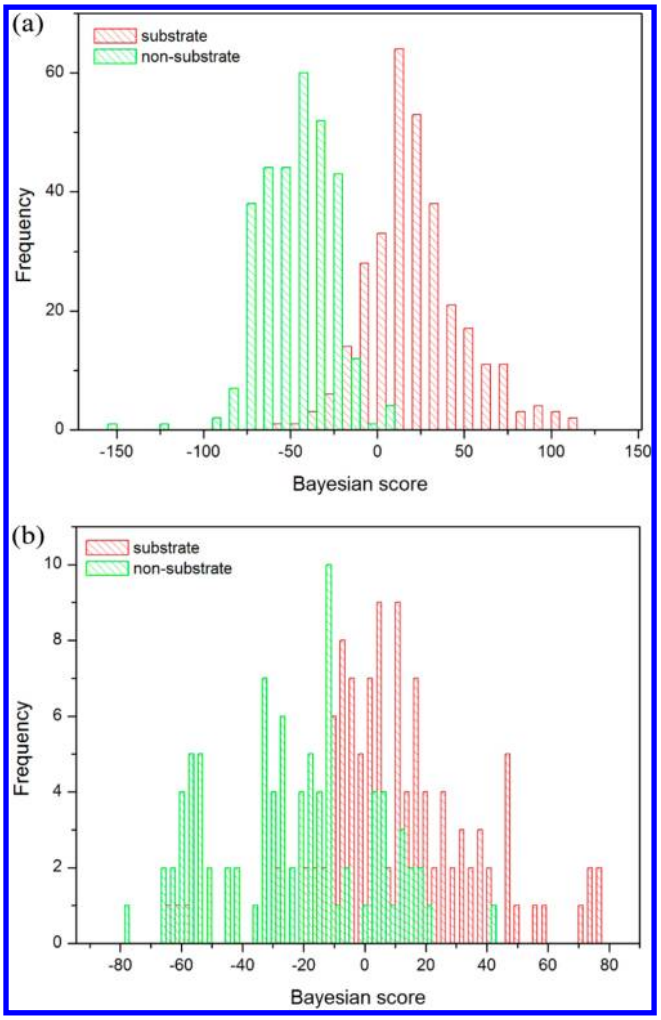


Figure 3. Distributions of the Bayesian scores predicted by the best Bayesian classifier based on the molecular properties, topological descriptors, and ECFC_10 fingerprint set for the substrate and nonsubstrate classes in (a) the training set and (b) the test set. The Bayesian scores for the training set were obtained by using the 5-fold cross-validation test.

