DOI: 10.1002/minf.201400121

A Machine Learning Approach to Explain Drug Selectivity to Soluble and Membrane Protein Targets

Eva Freyhult,[a] Mats G. Gustafsson,[b] and Helena Strömbergsson*[b]

Abstract: Improved understanding of the forces that determine drug specificity to their targets is important for drug design and discovery, as well as for gaining knowledge about molecular recognition. Here, we present a machine learning approach that includes all approved drugs with a known protein target. The drugs were characterized using easily interpretable physico-chemical descriptors. Employing the Random Forest method, we were able to predict whether a drug binds to a soluble or membrane protein with an average accuracy of 84 % and an average area under curve of 0.91. The high average performance suggests that there exist some general physico-chemical differences between drugs that bind to membrane and soluble protein targets. Variable importance measures in combination with permutation tests were used to find the most influential descriptors. This resulted in six outstanding descriptors, that all involve drug flexibility and lipophilicity, suggesting that drugs binding to membrane protein targets are in general more flexible and lipophilic, and conversely, drugs binding to soluble protein targets are more rigid and hydrophilic. With the notion that ligands in general are blueprints of their protein pockets, we may also draw general conclusions about the protein-pocket properties which may add to the understanding of molecular recognition.

Keywords: Drugs · Machine learning · Drug selectivity · Targetome · Protein-ligand recognition

1 Introduction

Drugs are typically small organic compounds that interact with their molecular targets. Drug targets have been categorized into larger groups such as enzymes, receptors and ion channels.[1] The large majority of targets consists of proteins while only a small percentage of the drug targets are non-proteins, such as nucleic acids and lipids.[1] The interaction between a drug and its protein target depends on properties of both parties according to the idea of induced fit,[2] where both the drug and the protein target adjust their shapes in order to fit to one another.

The three dimensional (3D) structure of a protein is largely dependent on the environment the protein resides in as well as the composition of amino acids of the polypeptide chain. Proteins are broadly classified into two groups. The first group consists of (water) soluble proteins that can either be globular or fibrous, such as enzymes and collagens.[3] The second group consists of membrane proteins, such as G-protein coupled receptors and porins.[3] Moreover, proteins have been classified by primary amino-acid sequence,[4] and 3D structure,[5,6] using machine-learning techniques. Due to laboratory-technical difficulties, there are still relatively few 3D structures determined for proteins that reside in the membrane. Binding pockets have therefore mostly been studied on soluble proteins,[7] and they vary in terms of size, shape, and hydrophobicity,[8] The pocketome of the drug targets has been studied lately by Kufareva et al.[7] The pocketome is computed from all known 3D protein structures and classifies binding pockets by their ligand-binding features.

The hypothetical perfect drug should bind tightly and selectively to its target where it ultimately leads to a desired change of the biochemical processes in the body. A large number of other requirements need to be fulfilled as well such as acceptable ADME (absorption, distribution, metabolism, excretion) properties, and limited adverse side effects including toxicology, and chemical stability.[9] However, unspecific binding by drugs to other biomolecules is very common,[10] which often causes unwanted and/or unexpected side effects. The physico-chemical features of all known drugs, i.e. the so-called drug space has been exten-

[a] *E. Freyhult*
Division of Cancer Pharmacology and Computational Medicine,
Department of Medical Sciences, Bioinformatics Infrastructure for
Life Sciences, Science for Life Laboratory, Uppsala University,
Uppsala Academic Hospital
SE-751 85 Uppsala, Sweden

[b] *M. G. Gustafsson, H. Strömbergsson*
Division of Cancer Pharmacology and Computational Medicine,
Department of Medical Sciences, Uppsala University, Uppsala
Academic Hospital
SE-751 85 Uppsala, Sweden
**e-mail: helena.strombergsson@medsci.uu.se*

Supporting Information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201400121.

sively studied[11] as well as the enlarged space of natural products.[12] Rosén et al.[12] used a subset of physicochemical descriptors to visualize the space of natural products which evidently is much larger that the space of all approved drugs.

Due to the shortage of 3D structure data from membrane-spanning proteins,[13] computational methods used for function prediction have traditionally been based on 3D data for soluble proteins,[14] and amino-acid-sequence data for membrane proteins.[15] This makes cross-comparison of the result very difficult between the two major protein categories, although there are studies that include the drug-target space of both membrane and soluble proteins.[16,17] In general, it is necessary to use very general methods to describe features of proteins since the complex information obtained from 3D structures is not available for the large majority of membrane proteins. For example, Yamanishi et al.[17] use chemical substructures to define the chemical space and amino acid alignment scores to define the target space, while Strömbergsson and Kleywegt[16] used general features of the amino acid chain such as percentage hydrophobic amino acids. Such methods to describe proteins are useful for certain applications but provides very little information on the actual binding pocket of the protein.

