Journal of Pharmaceutical Sciences 110 (2021) 301-313

Contents lists available at ScienceDirect

Journal of Pharmaceutical Sciences

journal homepage: www.jpharmsci.org

Pharmacokinetics, Pharmacodynamics and Drug Transport and Metabolism

Predicting the Permeability of Macrocycles from Conformational Sampling — Limitations of Molecular Flexibility

Vasanthanathan Poongavanam [a], Yoseph Atilaw [a], Sofie Ye [a], Lianne H.E. Wieske [a], Mate Erdelyi [a], Giuseppe Ermondi [b], Giulia Caron [b, *], Jan Kihlberg [a, *]

[a] Department of Chemistry - BMC, Uppsala University, SE-75123 Uppsala, Sweden
[b] Department of Molecular Biotechnology and Health Sciences, University of Torino, Quarello 15, 10135 Torino, Italy

ARTICLE INFO

Article history:
Received 13 August 2020
Revised 23 October 2020
Accepted 26 October 2020
Available online 29 October 2020

Keywords:
Permeability
Machine learning
Nuclear magnetic resonance (NMR) spectroscopy
Quantitative structure-property relationship(s) (QSPR)
Membrane translocation
Macrocycle

ABSTRACT

Macrocycles constitute superior ligands for targets that have flat binding sites but often require long synthetic routes, emphasizing the need for property prediction prior to synthesis. We have investigated the scope and limitations of machine learning classification models and of regression models for predicting the cell permeability of a set of *de novo*-designed, drug-like macrocycles. 2D-Based classification models, which are fast to calculate, discriminated between macrocycles that had low-medium and high permeability and may be used as virtual filters in early drug discovery projects. Importantly, stereo- and regioisomer were correctly classified. QSPR studies of two small sets of comparator drugs suggested that use of 3D descriptors, calculated from biologically relevant conformations, would allow development of more precise regression models for late phase drug projects. However, a 3D permeability model could only be developed for a rigid series of macrocycles. Comparison of NMR based conformational analysis with in silico conformational sampling indicated that this shortcoming originates from the inability of the molecular mechanics force field to identify the relevant conformations for flexible macrocycles. We speculate that a Kier flexibility index of $\leq$10 constitutes a current upper limit for reasonably accurate 3D prediction of macrocycle cell permeability.

Introduction

Macrocycles are attracting major interest in efforts to "drug" targets which cannot be modulated by small molecules that comply with the Lipinski's rule of 5 (Ro5).[1,2] A comprehensive investigation of drugs and clinical candidates residing in beyond rule of 5 (bRo5) space highlighted that the unique ability of macrocycles to adopt disk- and sphere-like shapes makes them ideal for binding to targets that have flat and groove-shaped binding sites,[2] i.e. targets that are difficult to modulate with Ro5 compliant compounds. However, macrocyclic drugs often require long synthetic routes,[3] which emphasizes the importance of developing methods for prediction of properties such as cell permeability before initiating synthesis.

Passive cell permeability is a complex process which involves desolvation when the drug leaves the extracellular aqueous environment, followed by interactions with the negatively charged phospholipid head groups before it enters the hydrophobic membrane interior.[4] Then, this sequence of events is reversed as the drug enters the cytosol. Each of these steps is affected to different extents by the drug's molecular properties.[4] For instance, the polarity of the compound which can be described by its 3D polar surface area (PSA) is a major determinant of the kinetics of the desolvation step. The size of the compound, approximated by the radius of gyration (R_{gyr}), governs the rate of diffusion across the membrane, while the lipophilicity (cLogP or cLogD) is of major importance for the thermodynamics of the permeation process.

Quantitative structure-permeability relationship (QSPR) methods are widely used to model permeability in drug discovery. They rely on statistical relationships derived from experimental permeability data and physicochemical descriptors, such as the PSA, R_{gyr} and cLogP/D, calculated for a training set of compounds.[5] An alternate approach is to develop models that are more directly based on the physics of the underlying processes.[6–9] Physics-based models have provided an in-depth understanding of how small sets of macrocyclic hexa-, hepta- and decapeptides composed of neutral and lipophilic residues may cross cell membranes.[8–13] These models have been developed using substantial computational and experimental resources and are based

* Corresponding authors.
E-mail addresses: giulia.caron@unito.it (G. Caron), jan.kihlberg@kemi.uu.se (J. Kihlberg).

https://doi.org/10.1016/j.xphs.2020.10.052

on either a low-dielectric[8] or a congruent[11] conformation as the permeating species. Alternatively, ensemble-based solvent accessible surface area[13] and ensemble-based 3D polar surface area[14] have been found to be good predictors of the permeability of cyclic peptides and semipeptidic macrocycles, respectively.

Previously, we have investigated a set of more than 200 non-peptidic, *de novo*-designed macrocycles inspired by natural products to obtain knowledge about the cell permeability of drug-like macrocycles.[15] QSPR models of good predictive power were obtained, as well as information on how substructural features ranging from the numbers of hydrogen bond donors and acceptors to different functional groups and substituents affected permeability. However, the cell permeabilities of the members of stereo- and regioisomeric series, which sometimes differed by an order of magnitude, could not be predicted by the QSPR models which were based on 2D descriptors. Instead, manual scoring of the overall polarity, the degree of intramolecular hydrogen bonding and general steric shielding of polar groups in the ensembles of low-energy conformations of the isomers allowed a qualitative ranking of their permeability. However, such manual studies are time consuming and may not always be applicable for other regio- and stereoisomeric series. Consequently, they are less suitable for incorporation in an industrialized drug discovery process than QSPR models.

Herein we have first investigated the ability of methods based on machine learning to provide a rapid and accurate classification of macrocycles as either having low-medium or high permeability across Caco-2 cell monolayers. Such methods could be of significant value for rapid decision making in design of macrocycles in the early phases of drug discovery projects. Then we have investigated the scope and limitations in prediction of cell permeability for stereo- and regioisomers using 3D descriptors calculated for conformers obtained by sampling using the distance-geometry based software OMEGA. Such an approach (3D-QSPR) would be of use in lead optimization of series of macrocycles; in a previous work[16] we showed that models based on the 3D PSA of experimentally determined conformers performed better than those based on the 2D descriptor TPSA. We based the current investigation on a subset of the 200 non-peptidic, *de novo*-designed macrocycles that only displayed passive cell permeability. Two sets of drugs were used as comparator sets, one being a set of non-macrocyclic (linear) drugs, most of which comply with the Ro5, while the other one consists of both macrocycles and non-macrocycles residing in bRo5 space.

Materials and Methods

General Description of Compound Sets

We based this study on three sets of compounds, *i*) a set of *de novo*-designed macrocycles inspired by natural products that has been prepared by diversity oriented synthesis (the DOS macrocycle set),[15] *ii*) a set of approved, non-macrocyclic drugs (the linear drug set)[17] and *iii)* a set of approved drugs all of which reside in the beyond rule of 5 chemical space (the bRo5 drug set).[16]

The ***DOS macrocycle set*** consists of 70 of the just over 200 macrocycles for which the permeability across Caco-2 cell monolayers was determined earlier under consistent conditions.[15] As discussed in the Introduction, models for the cell permeability of the full set of >200 macrocycles have been reported, as well as in-depth studies of a few of its series.[15] Herein we selected only those macrocycles that did not display any residual efflux (efflux ratio <2) in the presence of a cocktail of inhibitors of the three major efflux transporters, i.e. a set of macrocycles associated with passive cell permeability. Their permeability was fairly evenly distributed over close to two orders of magnitude, (Supplementary Fig. 1, Supplementary Table 1). The majority, i.e. 52 of the 70 macrocycles, can be grouped in seven series with compounds in each series differing only in their stereo- and regiochemistry (Fig. 1a, series A-G; Supplementary Table 1). The remaining 18 macrocycles are singletons (called the "Unique" series (U), Supplementary Table 1) and display large structural variation.

The ***linear drug set*** includes 79 non-macrocyclic compounds obtained from a previous study (Fig. 1b).[17] From the original dataset the following compounds were excluded: adefovir, acarbose and

