Using a graph convolutional neural network model to identify bile salt export pump inhibitors

Mohamed Diwan M. Abdulhameed[1,2], Ruifeng Liu[1,2], and Anders Wallqvist[1]**

[1]Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Development Command, Fort Detrick, MD, USA

[2]The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD, USA

*Corresponding Authors:
MDMA: E-mail: mabdulhameed@bhsai.org

AW: E-mail: sven.a.wallqvist.civ@health.mil

DoD Biotechnology High Performance Computing Software Applications Institute
Telemedicine and Advanced Technology Research Center
U.S. Army Medical Research and Development Command
ATTN: FCMR-TT, 504 Scott Street
Fort Detrick, MD 21702-5012
Tel: (301) 619-1989; Fax: (301) 619-1983

ABSTRACT

The bile salt export pump (BSEP) is a key transporter involved in the efflux of bile salts from hepatocytes to bile canaliculi. Inhibition of BSEP leads to accumulation of bile salts within the hepatocytes, leading to possible cholestasis and drug-induced liver injury. Screening for and identification of chemicals that inhibit this transporter aid in understanding the safety liabilities of these chemicals. Moreover, computational approaches to identify BSEP inhibitors provide an alternative to the more resource-intensive, gold-standard experimental approaches. Here, we used publicly available data to develop predictive machine learning models for identification of potential BSEP inhibitors. Specifically, we analyzed the utility of a graph convolutional neural network (GCNN)-based approach in combination with multi-task learning to identify BSEP inhibitors. Our analyses showed that the developed GCNN model performed better than the variable-nearest neighbor and Bayesian machine learning approaches, with a cross-validation receiver operating characteristic area under the curve of 0.86. In addition, we compared GCNN-based single-task and multi-task models and evaluated their utility in addressing data limitation challenges commonly observed in bioactivity modeling. We found that multi-task models performed better than single-task models and can be utilized to identify active molecules for targets with limited data availability. Overall, our developed multi-task GCNN-based BSEP model provides a useful tool for prioritizing hits during early drug discovery and in risk assessment of chemicals.

KEYWORDS: Bile salt export pump, inhibitors, graph convolutional neural network, multi-task learning, QSAR, ADMET, cholestasis, DILI

1. INTRODUCTION

The bile salt export pump (BSEP; gene symbol *ABCB11*) is a member of the ATP-binding cassette transporter family and is an important cell-membrane protein that regulates the efflux of bile salts from hepatocytes to bile canaliculi (**Figure 1A**).[1] Bile salts play a key role in the digestion of fatty substances, and nearly 90% of bile salts are reabsorbed from the intestines and shuttle back to the hepatocytes through entero-hepatic circulation.[2, 3] BSEP acts as the rate-limiting step in bile formation and is essential for normal liver function and maintenance of bile flow.[4] Patients with a genetic mutation that results in loss of BSEP function are known to develop progressive familial intrahepatic cholestasis type 2 (PFIC2), a genetic disorder that is marked by cholestasis within ~3 months after birth and can lead to death at a young age (<30 years) if left untreated.[4] Drug- or toxic chemical-induced inhibition of BSEP is now recognized as the molecular initiating event for the cholestasis adverse outcome pathway, i.e., inhibition of BSEP is causally related to the adverse outcome.[5] When BSEP is inhibited, intracellular bile acid levels rise to cytotoxic concentrations. (**Figure 1B**).[6] The degree of cytotoxicity is determined by the hydrophobicity of the bile acid, with more hydrophobic bile acids being more toxic.[7, 8] This can result in hepatocyte injury through multiple mechanisms, including oxidative stress, mitochondrial damage, apoptosis, and necrosis.[5, 7, 9] For example, troglitazone, an approved anti-diabetic drug, was withdrawn from the market due to the drug-induced liver injury (DILI) caused by BSEP inhibition.[10] Furthermore, BSEP inhibition is considered as an indicator of the DILI potential of drugs.[11] Due to the physiological significance of BSEP and its role in adverse health effects, the European Medicines Agency now recommends *in vitro* screening for BSEP inhibition as part of the evaluation of new drugs.[12, 13] Overall, there is a growing emphasis on earlier screening and identification of BSEP inhibition potential of new hits/lead molecules as it

will help avoid costly, late-stage failures during drug development and support risk assessments of chemicals.

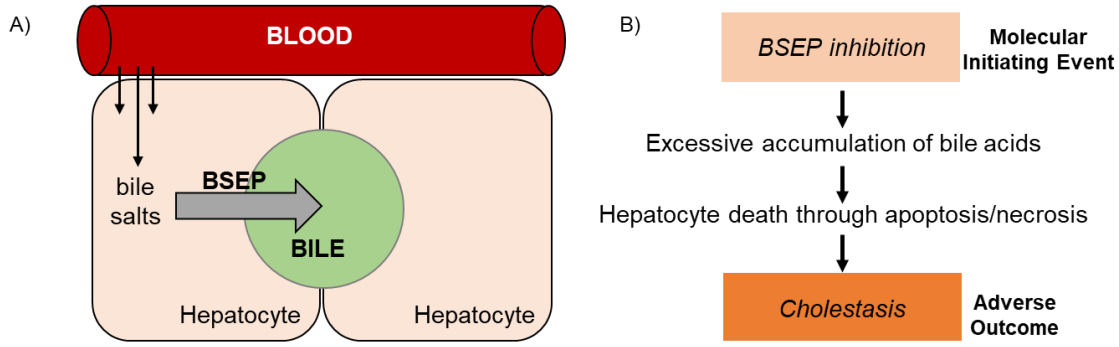

Figure 1. A) Schematic illustration of the role of bile salt export pump (BSEP), the primary transporter involved in the efflux of bile salts from hepatocytes to bile canaliculi. Bile salts are either synthesized in the hepatocytes or re-absorbed from the blood through entero-hepatic circulation, and BSEP performs the efflux of bile salts from the hepatocytes. B) Summary of BSEP inhibition leading to cholestasis through excessive accumulation of cytotoxic bile acids within the hepatocytes. BSEP inhibition acts as the molecular initiating event in the cholestasis adverse outcome pathway.

