JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 1253−1268

pubs.acs.org/jcim

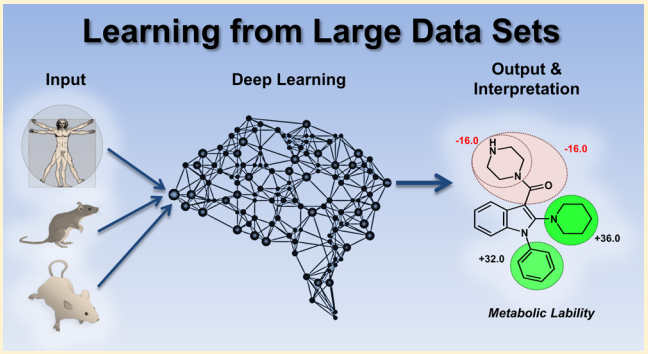
Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets

Jan Wenzel,*[,†] Hans Matter,[‡] and Friedemann Schmidt[†]

Sanofi-Aventis Deutschland GmbH, [†]R&D, Preclinical Safety, Global Operations Germany and [‡]R&D, Integrated Drug Discovery, Industriepark Höchst, 65926 Frankfurt am Main, Germany

Supporting Information

ABSTRACT: Successful drug discovery projects require control and optimization of compound properties related to pharmacokinetics, pharmacodynamics, and safety. While volume and chemotype coverage of public and corporate ADME-Tox (absorption, distribution, excretion, metabolism, and toxicity) databases are constantly growing, deep neural nets (DNN) emerged as transformative artificial intelligence technology to analyze those challenging data. Relevant features are automatically identified, while appropriate data can also be combined to multitask networks to evaluate hidden trends among multiple ADME-Tox parameters for implicitly correlated data sets. Here we describe a novel, fully industrialized approach to parametrize and optimize the setup, training, application, and visual interpretation of DNNs to model ADME-Tox data. Investigated properties include microsomal lability in different species, passive permeability in Caco-2/TC7 cells, and logD. Statistical models are developed using up to 50 000 compounds from public or corporate databases. Both the choice of DNN hyperparameters and the type and quantity of molecular descriptors were found to be important for successful DNN modeling. Alternate learning of multiple ADME-Tox properties, resulting in a multitask approach, performs statistically superior on most studied data sets in comparison to DNN single-task models and also provides a scalable method to predict ADME-Tox properties from heterogeneous data. For example, predictive quality using external validation sets was improved from R^2 of 0.6 to 0.7 comparing single-task and multitask DNN networks from human metabolic lability data. Besides statistical evaluation, a new visualization approach is introduced to interpret DNN models termed "response map", which is useful to detect local property gradients based on structure fragmentation and derivatization. This method is successfully applied to visualize fragmental contributions to guide further design in drug discovery programs, as illustrated by CRCX3 antagonists and renin inhibitors, respectively.



INTRODUCTION

Optimizing ADME-Tox properties (absorption, distribution, excretion, metabolism, and toxicity) is an integral process in pharmaceutical lead discovery and optimization. Some important parameters of in vivo experiments, such as pharmacokinetics and certain toxic effects, can be estimated early on by these properties. Therefore, these data support decision making, whether a compound with acceptable potency can effectively act as drug candidate in vivo, whether a dose is sufficient for the desired biological effect, and which safety margin has to be considered taking into account pharmacodynamics results.

This essential role of ADME-Tox has led to significant efforts to miniaturize, validate, and harmonize appropriate in vitro assays in pharmaceutical settings. For example, the Caco-2 cell line system often serves as model for passive permeability and active transport. Metabolic stability in different species can be explored using homogenized liver microsomes from human and other species. Novel compounds are routinely tested to identify critical liabilities for prioritizing series and compounds with favorable properties.

While the experimental capacity and the prediction of in vitro ADME-Tox assays is often limited, in silico ADME-Tox models can overcome this shortcoming; in silico models also can be applied to virtual molecules and are important tools for multidimensional compound design.[1−5] A tight integration of in vitro and in silico approaches allows extraction of structure−activity relationships (SAR) for design. Since modifications to lead structures are often incremental for balancing multiple properties, regression-type models producing a continuous prediction are powerful to rank-order (rather than only classify) new proposals. To this end, considerable efforts have been undertaken to develop predictive in silico models using machine learning techniques. For novel chemical series, "global" in silico models are useful, which cover a broad

Special Issue: Machine Learning in Drug Discovery

Received: November 6, 2018
Published: January 7, 2019

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

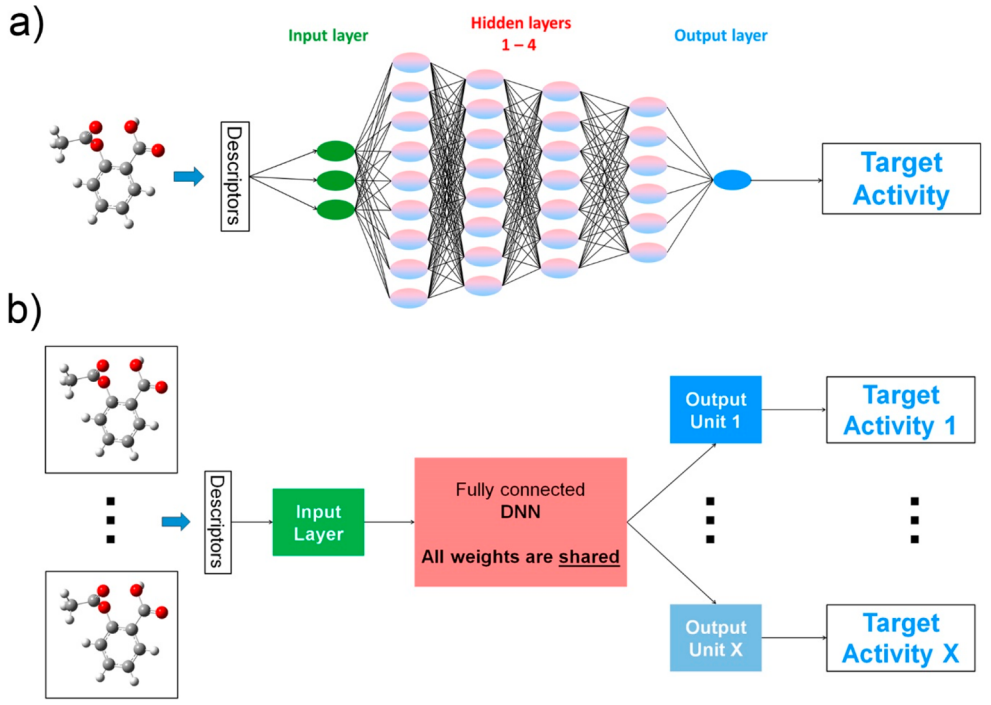


Figure 1. (a) Schematic illustration of a deep neural network with four hidden layers and a single-task output. Starting from molecules given in their 2D-structure, chemical molecular descriptors are computed that are employed as input features. The neurons are sequentially fully connected layer-wise from input units and hidden layers to the final output, which corresponds to a target activity. (b) Schematic illustration of a multitask deep neural net. There are *X* output units corresponding to *X* data sets and targets. Chemical molecular descriptors are computed for every data set and feed into the input layers, respectively. The input units are fully connected to a DNN with several hidden layers. All weights of the hidden layers are shared between the tasks, while the weights corresponding to the output units are unique.

chemical space and are not limited to a single chemotype. Along these lines, global in silico models can be constantly curated, for example, by integrating novel data.[6] Although linear models might produce simpler, interpretable models, it has been appreciated that nonlinear modeling often results in more accurate models.[5,7]

Recently, deep learning[8,9] has emerged as promising addition to machine learning for large data sets with many descriptors. Pioneering work in this area was reported, for example, by Hochreiter and co-workers.[10−15] Deep learning (or deep neural networks, DNNs) denotes complex nonlinear models consisting of neural nets with more than one hidden layer. Improved hardware, advances to avoid overfitting of neural nets, and distribution on widely accessible platforms with improved training algorithms allowed for successful applications of DNNs in artificial intelligence areas, such as computer vision and language processing.[8]

DNNs were also applied in computational chemistry[16,17] and for generating quantitative structure−activity relationship (QSAR) models.[18−20] Key applications in medicinal chemistry have been recently summarized.[21−23] The usage of DNNs for QSAR in industry discovery was pioneered by Sheridan and co-workers.[18−20] They demonstrate that DNNs often result in models with similar or improved predictivity compared to other algorithms such as random forest (RF)[24] on project data for relevant targets and antitargets.[18]

An important finding is that multitask DNN models, trained simultaneously from multiple biological properties, can outperform DNNs trained separately (single-task DNN) for those data sets.[18] Further work by Xu et al. suggested that multitask DNN models incorporate a "signal" from similar structures in training sets of other, sometimes unrelated target models.[20] However, whether transferred information improves or deteriorates the predictivity of a model depends on the correlation of biological properties that were combined. Therefore, the authors suggest a workflow to use multitask DNNs with prior domain knowledge to select training sets with correlated activities.[20]

Korotcov et al.[25] showed that DNN outperforms other machine learning techniques for some ADME-Tox properties. Other groups describe models for cytochrome P450s[26] and toxicity.[13,27]

In this study, we build on previous DNN reports[13,18,20,21,28,29] to develop a scalable, predictive ADME-Tox framework for pharmaceutical compounds using the TensorFlow[30] platform for development of predictive single-task and multitask DNN models for large to medium-sized ADME-Tox data sets. Our multitask DNN approach is founded on alternate training rather than joint training of properties, with the focus to support prospective application. We discuss critical aspects of DNN model building, present our approach for parameter optimization on GPUs, and provide suggestions for hyperparameter range scanning.

To study "gap-filling" capabilities of DNNs, we systematically analyze single-task versus multitask models for cross-species metabolic lability data. Subsequently the combination of mechanistically unrelated data sets using multitask DNN models is explored by combining Caco-2 permeability with either metabolic lability or logD(pH 7.4) data.

Finally we introduce a visualization concept termed "response maps" to serve practical aspects of DNN model interpretation in the framework of drug discovery programs. Our model-independent approach involves the incremental modulation of chemical structures by systematically removing

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

or adding substituents and small side chains. Resulting DNN predictions for a parent molecule and its virtual matched pairs (VMPs) provide an estimation of the impact of a substitution to the predicted property. This intuitive approach follows the work of Polishchuk et al.[31] By adding or changing chemical substituents, we introduce a property "response map" concept for exploring the sensitivity of machine learning models for particular substitutions.

