Accurate models for P-gp drug recognition induced from a cancer cell line cytotoxicity screen

*Jurica Levatić[a], Jasna Ćurak[a], Marijeta Kralj[a,b], Tomislav Šmuc[a,c], Maja Osmak[d], Fran Supek[a,c],**

[a] BioZyne Ltd, Bijenička 54, 10000 Zagreb, Croatia; [b] Division of Molecular Medicine, Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia; [c] Division of Electronics, Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia; [d] Division of Molecular Biology, Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia.

ABSTRACT. P-glycoprotein (P-gp, MDR1) is a promiscuous drug efflux pump of substantial pharmacological importance. Taking advantage of large-scale cytotoxicity screening data involving 60 cancer cell lines, we correlated the differential biological activities of ~13 000 compounds against cellular P-gp levels. We created a large set of 934 high-confidence P-gp substrates or non-substrates by enforcing agreement with an orthogonal criterion involving P-gp overexpressing ADR-RES cells.

A Support Vector Machine (SVM) was 86.7% accurate in discriminating P-gp substrates on independent test data, exceeding previous models. Two molecular features had an overarching influence: nearly all P-gp substrates were large (>35 atoms including H) and dense (specific volume <7.3 $Å^3$/atom) molecules. Seven other descriptors and 24 molecular fragments ("effluxophores") were found enriched in the (non)substrates and incorporated into interpretable rule-based models.

Biological experiments on an independent P-gp overexpressing cell line, the vincristine-resistant VK_2, allowed us to re-classify six compounds previously annotated as substrates, validating our method's predictive ability. Models are freely available at http://pgp.biozyne.com.

INTRODUCTION

The P-glycoprotein (P-gp) is the protein product of the ABCB1 gene, also known as MDR1 (multidrug resistance 1) that belongs to the ATP-binding cassette superfamily of membrane transporters. Its main biological function is to protect cells against various potentially harmful xenobiotics and other cytotoxic compounds, which it expels out of the cell. P-gp is abundant in a number of cells in human organs where it plays a secretory role (e.g. in the small intestine, liver and kidney), or acts as a barrier as in the brain endothelial capillaries where it contributes to blood-brain-barrier function.[1]

Unlike most other membrane transporters, P-gp is known for its promiscuity - it transports a wide range of chemically and pharmacologically unrelated substrates, including small molecules such as carbohydrates and organic cations, and macromolecules such as proteins and polysaccharides.[2] Accordingly, the pump can also expel medical therapeutics out of the cell, thus reducing their desired effect.[3] In particular, many cancer chemotherapeutics, such as vinca alkaloids, anthracyclines, epipodophyllotoxins and taxanes, are known to be P-gp substrates. Additionally, many human cancers overexpress the ABCB1 gene resulting in multi-drug resistant cancers.[4] Furthermore, P-gp not only causes cancer chemotherapy failure but also greatly influences the general pharmacokinetic parameters of clinically important therapeutics for other diseases.[5] It is these reasons that prompted our work to distinguish and evaluate P-gp substrates from non-substrates early in the drug discovery pipeline in hopes of streamlining compounds for future therapeutic development.

Computational Quantitative Structure-Activity Relationship (QSAR) models for P-gp substrate specificity provides a fast and cost-efficient means to achieve this goal, and were therefore the subject of many past research efforts. A pinnacle example is the study by Penzotti *et al*.,[6] who collected a dataset of 195 P-gp substrates and non-substrates from various sources (e.g. Seelig[7]), and reported a classification accuracy of 80% on the training set, and 63% on an independent test set using an ensemble of pharmacophore models. In later studies, the Penzotti dataset was often re-used or served as a starting point for data collection. Publicly available datasets used in modeling P-gp substrate specificity[8–11] have sizes measured in several hundreds of molecules, up to 332 compounds collected by Wang *et al*.[10] It should be noted that these datasets have significant overlap (see Figure 1), because they tend to use the same data sources.

Importantly, these sources rely on different experimental assays to determine whether a compound is a P-gp substrate, which may cause variability in the results. As discussed previously,[11–13] experimental assays may be performed under different conditions (e.g. concentration of compound) and often carry biases inherent to each method. Most prominently, the common ATPase and calcein-AM assays cannot differentiate between P-gp substrates and inhibitors; additionally, the compounds' passive membrane permeability influences the experiment outcome.[14] Discordance between experimental measurements was addressed in the recent Bikadi *et al.* study[11] by extensive curation - a compound was accepted only if more studies confirmed its classification, yielding an arguably high-quality dataset. Nevertheless, the 197 compounds therein were often classified based on assays in the less reliable group, according to the systematization in Didziapetris *et al.*[12]: approx. ⅓ of the compounds relied on the ATPase assay, and ½ of the compounds on the calcein-AM assay results. For instance, doxorubicin, considered a P-gp substrate by many[15,16] is here classified as a P-gp non-substrate.[14]

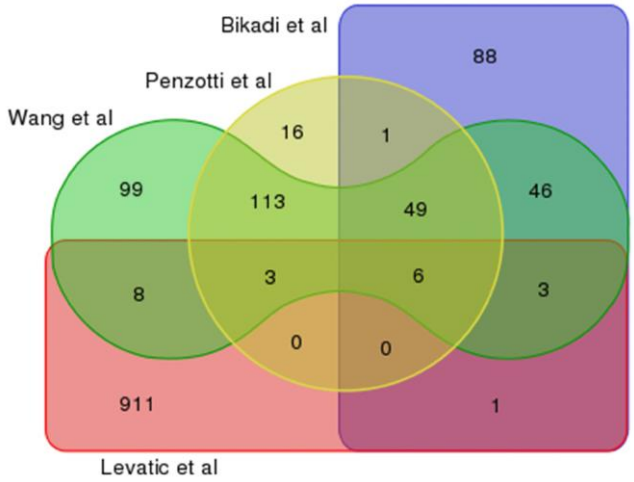


Figure 1. Overlap of the publicly available datasets used in recent P-gp studies modeling P-gp substrate recognition and the dataset reported in this study. The Venn diagram[17] shows unique compounds via exact matching of canonicalized SMILES after clearing stereochemistry information; on few occasions, this caused multiple stereoisomers within a dataset to be represented as a single molecule.

In contrast to the largely re-used sets of P-gp substrates and non-substrates, previous P-gp specificity studies employed very different computational approaches in generating molecular descriptors and modeling their relationship to the biological response. For instance, Cabrera *et al.* used topological substructural descriptors in combination with linear discriminant analysis reporting an accuracy of 77% on an independent test set.[9] De Cerqueira Lima tested various

combinations of learning methods (k-NN, Decision Tree, binary QSAR and SVM) and descriptor sets (molecular connectivity indices, atom pair descriptors, and descriptors from the VolSurf and MOE software) with the best model - SVM with VolSurf - achieving a test set accuracy of 80%.[18] The different learning methods were also compared across various QSAR datasets, including P-gp substrate specificity, yielding accuracies in the 70-80% range for the tested methods on P-gp data.[19] Recently, several researchers[8,10,11,20] employed SVM combined with various descriptor types and reported similarly high test set accuracies. These studies demonstrate how SVMs tend to have comparable or better predictive performance in comparison to other methods, as evidenced also in their widespread use in QSAR studies of diverse molecules,[21] for example applied to series of peptides[22] or crown ethers.[23]

The properties of a molecule thought to be relevant for P-gp recognition include a wide array of structural features, commonly including molecular weight and/or volume, the number and spatial arrangement of hydrogen bond acceptors, polar (or total) surface area, polarizability, molecular charge, and aromaticity.[9,12] There is, however, no consensus about which of these features is most important, or if some are important at all. The prime example is the LogP value, a measure of hydrophobicity with wide impact on ADME-Tox properties in general,[24] which was claimed in multiple studies to be either very important,[10,25,26] or of little relevance[12,20,27] for recognizing P-gp substrates. The discrepancies may be caused by (a) small datasets with patchy coverage of the molecular space; (b) data from unreliable experimental assays, where a feature's relevance might reflect a bias in the assay; and (c) confounding variables which were not controlled for, e.g. the perceived relevance of LogP might be due to the importance of aromatic rings which strongly contribute to the LogP.

In this study we present a novel and comprehensive set of P-gp substrates and non-substrates. The dataset is composed of 934 compounds, making it the largest publicly available collection, almost three times the size of the previous most comprehensive dataset.[10] Furthermore, our data shows strong agreement with previous substrate/non-substrate assignments, but low overlap with the molecular structures in past datasets. The data collection was performed in a systematic manner from the large-scale compound cytotoxicity screening against 60 human cancer cell lines at the US National Cancer Institute Developmental Therapeutics Program (NCI-DTP), while simultaneously drawing on the extensive knowledge about the expression level of the cellular target - the P-gp. The compounds-of-interest were carefully selected among thousands available

at the NCI-DTP database and labeled as a P-gp substrate or non-substrate only if they passed a stringent check, requiring two independent assessments of correlation of ABCB1 levels to cytostatic activity.

From this large dataset we derive a simple, interpretable rule based on two physical features of the compounds - atom count and specific volume - which guarantees that a compound is a non-substrate of P-gp with 89% precision, while still recovering most of the non-substrates. This rule proved to be generally valid, exhibiting a high precision for non-substrates on two independent datasets. Additionally, we find 24 molecular fragments strongly enriched in the substrates or the non-substrates after controlling for the effects of atom count and specific volume. Finally, we integrated a number of diverse molecular descriptors to train a SVM classification model with high accuracy and considerably improved statistical support over the previous studies. We experimentally validated the model by testing growth inhibition of six compounds on a P-gp overexpressing cell line not used in the training database. A free web server, available at http://pgp.biozyne.com, was developed to give the scientific community an easy possibility to apply our SVM model to classify their compounds as P-gp substrates or non-substrates.

RESULTS

Detecting P-gp substrates from cancer cell line screening data. Since P-gp acts by extruding substrates out of the cell, ABCB1-overexpressing cell lines should require a higher drug concentration to cause growth inhibition. The NCI-DTP cancer cell line panel (NCI-60) is heterogeneous in ABCB1 gene expression levels (Figure 2a) and accordingly ABCB1 expression levels should inversely correlate with cytotoxicity for P-gp substrates, but not for P-gp non-substrates. Previous work has shown that such a correlation can be successfully exploited for identifying P-gp substrates.[28–30] The NCI-60 contains two cell lines that were important for this study: the OVCAR-8 and NCI/ADR-RES[a], both ovarian cells derived from the same individual.[32,33] The key differential feature between the two cell lines is in the high ABCB1 expression in the NCI/ADR-RES cell line (Figure 2c). We have combined eleven previous measurements of ABCB1 mRNA levels across all cell lines using principal components analysis (see Methods for details) and summarized the outcome in Figure 2a. The NCI/ADR-RES cell

[a] previously named MCF-7/ADR-RES[31]

line expresses by far the highest level of ABCB1 gene, whereas OVCAR-8 cell has a low expression level; 18th cell line out of 60 in the NCI-60 panel. Consequently, if a compound is less cytotoxic for the NCI/ADR-RES cell line in comparison to the OVCAR-8, it is plausible that the compound is a P-gp substrate. Conversely, if the activity is similar, it suggests that P-gp did not interfere with the compound, so it is likely a P-gp non-substrate. For instance, an ~80-fold higher concentration of the known P-gp substrate vincristine is required for NCI/ADR-RES growth inhibition, than for OVCAR-8 cells (Figure 2b).

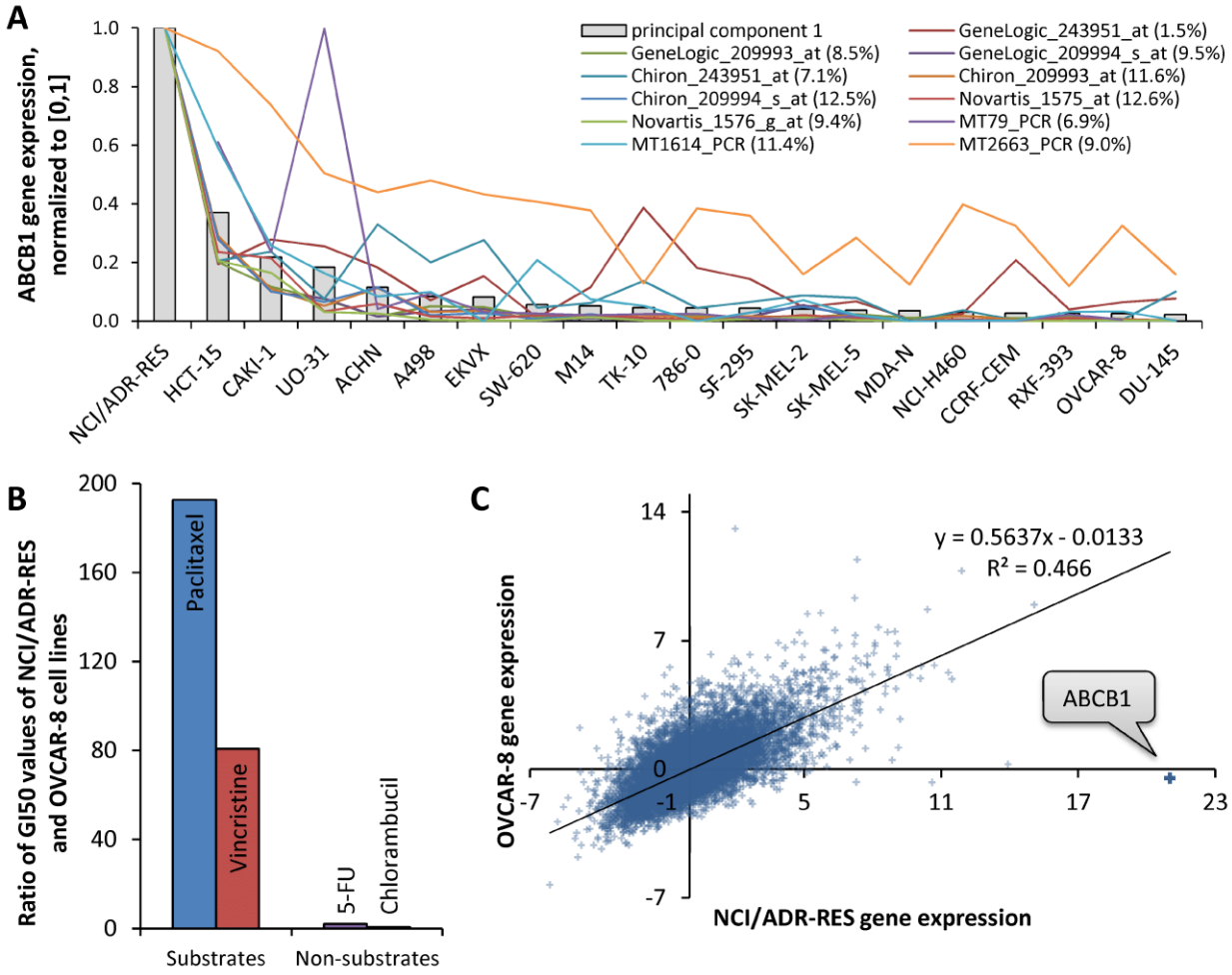


Figure 2. (A) Eleven measurements of ABCB1 mRNA levels using microarrays or quantitative PCR across cell lines (colored lines), and a summary of the measurements (PC1, grey bars; see Methods). The PC1 weighted average of the measurements, where the contribution of each mRNA level measurement (numbers in parentheses) depends on how well it agrees with the consensus over cell lines. The top 20 cell lines by PC1 are shown; all remaining 40 cell lines have very low normalized PC1<0.022. (B) Ratios of cytotoxic activities on NCI/ADR-RES and OVCAR-8 cell lines of known P-gp substrates and non-substrates. The y axis shows the GI50 ratio (the compound

concentration causing 50% growth inhibition) of the NCI/ADR-RES cell line, and the OVCAR-8 GI50. Higher ratios indicate that ADR-RES is more resistant to the compound. (C) A comparison of the mRNA levels between all genes in the NCI/ADR-RES and the OVCAR-8 cell lines indicates a good general correlation, consistent with the former cell line being derived from the latter. The ABCB1 mRNA, encoding the P-gp protein, is highly abundant in NCI/ADR-RES, but not in OVCAR-8 cells.

Thus, differential compound cytotoxicity in the NCI/ADR-RES and OVCAR-8 cell lines constitutes a criterion for detecting P-gp substrates (herein, the "difference" criterion). By analogy, the variability in P-gp expression in the remaining 58 cell lines in the NCI-60 panel (Figure 2a) can be used to establish a second, independent criterion to detect P-gp substrates, where an anti-correlation of P-gp levels and cytotoxic activity across the 58 cell lines (the "correlation" criterion) is expected. The compounds that hold true to both criteria are likely to be P-gp substrates. Conversely, if a compound shows this relationship for neither criterion, it is likely a P-gp non-substrate.

Constructing a comprehensive dataset. Enforcing the agreement of both criteria, we extracted a dataset of 958 compounds (471 substrates and 487 non-substrates) from the measurements on the NCI-60 cell line panel. We chose the stringency for declaring substrates to maximize the statistical support of the agreement between the "difference" and "correlation" criteria (Figure 3; see Methods for details). The threshold for declaring a substrate was set to the top 8% of the "difference" and "correlation" values, yielding 471 substrates that satisfy both criteria. The threshold for the non-substrates was chosen to strike a balance between the sizes of the substrate and the non-substrate class, selecting the mid-20% (centered around zero) of the molecules by both criteria (Figure 3).