Because experimental screening approaches are resource intensive and time consuming, it is not feasible to perform high-volume screening of large chemical databases for BSEP during the early hit identification/prioritization stage of drug discovery.[14] Computational approaches provide an alternative to experimental screening approaches and can be grouped into two categories: *1*) structure based and *2*) ligand based.[15] Structure-based approaches focus on the protein structure of BSEP and molecular docking calculations. Jain et al. have performed docking analyses using a homology model of BSEP[12] and observed that combining ligand-based models with docking resulted in better performance when classifying BSEP inhibitors than utilizing a docking

approach alone.[12] In general, structure-based approaches are challenging when considering protein flexibility, identifying varied ligand-binding sites, and accurately characterizing the effect of solvent and membrane-mediated interactions.[15] Ligand-based approaches do not have these limitations, but rely on the availability of bioactivity data for a set of ligands.[16-18] Hence, ligand-based approaches are more commonly used to develop BSEP classification models.[14, 19-23] Hirano et al. reported the first ligand-based model for BSEP,[19] developing multiple linear regression models using a small dataset of 38 compounds. They reported a coefficient of determination (R^2) of 0.952 for their training set, but no test set or cross-validation set was evaluated.[19] Warner et al. developed a support vector machine model using 196 AstraZeneca in-house descriptors, and their best model had an accuracy of 87%.[20] Their work also highlights the need for additional descriptors and machine learning models, as they showed that the use of simple molecular descriptors, such as molecular weight and logP, alone can lead to false negatives.[20] The data used in their work were disclosed in the paper, but the model and in-house descriptor calculation program are not available.[20] Montanari et al. developed a random forest model using 838 compounds shared internally by AstraZeneca as part of the eTox project.[21] They utilized the commercial Molecular Operating Environment (MOE) software for descriptor calculation, and their best model had an accuracy of 80% during test set evaluation.[21] Their model is provided as a KNIME workflow but requires the commercial MOE software for descriptor calculation, and the data used for model development are not publicly available.[21] McLoughlin et al. utilized a proprietary GSK dataset and developed classification (neural network and random forest) and regression models for BSEP models using their ATOM Modeling Pipeline (AMPL).[14] They also developed an open model using public data, but it was evaluated using the proprietary dataset and requires the use of their AMPL platform.[14] The

model provided an accuracy of 77% when examining an external test set.[14] More recently, Rodríguez-Pérez et al. reported the development of a BSEP classification and regression model using in-house Novartis BSEP inhibition assay data.[22] Their extreme gradient boosting classification model has a balanced accuracy of 69% on the calibration set.[22] Overall, the previous computational modelling work either lacks publicly available data for model development and evaluation or lacks easily accessible models. Significantly, all the previous models utilize molecular descriptors or fingerprints to represent the compounds during model building. There are no reports so far that utilize an alternative graph convolutional neural network (GCNN) approach for model building. The GCNN is a recent development in the field of cheminformatics where the molecular structure is learned in an automated manner, in contrast to previous fingerprint-based approaches that require pre-defined sets of chemical sub-structures/functional groups.[24, 25] A detailed evaluation of the computational models using benchmark datasets showed better performance of the GCNN approach over traditional fingerprint/descriptor-based approaches.[24] While other studies have shown that multi-task models perform better than single-task models,[26, 27] so far the utility of GCNN and multi-task approaches in BSEP modeling has not been studied.

In this work, we developed GCNN models to predict BSEP inhibitors and performed a detailed comparison between single- and multi-task models. First, we collected publicly available BSEP inhibition data from BindingDB. Next, we utilized ChemProp,[24] a publicly available tool, to develop GCNN models for BSEP and compared the performance of GCNN models with variable-nearest neighbor and Bayesian approaches. We found that the GCNN model outperformed the other types of models. Then, we collected datasets associated with the blood-brain barrier (BBB) and the human ether-à-go-go-related gene (hERG) to develop multi-task

models. We also evaluated whether multi-task models are useful in addressing data limitation challenges, a common problem in bioactivity modeling. We found that multi-task models consistently performed better than single-task models in predicting the activity of evaluation test sets. Finally, we evaluated the effect of additional datasets on the performance of multi-task models. We found that the addition of human immunodeficiency virus protease (HIVpro) further helped to improve model performance. This agrees with the literature finding that many HIVpro inhibitors are known to inhibit BSEP and other transporters. Overall, the developed multi-task GCNN-based BSEP models provide a rapid computational method for safety risk assessment during early drug discovery.

2. METHODS

2.1. Dataset and pre-processing

We collected 1,689 compounds with publicly available BSEP bioactivity data from BindingDB (accessed on 5-19-2022).[28] BindingDB curates the activity data for each target from various literature sources, including patents.[28] We used Pipeline Pilot (Version 18.1.100.11) to pre-process the molecules, removing duplicate compounds, salts, and mixtures and standardizing the molecules.[29] Standardization refers to a molecule pre-processing step wherein proper bond order, aromaticity, and hydrogens are assigned.[30] As suggested in the International Transporter Consortium workflow on BSEP inhibition in drug discovery, we used a half-maximal inhibitory concentration (IC_{50}) cut-off of 25 μM,[4] designating compounds with IC_{50} values <25 μM as inhibitors and >100 μM as non-inhibitors of BSEP. We excluded compounds with IC_{50} values between 25 and 100 μM from our analysis. Our final dataset consisted of 925 compounds with 152 BSEP inhibitors and 773 non-inhibitors (**Table S1, Supporting Information**).

2.2. Molecular properties, chemical space, and scaffold analysis

In order to understand the chemical space associated with BSEP inhibitors, we used Pipeline Pilot and calculated seven physicochemical properties, namely, molecular weight, log of the octanol/water partition coefficient (AlogP), number of rings, number of rotatable bonds, number of hydrogen bond acceptors, number of hydrogen bond donors, and molecular polar surface area. We used R statistical software[31] to generate box plots and compared the differences in distribution of physiochemical properties between the BSEP inhibitors and non-inhibitors. We evaluated the chemical space of each BSEP inhibitor by comparing it to the chemical space of approved drugs using principal component analysis. Specifically, we collected 1,150 approved

drugs from DrugBank and calculated the same seven physiochemical properties as above.[32] We then used R package prcomp to perform the principal component analysis.[33] In addition, we used the Pipeline Pilot component "scaffold frequency analysis" to identify the frequently occurring scaffolds among the collected BSEP inhibitors. This tool calculates Bemis-Murcko scaffolds and calculates their frequency in the given dataset.[34]

