MycPermCheck: the Mycobacterium tuberculosis permeability prediction tool for small molecules

Benjamin Merget¹, David Zilian¹, Tobias Müller² and Christoph A. Sotriffer¹,*

¹Institute of Pharmacy and Food Chemistry, University of Würzburg, D-97074 Würzburg, Germany and ²Department of Bioinformatics, Biocenter, University of Würzburg, D-97074 Würzburg, Germany

ABSTRACT

Motivation: With >8 million new cases in 2010, particularly documented in developing countries, tuberculosis (TB) is still a highly present pandemic and often terminal. This is also due to the emergence of antibiotic-resistant strains (MDR-TB and XDR-TB) of the primary causative agent Mycobacterium tuberculosis (MTB). Efforts to develop new effective drugs against MTB are restrained by the unique and largely impermeable composition of the mycobacterial cell wall.

Results: Based on a database of antimycobacterial substances (CDD TB), 3815 compounds were classified as active and thus permeable. A data mining approach was conducted to gather the physico-chemical similarities of these substances and delimit them from a generic dataset of drug-like molecules. On the basis of the differences in these datasets, a regression model was generated and implemented into the online tool MycPermCheck to predict the permeability probability of small organic compounds.

Discussion: Given the current lack of precise molecular criteria determining mycobacterial permeability, MycPermCheck represents an unprecedented prediction tool intended to support antimycobacterial drug discovery. It follows a novel knowledge-driven approach to estimate the permeability probability of small organic compounds. As such, MycPermCheck can be used intuitively as an additional selection criterion for potential new inhibitors against MTB. Based on the validation results, its performance is expected to be of high practical value for virtual screening purposes.

Availability: The online tool is freely accessible under the URL http://www.mycpermcheck.aksotriffer.pharmazie.uni-wuerzburg.de

Additional information: Supplementary data are available at Bioinformatics online.

Received on July 9, 2012; revised on September 25, 2012; accepted on October 22, 2012

1 INTRODUCTION

Despite its ancient origin and the successes of antibacterial chemotherapeutics introduced >50 years ago, tuberculosis (TB) is still one of the most widespread and abundant infectious diseases. In 2010, 8.8 million new cases were documented, particularly in developing countries (World Health Organization, 2012). Since the discovery of Mycobacterium tuberculosis (MTB) as the primary causative agent of TB, a complex first-line treatment was developed based on the prodrug isoniazid. However, because of the emergence of multi- and extensively drug-resistant strains (MDR-TB and XDR-TB), the design of novel potent inhibitors is an ongoing challenge (Koul et al., 2011).

An important natural defense mechanism of MTB is its thick and waxy cell wall providing a first powerful barrier against antibiotic drugs. It consists of a peptidoglycan-arabinogalactan-mycolic acid core as well as the outmost layer, the so-called capsule (Draper and Daffé, 2005; Rastogi et al., 1986). It has been shown that not only hydrophilic agents, but also lipophilic agents may have severe problems passing the permeability barrier of the cell wall, owing to the unusually low fluidity of the lipid bilayer (Liu et al., 1996).

Without the ability to penetrate this cell wall, even most potent inhibitors of validated mycobacterial drug targets like InhA (Quemard et al., 1995) will not have any efficacy. Unfortunately, data about mycobacterial permeability properties of chemical compounds are hardly available, hampering the development of knowledge-based methods for permeability estimation. However, as in most of the cases a compound must permeate the mycobacterial cell wall to show antimycobacterial activity, it is reasonable to infer an ability to pass this barrier for compounds active against mycobacteria. In 2010, Ekins and colleagues developed a collaborative database (CDD TB) of >200 000 compounds, which had been tested for antibiotic activity against MTB (Ekins et al., 2010). Over 3800 structures showed growth inhibition of ≥90% at a concentration of 10 μM. Obviously, these compounds have sufficient permeability to be active against MTB and may, thus, be used as a knowledge base for analyzing permeability-determining features. Accordingly, an extensive data mining approach based on the physico-chemical properties of this dataset was performed with the subsequent development of a regression model. With this knowledge-based classification system, the permeability of potential new compounds against MTB can be estimated with high accuracy and quantified comfortably.