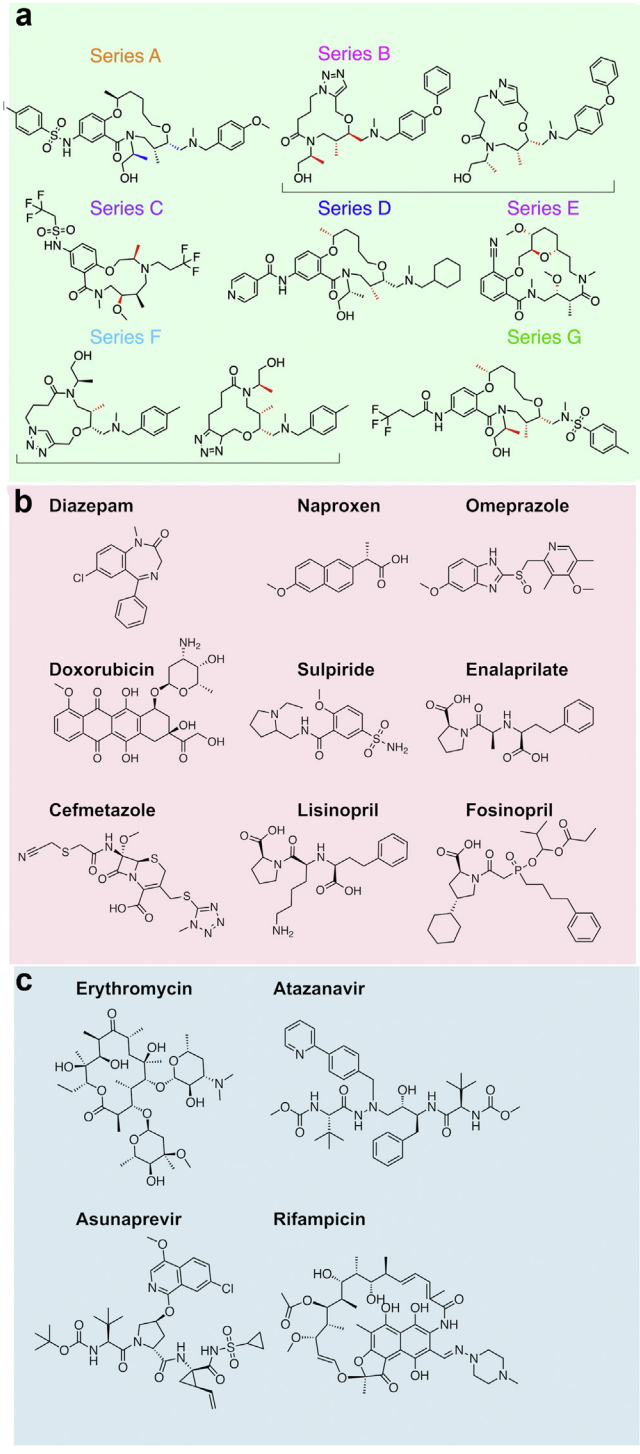


Fig. 1. Chemical structures of (a) members of each of the seven series that make up the majority of the DOS macrocycle set, (b) selected members of the linear drug set for which the number of rotatable bonds increases from 1 to 15, and (c) a member of each of the four classes of drugs in the training set of the bRo5 drug set.

digoxin in agreement with the original report,[17] methotrexate since it is involved in active transport mechanisms,[18,19] and azithromycin and erythromycin since they are macrocycles. Caco-2 cell permeability was determined under consistent conditions and the data is distributed over close to three orders of magnitude (Supplementary Fig. 1, Supplementary Table 2).[16] The final dataset was split in two subsets: non-flexible (n = 49, called DS1) and flexible (n = 30, named DS2) according to the number of rotatable bonds (NRotB <6 for DS1, ≥6 for DS2). Herein, it was used as a comparator compound set to investigate if the impact of flexibility on modelling of cell permeability of non-macrocyclic drugs is like that of macrocycles.

The ***bRo5 drug set*** includes 18 macrocyclic and non-macrocyclic drugs in bRo5 space that were extensively investigated in a previous study (Fig. 1c, Supplementary Table 3).[16] The efflux inhibited Caco-2 cell permeability correlated strongly ($r^2 = 0.90$) with the minimum solvent-accessible 3D polar surface areas (Min SA 3D PSA) calculated from the crystal structure conformations of the ten drugs in the training set of the bRo5 drug set. In addition, the correlation between the permeability and Min SA 3D PSA of the drugs in the training set predicted the permeabilities of the eight drugs in the test set well (RMSE = 0.71). We used this as a reference set to compare models of cell permeability based on experimental knowledge of the compounds' 3D conformations to models obtained using conformational sampling.

Overall the DOS macrocycle set is the main compound set investigated in this paper, the linear drug set is a comparator set of non-macrocyclic approved drugs, most of which comply with the Ro5, while the bRo5 comparator set consists of both macrocycles and non-macrocyclic drugs.

Characterization of the three compound sets using the four descriptors of Lipinski's rule of 5[20] (molecular weight, MW; the number hydrogen bond acceptors and donors, HBA and HBD; and the calculated lipophilicity, cLogP) and the descriptors of Veber's rule[21] (the topological polar surface area, TPSA, and the number of rotatable bonds, NRotB) reveals differences and similarities between the three sets (Fig. 2). The linear drug set has the lowest MW distribution, while the DOS macrocycles and the bRo5 drug set have increasingly higher MW distributions (Fig. 2a). The HBA and TPSA distributions for the three compound sets reflects their MW distributions, i.e. compounds having higher MWs had higher numbers of HBAs and greater values for TPSA (Fig. 2c and e). Apart for some compounds in the bRo5 drug set most compounds in the three sets had values below the upper cut-offs of the Ro5 and Veber's rule (HBA ≤10 and TPSA <140). With very few exceptions values for cLogP and HBD fall below their upper Ro5 cut-offs (≤5) for the three compounds sets (Fig. 2b and d). However, the bRo5 drug set has these two descriptors shifted towards somewhat higher values, while the DOS macrocycle set has fewer HBDs than the other two sets. NRotB falls below or at 10 (the upper limit in Veber's rules) for most compounds in the three sets, but with the DOS macrocycle set and the bRo5 drug sets having compounds just outside or quite far outside this upper limit, respectively (Fig. 2f).

3D Structures and Conformational Sampling

The Simplified Molecular-Input Line-Entry System (SMILES) codes of all the compounds were obtained and converted into 3D structure using CORINA (version 3.2) from Molecular Networks GmbH.[22,23] Special care was taken while converting the SMILES to the 3D structures to make sure that the stereochemistry for all compounds was correctly annotated according to the original source.[15] Uncharged (i.e. protonated acids and deprotonated bases) and charged states (at pH 7.4) were generated for the compounds using the Wash tool from MOE software (Molecular Operating Environment, version 2015.10).[24] Both uncharged and charged structures were submitted to conformational sampling using the OMEGA tool from OpenEye,[25] and a minimum energy conformation was calculated using CORINA. Both sets of conformations from OMEGA were generated in chloroform ($\varepsilon = 4.8$, Sheffield solvation model[26]) to mimic the apolar portion of cell membranes. The conformational sampling methodology and its implementation have been described elsewhere.[25,27,28] A RMSD-based clustering procedure implemented in the Diverse Subset tool from the MOE suite was then applied to reduce the number of conformers arising from conformational sampling. Partial charges were calculated on the final conformers using the PM3 semi-empirical method implemented in Spartan (v1.1.4).[29]

Molecular Descriptors

A set of 293 2D and 3D-molecular descriptors[30] were computed for the conformation obtained by CORINA, and on the conformers from conformational sampling with OMEGA. Computed descriptors include atom and bond counts, adjacency, distance matrix descriptors, Kier and Hall connectivity, kappa shape indices, pharmacophore feature descriptors, partial charge descriptors, surface area, volume, shape descriptors and MOE-based vsurf descriptors.[30,31] Some additional descriptors were also calculated. The lipophilicity distribution coefficient (logD at pH = 7.4) was computed using the MarvinView (v18.15.0) tool. The virtual logP arising from a molecular lipophilicity potential (logP (MLP)) was obtained with VEGA ZZ.[32,33] The three-dimensional solvent accessible polar surface area (SA 3D PSA) was calculated with PyMOL v1.7.4 as described before.[16]

Conformer-Dependent Descriptor Sets

Seven sets of descriptors were generated for the DOS macrocycle and the linear drug set. The first two were based on the 2D-structure (named "2D") and the single minimum energy conformer from CORINA (named "3D"). The remaining five descriptor sets were calculated on selected conformers arising from the conformational ensembles from OMEGA, i.e. on the minimum energy conformer (named "MEC"), the conformer with the lowest solvent accessible 3D PSA (named "MinPSA"), the conformer having the median PSA (named "MedPSA"), the conformer with the median radius of gyration (named "MedRgyr") and a virtual conformer based on the median value for all descriptors (named "Median"). These were chosen as we recently found that selection of conformations based on their polar surface area or radius of gyration provided a better approximation of the biologically relevant conformations than an energy-based selection.[28] The overall workflow of conformational search and descriptor calculation is shown in Fig. 3.

Training and Test Sets

The DOS macrocycle set was divided into training and test sets by applying a multivariate analysis based on a principal component analysis (PCA) followed by D-optimal onion design (DOOD).[34] Briefly, DOOD is a multivariate method for selecting representative compounds from a chemical space defined by the molecular properties and is based on a score vector obtained from principal component analysis (PCA). According to the score vector, the dataset is split into different layers and from each onion-like layer training compounds are chosen. In this study, the score was calculated with SIMCA (version 10.5, Umetrics) and the D-optimal onion design experiment was performed using the MODDE (version 7.0, Umetrics). The list of compounds used in the training (47) and test (23) set is provided in the Supplementary Table 1. The