2.3. Model building

2.3.1. Variable-nearest neighbor models

The variable-nearest neighbor (v-NN) method, a variant of the k-nearest neighbor (k-NN) method, is widely used to develop quantitative structure–activity relationship (QSAR) models[17, 35-37] and addresses the limitation associated with the k-NN method by using a structural similarity criterion.[35] The predicted biological activity (y) is a weighted average across structurally similar neighbors:

$$y = \frac{\sum_{i=1}^{v} y_i e^{-\left(\frac{d_i}{h}\right)^2}}{\sum_{i=1}^{v} e^{-\left(\frac{d_i}{h}\right)^2}}, \qquad d_i \leq d_0, \tag{1}$$

where d_i denotes the Tanimoto distance between a query molecule for which a prediction is made and a molecule i of the training set, y_i represents the experimentally measured activity value of molecule i, v denotes the total number of molecules in the training set that satisfy the condition $d_i \leq d_0$, h represents a smoothing factor which dampens the distance penalty, and d_0 denotes a Tanimoto-distance threshold beyond which two molecules are no longer considered to be sufficiently similar to be included in the average. We set the y_i value to 1 for predicting BSEP

inhibitors and to 0 for predicting non-inhibitors. The v-NN method has two adjustable parameters that influence performance: the Tanimoto-distance threshold d_0 and the smoothing factor *h*. In order to enable comparison across different machine learning approaches, we set the Tanimoto-distance threshold d_0 to 1. To identify structurally similar compounds, we used RDKit Morgan circular fingerprints with a radius of two chemical bonds.[38] We implemented the v-NN model development framework as KNIME pipeline, and it is available on a public web server as a vNN-absorption, distribution, metabolism, excretion, toxicity (ADMET) platform (https://vnnadmet.bhsai.org/vnnadmet/).[36]

2.3.2. Bayesian models

We used Naïve Bayes learner node in KNIME for building Bayesian classification models,[39] another popular QSAR approach widely used in ADMET studies.[40-42] Details of the Bayesian classifier approach have been described earlier.[16] Briefly, this approach uses Bayes' theorem and a "learn-by-example" model to predict the likelihood that a given compound is active.[16] It calculates the frequency of occurrence of each molecular feature among the inhibitors compared with all compounds in the dataset, and generates as output a Laplacian-adjusted probability estimate, which provides the likelihood of compounds being from the inhibitor set.[16] The KNIME protocol uses a RDKit Morgan circular fingerprint with a radius of two chemical bonds as an input feature for this model.[38]

2.3.3. Graph convolutional neural network models

2.3.3.1. GCNN single-task models

We used the publicly available ChemProp program to develop our GCNN models.[24] ChemProp is an open-source Python software package that uses a directed message passing neural network

(D-MPNN) to generate molecular descriptors.[24] ChemProp operates in two steps, namely, the message passing phase (graph encoder) and the readout phase (feed-forward neural network).[24] The structure of the compounds in the training set is learned in the message passing phase, and the activity prediction is performed in the readout phase. This approach differs from fingerprint-based approaches in that the molecular representation is learned by the program automatically and need not be pre-defined as in the case of chemical fingerprints. D-MPNN treats the structure of the molecule as a graph, where each atom is a node and each bond is an edge. These nodes and edges have associated feature vectors representing the identity of the respective atoms and bonds. D-MPNN iteratively updates the associated features based on the neighboring node/edge information in each convolution operation and finally creates the learned representation of the compound using a built-in aggregation function that collects the final updated atom-level and bond-level features.[43] The feed-forward neural network uses this learned representation as the input feature vector and predicts the activity of the compound.

The program takes a list of simplified molecular-input line-entry system (SMILES) and associated activity values, given as 1 and 0 for active and inactive molecule, respectively, in csv format as input. In this work, we chose "classification" as the modeling type and used five-fold cross-validation for the "number of folds" option. We developed 10 ensemble models in each run and used 30 epochs. We used the default values for the rest of the model development parameters: a depth value of three, i.e., the number of message passing steps in D-MPNN, the ReLu activation function, 300 hidden neurons, and two layers for the feed-forward neural network. We created the single-task GCNN models for BSEP activity data only.

2.3.3.2. GCNN multi-task models

We also used ChemProp to run multi-task learning models, where one neural network is used to make predictions for multiple properties at the same time. To perform multi-task learning, we first collected two additional datasets associated with BBB permeability and hERG. We used the BBB data provided by Roy et al.[44] and the hERG data provided by Schyman et al.[36] We pre-processed the datasets and removed duplicates and inconsistent compounds. After pre-processing the data, we had 3,439 compounds for BBB (2,489 actives and 950 inactives) and 645 compounds for hERG (271 actives and 374 inactives). In addition to comparing the performance between single-task learning and multi-task learning GCNNs, we also wanted to evaluate the utility of multi-task learning in situations where there are limited activity data available for training. In order to perform this analysis, we randomly split the data 10 times into training and test sets. Each of the training sets was further split into multiple smaller subsets (10%, 20%, 30%, 50%, and 80%). Then, we developed multi- and single-tasks models and evaluated model performance on the same test set from the corresponding initial split.

Next, we evaluated the influence of additional datasets on multi-task learning performance. For this, we collected two additional bioactivity datasets associated with 3-phosphoinositide dependent kinase-1 (PDK-1) and HIVpro from public databases. We collected PDK-1 bioactivity data from BindingDB[28] (accessed on 8-11-2022) and HIVpro data from ChEMBL (accessed on 8-11-2022).[45] We pre-processed the datasets and removed duplicates and inconsistent compounds. After pre-processing the data, we had 952 compounds for PDK-1 (788 actives and 164 inactives) and 2,427 compounds for HIVpro (2,129 actives and 298 inactives).

2.4. Performance evaluation

We carried out five-fold cross-validation as well as a common external test set-based validation to validate and compare model performance. In the five-fold cross-validation procedure, we split the dataset into five groups and left one group out; subsequently, we used the model built from the compounds in the remaining four groups to predict the compounds in the left-out group. Once we completed this prediction cycle by leaving out each of the five groups, we calculated the model evaluation parameter, the receiver operating characteristic (ROC) area under the curve (AUC). To compare different models' performance using the same external test set, we randomly split the data into training and test sets. We used the training set to develop the model using different approaches, such as v-NN, Bayesian, and GCNN, and calculated model performance based on their activity prediction on the common external test set. We used an R script to calculate performance evaluation parameters using a common external test set. We calculated the following metrics: Matthews correlation coefficient (MCC); sensitivity (also known as the recall or true positive rate), the ability to correctly predict positive results; specificity (also known as the true negative rate), the ability to correctly predict negative results; and accuracy, the total percentage correctly predicted. These parameters are defined as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad , \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad , \tag{3}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad . \tag{4}$$

where TP refers to true positive, TN to true negative, FP to false positive, and FN to false negative. We generated ROC and precision-recall (PR) curves and calculated the ROC-AUC and PR-AUC, respectively.

3. RESULTS AND DISCUSSION

BSEP is the key hepatic transporter for efflux of bile salts into bile fluid, and inhibition of this transporter leads to accumulation of bile acids within the hepatocytes, leading to cell death and eventually resulting in cholestasis and liver injury. Earlier prediction of BSEP inhibition potential during the drug discovery process can help screen out drugs with potential liver toxicity liabilities. In this work, we utilized publicly available BSEP data and developed GCNN models that can predict the potential of a chemical to inhibit BSEP.