2 SYSTEM DESCRIPTION

2.1 Datasets

The MLSMR dataset (Ananthan et al., 2009) of the CDD TB database (Ekins et al., 2010; Holman et al., 2009) was filtered for compounds that showed a mycobacterial growth inhibition of ≥90% at 10 μM and a molecular weight <500 Dalton. This step reduced the total number of considered molecules to 3815 chemical structures. All compounds were converted to 3D structures with the program Corina (available from Molecular...
Networks GmbH, Erlangen, Germany) (Sadowski et al., 1994). These structures were processed by the tool LigPrep (Version 2.3, Schrödinger, LLC, New York, NY, 2009) for protonation (at pH 7.0 ± 2.0), stereoisomerization, tautomerization and subsequent energy minimization. Physico-chemical descriptors were then calculated for each molecule with Schrödinger QikProp (Version 3.4, Schrödinger, LLC, New York, NY, 2011). Compounds with incomplete descriptor data were removed, leaving 3727 structures. This dataset is hereinafter referred to as Actives.

The foundation of this work is the assumption that a compound must sufficiently well permeate the mycobacterial cell envelope (consisting of cell wall, periplasm and inner membrane) to unleash its effect within the target cell. Therefore, the dataset Actives can be classified as ‘permeable’ (i.e. the corresponding compounds have sufficient permeability to be active). Far more difficult is the generation of a sufficiently large ‘impermeable’ (negative) dataset, as only few studies regarding the permeability of mycobacteria are available (e.g. Hong and Hopfinger, 2004; Jarlier and Nikaido, 1990; Laneelle and Daffe, 2009; Trias and Benz, 1994). Simply taking the inactive compounds from MTB activity tests is obviously not possible, as a lack of permeability may not be the only reason for inactivity. This issue can be addressed by collecting compounds that are active against MTB targets in target-based (e.g. enzymatic) assays, but inactive in a whole-cell MTB assay. This approach was indeed followed herein to generate a validation dataset (cf. Section 2.4). The number of compounds obtainable by this way is, however, by far not sufficient for data mining and training-set generation. Accordingly, randomly drawn datasets of drug-like small molecules were used as ‘negative’ data. These should allow to determine whether the ‘permeable’ substances show any significant differences with respect to random drug-like compounds. For this purpose, the drug-like subset of the ZINC database [Irwin and Shoichet (2005), newest version ZINC12] was processed in the same manner as the Actives. Thereby, an extensive table of physico-chemical properties of a randomly distributed dataset of drug-like molecules was obtained. This dataset is hereinafter referred to as ZINC. An overview of the datasets used in this survey, as well as information about the chemical diversity of the Actives dataset, is available in Supplement S1.

2.2 Descriptor selection and visualization

Pairwise Mann-Whitney-U-tests of Actives against ZINC (several sets of 100 randomly chosen structures each) were performed for each of the 51 QikProp descriptors. The tests showed consistent results regarding their P-values. Figure 1 depicts the distribution of the calculated P-values using the R package BioNet (Beisser et al., 2010; R Development Core Team, 2011) for one representative test set including a fitted beta and uniform distribution. Under the null hypothesis, the P-values are uniformly distributed representing only noise. The remaining part of the P-values describes the signal distribution. The fitted beta-uniform-mixture model (Pounds and Morris, 2003) shows a strong signal of significant differences in the physico-chemical properties of Actives and ZINC. Based on a descriptive representation of the 51 distributions (data not shown) and a common understanding of physico-chemical descriptors for drug development, five QikProp descriptors (P < 0.001) were further considered:

- FOSA: The hydrophobic part of the solvent accessible surface area (saturated carbon and attached hydrogen atoms);
- QPlogPo:w: The logarithm of the calculated octanol/water partition coefficient (hereinafter called logP);
- PISA: The π-interacting part of the solvent accessible surface area;
- accepHIB: The number of H-bond acceptors;
- glob: The generic spherical surface to molecule surface ratio.

Other common molecular descriptors (e.g. molecular weight or the number of H-bond donors) were not considered for model derivation, mostly due to insufficient differences between the datasets with respect to these descriptors. To increase statistical significance, in all of the following randomly chosen datasets, the size was increased to 1000 per group. Figure 2 illustrates the distribution of the five selected descriptors for a representative randomly chosen test set of Actives as well as a randomly chosen ZINC-test set of equal size. The non-overlapping box plots notches show significant differences in the medians of the distributions of these five descriptors for the two datasets.

A first impression whether a potential new inhibitor might show descriptor values typical for permeable compounds.
can be gained from a comparison with the distribution of the descriptors in the *Actives* dataset. For this purpose, four borders have been defined to better delimit the physico-chemical space of the permeable substances: *upper*, *up*, *low* and *lower* (Table 1).