We further removed multiple copies of identical or nearly identical compounds, reducing the dataset to 934 compounds, of those 448 substrates and 486 non-substrates (Table S1; see Methods for details). If both compounds from these (near-)identical pairs of compounds had been included, this might have resulted in overly optimistic error estimates when constructing QSAR models on the data. Our method for declaring P-gp substrates and non-substrates agrees well with the substrate/non-substrate annotations in a previous dataset by Penzotti *et al*.[6]: 10 out of the 10 compounds in our dataset that have a counterpart in Penzotti *et al*. agree with that counterpart. Importantly, there is little overlap between the two sets in terms of the compound identities (Figure 3), implying that most of the compounds in our dataset (911/934, 97.5%) are

novel with respect to any of the three previous datasets considered here (Figure. 1). Nevertheless, the coverage of chemical molecule space is largely similar as the past datasets (Figure S1), indicating that a model derived from this data would be generally applicable to various sets of compounds.

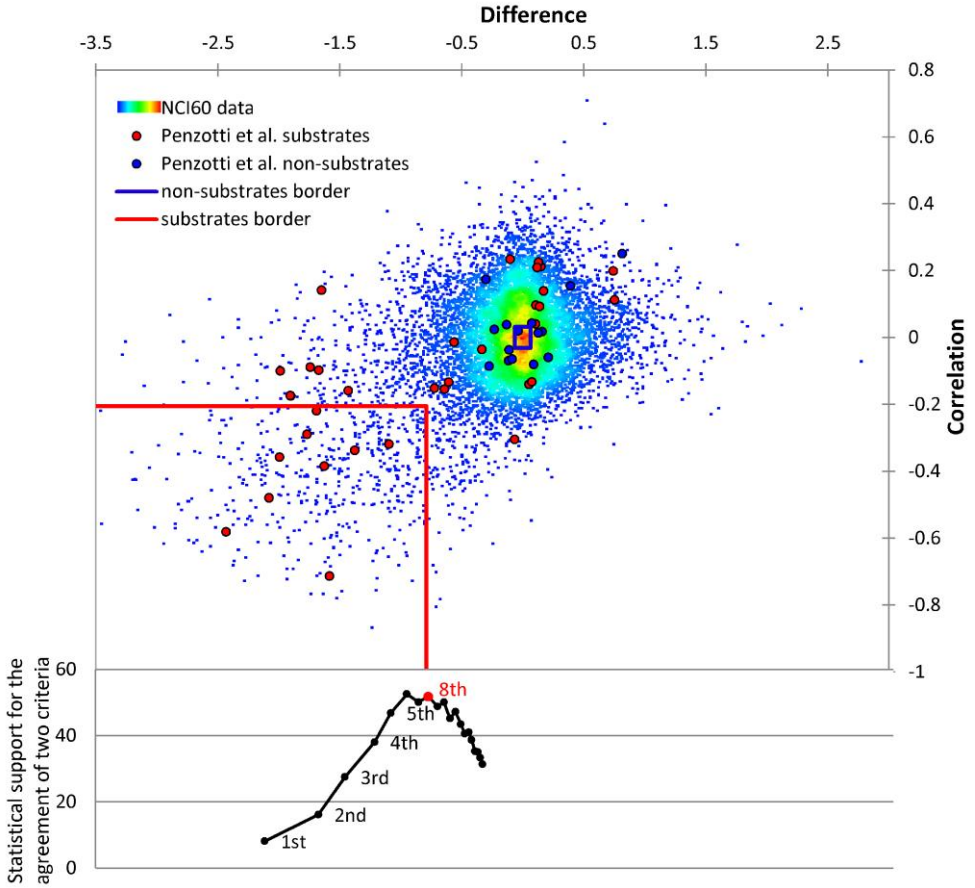


Figure 3. Constructing a high-confidence set of P-gp substrates and non-substrates from the set of 11 739 compounds tested on the NCI-60 panel. The two criteria for detecting P-gp substrates and non-substrates are: (i) the difference of cytotoxicity (log GI50) between the NCI/ADR-RES and OVCAR-8 cell lines, x axis, and (ii) correlation of the cytotoxicity and the ABCB1 gene expression over 58 cell lines, y axis. Areas of the plot with high density of points (compounds) are denoted by colored shading, from blue to red. The P-gp substrates are expected to show negative correlation and have negative differences, whereas non-substrates should show no such correlation or difference. The cutoffs for the optimal dataset (red lines for substrates, blue square for non-substrates) were selected to maximize the statistical support for the agreement between the two criteria as measured by Fisher's exact test P-value (bottom panel, y axis shows negative log10 of P-value, higher numbers indicate better support), while ensuring balanced class proportions, yielding 471 substrates and 487 non-substrates. The points in the bottom panel denote cutoffs at the 1st, 2nd, 3rd etc. percentile of the "correlation" and "difference" values. Overlapping compounds from the Penzotti *et al*. data[6] are shown as large blue and red dots.

An accurate model for recognizing P-gp substrates. We opted to use the SVM classifier with the Radial Basis Function (RBF) kernel, found to be the most appropriate method in former P-gp substrate classification studies.[8,10,11,20] Molecules were represented using 183 2D molecular descriptors generated with the Chemistry Development Kit (CDK).[34] Prior to training the SVM models, 120 randomly selected compounds were set aside as an independent test set, whereas the remaining 814 compounds served as a training set.

The SVM model performed well in predicting P-gp substrates, with an accuracy of 88.2% and AUC of 0.95 in cross-validation, indicating that the substrate/non-substrate classes are highly internally consistent with respect to the structural features of the molecules. The model showed a similar predictive performance on the test set (86.7% accuracy, AUC = 0.94) indicating a strong ability to generalize to unseen data. When compared with self-reported measures of model accuracy from six previous studies, our cross-validation accuracy is higher (Table 1), 88% vs. to our knowledge, at most, 81% previously. Our test set accuracy also exceeds previous models that have been evaluated on a fully independent test set (Table 1). Importantly, given an increase in accuracy combined with a considerably larger dataset (number of compounds), the statistical support for our model's predictions is orders of magnitude improved over past modeling efforts (Table 1). We additionally evaluated our model against two external validation sets - test sets exactly as used in Penzotti *et al*.[6] and Bikadi *et al*.[11] studies, and obtained accuracies of 64.7% and 78.1%, respectively. These numbers are close to the accuracies obtained on the same test sets in the original studies: 63% in Penzotti *et al*. and 75% in Bikadi *et al*., confirming the ability of our model to generalize to unseen data.

Table 1. Predictive accuracy of the SVM and rule-based models in this study compared to self-reported accuracies from previous studies.

	Cross Validation					Independent Test Set				
Model	Acc	MCC	NS Prec[a]	NS Recall[a]	P-value[b]	Acc	MCC	NS Prec[a]	NS Recall[a]	P-value[b]
our models										
Full SVM model	88%	0.76	86%	89%	$6*10^{-118}$	86%	0.74	82%	90%	$3*10^{-17}$
nAtom-specVol rule	74%	0.53	88%	57%	10^{-55}	66%	0.39	82%	49%	$2*10^{-5}$

descriptors based rule	80%	0.60	81%	80%	$3*10^{-69}$	77%	0.53	79%	77%	$5*10^{-9}$
effluxophore rule	78%	0.57	83%	74%	10^{-63}	74%	0.50	83%	66%	$4*10^{-8}$
previous work - models with independent test sets										
Penzotti *et al.*	80%	0.64	97%	71%	$3*10^{-16}$	62%	0.32	78%	50%	$3*10^{-2}$
Xue *et al.*	79%	0.60	76%	79%	$3*10^{-13}$	80%	0.48	67%	57%	$3*10^{-2}$
Cabrera *et al*.	80%	0.60	77%	78%	$5*10^{-15}$	78%	0.54	77%	78%	10^{-3}
Bikadi *et al*.	80%	0.60	80%	79%	10^{-14}	79%	0.57	76%	81%	$4*10^{-3}$
previous work - test sets not independent										
Huang *et al*.[c]	81%	-	-	-	-	90%[c]	0.80[c]	89%	89%	$4*10^{-7}$
Wang *et al*.[d]	75%	-	-	-	-	88%[d]	0.73[d]	92%	73%	$4*10^{-16}$

[a] The precision and recall scores are given for the non-substrate class. [b] The P-values are determined using a Fisher's exact test on the cross-validation or test set confusion matrices. [c] Huang *et al*.[20] used a feature selection scheme that chooses a set of descriptors to maximize the accuracy mostly on the test set (there called 'validation set', see Eq. 11 in Huang *et al*.[20]). Therefore, the reported accuracy cannot be considered to come from an independent test set since the modeling parameters were adjusted to better fit this test set. This is also evident in the reported test set accuracy greatly exceeding the cross-validation accuracy; here, the latter is likely to be a better estimate of model performance on unseen data. Consequently, the test set precision and recall are likely biased upwards. [d] In Wang *et al*.,[10] similarly to the Huang *et al*. case described above, descriptors appear to have been selected based on a set of compounds that included the test set, meaning the test set was not evaluated independently of the model training procedure. Additionally, Wang *et al*.[10] use a non-random train/test partitioning scheme (see Methods in Wang *et al*.[10]) that ensures that each compound in the test set has another molecule very similar to it in the training set. This is likely to bias the estimates of test accuracy upwards as such method sampling does not evaluate the models' ability to generalize across diverse compounds. Again, these factors would explain the unusual finding of a test set accuracy being significantly above the cross-validation accuracy.

Dominant role of molecular size and specific volume. Since the interpretability is among the main prerequisites of a practically usable QSAR model, we sought to develop a model relying on few molecular features, thus generating easily understandable guidelines which can be taken into consideration by medicinal chemists in the process of designing novel pharmaceuticals. By performing forward attribute selection using the SVM, we selected a combination of two most substrate/non-substrate discriminating features: the number of atoms (*nAtom*) and the volume of

a molecule (*VABC*, Table 2). The *nAtom* is significantly more predictive ($P = 3*10^{-4}$, t-test) than the next best CDK descriptor *apol* (sum of the atomic polarizabilities); similarly, the *nAtom+VABC* combination is more predictive than the *nAtom+nAtomP* (number of atoms in the largest pi chain) combination ($P = 5*10^{-4}$, t-test). These features are directly related to the ones previously reported to be important for P-gp-mediated transport: size, weight or bulkiness of molecule.[9,12,20,35] The molecular van der Waals volume[36] (*VABC*) is calculated as the sum of the contributions of all atoms, with negative contributions of the number of bonds involving non-hydrogen atoms and additionally for the number of rings, particularly for aromatic rings:

Eq. 1. V_{vdW} (Å^3/molecule) = ∑all atom contributions - $5.92N_B$ - $14.7R_A$ - $3.8R_{NA}$

where N_B stands for the number of bonds, R_A for the number of aromatic rings, R_{NA} the number of non-aromatic rings; atom contributions from Zhao *et al*.[36] are given in Table S2. Thus, *VABC* is highly correlated to the number of atoms (*nAtom*), allowing us to create a normalized version: "specific volume" (*specVol*), the volume per atom for the given compound. Note that this "specific volume" is not, in fact, a measure of molecular size, as it does not appreciably correlate to either *nAtom*, molecular weight or *VABC* (Figure S2); the *specVol* descriptor should rather be understood as a reciprocal of molecular density. An SVM model trained with only *nAtom* and *specVol* features yields a cross-validation accuracy of 75.4% and an AUC of 0.83. Plotting *nAtom* versus *specVol* (Figure 4) reveals that a simple rule combining the two variables can separate the majority (recall = 57.7%) of non-substrates with 88.7% precision (Figure 5a): compounds with less than 35 atoms (including H) or with specific volume more than 7.3 Å^3/atom are very likely to be non-substrates. Conversely, almost all P-gp substrates are large and dense molecules. The same rule can be successfully applied to compounds from other studies, yielding 78.0% precision for detecting non-substrates on the Penzotti *et al*.[6] data, and 76.3% precision for non-substrates on Bikadi *et al*.[11] data.

Table 2. The most informative single descriptors for P-gp substrate recognition in a SVM classification model.

CDK descriptor name	Description	Accuracy (std. dev.)[a]	AUC (std. dev.)[a]
nAtom	The number of atoms in a molecule (including H)	70.8% (0.2%)	0.796 (0.001)
apol	The sum of the atomic polarizabilities (including implicit hydrogens).	70.2% (0.1%)	0.787 (0.001)
bpol	The sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule (including implicit hydrogens)	69.2% (0.1%)	0.780 (0.001)
MLogP	Mannhold LogP;[37] a rough approximation of the LogP from the number of carbon and heteroatoms	68.2% (0.6%)	0.774 (0.001)
Additional descriptor			
JChem *LogP*[b]	LogP calculated with Instant JChem software[38]	58.0% (0.1%)	0.632 (0.002)

[a] The average and standard deviation are given for the accuracy (% correctly classified compounds) and the AUC scores obtained in 5 runs of 4-fold cross-validation testing of the SVM trained on single-descriptor datasets. [b] In addition to the four top ranked CDK descriptors (by cross-validation AUC, top part of table), a highly accurate LogP estimate[38] is included for comparison.

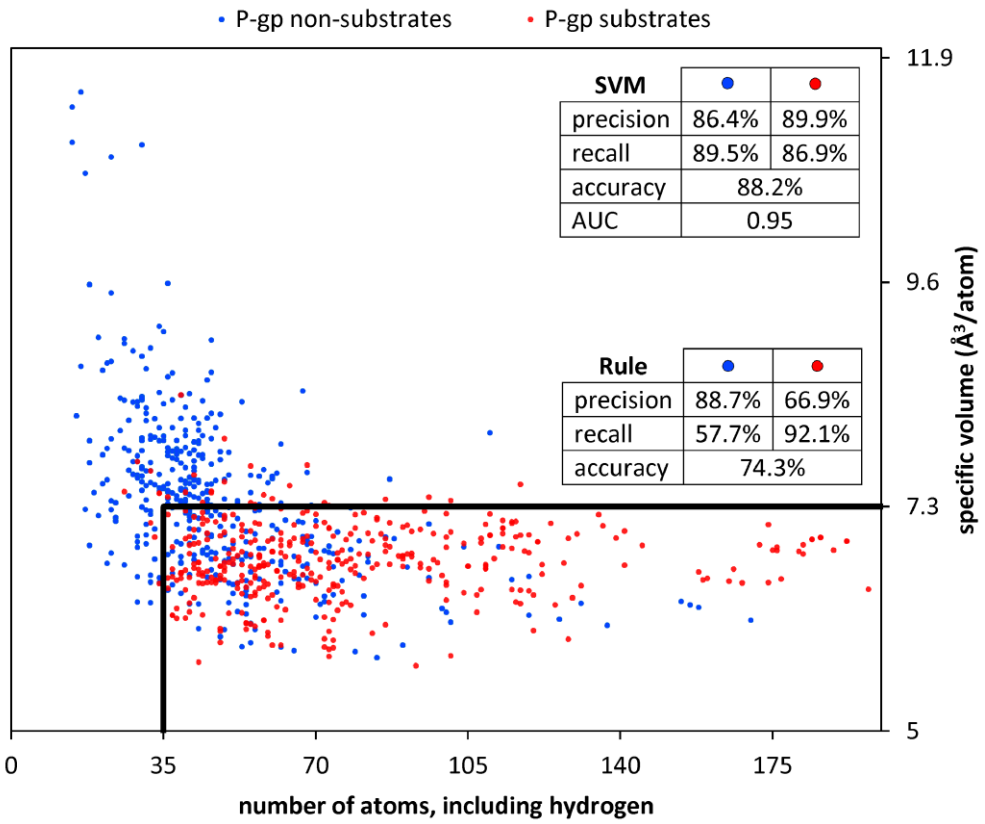


Figure 4. The number of atoms (*nAtom*) and the specific volume (*specVol*) discriminate the P-gp substrates from non-substrates. The two descriptors are complementary: by combining them into a simple rule (thick lines on the plot), the majority of non-substrates can be distinguished as compounds with *nAtom*<35 or *specVol*>7.3 Å^3/atom. All training set compounds are plotted as points, with the class denoted by the color. A single outlying non-substrate at (244, 6.423) is omitted for clarity. The overlaid tables contain cross-validation performance scores of the SVM classifier trained on the full set of 183 CDK descriptors (upper table), and of the *nAtom-specVol* rule shown on the plot (lower table).

Note that the dominant influence of the atom count and the specific volume in the model does not preclude that other molecular features have some bearing on the compounds' propensity for being a P-gp substrate; the rest of the CDK descriptors, when combined, are necessary to explain the difference in accuracy from 75.3% for the simple *nAtom-specVol* model to 88.2% in the complete model. An analysis of which single descriptors contribute most to accuracy of an SVM model when added on top of the *nAtom* and *specVol* (Table 3) has highlighted the importance of the number of atoms in largest pi chain (*nAtomP*), which alone contributes another 3.2% of accuracy. A rule-based model that includes the features complementary to *nAtom* and *specVol*

(Table 3) takes advantage of two extra descriptors to improve accuracy by 5% while remaining simple and interpretable (Figure 5b).

Table 3. Descriptors that best complement the number of atoms and molecular volume in SVM models.