3.1. Analysis of molecular properties and chemical space of BSEP inhibitors

After pre-processing the data and removing duplicates, we obtained a final dataset of 925 compounds with 152 BSEP inhibitors and 773 non-inhibitors (**Table S1, Supporting Information**). We evaluated the variation between the BSEP inhibitors and non-inhibitors in terms of the following seven physiochemical properties: molecular weight, AlogP, number of rings, number of rotatable bonds, polar surface area, and hydrogen bond acceptors and donors. **Figure 2** and **Figure S1** (**Supporting Information**) show boxplots with the medians and quartiles of these seven properties. We used the non-parametric Mood's median test to evaluate whether the median for each physiochemical property of BSEP inhibitors and non-inhibitors was significantly different. Our analysis showed that, with the exception of hydrogen bond donors ($p > 0.1$), all other properties were significantly different between BSEP inhibitors and non-inhibitors. For example, the mean and median molecular weights of BSEP inhibitors were 522 and 453, respectively, whereas those for BSEP non-inhibitors were 337 and 301 (**Figure 2** and **Table S2**, **Supporting Information**). This result agrees with the previous work of Pedersen et

al., which reported significant differences between BSEP inhibitors and non-inhibitors in terms of lipophilicity/hydrophobicity and size.[46]

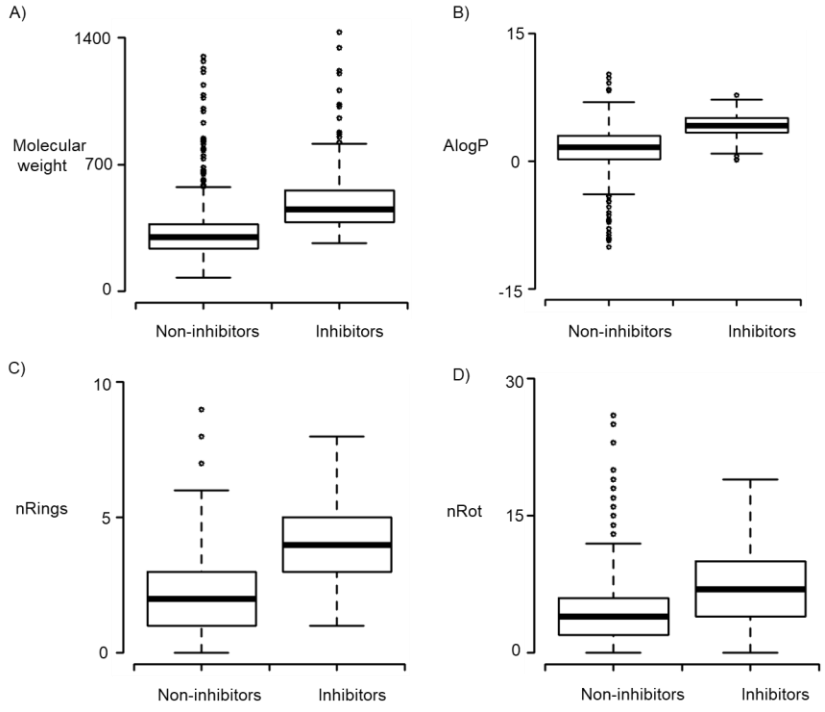

Figure 2. Boxplots showing the distribution of four molecular properties among bile salt export pump (BSEP) inhibitors and non-inhibitors. A) Molecular weight. B) Log of the octanol/water partition coefficient (AlogP). C) Number of rings (nRings). D) Number of rotatable bonds (nRot).

We utilized these seven physiochemical properties and evaluated the mapping of the chemical space associated with BSEP inhibitors with respect to that of approved drugs. **Figure 3A** shows the comparison of the chemical space of these known BSEP inhibitors with those of approved drugs, indicating that most of the BSEP inhibitors occupy a similar chemical space as that of approved drugs. This demonstrates that many BSEP inhibitors have drug-like properties, and utilization of simple physiochemical properties will not be able to separate BSEP inhibitors from

other compounds when screening drug-like compound databases. The large overlap of chemical spaces between these compound classes supports the need to develop machine learning models that capture structural features that differentiate these two classes. We evaluated the frequently occurring scaffolds among the known BSEP inhibitors. We found that scaffolds, such as dihydropyridines, dihydropyrans, piperazines, pyrazoles, and tetradecahydro-cyclopentaphenanthren-3-one, are more frequently found among BSEP inhibitors (**Figure 3B**).

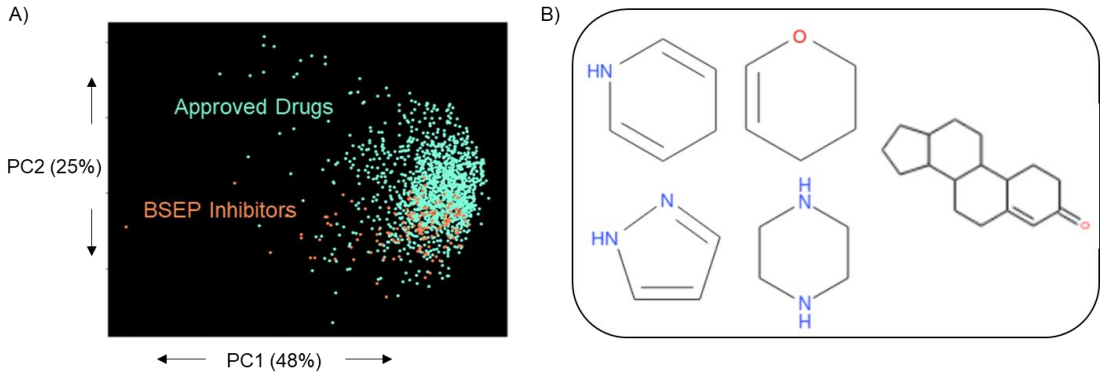

Figure 3. A) Chemical space analysis of bile salt export pump (BSEP) inhibitors (brown) compared with approved drugs (cyan) shows overlap of both datasets. B) Frequently occurring scaffolds among the known BSEP inhibitors.

3.2. Comparison of graph convolutional neural network and other machine learning models

Machine learning models are an integral part of the drug discovery process and are widely used for predicting various activity endpoints. Our group has developed many predictive models for various ADMET endpoints using v-NN and Bayesian approaches.[16, 35-37] These traditional machine learning approaches typically use fingerprints to represent the structure of the compounds. In particular, circular fingerprints are considered the *de facto* standard in

representing chemical structures for developing machine learning models.[16] This approach captures the presence of pre-defined sets of chemical substructures/functional groups to represent the chemical structure. More recently, GCNN-based approaches have provided an alternative to the traditional fingerprint approach and allowed us to learn the chemical structures in an automated manner.[24] The learned representation of the molecule can then be used with a feed-forward neural network to predict the activity of the compounds. Yang et al. have performed a detailed analysis of benchmark datasets and showed that GCNNs perform better than other machine learning approaches for predicting the properties of compounds.[24] GCNNs were successfully used to discover new compounds with anti-bacterial activity as well as to predict various ADMET endpoints.[25, 47] So far, the utility of GCNNs in predicting BSEP inhibition potential of chemicals has not been studied. In this work, we focused on developing a GCNN model for BSEP using the open-source ChemProp program.