THEORY

This section introduces key aspects of deep learning, while detailed reviews are provided in the literature.[8,9,23,32−34]

Deep Neural Networks. A neural network is based on units called neurons that are interconnected via functions mapping an input vector to an output vector. Hence each neuron receives multiple input vectors and generates a single output vector. The input vectors are associated and controlled with a weight and a bias term, which are optimized in the learning process. Furthermore, each neuron can be associated with an activation function leading to an enhanced learning behavior. The most prominent activation function is the rectified linear unit (relu) function, which has advantageous characteristics for ADME-Tox data such as sparse activation, biological plausibility, mathematical gradient propagation, and efficient computation.[8,18,35]

Deep neural networks consist of different layers (row of neurons): an input layer containing descriptors or input features, an output layer describing the target value predictions, and several intermediate layers (hidden layers) consisting of a few neurons (Figure 1a). Given a raw input (e.g., molecular descriptors), the DNN extracts the feature representations needed for target prediction automatically without human engineering. Hence in deep learning method procedures, raw input data is transformed from layer to layer into more abstract levels leading to the capability to learn very complex functions.

Using training data, the known values are used to establish an error or cost function, E, to measure the current loss of the network. For regression, the mean square error function is commonly employed. To optimize the DNN, one has to minimize E using the backpropagation algorithm.[36] Besides classical stochastic gradient descent (SGD) algorithms, there are many variants available to optimize a DNN. In this context, the SGD with momentum (SGDM) and the per-parameter adaptive rate methods like the Adam or RMSProp algorithms provide accurate results in combination with fast convergence.[18,32,36−40]

A risk of DNNs is overfitting due to the large number of interconnected parameters with respect to training elements. To avoid overfitting, the network complexity can be tuned (less neurons or hidden layers); alternatively the training data may be enriched, for example, by repetitive presentation of samples. Furthermore, regularization techniques like dropout and L2 weight decay have proven helpful to avoid overfitting.[40,41] In order to find the best set of weights before overfitting, it is recommended to apply early stopping. For this purpose, a reasonably sized test data set is essential, which is not used for training. After each epoch, the predictive quality of the network is controlled using the test set. If the test set performance declines, the optimization is stopped, and only the most accurate model based on the test set is kept. Validation of predictive performance is then done on an external validation data set that was kept aside.

Multitask Learning. One advantage of neural nets is their ability to operate on multiple endpoints and data sets simultaneously. Experimental assays or data sets corresponding to different biological endpoints can therefore be combined in a single model to improve model quality by learning shared models with shared feature extraction. Single-task DNNs employ a single output unit corresponding to one endpoint, whereas multitask networks have multiple output units corresponding to each endpoint, respectively. All hidden layers are shared in multitask DNNs, while the individual output and input units specifically correspond to one target data set (Figure 1b). Multitask learning therefore operates partially by "gap filling" of sparsely populated data sets. Abstract features from descriptors are learned in multitask models, but in contrast to single-task models they are transparently shared between individual tasks. Since even complex ADME-Tox properties can exhibit correlations, such a network can learn interconnections between targets.[19,25,28,29,34,42−44]

METHODS

Multitask Training. Training of multitask networks is straightforward. Generally, there are two classes of training algorithms, namely, alternate or joint training. In the first case, gradients are computed only for one output unit using the data set. After a defined number of iterations, the optimizer switches to the next output unit adopting the optimized weights from the shared layers or neurons of the previous step. This cycle can be repeated for a fixed number of iterations. Hence, in alternate training only the weights on shared network layers and the corresponding output unit are updated during each iteration, while all weights of the other output units are kept frozen. In contrast to that, all output units are trained at the same time when employing joint training. The gradient is built on a global joint output unit, which is generated by summation of the current loss functions of all individual output units.[45]

An advantage of alternate training is that different data sets with only small overlap of compounds can be combined. Hence, alternate multitask learning could transfer features from rich data sets to sparse ones without the need to fill all training data matrices. However, if one data set is significantly larger than the others, a predictivity bias towards the larger data sets is to be expected. This bias can be overcome to a certain degree, just by increasing the number of iterations of the smaller data set in a global iteration.

DNN Implementation. Calculations were performed on machines with NVIDIA Tesla M40 and P40 GPU cards and 24 GB GPU memory per card. DNN models were trained using Tensorflow[30] 1.3 and 1.4 in combination with Keras[46] 2.0.8. Single-task models were computed using the tensorflow.contrib.learn environment as well as Keras. For multitask networks, the Keras API has been implemented.

Nodes in multitask DNN models are fully connected, and the network structure is sequentially linear using shared input and hidden layers with task-unique output units. Our implemented training algorithm of multitask DNNs using alternate training works as follows:

1. Setup network structure, optimizer, loss functions, and hyperparameters. For every task, which is related to a data set with a corresponding target/endpoint, a separate Keras model is created and compiled.

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

Table 1. Overview of Public Data Sets from the ChEMBL v23 Database

data set	target	total compounds	training set size	test set size	ext. validation set size	total AP-DP descriptors
CHEMBL_hMC	human microsomal clearance	5348	4020	801	527	6937
CHEMBL_rMC	rat microsomal clearance	2166	1636	331	199	4849
CHEMBL_mMC	mouse microsomal clearance	790	617	117	56	4408
CHEMBL_Caco-2	Caco-2_Papp permeability	2582	1938	398	246	7433

2. Training is performed with loops over global iterations, that is, an iteration within all tasks included in the multitask DNN is trained for x epochs. Before switching to the next task, the current weights and biases are saved.
3. Afterwards, the optimizer switches to the next task. Before training starts, weights of the previous model are loaded. Since output units are unique, they remain unaffected.
4. This cycle is repeated for y global iterations. Using the Keras callback function "save_best_only" ensures convergence. The model with best performance on the test set is kept and can be evaluated. Thus, for every task a model is created with best performance for this task.

This procedure reduces the bias towards tasks with rich data sets, because every data set is trained for x epochs during one global iteration without affecting weights of other tasks instantly.

Generating Regression Tree Models Using Cubist. Reference in silico ADME-Tox models based on Cubist regression trees were generated using the program Cubist[47,48] constructing a rule-based decision tree as consensus of five individual trees. The detailed model building procedure is described elsewhere.[3,49]

Chemical Descriptors and Metrics. Molecular descriptors are computed using Dragon version 1.4,[50] MOE,[51] or RDKit version 2017.09.1.[52] Dragon computes many different descriptors, for example, physicochemical properties and structural and fragment information. Moreover RDKit derived atom pairs (APs)[53] and pharmacophoric donor−acceptor pairs (DPs)[54] were used, as suggested by Ma et al.[18] The combined AP-DP descriptor (or feature) set is described as

$$\text{feature} = \text{atom type } i - (\text{distance in bonds}) - \text{atom type } j \quad (1)$$

In the case of AP, i and j are referred to structural properties like elements, number of non-hydrogen neighbors etc., while i and j in the case of DP correspond to pharmacophore features like cation, anion, polar, hydrophobic, etc. An algorithm is used to compute bits describing one feature, and the final feature is a counting of the appearance of one bit in a molecule. In general, descriptors are normalized using range scaling:[55]

$$F_{ik}^{n} = \frac{F_{ik} - F_{k}^{\min}}{F_{k}^{\max} - F_{k}^{\min}} \quad (2)$$

where F_{ik}^{n} is the ith ($i = 1, ..., M$) normalized feature-value of descriptor k ($k = 1, \cdots, N$), while F_{ik} is the non-normalized one. $F_{k}^{\max}$ and $F_{k}^{\min}$ are the maximum and minimum values of descriptor k, respectively. This feature transformation provides a normalized set of features ranging between 0 and 1. Normalization parameters for models are stored for predictions.

DNN regression model results are reported using the common coefficient of determination (R^2) metric following Ma et al.[18]

Data Preparation. Chemical structures were desalted and normalized with respect to stereochemistry, charge, and tautomeric form, using Pipeline Pilot[56] v9.2. Canonical 3D geometries and hydrogens were generated using Corina.[57,58] Compounds with metal complexes, reactive fragments, or macrocycles or without carbon were rejected. Consistently in this article, we are using the term "test set" for data on which hyperparameter selection and the choice of "early stopping" during the training is done. Or in other words, the test set is for the validation of the model metrics during training to prevent overfitting. "External validation sets" are used to validate the model qualities after training using external data, which was not used to train or validate the model during training. In the case of public data, the final data set was randomly split into three classes: 75% were assigned to the training set, 15% to the test set, and 10% to external validation. Company data sets were randomly split into 75−80% training data, while the remaining compounds were kept as test set. Additionally, data from time-series splits were used as external validation sets.[59]

Public Data Set. Public data sets for metabolic clearance and passive permeability in Caco-2 cells were extracted from ChEMBL[60,61] v23. Raw data was obtained by keyword search in the assay description field, the resulting assay list was manually refined (see Supporting Information). Passive permeability was collected from apparent permeability (P_{app}) values. Clearance data was standardized in units of $mL \cdot min^{-1} \cdot g^{-1}$ and split by species. For each species, the data set was merged using canonical SMILES,[62−64] the standard deviation was used to keep data following stddev(CL) < 20 $mL \cdot min^{-1} \cdot g^{-1}$. Table 1 summarizes statistics and descriptors of the data sets from ChEMBL.

Company Data Set. Sanofi data were combined from different in-house laboratories employing a harmonized assay protocol. Metabolic lability measurements are performed using human, rat, and mouse liver microsomes in the presence of NADPH at 37 °C. Compounds are incubated for 20 min with hepatic microsomal fractions from human or other preclinical species. The remaining parent compound is quantified by LC−MS/MS technique. The reported metabolic lability [hLM/mLM/rLM%] corresponds to compound metabolism at the end of each experiment. Moreover samples with NADPH-independent metabolic lability were excluded. Compounds were omitted with standard deviations >10% in absolute values.

The distribution coefficient (logD) at pH 7.4 is determined by standardized HPLC using a Phenomenex GeminiNX C18 column and UV detector. Stock compound is vacuum-dried and prepared as sample solution in DMSO with 1:1 acetonitrile/water. A linear solvent gradient is applied over 6 min, ranging from 95% 5 mM MOPS/$(CH_3)_4N{\cdot}OH$ buffer at pH 7.4 to 5% MOPS/$(CH_3)_4N{\cdot}OH$ buffer at pH 7.4)/95% acetonitrile. The sample retention times are calibrated against basic, acidic, and neutral control. LogD is then estimated from retention times.