In accordance with the induced-fit theory, the drug (or more generally the ligand), is a blueprint of its protein-binding pocket. Ligand size obviously matters with large ligands requiring large pockets and vice versa. Hydrophobic parts of the binding pocket attract hydrophobic parts of the ligand. Aromatic interactions, hydrogen bonds, ionic bonds, and metal complex ions can only be possible if there are proper amino acids in the binding pocket to interact with. With the assumption that the shape and physico-chemical nature of binding pockets depend on protein fold, which in turn, depends on the amino acid sequence and the environment of the protein, it is reasonable to assume that there are some general differences between ligands that bind to soluble proteins and ligands that bind to membrane proteins. If such a bias exists, it would reflect general differences between binding pockets of soluble and membrane proteins.

In this study, we have investigated whether there exists a selectivity between drugs binding to soluble protein targets and drugs binding to membrane protein targets. To this end, all drugs, approved by the Food and Drug Administration, stored in the database DrugBank[18] were retrieved along with their drug targets. Approved drugs generally have an extensive target annotation, which is preferable in this case, although there are a number of other protein-ligand databases such as BindingMoad[19] and CREDO.[20] To overcome the shortage of 3D structures of membrane-binding proteins, and the crudeness of protein descriptors applicable on proteins without available 3D information, the focus is instead on the chemical nature of all approved drugs. A number of physico-chemical descriptors were generated from the 3D structure of each drug, and models were induced by machine-learning methods. The resulting models were able to distinguish between drugs that bind to soluble proteins and drugs that bind to membrane proteins with an average accuracy of 84% and an average AUC (area under curve) of 0.91. The descriptors used in this study are easily interpreted, which provide hints of general properties important for drug-protein recognition in soluble and membrane proteins.

2 Materials and Methods

2.1 Drug Descriptor Computation

Two-dimensional structures of all approved drugs were obtained from DrugBank[18] as sdf files on 15 June 2012. The sdf file format is suited for small molecules and do not allow for large approved drugs such as antibodies or peptides. In all, there were 1424 structure files. To make descriptor computation possible, each structure file has to contain one single compound. Therefore, each file was inspected manually to find entries with multiple structures. When such a file was found, identical compounds were removed so that each structure file contained only one single compound. Similarly, very small structures such as salts or buffers that are common experimental residuals were removed. If it was not possible to identify one of the compounds as the single active drug, the entire entry was removed from the data set. According to Fourches et al.,[21] entries that contain more than one compound should be removed. We choose to keep entries with one large compound and one small organic and inorganic compound when the small compound was a commonly used salt or buffer, since those were expected to be experimental residuals. Hydrogens were added to the compounds with OpenBabel[22] and 3D structures were generated with Corina.[23] The removal of entries not suited for descriptor computation resulted in 1370 remaining compounds. Descriptors were computed with Dragon 5.5[24] that generates in all 3224 chemical descriptors.[25]

2.2 Class Assignment of Drug Targets

Information on all drugs were retrieved from DrugBank[18] on 15 June 2012 as drug cards, which is a format specified by the database. Each drug card contains extensive information on drug properties including a mapping to drug targets. The number of all approved drugs was 1578. To obtain a data set that contains small molecule drugs that bind only to proteins, a number of exclusion criteria were set up. Drugs that were proteins or peptides were excluded, as well as nutraceuticals. Moreover, small molecule drugs that bind to non-proteins such as nucleic acids or lipids were excluded. The exclusion criteria resulted in a set of 1251 drugs. These drugs were to be categorized as binding to soluble protein targets, membrane protein targets or both types of targets. In order to achieve this goal, the target mapping for each drug was retrieved from the DrugBank drugcard. This mapping contains information on whether the drug targets has any transmembrane regions. A protein was assigned to the membrane target class if any transmembrane regions were annotated to the protein amino-acid chain. This criterion does not take into account that membrane proteins display a large variation in terms of membrane integration. For instance, G-protein-coupled receptors reside almost entirely within the plasma membrane with their seven alpha helices, while for instance tyrosine kinase receptors have a soluble part that is larger than their transmembrane domain. With the shortage of solved 3D structures of membrane proteins, it is however currently impossible to know whether a given drug binds to a transmembrane part or to a soluble part of a membrane protein. Therefore, all proteins that have a transmembrane region are assigned to the same drug-target class. According to this rational, if a drug binds to proteins that all have transmembrane regions, the drug was assigned to the membrane-protein-target class. Similarly, if a drug binds to proteins that are all soluble, the drug was assigned to the soluble protein-target class. If a drug binds to both protein types, the drug was assigned to a third ambiguous class of drugs.

2.3 Data Sets and Feature Selection

The chemical descriptors computed by Dragon and the target-preference outcomes were merged into a data set that contains in all 1212 entries. Each entry consists of a drug described by 3224 chemical descriptors (of which 2243 have a non-zero variance and a non-missing value for at least 20% of the entries) and an outcome that assigns each drug to the membrane protein target class, the soluble protein target class, or the ambiguous class. The distribution across these three classes is as follows: 241 ambiguous, 620 membrane, and 351 soluble. A subset of 34 chemical descriptors was selected for their interpretability as well as for properties important for drug design such as size, flexibility and lipophilicity (Table 1). The assembly process of the data set is illustrated in Figure 1. The set of 34 selected descriptors was further reduced by identifying clusters (hierachical clustering, complete linkage method, distance measure 1-|Spearman correlation|) of highly correlated descriptors (Spearman's $\rho \geq 0.8$) and for each such cluster keeping only one descriptor. Figure 1 shows the Spearman correlations (the ambiguous proteins were excluded when correlations were calculated) between the 34 descriptors and a dendrogram illustrates the hierachical clustering. After this procedure 17 descriptors remained. Most of these 17 descriptors have only weak correlations, with one exception, namely the rather strong correlation between nBnz and nCar (these two descriptors are both kept in the reduced data as they were not clustered into the same cluster by the complete linkage algorithm).

