UNIVERSITÀ DEGLI STUDI DI TORINO

JOURNAL OF COMPUTER-AIDED MOLECULAR DESIGN, 29 (4), 2015

DOI: 10.1007/s10822-015-9829-4

Noname manuscript No.
(will be inserted by the editor)

Prediction and Interpretation of the Lipophilicity of Small Peptides

Alessia Visconti · Giuseppe Ermondi* · Giulia Caron · Roberto Esposito

Received: date / Accepted: date

Abstract Peptide-based drug discovery has considerably expanded and solid *in silico* tools for the prediction of physico-chemical properties of peptides are urgently needed. In this work we tested some combinations of descriptors/algorithms to find the best model to predict $\log D_{\text{oct}}$ of a series of peptides. To do that we evaluate the models statistical performances but also their skills in providing a reliable deconvolution of the balance of intermolecular forces governing the partitioning phenomenon. Results prove that a PLS model based on VolSurf+ descriptors is the best tool to predict $\log D_{\text{oct}}$ of neutral and ionised peptides. The mechanistic interpretation also reveals that the inclusion in the chemical structure of a HBD group is more efficient in decreasing lipophilicity than the inclusion of a HBA group.

Keywords Lipophilicity · PLS · SVR · VolSurf+ descriptors

This work has been supported by Ateneo Compagnia di San Paolo-2012-Call 2, LIMPET project.

Alessia Visconti
Department of Genomics of Common Disease, Imperial College London, Du Cane Road, W12 ONN London, United Kingdom.
E-mail: a.visconti@imperial.ac.uk

Giuseppe Ermondi
Molecular Biotechnology and Health Sciences Department, University of Torino, Via Quarello 15, 10135 Torino, Italy.
E-mail: giuseppe.ermondi@unito.it

Giulia Caron
Molecular Biotechnology and Health Sciences Department, University of Torino, Via Quarello 15, 10135 Torino, Italy.
E-mail: giulia.caron@unito.it

Roberto Esposito
Department of Computer Science, University of Torino, Corso Svizzera 185, 10149 Torino, Italy.
E-mail: roberto.esposito@unito.it

Introduction

The decreasing number of approved drugs produced by the pharmaceutical industry demands alternative approaches to improve R&D productivity. In general terms, the suite of currently available drugs can be divided into two categories: small molecule drugs (typically weighting less than 500Da) with a high oral bioavailability, and much larger biologics (typically weighting more than 5000Da) that need to be delivered via injection. The time has now come to look for new approaches exploring molecules fitting between these two molecular weight extremes. The goal is to combine the advantages of small molecules (cost, conformational restriction, membrane permeability, metabolic stability, oral bioavailability) with those of proteins (natural components, target specificity, high potency) [1]. Peptides have weights allowing the exploration of this line of research and peptide-based drug discovery approaches can thus be serious options to improve R&D productivity [2]. As an example let's think at the major role of peptides in studying and modulating Protein-Protein Interactions (PPIs) [3].

However, this line of research has a chance to be successful only if the lessons learnt by traditional medicinal chemistry on small organic molecules are wisely applied. In particular, the prediction of Absorption, Distribution, Metabolism, Excretion and Toxicity (ADME-Tox) behavior of drug candidates in the earliest stage of the drug discovery process has to be adapted to peptidic structures [4–6]. For ADME-Tox prediction, the physico-chemical profiling of peptides is a mandatory step [7]. The determination of the physico-chemical profile of a drug candidate mainly consists of the determination of ionization, solubility, lipophilicity, and permeability properties. Lipophilicity represents the affinity

of a molecule (or a moiety) for a lipophilic environment and, for a neutral compound, it is expressed as the logarithm of the partition coefficient P_{oct} (the molar concentration ratio of a single species between octanol and water at equilibrium). If the molecule is ionisable, then the pH of the aqueous phase will influence the concentrations of ionised and neutral forms of the molecule. The term log D_{oct} is used to reflect the pH dependent lipophilicity of a drug. log D_{oct} refers to the logarithm of the distribution coefficient, D, which is defined as the ratio between the concentrations of all species (neutral and ionized) in octanol and the concentration of all species in water.

Since peptides are often ionized at physiological pH, log D_{oct} is the most relevant lipophicity descriptor that should be evaluated in peptide-based drug discovery. log D_{oct} is in fact of the utmost relevance in ADME-Tox studies [7]. For instance the determination of log D_{oct} is expected to be really important in the prediction of the hydrophobic depoting of peptides (a strategy to prevent too fast renal clearance) and in the description of peptide cellular uptake [8]. Brain-targeted peptide delivery is also strongly dependent on lipophilicity [9].

The very recent interest of medicinal chemists for peptides explains the poor number of studies reported in the literature so far to predict their lipophilicity [9–12]. Nowadays, peptide lipophilicity prediction is mostly performed by using algorithms developed for organic molecules. Moreover, most of these algorithms refer to log P_{oct} rather than log D_{oct} and the conversion of log P_{oct} into log D_{oct} through pK_a values when feasible has to be carefully performed [13].

There is therefore a need for new methods that predict peptides log D_{oct}. Moreover, the knowledge of the dominant intermolecular interactions between the solute and the system is also required for designing new and more powerful peptide-based drugs [14].

The design and implementation of methods to predict and interpret log D_{oct} of peptides is not an easy task. The difficulties are mainly due to the fact that ionization-related problems are rarely rigorously considered by molecular descriptors. Moreover only a relatively small number of experimental log D_{oct} values are available for peptides (especially when compared to the abundance of log D_{oct} data for small organic molecules).

This study describes a method for predicting and mechanistically interpreting log D_{oct} of small peptides.