CDK descriptor name	Description	Excess accuracy[a]	Excess AUC[a]
nAtomP	The number of atoms in the largest pi chain	+3.2%	+0.047
khs.sssCH	Kier-Hall smarts: [CD3H](-*)(-*)-*. The number of carbon atoms bound to three non-hydrogen atoms with single bonds	+3.4%	+0.043
khs.aaCH	Kier-Hall smarts: [C,c;D2H](:*):*. The number of aromatic carbon atoms bound to two non-hydrogen atoms	+1.9%	+0.039
HybRatio	Hybridization ratio: nsp3/(nsp3 + nsp2), considering carbon atoms only	+2.8%	+0.038
naAromAtom	The number of aromatic atoms of a molecule	+2.1%	+0.035
khs.ssssC	Kier-Hall smarts: [CD4H0](-*)(-*)(-*)-*. The number of carbons bound to exactly four non-hydrogens	+1.8%	+0.034
nAromBond	The number of aromatic bonds of a molecule	+1.9%	+0.034
additional descriptor			
JChem LogP[b]	LogP calculated with Instant JChem software[38]	+4.7%	+0.046
	Baseline:	75.9%	0.827

[a] Including these descriptors leads to the highest increase in SVM cross-validation accuracy (% correctly classified) and AUC scores when examined in addition to the baseline dataset which contained only the number of atoms (*nAtom*) and molecular volume (*VABC*). [b] In addition to the seven top ranked CDK descriptors (by excess cross-validation AUC, top part of table), a highly accurate LogP estimate[38] for comparison.

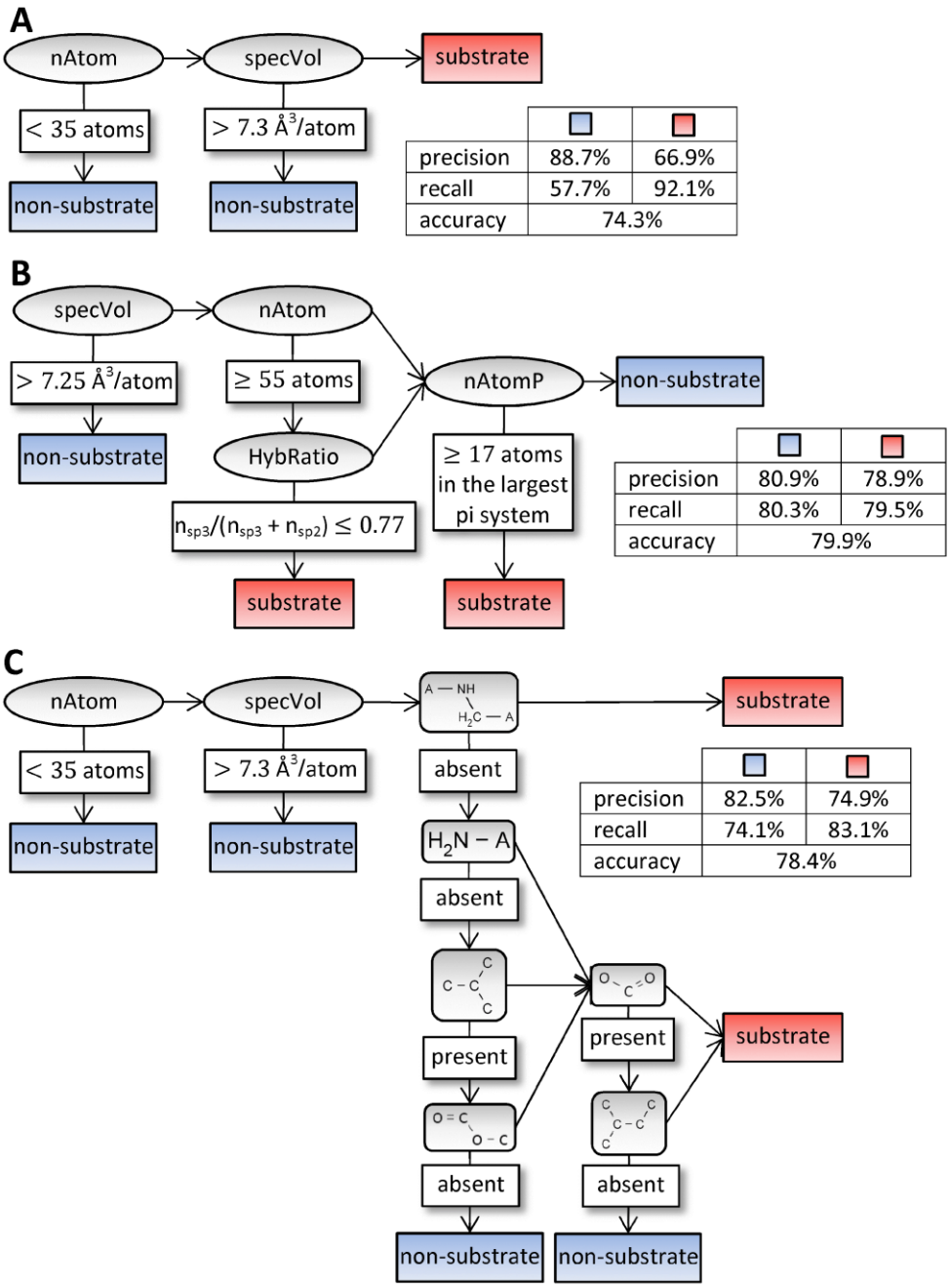


Figure 5. Rule-based models describing P-gp substrate specificity. The basic *nAtom-specVol* based rule for precise non-substrate separation (A) can be further enhanced with molecular descriptors that best complement molecular size and volume (Table 3), or with molecular fragments found to be significantly enriched at P-gp substrates or non-substrates (Table S3) which could not be separated by *nAtom* and *specVol*, giving more accurate models for P-gp substrate recognition (a molecular descriptors based rule B, and fragment based rule C). The symbol "A" at molecular fragments structures denotes any non-hydrogen atom.

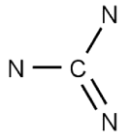
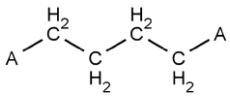
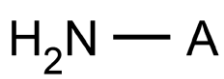
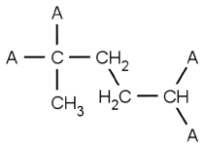
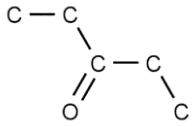
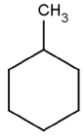
To more thoroughly evaluate the importance of molecular hydrophobicity, in addition to the CDK LogP estimates, we tested a highly accurate ChemAxon LogP estimate[38] and found it significantly predictive, but only after having controlled for the *nAtom* and *specVol* (Table 3). Alone, this descriptor is not highly predictive (Table 2), therefore the influence of LogP, while notable, is secondary to molecular size and specific volume.

Molecular fragments contributing to P-gp efflux. While the overall molecular properties, such as the size and density, greatly affect the P-gp substrate propensity of a molecule, it is possible that specific structural fragments could further fine-tune the molecular interactions with P-gp and modulate the efflux. The large size of our dataset enables us to test this hypothesis and search for such fragments. To this end, we created a subset of the substrates and non-substrates matched by their *nAtom* and *specVol* properties (see Methods), arriving at 376 compounds; this set roughly corresponds to the area in the *nAtom-specVol* plot (Figure 4) where the two classes overlap. The matching procedure allows us to find pharmacophores that facilitate (or prevent) P-gp efflux beyond their contribution to the *nAtom* or *specVol*. We found 7 such 'effluxophores' and 17 'anti-effluxophores' which are at least 2x enriched in either group ($P<0.002$ by Fisher's exact test, corresponds to false discovery rate=7.3%); Table 4, Table S3. The effluxophores contain amines, particularly secondary amines linked to short aliphatic chains, and aromatic rings, while the anti-effluxophores tend to describe aliphatic rings, or short saturated aliphatic chains with branching points or with oxygens (carbonyl or ether).

Based on the enrichment of these fragments in the substrate or non-substrate classes, we infer that excluding an effluxophore (or including an anti-effluxophore) during compound design and synthesis would lead to an increased probability of making a drug impervious to P-gp efflux. To computationally verify this principle, we tested whether including the information on occurrence of these molecular fragments improves the rule to discern substrates from non-substrates, while retaining its interpretability. Indeed, the rules including (anti)effluxophores have led to a substantial increase in recall to 74.1% (effluxophore rule, Figure 5c) for non-substrates, compared to the baseline recall of 57.7% for the *nAtom-specVol* rule (Figure 5a), while retaining precision at a high level (Table 1). Importantly, the merit of the simple *nAtom-specVol* rule should be viewed through its ability to summarize the principal determinants of small molecule recognition by P-gp, rather than through its predictive power. The extended rule models, or the

more complex SVM model, would serve better in scenarios where potential drug candidates are evaluated for P-gp substrate-likeness and high accuracy is desired.

Table 4. Examples of molecular fragments strongly enriched in P-gp substrates or non-substrates.

Substructures enriched in substrates			
 **p-value**[a]: 10^{-6} **enrichment**: ∞ **95% CI**[b]: N/A **excess accuracy**[c]: +2.59% **excess AUC**[c]: +0.020	 **p-value**: $2 \cdot 10^{-3}$ **enrichment**: 5.67 **95% CI** = [1.69, 19.02] **excess accuracy**: +0.84% **excess AUC**: +0.006	 **p-value**: $8 \cdot 10^{-4}$ **enrichment**: 3.71 **95% CI** = [1.65, 8.35] **excess accuracy**: +1.14% **excess AUC**: +0.006	 **p-value**: 10^{-5} **enrichment**: 3.21 **95% CI** = [1.83, 5.65] **excess accuracy**: +1.37% **excess AUC**: +0.012
Substructures enriched in non-substrates			
 **p-value**: $8 \cdot 10^{-4}$ **enrichment**: 14.0 **95% CI** = [1.86, 105.4] **excess accuracy**: +2.07% **excess AUC**: +0.023	 **p-value**: $2 \cdot 10^{-3}$ **enrichment**: 7.50 **95% CI** = [1.74, 32.34] **excess accuracy**: +1.22% **excess AUC**: +0.012	 **p-value**: $3 \cdot 10^{-5}$ **enrichment**: 6.50 **95% CI** = [2.31, 18.26] **excess accuracy**: +3.19% **excess AUC**: +0.028	 **p-value**: $9 \cdot 10^{-4}$ **enrichment**: 3.22 **95% CI** = [1.57, 6.62] **excess accuracy**: +2.89% **excess AUC**: +0.024

[a] P<0.002 by Fisher's exact test is required for declaring enrichment. Four fragments with the highest enrichment shown for substrates and for non-substrates; a more comprehensive list is in Table S3. The enrichment was tested in sets matched by *nAtom* and *specVol*, thus these fragments contribute to the substrate-non-substrate distinction beyond their influence on *nAtom* and *specVol*. The letter 'A' in molecular fragment structures denotes any non-hydrogen atom. [b] The 95% confidence interval for enrichment of this fragment could not be determined as the fragment was completely absent from all non-substrates in the matched sets. [c] Information about presence or absence of single molecular fragments improves the predictive performance of the SVM model when added to the simple *nAtom-specVol* model, and hence is valuable in discrimination between P-gp substrates and non-substrates.

Experimental validation using independent cell lines. To experimentally validate the utility of our SVM model in predicting P-gp substrate propensity, we employed an independent biological model system, consisting of a pair of cell lines not present in the NCI-60 screen: the commonly used HEp-2 cell line, which was previously[39] used to obtain a vincristine-resistant

derivative, the VK_2 cell line. We found high levels of the P-gp protein in VK_2 cells (Figure 6a), and, accordingly found these VK_2 cells were more resistant to known P-gp substrates vincristine, paclitaxel, and colchicine (Figure 6b) compared to the HEp-2 cells lines. This resistance could be abolished using P-gp inhibitors verapamil and CP100356 (Figure 6b). However, when treating the VK_2 cells with verified P-gp non-substrates such as 5-FU, mercaptopurine and chlorambucil, we observed similar cytotoxic effects in both VK_2 and HEp-2 (Figure 6c), where the addition of verapamil had no effect (Figure 6c). These positive and negative controls demonstrate that measuring growth inhibition in the HEp-2/VK_2 cell line pair is a good experimental assay for validating P-gp substrate propensity of compounds. Notably, while a negative result in this assay guarantees a P-gp non-substrate, the converse is not necessarily true: VK_2 might have other resistance mechanisms in addition to P-gp overexpression which would then yield a false positive result. We thus use the HEp-2/VK_2 comparison assay only to verify the non-substrate predictions of our computational SVM model.

To validate the accuracy of our model, we chose among the compounds that our SVM model predicts to be P-gp non-substrates, but that were labeled as P-gp substrates in previous investigations[6,11]. First, we chose three compounds with very high-confidence SVM predictions of being non-substrates: tramadol (SVM probability for non-substrate = 98.9%), estrone (98.8%) and digitoxigenin (98.7%); Figure 6d. In these three compounds, our cytotoxicity assay results show that VK_2 and HEp-2 exhibit similar sensitivity, suggesting they are, in fact, not P-gp substrates (Figure 6d). Accordingly, verapamil addition did not affect VK_2 drug sensitivity (Figure 6d).

Second, we chose three commonly used chemotherapeutics; Figure 6e. Previous studies[6,11] have labeled the topoisomerase poison topotecan as a P-gp substrate. Here, we test two of its close analogs that our SVM model predicts to be non-substrates: camptothecin and SN-38 (administered to the cells as the pro-drug irinotecan). We also included another topoisomerase inhibitor - etoposide - previously labeled as a P-gp substrate,[6,11] but predicted to be a non-substrate by the SVM. In all test cases, there was no difference in activity against HEp-2 and VK_2 cells, and co-administering verapamil with the drugs had no effect on VK_2 cell growth (Figure 6e), thus confirming our prediction. In summary, we have experimentally validated our SVM model's ability to recognize P-gp non-substrates and provided evidence that the previous assignment of given six compounds to the P-gp substrate class should be reconsidered.

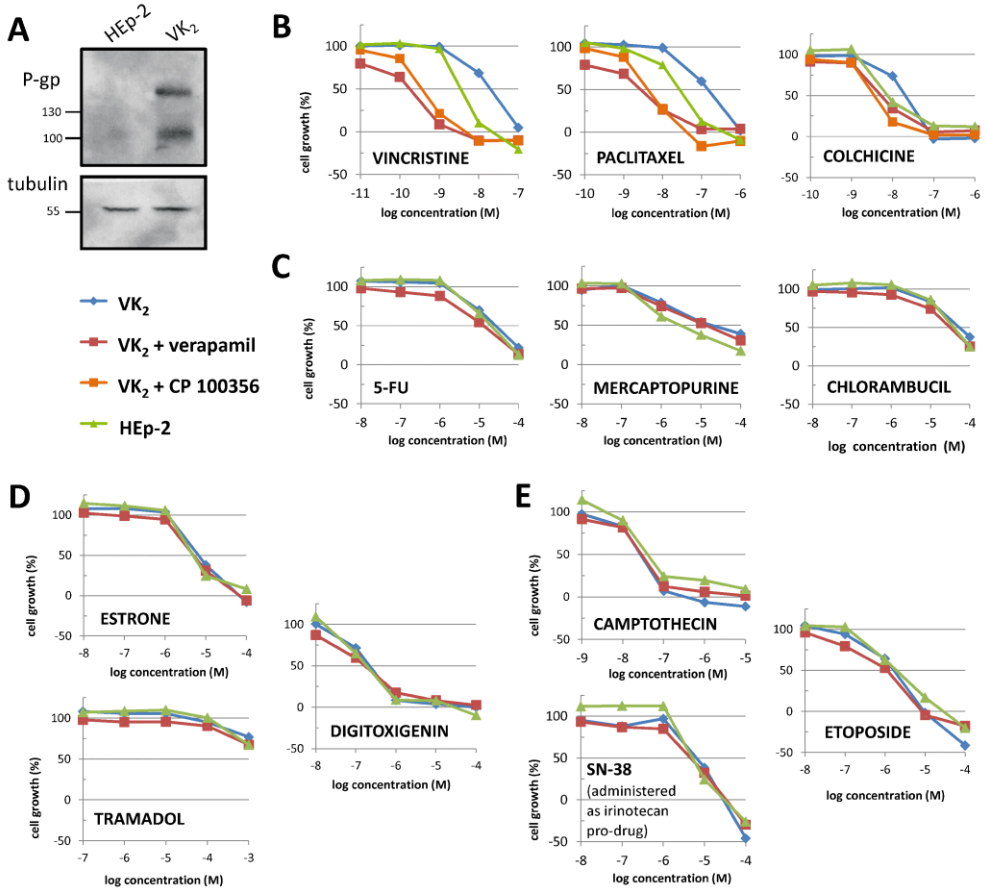


Figure 6. Sensitivity of the P-gp overexpressing cell line VK_2 to P-gp substrates and non-substrates. (A) The vincristine-resistant cell line VK_2[39] strongly overexpresses the P-gp protein, relative to its parental HEp-2 cell line. (B-E) Dose-response curves for VK_2 and HEp-2 cell lines grown with or without P-gp inhibitors verapamil (all panels) or CP 100356 (panel B). All points are averages of at least two biological replicates, each done in a technical quadruplicate. (B) Known P-gp substrates vincristine, paclitaxel and colchicine. (C) Known P-gp non-substrates 5-FU, mercaptopurine and chlorambucil. (D) Three compounds, estrone, tramadol and digitoxigenin, selected for having a high-confidence non-substrate prediction by our SVM model, but labeled as a P-gp substrates in previous datasets .[6,11] (E) Three anticancer drugs selected for being of broad therapeutic interest, predicted to be P-gp non-substrates by our SVM model: camptothecin and SN-38, analogues of the drug topotecan, labeled as a P-gp substrate in previous datasets,[6,11] and the similarly labeled etoposide.

A P-gp Web server. To facilitate the re-use of our P-gp substrate propensity models by colleagues working in the field and medicinal chemists in general, we have implemented the SVM model, rule-based models and effluxophore detection in a Web server at http://pgp.biozyne.com. The Biozyne P-gp server is free for non-commercial use.