The borders *up* and *low* are defined by the 75 and 25% quantile of the training dataset, respectively. *Upper* and *lower* represent 75 and 25% quantile /C6 half the interquartile range, respectively.

2.3 PCA and logistic regression method

Although the value mapping of each descriptor for a compound of interest is useful for later interpretation of results, a reliable prediction of the permeability cannot be achieved this way. Thus, the permeability prediction approach is based on multivariate statistics. First, 25 principal component analyses (PCAs) were performed based on the five chosen descriptors using random test sets of 1000 permeable substances of the *Actives* dataset and 1000 substances of the *ZINC* dataset. Then, the resulting coordinates were projected to the first principal component. All PCAs showed coherent results: each time a one-dimensional representation of principal component 1 (PC1) showed the best splitting of the two groups *Actives* and *ZINC*. Thus, by reducing the multi-dimensional information space to only the first principal component (42.4% information content), it is possible to achieve a maximum separation of these two groups. All PCA analyses were performed with the *vegan* R package (Oksanen et al., 2011).

The PC1 coordinates of one representative PCA were then used to generate a logistic regression model [Fig. 3; figure created with the R package *popbio* (Stubben and Milligan, 2007)] using R (R Development Core Team, 2011). The obtained logistic regression function follows:

$$P(z) = \frac{1}{1 + e^{-z}}$$

with $z = f(x) = \beta \cdot x$

Table 1. Borders of the five chosen descriptors based on the distributions of the descriptors in the complete *Actives* dataset, as further described in the text

<table>
<thead>
<tr>
<th>FOSA</th>
<th>logP</th>
<th>PISA</th>
<th>accepHB</th>
<th>glob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper</td>
<td>362.95</td>
<td>5.329</td>
<td>430.66</td>
<td>7.125</td>
</tr>
<tr>
<td>Up</td>
<td>272.23</td>
<td>4.479</td>
<td>355.49</td>
<td>6.000</td>
</tr>
<tr>
<td>Low</td>
<td>90.80</td>
<td>2.779</td>
<td>205.16</td>
<td>3.750</td>
</tr>
<tr>
<td>Lower</td>
<td>0.09</td>
<td>1.929</td>
<td>129.99</td>
<td>2.625</td>
</tr>
</tbody>
</table>

FOSA and PISA are measured in Å².
are calculated according to the existing PCA of the training data, (ii) then, the coordinate of PC1 is used as input \( x \) for the logistic regression model. As a result, the user receives a calculated probability \( 0 < P(x) < 1 \) of a compound to be classified as permeable.

### 2.4 Evaluation

For evaluation of the logistic regression model, the ChEMBL database (Gaulton et al., 2011) was browsed for antimycobacterially active compounds with a minimal inhibitory concentration (MIC) \( \leq 10 \mu M \), yielding a total of 771 permeable structures (Permeables) absent in the training dataset Actives. The compounds were prepared the same way as the compounds of the training set (3D conversion, protonation, stereoisomerization, tautomerization and energy minimization). After descriptor calculation with QikProp, MycPermCheck was used with the option Calculate Mean of all Isomeric Forms (as described in the next section). The calculated permeability probabilities show a median of 0.987 (±0.013 median absolute deviation). Hence, MycPermCheck yields valid predictions for these antimycobacterial and, thus, permeable substances.

To further evaluate MycPermCheck with biological real-life data, the intersection of two different datasets was generated: first, the CDD TB (Ekins et al., 2010) was filtered for substances which show <10% antimycobacterial activity at 10 \( \mu M \), yielding >190000 compounds. Simultaneously, the ChEMBL database (Gaulton et al., 2011) was browsed for assays against MTB targets and filtered for structures marked as active within the database according to their half-maximal inhibitory concentration (IC\textsubscript{50} value) or enzymatic inhibition constant (K\textsubscript{i} value) of \( \leq 10 \mu M \). On the basis of their International Chemical Identifiers (InChI strings), the intersection of the two datasets was established, yielding 22 compounds. As an additional filter criterion, an all versus all similarity matrix was generated based on the atom pair similarity of these compounds using ChemMineR (Cao et al., 2008; Carhart et al., 1985). A compound showing >80% similarity to another compound was removed. One structure (CHEMBL592772) was affected, yielding a final number of 21 compounds with low IC\textsubscript{50} or K\textsubscript{i} values (i.e. activity against an MTB target in an \textit{in vitro} enzyme assay), but without antimycobacterial activity. Based on the assumption that the most likely reason for the inactivity of these compounds against MTB is their inability to penetrate the mycobacterial cell wall, this dataset should be a collection of impermeable compounds. The 21 compounds (Impermeables) (see structures 1M1-1M21 in Supplement S2) were prepared the same way as the Permeables and the compounds of the training set. The calculated QikProp descriptors were then processed by MycPermCheck, again with the option Calculate Mean of all Isomeric Forms. The obtained permeability probabilities show a median of 0.188 (±0.188 median absolute deviation).

