Bioinformatics, 32(8), 2016, 1163–1169
doi: 10.1093/bioinformatics/btv725
Advance Access Publication Date: 10 December 2015
Original Paper

OXFORD

Structural bioinformatics

Prediction of the permeability of neutral drugs inferred from their solvation properties

Edoardo Milanetti[1], Domenico Raimondo[1] and Anna Tramontano[1,2,3,*]

[1]Department of Physics, Sapienza Università di Roma, Rome 00185, Italy, [2]Institute Pasteur – Fondazione Cenci Bolognetti, Rome, 00161, Italy and [3]Center for Life Nano Science @Sapienza, Istituto Italiano di Tecnologia, Sapienza Università di Roma, Rome, 00185, Italy

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on 10 September 2015; revised on 12 November 2015; accepted on 4 December 2015

Abstract

Motivation: Determination of drug absorption is an important component of the drug discovery and development process in that it plays a key role in the decision to promote drug candidates to clinical trials. We have developed a method that, on the basis of an analysis of the dynamic distribution of water molecules around a compound obtained by molecular dynamics simulations, can compute a parameter-free value that correlates very well with the compound permeability measured using the human colon adenocarcinoma (Caco-2) cell line assay.

Results: The method has been tested on twenty-three neutral drugs for which a consistent set of experimental data is available. We show here that our method reproduces the experimental data better than other existing tools. Furthermore it provides a detailed view of the relationship between the hydration and the permeability properties of molecules.

Contact: anna.tramontano@uniroma1.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The study of drug absorption is of critical importance in the development of effective drugs. The path of a drug from the site of administration to its target cells or compartments implies the crossing of several semipermeable cell membranes, therefore it is relevant to be able to predict whether and to which extent a molecule can pass through the cell membranes.

Passive permeation of drugs through the biological cell membranes is obviously strongly dependent on the molecule physicochemical properties (Meanwell, 2011). It has been established that the acid–base character of the molecule (which influences the charge of the molecule at the specific pH), its lipophilicity (which affects its partition between aqueous and lipid environments) and solubility are the most relevant parameters to take into account. These parameters are well described by the molecule hydropathy profile (Siew *et al.*, 2012; Smith *et al.*, 2010). A more lipophilic drug is more likely to effectively cross the hydrophobic phospholipid bilayer. On the other hand, extremely hydrophobic molecules, insoluble in aqueous body fluids, might be poorly absorbed (Frenkel *et al.*, 2005). In summary, there should be an appropriate balance between the hydrophobicity and hydrophilicity of a molecule (Ghuman *et al.*, 2005; Seelig *et al.*, 1994; Waring, 2009).

From an experimental point of view, data on permeability can be obtained by in situ and/or in vivo animal studies, but these are time consuming and expensive experiments and therefore only performed towards the end of the drug development process. Efforts have therefore focused on the development of in vitro permeability assays that can mimic the relevant characteristics of in vivo absorption. Among these, there are the Parallel Artificial Membrane Permeability Assay (PAMPA) (Avdeef *et al.*, 2007), the human colon adenocarcinoma (Caco-2) cell line assay (Artursson *et al.*, 2001), the Madin-Darby Canine Kidney (MDCK) cell assay (Irvine *et al.*, 1999), the rat duodenal immortalized cell line assay (2/4A1 cell) (Tavelin *et al.*, 2003), and the rat everted gut sac assay (Bohets *et al.*, 2001). All of them are routinely used for the preliminary assessment of drug permeability. In particular, the Caco-2 cell is

probably the most extensively characterized cell-based model and the most popular both in the pharmaceutical industry and in academia (Balimane *et al.*, 2006). It has been shown that this model can effectively predict the human initial drug absorption (Artursson and Karlsson, 1991) because it reflects the transport of the drug across a cell membrane rather than the interaction of the drug with the lipid bilayer (Hou *et al.*, 2006).

The membrane permeability for a given compound is usually estimated from its partition coefficient, logP, defined as the logarithm of the relative concentration of the molecule when it partitions between a two-phase system, usually water and octanol, where the latter is assumed to have a lipophilicity comparable to that of a cell membrane (Artursson *et al.*, 2001; Seddon *et al.*, 2009).

From the theoretical point of view, many computational approaches have been developed to infer drug properties, such as bioavailability, aqueous solubility, initial absorption, plasma-protein binding and toxicity (van de Waterbeemd and Gifford, 2003). These are often related to features such as molecular size, hydrophobicity, or number of hydrogen bonds established by the compound with water molecules (since these bonds need to be broken to allow the molecule to pass the membrane) (Hou *et al.*, 2004).

In general, permeability may be estimated in terms of the free energy barrier that the drug should overcome when crossing the membrane, which is usually predicted from computationally intensive molecular dynamics simulations of the translocation process (Carpenter *et al.*, 2014; Meng and Xu, 2013). Some methods compute the Polar Surface Area (PSA) of the drug to predict its permeability under the assumption that this parameter correlates with the hydrogen-bonding pattern in the aqueous solvent of the molecule and therefore with the energy cost of transferring the molecule from the solvent to the membrane (Kelder *et al.*, 1999; Stenberg *et al.*, 1999).