Table 1. A set of 34 easily interpretable physicochemical descriptors manually selected for model induction.

Descriptor	Meaning
MW	Molecular weight
Sv	Sum of van der Waals volume
Se	Sum of Sanderson electronegativities
Sp	Sum of atomic polarizability
Mv	Mean atomic van der Waals volume
Me	Mean atomic Sanderson electronegativity
nAT	Number of atoms
nSK	Number of non hydrogen atoms
nBT	Number of bonds
nBO	Number of non hydrogen bonds
nBM	Number of multiple bonds
ARR	Aromatic ratio
nCIC	Number of rings
RBN	Number of rotatable bonds
RBF	Rotatable bond fraction
nDB	Number of double bonds
nAB	Number of aromatic bonds
nC	Number of carbon atoms
nN	Number of nitrogen atoms
nO	Number of oxygen atoms
nX	Number of halogen atoms
nBnz	Number of benzene-like rings
nCar	Number of aromatic carbons (Sp2)
nRCONH2	Number of primary amides
nROH	Number of hydroxyl groups
nArOH	Number of aromatic hydroxyl
nHDon	Number of donor atom for hydrogen bonds (N and O)
nHAcc	Number of acceptor atom for hydrogen bonds (N, O, and F)
Ui	Unsaturation index
Hy	Hydrophilic factor
AMR	Ghose-Crippen molar refractivity
TPSA(NO)	Topological polar surface area (N and O contributions)
TPSA(Tot)	Topological polar surface area (N, O, S, and P contributions)
ALOGP	Ghose-Crippen octanol-water partition coefficient

2.4 Classification Models

RandomForest[26] classification was used to predict class identity, i.e. membrane, soluble or ambiguous. The Random Forest implementation in the R-package randomForest was used with default settings (ntree = 500 and mtry = sqrt(p), where p is the number of descriptors). Classification models were built both to discriminate between all three classes and to discriminate only between the two classes membrane and soluble (excluding all ambiguous protein binders). Models based on all 3224 descriptors, only the selected 34 descriptors, as well as only the 17 uncorrelated descriptors, were built and evaluated.

To assess the performance of the models, the data were divided in training and test-data sets. A Random Forest model was designed using the training data and the performance of the resulting model was assessed on the test data set. The modeling was performed in 10 different 5-fold cross validations (CVs) and the performance measures

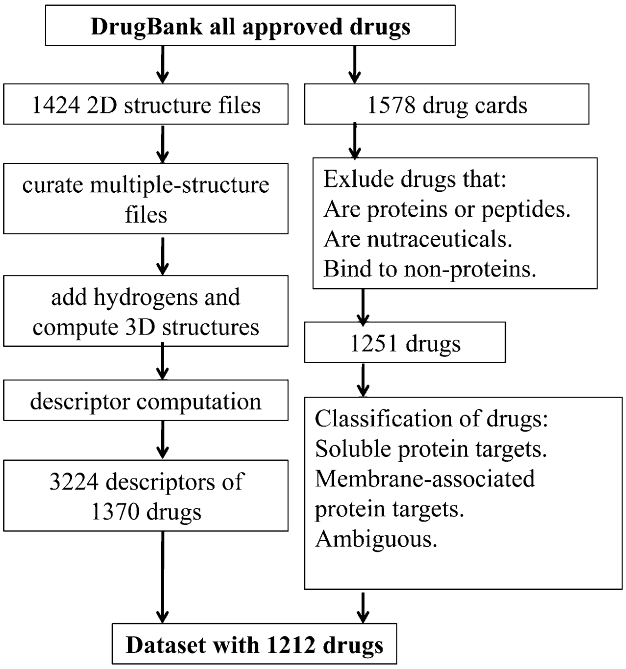

Figure 1. The assembly process of the data set used in this study.

average area under curve (AUC) and average accuracy was computed over the resulting 50 test sets.

In addition, null distributions for average AUC, average accuracy as well as average variable importance measures were determined by means of $N=1000$ permutations. In each permutation the class identities were permuted and then the full cross validation procedure was performed. The null distributions were used to compute permutation *p*-values. Here $p=(n+1)/(N+1)$, where n is the number of permutations generating a performance estimate (AUC or accuracy) at least as high as when the data is unpermuted. Note that the smallest possible *p*-value is $1/(N+1)$.