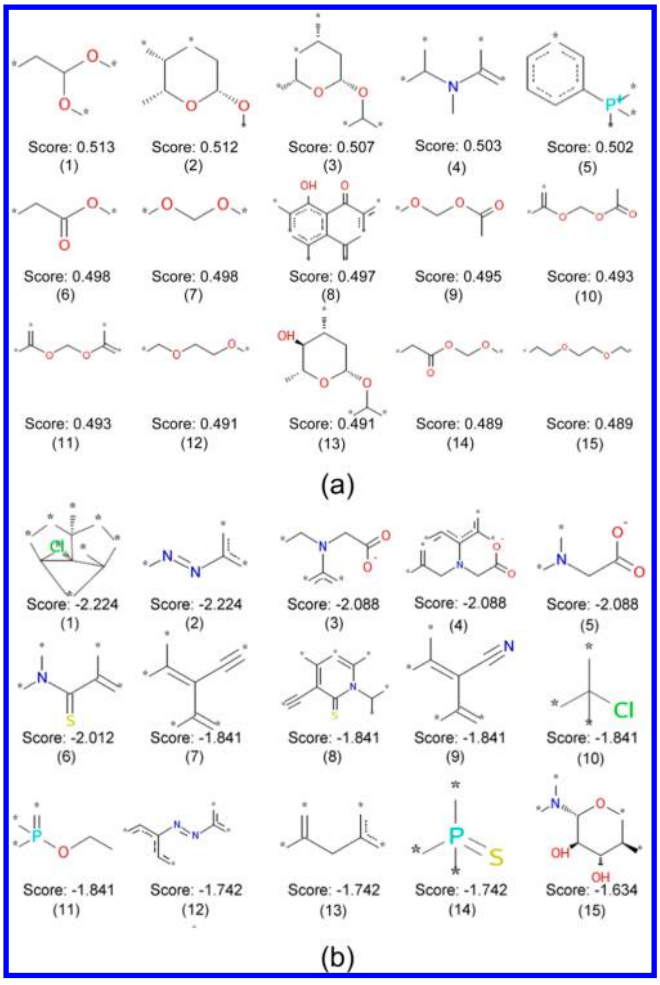


Figure 4. (a) The top 15 good and (b) 15 bad fragments for P-gp substrate-likeness identified by the best Bayesian classifier.

4. Analysis of the Important Fragments Given by Naive Bayesian Classifier. The learning process of Bayesian modeling generates the weight (or score) for each feature present in samples using a Laplacian-adjusted probability estimate, and then the importance of each fragment can be quantitatively evaluated. The fragments with high scores contribute positively to substrate-likeness, and those with low scores have negative influence on substrate-likeness. Here, the top 15 good fragments and the top 15 bad fragments ranked by the Bayesian scores are listed in Figure 4.

By analyzing the top 15 good fragments shown in Figure 4a, we observed that 13 out of 15 fragments have 2 or 3 oxygen atoms. Moreover, these two oxygen atoms are connected by 2−4 bonds. Intriguingly, our observations are in good agreement with the results reported by Seelig.[56] By comparing the structures of a hundred P-gp substrates, Seelig found that a set of well-defined structural elements are important for the interaction with P-gp, and these elements are formed by two (type I unit) or three electron donor groups (type II unit) with a fixed spatial separation. Type I units consist of two electron donor groups with a spatial separation of ∼2.5 Å, and type II units contain either two electron donor groups with a spatial separation of ∼4.6 Å or three electron donor groups with a spatial separation of the outer two groups of ∼4.6 Å. In these good fragments shown in Figure 4a, 11 of them (fragments 1, 2, 3, 6, 7, 8, 9, 10, 11, 13, 14) contain type I units, and one (fragment 13) contains a type II unit. It is quite possible that the oxygen atoms in most of these important fragments serve as H-bond acceptors and form stable H-bonds with the H-bond donors with specific spatial arrangement in P-gp. As demonstrated in Figure 1, the distributions of n_{HBA} for the substrates and nonsubstrates in our data set are not greatly distinct. Therefore, not all H-bond acceptors in P-gp substrates are valuable, and only the H-bond acceptors arranged in distinct spatial patterns are critical for the interactions between P-gp and substrates. In these top 15 good fragments, it appears that 2 of them (fragments 12 and 15) have at least two oxygen atoms while not belonging to either type I or type II unit. However, these two fragments are quite flexible, and the two oxygen atoms in fragment 12 or 15 can approach to ∼2.5 Å through conformational adjustment. Therefore, fragments 12 and 15 may also contain the type I unit defined by Seelig.[56] Moreover, most good fragments in Figure 4a are flexible. The compounds with these flexible fragments may adjust their conformations easily, and are more likely to form favorable interactions with P-gp.

dx.doi.org/10.1021/mp400450m | Mol. Pharmaceutics 2014, 11, 716−726

The top 15 bad fragments with the most negative scores are depicted in Figure 4b. Out of the top 5 fragments, three structures (fragments 3, 4, and 5) have a negatively charged carboxyl group. It is believed that the electrostatic repulsion between the negatively charged group and the negatively charge residues in P-gp may be unfavorable for substrate binding to P-gp. Furthermore, most unfavorable fragments carry carbon−carbon double bonds and carbon−carbon triple bonds, and besides, four fragments contain complicated ring systems. All these fragments can reduce the flexibility of a molecule, thus bringing more difficulties for the molecule to orientate into the binding site and form compact interactions with P-gp. In addition, the majority (9/15) of the bad fragments have nitrogen atoms. Among them, 5 fragments only have nitrogen atoms but no oxygen atoms. The nitrogen atoms in most fragments are linked to conjugated or aromatic systems, and therefore may give negative contribution to the flexibility of a molecule. Several bad fragments found in our study were also previously reported.[23] For example, Wang et al. found that a tertiary nitrogen linked to an alkoxy group in a ring system (fragment 15), and that thial group (fragments 6 and 8) and nitrile group (fragment 9) have a much higher frequency in nonsubstrates.[23]