To achieve this we apply different combinations of descriptors and algorithms.

Among descriptors reviewed by Mannhold and coworkers [15], we believe that VolSurf+ descriptors are the most suited to apply to peptides. In fact, VolSurf+ descriptors (see Methods for definition) are convenient for ionized species, easy to interpret and are obtained from the GRID Force-Field [16], a program originally developed for studying proteins and peptides. Nevertheless we also experiment more standard 2D descriptors implemented in the MOE software.

Partial Least Squares (PLS) and Support Vector for Regression (SVR) algorithms were applied to build the models described in this study. PLS algorithm was chosen since it is largely used in medicinal chemistry [17]. Machine learning tools on the other hand are largely used by biologists and have also been shown to have potential utility in the modelling of pharmaceutical problems [18]. For these reasons we also experiment a machine learning strategy based on the SVR algorithm.

Traditionally both PLS and SVR approaches have been characterised by difficulties in the interpretation of their regression equations. Here we also provide graphical tools to deconvolute the mechanistic information content of PLS and SVR models. These tools were used both to rank models which sport similar statistical performances and to extract practical information to be used in drug discovery programs. To the best of our knowledge only one paper comparing models performances both on prediction and interpretation has been published so far [19] whereas most papers limit the comparison to models statistical quality. (*e.g.,* [20]). This result provides an additional value to the study.

Methods

Dataset Preparation

A dataset of 176 small peptides was put together using the information provided by the literature. All dataset information are reported in Table S1 (Supplementary Information). Experimental log D_{oct} values were measured at pH = 7.4. The studied peptides have a maximum length of 6 amino acids and both linear and cyclic structures are present. Some linear peptides have C-terminal and N-terminal amino acids protected with standard groups. According to MoKa 2.5.4 (http://www.moldiscovery.com), 50 peptides are mostly ionized at pH = 7.4, 16 are positively charged whereas 11 are negatively charged. 23 zwitterionic structures are present.

We assume that small peptides, as those belonging to our dataset, can be structurally considered as small organic molecules and thus that it is legit to use standard building tools for preparing their 3D structures. This approach is indirectly supported by the fact that *in silico* tools for building peptides (*e.g.*, Pep-Fold [21] and I-Tasser [22]) are customarily used to build larger

structures (*i.e.*, with 9 or more amino acids). In this study we used the Protein Builder tool implemented in MOE (version 2012.10) [23] to build the peptides. Amino acids were chosen from the panel of the MOE tool and peptides in extended conformation were obtained. When necessary, the peptides were modified adding organic groups and adjusting ionisable functions according to MoKa predicted ionization state. Finally, the peptides were minimized using Molecular Mechanics tools present in MOE (force field MMFF94x). For comparative purposes we built a second dataset of the investigated peptides using the MOE Rebuild3D tool.

3D structures built with MOE were saved as MOL2 files (the dataset is downloadable in sdf format from `https://sites.google.com/site/cassmedchem/projects/limpet`) and submitted to VolSurf+ (version 1.0.7.l, http://www.moldiscovery.com) for the calculation of 82 VolSurf+ descriptors [24]. We used the default settings and four probes: OH2, DRY, N1 and O, that mimic respectively water, hydrophobic, hydrogen bond acceptor and hydrogen bond donor interaction of the compounds with the environment (see Table S2 in the Supplementary Information).

For comparative purposes peptides were also characterized by means of 82 2D-MOE descriptors (their list is reported in the Supplementary Information, Table S3).

The dataset of 176 peptides was randomly split in a training set of 132 peptides and a test set of 44 peptides. In all the experiments, $\log D_{\mathrm{oct}}$ values were considered dependent variables (Y) and a relation between Y and the VolSurf+ and 2D-MOE descriptors (X) was sought.

One of the peptides, namely KPWtLL, was detected as an outlier and thus removed from the dataset. The reason for this anomalous behavior is due to the complex ionization profile of this peptide that prevents the identification of a single dominant electrical species at pH = 7.4. A complete study of the ionization profile of KPWtLL goes beyond the aim of the paper.

In this study, lipophilicity data were manually checked by the authors (*e.g.* comparable experimental conditions in the original paper). Lipophilicity values lower than -3 were discarded from the study since they could potentially suffer from experimental uncertainty.

PLS

Partial Least Squares is a wide class of methods for modeling relations between sets of observed variables by means of latent variables. It is able to handle regression and classification tasks as well as allowing dimension reduction techniques and modeling tools based on its results. The underlying assumption of all PLS methods is that the observed data is generated by a system or process which is driven by a small number of latent (not directly observed or measured) variables [25]. Projections of the observed data to its latent structure by means of PLS was developed by Wold and co-workers [26]. In its general form PLS creates orthogonal score vectors (also called latent vectors or components) by maximizing the covariance between different sets of variables.

PLS can be naturally extended to regression problems. The predictor, x_i, and predicted (response) variables, y_i, are each considered as a block of variables. PLS regression finds components from vectors x_i that are also relevant for y_i. Specifically, PLS regression searches for a set of components (called latent variables, LV) that performs a simultaneous decomposition of x_i and y_i with the constraint that these components explain as much as possible of the covariance predictor and predicted variables.

In this study we used the PLS tools implemented in VolSurf+. The number of LVs is chosen on the basis of the maximum value of the cross validation Q^2 parameter.

The mechanistic interpretation of regression models is generally obtained through a plot showing the relative importance of descriptors contribution. In the case of PLS models this is achieved by means of Variable Importance in the Projection (VIP) plots [26]. VIPs values are regarded as valuable tools in interpreting PLS models since they are able to take into account both the correlations with the target variable Y as well as the correlations within the X descriptors.