2.5 Missing Values

Some of the drugs have missing values for some of the descriptors. Descriptors that are missing for more than 80% of all the drugs are excluded in the modeling procedure. This resulted in the removal of 19 descriptors. The remaining missing values are imputed based on training data (i.e. inside the cross validation) using a k nearest neighbor imputation, where the missing value is replaced by the mean of values of the 5% closest neighbors. The distance is computed by Euclidean distance after centering the descriptors and scaling to unit variance.

2.6 Variable Importance

The standard Random Forest algorithm employed computes a variable importance for each variable expressed as the mean decrease gini index, which is a measure of node purity. The algorithm can also compute another variable importance, the so-called mean decrease accuracy. However, this measure requires much more computation time and in addition it has been shown to be less robust than the gini importance.[27] It has been pointed out that random forests importance measures can be biased when the variables have different scales or number of categories.[28] To handle this problem permutation p-values based on the gini index were computed for each variable by comparing its average variable importance for a variable with the corresponding null distribution (as computed by permutations).

3 Results and Discussion

The assembled drug-target data contained originally 1212 drugs that either bind to membrane protein targets, soluble protein targets, or both soluble and membrane targets (the ambiguous class). All drugs were characterized by 3224 descriptors of various types. A large part of the full set of 3224 descriptors are difficult to interpret. Therefore, Random Forest prediction models were designed using both this complete descriptor set as well as using a manually selected set of 34 easily interpretable physico-chemical descriptors (Table 1). It is important to note that a manual selection in general causes pre-determined results in the sense that only certain properties are included in the model. There is a possibility that physicochemical features important for target selectivity has been omitted and thus not considered by the model. However, since the small 34 descriptor model perform equally well as the large 3224 descriptor model, it is likely that at least the most important features important for drug selectivity to membrane bound and soluble targets are included in the final model. The small set of 34 descriptors was further reduced to only 17 descriptors using Spearman correlations and hierarchical clustering as illustrated in Figure 2 and described in materials and methods.

In addition to the manually selected descriptor set, subsets of 17 randomly selected uncorrelated descriptors were evaluated as well.

Random forest prediction models were designed using all three descriptor sets, for both two-class and three-class problems. The two-class models were designed to discriminate only between drugs that bind to membrane and soluble targets. Similarly, the three-class prediction models were designed to discriminate between drugs that bind to membrane-bound, soluble and ambiguous targets. The main purpose of the three-class models was to assess if also the ambiguous class had unique properties and hence was possible to predict. The resulting performance measures for these models are shown in Table 2. These results show that models based on two classes, as well as three classes, were able to distinguish between drugs binding to membrane and soluble targets. The accuracies (probability