First, we evaluated three different machine learning approaches, i.e., v-NN, Bayesian, and GCNN. The first two methods use standard circular fingerprints to represent chemical structures, and the GCNN method employs graph-derived automatically learned molecular representation of the chemical structures. We carried out a five-fold cross-validation analysis using the three approaches. **Figure 4** shows that the GCNN approach had the best overall performance in terms of ROC-AUC compared to either the Bayesian or v-NN approach. We also evaluated the predictive ability of the model using a common external test set and calculated the ROC-AUC for the three approaches (**Figure S2**, **Supporting Information**).

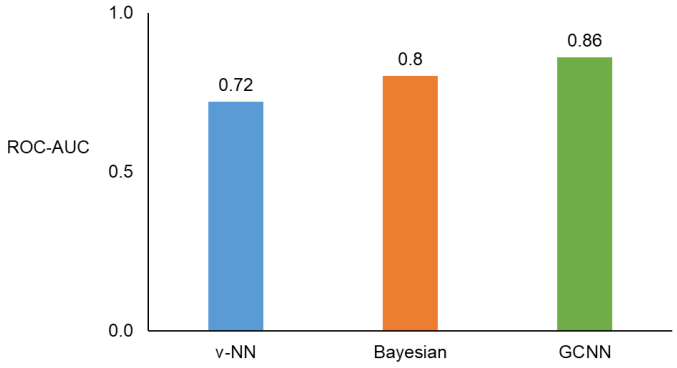

Figure 4. Performance of graph convolutional neural network (GCNN), Bayesian, and variable-nearest neighbor (v-NN) models for bile salt export pump (BSEP) data. The GCNN model performed better than the other two approaches. ROC-AUC, receiver operating characteristic area under the curve.

3.3. Analysis of single-task and multi-task learning GCNN models

GCNNs can be used to develop single- or multi-task learning models.[48] In the above model comparison analysis, we used only the BSEP data and developed the prediction model, i.e., single task. Previous reports have shown that multi-task models perform better than single-task models.[26, 27] Although deep learning/GCNN approaches are successfully used in many real-world problems, such as image classification, voice recognition, and self-driving cars, their utility is limited in biomedical and drug discovery problems.[49, 50] The main reason for this is limited data availability. For example, organic anion transporter-1 is another transporter similar to BSEP that is recommended for evaluation during the drug discovery process by regulatory agencies,[51] but a search of ChEMBL showed that only 50 compounds with bioactivity values are available for this transporter. This impedes the development of machine learning models for this transporter and poses a significant challenge for developing computational profilers/tools. Providing a possible

path forward, multi-task models have been reported to be useful in addressing such data limitation challenges.[52]

In this work, in addition to comparing single- and multi-task BSEP models, we extensively evaluated the utility of the multi-task model approach to address the data limitation challenge. To this end, we included related targets with sufficient data and developed multi-task models, using BSEP data together with BBB permeability and hERG inhibition data. BBB permeability involves multiple transporters, and hERG is a channel located in the cell membrane. We performed 10 random splits of the data into training and test sets to make sure that model performance was not influenced by the data composition in random splits. Each of the training sets was further split into multiple smaller subsets (10%, 20%, 30%, 50%, and 80%). We then developed multi-task and single-tasks models and evaluated model performance on the same external test set from the corresponding initial split (**Figure 5**). We developed 10 ensemble models for each training dataset and optimized parameters with a five-fold cross-validation. Overall, we generated 6,000 GCNN models as part of this analysis. The average ROC-AUC for the multi-task models was higher than for the single-task models when the data availability was below 80% (**Figure 6**). The average PR-AUC values for multi-task models with 10%, 20%, 30%, 50%, and 80% subsets were 0.38, 0.43, 0.42, 0.55, and 0.63, respectively, and for single-task models were 0.16, 0.27, 0.32, 0.39, and 0.55, respectively (**Figure 6**). Similarly, the average MCC value for the multi-task models was consistently better than for the single-task models for all subsets of training data (**Figure 6**). Our results show that multi-task models consistently performed better than single-task models even with limited data availability.

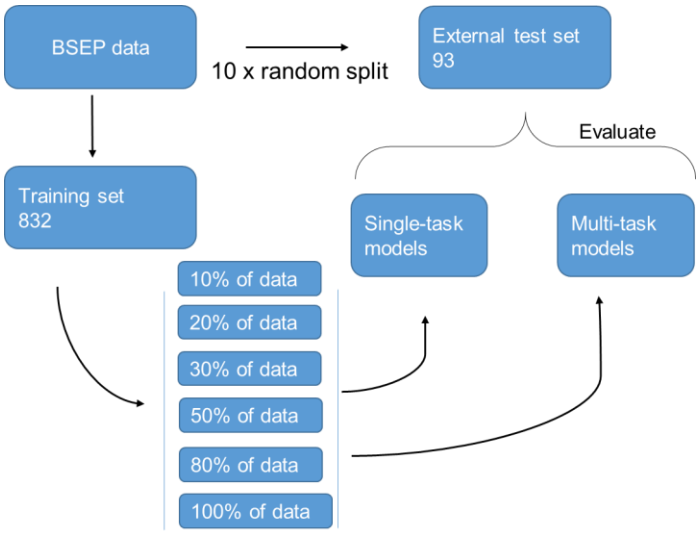

Figure 5. Schema for single-task and multi-task model comparison. The bile salt export pump (BSEP) data were split into training and external test sets. The training set was further split into six subsets of data from 10% to 100%. Each of the subsets was used to develop single-task and multi-task models. Both models were evaluated using the same common external test created in the first step of the process.