Support Vector Regression

Support Vector Machines are widely successful machine learning tools originally developed for two-class classification problem [27]. Support Vector Regression (SVRs) [28] extend SVMs to cope with the regression problem.

In a regression problem, we are given a dataset D, where x_i is the input vector and y_i is the response variable of the i-th example instance. In SVR, the goal is to find a function that has at most ϵ deviation from the targets y_i for all the training data.

In this study we used the SVRs implemented in the *e1071* R package [29]. We experiment with several settings, using different *kernels* (namely a linear, a polynomial, and a radial kernel) and *regularisation parameters*. Technical details about SVRs, kernels and regularisation parameters can be found in the Supplementary Methods.

While SVRs are considered one of the best off-the-shelf machine learning tools available today, it is also

recognized that they produce hard-to-interpret regressors. Recent research focused on interpreting the SVRs results by leveraging the support vectors (*e.g.*, [19]) and trying to find the characteristics that made them important for the regression. It is worth mentioning that, for the particular case of linear kernels, the w vector (*i.e.*, the coefficients of the linear model) can be reconstructed from the support vectors as $w = \sum_{x_i} \alpha_i x_i$. The set of weights w can be then used to assess the "importance" of a set of features by examining their contribution to the output of the model. However, since different descriptors may have different scales, the magnitude of the descriptor weight alone may not be representative of how important it is for the output. To cope with this difficulty, we define the *Average Contribution* (AC) of the j-th descriptor on dataset D as the quantity:

$$\mathrm{AC}_j(D) = \frac{\sum_{i=1}^{m} x_{ij} w_j}{\sum_{i=1}^{m} \sum_{j=1}^{n} |x_{ij} w_j|}$$

where x_{ij} is the value of the j^{th} descriptor of the i^{th} example in dataset D. As its name implies, $\mathrm{AC}_j(D)$ is the average contribution of the feature j to the value of the regression function when evaluated on dataset D. The numerator sums all the contributions of feature j on the regression output. The denominator is simply the sum of the absolute value of each contribution. The $\mathrm{AC}_j(D)$ ranges in $[-1, 1]$; it equals 1/-1 when the output of the regression function (minus the intercept) can be evaluated using only the j-th feature; it equals 0 when the feature j does not contribute to the result. In the very common case where the SVR package rescales the data before learning (as it is the case with the *e1071* R package used in this paper) to properly compute the $\mathrm{AC}_j(D)$ values, it is necessary to use a modified set of weights (or, alternatively, to rescale each piece of new data before using it with w). Technical details about how the *Average Contribution* were evaluated can be found in the Supplementary Methods.

Results

Models building and validation

In this study log D_{oct} values were considered dependent variables (Y) and a relation between Y and either the 82 VolSurf+ descriptors or the 82 MOE descriptors (X) was sought using both PLS and SVR algorithms.

In the PLS experiments we firstly selected the best performing model in terms of number of LVs (here six) using a cross validation procedure over the training set. Then we validated the model over the test set. Table 1 reports the generalization performances of the selected model over log D_{oct}.

The R^2 and the RMSE values show very good performances for both sets of descriptors and almost no overfitting (*i.e.*, performances measured by RMSE and R^2 over the test set are very close to those measured over the training set). Graphs of predicted versus actual values on both the training and test data are shown in Figure 1 (A) and (B).

In the SVR experiments we experimented with several kernels and selected the best performers. Then these models were evaluated over the training and the test set using the same procedure used for the PLS algorithm (see Supplementary Methods for details). Table 2 summarises SVR results. The three selected kernels (*i.e.* linear, polynomial, and radial kernels) have similar performances sporting good generalization performances for both series of descriptors. All kernels show a modest amount of overfitting for 2D-MOE descriptors and a slightly larger amount of overfitting for VolSurf+ descriptors. Since all kernels performed almost equally well, the simplest model (the linear) has to be preferred over the others. Graphs of predicted (linear kernel) vs actual values on both the training and test data are shown in Figure 1 (C) and (D).

Summing up, all the combinations of descriptors/algorithms performed very well in the prediction of log D_{oct}.

For the sake of comparison, we also performed the corresponding analysis on log P_{oct}. As expected all the log P_{oct} models show excellent performances with any combination of descriptors/algorithms (as an example PLS results are shown in Table S4, Supplementary Information).

Mechanistic interpretation

The four log D_{oct} models described above are substantially equivalent from a statistical point of view and thus a criterion has to be found to prove the superiority of one of them over the others. To do that we decided to rank the models on the basis of their propensity to be interpreted from a mechanistic point of view. This strategy is uncommon but of relevance in the drug discovery practice (see Discussion).

The extraction of the balance of intermolecular forces governing the partitioning phenomenon was obtained through the analysis of *ad hoc* plots. In the case of the PLS models we used the VIPs plots to show the relative importance of the descriptors. Figure 2 (panel (A) and (B)), reports the VIPs plots for the PLS models built with the two set of descriptors (namely, 2D-MOE and VolSurf+).

As detailed in the Methods section, SVR are based on the so-called support vectors: example instances (here

Table 1: PLS results for log D_{oct} models. Each result corresponds to the evaluation of the R^2 and $RMSE$ for PLS models with six latent variables. The first line corresponds to models learnt and tested over the training set, the second line corresponds to models learnt on the training set and tested on the test set.