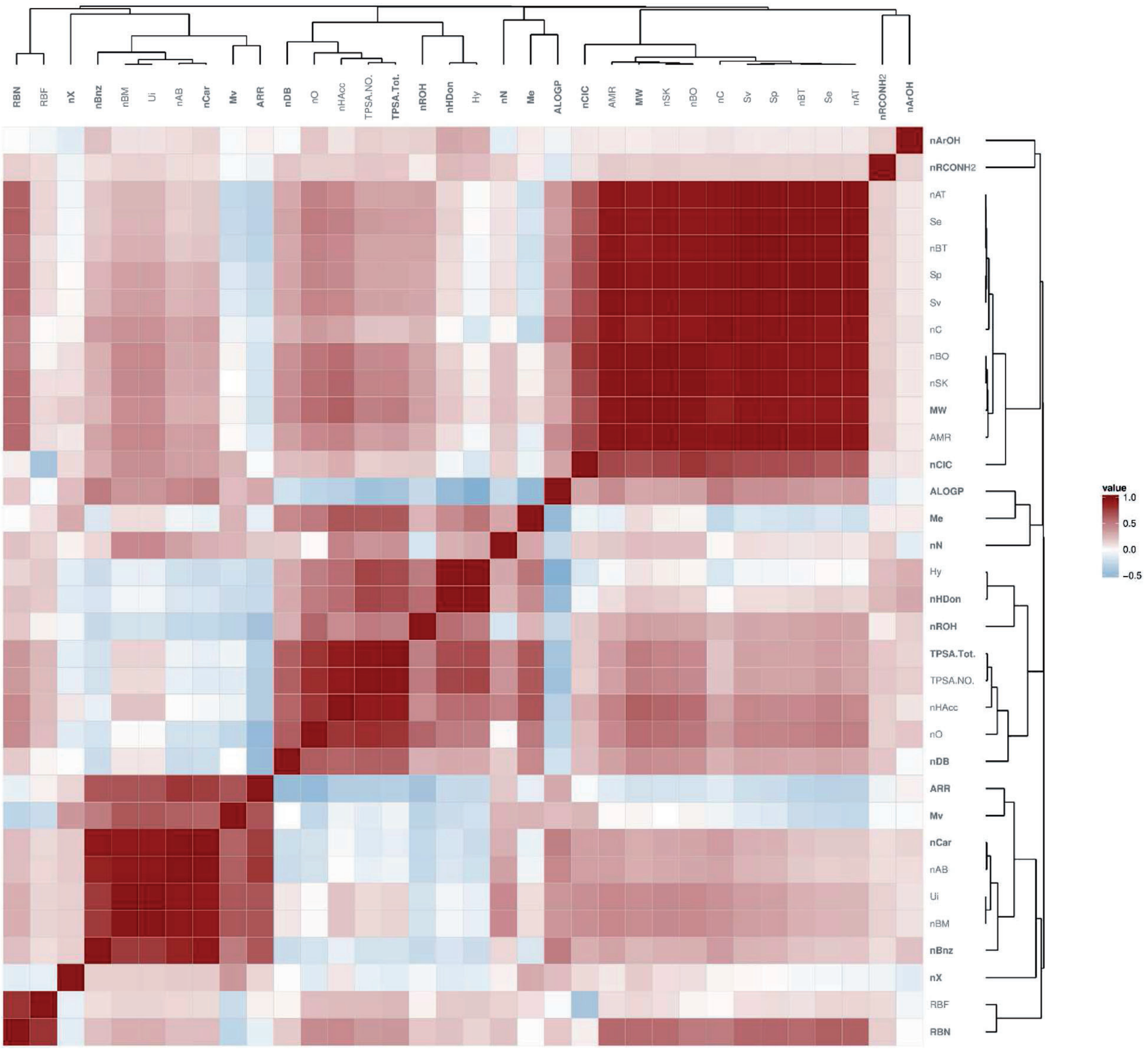

Figure 2. Spearman correlations between the 34 selected descriptors. The descriptors are clustered using complete linkage and 1-|Spearman correlation| as distance measure. In clusters with a Spearman correlation greater than or equal to 0.8 between all descriptors one representative descriptor is selected (marked in bold) to be included in the reduced data set of only 17 descriptors.

Table 2. Performances of models based on the full data set with 3224 descriptors, the reduced data set with 34 descriptors, and the further reduced data set with 17 uncorrelated descriptors as well as 1000 random sets of 17 uncorrelated descriptors. The three types of models are performed both to predict the three classes ambiguous (A), soluble (S) and membrane (M) as well as only the two classes soluble (S) and membrane (M). The performance measures reported in the table are all mean values (computed over the 50 test sets) of accuracy, AUC (area under ROC-curve) and class-specific accuracies for each of the predicted classes. For the random sets the median of the mean performance values are reported, and in parenthesis an interval in which 95% of the mean performance values lie.

	Three-class models				Two-class models			
Number of descriptors	3224	34	17	17 random	3224	34	17	17 random
Accuracy	0.69	0.68	0.68	0.64 (0.60,0.67)	0.84	0.84	0.84	0.79 (0.75,0.82)
AUC					0.92	0.90	0.91	0.86 (0.80,0.89)
Accuracy^M	0.88	0.85	0.86	0.86 (0.83,0.89)	0.92	0.91	0.92	0.90 (0.88,0.92)
Accuracy^S	0.66	0.65	0.67	0.56 (0.46,0.63)	0.70	0.71	0.72	0.59 (0.51,0.67)
Accuracy^A	0.26	0.26	0.25	0.19 (0.11,0.26)				

of making a correct classification of a random example) of the two-class models were in general higher which is expected simply because it is more difficult to find good decision boundaries for a three class problem than a two class problem. One should also note that the examples of the ambiguous class have features that make them bind to both target types making this class more challenging to define than the other two. The class-specific accuracies (probability of making a correct classification of a random example from the specific class) for the soluble and membrane-bound classes were in general high and higher for the two-class models than for the three-class models for all descriptor set sizes. However, the class-specific accuracy for the ambiguous class was in general low.

The two-class models were all able to predict if a drug binds to a soluble or membrane-bound protein with high accuracy. The models based on only 17 easily interpretable physico-chemical descriptors can distinguish between membrane targets and soluble protein targets with an accuracy of 84% and an average AUC of 0.910 (Table 2). It is notable that the results from large 3224-descriptor set performs just as well as the small 17-descriptor set. It is however important to note that the performance evaluation of this model is limited to drugs that a priori are known to not belong to the ambiguous class.

The 2 class model based on the 17 selected descriptors (trained on all non-ambiguous examples) was used to predict the ambiguous drugs as either membrane or soluble protein binding. 69% of the ambiguous are classified as membrane and 22% as soluble protein binding. To check the validity of these classifications a one-sided Mann-Whitney test was adopted. The average fraction of soluble protein targets was indeed higher among soluble than mebrane binders ($p=0.0001$).

The random descriptor sets show a reasonable performance (see Table 2), but not on the same level as the models based on the full data set or the manually selected descriptors.

The performance of the models based on the 17 selected descriptors was further evaluated in a 1000-fold permutation test with a random assignment of class identity in the same proportions as in the real data set. This resulted in an average accuracy of 59% (95% of all performance estimates were found in the interval [56%, 61%]) and an average AUC of 50% (with 95% of all the estimates found in the interval [47%, 54%]) for the permuted models. This permutation result strongly indicates that the original model performed better than what can be expected by random. The high class-specific accuracies, (Table 2) further support the predictive ability of the two-class models.