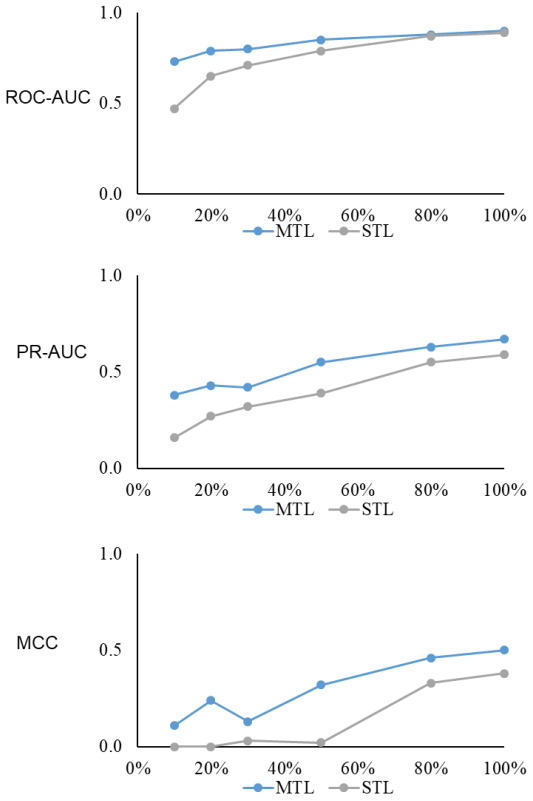

Figure 6. Comparison of single-task learning (STL) and multi-task learning (MTL) models using various subsets of bile salt export pump (BSEP) data. We evaluated model performance using 10%, 20%, 30%, 50%, 80%, and 100% of the training data and calculated the receiver operating characteristic area under the curve (ROC-AUC), precision-recall area under the curve (PR-AUC), and Matthews correlation coefficient (MCC).

Next, we evaluated the influence of additional datasets on the performance of the multi-task models. We collected two additional bioactivity datasets associated with PDK-1 and HIVpro. These two proteins were selected as they belong to a different enzyme class and represent a non-human target. After pre-processing the data, we had 952 compounds for PDK-1 and 2,427 compounds for HIVpro. We developed single-task and multi-task models and evaluated their performance. For the multi-task models, we tested the influence of different additional datasets

along with BSEP (BBB+hERG, PDK-1+HIVpro, and BBB+hERG+PDK-1+HIVpro). We carried out extensive analyses as described above, repeating the process with 10 random splits, and developed 10 ensemble models for each training dataset. We evaluated model performance using ROC-AUC, PR-AUC, and MCC. We found that multi-task models consistently performed better than single-task models, and additional datasets did not decrease the performance of the multi-task models (**Figure 7**). With an average PR-AUC of 71%, the combined dataset (BBB+hERG+PDK-1+HIVpro) had the best performance, as compared to an average PR-AUC of 55% for single-task models and an average PR-AUC of 64% for other combinations (**Figure 7**). Our results show that even if the additional data are from an unrelated protein, such as PDK-1, they can help improve multi-task model performance as they can be used to identify more inactive compounds in the dataset. We further searched through the literature and found that HIVpro inhibitors are indeed known to have cross-reactivity/inhibition of human transporter proteins, including BSEP.[53] This further provides a rationale for the improved performance of our final combined model using this additional dataset.

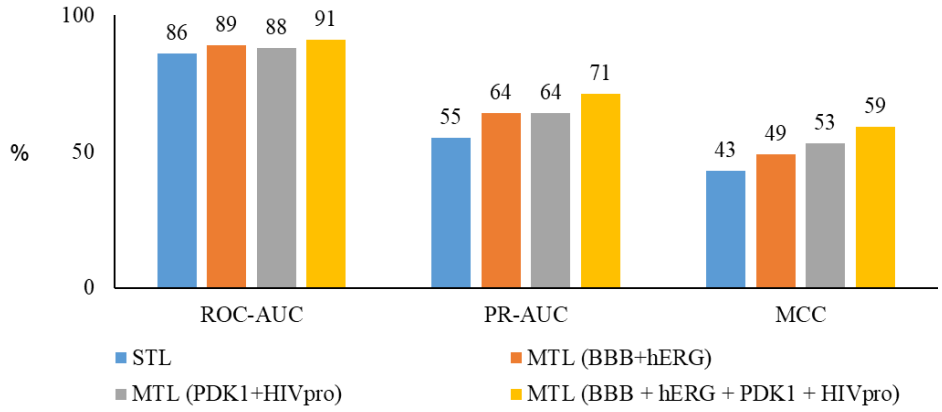

Figure 7. Comparison of the predictive performance of single-task learning (STL) and different multi-task learning (MTL) models. We developed bile salt export pump (BSEP) MTL models using different additional datasets, including blood-brain barrier (BBB) permeability, human ether-à-go-go-related gene (hERG) inhibition, 3-phosphoinositide dependent kinase-1 (PDK-1) inhibition, and human immunodeficiency virus protease (HIVpro) inhibition. We found that the multi-task models performed better than the single-task models. Of the multi-task models, the model developed with all four datasets (BBB+hERG+PDK1+HIVpro) performed better than those developed with the other two datasets. ROC-AUC, area under the receiver operating characteristic curve; PR-AUC, area under the precision-recall curve; MCC, Matthews correlation coefficient.

Overall, we developed a predictive, multi-task GCNN model for BSEP that can be used for screening of large chemical databases. It should be noted that the utilization of public data enables the development of an openly available model, but it may not be comprehensive of the available chemical space, which could limit its utility. Additionally, while the model has shown promising performance in cross-validation, its real-world predictive accuracy may be influenced by other factors not accounted for in the model, such as drug metabolism or species differences.

Understanding these limitations will ensure the proper interpretation and use of the model in drug discovery and risk assessment.

4. CONCLUSIONS

In this work, we developed a GCNN model for predicting the potential of a chemical to inhibit BSEP, an important transporter that plays a role in drug-induced liver injury. Our analysis of BSEP's molecular properties showed that BSEP inhibitors are more hydrophobic and have a larger molecular weight compared to non-inhibitors and occupy a similar chemical space as that of approved drugs. We found that our GCNN model performed better than the other machine learning approaches we evaluated, and we showed that multi-task models consistently performed better than single-task models. Specifically, we showed the utility of multi-task learning models to address data limitation challenges. Thus, we developed multi-task GCNN-based BSEP models that allow for a rapid computational screening of liver safety risk assessment during the early drug discovery stages.

FUNDING

This work was supported by Defense Threat Reduction Agency Grant CBCall14-CBS-05-2-0007. The Henry M. Jackson Foundation was supported by the U.S. Army Medical Research and Development Command under Contract No. W81XWH20C0031.

ACKNOWLEDGMENT

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army, the U.S. Department of Defense, or The Henry M. Jackson Foundation for Advancement of Military Medicine, Inc. This paper has been approved for public release with unlimited distribution. The authors gratefully acknowledge the assistance of Zhen Xu in creating KNIME workflows and Ms. Maria Kuhrmann in editing the manuscript.

AUTHOR CONTRIBUTIONS

MDMA, RL, and AW designed the study. Computations and data analysis were performed by MDMA. The first draft of the manuscript was written by MDMA. All authors commented on previous versions of the manuscript and approved the final manuscript.

SUPPORTING INFORMATION

Table S1: BSEP data of 925 compounds with SMILES and activity annotation.