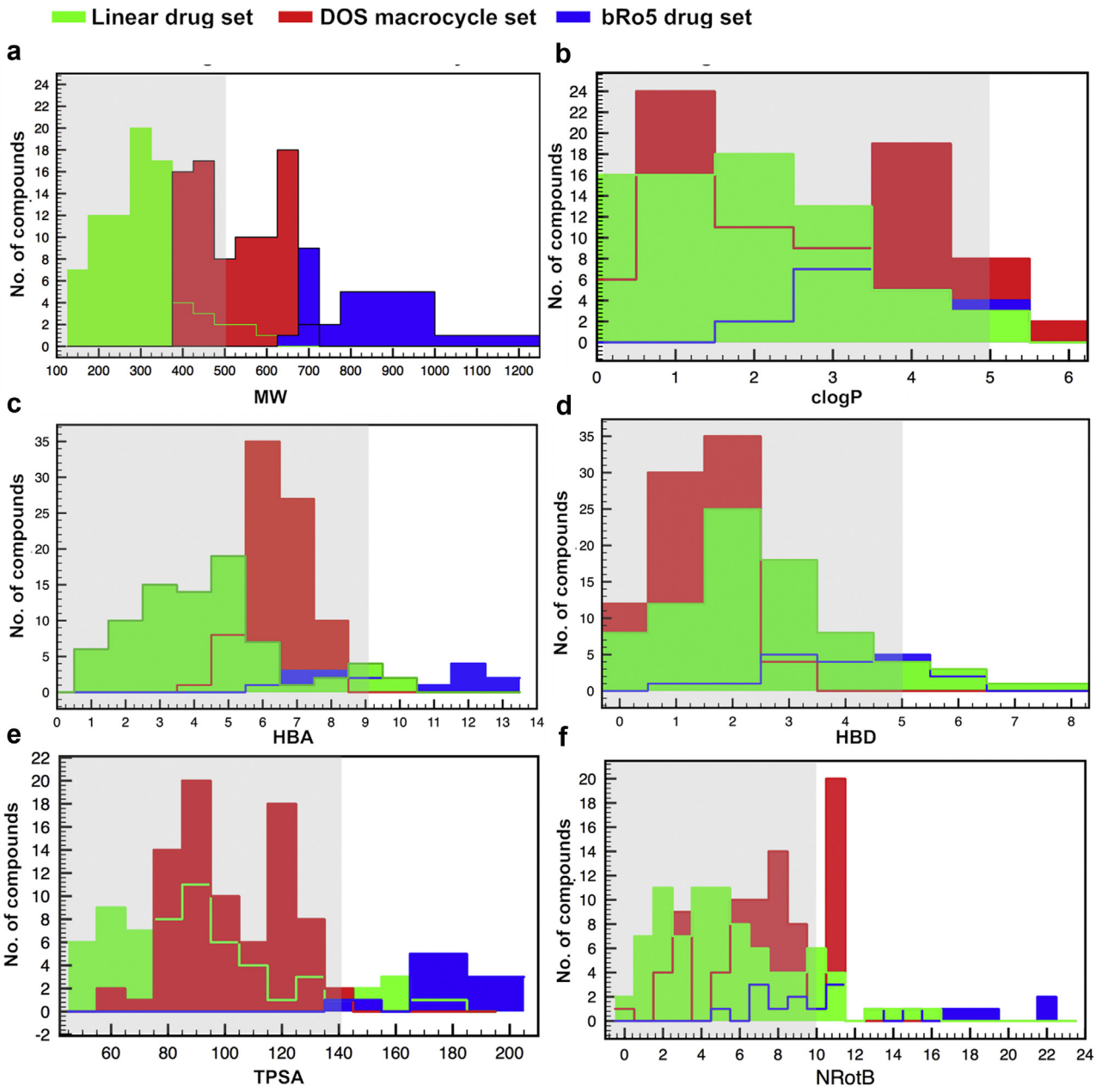


Fig. 2. Distribution of (a) molecular weight (MW), (b) calculated lipophilicity (cLogP), (c and d) the number of hydrogen bond acceptors and donors (HBA and HBD), (e) the topological polar surface area (TPSA), and (f) the number of rotatable bonds (NRotB) for the members of the three sets of compounds investigated. Calculated descriptors are plotted in green for the linear drug set, in red for the DOS macrocycle set and in blue for the bRo5 drug set. Values that adhere to the Lipinski's rule of 5[20] and Veber's rule[21] are marked with grey shading.

linear drug set was also split into training and test set using the "Diverse Subset" tool provided in the MOE suite.

Classification Models

Random Forest (RF) as implemented in the WEKA v3.8 data mining tool was used for the binary classification.[35] The random forest model was built with 10 trees and 1 seed (default setting), no additional parameters were set as tuning of other parameters did not improve the quality of the model. Method details have been described elsewhere.[36,37] To evaluate the quality of the classification, the Matthews Correlation Coefficient (MCC)[38] was used. It takes into account true positives and negatives and returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction, and −1 the worst possible prediction. In general, MCC values greater than 0.4 are considered to be predictive,[39]

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

where TP are true positives, TN are true negatives, FP are false positives, and FN are false negatives. In addition, each class (positive or negative) is assessed with specificity (or true negative rate, specificity = TN/(TN + FP)) and sensitivity (true positive rate, sensitivity = TP/(TP + FN)), defined by the proportion of correctly

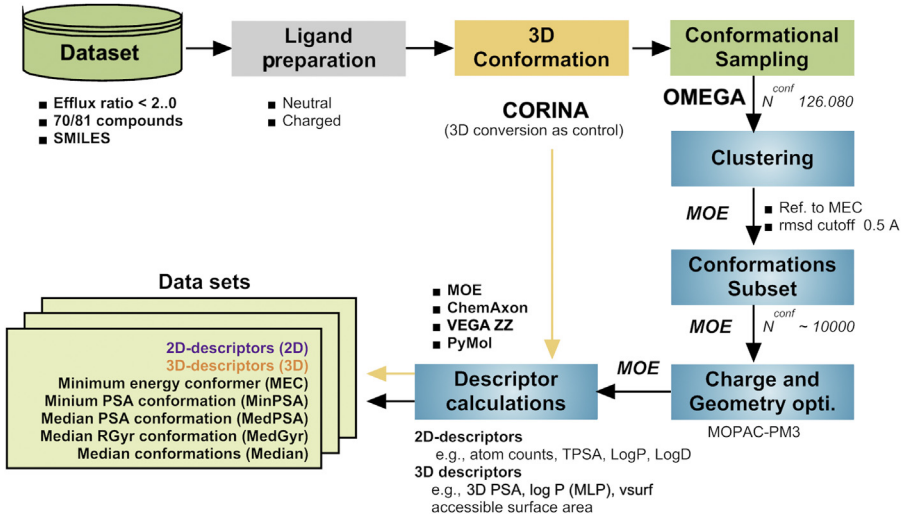


Fig. 3. Workflow to generate the data matrix.

classified negative and correctly classified positive class by five-fold cross validation or test set validation, respectively. For each dataset, suitable descriptors were chosen based on the automatic variable selection procedure (CfsSubsetEval-BestFirst) as implemented in the Weka software.[35,40] CfsSubsetEval combined with the BestFirst algorithm has been shown to be a better attribution selection method as compared to others.[37]

Regression Models

Multiple linear regression (MLR) models and statistics were obtained with QSARINS v.2.2.2 (www.qsar.it).[41] To select variables, we used the genetic algorithm (GA) tool implemented in the software using the following parameters: descriptors limit: 5, Pop size: 50, 5000 generation/size (iteration), mutational rate 50. All models were obtained after data normalization. For any model we provide: correlation coefficient (R^2), adjusted correlation coefficient (R^{2adj}), s (standard error of estimate), F (Fisher value), RMSE (root mean square error), MAE (Mean Absolute Error), PRESS (Predictive Residual Sum of Squares) and Q^2 (Explained variance in prediction, Leave- One-Out cross validation). Definitions for each term are described elsewhere.[42,43]

NMR Spectroscopy

The NMR spectra of **G16** ($CDCl_3$ and DMSO-d_6) and **E2-enant** (DMSO-d_6) were recorded at 25 °C on an 800 MHz BRUKER Avance III HD NMR spectrometer equipped with a 5 mm TCI cryogenic probe, while the spectra of **E2-enant** in $CDCl_3$ and D_2O were recorded on a 600 MHz Bruker Avance Neo NMR spectrometer with a 5 mm TCI cryogenic probe. The assignments were deduced using HSQC, HMBC, NOESY, TOCSY and COSY experiments. A single set of sharp peaks was observed for both macrocycles, revealing that they populate conformational ensembles in which individual conformations are separated by low energy barriers (<1 kJ/mol). The structure of **G16** is found in Supplementary Table 1, while **E2-enant** is the enantiomer of E2 (Supplementary Table 1). **E2-enant** was used because of lack of material for the four macrocycles in series E.

NOESY buildups were acquired with 7 mixing times (100, 200, 300, 400, 500, 600 and 700 ms), without solvent suppression, with the relaxation delay set to 2.5 s, and using 16 transients with 512 and 2048 points collected for the f1 and f2 dimensions, respectively. The NOE peak intensities were calculated according to ([cross peak1 × cross peak2]/[diagonal peak1 × diagonal peak2]).$^{0.5}$ For the calculation of the initial buildup rates (σ_{ij}) a minimum of 4 mixing times giving a linear ($R^2 \geq 0.95$) build-up were used. Interproton distances (r_{ij}) were calculated according to the equation $r_{ij} = r_{ref}(\sigma_{ref}/\sigma_{ij})^{(1/6)}$ using geminal methylene protons (1.78 Å) as internal distance reference.

NAMFIS Analysis

Conformation ensembles of **G16** and **E2-enant** were generated using the Monte Carlo conformational search algorithm with intermediate torsion sampling, 50 000 Monte Carlo steps followed by molecular mechanics energy minimization with Macromodel (v12.1), with the RMSD cut-off set to 2.0 Å, as implemented in the Schrödinger Suite. For each energy minimization, the Polak-Ribière type conjugate gradient (PRCG) with a maximum of 5000 iterative steps was used. Conformations within 42 kJ/mol from the global minimum were kept. Conformational searches were done using the five force fields OPLS, OPLS-2005, OPLS3e, AMBER* and MMFF, each with the GB/SA implicit solvation model which represent apolar ($\varepsilon = 4.8$) and polar ($\varepsilon = 80.0$) environments, respectively. Subsequently, the ensembles from the conformational searches using different force fields were combined, and redundant conformations were eliminated (non-hydrogen atom RMSD cutoff set to 2.0 Å (**G16**) and 1.0 Å (**E2-enant**)).

Solution ensembles were determined by fitting the experimentally measured distances and coupling constants to those back-calculated from computationally predicted conformations using the NAMFIS algorithm.[44,45] Distances to methylene protons were treated according to the equation $d = (((d_1^{-6}) + (d_2^{-6}))/2)^{-1/6}$, and to methyl protons according to $d = (((d_1^{-6}) + (d_2^{-6}) + (d_3^{-6}))/3)^{-1/6}$.