Other popular methods are the Quantitative Structure-Property Relationship (QSAR) analysis (Yu and Adedoyin, 2003), Multiple Linear Regression (MLR), Partial Least Square (PLS), Linear Discriminant Analysis (LDA), Artificial Neutral Networks (ANNs), Genetic Algorithms (Gas), Support Vector Machines (SVMs) and the 'Lipinski rule of five' (Lipinski, 2000). In particular, the Lipinski's rule takes into account different features to assess whether a compound is likely to be cell membrane permeable and easily absorbed by the body on the basis of the following criteria: molecular weight of the compound lower than 500; logP lower than 5; number of hydrogen bond donors (usually the number of hydroxyl and amine groups in a drug molecule) lower than 5; number of groups that can accept hydrogen atoms to form hydrogen bonds (estimated by the number of oxygen and nitrogen atoms) lower than 10.

In this work we describe a new method based on an estimate of the hydropathy and charge distribution of a compound deduced from the distribution and orientation of the water molecules around it. We have already successfully used a similar approach to estimate the hydrophobicity of the twenty natural amino acids (Bonella *et al.*, 2014). Here we show that, when applied to a set of 23 drugs, neutral at physiological pH, to compute their hydrophobicity and charge distribution, the method can effectively predict their ability to cross the plasma membrane.

Our dataset only includes neutral compounds since these are well known to mainly use passive transport to cross the phospholipid bilayer of the cell membrane (Neuhoff *et al.*, 2003, 2005; Seelig, 2007) and therefore their diffusion and permeability is essentially related to their chemico-physical properties that is what our method can infer.

2 Methods

We analyzed the hydration of small solutes by investigating the changes in the structure of the dynamic hydrogen bond network formed by the water molecules surrounding them as well as their orientation as obtained by Molecular Dynamics (MD) simulations.

2.1 Molecular dynamics

All simulations were performed using NAMD 2.7b1 (Phillips *et al.*, 2005) and the CHARMM force field was used for the investigated compounds (MacKerell *et al.*, 1998). In each simulation a single solute molecule was located in a cubic simulation box (with imposed periodic boundary conditions) filled with TIP4P rigid water molecules (Abascal and Vega, 2005). Each simulation contained a single copy of the compound and the size of the box varied in a range of 56–62 Å depending on the compound considered. The topologies and parameters for the small molecule compounds were obtained via the SwissParam server (Zoete *et al.*, 2011) [www.swissparam.ch] that generates molecules topologies and parameters for small organic compounds in a functional form that is compatible with the CHARMM force field.

The Particle Mesh Ewald (PME) method was used to calculate the electrostatic interactions. Each simulation was run for 1.5 ns. A 1 fs time step was used and the coordinates were retrieved every 0.5 ps. All simulations were performed at $T = 310$ K, and the system was thermostated using Langevin dynamics. The simulations were performed also at constant pressure using a modified Nosé-Hoover method in which Langevin dynamics is used to control fluctuations in the barostat (Hoover, 1985). More details about the molecular dynamics simulation parameters are available at: http://arianna.med.uniroma1.it/neutraldrugs/.

2.2 Dataset

We used a sample set of structurally diverse, small molecular weight drugs analyzed by Yazdanian *et al.* (1998) for which in vitro Caco-2 cell permeability data is available. We selected 23 neutral drugs at pH 7.4 from this dataset.

The advantage of selecting this specific dataset is that the data have been obtained in the same experimental conditions. To verify how representative our dataset is, we collected data for 131 compounds available in the literature for a total of 277 Caco-2 cell permeability values (different values have been obtained for a number of these drugs in different experimental conditions) (Artursson, 1990; Artursson and Karlsson, 1991; Artursson and Magnusson, 1990; Augustijns *et al.*, 1996; Aungst *et al.*, 2000; Chong *et al.*, 1997; Collett *et al.*, 1996; Gres *et al.*, 1998; Haeberlin *et al.*, 1993; Hilgendorf *et al.*, 2000; Hou *et al.*, 2004; Hovgaard *et al.*, 1995; Lentz *et al.*, 2000; Liang *et al.*, 2000; Rubas *et al.*, 1993; Ruiz-Garcia *et al.*, 2002; Saha and Kou, 2002; Schipper *et al.*, 2001; Wu *et al.*, 2000; Yee 1997; Zhu *et al.*, 2002) and compared both their Caco2 experimental values (Supplementary Figure S1) and their structural features.

To estimate the latter, we computed the structural dissimilarity of our selected compounds and compared it with that of the 131 compounds. To this end, we used the ChemMine tool (Backman *et al.*, 2011) that takes into account parameters such as partition coefficient, rule-of-five, partial charges, fingerprint calculation and more (for a detailed description of the features see 'http://www.ra.cs.uniuebingen.de/software/joelib/tutorial/descriptors/descriptors.html'). We used these values to perform a clustering analysis, using the 'hclust' function of R software package (Ihaka and Gentleman,

1996) [http://www.R-project.org], the results of which are shown in Supplementary Figure S2.

As it can be seen, the 23 compounds from the Yazdanian *et al.* (1998) dataset, selected for the analysis, span quite uniformly about 85% of the available range both in terms of Caco2 values and of structural features. Some regions of the feature space are less well represented in our dataset (left most branch of the tree in Supplementary Figure S2). These are all compounds with a rather large molecular weight (above 500 Da). This might imply that our method might behave differently for very large compounds (that in any case are usually excluded a priori as leads because of their size).