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

Table 2. Overview of Company Data Sets of Microsomal Metabolic Lability [Total %] Corresponding to Different Species

data set	species	total compounds	training set size	test set size	ext. validation set size	total AP-DP descriptors
dLM	dog	1056	814	156	86	5040
gpLM	guinea pig	1533	1159	239	135	4439
hLM	human	57635	44997	11245	1393	9717
maLM	macaque	588	461	86	41	3615
mLM	mouse	48242	38545	9581	116	9564
monLM	monkey	246	195	35	16	3484
rabLM	rabbit	553	434	82	37	4259
rLM	rat	51355	40999	10225	131	9502

Table 3. Overview of Company ADME-Tox Data Sets

data set	target	total compounds	training set size	test set size	ext. validation set size	total AP-DP descriptors
Caco-2/TC7	passive permeability	46440	34688	8595	3157	8595
logD	distribution coefficient	81309	61066	12236	8007	9879

Passive permeability is measured by a Caco-2/TC7 cell monolayer assay, which is conducted in 24-well (low throughput) or 96-well format for high throughput at 20 μM compound concentration. On the basolateral side, 5% BSA is added at pH 7.4, on the apical side, 0.5% BSA at pH 6.5. Samples on the apical and basolateral side are taken after 2 h and analyzed by LC/MS. PTmax is calculated by PTotal, and corrected by compound recovery. Threshold for well permeable compounds is considered as PTmax >20 nm/s ($=20 \times 10^{-7}$ cm/s). For permeability, all samples with recovery >50% were considered. Table 2 summarizes internal microsomal metabolic lability data, while Table 3 summarizes data and descriptors for the Caco-2/TC7 and logD data sets, respectively.

Hyperparameters and DNN Network Architectures. DNNs exhibit a large magnitude of adjustable independent parameters and options, called hyperparameters (HPs). The most important HPs can be classified as network architecture parameters and training-control parameters. The first group contains information about the design of the neural network, for example, the number of hidden layers, number of neurons per layer, structure (linear pyramidal or nonlinear), etc., while parameters like optimization algorithms, activation functions, and rates like learning, momentum, dropout rates, etc. are used to control the training.

Finding the best set of HPs is crucial for model quality. Unfortunately, no set of HPs works best for any data set but must be optimized depending on the size and type of the training as well as the kind and amount of molecular descriptors. The following "grid-search" like approach was used to determine a reliable set of HPs. At first, we are looking for optimal training-control parameters and freeze the network architecture using an initial guess of network parameters.[18,32] Learning rate, momentum rate, and L2 rate are scanned in a combinatorial way using different optimization algorithms like SGDM or Adam and different neuron activation functions like relu and sigmoid. Models with all possible combinations are computed and evaluated. For every optimization algorithm, the best result is chosen and then recomputed using different dropout rates, which are selected similar to the other rates. Using the best training parameters, a systematic scan of the network architecture is performed in a second step. Here, we scan different combinatorial possibilities of numbers of hidden layers/units with up to four hidden layers, since no significant improvements could be found by additional layers.

Deep learning is a stochastic method, because of randomness in initialization of the weights, stochastic optimization algorithms, and random dropout. Reproducibility of convergence behavior is antagonized by use of GPU architecture, which adds further variability through NVIDIA cuda libraries.[65] In our experience, deviations in the models can be large, in particular for smaller data sets (<10 000 molecules).

Here, we do not want to provide detailed investigation of HP optimization. In general, we can confirm findings in the literature and thus summarize our experiences with the considered HPs in this article and recommendations in Table 4. For every individual model discussed in this article, the best HP settings are provided in the Supporting Information.

Model Interpretation. All models are characterized by global statistics, i.e. R^2 for the presented models in this paper, using training, test, and external validation data. We also introduce a local property analysis approach termed "response

Table 4. Recommendation and Overview of Hyperparameters to Design and Control DNNs[a]

Network Architecture	
single-task or multitask	depends on available data
linear or nonlinear structure	linear feed-forward (fully connected)
number of hidden layers	3−4; more yields no significant improvement
number of hidden units per layer	pyramidal structure
	number of units of the first hidden layer depends on the size of the incoming descriptor matrix
Training-Control Parameters	
optimizer	Momentum or Adam
activation functions	relu, sigmoid in final unit classification
learning rate	0.001−0.0005
learning rate decay	10^{-6}
momentum rate	0.80−0.95
L1/L2 weight regularization rate	0.0
dropout rate	0.2−0.5, 0.1 input layer
minibatch size	depending on the size of the data set; 512 to 1024 or full-batch seem to be a good choice for our models

[a]Values are based on our experiences. Note that there can be exceptions and values can differ depending on size and type of the training set.

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

map", depicting property changes following fragmentation or derivatization of the parent structure.

For derivatization, molecular derivatives are generated by a Pipeline Pilot workflow. These derivatives are generated by systematic replacement of individual atoms in a target molecule. A number of "probe" reactions is defined for that purpose, including the replacement of hydrogen by hydroxyl function (−OH, partially charged "negative" probe), replacement of hydrogen by amine function ($-NH_2$, partially charged "positive" probe), replacement of hydrogen by fluorine (F, hydrophobic probe), and replacement of hydrogen by nitrile (−CN, polar probe). Derivatives are systematically generated, properties of the virtual molecules are predicted, and the difference between parent molecule and its derivative is visualized in a "response map" of the molecule. The response map is an alignment-free 2D or 3D representation of the molecule, color-coded by the property response to a given derivatization, which is expressed as the numerical difference between predicted properties of parent molecule and its derivative.

For fragmentation, a model-independent approach related to Polishchuk et al.[31] was implemented. Interesting structures for analysis are modified by systematically removing substituents and small side chains. While rings are kept intact, only exocyclic single bonds are split and open valences filled with hydrogen atoms. The maximum small substituent size is 12 heavy atoms. For each virtual matched pair (VMP) formed between the parent and modified structure, the difference from predictions using a corresponding model are mapped to the parent structure (Figure 6). This interpretation allows estimating the importance of a particular substituent for the SAR reflected by the statistical model. The entire approach is implemented in SPL using SybylX (version 2.1.1).[66]

BUILDING, OPTIMIZATION AND STATISTICAL RESULTS OF DNN ADME-TOX MODELS

Model Optimization. As quality of a DNN model depends on descriptors, HPs and statistical effects due to GPU usage, we first discuss model optimization using company data for human metabolic lability of 57 635 compounds.

Starting with finding the best training-control parameters, a series of 184 models with 9717 AP-DP descriptors were built using different combinations of learning rates ranging between 0.0001 and 0.1 and momentum rates between 0.8 and 0.95. Furthermore, relu and sigmoid activation functions were tested as well as the SGD, SGDM, Adam, and RMSProp optimizer. The investigated dropout rates range between 0.0 and 0.7. The Adam optimizer in combination with the relu activation function and learning rates between 0.0001 and 0.0005 performs best, leading to squared test set correlation coefficients R^2 ranging between 0.610 and 0.647. The training procedure does not converge, while learning rates $\gg$0.0005 or sigmoid activation functions are applied in conjunction with the Adam optimizer with this data set. However, results obtained by the Adam optimizer using the SGDM optimizer with larger learning rates between 0.01−0.5 provide similar results with R^2 values around 0.64. The best results using the RMSProp and the SGD optimizer produce acceptable test set correlations with R^2 values of 0.620 and 0.632, respectively. In contrast, sigmoid activation functions produce poor results, for example, for SGDM with a learning rate of 0.1 and momentum rate of 0.85; here model training does not converge. Furthermore, the procedure is sometimes sensitive to dropout rates. For example, for the SGDM optimizer, R^2 values range reliably between 0.642 and 0.644 using optimized SGDM parameters with dropout rates indicated above. Using the Adam optimizer, the R^2 values vary slightly between 0.637 and 0.647.

To find the best network architecture, an ensemble of 384 models was built using different combinations of hidden layers (1−5) and units per layer (100−6000). Here, the network exhibiting three hidden layers with 6000−500−100 units provided the best result (R^2 = 0.647). This result agrees with that of Angermueller et al.,[32] who propose a pyramidal layer/neuron structure as an optimal network setting for DNN models. Since optimized training-control parameters have been used, the influence of the network architecture on the statistical predictivity is small ranging from R^2 values of 0.558 to 0.647 with a mean R^2 value of 0.626.

To analyze descriptor influences, this procedure was repeated using a subset of 612 Dragon descriptors. Here, the best model exhibits a R^2 value of 0.552 for test set, which is lower than for the model using AP-DP descriptors. Table 5 summarizes test and external validation R^2 values of the best models computed with single-task DNN using Dragon and AP-DP descriptors. For comparison, the R^2 values from a reference Cubist model for the same data set are provided, which is performing slightly better than the DNN model using the same Dragon descriptor set. However, a DNN model with AP-DP descriptors shows improved performance. Unfortunately, a Cubist model for this large descriptor matrix cannot be developed for direct comparison. The use of AP-DP descriptors was for us very useful for medium-sized and large data sets. However, the DNN parametrization in Table 4 provides a reasonable first guess leading to predictive models with stable convergence. It might be applied as balanced parameter set for general ADME-Tox modeling in accordance with other recommendations.[18,32]

Table 5. Comparison of Test and External Validation R^2 of Different Company Human Metabolic Lability (hLM) Models Computed with Single-Task DNN and Cubist Regression Trees Using Dragon Descriptors[a]

data set	Cubist/Dragon	DNN/Dragon	DNN/AP-DP
test set [R^2]	0.593	0.552	0.647
ext. validation set [R^2]	0.533	0.509/0.436	0.597/0.497

[a]For DNNs, results obtained using AP-DP descriptors are also shown. The statistically best R^2 are shown. Note that there two different external validation sets for hLM data.