The output solution ensembles were validated using standard methods, that is through evaluation of the reliability of the conformational restraints by the addition of 10% random noise to the experimental data, by the systematic removal of individual restraints, and by comparison of the experimentally observed and back-calculated distances.[45]

Results and Discussion

Classification Models for Cell Permeability

Machine learning (ML) methods are finding increasing use due to their efficiency in uncovering patterns in complex and

multifactorial datasets for which the underlying mechanistic understanding may be poor. We built random forest (RF)-based classification models for the cell permeability of the DOS macrocycle set and used the linear drug set as a comparator set.[37,46] The threshold value for distinguishing compounds having low-medium permeability across Caco-2 cell monolayers from those having high permeability was set at 10×10^{-6} cm/s as this threshold has been used in industrial drug discovery projects.[4]

For the ***DOS macrocycle set*** excellent permeability models were obtained based only on 2D descriptors and their quality improved when the macrocycles were treated as uncharged (>90% accuracy, MCC = 0.80, 5-fold CV) instead of as charged species (Fig. 4a, Supplementary Table 4). When considering a balanced prediction of highly permeable and low-medium permeable compounds, 2D-descriptors for the uncharged form again performed better than those of the charged form. The overall accuracy of high versus low-medium permeable was found to be 94 and 87%, respectively, for the uncharged form as compared to 73 and 33% for the charged form. Test set prediction confirmed the statistical quality of the model (n = 23, MCC = 0.73) (Supplementary Table 5). It is important to note that the model based on 2D descriptors for the uncharged macrocycles succeeds in the successful classification of the stereo- and regioisomeric macrocycles in series D-G, with only one exception (Fig. 4b). The macrocycles in these series have permeabilities ranging from low-medium to high, and the single misclassified macrocycle has a permeability very close to the threshold value. The macrocycles in series A-C, that all have high permeabilities were also correctly predicted, just as all macrocycles in the unique (U) series which have low-medium as well as high permeabilities.

Classification models were also built using descriptors calculated from the different 3D conformations (3D, MEC, MinPSA, MedPSA, MedGyr and Median) for uncharged and charged macrocycles in the same manner as for the 2D-models (Supporting Information, Supplementary Tables 4 and 5). Perhaps surprisingly, the use of 3D descriptors decreased the statistical quality of the models (Fig. 4a). Although >70% of the compounds were correctly predicted by most models, 3D descriptor-based models suffer from a high false-positive rate. It is possible that the poor performance of the 3D models is due to that the DOS macrocycle set is composed of a large number of stereo- and regioisomers and/or that the selected conformations were not relevant for permeability. The final 2D (and 3D) models were mainly governed by lipophilicity and polar surface area (Supplementary Table 6).

For the ***linear drug set***[17] we focused on the four descriptor sets that provided good or acceptable models for the DOS macrocycle set (2D, 3D, MedPSA, and MedGyr), as well as the MEC set, and built models for the drugs in their uncharged form. In total 79 compounds (55 high permeable and 24 low permeable) were used for development of classification models. The best model was again based on 2D descriptors, and classified more than 78% and 88% of the drugs in the training and test sets, respectively, correctly (MCC = 0.47 and 0.69, respectively, 5-fold cross-validation) (Fig. 4c, Supplementary Table 7). In this case, 3D-descriptor based models performed almost as well as the model based on 2D descriptors, i.e. with an overall accuracy of 71–76% (67–79% for the test set) and a MCC of 0.30–0.42 (0.39–0.50 for the test set) (Supplementary Table 7). The improved performance of the 3D models, as compared to the DOS macrocycle set, may be due to that the linear drug set contains fewer stereocenters and/or that conformational sampling is more successful than for the DOS macrocycle set.

Overall, the results for these two sets of compounds suggest that machine learning-based classification models can be used as a virtual permeability filter to distinguish low and high permeable compounds at the stage of design in early drug discovery projects. It

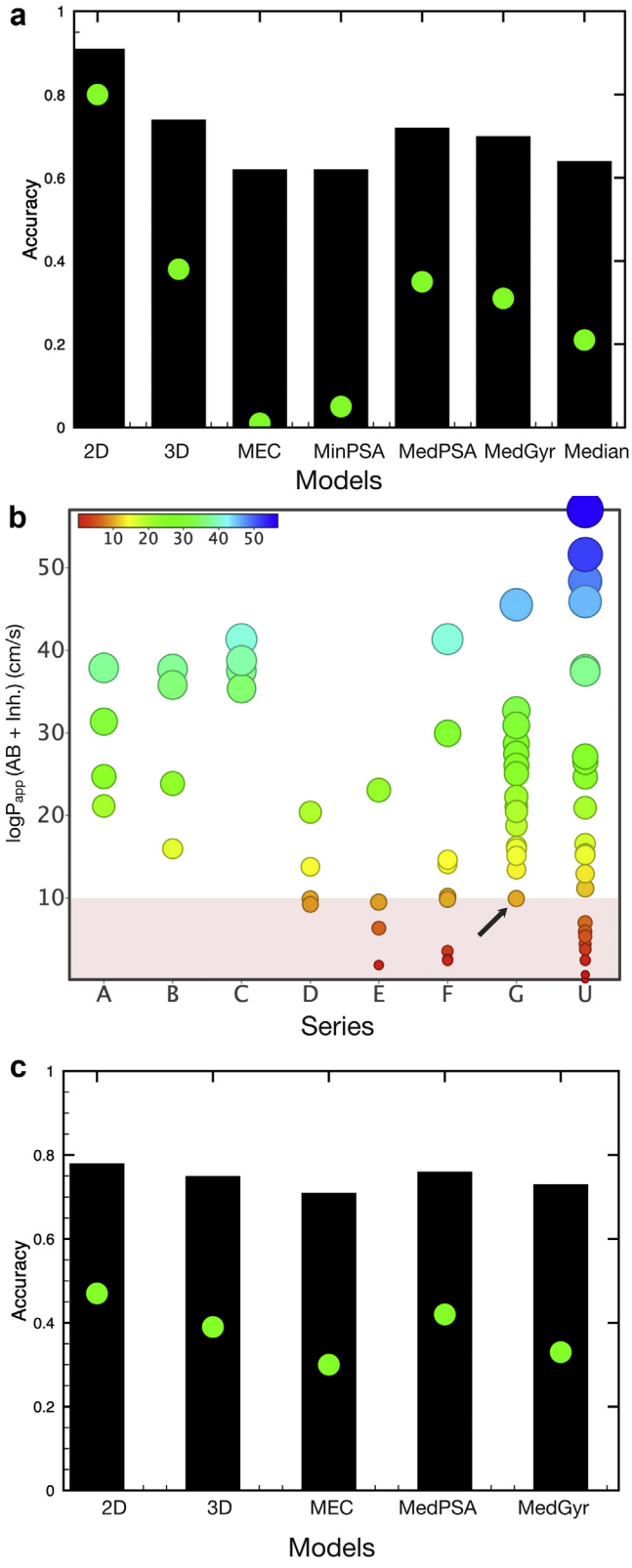


Fig. 4. Random-Forest classification of high and low-medium permeable compounds for the uncharged form of the compounds in (a) the DOS macrocycle set and (c) the linear drug set. The black bars indicate the accuracy with which the permeability of the compounds in each of the two sets is predicted, as determined by 5-fold cross validation. The MCC value (range −1 to +1) of each model is indicated by the green circles. (b) Plot of permeabilities for the members of the DOS macrocycle set. Macrocycles having low-medium permeabilities are denoted with smaller, red-orange circles, those having high permeabilities have larger, yellow-green-blue circles. The misclassified macrocycle is indicated with an arrow.

is important to note that models based on 2D-descriptors, which are faster to calculate, performed better than the more time-consuming 3D models for both sets of compounds. Interestingly, the difference in performance between the 2D and 3D models was greater for the DOS macrocycle set than for the linear drug set. Potentially, this is caused by that it is more difficult to sample biologically relevant conformational space for macrocycles than for linear compounds.

Experimental 3D Structural Information Improves Cell Permeability Models

We have reported that excellent models for efflux inhibited permeability across Caco-2 cells were obtained for the bRo5 drug set when knowledge about conformational preferences and 3D structural information was incorporated in building of the models.[16] Thus, cell permeability was highly correlated to the minimum solvent accessible 3D PSA (Min SA 3D PSA) calculated from multiple crystal structures of each drug (Fig. 5a, black dots, $r^2 = 0.90$), whereas the correlation to TPSA was poor ($r^2 = 0.36$).[16] Ideally, cell permeability should be modelled prior to synthesis, but identification of the permeating conformation(s) by conformational sampling is often non-trivial for macrocycles and drugs in bRo5 space.[28] We therefore used the bRo5 drug set to assess if calculated conformations can be used for prediction of cell permeability, and how accurate such models may be. The correlation between cell permeability and the Min SA 3D PSA of the conformations obtained by conformational sampling using OMEGA was weaker than when crystal structure conformations were used (Fig. 5a, red dots, $r^2 = 0.52$), but still stronger than the correlation with TPSA ($r^2 = 0.36$).