All compound three-dimensional coordinates were downloaded from the free public database ZINC (Irwin *et al.*, 2012) [zinc.docking.org]. For this study the following small molecule compounds were chosen: Griseofulvin, Aminopyrine, Piroxicam, Diazepam, Nevirapine, Phenytoin, Testosterone, Progesterone, Clonidine, Corticosterone, Estradiol, Hydrocortisone, Dexamethasone, Scopolamine, Zidovudine, Urea, Uracil, Sucrose, Hydrochlorothiazide, Mannitol, Ganciclovir, Acyclovir and Chlorothiazide (Table 2). Of importance, they cover a wide range of permeability values (Pcaco-2), from 36.6×10^{-6} cm/s to 0.19×10^{-6} cm/s and are as evenly distributed as possible (see Table 2).

2.3 Data analysis

The results of the molecular dynamics simulations of each molecule are used to evaluate the orientation of the water molecules in the first and second hydration shell, being the first related to the hydrophilic and the second to the hydrophobic characteristics of the compound, respectively (see ref (Bonella *et al.*, 2014) for details).

We represent each water molecule as a tetrahedron, where an sp3-hybridized oxygen atom lies at the center and two hydrogen atoms and two lone pair electrons point to the vertices. Each water molecule can then form up to four hydrogen bonds with other water molecules. According to this model of the water molecule, we can define four Hydrogen Bond Vectors (HBVs) and one dipole vector (Fig. 1). The HBVs are defined as the lines connecting the oxygen atom and the vertices of the tetrahedron (in blue in the Fig. 1). The dipole vector (in red in the Fig. 1) lies along the bisectrix of the angle formed by the oxygen and the two hydrogen atoms.

We can define the angles related to hydrogen bond orientations (θ_{b1}, θ_{b2}, θ_{b3} and θ_{b4}) as those formed by the straight line linking the solute atom with the oxygen atom of the nearest water molecule and the hydrogen bond vector (for clarity, only one of the four angles is represented in blue in Fig. 1). Similarly, we can define the angle θ_d related to the orientation of the dipole vector as the angle formed by the straight line connecting a solute atom (S in Fig. 1) to the oxygen atom of the closest water molecule (in black) and the dipole vector of the molecule itself (in red). The different orientations of the water molecules around a solute can be used to analyze the compound hydrophilicity and hydrophobicity. In fact a water molecule in the vicinity of a hydrophobic solute positions one of the faces of the tetrahedron toward the solute. On the other hand, for a hydrophilic solute, a water molecule reorients to point toward the compound with one of its vertices. We need to take the dipole vector into account because the four vertices of the tetrahedron representing the waters are equivalent in our model and therefore it would be impossible to distinguish between positive and negative partial charges without considering θ_d.

At each step of the molecular dynamics simulation, we can measure the values of the five angles (θ_{b1}, θ_{b2}, θ_{b3}, θ_{b4} and θ_d) and the distance R (Å) between each water molecule and the nearest solute atom and compute the probability of finding a water molecule with a given orientation and around at a given distance from the solute atoms.

The hydropathy and charge distribution properties are computed from the conditional probability density of the waters in the appropriate intervals of the angles and distances described before. We can build two three-dimensional histograms for each simulation; the first reports the conditional probability density $P(\theta_{bi}|\mathrm{R})$ (for $i = 1, 2, 3, 4$), the second is the conditional probability density $P(\theta_d|R)$. R is defined as the distance between each solute atom and the oxygen atom of the nearest water molecule. The histogram distance and angle bins were set to 0.05 Å and 1°, respectively (Bonella *et al.*, 2014).

2.4 Molecular descriptors

The analysis of the conditional probability density distributions allows us to compute four indices, named I_y, I_n, I_+ and I_-, obtained by summing the intensity of the peaks in the appropriate angle and distance range.

As described in more detail in our previous work (Bonella *et al.*, 2014), the distribution $P(\theta_{bi}|R)$ permits to distinguish between the hydrophilicity and hydrophobicity of a compound on the basis of the probability values observed in the first and second hydration shell, respectively. Intuitively, this is justified by the fact that a polar solute will establish Coulomb interactions with the closest water molecules and this situation will contribute to the peaks observed in the first hydration shell of the hydrogen bond histogram, while a hydrophobic (or apolar) solute will cause the waters to orient themselves as to maximize the number of hydrogen bonds with neighboring waters, forming a cage around the solute, and will contribute to peaks in the second hydration shell in the hydrogen bond histogram.