5. Analysis of the Misclassified Molecules. The best Bayesian classifier provides satisfactory prediction (GA = 83.5%) for the 200 molecules in the test set. A total of 33 molecules (11 substrates and 22 nonsubstrates) still cannot be well predicted. The following reasons may be responsible for the misclassification. First, the quality of the data set may be a primary source of misclassification. In this work, we collected the data from hundreds of literature sources. The assays and standards of classifying a molecule as a P-gp substrate or nonsubstrate are certainly not completely identical. For example, buprenorphine was collected as a nonsubstrate, but predicted as a substrate by the Bayesian classifier. Regarding Suzuki's study, buprenorphine is at least in part transported into the brain across the blood−brain barrier (BBB) via a P-gp-mediated efflux transport system and, therefore, is believed to be a P-gp substrate.[60] In Hassan's work, both *in vitro* (P-gp ATPase and monolayer efflux) assays and *in vivo* (tissue distribution and antinociceptive monitoring in P-gp deficient/competent mice) assays were employed to evaluate the P-gp affinity status of buprenorphine. Buprenorphine was negative in all assays and thought most likely not a P-gp substrate.[33] Obviously, for the same molecule, different assays produced conflicting data. In our study, the experimental data is collected from a variety of sources and the diversity of the experimental data will increase the data uncertainty. Second, the misclassification of some molecules is perhaps caused by the intrinsic disadvantage of the theoretical models. The Bayesian classifier based on fingerprints can highlight the important substructures that have positive or negative contributions for P-gp substrate-likeness, but it cannot offer the information if the favorable substructure in a molecule satisfies the conformational requirement to form favorable interactions with P-gp. Third, even if a much larger data set was employed for model development, it is quite possible that some fragments essential for substrate-likeness are not included in the training set molecules, and then the molecules containing these undetected important fragments in the test set cannot be correctly predicted.

CONCLUSIONS

In this study, we compiled a much larger data set than previous studies for predicting P-gp substrates, including 423 P-gp substrates and 399 nonsubstrates. The distributions of eight important physicochemical properties for 423 P-gp substrates and 399 nonsubstrates were then compared. The results from the analysis illustrate that molecular solubility and molecular weight are more informative than the other six properties for differentiating substrates from nonsubstrates. Furthermore, the distributions of the important physicochemical properties for 735 P-gp inhibitors and 423 substrates were compared. The comparison study shows that the P-gp inhibitors are significantly more hydrophobic than the P-gp substrates while the substrates tend to have more H-bond donors than the inhibitors. Finally, the naive Bayesian classification technique was employed to develop the classifiers to distinguish P-gp substrates from nonsubstrates. The best Bayesian classifier achieves a global accuracy of 91.2% for the training set and a global accuracy of 83.5% for the test set. The critical fragments favorable and unfavorable for P-gp substrate-likeness highlighted by the Bayesian classifiers were analyzed and discussed, which affords a deeper understanding of the molecular basis of substrate−transporter interactions.

ASSOCIATED CONTENT

Supporting Information

The complete list of literature references. This material is available free of charge via the Internet at http://pubs.acs.org. The whole data set can be obtained from the supporting Web site: http://cadd.suda.edu.cn/admet.

AUTHOR INFORMATION

Corresponding Author

*E-mail: tingjunhou@hotmail.com or tingjunhou@zju.edu.cn. Phone: +86-512-65882039.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This study was supported by the National Science Foundation of China (21173156), Specialized Research Fund for the Doctoral Program of Higher Education (20123201110017), the National Basic Research Program of China (973 program, 2012CB932600), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD)

REFERENCES

(1) Childs, S.; Ling, V. The MDR superfamily of genes and its biological implications. *Important Adv. Oncol.* **1994**, 21−36.

(2) Dean, M.; Allikmets, R. Evolution of ATP-binding cassette transporter genes. *Curr. Opin. Genet. Dev.* **1995**, *5*, 779−785.