Predictive Single-Task DNN Models from Large Data Sets. Metabolic clearance DNN models can be derived using ChEMBL public data and similar parametrization (see Table 6). The SGDM optimizer provided results for microsomal clearance data. The obtained human microsomal clearance (hMC) model computed using AP-DP descriptors exhibits R^2 values of 0.586 for both test and external validation set comparable to the reference Cubist Dragon model. For rodent microsomal clearance (rMC, mMC), the models are significantly different. The rat rMC model results in R^2 values of 0.709 and 0.751 for test and external validation, respectively, indicating improved predictivity compared to hMC. In contrast the mouse mMC model exhibits R^2 values of 0.493 and 0.398, which indicates poor performance. The data set size drops from 5348 (hMC) to 2166 (rMC) and 790 (mMC). Besides

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

Table 6. Overview of Test and External Validation R^2 of Different Models Computed with Multitask and Single-Task DNNs Using Public Data from CHEMBL[a]

data set	test (ext. validation) [R^2]		
	DNN single-task	DNN multitask	
		microsomal clearance	microsomal clearance + Caco-2_Papp
CHEMBL_hMC	0.586 (0.586)	0.610 (0.566)	0.624 (0.574)
CHEMBL_rMC	0.709 (0.751)	0.729 (0.771)	0.722 (0.783)
CHEMBL_mMC	0.493 (0.398)	0.563 (0.475)	0.575 (0.486)
CHEMBL_Caco-2	0.576 (0.542)		0.500 (0.542)

[a]The statistically best R^2 are shown. Values in parentheses correspond to external validation data.

microsomal clearance, we used 2582 ChEMBL compounds to develop a Caco-2/P_{app} permeability model with R^2 values of 0.576 and 0.542 for test and external validation sets, respectively.

Consequently, predictive ADME-Tox models indeed can be derived using public data from various laboratories and nonstandardized experiments, although this introduces noise (Table 7). The hLM, rLM, and mLM models built with company data are based on the same parameter sets and can be directly compared. These models exhibit R^2 values >0.68 for test sets and ∼0.8 for distinct external validation sets, which were collected from novel laboratory data some months after model building. The Caco-2/TC7 model provides R^2 values of 0.715 and 0.621, respectively. Furthermore, a logD single-task model led to predictive R^2 values >0.86 using corporate data. This underlines how harmonized experimental conditions can influence the predictive quality of a DNN model.

Table 7. Overview of Test and External Validation R^2 of Different Models Computed with Single-Task DNN Using Company Compounds and AP-DP Descriptors[a]

data set	species	test set [R^2]	validation set [R^2]
metabolic lability (LM)	human	0.647	0.597/0.497
	mouse	0.689	0.780
	rat	0.681	0.825
Caco-2/TC7	human	0.715	0.621
logD		0.868	0.863

[a]The statistically best R^2 are shown.

Combination of Not Directly Correlated Data Sets Using Multitask DNNs. To study "gap-filling" capabilities of DNNs, we investigated multitask models embedding multiple cross-species metabolic clearance/lability and Caco-2-permeability data. First, we combined the correlated ChEMBL metabolic clearance data (hMC, rMC, mMC) applying alternate training plus the optimized parameters from our previous single-task models. Next, Caco-2/P_{app} permeability is added to study a mechanistically unrelated data set. Table 6 summarizes resulting R^2 values and a comparison with the single-task models.

For all species, multitask training improves R^2 values for the test sets. In comparison to single-task results, the R^2 value for hMC is improved by 4.1%, while the improvement for rMC and mMC is 2.8% and 14.2%, respectively. Looking at the external validation set metrics, rMC and mMC improved R^2 by 2.7% and 19.3%, while the R^2 for the hMC external validation set is lowered by −3.4% in the multitask model. For the more complex multitask microsomal clearance/Caco-2/P_{app} DNN model, a test set R^2 improvement for all three microsomal clearance species is observed in comparison to single-task models (hMC = 6.5%; rMC = 1.8%; mMC = 16.6%). For hMC and mMC, the R^2 value is improved compared to the multi-microsomal clearance model, while the rMC value is slightly decreased. The external validation set R^2 values are improved for all three species taking the multi-microsomal clearance model as baseline. Compared to the single-task baseline, only the external validation set R^2 value of hMC is slightly decreased (hMC = −2.0%; rMC = 4.3%; mMC = 22.1%). While the microsomal clearance models benefit from adding Caco-2/P_{app} data, the test set R^2 values of the Caco-2/P_{app} multitask model is significantly lowered by −13.2% compared to single-task. Interestingly, the external validation set R^2 value is not influenced by this multitask approach. Therefore, the combination of mechanistically unrelated properties in a multitask model can improve predictivity, because hidden correlations between data are possibly exploited. In our investigation, microsomal clearance models were improved by multitask training, while combining with microsomal clearance is not useful for Caco-2/P_{app} data.

These studies were repeated using the company data (Table 8). The combination of human, rat, and mouse metabolic lability data in one multitask network leads to a significant improvement of the correlation coefficients compared to single-task models. For hLM, the test set R^2 value is enhanced by 8.5%, while the improvements for mLM and rlm are 13.8% and 12.9%. For the external validation sets, the influence of compound overlap in the data sets was also investigated. The

Table 8. Comparison of Test and External Validation R^2 of Different Models Computed with Multitask and Single-Task DNNs Using Company Compounds[a]

data set	test (ext. validation) [R^2]			
	DNN single-task	DNN multitask		
		metabolic lability	metabolic lability + Caco-2/TC7	Caco-2 + logD
human hLM	0.647 (0.597/0.497)	0.702 (0.689/0.517)	0.710 (0.703/0.569)	
mouse mLM	0.689 (0.780)	0.784 (0.799)	0.782 (0.827)	
rat rLM	0.681 (0.825)	0.769 (0.779)	0.775 (0.812)	
Caco-2/TC7	0.715 (0.621)		0.682 (0.614)	0.703 (0.623)
logD	0.868 (0.863)			0.852 (0.847)

[a]The statistically best R^2 are shown. Values in parentheses correspond to external validation data. Note that there are two different external validation sets for hLM data.

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

external validation sets consists of compounds measured after model generation. In the case of hLM, there are two external validation sets based on a time split. External validation set A includes compounds that are available in training and test sets of rLM and mLM, while external validation set B is totally external. For rLM and mLM, only external validation set B is available. It turns out that set A is improved by 15.4% compared to the single-task baseline, while set B is only improved by 4.0%. This experiment demonstrates the "gap-filling" capability of multitask networks during training, that is, the transfer of abstract features that are available in one data set to another data set, where this feature is missing. Since compounds in external validation set A are available in training sets of rLM and mLM, features can be learnt and transferred to the hLM model. The improvement of set B (newest compounds based on time-split) further shows the strength of possible improvements using multitask networks for totally external data. In the case of mLM, the external validation set R^2 value is improved by 2.4%, but the value for rLM is decreased by −5.6%. Turning to the metabolic lability/Caco-2/TC7 (ML_Caco-2) model, the R^2 values for the three metabolic lability species can be further improved for both test and external validation using the multi-metabolic lability model as reference. In particular, external validation set B for hLM is significantly improved by 10.1%. However, the Caco-2/TC7 model is not improved using the multitask technique. The test set R^2 value is decreased by −4.6%, but the external validation set quality is at the same level as the single-task model. Another example showing no improvement in the case of multitask networks by combining mechanistically not related data sets is provided by the combination of logD and Caco-2/TC7. The Caco-2/TC7 test set R^2 value is decreased by −1.7% and logD is decreased by −1.8%. However, the external validation set R^2 for Caco-2/TC7 is at the same level for both multi- and single-task, while the value for logD is also decreased by −1.9% compared single-task models.

The models derived using public data show exactly the same trends as models trained with company data. Combining mechanistically related data (metabolic clearance/lability) of different preclinical species provides statistical model improvements compared to single-task training. When adding a mechanistically unrelated data like Caco-2, the metabolic clearance/lability models are improved, while the Caco-2 predictivity is slightly lowered. Overall, the availability of harmonized data sets from company standardized experiments is beneficial for model quality.

Metabolic lability DNN models for sparse data sets. Since we could successfully improve human, rat and mouse metabolic lability models using company data, we extended this model with data from rarely tested preclinical species like dog, guinea pig, macaque, monkey and rabbit. While each of the hLM, rLM and mLM data sets consist of ∼50000 compounds, the number of compounds for those data sets ranges from 246 (monkey) to 1533 (guinea pig). Single-task and multitask results are given in Table 9 and illustrated in Figure 2. The multitask model is trained with the same HP setting as obtained from the HP optimization of the hLM model. Looking at the single-task results, all investigated preclinical species data sets provide statistically predictive models with $R^2 > 0.5$ for test and external validation sets except for rabbit and guinea pig data. For test sets, all R^2 values are improved from single-task to multitask networks. In particular, dog, macaque and rabbit models are significantly improved by ∼30%. In contrast, external validation of monkey and rabbit data sets led to lowering of R^2 values by −24.3% and −16.2%, respectively. As monkey and rabbit external validation sets consists of only 16 and 37 compounds, respectively, the results are not significant. Eventually, this study suggests that relevant features and information from larger sets could be transferred to the smaller sets enhancing model quality. Even in the single-task case, the DNN technology enables predictive models using smaller data sets.

Table 9. Comparison of Test and External Validation R^2 of Different Species Metabolic Lability Models Computed with Multitask and Single-Task DNNs

	test set [R^2]		ext. validation set [R^2]	
species	DNN single-task	DNN multitask	DNN single-task	DNN multitask
dog	0.570	0.759	0.723	0.701
guinea pig	0.502	0.583	0.493	0.532
human	0.652	0.709	0.642	0.686
macaque	0.614	0.805	0.648	0.737
monkey	0.636	0.666	0.783	0.593
mouse	0.692	0.778	0.695	0.785
rabbit	0.448	0.567	0.608	0.509
rat	0.686	0.760	0.682	0.760

APPLICATIONS IN DRUG DISCOVERY PROJECTS

Practical experience shows that global machine learning models for ADME-Tox properties often have limited use to describe local, series-specific SAR. Particularly, if the model uses consensus, ensemble or averaging, predictions for close analogs are often pushed toward the mean value of the predicted property for the chemotype, or even the mean value of the training data set. This tendency limits the model capability to detect property differences that can be exploited for lead optimization. Multidimensional design requires balancing multiple ADME-Tox properties which is often only achievable by minor chemical modifications rather than "radical" changes. Therefore, it is important that models properly capture local differences in chemical series and reliably predict modifications. The performance of our models is illustrated here using two chemical series in detail. In this section, we use the global hLM model gained from multitask learning together with rLM, mLM, and Caco2/TC-7 data as well as the global single-task and the global multitask Caco2/TC-7 model. For logD, the global single-task model is employed.