Inspired by the finding that a better model for cell permeability was obtained based on the Min SA 3D PSA of sampled conformations for the bRo5 drug set than on their TPSA, we investigated the importance of incorporating 3D structural information in modelling permeability of the linear drug set. For this set $\geq$2 crystal structures had been reported in the CSD and PDB for 12 of the drugs (Supplementary Tables 8 and 9). Also for this set of drugs better models for cell permeability were obtained based on the Min SA 3D PSA of the experimentally determined conformations ($r^2 = 0.53$), or sampled conformations ($r^2 = 0.43$), than when TPSA was used ($r^2 = 0.10$) (Fig. 5b and Supplementary Fig. 2). However, the quality of the model based on experimentally determined conformations is significantly lower than that of the bRo5 drug set.

Regression Models of Cell Permeability for the DOS Macrocycle Set

Since accurate predictions of permeability are desired in the lead optimization process, QSPR strategies were applied to the DOS macrocycle set. Most compounds in this set belong to series of stereo- and regioisomers and it is therefore attractive in efforts to unravel how incorporation of 3D structural information impacts cell permeability modelling. We investigated both whether global regression models of cell permeability could be determined for this compound set, as well as if models for individual series could be obtained.

Global Models

Since polarity and lipophilicity have been shown to be strongly correlated to permeability,[4] a relationship between permeability and calculated polar surface area or lipophilicity was first looked for (Supplementary Table 10). However, no significant correlations were found for the neutral form of the macrocycles independent of if 2D or 3D versions of the two descriptors were used ($r^2 = 0.04$; 0.03). Then MLR models for cell permeability were built using QSARINS for the seven sets of descriptors described in the Materials and Methods section, i.e. 2D, 3D, MEC, MinPSA, MedPSA, MedGyr and Median. The best model was obtained for the uncharged form of the macrocycles using the Median set, but the test set validation failed (RMSE = 0.35). The other six models were of slightly lower quality, and attempts to build models for the charged forms of the macrocycles did not provide any improvement (Supplementary Table 11).

Modelling Series

The difficulties to obtain a global permeability model for the entire DOS macrocycle set could e.g. originate from that flexibility and/or protonation state varies between the series. To investigate the influence of variation between series correlations between cell permeability and polar surface area or lipophilicity were investigated for the uncharged form of the compounds in the seven series of macrocycles (A-G, cf. structures in Fig. 1). Strong correlations ($r^2 > 0.7$), having the expected negative slope, were found between the cell permeability of the four macrocycles of series E and the SA

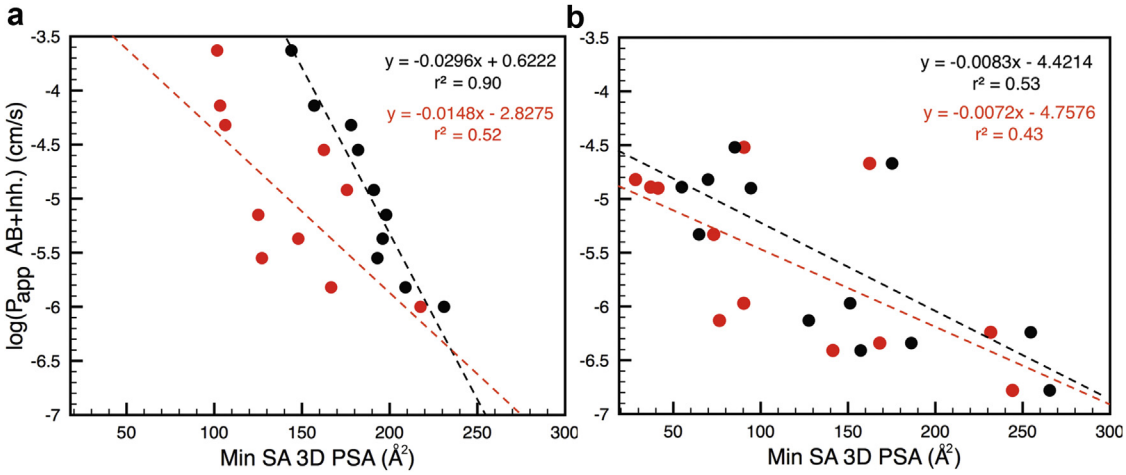


Fig. 5. Correlation between efflux inhibited cell permeability across Caco-2 cell monolayers [log (P_{app} AB-Inh)] and the minimum solvent accessible 3D PSA (Min SA 3D PSA) calculated for (a) ten compounds from the bRo5 drug dataset[16] and (b) twelve compounds from the linear drug dataset. The compounds included in the correlations had at least two crystal structures deposited in the PDB or CSD that differed by an RMSD $\geq$0.75 Å. Min SA 3D PSAs were calculated from the crystal structures of the compounds for the correlations in black, while the correlations in red are the Min SA 3D PSA conformations calculated by OMEGA.

3D PSA of their MinPSA, MedPSA and MedGyr conformations (Fig. 6). In addition, permeability correlated positively with the calculated $LogP_{(o/w)}$ and the LogP (MLP) of the MedGyr conformation for this series. No, or poor, correlations were found for the other six series (Supplementary Table 10), with difficulties being particularly evident for series G (Fig. 6).

NMR Derived Solution Ensembles of DOS Macrocycles

Based on the above analysis of the individual stereo- and regiosomeric series in the DOS macrocycle set we hypothesized that the difficulties in modelling their cell permeability originated from that the flexibility of these series prevents the prediction of the relevant

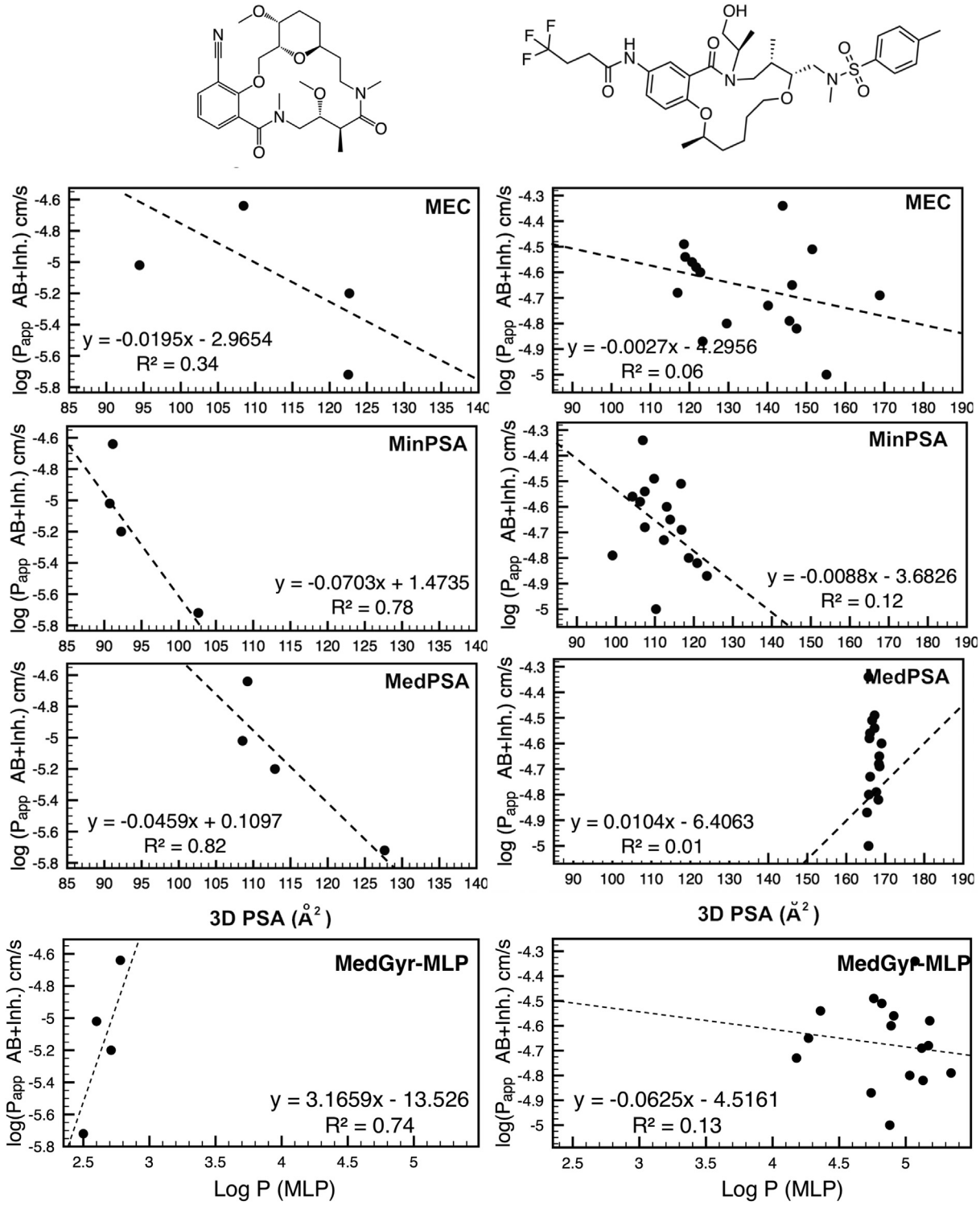
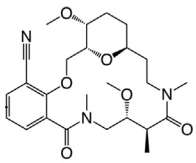
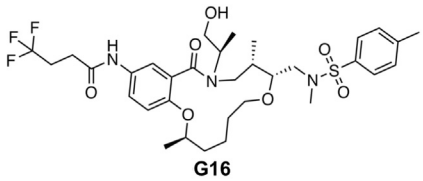


Fig. 6. Correlations between the cell permeability of the compounds in series E (left panels) and G (right panels) of the DOS macrocycle set and SA 3D PSA calculated for the MEC, MinPSA and MedPSA conformations of these macrocycle (rows 1–3). The correlation between permeability and the lipophilicity (log P (MLP)) for the two series is also shown (bottom row). Examples of structures from each series are shown at the top.