	MOE descriptors		VolSurf+ descriptors	
	R^2	RMSE	R^2	RMSE
Training set	0.891	0.483	0.882	0.511
Test set	0.872	0.441	0.841	0.592

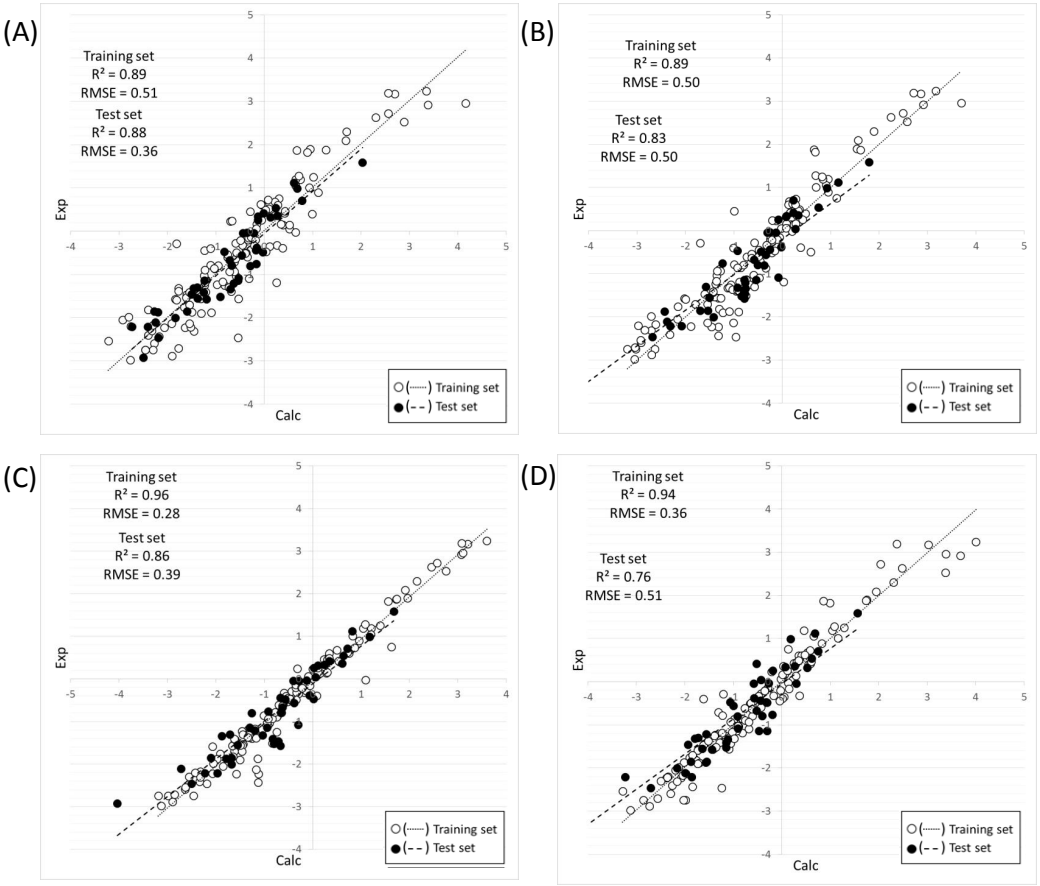

Fig. 1: Plot of experimental versus calculated log D_{oct} values. (A) PLS – 2D-MOE (B) PLS – VolSurf+ (C) SVR – 2D-MOE (D) SVR – VolSurf+. White and black dots represent the training and the test set respectively. SVR plots refer to the linear kernel.

peptides) that are particularly important to solve the regression problem. Indeed, the model built by SVR algorithms is a linear combination of support vectors: a representation very important for algorithmic purposes, albeit hard to understand. To overcome this problem we introduced the *AverageContribution* (AC). Figure 2 (panel (C) and (D)) reports the AC plots extracted from the SVR models.

In Figure 2 (B) and (D), the bars of the 82 VolSurf+ descriptors are color-coded according to the molecular properties they relate to. Specifically, descriptors associated to the size of the molecules are in green, descriptors related to the interaction with water are in cyan, descriptors associated to hydrophobicity are in yellow, descriptors for the hydrogen bonding donor (HBD) properties of the peptides are in red, descriptors for the hydrogen bonding acceptor (HBA) properties are in blue, and descriptors not included in any of the previous classes are in grey.

VolSurf+ VIPs plots (Figure 2 (B)) reveal that the descriptors related to the molecular size (green bars) are important. The positive sign of most of them indicates

Table 2: SVR training and generalization performances. The table reports the training and the generalization performances of the best three models according to the model-selection step: the linear, the polynomial (degree 1) and the radial kernels. Columns report: the kernel function used to build the model, the γ regularization constant (when applicable, n is the size of the training set), and, for each dataset, and the R^2 and RMSE scores.

		MOE descriptors				VolSurf+ descriptors			
		Training		Test		Training		Test	
Kernel	γ	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
linear	–	0.960	0.292	0.863	0.386	0.941	0.357	0.758	0.513
polynomial	1/n	0.902	0.458	0.822	0.440	0.881	0.507	0.835	0.423
radial	1/n	0.969	0.258	0.904	0.322	0.953	0.317	0.790	0.478