Fifty combined datasets of the 21 Impermeables and 21 randomly chosen Permeables were then created to perform a multiple Receiver Operating Characteristic (ROC) analysis with the R package \textit{ROCR} (Sing et al., 2005) (Fig. 4a). The single ROC curves were averaged by true-positive rate (black curve) as well as by threshold (colored curve). The color scale illustrated in Figure 4 represents the actual permeability probability that is used as a sliding threshold for establishing the true-versus-false positive rate. Error bars indicate the standard deviation of the true-positives–averaged curve. The gray dashed curves illustrate the single ROC analyses. The thick black curve shows the ROC curve averaged by true-positive rate, whereas the thick colored curve represents the calculated average by threshold. Error bars indicate the standard deviation of the true-positives–averaged curve. The dashed angle bisector illustrates a uniform rise of the true-positive and false-positive rate, equivalent to a random model. (b) ROC analysis of calculated permeability probabilities for the evaluation dataset of 19 InhA inhibitors and 21 Impermeables. The true-positive rate is plotted against the false-positive rate for a rising threshold of the calculated permeability probability (indicated by the color scale). The gray dashed curves illustrate the single ROC analyses. The thick black curve shows the ROC curve averaged by true-positive rate, whereas the thick colored curve represents the calculated average by threshold. Error bars indicate the standard deviation of the true-positives–averaged curve. The dashed angle bisector illustrates a uniform rise of the true-positive and false-positive rate, equivalent to a random model. Both ROC curves show a clear enrichment of permeable compounds at the top of the permeability-ranked list.
false-positive rate and, hence, the ROC curve. The average ROC curve shows a fast increase of the true-positive rate without producing an equivalent amount of false positives, indicating that reliable and practically useful results can be obtained with MycPermCheck for randomly selected permeable molecules. At a false-positive rate of \( \beta = 10\% \) (a specificity of \( 1 - \beta = 90\% \)), a true-positive rate (sensitivity) of \( 67.2 \pm 9.5\% \) (SD) is achieved (i.e. over two-thirds of all true positives already appear at this cut-off). A less strict false-positive rate of 25\% yields a higher sensitivity of \( 72.2 \pm 8.9\% \) (SD). At a permeability probability cut-off of 0.816, a specificity of 90\% is obtained, whereas a cut-off of 0.553 matches a specificity of 75\%. These two cut-offs (rounded to 0.82 and 0.55, respectively) form the basis of the traffic-lights color code of the web program output, as described in Section 2.5.

For evaluation of the logistic regression model for three well-studied classes of inhibitors of the mycobacterial enzyme enoyl acyl carrier protein reductase (InhA), the chemical structures of 20 mycobacterial inhibitors (not present in the dataset Actives) were extracted from the literature (Freundlich et al., 2009; He et al., 2006, 2007; Luckner et al., 2010; Muddassar et al., 2016; Sullivan et al., 2006). Again, the structures were filtered for atom pair similarity <80\%. After removing one compound (SPP), 19 mycobacterial inhibitors remained in this test set (see structures P1–P19 in Supplement S2). These inhibitors cover a broad chemical range from triclosan and its derivatives (diphenyl ethers) to arylamides and pyrrolidine carboxamides. Again, the compounds were prepared as before (3D conversion, protonation, stereoisomerization, tautomerization, energy minimization and QikProp descriptor calculation). The calculated permeability probabilities show a median value of 0.999 (±0.001 median absolute deviation). Therefore, MycPermCheck yields valid predictions for these permeable substances.