Table 1
Conformational Ensembles of Macrocycles **E2-enant** and **G16** in Apolar and Polar Solutions as Determined by NAMFIS Analysis.[a]





Cpd	$CDCl_3$		DMSO-d_6		D_2O	
	Conf. No	Molar Fraction (%)[b]	Conf. No	Molar Fraction (%)[b]	Conf. No	Molar Fraction (%)[b]
E2-enant	*1*	*4*	*1*	*6*	**3**	**81**
	2	*11*	*2*	*3*	7	2
	3	**81**	**3**	**76**	8	3
	4	3	5	11	9	4
			6	4	10	7
					11	3
G16	1	3	*3*	*22*		
	2	11	6	3		
	3	*34*	7	6		
	4	49	8	4		
	5	4	9	7		
			10	27		
			11	9		
			12	6		
			13	3		
			14	6		
			15	3		

[a] Conformations populated in more than one solvent are marked with italics and bolded values.
[b] Population in % of the indicated solution conformer. Conformers having populations ≤1% have been discarded.

conformations. Inspection of the structures of the macrocycles in series E and G, i.e. the series displaying the best and worst correlations to permeability (Fig. 6), suggest that series G is much more flexible because of its three side chains and probably also due to a more flexible macrocyclic ring. The difference in flexibility is supported both by the number of rotatable bonds (NRotB) and by the Kier flexibility index[47] (Φ) of the macrocycles in these two series (Supplementary Fig. 3). To get experimental insight into the flexibility of the macrocycles in series E and G, we determined the solution conformational ensembles for one macrocycle from each series, **E2-enant** and **G16** (Table 1, Supplementary Table 1), by deconvolution of time-averaged NMR data using the NAMFIS algorithm (cf. Supplementary Tables 12-18 for **E2-enant** and Supplementary Tables 19-24 for **G16**).[44] The enantiomer of macrocycle E2 was used as insufficient amounts of material was available for the four compounds in series E. The NAMFIS method was chosen as it has previously been successfully applied for the description of the solution ensembles of diverse sets of macrocyles,[48–50] some of them similar to **G16** and **E2-enant** in size and flexibility. As the conformations responsible for cell permeability are expected to be found among those in the solution ensembles,[48] the solution ensembles of **E2-enant** and **G16** were then compared to the sampled ensembles used for prediction of cell permeability.

Description of Solution Ensembles

Macrocycle **E2-enant** from series E populated from four to six conformations in the increasingly polar solvents $CDCl_3$, DMSO-d_6 and D_2O (Table 1). Chloroform ($\varepsilon = 4.8$) was used to mimic the cell membrane ($\varepsilon = 3.0$),[4] while DMSO and water mimicked the plasma and cytosol. Conformation number 3 is the major one and represents approximately 80% of the ensemble in each solvent, while each of the minor ones represent ≤11% of the ensembles. The pairwise RMSD values between the most different conformations ranged from 2.5 to 3.2 Å in the three solvents (Supplementary Table 25), confirming a low flexibility for **E2-enant**. In contrast, macrocycle **G16** from series G is more flexible. In DMSO-d_6 G16 populates two major (number 3 and 10) and nine minor conformations, with a pairwise RMSD value of 5.26 Å between the most different conformations (Supplementary Table 26). G16 populates two major (number 3 and 4) and three minor conformations in $CDCl_3$, and the RMSD value between the most different conformations was 4.95 Å. G16 had a too low solubility to allow determination of its solution ensemble in D_2O.

Structural Comparison of Solution and Sampled Ensembles

The experimentally determined solution ensembles for **E2-enant** and **G16** were compared to ensembles obtained by conformational sampling using OMEGA within a 25 kcal/mol energy window in polar ($\varepsilon = 80$) and apolar ($\varepsilon = 4.8$) implicit solvents (Fig. 7). For **E2-enant** the predicted minimum energy conformation (MEC) in each of the two implicit solvents was similar (RMSD ≤2 Å)[28,51] to the experimentally determined conformations in $CDCl_3$ and DMSO-d_6, respectively (Fig. 7a), indicating that OMEGA predicts relevant solution conformers for this macrocycle. An increasing number of conformations sampled at higher energies were also similar to the conformations in the experimental ensembles. For the more flexible **G16** only one of the minor conformations in the DMSO ensemble (number 12, 6%) was similar to the predicted MEC in an implicit polar solvent (RMSD ≤2 Å, Fig. 7b). In $CDCl_3$ a predicted conformation similar to minor conformation 2 (11%) was found 1 kcal/mol above the MEC. Predicted conformations similar to the other solution conformations of **G16** were found only at energies ≥5 kcal/mol above the MEC, revealing the difficulties in predicting conformations for this flexible macrocycle. In fact, conformation 4 in $CDCl_3$ and 3 in DMSO, i.e. one of the two major conformations in each solution, was sampled only at

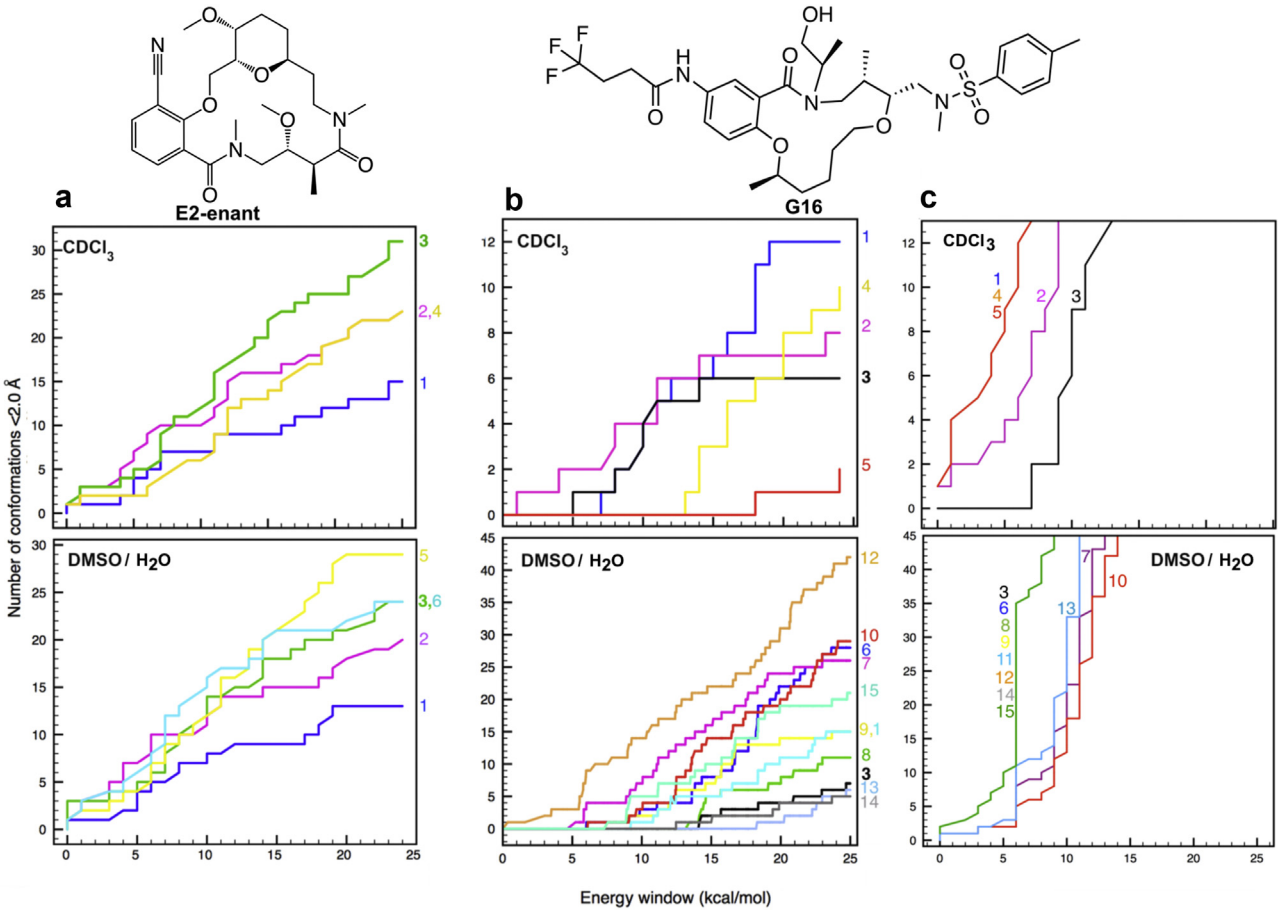


Fig. 7. Enrichment of solution conformations of **E2-enant** (a) and **G16** (b and c) in $CDCl_3$ and DMSO-d_6 in conformational ensembles generated with OMEGA in apolar ($\varepsilon = 4.8$, top panels) and polar ($\varepsilon = 80$, bottom panels) implicit solvents. For **G16** panel b refers to the overall macrocycle, while panel c refers to the macrocycle core, defined as the heavy atoms in the macocycle ring and the first attached heavy atom of the substituents. An RMSD value of ≤2.0 Å was used as a cut-off for similarity. The numbers of the experimental conformations are indicated on the right Y-axis.

energies 13−14 kcal/mol above the MEC. The conformations of the core of **G16** was predicted better by conformational sampling; the MECs resembled the cores in all conformations but that of number 3 in $CDCl_3$ (Fig. 7c, Supplementary Table 27). This highlights that the cores are more rigid, and that the side-chains of G16 are the main sources of flexibility and thereby of the difficulties to sample the biologically relevant conformations.