Strategies for designing P-gp (non)substrate compounds. The most accurate way to predict P-gp substrate propensity is using our SVM classifier (Table 1) that combines many descriptors in a complex model. Medicinal chemists could optimize their existing compound library by designing novel variants *in silico* and submitting them to the SVM (implemented on our Web server, or elsewhere) to quickly predict if the alternative structures are substantially less/more likely to be a P-gp substrate. Then, the hypothetical structures with the desired properties can be synthesized and tested.

Alternatively, a simpler approach would be to add, remove, or modify an "effluxophore" functional group (Table 4, Table S3) from an existing structure. Further extrapolating this concept, molecules could also be modified to optimize the overall molecular features -- a prominent representative being the specific volume, *specVol* (Figure 5) -- to alter the P-gp propensity.

To investigate the feasibility of these design principles, we searched for pairs of compounds that have substantial structural similarity overall (Tanimoto coefficient ≥ 0.85), while differing specifically in one of the effluxophores; we found 10 such pairs (examples in Figure 7A and all pairs in Table S4). Similarly, we searched for pairs that are overall similar but differ in their respective *specVol* value; 6 pairs were identified (Figure 7C, Table S5). One member of each pair was chosen amongst 448 known P-gp substrates in our training set, while the other, matched compound in the pair was chosen from any of the 12998 compounds available from the NCI-60 screen.

Within these structure-matched pairs, we compared the compounds' differential biological effect on the ADR-RES *versus* OVCAR-8 cell line, as well as across the other 58 cell lines. The compounds lacking a P-gp efflux promoting group consistently displayed both (a) a weaker differential activity on ADR-RES (P=0.0002, paired t-test, two-tailed, Fig. 7B), and (b) a weaker correlation over 58 cell lines, when compared their counterparts that do possess an efflux-promoting group (P=0.06, Fig. 7B). An analogous relationship is readily observable where increasing the *specVol* of a P-gp substrate consistently decreases the P-gp's propensity for the compounds judged by both biological criteria (P=0.02 and 0.013 in Fig. 7D). Thus, introducing or abolishing the functional groups (Table 4, Table S3), and changing the molecular specific volume, are two viable strategies to circumvent (or promote, if so desired) P-gp targeting.

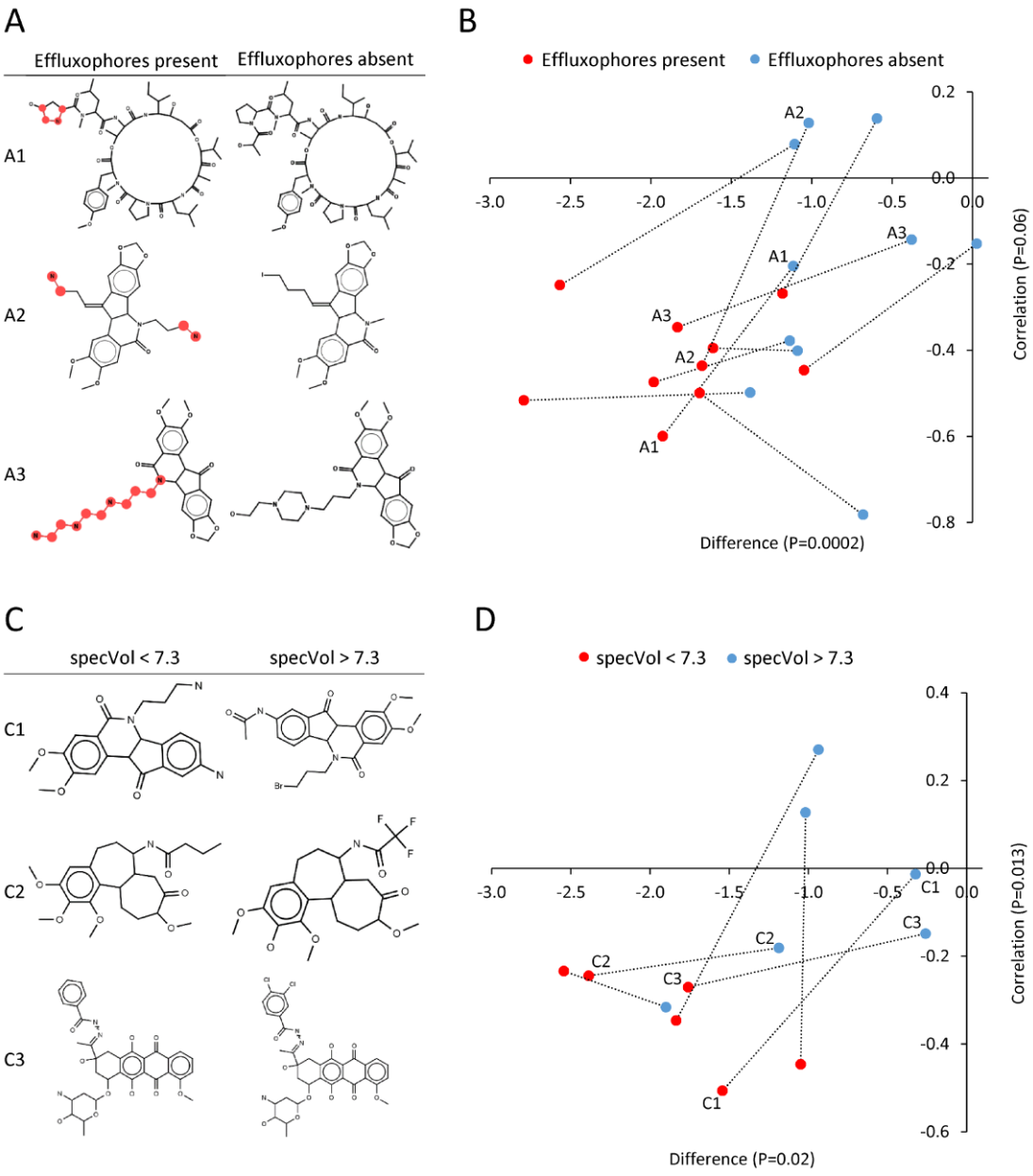


Figure 7. Structurally similar pairs of compounds which differ specifically in one of the effluxophores (A) show differential biological activity (B). The compounds lacking a P-gp efflux-promoting group consistently display a weaker differential activity on ADR-RES *vs*. OVCAR-8 cell lines (″difference", *x* axis) and a weaker correlation with P-gp expression over 58 cell lines ("correlation", *y* axis). Similarly, within pairs of structurally similar compounds which differ in the value of the *specVol* descriptor (C), increasing the *specVol* of a P-gp substrate consistently decreased the P-gp's affinity for the compounds, as judged by both biological criteria (D). P values are by paired t-test.

DISCUSSION AND CONCLUSION

Prior QSAR studies that developed computational models to discriminate the P-gp substrates from non-substrates depended on relatively small compound training sets that were largely re-used from one study to the next (Figure 1). It is perhaps hard to expect a significant step forward in understanding the regularities underlying a compound's recognition by P-gp if only minor improvements have been made in the publicly available sets of compounds, even though the modeling approaches have become increasingly sophisticated and accurate.

Here, we aim to address this situation by mining the most comprehensive public database of biological activity of compounds against a panel of 60 human cancer cell lines, the NCI-DTP database. Extensive information is available from this database including ABCB1 mRNA levels that were derived through multiple independent experimental measurements. We used this information to search for compounds whose biological activity strongly anti-correlates to ABCB1 expression (the putative substrates), or is completely uncorrelated to ABCB1 expression (the putative non-substrates). In principle, the approach to create a QSAR compound series from existing cell line cytotoxicity screening data could be applied to other target genes as well, given there is sufficient information of the gene expression across multiple cell lines. Our work thus exemplifies an useful scheme of linking the biological characteristics of cell lines (here, the expression data), the cells' response to compounds and the compounds' structural features. Recently, two comprehensive screens using hundreds of cell lines (~10x larger than the NCI-60 panel) have been performed, while simultaneously characterizing the genomic, epigenetic and transcriptomic properties of the cells - the "Cancer Cell Line Encyclopaedia",[40] and the "COSMIC Cell Line Project".[41] As more compounds are processed in these new screens, they will become invaluable for QSAR studies targeted to many different genes or pathways with relevance to cancer and other disease.

Here, we paid special attention to the P-gp overexpressing NCI/ADR-RES multidrug resistant cell line which, conveniently, has its non-resistant counterpart (the OVCAR8) also present in the NCI-60 panel.[32,33] Thus, the NCI/ADR-RES - OVCAR8 pair presents a unique opportunity to establish a second criterion to verify the putative substrates/non-substrates detected by the biological activity signature on the other 58 cell lines.

Our data mining approach to constructing this P-gp substrate dataset could be considered as a more general case of the use of cytotoxicity assays, widely used to test P-gp substrates (reviewed

in Szakács *et al*.[5] and Didziapetris *et al*.[12]). Therein lies a limitation of our approach: assays from this class are applicable only to compounds that have some cytotoxic activity. In particular, we require at least a small part of the NCI-60 cell line panel (12/60 cell lines) to have growth inhibitory activity above the detection threshold (typically meaning that $GI50<10^{-4}$ M; see Methods). Note that this requirement is not too stringent: some commonly used (not anti-cancer) drugs meet this criterion, including the antidepressant nortriptyline, the antispasmodic cyclobenzaprine, and the anti-protozoal emetine. Nevertheless, given that our SVM and rule-based models were trained on compounds with cytotoxic activity, it seems prudent to recommend their application on other cytotoxic compounds. In this context, development of QSAR models that would help direct the design of antitumor drug candidates to safeguard them against P-gp activity seems timely, given the recent biological research on 'cancer stem cells' that cause relapses of cancer after therapy due to their resilience against many drugs in clinical use today,[42] in-part due to P-gp activity.[43]

Our approach aims to take advantage of the natural variability in P-gp levels across the NCI-60 panel compared to previous work where a cell line may be engineered to over- or under-express the ABCB1 gene to detect P-gp substrates. Such an approach was validated in a previous study[28] precisely on the NCI-60 cells, where mRNA levels were measured by quantitative PCR for ABCB1 and other drug transporters, and six compounds with strong ABCB1 mRNA-cytotoxic anti-correlations were experimentally verified as novel P-gp substrates. The authors noted that relying on mRNA levels as a proxy for P-gp activity has its caveats. For instance, mRNA levels are not linearly indicative of protein levels, there are additional complex layers of regulation between mRNA and translation into the protein product,[44] such as mRNA secondary structure and codon biases;[45] however, the mRNA and protein levels do agree quite well for ABCB1/P-gp across the NCI-60 panel.[28] Additionally, given that ABCB1 is a member of the large ABC transporter family, some substrate specificity overlap is expected between other family members. Consequently, the expression levels of other ABC transporters may confound our analysis. Still, this effect should be mild since ABCB1 has a dominant role in drug efflux over other transporters:[28] the other two relevant multidrug transporters ABCC1 (MRP1) and ABCG2 (MXR or BCRP) exhibited weaker correlations of mRNA and activity profiles than the ABCB1.[28] Furthermore, given our set of eleven ABCB1 mRNA level measurements (this includes the quantitative PCR measurements from Szakács *et al*.[28] but also 10 additional ABCB1 mRNA data

points per cell line), we found a striking difference in the mRNA levels of ABCB1 between the NCI/ADR-RES and OVCAR-8 cells, but not other ABC transporters (Figure 2c). The strongly elevated ABCB1 expression is supported by a previous analysis of gene copy numbers in various cell lines,[46,47] that found a sharp peak in gene amplification in the region of chromosome 7 corresponding to the location of the ABCB1 gene (at ~87 megabases) in the ADR-RES cells,[48] but not in the OVCAR-8 cells.[49]

By exploiting the sizeable NCI-DTP database containing tens of thousands of compounds, together with the knowledge of the cellular expression levels of the ABCB1 gene, we arrived at a dataset of 934 P-gp substrates and non-substrates, the largest publicly available collection. The size of the dataset allows us to search for principal determinants of P-gp recognition in the structure of the molecules with greatly increased statistical support over the previous attempts. This point might be especially relevant given that the molecular features previously described as characteristic of P-gp substrates are typically inter-correlated and thus a larger dataset allows them to be disentangled with more confidence. Our feature selection scheme that employed the SVM has singled out the size of the molecule, measured by the number of atoms (*nAtom*), as the most informative single descriptor for P-gp substrate recognition. Among the other descriptors, the one that complemented *nAtom* best was found to be the molecular volume, or equivalently, its normalized version, the specific volume (*specVol*, in $Å^3$/atom; note that this is equivalent to reciprocal density).

The salient role of the *nAtom* descriptor is not surprising, given that many previous studies have highlighted features describing various aspects of the size of the molecule: molecular weight,[12,20,50] surface area,[27,51] or bulkiness.[9,35] To our knowledge, the predictive value of the *specVol* - the specific van der Waals volume per atom (uncorrelated to overall volume, Figure S2) - for P-gp recognition was not previously discussed. This property is easily calculated using a simple formula for determining molecular volume (Eq. 1[36]) which, however, does reflect other molecular features said to be relevant for P-gp transport in prior research. Most prominently, the formula for volume has a strong contribution of the number of aromatic rings, and aromaticity of molecules has been discussed as relevant.[6,52,53] Next, the volume also reflects flexibility of the molecule: as seen from the formula (Eq. 1), rings (even if not aromatic) reduce both the calculated volume (thus also the *specVol*) and the flexibility. Finally, the number of hydrogen bond donors and/or acceptors (N and O), widely cited as important for P-gp recognition,[7,12,50,54]

also has a bearing on the *specVol* as N and O have a lower contribution to the molecular volume than C (Table S2). Since all of the above variables contribute to the specific volume of the molecule, and are thus correlated to a degree, it is not straightforward to determine the one most likely to be the main causal factor behind P-gp recognition. A larger dataset that offers more statistical support for analyses, combined with a state-of-the-art machine learning method - the SVM - allows us to get a step closer towards an accurate model: in our data, the number of atoms and the specific volume has proven to be the principal determinants of P-gp recognition. A simple, convenient rule derived from these features separates the majority of the P-gp non-substrates with very high precision (Figure 4): compounds with *nAtom*<35 or *specVol*>7.3 $Å^3$/atom are very likely to be non-substrates.

Still, these two variables are not sufficient to reproduce the accuracy of the SVM model using the full set of 183 molecular descriptors (Table 1), meaning other molecular features contribute to classification of P-gp substrates, in addition to *nAtom* and *specVol*. The most relevant single descriptors (Table 2) are connected with size of conjugated systems and aromaticity - possibly indirectly describing condensed aromatic rings - followed by the degree of branching and the saturation in the molecule. A feature that deserves special mention is the LogP. The importance of LogP for discriminating P-gp substrates from non-substrates in QSAR studies has been contentious: it was claimed to be either very important,[10,25,26] or of little relevance.[12,20,27] It is possible that LogP could have been declared as relevant in past studies, when in fact it found as such only as it was correlated to some other feature relevant for P-gp recognition, for instance aromaticity. Closer examinations of the mechanism of P-gp activity have indicated that P-gp recognizes and binds substrates that are dissolved in the membrane, implying that the membrane solubility, and consequently the LogP, determines substrate recognition.[25,55] On the other hand, ligand docking simulations into the mouse P-gp crystal structure[56] have indicated that the hydrophobicity has little bearing on the thermodynamics of transferring a ligand from water to the P-gp binding site. Our data supports the importance of LogP for discrimination of P-gp substrates from non-substrates, but only after controlling for the number of atoms and specific volume. As a single feature, LogP was not highly informative; thus, the number of atoms and the specific volume are dominant in distinguishing the P-gp substrates.

While the global physicochemical properties of molecules, such as the size, the aromaticity, or the number of hydrogen bond donors/acceptors, may prove to be sufficient to roughly describe

how P-gp recognizes its substrates, there are finer points to be made to more accurately characterize the process of recognition.[50] For instance, the P-gp substrates may be recognizable by a spatial separation of the electron donor groups,[7] or possibly through a more complex definition of P-gp substrate pharmacophore points involving hydrophobic groups, aromatic rings, H-bond acceptors and donors.[54,57] Recent studies have taken a more statistically-inspired approach, searching for molecular fragments enriched in substrates or non-substrates.[10,19,58] Such fragments can be easily and quickly identified from databases of 2D structures without needing to know or simulate the spatial conformation(s) of the compound. Additionally, a simple set of rules that includes fragments to avoid or to include is easier to follow for chemists in the stage of molecular design. We have thus followed a similar path in discovering P-gp 'effluxophores' and 'anti-effluxophores' (Table 4, Table S3) by enrichment in substrates and non-substrates, respectively, with two important differences over past work. First, we searched for enriched fragments only in substrate-non-substrate pairs matched in two global properties of molecules found to be important for recognition (*nAtom* and *specVol*), thus controlling for the confounding effect of these properties. In other words, the fragments we identified are guaranteed to contribute to P-gp recognition independently of *nAtom* and *specVol*. Second, a larger dataset allows us to impose a strict check for statistical significance of the enrichment, ensuring a low false positive rate of fragments detected to be relevant. Third, we verified that these features (presence of absence of fragments) actually are important for discrimination of P-gp substrates from non-substrates by including them in the baseline *nAtom-specVol* SVM model: the simple, interpretable features contribute significantly to the classification accuracy (Table 4). A subset of the fragments can be selected to refine the simple *nAtom-specVol* rule (Figure 5), providing a more accurate yet still straightforward guide for synthesis of compounds more likely to be non-substrates.