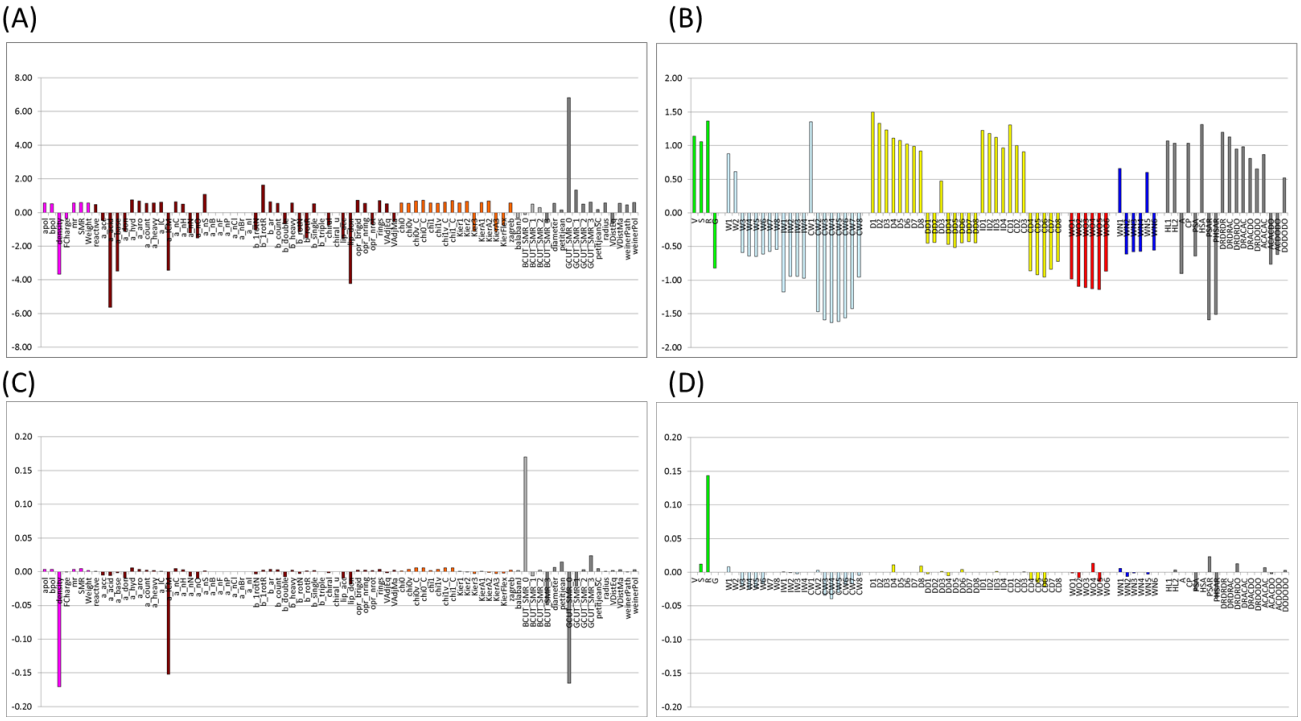

Fig. 2: Models interpretation (log D_{oct} values). (A) PLS – 2D-MOE (B) PLS – VolSurf+ (C) SVR – 2D-MOE (D) SVR – VolSurf+. In (A) and (C) 2D-MOE descriptors are color-coded according to their origin. In (B) and (D) VolSurf+ descriptors associated to the Size of the molecules are in green, descriptors related to the interaction with water are in cyan, descriptors associated to hydrophobicity are in yellow, descriptors for the hydrogen bonding donor (HBD) properties of the peptides are in red, descriptors for the hydrogen bonding acceptor (HBA) properties are in blue, descriptors not included in any class are in grey. SVR plots refer to the linear kernel.

that, as expected, the larger the molecule, the higher the lipophilicity. The hydrophobicity descriptors contributions (yellow bars) are partly positive and partly negative. The positive term could be due to the hydrophobic interactions between the probe and the apolar regions of the molecules. The negative contribution could be related to the interactions due to dipolarity/polarizability properties and occurring between the solute and the aqueous phase of the system. The contribution of the descriptors related with the interaction with water (cyan bars) are the most relevant but have opposite sign (negative) with respect to the green bars. This means that the higher the skills of the solute to interact with water, the higher its propensity to have low lipophilicity. The contribution of the grey bars (Others descriptors) describes molecular properties related to the balance/unbalance of polar regions and take into account the observation that if closely located two (or more) polar regions partially mask their polarity and thus enhance the lipophilicity. The solutes' HBA related descriptors (blue bars) are poorly significative.

The HBD related descriptors (in red) are negative and important.

The bars of the 82 2D-MOE descriptors (Figure 2 (A) and (C)) are color-coded according to their origin (see Table S3 Supplementary Material) since it is not trivial to make a classification on the basis on their significance. For this reason the VIP plot build on the 2D-MOE descriptors (Figure 2 (A)) is more difficult to be used for interpretative purposes. The most relevant descriptors are: density (negative), a_acid and a_base (negative), lip_don (negative) and GUT_SMR_0 (positive). The first could be related to molecular weight whereas the descriptors with the negative sign to HB properties.

Figure 2 (D) shows the relevance of the VolSurf+ descriptors in the SVR models. More evidently than in the case of the PLS experiments, the SVR plot supports the main role played by the descriptors associated to the molecular dimensions (green) in governing lipophilicity. The relevance of HB properties is poorly evident.

Figure 2 (C) shows the relevance of the 2D-MOE descriptors in the SVR models. The most relevant descriptors are four: density (negative), a_ICM (negative), BCUT_SMR_0 (positive) and GCUT_SMR_0 (negative). The negative sign of GCUT_SMR_0 (negative) is somewhat surprising compared to what found in Figure 2 (A).

Log P_{oct} and log D_{oct} are expected to express a different balance of intermolecular interactions. To verify this hypothesis we report in Figure 3 the VIPs plot for log P_{oct}. The relevance of the molecular size seems to decrease when passing from log P_{oct} to log D_{oct} whereas the reverse is true for the contribution of the descriptors associated to HBD properties of the peptides. The remaining descriptors are less influenced.

Discussion

Peptides are a class of organic compounds with chemical features that differ from traditional drugs. They are gaining a growing interest from pharma industries mostly because they can be highly target specific and thus could limit side effects. This evidence calls for specifically tailored drug discovery tools. In particular methods for an efficient prediction of peptides log D_{oct} are strongly needed.