CXCR3 Antagonists. A series of polar CXCR3 antagonists was recently described.[67,68] This series, which also bears a zwitterionic pattern on a flexible amino acid backbone, has shown high variability of ADME parameters, namely, logD, passive permeability, and metabolic lability, raising the need to perform multiobjective optimization on the scaffold decoration. While target affinity obviously is a key factor, it has been discussed elsewhere, and we solely focus on the analysis of ADME predictions.

The general correlation of experimental and predicted metabolic lability, passive permeability, and the lipophilicity coefficient logD7.4 was investigated using 199 CXCR3 antagonists, out of which 146 were part of the DNN training set. Additional 43 compounds were part of the model test set; further to that 10 external samples from the same series were used for external validation. The correlations obtained between

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

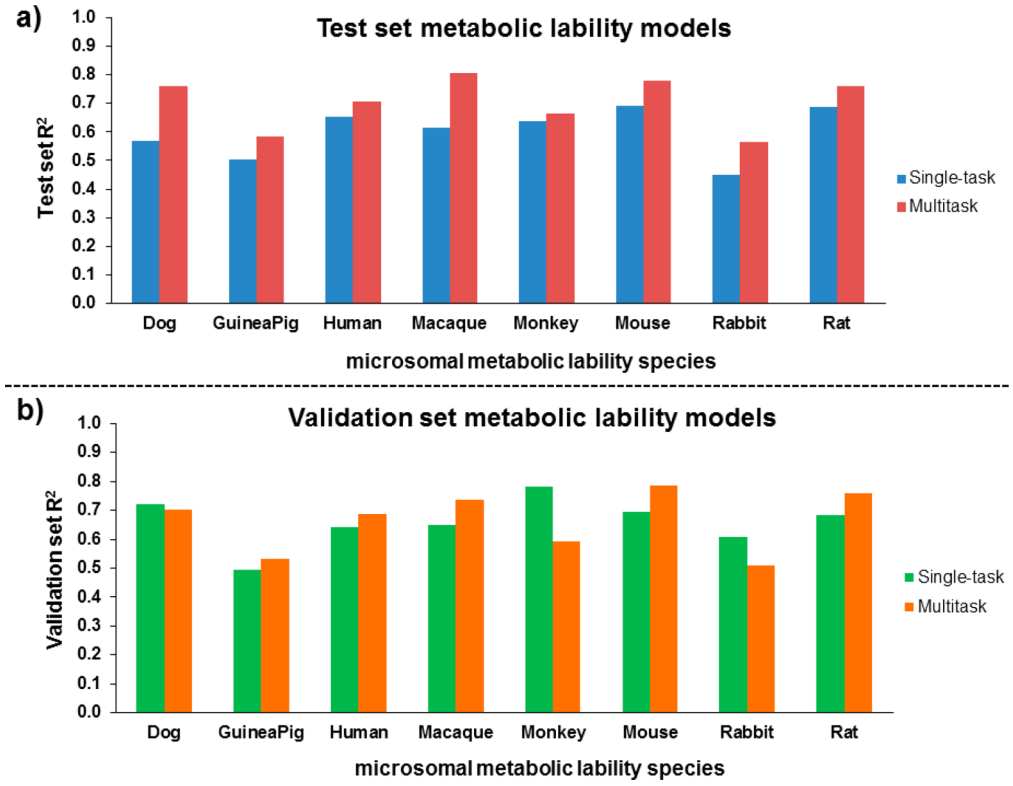


Figure 2. Overview of metabolic lability models of different frequently and nonfrequently tested preclinical species: (a) test set R^2 values; (b) external validation set R^2 values.

predicted and experimental data is excellent, while the model performs equally reliably in the high stability and the high lability regions. In the test set space, 3 outliers are found that are referenced by well-predicted close analogs and might therefore be subjected to further experimental external validation. Overall, the prediction of 10 external samples by the multitask model is slightly better than that by the single task DNN. Figure 3a,b shows an overview of the correlation obtained between experimentally determined and predicted ADME properties using the hLM DNN model, with the squared correlation coefficients for single task and multitask prediction of 10 external samples being 0.776 and 0.802, respectively. Furthermore, we observe a good correlation between experimental and predicted permeability data, which allows for prospective use. Similarly to the observations with the global data set, the predictivity of the single-task model slightly outperforms the multitask approach for Caco-2/TC7 permeability (Figure 3c,d). The correlation between predicted and experimental samples for logD is also encouraging (0.806).

Renin Inhibitors. A class of renin inhibitors based on a central indole-3-carboxamide or related azaindole scaffolds was recently described.[69−71] This data set was used to investigate the correlation of experimental and predicted ADME properties for a total of 48 renin inhibitors, with 35 training compounds, 8 in the test, and 4 in the external validation set. Again, a good correlation between predicted and experimental data was obtained (single-task DNN, R^2=0.96/0.57/0.62 for training set, test set, and external validation). The performance of the multitask model is significantly better compared to the single-task model (multitask DNN, R^2 = 0.96/0.71/0.79 for training set, test set, and external validation, Figure 4). A single compound, CHEMBL1825189, is an outlier in both models comparing predicted (62.5%) versus experimental human microsomal lability (29%). Since lability of this compound was found variable in microsomes and the sample showed low recovery in the permeability assay, it can be assumed that it may be sticking to the vial and experimental data may underestimate its true lability.

Model Interpretation. Any visualization of property trends for in silico models is useful to support interpretation and design. Therefore, we introduce the "response map" concept for exploring the sensitivity of machine learning models for a particular substitution and to identify favorable substitutions on the scaffold. While in the first "decoration" approach the molecule of interest is decorated with informative probe atoms, the second "split" method systematically removes substituents and smaller side chains by splitting exocyclic single bonds. For both cases, differences between virtual molecules and parent structures are then visualized.

Response Map−Decoration Approach for a CXCR3 Antagonist Series. Figure 5 displays response maps from the "decoration" approach for nefazodone (CHEMBL623), a highly labile but permeable molecule. Nefazodone is metabolically activated to a reactive and toxic quinone-imine derivative in human liver microsomes through oxidation by cytochrome P450 3A4, as a primary factor for its liver toxicity.[72] For visualization of the decoration response maps, a color ramp from green to red is used. The map in Figure 5a, depicts changes of four predicted properties following polar substitution of CHEMBL623, derived from the polar CN probe. Metabolic lability of this molecule in human or mouse microsomes is reduced by CN substituents, irrespective of the position. However, some positions are clearly favored. The model suggests a preference for the peripheral aromatic positions of the molecule. Both in human and in mice, the predicted response to CN is largest, reaching −37.7%/−34.4%

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

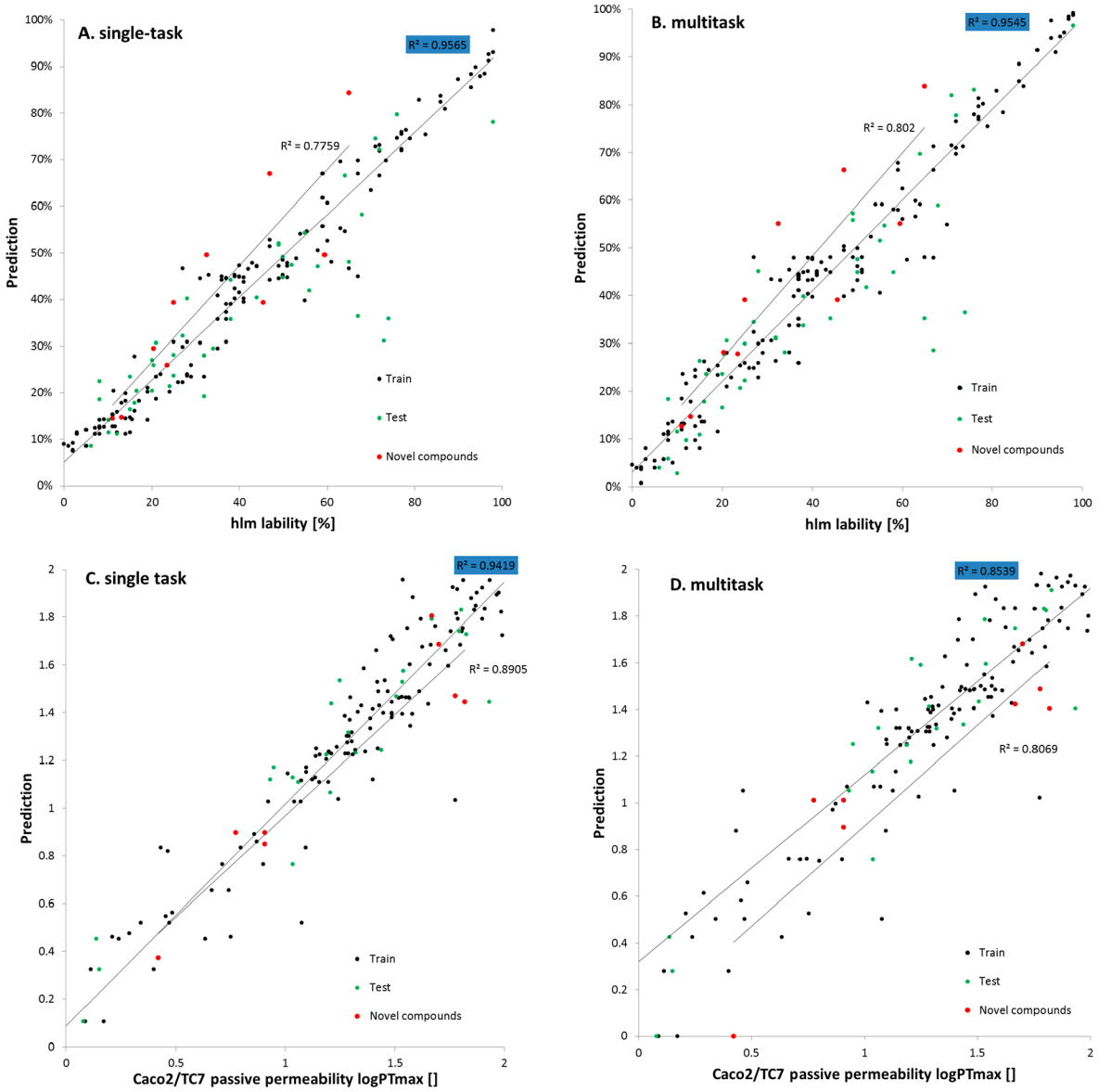


Figure 3. Overview of the correlation obtained between experimentally determined and predicted ADME-Tox properties of CXCR3 antagonist series using DNN models. (A) Single-task and (B) multitask results of the human metabolic lability model; (C) single-task and (D) multitask results of the Caco-2/TC7 passive permeability model.