In addition to predicting P-gp substrates, another related task is predicting compounds that inhibit P-gp activity, which have, for instance, been clinically investigated as adjunct therapy for some cancers.[59] The specific SVM model we have trained on a set of P-gp substrates and non-substrates cannot be used to predict P-gp inhibitors - past work has shown these are two separate modeling tasks.[60–62] In particular, many P-gp inhibitors are allosteric - they do not bind to the substrate site and are consequently unlikely to resemble a P-gp substrate.[60] Additionally, some P-gp substrates do not competitively inhibit the transport of other substrates.[61] Accordingly,

structural features were found to substantially differ between the compounds in the two sets: a simple unsupervised learning approach was successfully used to distinguish the P-gp substrates from the inhibitors.[62] For researchers interested in predicting P-gp inhibitors, a recent paper by Broccatelli *et al*.[63] introduces a large set of compounds, used to derive highly predictive QSAR models; our work aims to do the same for P-gp substrate prediction.

In conclusion, our work exploits a readily available source of high-throughput biological screening data on cancer cell lines to extract a novel set of P-gp substrates and non-substrates. The stringent internal consistency checks and an emphasis on maximizing statistical support in the compound selection process guarantees a high quality dataset that agrees with previously known P-gp substrates. The enlarged number of available compounds over previous publicly available datasets allowed us to construct an accurate and statistically sound SVM classification model, which we experimentally validate by using an independent biological test system.

A search for relevant features singled out the number of atoms and the specific atomic volume as the principal determinants of P-gp recognition: all P-gp substrates are large, dense molecules. In addition to these features, a number of molecular fragments - 'effluxophores' and 'anti-effluxophores' - were found to discriminate the P-gp substrates from the non-substrates. The derived rule-based models are interpretable, yet retain their accuracy, and can help direct synthesis of novel anticancer agents that would circumvent P-gp mediated resistance to therapy.

MATERIALS AND METHODS

NCI-DTP cell line screening data. We downloaded the September 2010 release of the human tumor cell line screen database from the National Cancer Institute's Developmental Therapeutics (NCI-DTP) program (http://dtp.nci.nih.gov/docs/cancer/cancer_data.html). From the database, we used the -logGI50 - GI50 is the concentration of each compound that retards cell growth by 50% - as a measurement of activity of a compound against a given cell line. The NCI-DTP typically tests a compound once on each of the 60 used cell lines. Some compounds may not have data for some cell line; we allow a maximum of 3 such missing measurements for a compound. Most compounds are tested at the largest concentration of 10^{-4} M; if the compound has very little or no activity at this concentration, the NCI-DTP database provides this biggest tested concentration as the 'default' GI50 value. Our dataset contained only the compounds where at least 10 cell lines had non-default measurements. In practice, this means that the

compounds with no appreciable cytostatic activity on at least 10 cell lines were filtered out, leaving a total of 12998 compounds of the initial ~43000. Due to the special importance of the OVCAR8 and NCI/ADR-RES cell lines, we imposed the additional condition that the measurements for these two cell lines must not be missing or default (i.e. no activity), leaving 11739 compounds. Some compounds may be tested in more than one experiment, in which case we take the average GI50 value over the experiments, except in the case of 'default' GI50 values which don't contribute to the averages. The GI50 datasets and scripts that perform the filtering are available on request from the authors.

Collecting and preprocessing ABCB1 expression data. The cellular expression levels of P-gp mRNA were determined using data downloaded from the August 2010 release of the NCI-DTP molecular target data available at http://dtp.nci.nih.gov/mtargets/download.html. In particular, we extracted P-gp expression measurements from the following Affymetrix microarray data sets: WEB_DATA_CHIRON.ZIP (probes 243951_at, 209993_at and 209994_s_at), WEB_DATA_GENELOGIC_U133.ZIP (same three probes), WEB_DATA_NOVARTIS.ZIP (probes 1575_at and 1576_g_at), and the mRNA levels measured by quantitative PCR from WEB_DATA_ALL_MT.ZIP (columns MT79,[29] MT1614[64] and MT2663[28]). In total, we had 11 P-gp mRNA level measurements for each cell line, with a small number of missing values. To summarize the 11 measurements into a single quantity, we performed a principal components analysis on this 11 P-gp expression x 60 cell lines table using the XLStat 2010 software (Addinsoft, Paris, France); Spearman correlations between cell lines were used. The first principal component (PC1) is the direction of strongest variability between cell lines in the full dataset and alone describes 51.9% variance of the 11 measurements. PC1 was used as a summary for P-gp expression level; it strongly correlated to the 10 of the 11 mRNA level measurements (r = 0.63-0.85, average r=0.75). Expectedly, the NCI/ADR-RES cell line had by far the highest PC1 value among the 60 cell lines (PC1=20.2), while the OVCAR8 cell line had a below-average P-gp expression at PC1=-0.5 (arbitrary units, average PC1 over cell lines=0.0, standard deviation=2.9). The cell lines from the NCI-DTP panel previously known to be naturally multi-drug resistant are HCT-15, UO-31 and TK-10,[65] and all have elevated PC1 values in our data: rank 2/60, 4/60 and 10/60, respectively. P-gp expression data for the cell lines is available on request.

Forming the "substrate" and "non-substrate" compound classes. The "substrate" and "non-substrate" classes were formed according to the "difference" and "correlation" criteria: difference of cytostatic activities (-log GI50 values) of NCI/ADR-RES and OVCAR-8 cell lines, and Pearson's correlation coefficient between the cytostatic activity and the ABCB1 gene expression (summarization of the eleven different measurements of ABCB1 mRNA levels, Figure 2a) over the remaining 58 cell lines in the NCI60 panel. The thresholds for the substrate class were iterated through: 1st, 2nd, 3rd etc. percentile of the data distribution for the "correlation" and "difference" criteria independently. In other words, compounds with "correlation" and "difference" above than n-th percentile were considered substrates. In the case of non-substrate class, the thresholds were set (centered around zero) so that the resulting number of non-substrate compounds approximately matched the number of substrates, ensuring balanced class proportions. At each iteration, the statistical support for the agreement between the two independent criteria was measured by the Fisher's exact test P-value. The final cutoff values were chosen at the 8th percentile for substrates and at the mid-20% of compounds for non-substrates, which is the point where the number of substrates and non-substrates is maximized, while keeping the Fisher's test P-value near-optimal (Figure 3). The resulting dataset consists of 958 compounds: 471 substrates and 487 non-substrates. Please refer to Figure 8 for a schematic overview of how the dataset was constructed.

Preprocessing of chemical structures. The compounds structures were downloaded from the NCI website (December 2010 release, [66]) and represented using SMILES strings. Prior to computation of the molecular descriptors, the SMILES were preprocessed using OpenBabel[67]: hydrogens were made implicit and salts were striped. In order to detect highly similar compounds in our dataset, pairwise Tanimoto similarity coefficients for Extended Chemistry Development Kit (CDK)[34] fingerprints were calculated. For each pair of compounds that share a Tanimoto similarity of 1 (identical or nearly identical molecules, for examples see Table S1), the one with the lower confidence was removed from the dataset. The confidence score was measured as the proximity to the reference point (minimum for substrates, zero for non-substrates) considering the "difference" and "correlation" criteria. In total, 24 compounds were removed from the dataset in this step. The compounds' structures in SMILES format and their substrate/nonsubstrate labels are listed in the Supporting Information (Table S6) or available for download at http://pgp.biozyne.com.

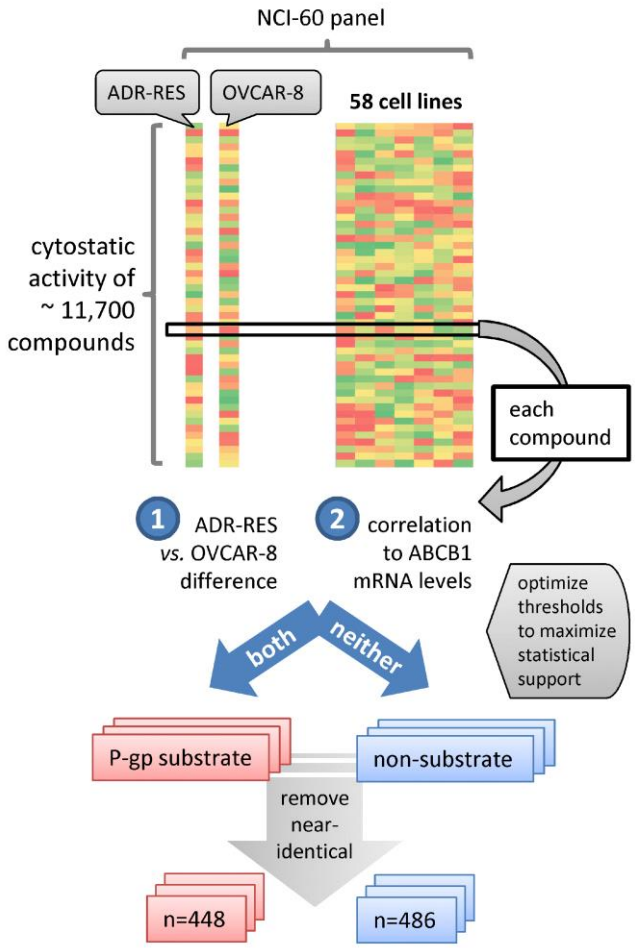


Figure 8. A schematic representation of the methods employed in our analysis, where the NCI-60 cancer cell line cytotoxicity screening database is combined with known ABCB1 expression levels and used to form a high-confidence dataset of 448 P-gp substrates and 486 P-gp non-substrates.

SVM training and performance measures. The molecules were represented by 183 2D molecular descriptors calculated with Chemistry Development Kit (CDK).[34] The 934 compounds were split into a test set (120 randomly chosen compounds) and a training set (814 compounds of which 2 were not accepted by CDK when calculating the descriptors, so the effective training set size was 812). To train the SVM model, we used the LIBSVM software[68] adapted for the Weka data mining environment.[69] Following the recommendation of LIBSVM authors,[70] we used with a Radial Basis Function (RBF) kernel, while optimizing the c (from 2^{-5} to 2^{20}) and gamma (from 2^{-15} to 2^{5}) parameters in a grid search procedure. The model with the best Area Under ROC Curve (AUC) was obtained for $c=2^{3}$ and $\gamma=2^{-2}$; AUC=0.95. For model validation 4-fold cross validation was used, and to get more stable results latter was repeated 5 times with different

random initialization. The RBF (or Gaussian) kernel, $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, is a popular kernel function which allows the SVM to represent complex nonlinear relationships in the data by effectively adding a smooth "bump" around each data point. The choice of kernel influences the predictions by defining the manner in which the data points are deemed similar or dissimilar in. To ensure that the choice of RBF kernel was optimal for our modelling effort, we also repeat the above-described grid search procedure for the polynomial kernel while varying its two parameters: degree (from 1 to 5) and c (from 2^{-5} to 2^{15}). The highest cross-validation AUC for the polynomial kernel was obtained for c=2^{-1} and degree=2; the AUC was = 0.93 and the accuracy 87.6%, both somewhat lower than the accuracy obtained with the RBF kernel (see above) and we thus choose to keep the RBF kernel SVM as our primary classifier. Reassuringly, the individual predictions obtained with the RBF kernel and the poly kernel agree very well: for the 120 test set compounds, the P-gp substrate probabilities obtained with the best RBF SVM model and the best poly SVM model correlate with R^2 = 0.90 (see Figure S3). Thus, the predictions are robust to the exact choice of the kernel and there is little kernel-dependent bias in our results.

The predictive performance of our models was evaluated based on the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) by the following performance measures: accuracy (Acc); precision and recall - class specific performance measures; the Matthews correlation coefficient (MCC) - provides a balanced evaluation of prediction, where MCC of 1 indicates perfect, and MCC of 0 random prediction. The measures are defined as follows:

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

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

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

The AUC score is the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance, AUC of 1 indicates perfect prediction, and AUC of 0.5 random guessing; for details, see Fawcett[71]]. Final scores are averages over five runs. Additionally, statistical significance of results was measured by P value of Fisher's exact test. The best model was selected according to the cross-validation AUC score, and then additionally evaluated on the independent test set. The test set was in no way used to decide on any parameters used during the training phase.

Feature selection. To select the most important molecular descriptors in P-gp substrate recognition, we used a forward feature selection scheme employing an SVM. The full, 183 descriptor dataset was divided into datasets containing single descriptors and evaluated according to the cross-validation scheme as described above. The AUC and the accuracy of the SVM trained on the single descriptor were a measure of relevance for the descriptor. In the second iteration, the best descriptor from the first iteration remained fixed, while all pairwise combinations with other descriptors added one by one were evaluated for cross-validation performance of the SVM model.

Enrichment of pharmacophores and training of rule-based models. To control for the two most important properties singled out by the feature selection scheme, the number of atoms (*nAtom*) and specific volume (*specVol*; in $Å^3$/atom), matched pairs of compounds were created as follows: for each substrate, a corresponding non-substrate that lies within a radius of 0.04 (pairwise substrate-non-substrate distances were calculated using [0,1] normalized *nAtom* and *specVol* attributes), with prerequisite that each compound can be used in substrate-non-substrate pairs at most once. For 376 compounds matched in such a way, presence of 4860 chemical substructures associated with bioactivity reported by Klekota and Roth[72] was determined using OpenBabel.[67] For each substructure present at least once the size of enrichment was calculated as ratio of frequencies of occurrences at substrate and non-substrate group. Finally substructures that are at least 2x enriched in one of the groups and are statistically significant ($P<0.002$ by Fisher's exact test) were considered important in P-gp mediated efflux. In total 7 substructures enriched in P-gp substrates, and 14 enriched in P-gp non-substrates met these criteria (Table 4, Table S3). Information about presence or absence of fragments which occur in at least 10% of compounds (38/376) was used to further enhance the *nAtom-specVol* rule (Figure 5a). The "Repeated Incremental Pruning to Produce Error Reduction" (RIPPER)[73] rule learning algorithm

implemented in Weka[74] was trained on subset of molecules with *nAtom*>=35 and *specVol*<=7.3 to give sets of rules which were then merged with *nAtom-specVol* rule, giving more accurate rule based model (Figure 5c). The molecular descriptors which best complement the molecular size and volume (Table 3) were used to train an additional rule based model (Figure 5b). A schematic overview of our modeling workflow, including the SVM model and others, is given in Figure 9.

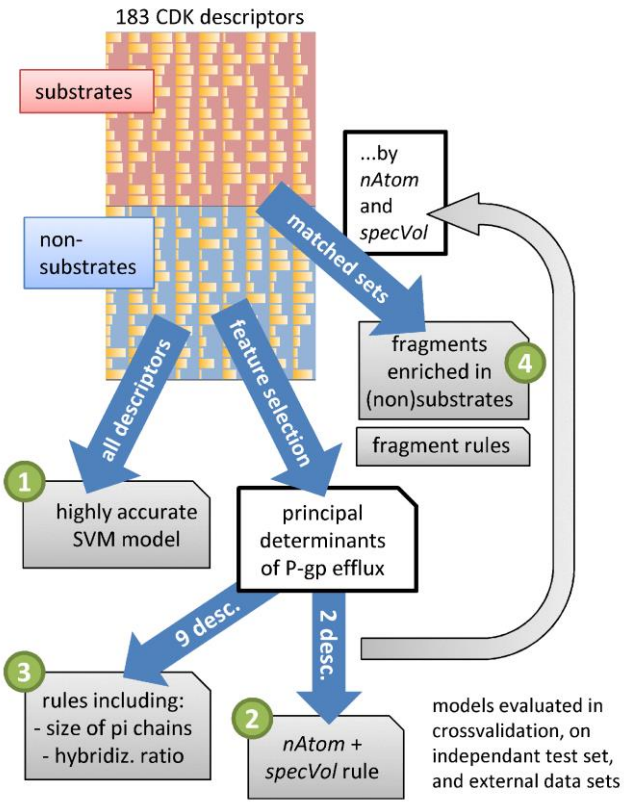


Figure 9. A schematic representation of the methods employed in our analysis, where the 934 compounds are characterized by CDK descriptors, used to derive a (1) highly accurate SVM model, the (2) simple *nAtom-specVol* rule, more complex and accurate (3) rule based models, and finally to find (4) molecular fragments enriched in substrates or non-substrates.

Experimental section

Cell Culture. The HEp-2 cell line, previously known as a larynx carcinoma cell line, but subsequently found to have been established from HeLa cells (www.atcc.org) was previously used to derive a vincristine-resistant counterpart called VK_2.[39,75] Both cell lines were maintained in RMPI (Sigma) supplemented with 10% decomplemented fetal bovine serum (iFBS, Sigma),

2mM L-glutamine (Lonza), 100 U/ml penicillin, 100 μg/ml streptomycin (Gibco), and were incubated at 37 °C with 5% CO_2 in a humidified incubator.

Drug Susceptibility Testing. HEp-2 cells were seeded at 2500 cells/well in 96-well plate while 3000 cells/well were used for the VK_2 line because of its longer doubling period. After 24 h, cells were treated with indicated compounds and concentrations where highest DMSO concentration at <1%. Following a 72 h incubation, an MTT assay (Sigma) was performed per standard protocol, see e.g. Supek *et al*.,[76] Ester *et al*.[77] The absorbance (A) of the microtiter plate was measured on a microplate reader at 570 nm where absorbance is directly proportional to the number of living, metabolically active cells. When indicated, 1 hour prior to drug addition, cells were incubated with the P-gp inhibitors verapamil at 10 μM[78] or 1 μM CP-100356,[79] which we determined to be the highest concentrations that do not affect cellular viability (data not shown). Each treatment was performed in technical quadruplicates, with at least 2 biological replicates.