Peptide-based models are generally built using datasets containing about 150 compounds. Two main reasons explain the use of such small datasets: *a)* the limited number of available peptides (at least in comparison with traditional organic drugs) and *b)* experimental data curation limits the number of reliable experimental measurements. The dataset used in this study has thus a standard size for this kind of study. Its applicability domain (*i.e.*, the response and chemical structure space in which the model makes predictions with a given reliability) is in line with the chemical features of new peptides of pharmaceutical interest recently reported in the literature. In fact, drug discovery research is widely focused on peptides having a length of 6-10 amino acids [30,31]. Finally the investigated dataset also contains a significant amount (about 28%) of peptides that at the experimental conditions (pH = 7.4) are in their ionized form (i.e zwitterionic, cationic and anionic). This is an essential requisite since building of a model able to predict lipophilicity of both neutral and ionised species is a challenging endeavour.

The prediction of peptides log P_{oct} is relatively easy to perform, as proven by the fact that log P_{oct} calculators implemented in commercial software (*e.g.*, MOE v.2012.10) provide good predictions of log P_{oct} values of the investigated dataset (see Figure S1). Conversely log D_{oct} predictors generally show poorer performances mostly due to the presence of ionised species (i.e. zwitterions but also anions and cations). As an example we show in Figure 4 (A) the modest predictive performance of ChemAxon log D (http://www.chemaxon.com), an established log D_{oct} model, in the prediction of the lipophilicity of the peptides of the dataset.

In this paper we used different combinations of descriptors (VolSurf+ and 2D-MOE) and algorithms (PLS and SVR) to build models for the prediction of log D_{oct}. We obtained a number of statistically equivalent models (the different performances are not relevant compared to the experimental error associated the measurements). All models performed very well and significantly better than models implemented in commercial software especially in the prediction of log D_{oct} of molecules that are ionized at pH = 7.4. This is shown in Figure 4 (B) taking the model PLS/VolSurf+ as an example. Figure 4 highlights the difficulties of the simultaneous prediction of lipophilicity for both neutral and ionised species experienced by ChemAxon log D but not by our method.

VolSurf+ descriptors are conformation dependent. We verified (Figure S2 of the Supplementary Information) that for this dataset a minor impact of the conformation variability is registered on the VolSurf descriptors and thus on the derived log D models.

The presence of statistically equivalent models provides the evidence that, for this particular dataset, the use of the PLS algorithm returns satisfactory results and that the application of the more complex SVR algorithms may be unnecessary. From a statistical point of view also the choice of descriptors seems to be poorly relevant.

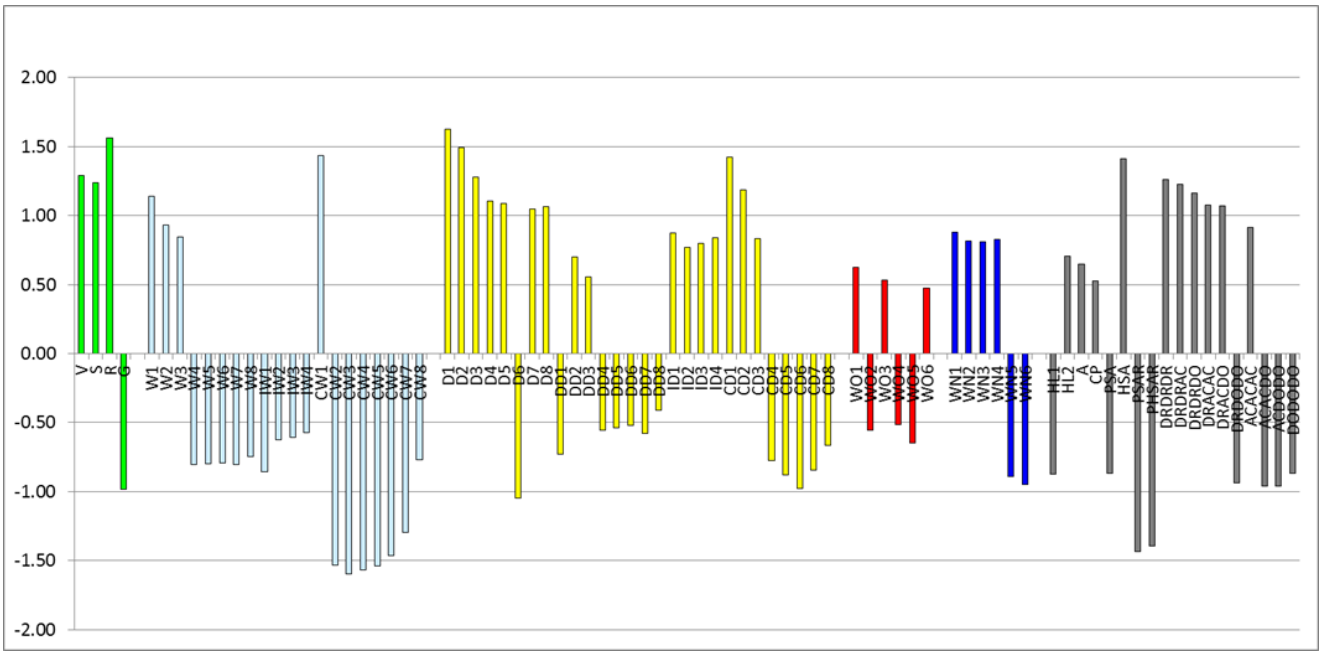

Fig. 3: Models interpretation (log P_{oct} values). PLS – VolSurf+. Descriptors associated to the Size of the molecules are in green, descriptors related to the interaction with water are in cyan, descriptors associated to hydrophobicity are in yellow, descriptors for the hydrogen bonding donor (HBD) properties of the peptides are in red, descriptors for the hydrogen bonding acceptor (HBA) properties are in blue, descriptors not included in any class are in grey.