The dipole probability density $P(\theta_d|R)$ in the first hydration shell takes into account which of the vertices of the tetrahedron

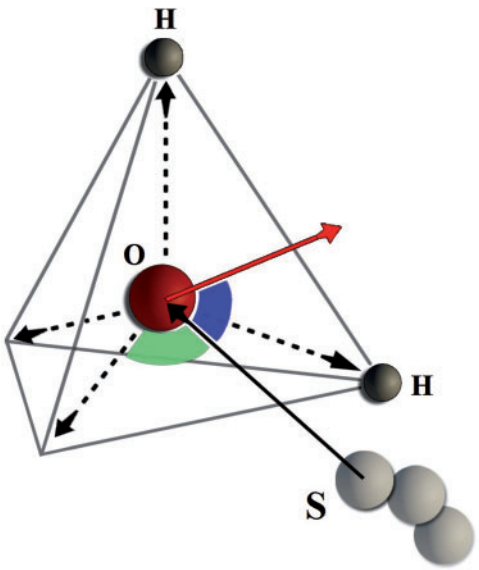

Fig. 1. Definition of the angles used in the analysis. We connect each solute atom (S in the figure) to the oxygen atom (red circle) of the closest water molecule and the same oxygen atom to each vertex of the water tetrahedron, thus defining the four angles, θ_{h1}, θ_{h2}, θ_{h3} and θ_{h4} (for clarity only one, in green, is shown in the figure). Hydrogen atoms are represented as dark grey circles. We also define the dipole vector of the water molecule (red arrow) and compute the angle θ_d between this vector and the line connecting the solute atom and the oxygen (in blue)

representing the waters (all equivalent in our model) is oriented towards the solute and therefore provides information about the electric charge (positive or negative) of the interacting solute atoms.

We define the compound hydrophilicity I_y and hydrophobicity I_n as the sum of the hydrogen bond probability densities, computed over the appropriate distance and angle range ($\Delta\theta$ and ΔR) in the first and second shell of hydration, respectively. The charge indices I_+ and I_- are defined as the sum, in the appropriate range, of the probability densities in the first shell of the distribution related to dipole moment (see Fig. 2). For more details, see ref. (Babiaczyk *et al.*, 2010; Bonella *et al.*, 2014).

As shown in Supplementary Figure S3a–c, the length of the MD simulation (1.5 ns) is sufficient to ensure convergence of the indices.

The scheme used to select the boundaries of the region ($\Delta\theta$ and ΔR) is based on Guassian fits. In particular, we performed a Gaussian fit of the probability distribution for both the first and second hydration shell along the θ axis (see Supplementary Fig. S4) and determined the average and standard deviation of the Guassian distributions for each of the compounds. The average of these values is used to compute the volume of each peak. A similar approach has been used to determine the range of integration along the R axis.

The analytical details of the scheme used to select the boundaries of the region ($\Delta\theta$ and ΔR) are described in the supporting information. The scripts for running the simulations and perform the analysis are available at: http://arianna.med.uniroma1.it/neutraldrugs/.

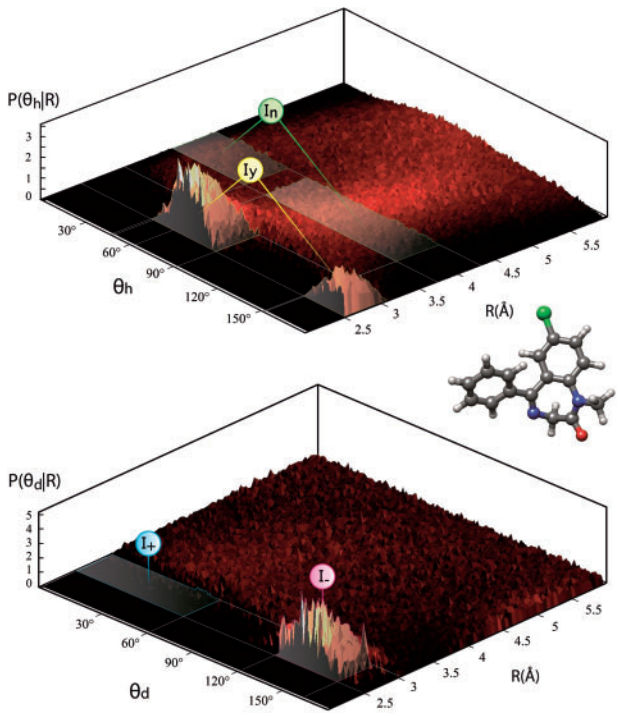

Fig. 2 Histograms of $P(\theta_{hi}|R)$ and $P(\theta_d|R)$ for Diazepam. In both histograms the cells highlighted in grey are used to calculate the sum of the conditional probability densities at each given angle and distance. In the $P(\theta_{hi}|R)$ histogram, the yellow arrows indicate the first and second component of the hydrophilic index related to the two peaks in the first hydration shell. The green arrows show the first and the second component of the hydrophobic peaks that are localized in the second hydration shell. In the $P(\theta_d|R)$ histogram, the blue arrow indicates the contribution of positive charge distribution. The pink arrow indicates the contribution of the negative charge distribution

2.5 Statistical analysis and comparison with other methods

The program used to analyze the molecular dynamics trajectories and to build the histograms was written in Fortran90. The R package (Ihaka and Gentleman, 1996) [http://www.R-project.org] was used to analyze the histograms. The same package was used to calculate the indices, perform the Gaussian fitting and the Multiple Regression Analysis (MRA), compute the Pearson's correlation coefficient, (r) and perform the cross validation analysis. The clustering analysis was performed using the Euclidean distance and via the 'hclust' function from the 'Stats' package of *R* (in particular, the 'average' method of the 'hclust' function was used).