(3) Higgins, C. F. ABC transporters: from microorganisms to man. *Annu. Rev. Cell Biol.* **1992**, *8*, 67−113.

(4) Lin, J. H.; Yamazaki, M. Clinical relevance of P-glycoprotein in drug therapy. *Drug Metab. Rev.* **2003**, *35*, 417−454.

(5) Aszalos, A. Drug-drug interactions affected by the transporter protein, P-glycoprotein (ABCB1, MDR1): I. Preclinical aspects. *Drug Discovery Today* **2007**, *12*, 833−837.

(6) Perez-Tomas, R. Multidrug resistance: retrospect and prospects in anti-cancer drug treatment. *Curr. Med. Chem.* **2006**, *13*, 1859−1876.

(7) Aller, S. G.; Yu, J.; Ward, A.; Weng, Y.; Chittaboina, S.; Zhuo, R. P.; Harrell, P. M.; Trinh, Y. T.; Zhang, Q. H.; Urbatsch, I. L.; Chang,

G. Structure of P-Glycoprotein Reveals a Molecular Basis for Poly-Specific Drug Binding. *Science* **2009**, *323*, 1718−1722.
(8) Adachi, Y.; Suzuki, H.; Sugiyama, Y. Comparative studies on in vitro methods for evaluating in vivo function of MDR1 P-glycoprotein. *Pharm. Res.* **2001**, *18*, 1660−1668.
(9) Chen, L.; Li, Y.; Yu, H.; Zhang, L.; Hou, T. Computational models for predicting substrates or inhibitors of P-glycoprotein. *Drug Discovery Today* **2012**, *17*, 343−351.
(10) Ekins, S.; Kim, R. B.; Leake, B. F.; Dantzig, A. H.; Schuetz, E. G.; Lan, L. B.; Yasuda, K.; Shepard, R. L.; Winter, M. A.; Schuetz, J. D.; Wikel, J. H.; Wrighton, S. A. Application of three-dimensional quantitative structure-activity relationships of P-glycoprotein inhibitors and substrates. *Mol. Pharmacol.* **2002**, *61*, 974−981.
(11) Ekins, S.; Polli, J. E.; Swaan, P. W.; Wright, S. H., Computational modeling to accelerate the identification of substrates and inhibitors for transporters that affect drug disposition. *Clin. Pharmacol. Ther.* **2012**.
(12) Ekins, S.; Ecker, G. F.; Chiba, P.; Swaan, P. W. Future directions for drug transporter modelling. *Xenobiotica* **2007**, *37*, 1152−1170.
(13) Ford, J. M.; Hait, W. N. Pharmacology of drugs that alter multidrug resistance in cancer. *Pharmacol. Rev.* **1990**, *42*, 155−199.
(14) Gleeson, M. P. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **2008**, *51*, 817−834.
(15) Didziapetris, R.; Japertas, P.; Avdeef, A.; Petrauskas, A. Classification analysis of P-glycoprotein substrate specificity. *J. Drug Targeting* **2003**, *11*, 391−406.
(16) Ecker, G.; Huber, M.; Schmid, D.; Chiba, P. The importance of a nitrogen atom in modulators of multidrug resistance. *Mol. Pharmacol.* **1999**, *56*, 791−796.
(17) Lima, P. D. C.; Golbraikh, A.; Oloff, S.; Xiao, Y. D.; Tropsha, A. Combinatorial QSAR modeling of P-glycoprotein substrates. *J. Chem. Inf. Model.* **2006**, *46*, 1245−1254.
(18) Crivori, P.; Reinach, B.; Pezzetta, D.; Poggesi, I. Computational models for identifying potential P-glycoprotein substrates and inhibitors. *Mol. Pharmaceutics* **2006**, *3*, 33−44.
(19) Cabrera, M. A.; Gonzalez, I.; Fernandez, C.; Navarro, C.; Bermejo, M. A topological substructural approach for the prediction of P-glycoprotein substrates. *J. Pharm. Sci.* **2006**, *95*, 589−606.
(20) Gombar, V. K.; Polli, J. W.; Humphreys, J. E.; Wring, S. A.; Serabjit-Singh, C. S. Predicting P-glycoprotein substrates by a quantitative structure-activity relationship model. *J. Pharm. Sci.* **2004**, *93*, 957−968.
(21) Vasanthanathan, P.; Haider, N.; Ecker, G. F. Fingerprint-based in silico models for the prediction of P-glycoprotein substrates and inhibitors. *Bioorg. Med. Chem.* **2012**, *20*, 5388−5395.
(22) Huang, J. P.; Ma, G. L.; Muhammad, I.; Cheng, Y. Y. Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm. *J. Chem. Inf. Model.* **2007**, *47*, 1638−1647.
(23) Wang, Z.; Chen, Y. Y.; Liang, H.; Bender, A.; Glen, R. C.; Yan, A. X. P-glycoprotein Substrate Models Using Support Vector Machines Based on a Comprehensive Data set. *J. Chem. Inf. Model.* **2011**, *51*, 1447−1456.
(24) Xue, Y.; Yap, C. W.; Sun, L. Z.; Cao, Z. W.; Wang, J. F.; Chen, Y. Z. Prediction of P-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1497−1505.
(25) Wang, Y. H.; Li, Y.; Yang, S. L.; Yang, L. Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. *J. Chem. Inf. Model.* **2005**, *45*, 750−757.