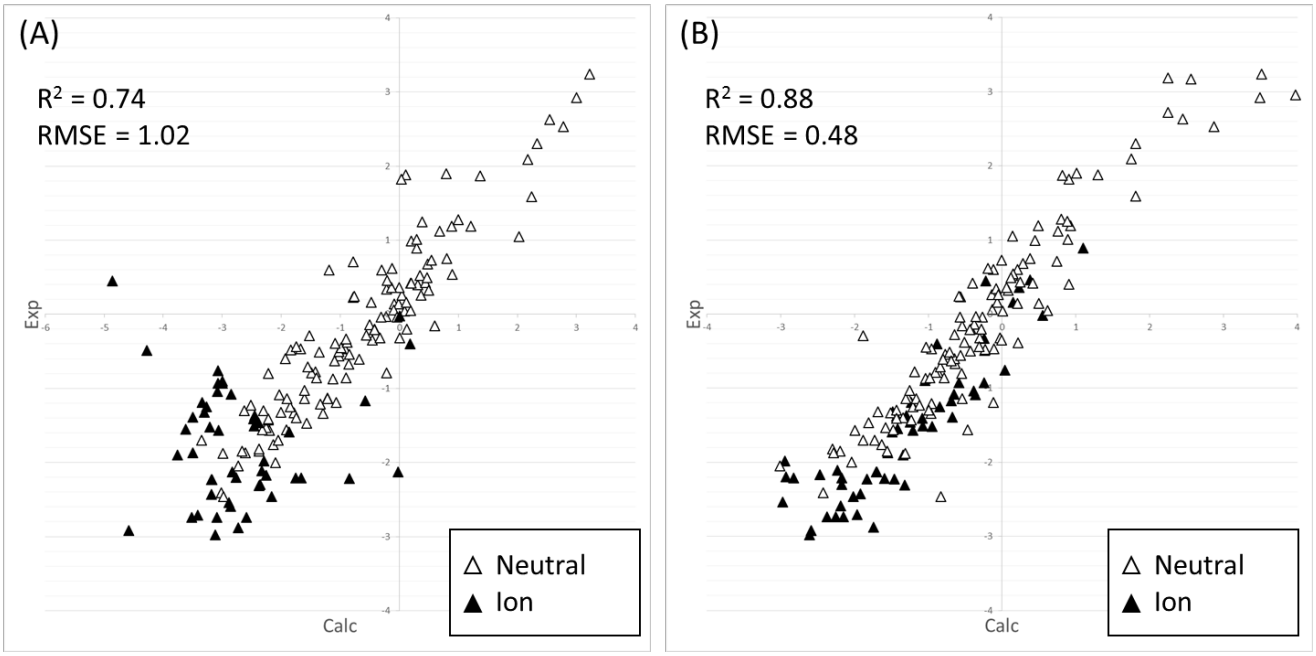

Fig. 4: Correlation between experimental and calculated log D_{oct} values. The calculated values are obtained by (A) ChemAxon log D implemented in Marvin and (B) PLS/VolSurf+ model described here.

To select the best model we propose a criterion based on the propensity of the models to unravel the balance intermolecular forces governing the phenomenon. In practice we state that, if similar statistics are verified, the best model is the one that gives the easiest and more convincing interpretation of the involved phenomenon (here partitioning). The mechanistic interpretation of the models was obtained through graphical analysis. Please note that both PLS and SVR models are poorly analysed in the literature for the mechanis-

tic information they include and thus this aspect is a novelty introduced by our study. VIPs plots are easily obtained whereas the $AC_j(D)$ plots were specifically implemented for this study and are available only for linear kernels. Moreover $AC_j(D)$ plots gives comparable information to VIPs for VolSurf+ descriptors but not for 2D-MOE descriptors.

VIPs plots showed that VolSurf+ descriptors are particularly suited for a mechanistic interpretation. Their strength lies on the homogeneous derivation from MIFs, that enables their classification in groups related to the physico-chemical properties they describe. VIPs based on VolSurf+ descriptors make it evident the dominant role played by the size of the peptides and HBD donor properties in governing their lipophilicity. This is in line with solvatochromic equations [32]. The deconvolution of the information content from the models based on 2D-MOE descriptors is more confusing because of their inhomogeneous derivation from various sources.

The mechanistic interpretation given by the VolSuf+ based PLS model also showed that the log D_{oct} model properly handles with ionized species. This information was extracted by comparison with the correspondent log P_{oct} VIPs plot. In particular the dominant effect played by the block of descriptors related to the size of the solutes seems to decrease when passing from log P_{oct} (Figure 4) to log D_{oct} (Figure 3B) whereas the reverse is true for the contribution of the descriptors associated to hydrogen bond donor properties of the peptides. This suggests that HBD are more efficient in decreasing lipophilicity than HBA. In fact for sp^3 amines the difference between log P_{oct} and log D_{oct} is generally larger than the same difference in acids. At a deeper analysis we verifed that this is true except for guanidines for which protonation causes only a modest decrease in lipophilicity.

Summing-up the best model we produced to predict log D_{oct} was obtained using VolSurf+ descriptors and the PLS algorithm. This model also provides some rules on how to increase/decrease the lipophilicty of peptides through structural modification. This is the result of putting together statistical analysis and mechanistic interpretation. The relevance of the latter is often underestimated in the literature by researchers despite the recommendation of some fields experts along these lines [14].