We compared our results with those of several other methods. In particular we computed, for each of the 23 compounds, the predicted permeability values according to the two methods described in ref. (Fujiwara *et al.*, 2002), based on a linear combination of molecular descriptors (Fuij_1), or including quadratic terms (Fuij_2). We also compared our results with those obtained by a linear regression (Hou) and a multiple linear regression (Guangli and Yiyu, 2006) (Gua_1) method. Finally we also used for comparison the Support Vector Machine based method (Gua_2) described in ref. (Guangli and Yiyu, 2006)

3 Results

In silico permeability prediction is consistent with available published data. We computed four indicators (I_y, I_n, I_+ and I_-) described in the Methods section for each of the drugs in our dataset. As explained in detail in the Methods section, these indices are derived from the conditional probability of finding a water molecule with a given orientation around the solute atoms estimated from the results of molecular dynamics simulations. In particular, the first two (I_y and I_n) provide information about the hydrophilic and hydrophobic properties of the compound and are computed from the probability values of finding water molecules in the first and second hydration shells, respectively. I_+ and I_- are related to the dipole orientation of the water molecules surrounding the analyzed compound and therefore to the effect of its positive and negative charges.

The values of the indices for the analyzed molecules are reported in Supplementary Table S1. Three of these parameter-free indicators (I_n, I_+ and I_-) correlate remarkably well with the permeability data while the I_y index shows a lower level of correlation.

We tested whether a combination of these indices can represent a good proxy for estimating the permeability of a molecule. To this end, we used a multiple linear regression algorithm as implemented in the *R* function 'lm' (Ihaka and Gentleman, 1996) to find the weights providing the best correlation with the Caco-2 experimental data. The tool also provides the probability *P*-value of a computed coefficient to be different from 0. We tested both linear and quadratic terms in the regression. The best correlation is obtained by a linear fit of the I_n and I_- indices (*P*-value < 0.001), while I_y and I_+ were found to contribute very little to the overall correlation (*P*-value > 0.05). This is consistent with the values of their correlation coefficients (see Table 1).

The regression model corresponding to the best fit is:

$$P_{pred} = (a*I_n) + (b*I_-) + c \quad (1)$$

where $a = 3.06$ (*P*-value $= 4.7 \times 10^{-7}$), $b = 0.04$ (*P*-value $= 2.6 \times 10^{-3}$) and $c = 3092$ (*P*-value $= 4.0 \times 10^{-7}$).

In Table 1 we also report the correlation between each index and the Caco-2 permeability values. As aspect the highest linear correlation value is between I_n index (hydrophobic index) and Caco-2 permeability value because a more lipophilic drug is more likely to effectively cross the hydrophobic phospholipid bilayer. More interesting is the correlation linked to positive charge distribution index I_+. It can be observed that the index with the highest value of negative correlation is I_+, indicating that most likely positive groups prevent uptake of compounds more than negative ones (see also Supplementary Fig. S5).

Table 2 reports the predicted Ppred permeability values obtaining using Eq. (1) for all the drugs considered and shows that they reproduce very well the experimental Caco-2 permeability values (Pearson's correlation coefficient, $r = 91\%$). We also performed a cross validation analysis by repeatedly leaving out 20% of the compounds (testing sets) and re-computing the coefficients of Eq. (1) on the remaining ones (training sets) as described in the Methods section. We iterated this procedure 10 000 times, randomly choosing the training set at each step. The predicted average values ($P_{\text{pred\_CV}}$) obtained for each drug in the test set are reported in Table 2. Once again, the correlation between prediction and experiment is very satisfactory (88%) (Fig. 3).

The coefficients of Eq. (1) are also very stable. Their average value and standard deviation obtained in the 10 000 cross validation runs are: $a = 3.064 \pm 0.194$, $b = 0.043 \pm 0.005$ and $c = -3091.866 \pm 195.453$.

The average difference between the predicted and experimental values is 4.7×10^{-6} cm/s. It is relevant to mention here that the threshold used to discriminate between low absorbance and high absorbance compounds is usually set to 8.0×10^{-6} cm/s (Castillo-Garit, *et al.*, 2008) and the data shown in Table 1 demonstrate that only in two cases (Acyclovir and Zidovudine) our method would significantly misclassify the compound.

In summary, Eq. (1) describes well the permeability properties of neutral compounds. It is worth noticing that the P_{pred} value is well balanced in the sense that it overestimates and underestimates the experimental values in a similar number of cases (11 and 12 respectively).

We compared our results with those of several other methods (as described in the Methods section) and the results are reported in Table 3 and Supplementary Figure S6a–e. It can be appreciated that the correlation between predicted and experimental values is higher for our method. The average error is lower than all other tested methods, but for the Gua_2 method (Guangli and Yiyu, 2006) that shows a very similar value.

Table 1. Correlation between the values of the indices in our dataset

	I_y	I_n	I_+	I_-	Caco-2
I_y	1	−0.05	−0.39	0.89	0.28
I_n		1	0.76	0.33	0.85
I_+			1	−0.76	−0.81
I_-				1	0.59

Also the correlation value between each index and Caco-2 experimental value is reported.