The performance of the two-class models were in general high and the models based on the small selected 17-descriptor set were just as good as the models based on the larger descriptor sets. Therefore, the small 17-descriptor two-class model was used to investigate which of the physicochemical properties of the drugs in the data set that determine selectivity to drug-target type. This was done by studying the so-called variable importance (VIMP) measure for each descriptor provided by the Random Forest learning algorithm which is the mean decrease gini index as described in the materials and methods section. The average VIMP for each descriptor was computed over the cross validation folds and this average was compared with the corresponding average VIMP measures for the permuted models. Figure 3 shows the average VIMP values and their corresponding permutation p-values for each of the 17 descriptors. The most important descriptors are nN, nROH, nDB, nCIC, nCar and nBnz, in the order of decreasing average VIMP. In this set, the *nCar* and *nCIC* variables have slightly smaller p-values but can still be considered important for drug target selectivity. The *nBnz* descriptors is the number of benzene-like rings in the drug molecule. Benzene rings are chemically very stable structures and help to stabilize the drug molecule as well as making it more lipophilic. The *nDB* descriptor is the number of double bonds. Double bonds occur most commonly between two carbons and reduce the flexibility of the drug. The *nROH* descriptor is the number of hydroxyl groups. Hydroxyl groups often act as a hydrogen donors which adds to the polarity of the compound. The *nROH* descriptor is highly correlated to the *nHDon* descriptor (number of hydrogen donors) according to the Spearman correlation presented in Figure 4. Due to the polarity of water, polar compounds are in general able to dissolve in water, which means that hydroxyl groups as such make the compound more water soluble. The *nN* descriptor is the number of nitrogen atoms. Nitrogen occurs in almost all drugs and commonly adds to the polarity of the compound depending on the fraction of nitrogen to the number of carbons. Small amines and amides are soluble in water since they are able to form hydrogen bonds with the water molecules while larger amines and amides are much more lipophilic in their nature. The *nCar* descriptor is the number of aromatic carbons which most commonly take part in benzene-like rings. Not surprisingly, the *nCar* descriptor is correlated with the *nBnz* descriptor. The *nCIC* descriptor is the number of rings and is correlated with the number of non-hydrogen bonds and the number of carbons. Both the number of non-hydrogen bonds and the number of carbons reflect the solubility of the molecule. A high number of carbons generally makes a molecule more lipophilic. To summarize, these six descriptors reflect the solubility and flexibility of the drugs that are part of this data set and it seems that, in general, these properties are the most important features that determine selectivity of drugs to membrane and soluble protein targets.

To gain knowledge on the direction of these descriptors the distributions of descriptor values in the membrane and soluble classes were visualized in frequency histograms, see Figure 4. The histograms and a corresponding Mann-Whitney U test (data not shown) show that the number of benzene rings and aromatic carbons are on average much higher in the membrane class than in the soluble class.