Results were analyzed according to slightly modified protocol used at the NCI-DTP (http://dtp.nci.nih.gov/branches/btb/ivclsp.html) as previously described.[76,77] Briefly, a positive percentage of growth (PG) indicates cell growth following drug treatment, where percent growth is relative to a negative control (no drug). A PG of 100% shows that the treated culture grew as well as the negative control, and 0% means growth has stopped upon adding the drug. A negative percentage indicates cytotoxicity following drug treatment where -100% shows no cells survived the treatment at the specific drug concentration.

Compounds. All compounds used for experimental testing were purchased from Sigma-Aldrich and were all declared by the seller to have purity 98% or higher (paclitaxel, 5-FU, mercaptopurine, chlorambucil, estrone, tramadol, digitoxigenin and etoposide), except irinotecan (>97%), vincristine (>95%), colchicine (>95%) and camptothecin (>90%).

Western Blotting. P-gp expression was analyzed by western blotting using the anti-Mdr antibody recommended for detection of P-gp and Mdr-3 of human origin (G-1, sc-13131, Santa Cruz Biotechnology). HEp-2 and VK_2 were lysed using lysis buffer (50 mM Tris (Sigma) pH 7.6, 150 mM NaCl (Kemika, Zagreb), 2 mM EDTA (Sigma) and 1% NP40 (BioRad), supplemented with protease inhibitors (Complete Mini protease inhibitors, Roche) for 30 min on ice. Cell extract was subsequently harvested and pelleted for 15 min at 13k rpm. Protein concentration was determined using the Pierce BCA Protein Assay Kit (Thermo Scientific) as per manufactures' instructions and proteins were fractionated by SDS-PAGE on 8% gel,

transferred to PVDF membrane and immunoblotted using Mdr antibody (1:1000 dilution) followed by a goat anti-mouse conjugated HRP secondary antibody (1:5000 dilution) (Cell Signaling). SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific) was used for developing blots.

ASSOCIATED CONTENT

Supporting Information. Example pairs of near-identical compounds cleaned from the initial dataset; Atom contribution for calculation of molecular van der Waals volume; Molecular fragments significantly enriched in P-gp substrates or non-substrates; The chemical space coverage of Penzotti *et al*., Bikadi *et al*. and dataset reported in this study; A lack of correlation of the "specific volume" (*specVol*) to common measures of molecular size in our dataset; Agreement of predictions between the default RBF kernel SVM and the alternative polynomial kernel for the 120 test set compounds; Pairs of structurally very similar compounds which differ specifically in one of the effluxophores; Pairs of structurally very similar compounds which differ in the value of specific volume; Set of P-gp substrates and non-substrates used for training and validation of the SVM model. This material is available free of charge via the Internet at http://pubs.acs.org

AUTHOR INFORMATION

Corresponding Author

*Tel: +385 1 4561 080. E-mail: fran.supek@irb.hr

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding Sources

This work was funded by Biozyne Ltd.

ABBREVIATIONS

P-gp, P-glycoprotein, also known as ABCB1 or MDR1; ADME-Tox, Absorption, Distribution, Metabolism, Excretion, and Toxicity; QSAR, Quantitative Structure-Activity Relationship; SMILES, Simplified Molecular-Input Line-Entry System, k-NN, k Nearest Neighbors; SVM, Support Vector Machine; NCI-DTP, US National Cancer Institute Developmental Therapeutics Program; NCI-60, NCI-DTP Cancer Cell Line Panel; LogP, Logarithm of Partition Coefficient; PC1, First Principal Component; GI50, Concentration of the Compound causing 50% Growth Inhibition; RBF, Radial Basis Kernel; CDK, Chemistry Development Kit; AUC, Area Under the Receiver Operating Characteristic Curve; nAtom, Number of Atoms; VABC, Molecular van der Waals Volume; apol, Sum of the Atomic Polarizabilities; nAtomP, Number of Atoms in the Largest Pi Chain; specVol, Specific Volume, Volume per Atom; TP, True Positive; TN, True Negative; FP, False positive; FN, False Negative; Acc, Classification Accuracy; MCC, Matthews Correlation Coefficient; PCR, polymerase chain reaction; DMEM, Dulbecco's modified Eagle's medium; FBS, fetal bovine serum; MTT, 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide; PVDF, polyvinylidene difluoride; DMSO, dimethyl sulfoxide; HPR, horseradish peroxidase; SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis.

REFERENCES

(1) Fromm, M. F. Importance of P-glycoprotein at blood–tissue barriers. *Trends Pharmacol. Sci.* **2004**, *25*, 423–429.

(2) Zhou, S.-F. Structure, function and regulation of P-glycoprotein and its clinical relevance in drug disposition. *Xenobiotica* **2008**, *38*, 802–832.

(3) Shoemaker, R. H. Genetic and Epigenetic Factors in Anticancer Drug Resistance. *JNCI J. Natl. Cancer Inst.* **2000**, *92*, 4–5.

(4) Szakács, G.; Paterson, J. K.; Ludwig, J. A.; Booth-Genthe, C.; Gottesman, M. M. Targeting multidrug resistance in cancer. *Nat. Rev. Drug Discovery* **2006**, *5*, 219–234.

(5) Szakács, G.; Váradi, A.; Özvegy-Laczka, C.; Sarkadi, B. The role of ABC transporters in drug absorption, distribution, metabolism, excretion and toxicity (ADME–Tox). *Drug Discov. Today* **2008**, *13*, 379–393.

(6) 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.

(7) Seelig, A. A general pattern for substrate recognition by P-glycoprotein. *Eur. J. Biochem.* **1998**, *251*, 252–261.

(8) 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.

(9) Cabrera, M. A.; González, I.; Fernández, C.; Navarro, C.; Bermejo, M. A topological substructural approach for the prediction of P-glycoprotein substrates. *J. Pharm. Sci.* **2006**, *95*, 589–606.

(10) Wang, Z.; Chen, Y.; Liang, H.; Bender, A.; Glen, R. C.; Yan, A. P-glycoprotein substrate models using support vector machines based on a comprehensive data set. *J. Chem. Inf. Model.* **2011**, *51*, 1447–1456.

(11) Bikadi, Z.; Hazai, I.; Malik, D.; Jemnitz, K.; Veres, Z.; Hari, P.; Ni, Z.; Loo, T. W.; Clarke, D. M.; Hazai, E.; Mao, Q. Predicting P-Glycoprotein-Mediated Drug Transport Based On Support Vector Machine and Three-Dimensional Crystal Structure of P-glycoprotein. *PLoS One* **2011**, *6*.

(12) Didziapetris, R.; Japertas, P.; Avdeef, A.; Petrauskas, A. Classification Analysis of P-Glycoprotein Substrate Specificity. *J. Drug Target.* **2003**, *11*, 391–406.

(13) Chen, L.; Li, Y.; Yu, H.; Zhang, L.; Hou, T. Computational models for predicting substrates or inhibitors of P-glycoprotein. *Drug Discov. Today* **2012**, *17*, 343–351.

(14) 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.

(15) Mechetner, E.; Kyshtoobayeva, A.; Zonis, S.; Kim, H.; Stroup, R.; Garcia, R.; Parker, R. J.; Fruehauf, J. P. Levels of multidrug resistance (MDR1) P-glycoprotein expression by human breast cancer correlate with in vitro resistance to taxol and doxorubicin. *Clin. Cancer Res.* **1998**, *4*, 389–398.

(16) Gottesman, M. M.; Fojo, T.; Bates, S. E. Multidrug resistance in cancer: role of ATP-dependent transporters. *Nat. Rev. Cancer* **2002**, *2*, 48–58.

(17) Bioinformatics & Systems Biology A Division of the Plant Systems Biology department, http://bioinformatics.psb.ugent.be/webtools/Venn/ (accessed Jan, 2012).

(18) De Cerqueira Lima, P.; Golbraikh, A.; Oloff, S.; Xiao, Y.; Tropsha, A. Combinatorial QSAR Modeling of P-Glycoprotein Substrates. *J. Chem. Inf. Model.* **2006**, *46*, 1245–1254.

(19) Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Model.* **2005**, *45*, 786–799.

(20) Huang, J.; Ma, G.; Muhammad, I.; Cheng, Y. Identifying P-Glycoprotein Substrates Using a Support Vector Machine Optimized by a Particle Swarm. *J. Chem. Inf. Model.* **2007**, *47*, 1638–1647.

(21) Ivanciuc, O. Applications of Support Vector Machines in Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Cundari, T. R., Eds.; John Wiley & Sons, Inc., 2007; pp. 291–400.

(22) Grdičak, M.; Supek, F.; Kralj, M.; Majer, Z.; Hollósi, M.; Šmuc, T.; Mlinarić-Majerski, K.; Horvat, Š. Computational structure–activity study directs synthesis of novel antitumor enkephalin analogs. *Amino Acids* **2010**, *38*, 1185–1191.

(23) Supek, F.; Ramljak, T. Š.; Marjanović, M.; Buljubašić, M.; Kragol, G.; Ilić, N.; Šmuc, T.; Zahradka, D.; Mlinarić-Majerski, K.; Kralj, M. Could LogP be a principal determinant of biological activity in 18-crown-6 ethers? Synthesis of biologically active adamantane-substituted diaza-crowns. *Eur. J. Med. Chem.* **2011**, *46*, 3444–3454.

(24) Gleeson, M. P. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *J. Med. Chem.* **2008**, *51*, 817–834.

(25) Seelig, A.; Landwojtowicz, E. Structure–activity relationship of P-glycoprotein substrates and modifiers. *Eur. J. Pharm. Sci.* **2000**, *12*, 31–40.

(26) Yasuda, K.; Lan, L.; Sanglard, D.; Furuya, K.; Schuetz, J. D.; Schuetz, E. G. Interaction of Cytochrome P450 3A Inhibitors with P-Glycoprotein. *J. Pharmacol. Exp. Ther.* **2002**, *303*, 323–332.

(27) Crivori, P.; Reinach, B.; Pezzetta, D.; Poggesi, I. Computational models for identifying potential P-glycoprotein substrates and inhibitors. *Mol. Pharm.* **2006**, *3*, 33–44.

(28) Szakács, G.; Annereau, J.-P.; Lababidi, S.; Shankavaram, U.; Arciello, A.; Bussey, K. J.; Reinhold, W.; Guo, Y.; Kruh, G. D.; Reimers, M.; Weinstein, J. N.; Gottesman, M. M. Predicting drug sensitivity and resistance: Profiling ABC transporter genes in cancer cells. *Cancer Cell* **2004**, *6*, 129–137.

(29) Alvarez, M.; Paull, K.; Monks, A.; Hose, C.; Lee, J. S.; Weinstein, J.; Grever, M.; Bates, S.; Fojo, T. Generation of a drug resistance profile by quantitation of mdr-1/P-glycoprotein in the cell lines of the National Cancer Institute Anticancer Drug Screen. *J. Clin. Invest.* **1995**, *95*, 2205–2214.

(30) Wallqvist, A.; Rabow, A. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. Linking the growth inhibition response from the National Cancer Institute's anticancer screen to gene expression levels and other molecular target data. *Bioinformatics* **2003**, *19*, 2212–2224.

(31) Ke, W.; Yu, P.; Wang, J.; Wang, R.; Guo, C.; Zhou, L.; Li, C.; Li, K. MCF-7/ADR cells (re-designated NCI/ADR-RES) are not derived from MCF-7 breast cancer cells: a loss for breast cancer multidrug-resistant research. *Med. Oncol.* **2011**, *28 Suppl 1*, S135–141.

(32) Roschke, A. V.; Tonon, G.; Gehlhaus, K. S.; McTyre, N.; Bussey, K. J.; Lababidi, S.; Scudiero, D. A.; Weinstein, J. N.; Kirsch, I. R. Karyotypic Complexity of the NCI-60 Drug-Screening Panel. *Cancer Res.* **2003**, *63*, 8634–8647.

(33) Garraway, L. A.; Widlund, H. R.; Rubin, M. A.; Getz, G.; Berger, A. J.; Ramaswamy, S.; Beroukhim, R.; Milner, D. A.; Granter, S. R.; Du, J.; Lee, C.; Wagner, S. N.; Li, C.; Golub, T. R.; Rimm, D. L.; Meyerson, M. L.; Fisher, D. E.; Sellers, W. R. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **2005**, *436*, 117–122.

(34) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.

(35) 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.

(36) Zhao, Y. H.; Abraham, M. H.; Zissimos, A. M. Fast Calculation of van der Waals Volume as a Sum of Atomic and Bond Contributions and Its Application to Drug Compounds. *J. Org. Chem.* **2003**, *68*, 7368–7373.

(37) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *J. Pharm. Sci.* **2009**, *98*, 861–893.

(38) Szegezdi, J.; Csizmadia, F. Prediction of distribution coefficient using microconstants. In; ChemAxon Ltd.: Anaheim, California, 2004.

(39) Osmak, M.; Beketić-Orešković, L.; Matulić, M.; Sorić, J. Resistance of human larynx carcinoma cells to cisplatin, γ-irradiation and methotrexate does not involve overexpression of c-myc or c-Ki-ras oncogenes. *Mutation Research Letters* **1993**, *303*, 113–120.

(40) Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A. A.; Kim, S.; Wilson, C. J.; Lehár, J.; Kryukov, G. V.; Sonkin, D.; Reddy, A.; Liu, M.; Murray, L.; Berger, M. F.; Monahan, J. E.; Morais, P.; Meltzer, J.; Korejwa, A.; Jané-Valbuena, J.; Mapa, F. A.; Thibault, J.; Bric-Furlong, E.; Raman, P.; Shipway, A.; Engels, I. H.; Cheng, J.; Yu, G. K.; Yu, J.; Aspesi, P.; de Silva, M.; Jagtap, K.; Jones, M. D.; Wang, L.; Hatton, C.; Palescandolo, E.; Gupta, S.; Mahan, S.; Sougnez, C.; Onofrio, R. C.; Liefeld, T.; MacConaill, L.; Winckler, W.; Reich, M.; Li, N.; Mesirov, J. P.; Gabriel, S. B.; Getz, G.; Ardlie, K.; Chan, V.; Myer, V. E.; Weber, B. L.; Porter, J.; Warmuth, M.; Finan, P.; Harris, J. L.; Meyerson, M.; Golub, T. R.; Morrissey, M. P.; Sellers, W. R.; Schlegel, R.; Garraway, L. A. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **2012**, *483*, 603–307.

(41) Garnett, M. J.; Edelman, E. J.; Heidorn, S. J.; Greenman, C. D.; Dastur, A.; Lau, K. W.; Greninger, P.; Thompson, I. R.; Luo, X.; Soares, J.; Liu, Q.; Iorio, F.; Surdez, D.; Chen, L.; Milano, R. J.; Bignell, G. R.; Tam, A. T.; Davies, H.; Stevenson, J. A.; Barthorpe, S.; Lutz, S. R.; Kogera, F.; Lawrence, K.; McLaren-Douglas, A.; Mitropoulos, X.; Mironenko, T.; Thi, H.; Richardson, L.; Zhou, W.; Jewitt, F.; Zhang, T.; O'Brien, P.; Boisvert, J. L.; Price, S.; Hur, W.; Yang, W.; Deng, X.; Butler, A.; Choi, H. G.; Chang, J. W.; Baselga, J.; Stamenkovic, I.; Engelman, J. A.; Sharma, S. V.; Delattre, O.; Saez-Rodriguez, J.; Gray, N. S.; Settleman, J.; Futreal, P. A.; Haber, D. A.; Stratton, M. R.; Ramaswamy, S.; McDermott, U.; Benes, C. H. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **2012**, *483*, 570–575.

(42) Gupta, P. B.; Onder, T. T.; Jiang, G.; Tao, K.; Kuperwasser, C.; Weinberg, R. A.; Lander, E. S. Identification of selective inhibitors of cancer stem cells by high-throughput screening. *Cell* **2009**, *138*, 645–659.

(43) Miller, S. J.; Lavker, R. M.; Sun, T.-T. Interpreting epithelial cancer biology in the context of stem cells: tumor properties and therapeutic implications. *Biochim. Biophys. Acta* **2005**, *1756*, 25–52.

(44) Vogel, C.; Abreu, R. de S.; Ko, D.; Le, S.-Y.; Shapiro, B. A.; Burns, S. C.; Sandhu, D.; Boutz, D. R.; Marcotte, E. M.; Penalva, L. O. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* **2010**, *6*.

(45) Supek, F.; Šmuc, T. On Relevance of Codon Usage to Expression of Synthetic and Natural Genes in Escherichia coli. *Genetics* **2010**, *185*, 1129–1134.

(46) Greenman, C. D.; Bignell, G.; Butler, A.; Edkins, S.; Hinton, J.; Beare, D.; Swamy, S.; Santarius, T.; Chen, L.; Widaa, S.; Futreal, P. A.; Stratton, M. R. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics* **2010**, *11*, 164–175.

(47) Cancer Genome Project Gene Copy Number Analysis, http://www.sanger.ac.uk/cgi-bin/genetics/CGP/cghviewer/CghHome.cgi (accessed Jul, 2012).

(48) Cancer Genome Project Gene Copy Number Analysis, NCI/ADR-RES Chromosome 7, http://www.sanger.ac.uk/cgi-bin/genetics/CGP/cghviewer/CghViewer.cgi?action=DisplayChromosome&tissue=breast&d=2&chr=7&id=6582 (accessed Jul, 2012).

(49) Cancer Genome Project Gene Copy Number Analysis, OVCAR-8 Chromosome 7, http://www.sanger.ac.uk/cgi-bin/genetics/CGP/cghviewer/CghViewer.cgi?action=DisplayChromosome&d=2&chr=7&id=6707 (accessed Jul, 2012).