Table 2. Experimental and predicted permeability values

Drug	$P_{\text{caco-2}}$	P_{pred}	$P_{\text{pred\_CV}}$
Griseofulvin	36.6	31.07	29.96 ± 1.09
Aminopyrine	36.5	37.41	37.62 ± 1.53
Piroxicam	35.6	24.52	23.46 ± 0.65
Diazepam	33.4	29.87	29.26 ± 1.03
Nevirapine	30.1	31.52	31.81 ± 1.04
Phenytoin	26.7	24.53	24.27 ± 0.75
Testosterone	24.9	21.89	21.64 ± 0.63
Progesterone	23.7	29.20	30.78 ± 1.12
Clonidine	21.8	21.94	21.90 ± 2.33
Corticosterone	21.2	15.67	15.01 ± 0.74
Estradiol	16.6	15.57	14.39 ± 1.88
Hydrocortisone	14	10.55	10.17 ± 0.81
Dexamethasone	12.2	10.51	10.33 ± 0.81
Scopalamine	11.8	21.93	22.79 ± 0.53
Zidovudine	6.9	13.61	14.12 ± 0.66
Urea	4.56	4.64	4.67 ± 0.98
Uracil	4.24	8.61	9.01 ± 0.77
Sucrose	1.7	−2.35	−3.31 ± 1.45
Hydrochlorothiazide	0.51	5.20	5.85 ± 0.91
Mannitol	0.38	−8.10	−11.42 ± 1.20
Ganciclovir	0.38	2.38	2.64 ± 1.10
Acyclovir	0.25	9.33	10.16 ± 0.65
Chlorothiazide	0.19	5.03	5.64 ± 0.91

The first column reports the drug name, the second reports the experimental values, the third (Ppred) the values obtained using Eq. (1). The last column reports the predicted values obtained in the cross validation test ($P_{\text{pred\_CV}}$).

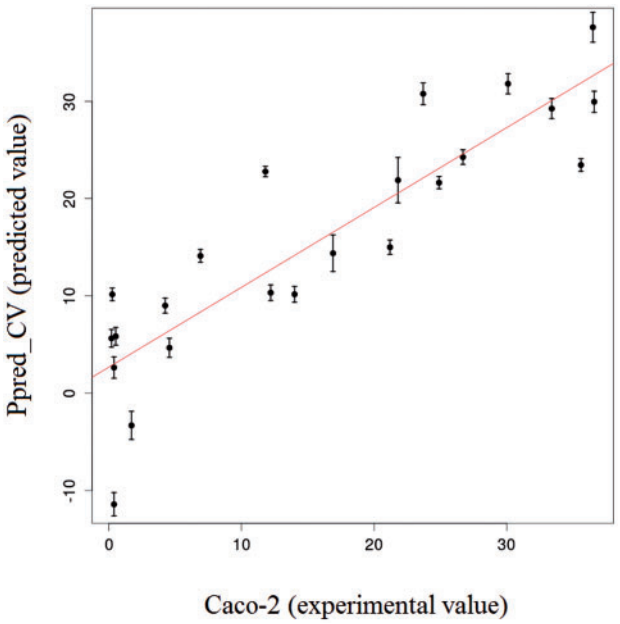

Fig. 3. Scatter plot correlating the predicted permeability values in the cross validation ($P_{\text{pred\_CV}}$) and their experimental Caco-2 values. For each compound the average predicted value and the standard deviation are reported

Table 3. Comparison of the results of the P_{pred} method with those obtained by a number of other predictors (described in Experimental section). P_{pred_CV} (using the test set data) also has been reported

	Fuij_1	Fuij_2	Hou	Gua_1	Gua_2	P_{pred}	P_{pred_CV}
r	0.66	0.62	0.80	0.85	0.79	0.91	0.88
R^2	0.43	0.39	0.64	0.72	0.63	0.83	0.78
Average error	8.2	9.9	4.8	6.2	4.8	4.3	
Ref.	(59)	(59)	(18)	(60)	(60)		

The goodness of fit parameters (r and R^2) are also shown.

4 Conclusion

We have shown here that an approach based on the simultaneous analysis of molecule hydrophobicity and charge distribution has the potential to accurately predict the passive plasma membrane permeability of neutral drugs. This method may be useful for investigating the mechanism of passive permeation of small neutral compounds since it can easily provide information on the role that every single atom plays on the hydration process.

Our P_{pred} indicator correlates very well with the experimentally determined Caco-2 permeability values and performs better than other available methods. Furthermore, it only requires the knowledge of the chemical structure of the compound. Given the cost and impact of late stage failures in drug development we believe that the relatively high computational cost of running the molecular dynamics simulations (an average of 48 hours on a 20 CPU server for each molecule) is not necessarily a relevant drawback of the approach.

As is the case also for several in vitro methods, our method cannot estimate the permeability of drugs that use an active uptake system. In these cases, additional techniques, such as docking the compounds to efflux/influx protein models, should be explored.

Acknowledgements

The authors would like to thank Prof. Antonello Mai for critical comments and discussion, Dr Claudio Graziani and Dr Jacopo Falleti for helping with the graphical representations.

Funding

KAUST Award No. KUK-I1-012-43 made by King Abdullah University of Science and Technology. Progetto di Ricerca di Università, anno 2014 - prot. C26A14RFYP, EPIGEN flagship Project and PRIN 20108XYHJS.

Conflict of Interest: none declared.

References

Abascal,J.L. and Vega,C. (2005) A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.*, **123**, 234505

Artursson,P. (1990) Epithelial transport of drugs in cell culture. I: a model for studying the passive diffusion of drugs over intestinal absorptive (Caco-2) cells. *J. Pharm. Sci.*, **79**, 476–482.