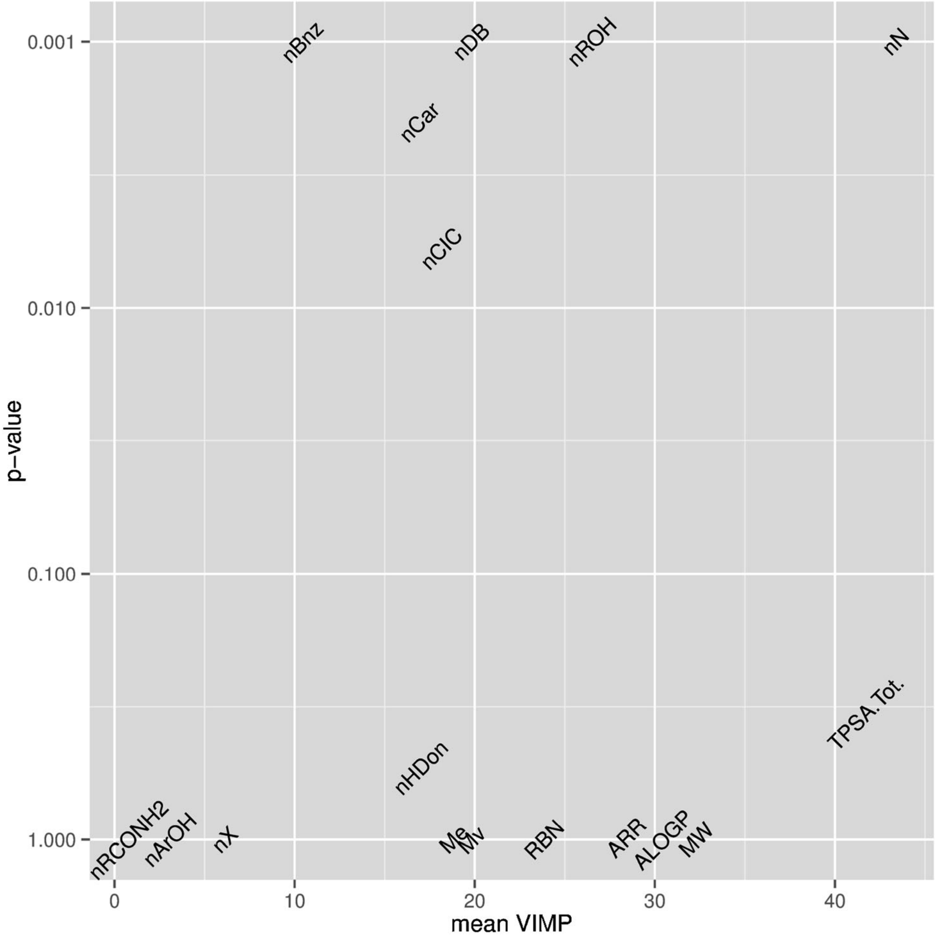

Figure 3. Permutation p-values for the 17 descriptors in the two-class model (excluding the ambiguous class) plotted versus the average VIMP. The most important descriptors are nN, nROH, nDB, nClC, nCar and nBnz.

Moreover, the average number of double bonds and hydroxyl groups are higher in the drugs binding to soluble proteins. The average number of aromatic rings are approximately equal in the two groups and the average number of nitrogens is only slightly higher in the membrane class. The fact that the number of nitrogens was shown to be very important in the multivariate models, although nN alone is not a very good discriminator, suggests that the number of nitrogens in combination with other descriptors is important.

4 Conclusion

In summary, our results indicate that drugs binding to protein targets that reside within or are anchored to the plasma membrane are more flexible and lipophilic, and conversely drugs that prefer soluble protein targets are more rigid and polar. According to the induced-fit theory and the general notion that ligands are blueprints of their binding pockets, we can also tentatively conclude that membrane protein targets have more flexible and lipophilic pockets while soluble protein targets have more rigid and hydrophilic pockets. It is well known that membrane proteins float in the plasma membrane. The lipids that the plasma membrane is composed of require that membrane proteins are composed mainly of lipophilic amino acids.[29] From this model, we may hypothesize that these lipophilic amino acids in general form a binding pocket that attract flexible and non-polar ligands. Similarly, soluble proteins that reside freely in a water-like environment, need to be less flexible in order to maintain their 3D shape and therefore attract more soluble and rigid drugs.

To our best knowledge, this is the first time that such tentative conclusions on the general nature of drugs and their binding pockets have been drawn from a data set encompassing all approved drugs with a known protein-drug target using machine learning and statistical methods. It is important to be aware of that the data set used in this study still only covers a small fraction of all possible ligands occurring in nature. The conclusions drawn from this data set therefore only applies to drug-like molecules. However, with the ever increasing production of genomic, proteomic, and biochemical data, we may in the near future, be able