A second ROC analysis was performed for a combined dataset of these 19 active substances and the previously detected 21 impermeable compounds (Fig. 4b). Regarding the highly active InhA inhibitors, MycPermCheck shows an even faster increase of true-positive results than for the randomized evaluation test sets. At a false-positive rate of \( \beta = 10\% \), a true-positive rate (sensitivity) of \( 84.2\% \) is achieved, whereas a false-positive rate of 25\% corresponds to a sensitivity of 94.7\%. The actual permeability cut-offs at these false-positive rates are slightly higher than those of the multiple ROC analysis of the randomized evaluation test sets (0.929 and 0.605, respectively). These results illustrate that MycPermCheck is applicable on inhibitors of the MTB target InhA.

### 2.5 Implementation

MycPermCheck is a freely accessible online tool. It is programmed entirely in perl, making use of the perl packages *Statistics::R* for file formatting and probability prediction with R and of the CGI-package for displaying browser contents. Usage of the program begins by accessing the startup screen (Fig. 5a). Here, the input data [a QikProp comma-separated values (CSV) file] must be chosen using the browser function of the website. The checkbox ‘Calculate Mean of Isomeric Forms’ defines whether all ‘isomeric’ forms of a compound (i.e. tautomers, protomers, stereoisomers, conformers, etc.; indicated and recognized by the same molecule name in the QikProp CSV file) should be considered and averaged, or whether, alternatively, only the first representative should be used for the calculation (box unchecked). Clicking the ‘Submit’ button submits the job to the instant calculation of the permeability probabilities.

Within few seconds, a list of the submitted compounds appears as a result, sorted either by the calculated permeability probability (default) or by the compound name (optional.
selection on submission). The list shows the calculated permeability probability in the first column after the compound name, followed by the single descriptor values (Fig. 5b; detailed list of evaluation data including structures in Supplement S2). For the single descriptor values, blue-scale colors are assigned based on the borders defined in Table 1: (i) if the value lies between the borders up and low, this state is colored light blue; (ii) a value between the borders up and upper or low and lower, respectively, is colored blue; (iii) a value below lower or above upper is illustrated by a dark blue coloring. This graphical representation represents the chemical similarity of a given compound to the training dataset Actives in terms of the five most relevant descriptors (see colored descriptor values in Fig. 5b). In contrast, the quality of each result is rated according to a simple and intuitive traffic-lights system: for highlighting the permeability probability, two borders have been defined based on the ROC analyses of the evaluation dataset (Fig. 4). The first cut-off of 0.82 corresponds to a false-positive rate of ~10%. Results with probabilities above this value (>0.82) are marked green. The second cut-off of 0.55 corresponds to a false-positive rate of ~25%. Results above this threshold are marked orange. Probabilities below 0.55 are colored red. A download function can be used to save all results in a CSV file for further processing by the user.

Besides the use of Maestro QikProp descriptors for estimating the permeability probability, MycPermCheck is also able to process CSV output files of the open-source java descriptor calculation package PaDEL-Descriptor (Yap, 2011). A complete evaluation of the PCA and regression model for PaDEL descriptor input is available in Supplement S3.

The program is accessible under the following website: http://www.mycpermcheck.aksotrieff.pharmazie.uni-wuerzburg.de

### 3 DISCUSSION

MycPermCheck is an intuitively accessible online tool for knowledge-based estimation of the permeability of potential antymycobacterial compounds with respect to the MTB cell wall. The program is based on a chemoinformatic data-mining approach without any assumptions regarding the uptake mechanism. It is, hence, generally applicable to drug-like compounds with a molecular weight <500 Dalton. With statistical significance, a training set of permeable compounds (Actives) could be delimited from randomly distributed drug-like molecules based on five physico-chemical descriptors in a principal component analysis. Based on the resulting first principal component, a logistic-regression model for estimating the permeability probability could be derived. Thereby, instead of hard cut-offs for molecular descriptor interpretation [as, for example, in Lipinski’s Rule of 5 (Lipinski et al., 1997)], a ‘more realistic and graded description of the continuum of compound quality’ is obtained, an advantage recently pointed out by Bickerton and colleagues in the context of their quantitative estimate of drug likeness (Bickerton et al., 2012).