References

1. Craik, DJ, Fairlie, DP, Liras, S, Price, D (2013) The Future of Peptide-based Drugs. Chem Biol Drug Des 81(1):136–147.
2. Vlieghe, P, Lisowski, V, Martinez, J, Khrestchatisky, M (2010) Synthetic therapeutic peptides: science and market. Drug Discov Today 15(1):40–56.
3. Benyamini, H, Friedler, A (2010) Using peptides to study protein-protein interactions. Future Med Chem 2(6):989–1003.
4. Bose, PP, Chatterjee, U, Hubatsch, I, Artursson, P, Govender, T, Kruger, HG, Bergh, M, Johansson, J, Arvidsson, PI (2010) In vitro ADMET and physicochemical investigations of poly-N-methylated peptides designed to inhibit Aβ aggregation. Bioorgan Med Chem 18(16):5896–5902.
5. Doak, BC, Over, B, Giordanetto, F, Kihlberg, J (2014) Oral Druggable Space beyond the Rule of 5: Insights from Drugs and Clinical Candidates. Chem Biol 21(9):1115–1142.
6. Milletti, F (2012) Cell-penetrating peptides: classes, origin, and current landscape. Drug Discov Today 17(15):850–860.
7. Lipiński, C, Lombardo, F, Dominy, BW, Feeney, P (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Delivery Rev 46:1–3.
8. Nestor, J, et al. (2009) The medicinal chemistry of peptides. Curr Med Chem 16(33):4399–4418.
9. Buchwald, P, Bodor, N (1998) Octanol-water partition: searching for predictive models. Curr Med Chem 5(5):353–380.
10. Hattotuwagama, CK, Flower, DR (2006) Empirical prediction of peptide octanol-water partition coefficients. Bioinformation 1(7):257.
11. Akamatsu, M, Yoshida, Y, Nakamura, H, Asao, M, Iwamura, H, Fujita, T (1989) Hydrophobicity of Di-and Tripeptides Having Unionizable Side Chains and Correlation with Substituent and Structural Parameters. Quant Struct-Act Rel 8(3):195–203.
12. Tao, P, Wang, R, Lai, L (1999) Calculating partition coefficients of peptides by the addition method. Mol Mod Annual 5(10):189–195.
13. Caron, G, Ermondi, G (2003) A comparison of calculated and experimental parameters as sources of structural information: the case of lipophilicity-related descriptors. Mini-Rev Med Chem 3(8):821–830.
14. Cherkasov, A, Muratov, EN, Fourches, D, Varnek, A, Baskin, II, Cronin, M, Dearden, J, Gramatica, P, Martin, YC, Todeschini, R, et al. (2014) QSAR Modeling: Where have you been? Where are you going to? J Med Chem .
15. Mannhold, R, Poda, GI, Ostermann, C, Tetko, IV (2009) Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. J Pharma Sci 98(3):861–893.
16. Goodford, PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 28(7):849–857.
17. Varmuza, K, Filzmoser, P, Dehmer, M (2013) Multivariate linear QSPR/QSAR models: Rigorous evaluation of variable selection for PLS. Comput Struct Biotechnol J 5.
18. Michielan, L, Moro, S (2010) Pharmaceutical perspectives of nonlinear QSAR strategies. J Chem Inf Model 50(6):961–978.
19. Hernández, N, Kiralj, R, Ferreira, M, Talavera, I (2009) Critical comparative analysis, validation and interpretation of SVM and PLS regression models in a QSAR study on HIV-1 protease inhibitors. Chemometr Intell Lab 98(1):65–77.

20. Liao, Q, Yao, J, Yuan, S (2006) SVM approach for predicting LogP. Mol Divers 10(3):301–309.
21. Maupetit, J, Derreumaux, P, Tuffery, P (2009) PEP-FOLD: an online resource for de novo peptide structure prediction. Nucleic Acids Res 37(suppl 2):W498–W503.
22. Roy, A, Kucukural, A, Zhang, Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5(4):725–738.
23. Inc., CCG. Molecular Operating Environment (MOE), version 2012.10. `http://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm`.
24. Cruciani, G, Crivori, P, Carrupt, PA, Testa, B (2000) Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. J Mol Struct - THEOCHEM 503(1):17–30.
25. Rosipal, R, Krämer, N (2006) Overview and recent advances in partial least squares. In Subspace, Latent Structure and Feature Selection, pages 34–51. Springer.
26. Wold, S, Sjöström, M, Eriksson, L (2001) PLS-regression: a basic tool of chemometrics. Chemometr Intell Lab 58(2):109–130.
27. Cortes, C, Vapnik, V (1995) Support-vector networks. Mach Learning 20(3):273–297.
28. Schölkopf, B, Smola, AJ (2002) Learning with kernels. "The" MIT Press.
29. Dimitriadou, E, Hornik, K, Leisch, F, Meyer, D, Weingessel, A (2011). e1071: Misc Functions of the Department of Statistics (e1071).
30. Rafi, SB, Hearn, BR, Vedantham, P, Jacobson, MP, Renslo, AR (2012) Predicting and improving the membrane permeability of peptidic small molecules. J Med Chem 55(7):3163–3169.
31. Mas-Moruno, C, Rechenmacher, F, Kessler, H (2010) Cilengitide: the first anti-angiogenic small molecule drug candidate. Design, synthesis and clinical evaluation. Anti-cancer Agent Me 10(10):753.
32. Abraham, MH, Acree Jr, WE, Leo, AJ, Hoekman, D, Cavanaugh, JE (2010) Watersolvent partition coefficients and LogP values as predictors for bloodbrain distribution; application of the Akaike information criterion. J Pharm Sci 99(5):2492–2501.