Polar Surface Area of Solution and Sampled Ensembles

The solvent accessible 3D polar surface area (SA 3D PDA) is a key determinant of cell permeability.[16] It showed only a small variation between the solution conformations of the rigid **E2-enant** (<25 $Å^2$ between conformations, Fig. 8a). PSA differences were also small between the ensembles sampled with OMEGA in apolar and polar environments. In addition, the SA 3D PSA of the sampled and solution ensembles overlapped well; in agreement with that SA 3D PSA could be used to develop good models for the permeability of series E. The SA 3D PSA of the flexible **G16** showed a significantly larger variation between conformations, reaching a difference of >70 $Å^2$. As for **E2-enant** the SA 3D PSA of the sampled ensembles differed little between an apolar and polar environment. However, for G16 the SA 3D PSA of the second most populated conformation in each solution ensemble was far outside of 25−75 percentiles and close to the minimum of the sampled ensembles. This illustrates the difficulties of conformational sampling in reproducing the properties of the solution ensembles of flexible compounds such as those in series G of the DOS macrocycle set.

Overall the NMR studies confirm that the macrocycles in series E are significantly less flexible than those of series G. The study also supports that the biologically relevant conformations are better reproduced by conformational sampling for the rigid macrocycles than for the more flexible ones. We therefore conclude that the difficulties in modelling the permeability for the DOS macrocycle set, and all but one of its series, originate from the failure of conformational sampling to identify their biologically relevant conformations.

Regression Models of Cell Permeability for the Linear Drug Set

We investigated if flexibility is a limiting factor for modelling of permeability also for non-macrocyclic compounds using the linear drug set. To this aim, the compounds were split in two classes according to the their flexibility as estimated by the number of rotatable bonds, i.e. one class that had NRotB <6 and one with NRotB ≥6. Then, QSPR models were built for the neutral forms of each class as described for the DOS macrocycles. Models built using only 2D descriptors were slightly better than models that also used 3D structural information; in particular in prediction of the permeability for the external test sets (Supplementary Table 28 and Supplementary Fig. 4). As expected, the more rigid compounds in the NRotB <6 class were slightly better modelled than the more flexible ones in the NRotB ≥6 class both for the 2D and 3D models.

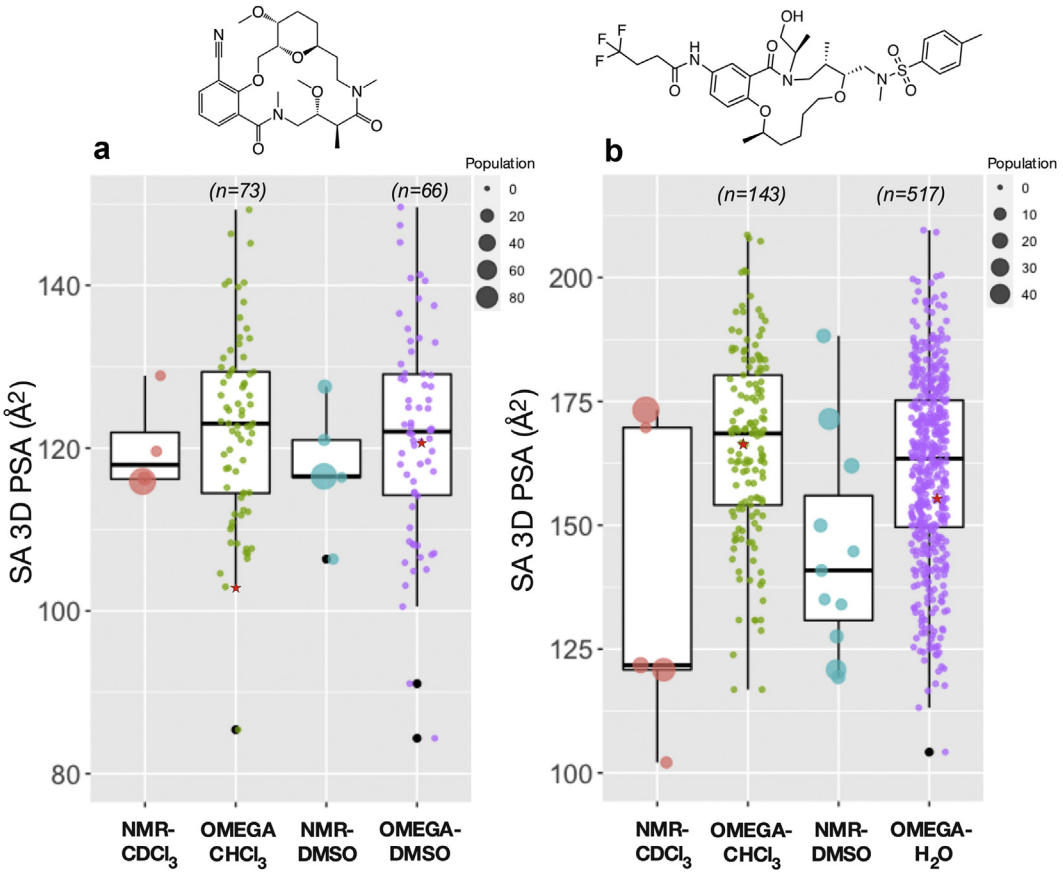


Fig. 8. Solvent accessible 3D polar surface area (SA 3D PSA) for **E2-enant** (a) and **G16** (b). The descriptor has been calculated for the conformations adopted in apolar ($CDCl_3$) and polar (DMSO-d_6) solutions, as determined by NMR spectroscopy, and for the conformations generated with OMEGA in apolar ($\varepsilon = 4.8$) and polar ($\varepsilon = 80$) implicit solvents. The size of each circle is representative of the population (in %) of each of the experimentally determined conformations. Boxplots show minimum and maximum values as whiskers, 25th and 75th percentiles as boxes, 50th percentiles as horizontal black bars. The MEC is indicated with a red star for each of the conformational ensembles generated with OMEGA, and the number of conformations (n) is given above each boxplot.

Conclusions

We found that machine learning-based classification models can be used to distinguish between low-medium and highly permeable compounds, both for the DOS macrocycle set and the linear drug set. Models based on 2D descriptors, which are fast to calculate, outperformed those based on more time-consuming sampling of 3D conformations. We conclude that 2D-based classification models are precise enough for use as virtual filters in early phases of drug discovery projects, when prioritizing compounds for synthesis from larger sets in similar chemical space. Importantly, the 2D machine learning-based models succeeded in the correct permeability classification of regio- and stereoisomeric macrocycles.

In the lead optimization phase better predictions of cell permeability are desired than a classification into low-medium or high. Earlier we have reported that the minimum solvent accessible 3D PSA of conformations determined by X-ray crystallography provided a more accurate model for cell permeability than the 2D descriptor TPSA for the 11 drugs in the bRo5 drug set.[16] Herein, we found this to be true also for a subset of the linear drug set. We also found that use of the conformations having the minimum solvent accessible 3D PSA[4,16] from conformational sampling provided reasonable permeability models both for the bRo5 drug set and for the subset of the linear drug set, supporting that 3D descriptors should be useful. Consequently, we tried to shed light on two major questions regarding regression permeability models for macrocycles, i.e. *i*) do 3D descriptors calculated from conformers provide better models than those based only on 2D descriptors, and *ii*) for which macrocycles can relevant conformations be predicted by conformational sampling? Unfortunately, global QSPR models based on 3D descriptors could not be developed for the 70 compounds in the DOS macrocycle set. However, a strong correlation between permeability and solvent accessible 3D PSA was found for one of the stereo- and regioisomeric series of this set.

Determination of the conformations populated in apolar and polar environments by NMR spectroscopy for a macrocycle from the well predicted series and one from a series which failed to give any permeability model confirmed that the well predicted series was more rigid. For the macrocycle from the well predicted, rigid series the sampled minimum energy conformations resembled the conformations in the solution phase ensembles (RMSD <2 Å). In contrast, conformations similar to those of the solution ensembles of the flexible series were usually found at energies significantly above the global minimum, i.e. at 5–15 kcal/mol above the minimum. It therefore appears that for flexible macrocycles, conformational sampling fails to identify the conformations that are essential for permeability. This arises from the inability of the force field to identify these conformations, but is not a result of incomplete sampling of relevant conformational space. This conclusion is in line with that from a recent study of ten drugs in bRo5 space.[28]

The well predicted series in the DOS macrocycle set has a Kier flexibility index of 9.3. Interestingly, the Kier flexibility indexes of the compounds in the linear drug set that were fairly well predicted based on the minimum energy conformations range up to 10. We

therefore speculate that a Kier flexibility index of approximately 10 constitutes a current upper limit for reasonably accurate conformational analysis using molecular mechanics forcefields for ranking of conformations. In an earlier study we reported that manual scoring of the overall polarity, intramolecular hydrogen bonding and steric shielding of polar groups in the low-energy conformations allowed ranking of the permeability of some of the more flexible series in the DOS macrocycle set.[15] These more flexible series are expected to behave as molecular chameleons that adapt their conformations to the environment in a manner which results in dynamic exposure of polar surface area, allowing compounds to display both high cell permeability and aqueous solubility.[15,16,52] The challenges posed by modelling of permeability for molecular chameleons also apply to flexible cyclic peptides, as reported previously by others.[10,14] We conclude that regression modelling of macrocycle cell permeability requires an analysis of the investigated dataset in terms of flexibility and its impact on the formation of intramolecular hydrogen bonds and other structural features masking polar regions, rather than a fast but a critical submission to computational tools.