for the phenol *p*-position and −27.0%/−39.3% for the aniline *p*-position. The predicted response for substitution in either of the para-positions is higher than for the corresponding ortho-positions, though both are electronically activated and thus have a high probability for cytochrome-mediated microsomal activation. In contrast, a weaker response in meta-position is consistent with the electronic structure of the aromatic ring, and a weaker response in the aromatic ortho-positions might be understood as a result of buriedness, leaving the *p*-proton fully exposed to a potential oxidation site. The responses to CN substitution in the central linker regions are much weaker, consistent with high buriedness of the negative polarity of a CN substituent here and also consistent with redundancy of the aliphatic protons that are accessible for metabolic activation. In essence, a peripheral decoration of the molecule with a negatively polarizable group seems most effective to reduce metabolic lability in both species. While those positional preferences are not explicitly trained into the models, they are implicitly transferred from existing structure−property relationships of the broad training data. Medicinal chemistry knowledge suggests that such derivatization might be paid off by changes in properties that are driven by compound lipophilicity. A local analysis of compound permeability (logPTmax) with the same CN probe suggests that moderate reduction of permeability could be expected by the eastern peripheral aromatic position and any aliphatic position around the triazolone core. All other positions are mostly invariant. Likewise, hydrophobicity (logD) is predicted to increase by peripheral aromatic and aliphatic substitution, leaving a −CN substituent exposed to the solvent, while the central linker positions are of less impact. Consequently, a medicinal chemistry strategy for ADME-driven optimization of CHEMBL623 could preferable focus on the exposed phenolether positions of this compound. Figure 5b displays the permeability response for the CHEMBL3700581 employing four probes of different polarity. Overall, the permeability

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

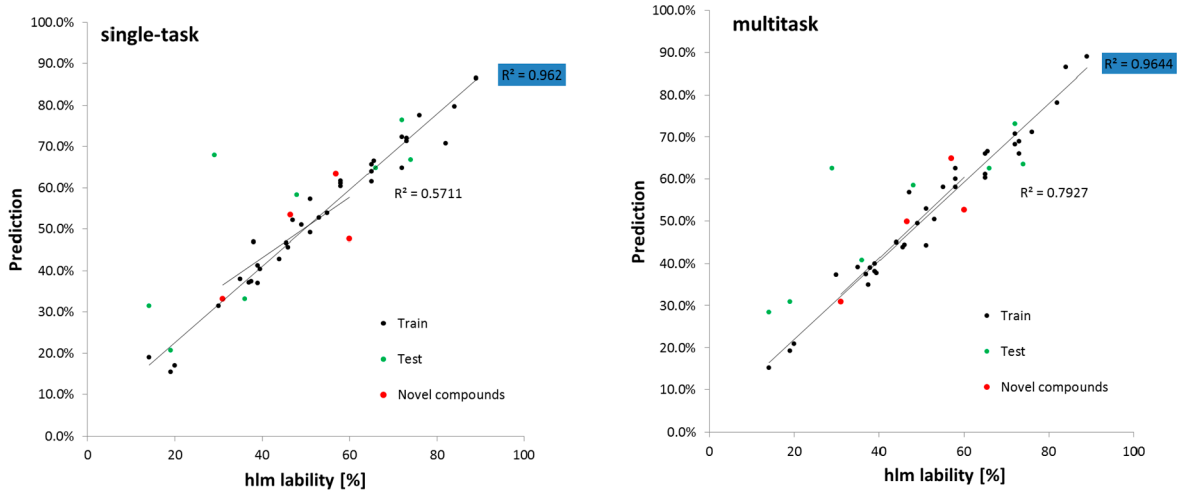


Figure 4. Overview of the correlation obtained between experimentally determined and predicted human metabolic lability properties of renin inhibitor series using DNN models: (left) single-task and (right) multitask results.

of the compound is compromised by substituents. While −F has negligible effects, increasing polarity negatively effects the passive permeability in the order −CN < −OH < $-NH_2$. For the polar, nonprotic −CN probe, a stronger detrimental effect is predicted for positions proximal to electronically activating heteroatoms. For the protic −OH probe, the predicted permeability reduction is slightly stronger for peripheral aromatic positions (phenol derivatives) versus aliphatic positions (alcohol derivatives). In contrast, the $-NH_2$ probe predicts massive deterioration of permeability for all kinds of aliphatic amines (formation of zwitterions). As shown, both polarity and regional substitution preferences are conserved and can be exploited for decoration of the molecule.

Response Map−Split Approach for a Renin Inhibitor Series. Results from the "split" approach are displayed in Figure 6 for members of the indole-3-carboxamide renin inhibitor series (Figure 6a) and CXCR3 antagonists (Figure 6b) and the DNN multitask hLM model (multi-DNN metabolic lability + Caco-2/TC7, Table 8). For visualization, filled circles with different green colors (increase of property) to red (decrease of property) are displayed, while numerical values next to a circle indicate prediction improvements for a virtual matched pair (VMP). Here a positive value plus green color indicate that a molecule without this particular substituent is predicted less metabolically labile in the hLM model. Therefore, the green substituent contributes to lability.

For renin inhibitors, six metabolically labile compounds from the data set are shown with ChEMBL IDs and predicted lability from 63.2% to 82.7%. The inspection for CHEMBL1269855 (upper left) suggests the indole-N1-phenyl and C2-piperidine substituents to be linked to metabolic lability, as removal tends to stabilize the molecule. In contrast, removal of the piperazine moiety at the indole-C3 position causes even higher lability. This trend is qualitatively similar in all displayed molecules. From inspection of Figure 6a, the lipophilic aromatic substituent at indole-C2, which is either a phenoxy or benzyl derivative, significantly influences metabolic lability ranging from +33.1% (CHEMBL1824931, upper middle) to +50.0% (CHEMBL1271282, lower middle). Interestingly, for CHEMBL1824950 (lower left), the methoxy-substituent attached to the upper piperazine moiety, also reveals a significant influence, which is not unexpected for this metabolically labile site. Therefore, this approach reflects SAR trends in accord with medicinal chemistry experience.

A similar analysis is provided in Figure 6b for six CXCR3 antagonists, which are metabolically labile (experimental and predicted). Inspection for CHEMBL623 (upper left) suggests the right piperazine chlorophenyl substituent (black circle, +89.7%) to be mainly linked to metabolic lability of this compound. In contrast, the phenoxy substituent on the left (green circle) displays a lower but marked influence on lability. Unfortunately the same trend can be observed from a detailed inspection of Figure 6b for all other examples. Here it is not possible to identify a single substituent to be exclusively linked to metabolic lability, but labile sites are located in all examples on the left side, on the right plus also at the upward oriented alkyl-substituent attached to the central tertiary amine. For example for CHEMBL390823 (middle right), the dihydrobenzofurane on the left, but also the 2-oxopropyl at the central amine and 1-ethyl-pyrrolidine-2,5-dione on the right influence metabolic lability with contributions ranging from +41.7 to +54.0. This suggests significant problems, when optimizing metabolic lability by adequate substitution of only one of these labile moieties.

In practice, the response map interpretation is useful for DNN and other models, as it is independent from model building. As DNN models often predict extreme values reliably, derived response maps can be informative for substitution positions, if the DNN training set sufficiently covers the available chemical space. This, however, might also limit its applicability, as DNN for ADME-Tox models requires training from large data sets to result in useful interpretations.

CONCLUSIONS

In this work, we presented our strategy to parametrize and optimize the setup, training, and interpretation for deep neural networks in the field of industrial ADME-Tox predictions. We used large and harmonized data sets from different public and company sources to support compound optimization linked to pharmacokinetics and safety. To prove our strategy, we investigated ADME-Tox models for metabolic lability/clearance, Caco-2 permeability, and logD(pH 7.4). We compared single-task models with our multitask approach;