(26) Penzotti, J. E.; Lamb, M. L.; Evensen, E.; Grootenhuis, P. D. J. A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. *J. Med. Chem.* **2002**, *45*, 1737−1740.
(27) Polli, J. W.; Wring, S. A.; Humphreys, J. E.; Huang, L.; Morgan, J. B.; Webster, L. O.; Serabjit-Singh, C. S. Rational use of in vitro P-glycoprotein assays in drug discovery. *J. Pharmacol. Exp. Ther.* **2001**, *299*, 620−628.
(28) Schwab, D.; Fischer, H.; Tabatabaei, A.; Poli, S.; Huwyler, J. Comparison of in vitro P-glycoprotein screening assays: recommendations for their use in drug discovery. *J. Med. Chem.* **2003**, *46*, 1716−1725.
(29) Ramu, A.; Ramu, N. Reversal of multidrug resistance by phenothiazines and structurally related compounds. *Cancer Chemother. Pharmacol.* **1992**, *30*, 165−173.
(30) Estrada, E.; Molina, E.; Nodarse, D.; Uriarte, E. Structural contributions of substrates to their binding to P-glycoprotein. A TOPSMODE approach. *Curr. Pharm. Des.* **2010**, *16*, 2676−2709.
(31) Scala, S.; Akhmed, N.; Rao, U. S.; Paull, K.; Lan, L. B.; Dickstein, B.; Lee, J. S.; Elgemeie, G. H.; Stein, W. D.; Bates, S. E. P-glycoprotein substrates and antagonists cluster into two distinct groups. *Mol. Pharmacol.* **1997**, *51*, 1024−1033.
(32) Tang-Wai, D. F.; Brossi, A.; Arnold, L. D.; Gros, P. The nitrogen of the acetamido group of colchicine modulates P-glycoprotein-mediated multidrug resistance. *Biochemistry* **1993**, *32*, 6470−6476.
(33) Hassan, H. E.; Myers, A. L.; Coop, A.; Eddington, N. D. Differential involvement of P-glycoprotein (ABCB1) in permeability, tissue distribution, and antinociceptive activity of methadone, buprenorphine, and diprenorphine: In vitro and in vivo evaluation. *J. Pharm. Sci.* **2009**, *98*, 4928−4940.
(34) Takara, K.; Sakaeda, T.; Yagami, T.; Kobayashi, H.; Ohmoto, N.; Horinouchi, M.; Nishiguchi, K.; Okumura, K. Cytotoxic effects of 27 anticancer drugs in HeLa and MDR1-overexpressing derivative cell lines. *Biol. Pharm. Bull.* **2002**, *25*, 771−778.
(35) Naito, M.; Hamada, H.; Tsuruo, T. d. ATP/Mg2+-dependent binding of vincristine to the plasma membrane of multidrug-resistant K562 cells. *J. Biol. Chem.* **1988**, *263*, 11887−11891.
(36) Cao, D.; Wang, J.; Zhou, R.; Li, Y.; Yu, H.; Hou, T. ADMET evaluation in drug discovery. 11. PharmacoKinetics Knowledge Base (PKKB): a comprehensive database of pharmacokinetic and toxic properties for drugs. *J. Chem. Inf. Model.* **2012**, *52*, 1132−1137.
(37) *SYBYL X1.1*; Tripos, Inc.: St. Louis, MO, USA, 2011.
(38) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(39) *Discovery Studio 2.5 Guide*; Accelrys Inc.: San Diego, 2009; http://www.accelrys.com.
(40) Hou, T.; McLaughlin, W. A.; Wang, W. Evaluating the potency of HIV-1 protease drugs to combat resistance. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 1163−1174.
(41) Hou, T. J.; Wang, J. M.; Zhang, W.; Wang, W.; Xu, X. Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Curr. Med. Chem.* **2006**, *13*, 2653−2667.
(42) Zhu, J. Y.; Wang, J. M.; Yu, H. D.; Li, Y. Y.; Hou, T. J. Recent Developments of In Silico Predictions of Oral Bioavailability. *Comb. Chem. High Throughput Screening* **2011**, *14*, 362−374.
(43) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.
(44) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.
(45) Muller, W. R.; Szymanski, K.; Knop, J. V.; Trinajstic, N. An algorithm for construction of the molecular distance matrix. *J. Comput. Chem.* **1987**, *8*, 170−173.
(46) Bonchev, D.; Boncev, D.; Chemiker, B.; Boncev, D.; Boncev, D.; Chemist, B. *Information theoretic indices for characterization of chemical structures*; Research Studies Press: Chichester, 1983; Vol. *5*.
(47) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(48) Kier, L. B.; Hall, L. H. *Molecular Connectivity Indices in Chemistry and Drug Research*; Academic Press: New York, 1976; Vol. *14*.
(49) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-property Modeling. *Rev. Comput. Chem.* **2007**, *2*, 367−422.
(50) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening* **2005**, *10*, 682−686.