Table S2: Mean values of seven molecular properties for BSEP inhibitors and non-inhibitors.

Figure S1: Boxplots showing the distribution of three molecular properties among BSEP non-inhibitors and inhibitors.

Figure S2. Performance of graph convolutional neural network, Bayesian, and variable-nearest neighbor models for BSEP data using a common external test set for all three approaches.

DATA/CODE AVAILABILITY

The datasets (.csv files) used in this study and the Python and R codes used in this study are freely available at https://github.com/BHSAI/BSEP_GCNN_model

REFERENCES

(1) Sohail, M. I.; Donmez-Cakil, Y.; Szollosi, D.; Stockner, T.; Chiba, P. The Bile Salt Export Pump: Molecular Structure, Study Models and Small-Molecule Drugs for the Treatment of Inherited BSEP Deficiencies. *Int J Mol Sci* **2021**, *22* (2).

(2) Kunst, R. F.; Verkade, H. J.; Oude Elferink, R. P. J.; van de Graaf, S. F. J. Targeting the Four Pillars of Enterohepatic Bile Salt Cycling; Lessons From Genetics and Pharmacology. *Hepatology* **2021**, *73* (6), 2577-2585.

(3) Goutam, K.; Ielasi, F. S.; Pardon, E.; Steyaert, J.; Reyes, N. Structural basis of sodium-dependent bile salt uptake into the liver. *Nature* **2022**, *606* (7916), 1015-1020.

(4) Kenna, J. G.; Taskar, K. S.; Battista, C.; Bourdet, D. L.; Brouwer, K. L. R.; Brouwer, K. R.; Dai, D.; Funk, C.; Hafey, M. J.; Lai, Y.; et al. Can Bile Salt Export Pump Inhibition Testing in Drug Discovery and Development Reduce Liver Injury Risk? An International Transporter Consortium Perspective. *Clin Pharmacol Ther* **2018**, *104* (5), 916-932.

(5) Vinken, M.; Landesmann, B.; Goumenou, M.; Vinken, S.; Shah, I.; Jaeschke, H.; Willett, C.; Whelan, M.; Rogiers, V. Development of an adverse outcome pathway from drug-mediated bile salt export pump inhibition to cholestatic liver injury. *Toxicol Sci* **2013**, *136* (1), 97-106.

(6) Liu, H.; Sahi, J. Role of Hepatic Drug Transporters in Drug Development. *J Clin Pharmacol* **2016**, *56 Suppl 7*, S11-22.

(7) Hofmann, A. F. The continuing importance of bile acids in liver and intestinal disease. *Arch Intern Med* **1999**, *159* (22), 2647-2658.

(8) Li, T.; Apte, U. Bile Acid Metabolism and Signaling in Cholestasis, Inflammation, and Cancer. *Adv Pharmacol* **2015**, *74*, 263-302.

(9) Faubion, W. A.; Guicciardi, M. E.; Miyoshi, H.; Bronk, S. F.; Roberts, P. J.; Svingen, P. A.; Kaufmann, S. H.; Gores, G. J. Toxic bile salts induce rodent hepatocyte apoptosis via direct activation of Fas. *J Clin Invest* **1999**, *103* (1), 137-145.

(10) Yang, K.; Kock, K.; Sedykh, A.; Tropsha, A.; Brouwer, K. L. An updated review on drug-induced cholestasis: mechanisms and investigation of physicochemical properties and pharmacokinetic parameters. *J Pharm Sci* **2013**, *102* (9), 3037-3057.

(11) Morgan, R. E.; Trauner, M.; van Staden, C. J.; Lee, P. H.; Ramachandran, B.; Eschenberg, M.; Afshari, C. A.; Qualls, C. W., Jr.; Lightfoot-Dunn, R.; Hamadeh, H. K. Interference with bile salt export pump function is a susceptibility factor for human liver injury in drug development. *Toxicol Sci* **2010**, *118* (2), 485-500.

(12) Jain, S.; Grandits, M.; Richter, L.; Ecker, G. F. Structure based classification for bile salt export pump (BSEP) inhibitors using comparative structural modeling of human BSEP. *J Comput Aided Mol Des* **2017**, *31* (6), 507-521.

(13) Kenna, J. G. Current Concepts in Drug-Induced Bile Salt Export Pump (BSEP) Interference. *Curr Protoc Toxicol* **2014**, *61*, 23 27 21-15.

(14) McLoughlin, K. S.; Jeong, C. G.; Sweitzer, T. D.; Minnich, A. J.; Tse, M. J.; Bennion, B. J.; Allen, J. E.; Calad-Thomson, S.; Rush, T. S.; Brase, J. M. Machine Learning Models to Predict Inhibition of the Bile Salt Export Pump. *J Chem Inf Model* **2021**, *61* (2), 587-602.

(15) AbdulHameed, M. D.; Chaudhury, S.; Singh, N.; Sun, H.; Wallqvist, A.; Tawa, G. J. Exploring polypharmacology using a ROCS-based target fishing approach. *J Chem Inf Model* **2012**, *52* (2), 492-505.

(16) AbdulHameed, M. D.; Ippolito, D. L.; Wallqvist, A. Predicting Rat and Human Pregnane X Receptor Activators Using Bayesian Classification Models. *Chem Res Toxicol* **2016**, *29* (10), 1729-1740.

(17) Schyman, P.; Liu, R.; Wallqvist, A. Using the Variable-Nearest Neighbor Method To Identify P-Glycoprotein Substrates and Inhibitors. *ACS Omega* **2016**, *1* (5), 923-929.

(18) Liu, R.; Madore, M.; Glover, K. P.; Feasel, M. G.; Wallqvist, A. Assessing Deep and Shallow Learning Methods for Quantitative Prediction of Acute Chemical Toxicity. *Toxicol Sci* **2018**, *164* (2), 512-526.

(19) Hirano, H.; Kurata, A.; Onishi, Y.; Sakurai, A.; Saito, H.; Nakagawa, H.; Nagakura, M.; Tarui, S.; Kanamori, Y.; Kitajima, M.; et al. High-speed screening and QSAR analysis of human ATP-binding cassette transporter ABCB11 (bile salt export pump) to predict drug-induced intrahepatic cholestasis. *Mol Pharm* **2006**, *3* (3), 252-265.

(20) Warner, D. J.; Chen, H.; Cantin, L. D.; Kenna, J. G.; Stahl, S.; Walker, C. L.; Noeske, T. Mitigating the inhibition of human bile salt export pump by drugs: opportunities provided by physicochemical property modulation, in silico modeling, and structural modification. *Drug Metab Dispos* **2012**, *40* (12), 2332-2341.