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

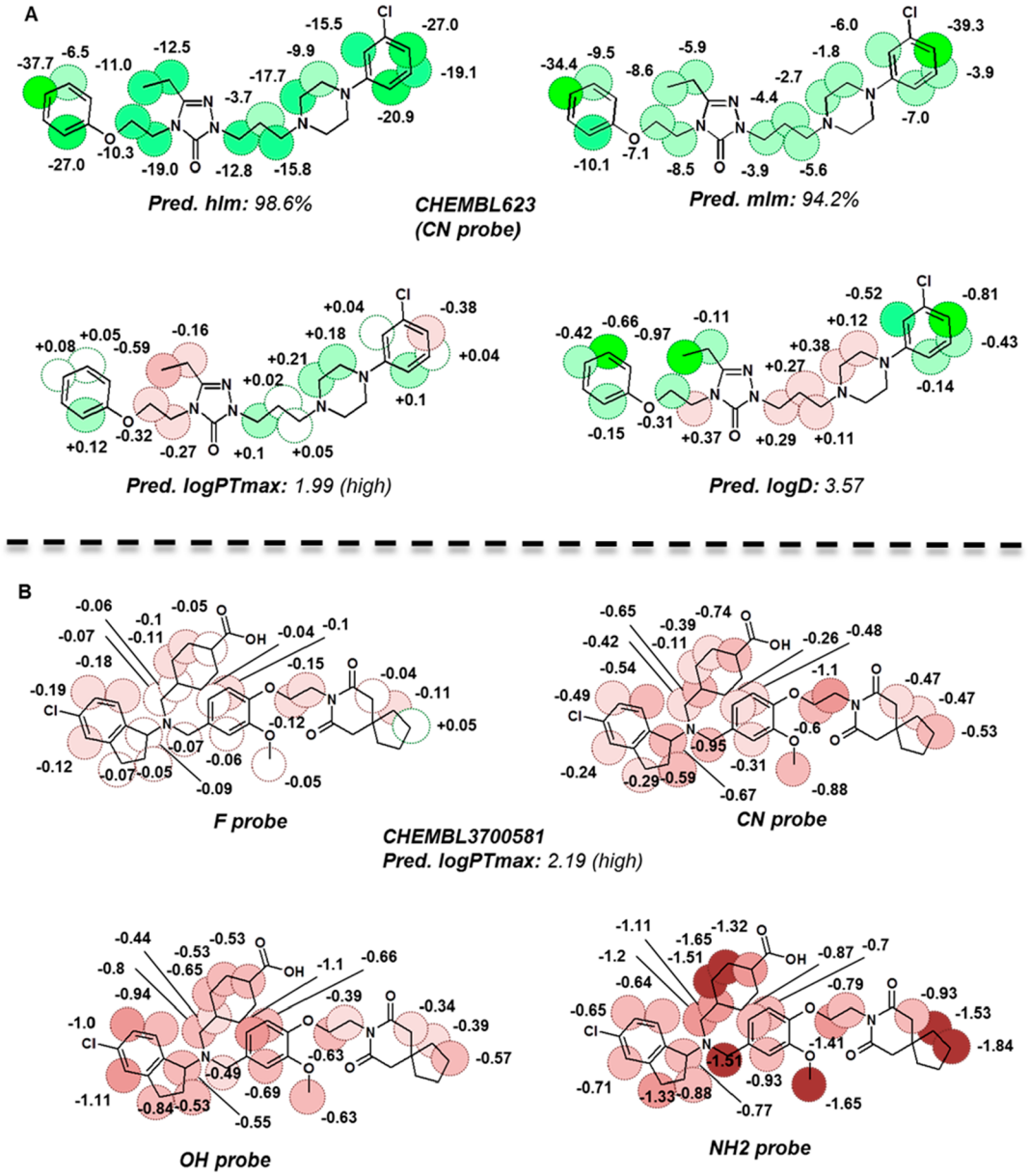


Figure 5. Property response maps resulting from the "decoration" response method, generated by systematic replacement of hydrogen atoms by a fragment probe. (A) Four different properties are visualized upon derivatization with the polar CN fragment probe in a 2d graph of nefazodone (CHEMBL623), that is, human metabolic lability, mouse metabolic lability, passive permeability, and logD Numeric trends of the properties upon derivatization at the respective position are denoted by numbers. (B) Four different probes (−CN/−F/−OH/−NH_2) are used to predict the substituent impact on passive permeability of CHEMBL3700581. The permeability response to fluorination is negligible, while polar and particularly protic derivatives are all associated with strongly reduced passive permeability.

different data sets are combined in one multitask model via iteratively alternate training.

As a first step, it was key to optimize the DNN network architecture and HP to control training generating a model with good prediction. Since there is no set of HPs for all problems and data sets, for example, hLM model ranged between R^2 values of 0.558 and 0.647, we had to scan different possibilities of HP combinations to find the optimal setting. Furthermore, our investigations revealed that DNN models built with internal company data outperform existent models derived by established machine learning methods or demonstrated at least the same level of quality.

The investigation of microsomal metabolic clearance or lability models related to human, rat, and mouse species provided information about applying single-task or multitask DNN models. We are able to build predictive and stable DNN single-task models with $R^2 > 0.5$ using public data from ChEMBL. The quality of the models can be improved up to 19.3% by applying alternate multitask training combining microsomal clearance data sets from different species. However, combining mechanistically unrelated data like microsomal clearance and Caco-2/P_{app} in a multitask framework does not necessarily lead to an improvement over single-task models. In this example, hidden correlations between the cross-data sets cannot be exploited toward a better predictivity in the case of Caco-2/P_{app}.

Using company data only, we rebuilt the respective ChEMBL models. It turns out that the statistically predictive

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

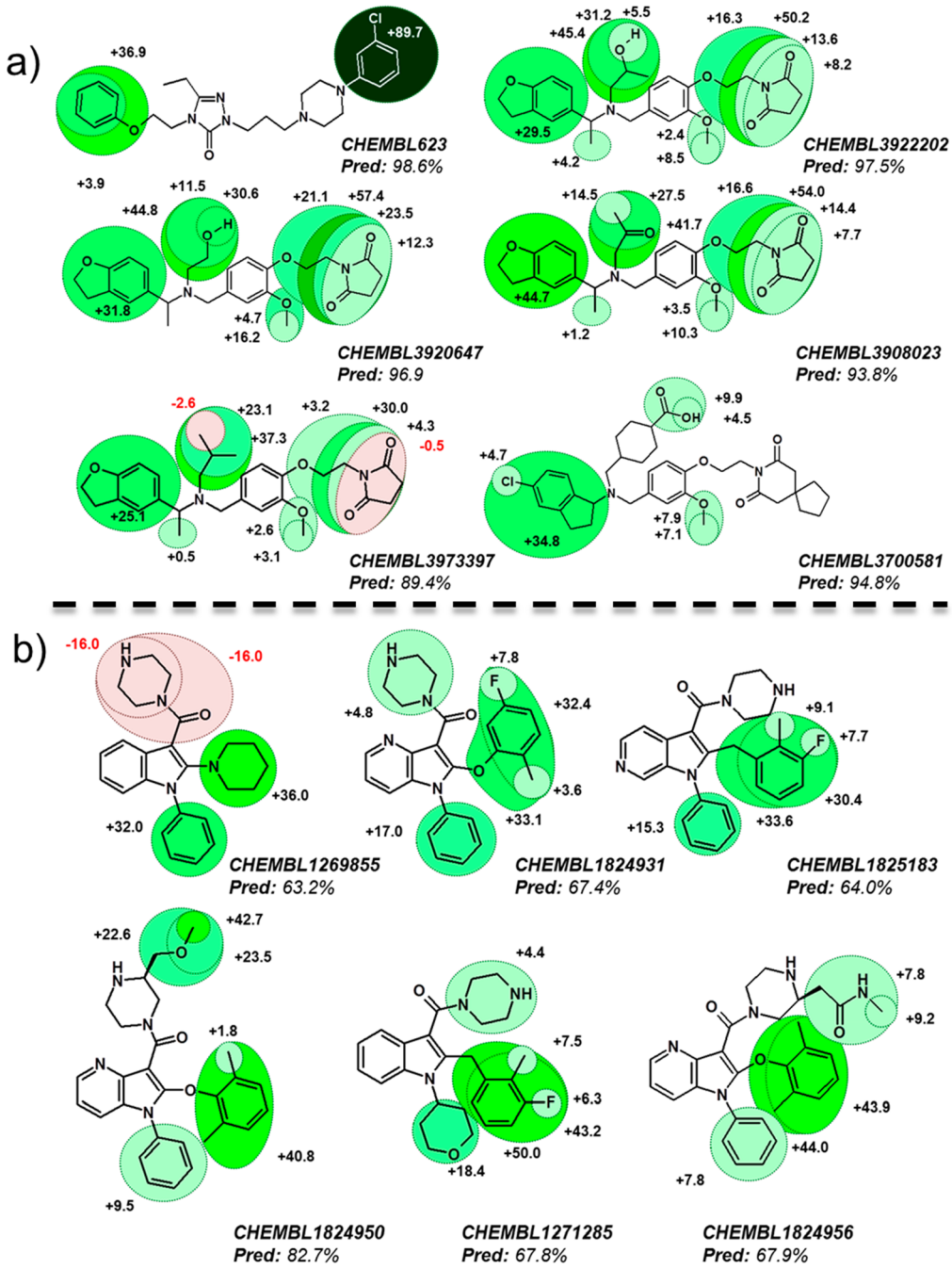


Figure 6. Color coded representation of derivatives from the "split" response map method obtained by systematically removing substituents and smaller side chains by splitting exocyclic single bonds. Differences between newly generated virtual molecules and parent structures are visualized for a series of indole-3-carboxamide based renin inhibitor (a) and CXCR3 antagonists (b). For each compound, the CHEMBL database ID and predicted metabolic lability using the investigated DNN multitask hLM model is given. For visualization, filled circles with different green colors (increase of property) to red (decrease of property) are displayed, while numerical values next to a circle indicate prediction improvements for a virtual matched pair (VMP) by subtraction.

quality of the company models is higher compared to models from public data, demonstrating that data sets using standardized experimental conditions and more data can influence DNN predictive quality. These findings prompted us to develop metabolic lability DNN models for less frequently tested species, showing that features and information from the larger sets can be transferred to the smaller sets enhancing the quality of a multitask model.

Finally, we presented our model-independent response map approach. We interpreted and visualized machine learning models via probing substituents or fragmenting parts of molecules and checked the predicted property response from the model. Besides more robust statistical analyses, these approaches provide useful information to compound design and further optimization options.

Collectively the models and interpretations provide an overview of our activities in the field of deep learning to predict ADME-Tox properties. Further progress, particularly with respect to sparse toxicology data, will be reported in a forthcoming publication. Single-task and multitask DNN

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

models demonstrate a reasonable level of quality and are employed to predict ADME-Tox properties for large, harmonized data sets. The models are integrated in our in silico ADME-Tox workflows for guiding multidimensional optimization programs.

With this approach and our case studies, we have contributed toward better understanding of the scope and practical limitations of DNN in medicinal chemistry. Additional effort is required to gain a realistic and complete picture on the value of artificial intelligence methods applied in drug discovery to support faster and more predictive design of new drugs in early research.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00785.

Table of best hyperparameter settings used to train the final DNNs discussed in this article, training, test and external validation data sets of ChEMBL models (ChEMBL-ID, SMILES and activity), lists of ChEMBL-IDs, Smiles, and experimental and predicted ADME properties of the renin inhibitor and CXCR3 antagonist series, and code snippets for alternate multitask training with Keras/Tesnorflow (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jan.wenzel@sanofi.com.

ORCID

Jan Wenzel: 0000-0002-6771-5567

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge many stimulating discussions with A. Czich, G. Hessler, S. Güssregen (all Sanofi), and S. Hochreiter and G. Klambauer (both JKU Linz). Further we thank S. Toulzac and K. Mi (Sanofi) for support and still growing enthusiasm in the field of machine learning. We particularly thank K. Mertsch, N. Griesang, A. Marker, and many other colleagues at Sanofi over recent years for discussions and experimental ADME data sets and M. Walden and P. Monecke (Sanofi) for providing physicochemical data sets.

■ REFERENCES

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

(2) Paul Gleeson, M.; Hersey, A.; Hannongbua, S. In-Silico ADME Models: A General Assessment of their Utility in Drug Discovery Applications. *Curr. Top. Med. Chem.* **2011**, *11*, 358−381.

(3) Baringhaus, K. H.; Hessler, G.; Matter, H.; Schmidt, F. Development and applications of global ADMET models: in silico prediction of human microsomal lability. In *Chemoinformatics for Drug Discovery*; Bajorath, J., Ed.; John Wiley & Sons, Inc.: 2013.

(4) Cumming, J. G.; Davis, A. M.; Muresan, S.; Haeberlein, M.; Chen, H. Chemical predictive modelling to improve compound quality. *Nat. Rev. Drug Discovery* **2013**, *12*, 948−962.