Acknowledgements

This work was funded by grants from the Swedish Research Council (grant no. 2016-05160; J.K. and 2016-03602; M.E.). We thank OpenEye scientific software and ChemAxon for providing free academic licenses and Dr. Paul C. Hawkins at OpenEye Scientific Software for insightful discussions during this project. The QSARINS free license was kindly provided by Prof. Paola Gramatica who is acknowledged by the authors.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.xphs.2020.10.052.

References

1. Villar EA, Beglov D, Chennamadhavuni S, et al. How proteins bind macrocycles. *Nat Chem Biol*. 2014;10(9):723-731.
2. Doak BC, Zheng J, Dobritzsch D, Kihlberg J. How beyond rule of 5 drugs and clinical candidates bind to their targets. *J Med Chem*. 2016;59(6):2312-2327.
3. Tyagi M, Begnini F, Poongavanam V, Doak BC, Kihlberg J. Drug syntheses beyond the rule of 5. *Chem Eur J*. 2020;26(1):49-88.
4. Guimaraes CRW, Mathiowetz AM, Shalaeva M, Goetz G, Liras S. Use of 3D properties to characterize beyond rule-of-5 property space for passive permeation. *J Chem Inf Model*. 2012;52(4):882-890.
5. Rafi SB, Hearn BR, Vedantham P, Jacobson MP, Renslo AR. Predicting and improving the membrane permeability of peptidic small molecules. *J Med Chem*. 2012;55(7):3163-3169.
6. Leung SS, Sindhikara D, Jacobson MP. Simple predictive models of passive membrane permeability Incorporating size-dependent membrane-water partition. *J Chem Inf Model*. 2016;56(5):924-929.
7. Bennion BJ, Be NA, McNerney MW, et al. Predicting a drug's membrane permeability: a computational model validated with in vitro permeability assay data. *J Phys Chem B*. 2017;121(20):5228-5237.
8. Rezai T, Bock JE, Zhou MV, Kalyanaraman C, Lokey RS, Jacobson MP. Conformational flexibility, internal hydrogen bonding, and passive membrane permeability: successful in silico prediction of the relative permeabilities of cyclic peptides. *J Am Chem Soc*. 2006;128(43):14073-14080.
9. Rezai T, Yu B, Millhauser GL, Jacobson MP, Lokey RS. Testing the conformational hypothesis of passive membrane permeability using synthetic cyclic peptide diastereomers. *J Am Chem Soc*. 2006;128(8):2510-2511.
10. Wang CK, Swedberg JE, Harvey PJ, Kaas Q, Craik DJ. Conformational flexibility is a determinant of permeability for cyclosporin. *J Phys Chem B*. 2018;122(8):2261-2276.
11. Witek J, Wang SZ, Schroeder B, et al. Rationalization of the membrane permeability differences in a series of analogue cyclic decapeptides. *J Chem Inf Model*. 2019;59(1):294-308.
12. Kamenik AS, Lessel U, Fuchs JE, Fox T, Liedl KR. Peptidic macrocycles - conformational sampling and thermodynamic characterization. *J Chem Inf Model*. 2018;58(5):982-992.
13. Ono S, Naylor MR, Townsend CE, Okumura C, Okada O, Lokey RS. Conformation and permeability: cyclic hexapeptide diastereomers. *J Chem Inf Model*. 2019;59(6):2952-2963.
14. Le Roux A, Blaise E, Boudreault PL, et al. Structure-permeability relationship of semipeptidic macrocycles-understanding and optimizing passive permeability and efflux ratio. *J Med Chem*. 2020;63(13):6774-6783.
15. Over B, Matsson P, Tyrchan C, et al. Structural and conformational determinants of macrocycle cell permeability. *Nat Chem Biol*. 2016;12(12):1065-1074.
16. Rossi Sebastiano M, Doak BC, Backlund M, et al. Impact of dynamically exposed polarity on permeability and solubility of chameleonic drugs beyond the rule of 5. *J Med Chem*. 2018;61(9):4189-4202.
17. Potter T, Ermondi G, Newbury G, Caron G. Relating Caco-2 permeability to molecular properties using block relevance analysis. *MedChemComm*. 2015;6(4):626-629.
18. Ng C, Xiao YD, Lum BL, Han YH. Quantitative structure-activity relationships of methotrexate and methotrexate analogues transported by the rat multispecific resistance-associated protein 2 (rMrp2). *Eur J Pharm Sci*. 2005;26(5):405-413.
19. Morrissey KM, Wen CC, Johns SJ, Zhang L, Huang SM, Giacomini KM. The UCSF-FDA TransPortal: a public drug transporter database. *Clin Pharmacol Ther*. 2012;92(5):545-546.
20. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev*. 2001;46(1-3):3-26.
21. Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem*. 2002;45(12):2615-2623.
22. Sadowski J, Gasteiger J, Klebe G. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J Chem Inf Comput Sci*. 1994;34(4):1000-1008.
23. Corina. 3D Structure Generator CORINA Classic. 3.2 ed.: Molecular Networks GmbH, Nuremberg, Germany.
24. MOE. *Molecular Operating Environment, Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7*. 2016.
25. Hawkins PCD, Wlodek S. *OMEGA, Version 3.0*. Santa Fe, NM: OpenEye Scientific Software; 2016.
26. Grant JA, Pickup BT, Sykes MJ, Kitchen CA, Nicholls A. A simple formula for dielectric polarisation energies: the Sheffield Solvation Model. *Chem Phys Lett*. 2007;441(1):163-166.
27. Hawkins PC, Nicholls A. Conformer generation with OMEGA: learning from the data set and the analysis of failures. *J Chem Inf Model*. 2012;52(11):2919-2936.
28. Poongavanam V, Danelius E, Peintner S, et al. Conformational sampling of macrocyclic drugs in different environments: can we find the relevant conformations? *ACS Omega*. 2018;3(9):11742-11757.
29. Spartan. *14' v.1.1*. 2 ed. Irvine: Wavefunction, Inc.; 2013.
30. Labute P. A widely applicable set of descriptors. *J Mol Graph Model*. 2000;18(4-5):464-477.
31. Cruciani G, Crivori P, Carrupt PA, Testa B. Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. *J Mol Struct*. 2000;503(1):17-30.
32. Pedretti A, Villa L, Vistoli G. VEGA - an open platform to develop chemo-bioinformatics applications, using plug-in architecture and script programming. *J Comput Aided Mol Des*. 2004;18(3):167-173.
33. Gaillard P, Carrupt PA, Testa B, Boudon A. Molecular lipophilicity potential, a tool in 3D QSAR - method and applications. *J Comput Aided Mol Des*. 1994;8(2):83-96.
34. Olsson I-M, Gottfries J, Wold S. D-optimal onion designs in statistical molecular design. *Chemom Intell Lab Syst*. 2004;73(1):37-46.
35. Frank E, Hall MA, Witten IH. *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*. 4 ed. Morgan Kaufmann; 2016.
36. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
37. Vasanthanathan P, Taboureau O, Oostenbrink C, Vermeulen NP, Olsen L, Jorgensen FS. Classification of cytochrome P450 1A2 inhibitors and non-inhibitors by machine learning techniques. *Drug Metab Dispos*. 2009;37(3):658-664.
38. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*. 1975;405(2):442-451.
39. Chohan KK, Paine SW, Mistry J, Barton P, Davis AM. A rapid computational filter for cytochrome P450 1A2 inhibition potential of compound libraries. *J Med Chem*. 2005;48(16):5154-5161.
40. Hall MA. *Correlation-based Feature Subset Selection for Machine Learning*. University of Waikato; 1998.
41. Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S. QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. *J Comput Chem*. 2013;34(24):2121-2132.
42. Gramatica P, Sangion A. A historical excursus on the statistical validation parameters for QSAR models: a clarification concerning metrics and terminology. *J Chem Inf Model*. 2016;56(6):1127-1131.
43. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci*. 2003;22(1):69-77.
44. Cicero DO, Barbato G, Bazzo R. NMR analysis of molecular flexibility in solution: a new method for the study of complex distributions of rapidly exchanging conformations. application to a 13-residue peptide with an 8-residue loop. *J Am Chem Soc*. 1995;117(3):1027-1033.

45. Nevins N, Cicero D, Snyder JP. A test of the single-conformation hypothesis in the analysis of NMR data for small polar molecules: a force field comparison. *J Org Chem*. 1999;64(11):3979-3986.
46. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *SIGKDD Explor Newsl*. 2009;11(1):10-18.
47. Caron G, Digiesi V, Solaro S, Ermondi G. Flexibility in early drug discovery: focus on the beyond-Rule-of-5 chemical space. *Drug Discov Today*. 2020;25(4):621-627.
48. Danelius E, Poongavanam V, Peintner S, Wieske LHE, Erdelyi M, Kihlberg J. Solution conformations explain the chameleonic behaviour of macrocyclic drugs. *Chem Eur J*. 2020;26(23):5231-5244.
49. Peng C, Atilaw Y, Wang J, et al. Conformation of the macrocyclic drug lorlatinib in polar and nonpolar environments: a MD simulation and NMR study. *ACS Omega*. 2019;4(26):22245-22250.
50. Dickman R, Danelius E, Mitchell SA, Hansen DF, Erdélyi M, Tabor AB. A chemical biology approach to understanding molecular recognition of lipid II by Nisin(1−12): synthesis and NMR ensemble analysis of Nisin(1−12) and analogues. *Chem Eur J*. 2019;25(64):14572-14582.
51. Hawkins PCD. Conformation generation: the state of the art. *J Chem Inf Model*. 2017;57(8):1747-1756.
52. Whitty A, Zhong M, Viarengo L, Beglov D, Hall DR, Vajda S. Quantifying the chameleonic properties of macrocycles and other high-molecular-weight drugs. *Drug Discov Today*. 2016;21(5):712-717.