Artursson,P. and Karlsson,J. (1991) Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (Caco-2) cells. *Biochem. Biophys. Res. Commun.*, **175**, 880–885.

Artursson,P. and Magnusson,C. (1990) Epithelial transport of drugs in cell culture. II: effect of extracellular calcium concentration on the paracellular transport of drugs of different lipophilicities across monolayers of intestinal epithelial (Caco-2) cells. *J. Pharm. Sci.*, **79**, 595–600.

Artursson,P. *et al.* (2001) Caco-2 monolayers in experimental and theoretical predictions of drug transport. *Adv. Drug Deliv. Rev.*, **46**, 27–43.

Augustijns,P. *et al.* (1996) Transport of artemisinin and sodium artesunate in Caco-2 intestinal epithelial cells. *J. Pharm. Sci.*, **85**, 577–579.

Aungst,B.J. *et al.* (2000) The influence of donor and reservoir additives on Caco-2 permeability and secretory transport of HIV protease inhibitors and other lipophilic compounds. *Pharm. Res.*, **17**, 1175–1180.

Avdeef,A. *et al.* (2007) PAMPA–critical factors for better predictions of absorption. *J. Pharm. Sci.*, **96**, 2893–2909.

Babiaczyk,W.I. *et al.* (2010) Hydration structure of the quaternary ammonium cations. *J. Phys. Chem. B*, **114**, 15018–15028.

Backman,T.W. *et al.* (2011) ChemMine tools: an online service for analyzing and clustering small molecules. *Nucleic Acids Res.*, **39**, W486–W491.

Balimane,P.V. *et al.* (2006) Current industrial practices of assessing permeability and P-glycoprotein interaction. *AAPS J.*, **8**, E1–13.

Bohets,H. *et al.* (2001) Strategies for absorption screening in drug discovery and development. *Curr. Top. Med. Chem.*, **1**, 367–383.

Bonella,S. *et al.* (2014) Mapping the hydropathy of amino acids based on their local solvation structure. *J. Phys. Chem. B*, **118**, 6604–6613.

Carpenter,T.S. *et al.* (2014) A method to predict blood-brain barrier permeability of drug-like compounds using molecular dynamics simulations. *Biophys. J.*, **107**, 630–641.

Castillo-Garit,J.A. *et al.* (2008) Estimation of ADME properties in drug discovery: predicting Caco-2 cell permeability using atom-based stochastic and non-stochastic linear indices. *J. Pharm. Sci.*, **97**, 1946–1976.

Chong,S. *et al.* (1997) Evaluation of Biocoat intestinal epithelium differentiation environment (3-day cultured Caco-2 cells) as an absorption screening model with improved productivity. *Pharmaceutical Research*, **14**, 1835–1837.

Collett,A. *et al.* (1996) Comparison of HT29-18-C1 and Caco-2 cell lines as models for studying intestinal paracellular drug absorption. *Pharm. Res.*, **13**, 216–221.

Frenkel,Y.V. *et al.* (2005) Concentration and pH dependent aggregation of hydrophobic drug molecules and relevance to oral bioavailability. *J. Med. Chem.*, **48**, 1974–1983.

Fujiwara,S. *et al.* (2002) Prediction of Caco-2 cell permeability using a combination of MO-calculation and neural network. *Int. J. Pharm.*, **237**, 95–105.

Ghuman,J. *et al.* (2005) Structural basis of the drug-binding specificity of human serum albumin. *J. Mol. Biol.*, **353**, 38–52.

Gres,M.C. *et al.* (1998) Correlation between oral drug absorption in humans, and apparent drug permeability in TC-7 cells, a human epithelial intestinal cell line: comparison with the parental Caco-2 cell line. *Pharm. Res.*, **15**, 726–733.

Guangli,M. and Yiyu,C. (2006) Predicting Caco-2 permeability using support vector machine and chemistry development kit. *J. Pharm. Pharm. Sci.*, **9**, 210–221.

Haeberlin,B. *et al.* (1993) in vitro evaluation of dexamethasone-beta-d-glucuronide for colon-specific drug delivery. *Pharm. Res.*, **10**, 1553–1562.

Hilgendorf,C. *et al.* (2000) Caco-2 versus Caco-2/HT29-MTX co-cultured cell lines: permeabilities via diffusion, inside- and outside-directed carrier-mediated transport. *J. Pharm. Sci.*, **89**, 63–75.

Hoover,W.G. (1985) Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, **31**, 1695–1697.

Hou,T. *et al.* (2006) Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Curr. Med. Chem.*, **13**, 2653–2667.

Hou,T.J. *et al.* (2004) ADME evaluation in drug discovery. 5. Correlation of Caco-2 permeation with simple molecular properties. *J. Chem. Inf. Comput. Sci.*, **44**, 1585–1600.

Hovgaard,L. *et al.* (1995) Drug delivery studies in Caco-2 monolayers. Synthesis, hydrolysis, and transport of O-cyclopropane carboxylic acid ester prodrugs of various beta-blocking agents. *Pharm. Res.*, **12**, 387–392.

Ihaka,R. and Gentleman,R. (1996) R: A language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.

Irvine,J.D. *et al.* (1999) MDCK (Madin-Darby canine kidney) cells: a tool for membrane permeability screening. *J. Pharm. Sci.*, **88**, 28–33.

Irwin,J.J. *et al.* (2012) ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.*, **52**, 1757–1768.