(5) Lombardo, F.; Desai, P. V.; Arimoto, R.; Desino, K. E.; Fischer, H.; Keefer, C. E.; Petersson, C.; Winiwarter, S.; Broccatelli, F. In Silico Absorption, Distribution, Metabolism, Excretion, and Pharmacokinetics (ADME-PK): Utility and Best Practices. An Industry Perspective from the International Consortium for Innovation through Quality in Pharmaceutical Development. *J. Med. Chem.* **2017**, *60*, 9097−9113.

(6) Muegge, I.; Bentzien, J.; Mukherjee, P.; Hughes, R. O. Automatically updating predictive modeling workflows support decision-making in drug design. *Future Med. Chem.* **2016**, *8*, 1779−1796.

(7) Polishchuk, P. Interpretation of Quantitative Structure−Activity Relationship Models: Past, Present, and Future. *J. Chem. Inf. Model.* **2017**, *57*, 2618−2639.

(8) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436−444.

(9) Schmidhuber, J., Deep Learning in Neural Networks: An Overview. arXiv:1404.7828, 2014, https://arxiv.org/abs/1404.7828.

(10) Hochreiter, S. Untersuchungen zu dynamischen neuronalen Netzen. Diploma; Technical University Munich, Munich, 1991.

(11) Hochreiter, S. Generalisierung bei Neuronalen Netzen geringer Komplexität. Ph.D. Dissertation; Technical University Munich, 1999.

(12) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput* **1997**, *9*, 1735−1780.

(13) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction using Deep Learning. *Front. Environ. Sci.* **2016**, *3*, 80.

(14) Unterthiner, T.; Mayr, A.; Klambauer, G.; Hochreiter, S., Toxicity Prediction using Deep Learning. arXiv:1503.01445, https://arxiv.org/abs/1503.01445, 2015.

(15) Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S., Self-Normalizing Neural Networks. arXiv:1706.02515, https://arxiv.org/abs/1706.02515, 2017.

(16) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. KDEEP: Protein−Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287−296.

(17) Ryan, K.; Lengyel, J.; Shatruk, M. Crystal Structure Prediction via Deep Learning. *J. Am. Chem. Soc.* **2018**, *140*, 10158−10168.

(18) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263−274.

(19) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57*, 2068−2076.

(20) 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*, 2490−2504.

(21) Ghasemi, F.; Mehridehnavi, A.; Perez-Garrido, A.; Perez-Sanchez, H. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. *Drug Discovery Today* **2018**, *23*, 1784−1790.

(22) Gawehn, E.; Hiss, J. A.; Brown, J. B.; Schneider, G. Advancing drug discovery via GPU-based deep learning. *Expert Opin. Drug Discovery* **2018**, *13*, 579−582.

(23) Gawehn, E.; Hiss, J. A.; Schneider, G. Deep Learning in Drug Discovery. *Mol. Inf.* **2016**, *35*, 3−14.

(24) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(25) Korotcov, A.; Tkachenko, V.; Russo, D. P.; Ekins, S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharmaceutics* **2017**, *14*, 4462−4475.

(26) Li, X.; Xu, Y.; Lai, L.; Pei, J. Prediction of Human Cytochrome P450 Inhibition Using a Multitask Deep Autoencoder Neural Network. *Mol. Pharmaceutics* **2018**, *15*, 4336−4345.

(27) Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L. Deep Learning for Drug-Induced Liver Injury. *J. Chem. Inf. Model.* **2015**, *55*, 2085−2093.

(28) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R., Multi-task Neural Networks for QSAR Predictions. arXiv:1406.1231, https://arxiv.org/abs/1406.1231, 2014.

(29) Kearnes, S.; Goldman, B.; Pande, V., Modeling Industrial ADMET Data with Multitask Networks. arXiv:1606.0879, https://arxiv.org/abs/1606.00879, 2016.
(30) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mane, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viegas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv:1603.04467, https://arxiv.org/abs/1603.04467, 2016.
(31) Polishchuk, P. G.; Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N. Universal Approach for Structural Interpretation of QSAR/QSPR Models. *Mol. Inf.* **2013**, *32*, 843−853.
(32) Angermueller, C.; Parnamaa, T.; Parts, L.; Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **2016**, *12*, 878.
(33) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep learning for computational chemistry. *J. Comput. Chem.* **2017**, *38*, 1291−1307.
(34) Ruder, S., An Overview of Multi-Task Learning in Deep Neural Networks. arXiv:1706.05098, https://arxiv.org/abs/1706.05098, 2017.
(35) Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*; Geoffrey, G., David, D., Miroslav, D., Eds.; PMLR: Proceedings of Machine Learning Research, 2011; Vol. *15*, pp 315−323.
(36) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533−536.
(37) Sutskever, I.; Martens, J.; Dahl, G.; Hinton, G. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*; Sanjoy, D., David, M., Eds.; PMLR: Proceedings of Machine Learning Research, 2013; Vol. *28*, pp 1139−1147.
(38) Bengio, Y. Practical recommendations for gradient-based training of deep architectures. arXiv:1206.5533, https://arxiv.org/abs/1206.5533, 2012.
(39) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. arXiv:1412.6980, https://arxiv.org/abs/1412.6980, 2014.
(40) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929−1958.
(41) Nowlan, S. J.; Hinton, G. E. Simplifying Neural Networks by Soft Weight-Sharing. *Neural Comput* **1992**, *4*, 473−493.
(42) Gong, P.; Ye, J.; Zhang, C. Robust Multi-Task Feature Learning. *KDD* **2012**, *2012*, 895−903.
(43) Kendall, A.; Gal, Y.; Cipolla, R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. arXiv:1705.07115, https://arxiv.org/abs/1705.07115, 2017.
(44) Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively Multitask Networks for Drug Discovery. arXiv:1502.02072, https://arxiv.org/abs/1502.02072, 2015.
(45) Godwin, J., Multi-Task Learning in Tensorflow, https://jg8610.github.io/Multi-Task/: 2016, (accessed 2018−12−21).
(46) Chollet, F.; et al. *Keras*, version 2.0.8, https://keras.io, 2005, (accessed 2018−12−21).
(47) Quinlan, J. R. Improved estimates for the accuracy of small disjuncts. *Mach. Learn.* **1991**, *6*, 93−98.
(48) Quinlan, J. R. In *Learning with Continuous Classes*, AI'92, 5th Australian Joint Conference on Artificial Intelligence, Singapore; Adams, A., Sterling, L., Eds.; World Scientific: Singapore, 1992; pp 343−348.
(49) Matter, H.; Anger, L. T.; Giegerich, C.; Güssregen, S.; Hessler, G.; Baringhaus, K.-H. Development of in silico filters to predict activation of the pregnane X receptor (PXR) by structurally diverse drug-like molecules. *Bioorg. Med. Chem.* **2012**, *20*, 5352−5365.
(50) Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *MATCH Commun. Math. Comput. Chem.* **2006**, *56*, 237−248.
(51) *MOE* (version 2011); Chemical Computing Group (CCG): Montreal, Canada, 2011.
(52) *RDKit: Open-source cheminformatics*, 2017.09.1; http://www.rdkit.org, (accessed 2018−12−21).
(53) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Model.* **1985**, *25*, 64−73.
(54) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(55) Faulon, J. L.; Bender, A. *Handbook of Chemoinformatics Algorithms*; CRC Press: 2010.
(56) *BIOVIA Pipeline Pilot* (version 9.2); Dassault Systèmes BIOVIA: San Diego: Dassault Systèmes, 2014.
(57) Sadowski, J.; Rudolph, C.; Gasteiger, J. The generation of 3D models of host-guest complexes. *Anal. Chim. Acta* **1992**, *265*, 233−241.
(58) *Corina*, version 3.491; Molecular Networks Inc.: Erlangen, Germany, 2006.
(59) Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783−790.
(60) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−D954.
(61) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* **2015**, *43*, W612−W620.
(62) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.
(63) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Model.* **1989**, *29*, 97−101.
(64) Weininger, D. SMILES. 3. DEPICT. Graphical depiction of chemical structures. *J. Chem. Inf. Model.* **1990**, *30*, 237−243.
(65) NVIDIA, CUDA Toolkit Documentation. 2018.
(66) *SybylX Molecular Modelling Package*, 2.1.1.; Certara: St. Louis, MO, 2018.
(67) Bata, I.; Tömösközi, Z.; Buzder-Lantos, P.; Vasas, A.; Szeleczky, G.; Balázs, L.; Barta-Bodor, V.; Ferenczy, G. G. II. Discovery of a novel series of CXCR3 antagonists with a beta amino acid core. *Bioorg. Med. Chem. Lett.* **2016**, *26*, 5429−5437.
(68) Bata, I.; Tömösközi, Z.; Buzder-Lantos, P.; Vasas, A.; Szeleczky, G.; Bátori, S.; Barta-Bodor, V.; Balázs, L.; Ferenczy, G. G. I. Discovery of a novel series of CXCR3 antagonists. Multiparametric optimization of N,N-disubstituted benzylamines. *Bioorg. Med. Chem. Lett.* **2016**, *26*, 5418−5428.
(69) Matter, H.; Scheiper, B.; Steinhagen, H.; Bocskei, Z.; Fleury, V.; McCort, G. Structure-based design and optimization of potent renin inhibitors on 5- or 7-azaindole-scaffolds. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 5487−5492.
(70) Scheiper, B.; Matter, H.; Steinhagen, H.; Bocskei, Z.; Fleury, V.; McCort, G. Structure-based optimization of potent 4- and 6-azaindole-3-carboxamides as renin inhibitors. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 5480−5486.
(71) Scheiper, B.; Matter, H.; Steinhagen, H.; Stilz, U.; Bocskei, Z.; Fleury, V.; McCort, G. Discovery and optimization of a new class of potent and non-chiral indole-3-carboxamide-based renin inhibitors. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 6268−6272.
(72) Kalgutkar, A. S.; Vaz, A. D.; Lame, M. E.; Henne, K. R.; Soglia, J.; Zhao, S. X.; Abramov, Y. A.; Lombardo, F.; Collin, C.; Hendsch, Z.

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268

S.; Hop, C. E. Bioactivation of the nontricyclic antidepressant nefazodone to a reactive quinone-imine species in human liver microsomes and recombinant cytochrome P450 3A4. *Drug. Metab. Dispos.* **2004**, *33*, 243−253.

DOI: 10.1021/acs.jcim.8b00785
J. Chem. Inf. Model. 2019, 59, 1253−1268