MycPermCheck was multiply tested on 50 evaluation datasets of 21 permeable compounds and a set of 21 impermeable compounds. With a standard deviation of true positives of ~9% for a specificity of both 90 and 75%, the average of the multiple ROC curves shows a robust prediction for randomly selected permeable compounds (cf. Fig. 4a). Moreover, MycPermCheck was tested on 19 highly active InhA inhibitors and 21 impermeable compounds, leading to a good enrichment of the permeable compounds in the range of the highest probability values and of the impermeable compounds in the range of the lowest probability values (cf. Fig. 4b and Supplement S2). In fact, among the top 10 compounds (1.000 ≥ P ≥ 0.999), no false positive is found. Moreover, the 10 lowest ranked compounds (0.002 ≥ P ≥ 0.000) are all true negatives, i.e. impermeables. Comparing the molecular structures and descriptor values of these two groups of top-ranked and lowest-ranked compounds indicates that with only few exceptions permeable compounds are characterized by a high PISA to FOSA ratio (i.e. the π-interacting surface area is generally much larger than the hydrophobic surface area), a logP of >4 and an acceptHB value <2. In contrast, the 10 lowest-ranked impermeable compounds show frequently larger FOSA than PISA values, have low logP values (often <1) and generally an acceptHB value of >5. The compounds with the highest permeability probability show indeed at least two aromatic ring systems to which small to moderately sized hydrophobic substituents and few H-bond acceptors are attached. The impermeable compounds, instead, have often only one (if any) aromatic ring system and—despite a significant hydrophobic surface area—a higher polarity and more H-bond acceptors (cf. compounds IM14-IM21 in Supplement S2).

These observations may provide some guidelines for ensuring mycobacterial permeability of designed compounds. Nevertheless, it is also clear that looking at single parameters only is not sufficient. In fact, simply aiming for descriptor values that are within the ‘up’ and ‘low’ borders defined in Table 1 does not ensure a high permeability probability. For example, compound IM6 shows three descriptor values within the light-blue range, one within the blue (logP) and only one within the dark-blue range (FOSA), yet the probability is only 0.472, and the compound is indeed impermeable. Conversely, compounds with high probabilities may also show descriptor values in the blue or dark-blue range, as, for example, the top three compounds P1, P2 and P3 (cf. Supplement S2). Thus, a ‘one-dimensional’ view focused at single descriptor values and their univariate statistics is indeed of little value. Instead, the correct combination and relative weighting of molecular properties is essential, as incorporated in the logistic regression model based on the first principal component: PISA should be larger than FOSA, logP not too small (rather >3) and acceptHB not too large (rather <5); the descriptor glob plays a minor role in modulating the probability.

Although the validation results of MycPermCheck illustrate a high predictivity, it is also clear that an absolute accuracy should not be expected. Considering the false positives IM1 and IM2, which both obtain probability values >0.9, their lack of permeability cannot be explained in terms of the descriptor values, as they fit the general trends observed for the truly permeable compounds. It should be kept in mind, however, that compounds of the Impermeables validation set are actually only assumed to be impermeable because of a lack of antimycobacterial activity despite an inhibitory effect in an in vitro target-based assay. Obviously, this lack of activity may also have other reasons than mere impermeability. Examples include the activity of efflux pumps and the in vivo degradation/inactivation of a compound. Accordingly, it cannot be ruled out that
IM1 and IM2 are indeed permeable, but inactive for other reasons. Considering false negatives, a few cases are observed as well. Of the 19 permeable compounds in the validation set, P18 and P19 obtain probability values <0.7. Although these compounds have larger FOSA than PISA values and 5-6 hydrogen bond acceptors, they show antimycobacterial activity.

These examples illustrate the limits of the approach, which (i) does not make any distinction with respect to the uptake mechanism and (ii) is not based on a dataset of experimentally proven impermeable compounds. Clearly, a sufficiently large dataset of compounds with known uptake mechanism or confirmed impermeability would be highly advantageous, both for the derivation of improved models as well as for a more reliable validation of the current model. Given the lack of such a dataset, MycPermCheck is an attempt to make best use of the available knowledge base.

Despite these shortcomings, MycPermCheck is expected to be of significant practical value for any (virtual) screening endeavor dedicated to antimycobacterial drug design. The validation results indicate that a clear enrichment of potentially permeable compounds and a highly reliable filtering of impermeable compounds (with $P<0.1$) can be achieved with this, to our knowledge, unique approach. Accordingly, MycPermCheck may serve as an additional selection criterion on virtual screening and as a utility for increasing the likelihood of obtaining permeable antimycobacterial compounds.

ACKNOWLEDGEMENT

We thank Maximilian Kuhn (University of Würzburg) for providing valuable preliminary work on the subject.

Funding: This work was supported by the German Research Foundation DFG (SFB 630).

Conflict of Interest: none declared.

REFERENCES