(50) Desai, P. V.; Sawada, G. A.; Watson, I. A.; Raub, T. J. Integration of in Silico and in Vitro Tools for Scaffold Optimization during Drug Discovery: Predicting P-Glycoprotein Efflux. *Mol. Pharmaceutics* **2013**, *10*, 1249–1261.

(51) Österberg, T.; Norinder, U. Theoretical calculation and prediction of P-glycoprotein-interacting drugs using MolSurf parametrization and PLS statistics. *Eur. J. Pharm. Sci.* **2000**, *10*, 295–303.

(52) Zamora, J. M.; Pearce, H. L.; Beck, W. T. Physical-chemical properties shared by compounds that modulate multidrug resistance in human leukemic cells. *Mol. Pharmacol.* **1988**, *33*, 454–462.

(53) Pawagi, A. B.; Wang, J.; Silverman, M.; Reithmeier, R. A. F.; Deber, C. M. Transmembrane Aromatic Amino Acid Distribution in P-glycoprotein: A Functional Role in Broad Substrate Specificity. *J. Mol. Biol.* **1994**, *235*, 554–564.

(54) Pajeva, I. K.; Wiese, M. Pharmacophore Model of Drugs Involved in P-Glycoprotein Multidrug Resistance: Explanation of Structural Variety (Hypothesis). *J. Med. Chem.* **2002**, *45*, 5671–5686.

(55) Seelig, A.; Landwojtowicz, E.; Fischer, H.; Li Blatter, X. Towards P-Glycoprotein Structure–Activity Relationships. In *Drug Bioavailability*; Waterbeemd, H. van de; Lennernäs, H.; Artursson, P., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA, 2004; pp. 461–492.

(56) Dolghih, E.; Bryant, C.; Renslo, A. R.; Jacobson, M. P. Predicting Binding to P-Glycoprotein by Flexible Receptor Docking. *PLoS Comput. Biol.* **2011**, *7*, e1002083.

(57) Li, W.-X.; Li, L.; Eksterowicz, J.; Ling, X. B.; Cardozo, M. Significance Analysis and Multiple Pharmacophore Models for Differentiating P-Glycoprotein Substrates. *J. Chem. Inf. Model.* **2007**, *47*, 2429–2438.

(58) Poongavanam, V.; Haider, N.; Ecker, G. F. Fingerprint-based in silico models for the prediction of P-glycoprotein substrates and inhibitors. *Bioorganic & Medicinal Chemistry* **2012**, *20*, 5388–5395.

(59) Kelly, R. J.; Draper, D.; Chen, C. C.; Robey, R. W.; Figg, W. D.; Piekarz, R. L.; Chen, X.; Gardner, E. R.; Balis, F. M.; Venkatesan, A. M.; Steinberg, S. M.; Fojo, T.; Bates, S. E. A Pharmacodynamic Study of Docetaxel in Combination with the P-glycoprotein Antagonist Tariquidar (XR9576) in Patients with Lung, Ovarian, and Cervical Cancer. *Clin. Cancer Res.* **2011**, *17*, 569–580.

(60) Maki, N.; Hafkemeyer, P.; Dey, S. Allosteric modulation of human P-glycoprotein. Inhibition of transport by preventing substrate translocation and dissociation. *J. Biol. Chem.* **2003**, *278*, 18132–18139.

(61) Rautio, J.; Humphreys, J. E.; Webster, L. O.; Balakrishnan, A.; Keogh, J. P.; Kunta, J. R.; Serabjit-Singh, C. J.; Polli, J. W. In Vitro P-Glycoprotein Inhibition Assays for Assessment of Clinical Drug Interaction Potential of New Drug Candidates: A Recommendation for Probe Substrates. *Drug. Metab. Dispos.* **2006**, *34*, 786–792.

(62) 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.

(63) Broccatelli, F.; Carosati, E.; Neri, A.; Frosini, M.; Goracci, L.; Oprea, T. I.; Cruciani, G. A novel approach for predicting P-glycoprotein (ABCB1) inhibition using molecular interaction fields. *J. Med. Chem.* **2011**, *54*, 1740–1751.

(64) Lu, X.; Gong, S.; Monks, A.; Zaharevitz, D.; Moscow, J. A. Correlation of nucleoside and nucleobase transporter gene expression with antimetabolite drug cytotoxicity. *J. Exp. Ther. Oncol.* **2002**, *2*, 200–212.

(65) Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **2006**, *6*, 813–823.

(66) The NCI Database Download Page, http://cactus.nci.nih.gov/download/nci/ (accessed Dec, 2012).

(67) O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 1–14.

(68) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27:1–27:27.

(69) EL-Manzalawy, Y.; Honavar, V. Integrating LibSVM into Weka Environment.

(70) Hsu, C.-W.; Chang, C.-C.; Chih-Jen, L. A Practical Guide to Support Vector Classification **2010**.

(71) Fawcett, T. An introduction to ROC analysis. *Pattern Recog. Lett.* **2006**, *27*, 861–874.

(72) Klekota, J.; Roth, F. P. Chemical Substructures That Enrich for Biological Activity. *Bioinformatics* **2008**, *24*, 2518–2525.

(73) Cohen, W. W. Fast Effective Rule Induction. In *Proceedings of the Twelfth International Conference on Machine Learning*; Morgan Kaufmann, 1995; pp. 115–123.

(74) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **2009**, *11*, 10–18.

(75) Osmak, M. Collateral resistance or sensitivity of human larynx carcinoma HEp2 cells resistant to cis-dichlorodiammineplatinum (II) or vincristine sulfate. *Neoplasma* **1992**, *39*, 197–202.

(76) Supek, F.; Kralj, M.; Marjanović, M.; Šuman, L.; Šmuc, T.; Krizmanić, I.; Žinić, B. Atypical cytostatic mechanism of N-1-sulfonylcytosine derivatives determined by in vitro screening and computational analysis. *Invest. New Drugs* **2008**, *26*, 97–110.

(77) Ester, K.; Supek, F.; Majsec, K.; Marjanović, M.; Lembo, D.; Donalisio, M.; Šmuc, T.; Jarak, I.; Karminski-Zamola, G.; Kralj, M. Putative mechanisms of antitumor activity of cyano-substituted heteroaryles in HeLa cells. *Invest. New Drugs* **2012**, *30*, 450–467.

(78) Lavie, Y.; Cao, H. t; Volner, A.; Lucci, A.; Han, T. Y.; Geffen, V.; Giuliano, A. E.; Cabot, M. C. Agents that reverse multidrug resistance, tamoxifen, verapamil, and cyclosporin A, block glycosphingolipid metabolism by inhibiting ceramide glycosylation in human cancer cells. *J. Biol. Chem.* **1997**, *272*, 1682–1687.

(79) Zhao, P.; Kunze, K. L.; Lee, C. A. Evaluation of time-dependent inactivation of CYP3A in cryopreserved human hepatocytes. *Drug Metab. Dispos.* **2005**, *33*, 853–861.

Supporting Information

Accurate models for P-gp drug recognition induced from a cancer cell line cytotoxicity screen

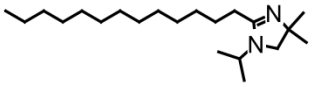
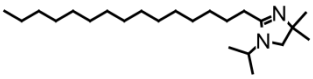
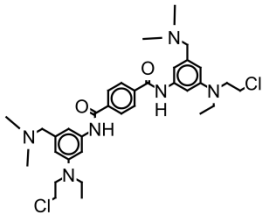
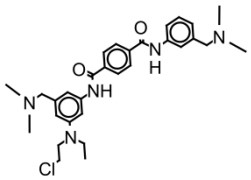
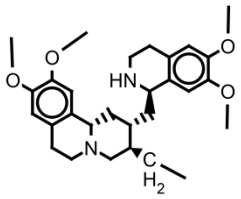
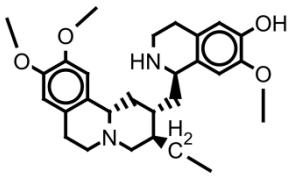
*Jurica Levatić[a], Jasna Ćurak[a], Marijeta Kralj[a,b], Tomislav Šmuc[a,c], Maja Osmak[d], Fran Supek[a,c],**

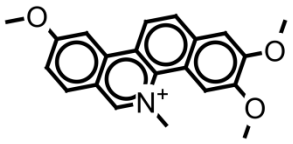
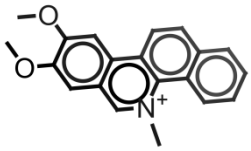
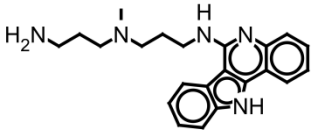
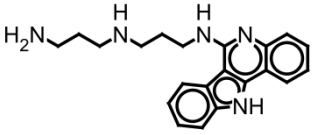
[a] BioZyne Ltd, Bijenička 54, 10000 Zagreb, Croatia; [b] Division of Molecular Medicine, Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia; [c] Division of Electronics, Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia; [d] Division of Molecular Biology, Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia.

*corresponding author: fran.supek@irb.hr

Table of contents:

Supporting Table S1. Example pairs of near-identical compounds cleaned from the initial dataset.

NSC Number	Compound retained[a]	NSC Number	Compound removed[a]
17743		17744	
670013		675252	
44185		32944	

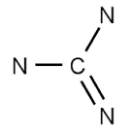
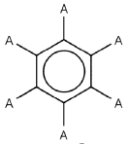
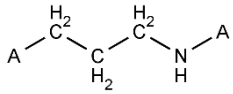
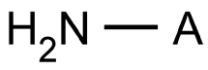
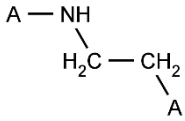
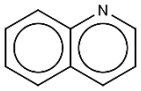
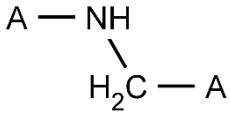
645305		645301	
736612		736611	

[a] Prior to training of classification models, the dataset was cleaned of compounds that share a Tanimoto coefficient of 1 to another compound in the training data, meaning they are structurally identical or near-identical compounds. Shown are five such example pairs, from each pair the compound with worse confidence based on "difference" and "correlation" criteria was removed from the dataset. In total, 24 compounds were removed in this step.

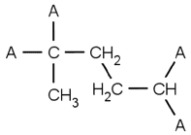
Supporting Table S2. Atom contributions for calculation of molecular van der Waals volume (Zhao *et al*., 2003, J. Org. Chem.).

atom	**volume(Å^3)**	**atom**	**volume(Å^3)**
H	7.24	P	24.43
C	20.58	S	24.43
N	15.60	As	26.52
O	14.71	B	40.48
F	13.31	Si	38.79
Cl	22.45	Se	28.73
Br	26.52	Te	36.62
I	32.52		

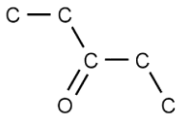
Supporting Table S3. Molecular fragments significantly enriched in P-gp substrates or non-substrates.

Substructures enriched in substrates[a]		
 **p-value:** 10^{-6} **enrichment:** ∞ **95% CI** = N/A **SMARTS:** NC(=N)N	 **p-value:** $2*10^{-3}$ **enrichment:** 5.67 **95% CI** = [1.69, 19.02] **SMARTS:** [!#1]c1c([!#1])c([!#1])c([!#1])c([!#1])c1[!#1]	 **p-value:** $8*10^{-4}$ **enrichment:** 3.71 **95% CI** = [1.65, 8.35] **SMARTS:** [!#1][CH2][CH2][CH2][NH][!#1]
 **p-value:** $1*10^{-5}$ **enrichment:** 3.21 **95% CI** = [1.83, 5.65] **SMARTS:** [!#1][NH2]	 **p-value:** $3*10^{-5}$ **enrichment:** 3.14 **95% CI** = [1.78, 5.54] **SMARTS:** [!#1][CH2][CH2][NH][!#1]	 **p-value:** $1*10^{-3}$ **enrichment:** 2.91 **95% CI** = [1.51, 5.60] **SMARTS:** c1ccc2ncccc2c1
 **p-value:** $2*10^{-5}$ **enrichment:** 2.88 **95% CI** = [1.73, 4.82] **SMARTS:** [!#1][CH2][NH][!#1]		

Substructures enriched in non-substrates[a]

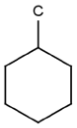


p-value: $8*10^{-4}$
enrichment: 14.0
95% CI = [1.86, 105.4]
SMARTS: [!#1][CH]([!#1])[CH2][CH2]C([!#1])([!#1])[CH3]

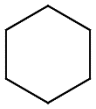


p-value: $2*10^{-3}$
enrichment: 7.50
95% CI = [1.74, 32.34]
SMARTS: CCC(=O)CC

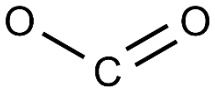
p-value: $3*10^{-5}$
enrichment: 6.50
95% CI = [2.31, 18.26]
SMARTS: C1CCCC1



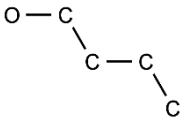
p-value: $9*10^{-4}$
enrichment: 3.22
95% CI = [1.57, 6.62]
SMARTS: CC1CCCCC1



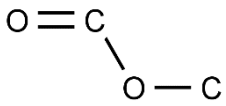
p-value: $4*10^{-4}$
enrichment: 3.09
95% CI = [1.62, 5.92]
SMARTS: C1CCCCC1



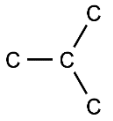
p-value: 10^{-4}
enrichment: 2.29
95% CI = [1.48, 3.54]
SMARTS: OC=O



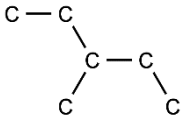
p-value: $4*10^{-5}$
enrichment: 2.21
95% CI = [1.50, 3.26]
SMARTS: CCCCO



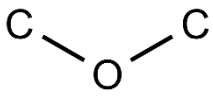
p-value: 10^{-3}
enrichment: 2.19
95% CI = [1.36, 3.52]
SMARTS: COC=O



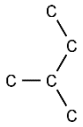
p-value: 10^{-5}
enrichment: 2.18
95% CI = [1.52, 3.12]
SMARTS: CC(C)C



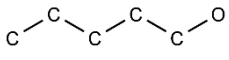
p-value: $6*10^{-4}$
enrichment: 2.17
95% CI = [1.39, 3.41]
SMARTS: CCC(C)CC



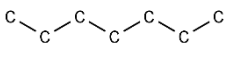
p-value: $4*10^{-5}$
enrichment: 2.13
95% CI = [1.47, 3.07]
SMARTS: COC



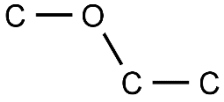
p-value: $5*10^{-4}$
enrichment: 2.04
95% CI = [1.36, 3.05]
SMARTS: CCC(C)C



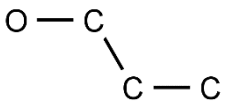
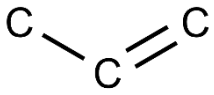
p-value: $5*10^{-4}$
enrichment: 2.04