Kelder,J. *et al.* (1999) Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.*, **16**, 1514–1519.

Lentz,K.A. *et al.* (2000) Influence of passive permeability on apparent P-glycoprotein kinetics. *Pharm. Res.*, **17**, 1456–1460.

Liang,E. *et al.* (2000) Mechanisms of transport and structure-permeability relationship of sulfasalazine and its analogs in Caco-2 cell monolayers. *Pharm. Res.*, **17**, 1168–1174.

Lipinski,C.A. (2000) Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharm. Toxicol. Methods*, **44**, 235–249.

MacKerell,A.D. *et al.* (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, **102**, 3586–3616.

Meanwell,N.A. (2011) Improving drug candidates by design: a focus on physicochemical properties as a means of improving compound disposition and safety. *Chem. Res. Toxicol.*, **24**, 1420–1456.

Meng,F. and Xu,W. (2013) Drug permeability prediction using PMF method. *J. Mol. Model.*, **19**, 991–997.

Neuhoff,S. *et al.* (2003) pH-dependent bidirectional transport of weakly basic drugs across Caco-2 monolayers: implications for drug-drug interactions. *Pharm. Res.*, **20**, 1141–1148.

Neuhoff,S. *et al.* (2005) pH-Dependent passive and active transport of acidic drugs across Caco-2 cell monolayers. *Eur. J. Pharm. Sci.*, **25**, 211–220.

Phillips,J.C. *et al.* (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, **26**, 1781–1802.

Rubas,W. *et al.* (1993) Comparison of the permeability characteristics of a human colonic epithelial (Caco-2) cell line to colon of rabbit, monkey, and dog intestine and human drug absorption. *Pharm. Res.*, **10**, 113–118.

Ruiz-Garcia,A. *et al.* (2002) Kinetic characterization of secretory transport of a new ciprofloxacin derivative (CNV97100) across Caco-2 cell monolayers. *J. Pharm. Sci.*, **91**, 2511–2519.

Saha,P. and Kou,J.H. (2002) Effect of bovine serum albumin on drug permeability estimation across Caco-2 monolayers. *Eur. J. Pharm. Biopharm.*, **54**, 319–324.

Schipper,N.G. *et al.* (2001) In vitro intestinal permeability of factor Xa inhibitors: influence of chemical structure on passive transport and susceptibility to efflux. *Pharm. Res.*, **18**, 1735–1741.

Seddon,A.M. *et al.* (2009) Drug interactions with lipid membranes. *Chem. Soc. Rev.*, **38**, 2509–2519.

Seelig,A. (2007) The role of size and charge for blood-brain barrier permeation of drugs and fatty acids. *J. Mol. Neurosci.: MN*, **33**, 32–41.

Seelig,A. *et al.* (1994) A method to determine the ability of drugs to diffuse through the blood-brain barrier. *Proc. Natl. Acad. Sci. U. S. A.*, **91**, 68–72.

Siew,A. *et al.* (2012) Enhanced oral absorption of hydrophobic and hydrophilic drugs using quaternary ammonium palmitoyl glycol chitosan nanoparticles. *Mol. Pharm.*, **9**, 14–28.

Smith,D.A. *et al.* (2010) The effect of plasma protein binding on in vivo efficacy: misconceptions in drug discovery. *Nat. Rev. Drug Discovery*, **9**, 929–939.

Stenberg,P. *et al.* (1999) Prediction of membrane permeability to peptides from calculated dynamic molecular surface properties. *Pharm. Res.*, **16**, 205–212.

Tavelin,S. *et al.* (2003) An improved cell culture model based on 2/4/A1 cell monolayers for studies of intestinal drug transport: characterization of transport routes. *Pharm. Res.*, **20**, 373–381.

van de Waterbeemd,H. and Gifford,E. (2003) ADMET in silico modelling: towards prediction paradise?. *Nat. Rev. Drug Discovery*, **2**, 192–204.

Waring,M.J. (2009) Defining optimum lipophilicity and molecular weight ranges for drug candidates-Molecular weight dependent lower logD limits based on permeability. *Bioorg. Med. Chem. Lett.*, **19**, 2844–2851.

Wu,X. *et al.* (2000) Atorvastatin transport in the Caco-2 cell model: contributions of P-glycoprotein and the proton-monocarboxylic acid co-transporter. *Pharm. Res.*, **17**, 209–215.

Yazdanian,M. *et al.* (1998) Correlating partitioning and caco-2 cell permeability of structurally diverse small molecular weight compounds. *Pharm. Res.*, **15**, 1490–1494.

Yee,S. (1997) In vitro permeability across Caco-2 cells (colonic) can predict in vivo (small intestinal) absorption in man–fact or myth. *Pharm. Res.*, **14**, 763–766.

Yu,H. and Adedoyin,A. (2003) ADME-Tox in drug discovery: integration of experimental and computational technologies. *Drug Discovery Today*, **8**, 852–861.

Zhu,C. *et al.* (2002) A comparative study of artificial membrane permeability assay for high throughput profiling of drug absorption potential. *Eur. J. Med. Chem.*, **37**, 399–407.

Zoete,V. *et al.* (2011) SwissParam: a fast force field generation tool for small organic molecules. *J. Comput. Chem.*, **32**, 2359–2368.