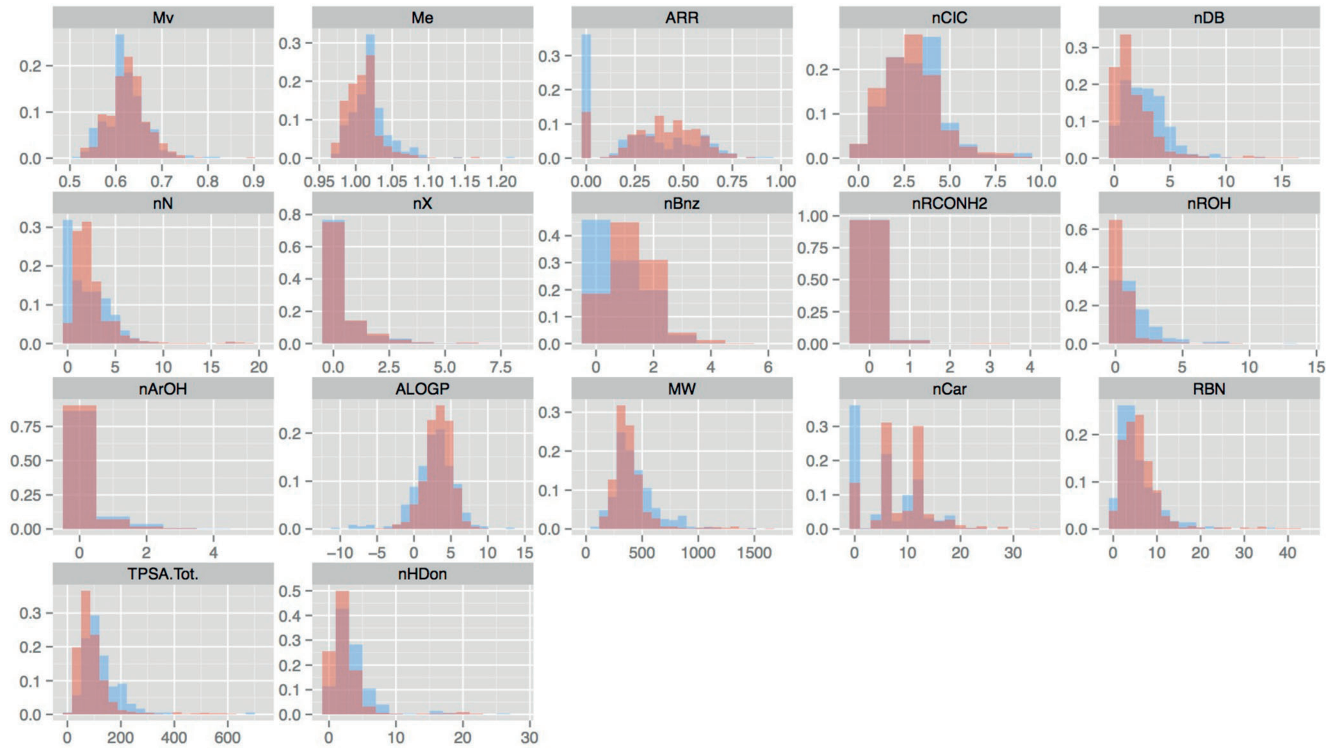

Figure 4. Distributions for the 17 selected descriptors. The soluble protein target class is shown in blue and the membrane protein target class in red. The membrane class is larger than the soluble class, and therefore the histograms are scaled so that the sum of the bins within each class is 1.

to further elucidate the physicochemical driving forced that determine drug-target selectivity.

Contributions

HS conceived of the approach. HS assembled the data set. EF performed the computational modeling. HS an EF drafted the manuscript. All authors contributed to the analysis of the results and the final manuscript writing.

Acknowledgements

Funding to finalize this work was provided by the *Hennerlöfska foundation*. The computations were performed on resources provided by *SNIC* through *Uppsala Multidisciplinary Center for Advanced Computational Science* (*UPPMAX*) under Project p2013054.

References

[1] P. Imming, C. Sinning, A. Meyer, *Nat. Rev. Drug Discov.* **2006**, *5*, 821–834.
[2] D. E. Koshland, *Proc. Natl. Acad. Sci. USA* **1958**, *44*, 98–104.
[3] D. Whitford, *Proteins Structure and Function*, Wiley, Chichester, **2005**.
[4] E. L. Sonnhammer, S. R. Eddy, R. Durbin, *Proteins* **1997**, *28*, 405–20.
[5] A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, *J. Mol. Biol.* **1995**, *247*, 536–40.
[6] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J. M. Thornton, *Structure* **1997**, *5*, 1093–108.
[7] I. Kufareva, A. V. Ilatovskiy, R. Abagyan, *Nucleic Acids Res.* **2012**, *40*, D535–D540.
[8] A. Kahraman, R. J. Morris, R. A. Laskowski, J. M. Thornton, *J. Mol. Biol.* **2007**, *368*, 283–301.
[9] L. Kroogsgaard-Larsen, P. Bunch, in *Drug Des. Discov.* (Eds: P. Kroogsgaard-Larsen, K. Stromgard, U. Madsen), CRC Press, Boca Raton, **2010**, pp. 1–14.
[10] M. Campillos, M. Kuhn, A. Gavin, L. J. Jensen, P. Bork, *Science* **2008**, *321*, 263–266.
[11] C. Lipinski, *J. Pharmacol. Toxicol. Meth.* **2000**, *44*, 235–249.
[12] J. Rosén, J. Gottfries, S. Muresan, A. Backlund, T. I. Oprea, *J. Med. Chem.* **2009**, *52*, 1953–1962.
[13] V. J. Haupt, S. Daminelli, M. Schroder, *PLoS One* **2013**, *8*, e65894.
[14] D. Lee, O. Redfern, C. Orengo, *Nat. Rev. Mol. Cell Biol.* 2007, 8, 995–1005.
[15] M. Punta, L. R. Forrest, H. Bigelow, A. Kernytsky, J. Liu, B. Rost, *Methods* **2007**, *41*, 460–474.
[16] H. Strömbergsson, G. J. Kleywegt, *BMC Bioinformatics* **2009**, *10*, S13.
[17] Y. Yamanishi, M. Kotera, M. Kanehisa, S. Goto, *Bioinformatics* **2010**, *26*, i246–i254.
[18] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, et al., *Nucleic Acids Res.* **2011**, *39*, D1035–1041.

[19] M. L. Benson, R. D. Smith, N. A. Khazanov, B. Dimcheff, J. Beaver, P. Dresslar, J. Nerothin, H. A. Carlson, *Nucleic Acids Res.* **2008**, D674–678.
[20] A. Schreyer, T. Blundell, *Chem. Biol. Drug. Des.* **2009**, *73*, 157–167.
[21] D. Fourches, E. Muratov, A. Tropsha, *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204.
[22] R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, E. L. Willighagen, *J. Chem. Inf. Model.* **2006**, *46*, 991–998.
[23] *Corina*, Molecular Networks, n.d.
[24] *Dragon*, Version 5.5 2007, Talete srl. n.d.
[25] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, **2000**.
[26] L. Breiman, *Mach. Learn.* **2001**, *45*, 5–32.
[27] M. L. Calle, V. Urrea, *Brief. Bioinform.* **2011**, *12*, 86–89.
[28] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, *BMC Bioinformatics* **2008**, *9*, 307.
[29] J. Nilsson, B. Persson, G. v. Heine, *Proteins* **2005**, *60*, 606–616.

Received: September 5, 2014
Accepted: October 29, 2014
Published online: January 8, 2015