(51) Chen, L.; Li, Y.; Zhao, Q.; Peng, H.; Hou, T. ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques. *Mol. Pharmaceutics* **2011**, *8*, 889−900.
(52) Berger, J. O. *Statistical decision theory and Bayesian analysis*; Springer: 1985.
(53) Sun, H. M. A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J. Med. Chem.* **2005**, *48*, 4031−4039.
(54) Wang, J. M.; Hou, T. J.; Xu, X. J. Aqueous Solubility Prediction Based on Weighted Atom Type Counts and Solvent Accessible Surface Areas. *J. Chem. Inf. Model.* **2009**, *49*, 571−581.
(55) Demel, M. A.; Kraemer, O.; Ettmayer, P.; Haaksma, E.; Ecker, G. F. Ensemble Rule-based Classification of Substrates of the Human ABC-transporter ABCB1 Using Simple Physicochemical Descriptors. *Mol. Inf.* **2010**, *29*, 233−242.
(56) Seelig, A. A general pattern for substrate recognition by P-glycoprotein. *Eur. J. Biochem.* **1998**, *251*, 252−261.
(57) Bain, L. J.; McLachlan, J. B.; LeBlanc, G. A. Structure-activity relationships for xenobiotic transport substrates and inhibitory ligands of P-glycoprotein. *Environ. Health Perspect.* **1997**, *105*, 812−818.
(58) Tian, S.; Wang, J.; Li, Y.; Xu, X.; Hou, T. Drug-likeness analysis of traditional Chinese medicines: prediction of drug-likeness using machine learning approaches. *Mol. Pharmaceutics* **2012**, *9*, 2875−2886.
(59) Wang, S.; Li, Y.; Wang, J.; Chen, L.; Zhang, L.; Yu, H.; Hou, T. ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. *Mol. Pharmaceutics* **2012**, *9*, 996−1010.
(60) Suzuki, T.; Zaima, C.; Moriki, Y.; Fukami, T.; Tomono, K. P-glycoprotein mediates brain-to-blood efflux transport of buprenorphine across the blood-brain barrier. *J. Drug Targeting* **2007**, *15*, 67−74.