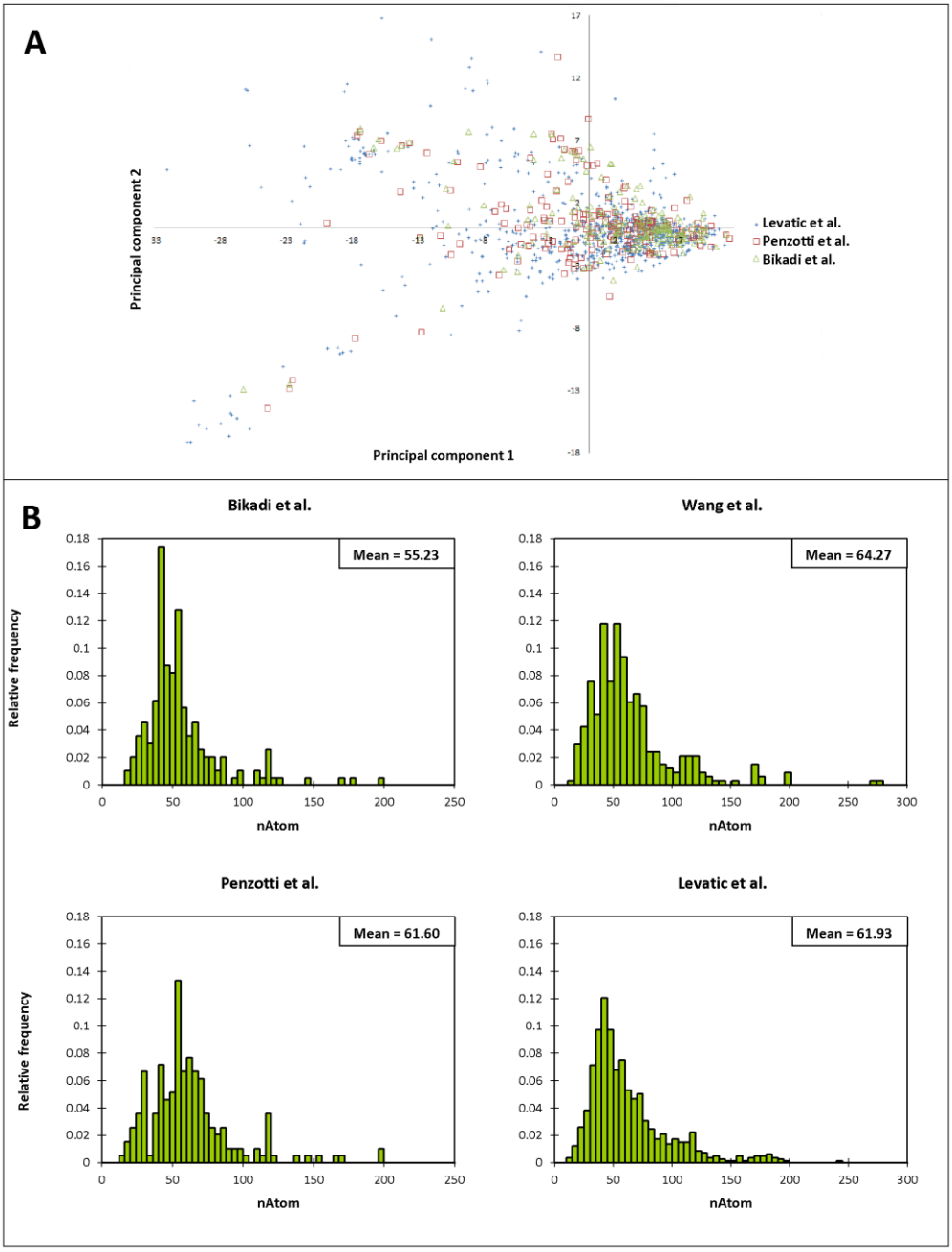
p-value: $3*10^{-4}$
enrichment: 2.03



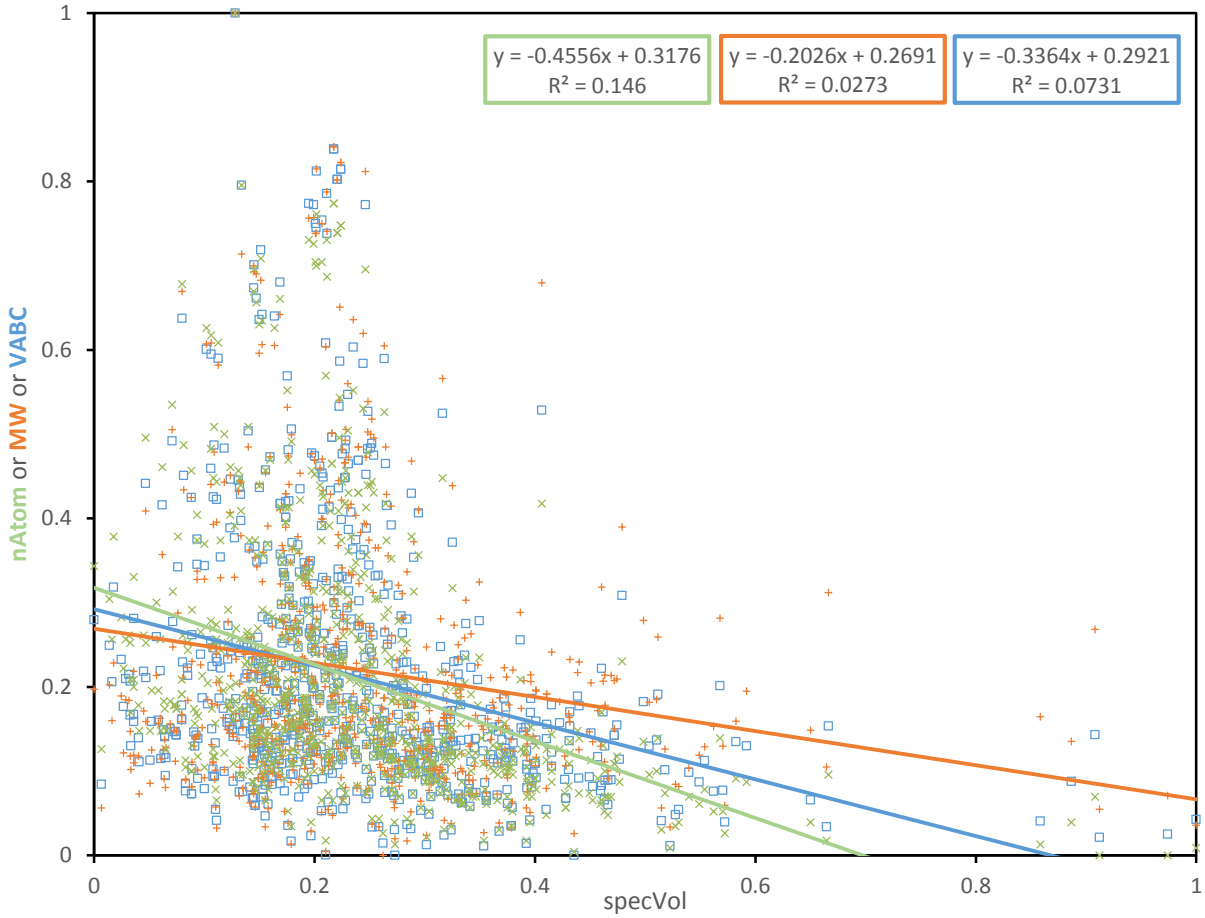
p-value: $2*10^{-4}$
enrichment: 2.03

95% CI = [1.36, 3.05] **SMARTS:** CCCCCO	**95% CI** = [1.38, 3.00] **SMARTS:** CCCCCCC	**95% CI** = [1.40, 2.95] **SMARTS:** CCOC
		
p-value: $4*10^{-5}$ **enrichment:** 2.03 **95% CI** = [1.44, 2.86] **SMARTS:** CCCO	**p-value:** $2*10^{-4}$ **enrichment:** 2.00 **95% CI** = [1.38, 2.90] **SMARTS:** CC=C	

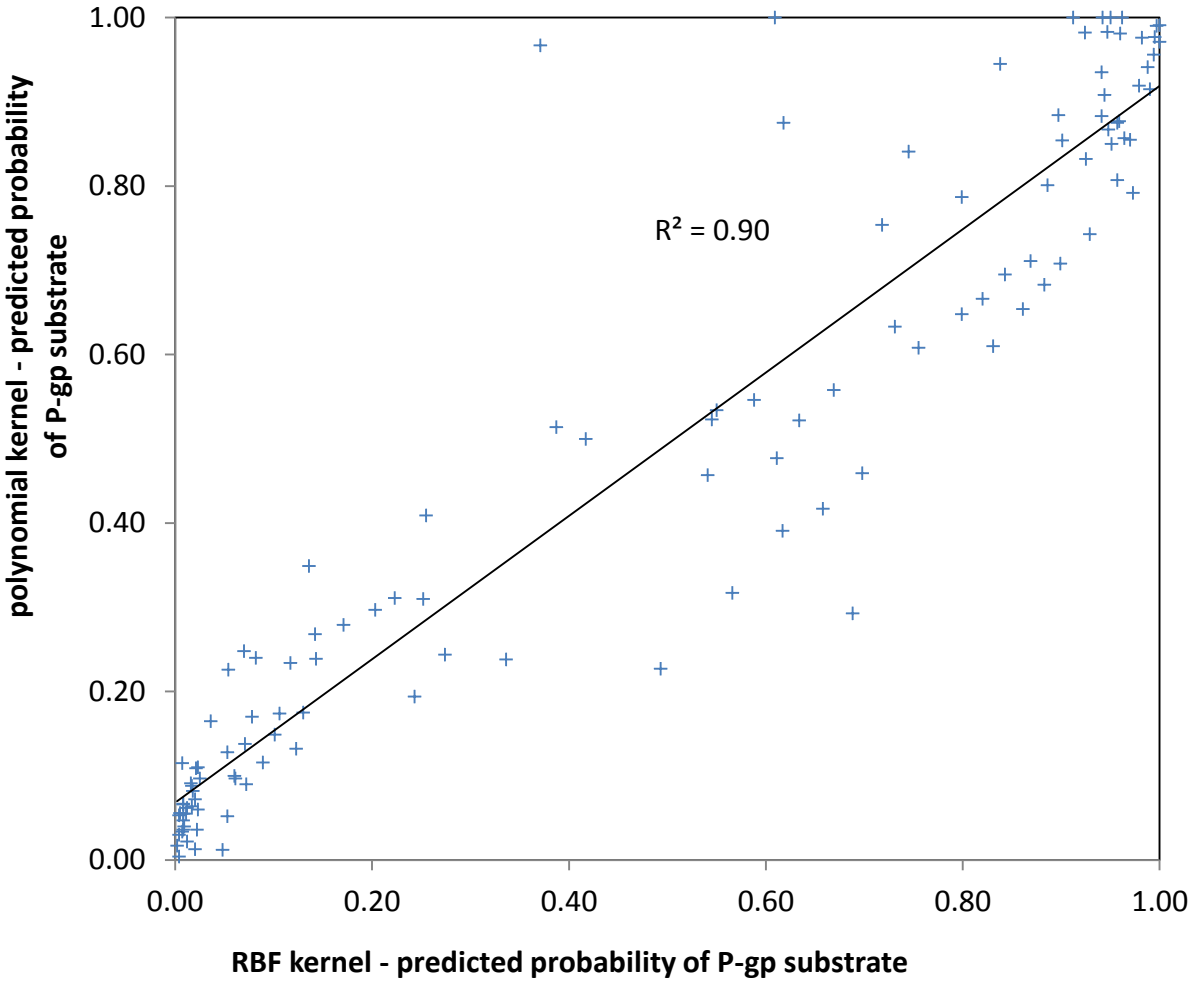
[a] At least 2x enrichment in either substrates or non-substrates, and additionally P<0.002 by Fisher's exact test, is required. The enrichment was tested in sets matched by *nAtom* and *specVol*, thus these fragments contribute to the substrate-non-substrate distinction beyond their influence on *nAtom* and *specVol*. The letter 'A' in molecular fragment structures means any non-hydrogen atom.



Supporting Figure S1. (A) The chemical space coverage of Penzotti *et al*., Bikadi *et al*. and our dataset, presented as a plot of the first two principal components of 183 CDK molecular descriptors used in this study. A single outlying point at (-18.69, 37.19) from Penzotti *et al*. dataset was omitted from the plot. (B) Comparison of distributions of number of atoms (*nAtom*) in our compound set to three recent studies (Bikadi, Wang and Penzotti), reveals that the distributions are very similar. Thus, our compound set is as representative of the drug-like molecules in the 30-80 atom range as the other previous P-gp compound sets.

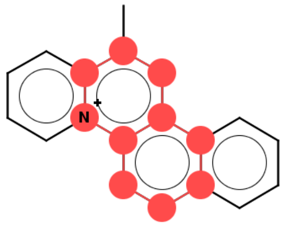
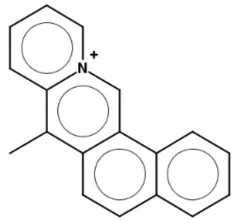
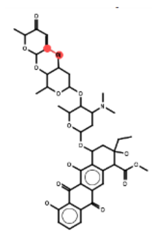
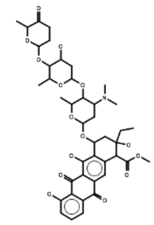
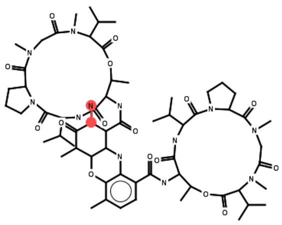
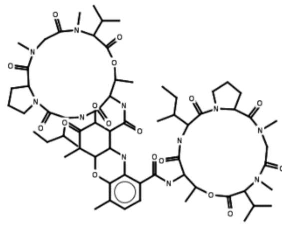
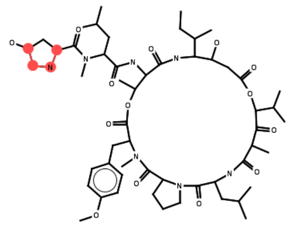
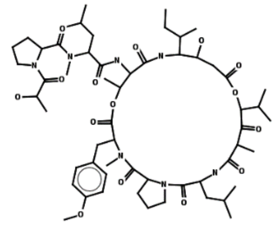
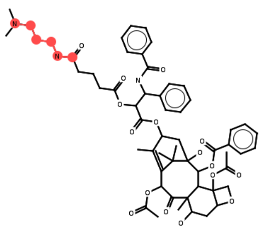
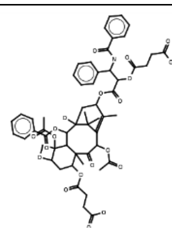
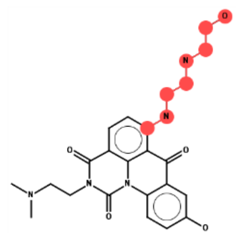
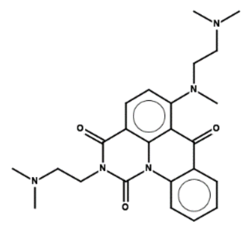
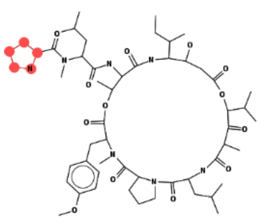
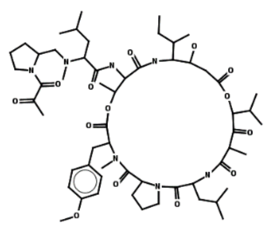
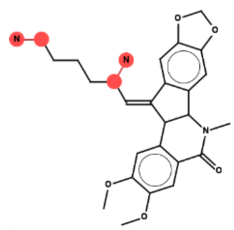
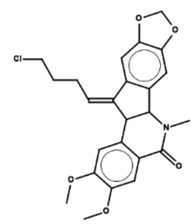
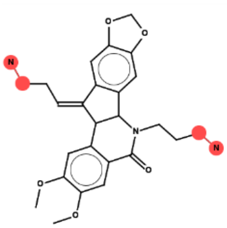
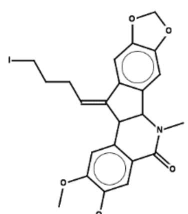
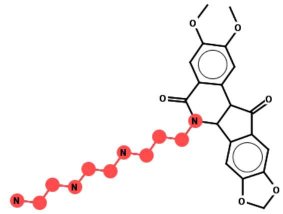
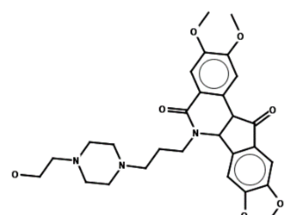


Supporting Figure S2. A lack of correlation of the "specific volume" (*specVol*) to common measures of size molecular size in our dataset. "*nAtom*" is the number of atoms in the compound, H included. "*MW*" is the molecular weight. "VABC" is the molecular volume in cubic Angström (Zhao *et al*., 2003, J. Org. Chem.). *MW*, *nAtom*, *VABC* and *specVol* are normalized to range from 0 to 1 so that they can be shown on the same plot.

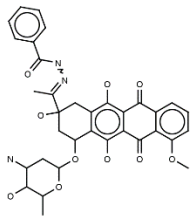
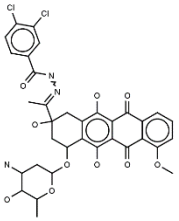
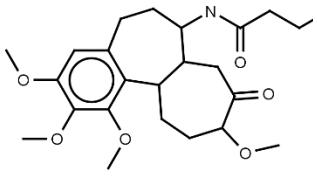
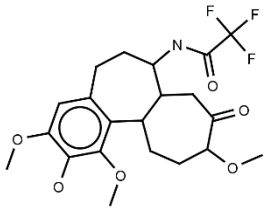
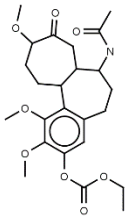
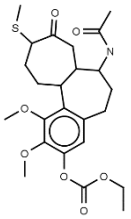
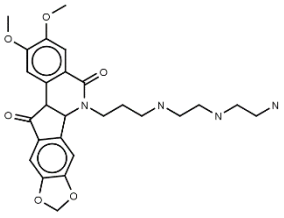
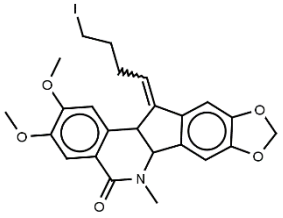
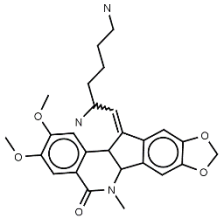
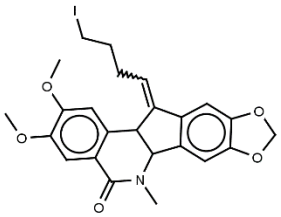
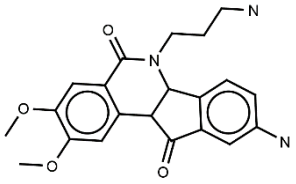
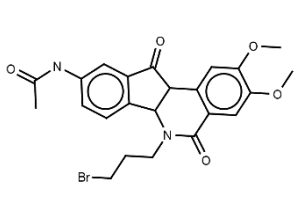


Supporting Figure S3. Agreement of predictions between the default RBF kernel SVM ($c=2^3$, $gamma=2^{-2}$), and the alternative polynomial kernel (degree=2, $c=2^{-1}$), for the 120 test set compounds.

Supporting Table S4. Pairs of structurally very similar (Tanimoto coefficient ≥ 0.85) compounds which differ specifically in one of the effluxophores (red) – molecular fragments significantly enriched in the P-gp substrate group (Table S3).

Effluxophores present	Effluxophores absent	Effluxophores present	Effluxophores absent
			
			
			
			
			

Supporting Table S5. Pairs of structurally very similar (Tanimoto coefficient ≥ 0.85) compounds which differ in the value of specific volume (*specVol*) – molecular descriptor important for P-gp propensity.

specVol < 7.3	*specVol* > 7.3	*specVol* < 7.3	*specVol* > 7.3
			
			
			

Supporting Table S6. Set of P-gp substrates (class 1) and non-substrates (class -1) used for training (t) and validation of (e) the SVM model.