(21) Montanari, F.; Pinto, M.; Khunweeraphong, N.; Wlcek, K.; Sohail, M. I.; Noeske, T.; Boyer, S.; Chiba, P.; Stieger, B.; Kuchler, K.; et al. Flagging Drugs That Inhibit the Bile Salt Export Pump. *Mol Pharm* **2016**, *13* (1), 163-171.

(22) Rodríguez-Pérez, R.; Gerebtzoff, G. Identification of bile salt export pump inhibitors using machine learning: Predictive safety from an industry perspective. *Artificial Intelligence in the Life Sciences* **2021**, *1*, 100027.

(23) Kotsampasakou, E.; Ecker, G. F. Predicting Drug-Induced Cholestasis with the Help of Hepatic Transporters-An in Silico Modeling Approach. *J Chem Inf Model* **2017**, *57* (3), 608-615.

(24) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing Learned Molecular Representations for Property Prediction. *J Chem Inf Model* **2019**, *59* (8), 3370-3388.

(25) Liu, K.; Sun, X.; Jia, L.; Ma, J.; Xing, H.; Wu, J.; Gao, H.; Sun, Y.; Boulnois, F.; Fan, J. Chemi-Net: A Molecular Graph Convolutional Network for Accurate Drug Property Prediction. *Int J Mol Sci* **2019**, *20* (14).

(26) Feinberg, E. N.; Joshi, E.; Pande, V. S.; Cheng, A. C. Improvement in ADMET Prediction with Multitask Deep Featurization. *J Med Chem* **2020**, *63* (16), 8835-8848.

(27) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure-Activity Relationships. *J Chem Inf Model* **2017**, *57* (10), 2490-2504.

(28) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* **2016**, *44* (D1), D1045-1053.

(29) Hassan, M.; Brown, R. D.; Varma-O'brien, S.; Rogers, D. Cheminformatics analysis and learning in a data pipelining environment. *Mol Divers* **2006**, *10* (3), 283-299.

(30) AbdulHameed, M. D. M.; Liu, R.; Schyman, P.; Sachs, D.; Xu, Z.; Desai, V.; Wallqvist, A. ToxProfiler: Toxicity-target profiler based on chemical similarity. *Computational Toxicology* **2021**, *18*, 100162.

(31) *R: A Language and Environment for Statistical Computing*; Vienna, Austria, 2022. https://www.R-project.org/ (accessed.

(32) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **2018**, *46* (D1), D1074-D1082.

(33) Venables, W. N.; Ripley, B. D.; Venables, W. N. *Modern applied statistics with S / W.N. Venables, B.D. Ripley*; Springer, 2002.

(34) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J Med Chem* **1996**, *39* (15), 2887-2893.

(35) Liu, R.; Tawa, G.; Wallqvist, A. Locally weighted learning methods for predicting dose-dependent toxicity with application to the human maximum recommended daily dose. *Chem Res Toxicol* **2012**, *25* (10), 2216-2226.

(36) Schyman, P.; Liu, R.; Desai, V.; Wallqvist, A. vNN Web Server for ADMET Predictions. *Front Pharmacol* **2017**, *8*, 889.

(37) Liu, R.; Schyman, P.; Wallqvist, A. Critically Assessing the Predictive Power of QSAR Models for Human Liver Microsomal Stability. *J Chem Inf Model* **2015**, *55* (8), 1566-1575.

(38) *RDKit: Open-source cheminformatics.*; https://www.rdkit.org (accessed.

(39) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter* **2009**, *11* (1), 26-31.

(40) Clark, A. M.; Dole, K.; Coulon-Spektor, A.; McNutt, A.; Grass, G.; Freundlich, J. S.; Reynolds, R. C.; Ekins, S. Open Source Bayesian Models. 1. Application to ADME/Tox and Drug Discovery Datasets. *J Chem Inf Model* **2015**, *55* (6), 1231-1245.

(41) Ekins, S.; Reynolds, R. C.; Kim, H.; Koo, M. S.; Ekonomidis, M.; Talaue, M.; Paget, S. D.; Woolhiser, L. K.; Lenaerts, A. J.; Bunin, B. A.; et al. Bayesian models leveraging bioactivity and cytotoxicity information for drug discovery. *Chem Biol* **2013**, *20* (3), 370-378.

(42) Singh, N.; Chaudhury, S.; Liu, R.; AbdulHameed, M. D.; Tawa, G.; Wallqvist, A. QSAR classification model for antibacterial compounds and its use in virtual screening. *J Chem Inf Model* **2012**, *52* (10), 2559-2569.

(43) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J Chem Inf Model* **2022**, *62* (9), 2101-2110.

(44) Roy, D.; Hinge, V. K.; Kovalenko, A. To Pass or Not To Pass: Predicting the Blood-Brain Barrier Permeability with the 3D-RISM-KH Molecular Solvation Theory. *ACS Omega* **2019**, *4* (16), 16774-16780.

(45) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* **2019**, *47* (D1), D930-D940.

(46) Pedersen, J. M.; Matsson, P.; Bergstrom, C. A.; Hoogstraate, J.; Noren, A.; LeCluyse, E. L.; Artursson, P. Early identification of clinically relevant drug interactions with the human bile salt export pump (BSEP/ABCB11). *Toxicol Sci* **2013**, *136* (2), 328-343.

(47) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *181* (2), 475-483.

(48) Kojima, R.; Ishida, S.; Ohta, M.; Iwata, H.; Honma, T.; Okuno, Y. kGCN: a graph-based deep learning framework for chemical structures. *J Cheminform* **2020**, *12* (1), 32.

(49) Liu, R.; Laxminarayan, S.; Reifman, J.; Wallqvist, A. Enabling data-limited chemical bioactivity predictions through deep neural network transfer learning. *J Comput Aided Mol Des* **2022**, *36* (12), 867-878.

(50) Peterson, J. C.; Uddenberg, S.; Griffiths, T. L.; Todorov, A.; Suchow, J. W. Deep models of superficial face judgments. *Proc Natl Acad Sci U S A* **2022**, *119* (17), e2115228119.

(51) Britz, H.; Hanke, N.; Taub, M. E.; Wang, T.; Prasad, B.; Fernandez, E.; Stopfer, P.; Nock, V.; Lehr, T. Physiologically Based Pharmacokinetic Models of Probenecid and Furosemide to Predict Transporter Mediated Drug-Drug Interactions. *Pharm Res* **2020**, *37* (12), 250.

(52) Cang, Z.; Wei, G. W. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput Biol* **2017**, *13* (7), e1005690.

(53) Griffin, L.; Annaert, P.; Brouwer, K. L. Influence of drug transport proteins on the pharmacokinetics and drug interactions of HIV protease inhibitors. *J Pharm Sci* **2011**, *100* (9), 3636-3654.

TOC GRAPHIC