AI for predicting chemical-effect associations at the universe level - `deepFPlearn`

Jana Schor*[1], Patrick Scheibe[2], Matthias Bernt[1], Wibke Busch[3], Chih Lai[4], and Jörg Hackermüller[1]

[1] *Young investigators group bioinformatics & transcriptomics, Department of molecular systems biology, Helmholtz Centre for environmental research - UFZ, Leipzig, Germany*
[2] *Department of Neurophysics, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*
[3] *Department of Bioanalytical Ecotoxicology, Helmholtz Centre for environmental research - UFZ, Leipzig, Germany*
[4] *Graduate Program in Software & School of Engineering, University of St. Thomas, Minnesota, USA*

Abstract

A plethora of chemical substances is out there in our environment, and all living species, including us humans, are exposed to various mixtures of these. Our society is accustomed to developing, producing, using and dispersing a diverse and vast amount of chemicals with the original intention to improve our standard of living. However, many chemicals pose risks, for example of developing severe diseases, if they occur at the wrong time in the wrong place. For the majority of the chemicals these risks are not known. Chemical risk assessment and subsequent regulation of use requires efficient and systematic strategies, which are not available so far. Experimental methods, even those of high-throughput, are still lab based and therefore too slow to keep up with the pace of chemical innovation. Existing computational approaches, e.g. ML based, are powerful on specific chemical classes, or sub-problems, but not applicable on a large scale. Their main limitation is the lack of applicability to chemicals outside the training data and the availability of sufficient amounts of training data. Here, we present the ready-to-use and stand-alone program `deepFPlearn` that predicts the association between chemical structures and effects on the gene/pathway level using deep learning. We show good performance values for our trained models, and demonstrate that our program can predict meaningful associations of chemicals and effects beyond the training range due to the application of a sophisticated feature compression approach using a deep autoencoder. Further, it can be applied to hundreds of thousands of chemicals in seconds. We provide `deepFPlearn` as open source and flexible tool that can be easily retrained and customized to different application settings at `https://github.com/yigbt/deepFPlearn`.

Keywords: Deep learning, toxicology, binary fingerprint, autoencoder, molecular structures

*Corresponding author.

1 Introduction

Humans and ecosystems are exposed to a vast number of chemicals that may threaten their health. While chemical products are important for maintaining our standard of living and chemicals form the building blocks of life, some chemicals are hazardous, and their safety needs to be thoroughly evaluated. The list of chemicals humans and ecosystems may be exposed to is vast, and the set of chemicals of anthropogenic origin is rapidly growing: The Chemical Abstracts Registry Service grew from 20 million in 2002 to currently 169 million unique chemical substances [1]. Certainly, many chemicals listed in this registry are not relevant for exposure since they are not used in larger quantities. However, between 30 000 and 100 000 chemicals are expected to be currently on the European market [2], [3, p. 238]. The number of substances detected in humans is of similar order of magnitude. More than 50 thousand substances have been compiled from scientific texts that reported detection in blood and are listed in the Blood Exposome DB [4]. Contamination with anthropogenic chemicals is of analogous concern in the environment [5, 6] and more than 3000 substances are currently listed in a database of emerging pollutants across Europe [7].
Environmental exposure poses a major risk for the health of humans and ecosystems. Environmental and life-style rather than genetic factors have been defined as the major cause for many chronic diseases [8] and the observed growth in the incidence of non-communicable chronic diseases can only be explained by changing environmental determinants [9, 10]. Exposure to chemicals has been a major concern for wildlife populations [11, 12, 13]. Up to 26% of aquatic species loss may be attributed to exposure to chemical mixtures [14], [3, p. 245], [15].

Risk assessment fails to keep up with the pace of chemical innovation. The range of chemicals humans

or ecosystems may be exposed to, and the observed hazards of individual chemicals clearly call for an efficient and effective risk assessment and subsequent thorough regulation of chemical use. Also, a "toxic-free environment" has been designated a key priority for the EU [16]. However, the European Environment Agency forecasts that exposure to mixtures of chemicals, including hazardous, will further increase [3, pp. 248–249]. Also, the throughput of regulatory processes is slow compared to the pace of chemical innovation. E.g. of the more than 21 000 chemicals registered under REACH, the European legislation regulating industrial chemicals, more than 1000 have been designated low priority for further work, because of no further concern, or already implemented risk management and more than 1800 have been designated high priority, because data generation was ongoing or risk management under consideration. However, for the vast majority of more than 18 000 substances, the need for suitable regulatory actions still needs to be determined [17]. It has been estimated that evaluation of a substance of potential concern under the REACH legislation takes 7–9 years, during which exposure may continue [3, pp. 248–249]. Also, approximately 70% of the registration dossiers that were evaluated have been found to be incomplete or not compliant [18]. In summary, traditional approaches to chemical regulation are not efficient enough to master the number and growth of substances on the market. Predictive *in silico* approaches that build on the integration of different data sources may contribute to solving this challenge – not necessarily by replacing experimental approaches, but at least by prioritizing substances for further evaluation.

***In silico* approaches for toxicity prediction.** A range of different approaches for predicting toxicity *in silico* have been developed, including Structural Alerts, Read across, and Quantitative Structure Activity Relationships (QSAR) [19]. Structural alerts and rule-based models constitute a simple but powerful approach to toxicity prediction and build on using individual chemical substructures as indicators for toxicity [20]. These methods rely on either human-expertise-based or data-derived rules and are easy to interpret. However, since there is no mechanism that enforces the completeness of rules, there is a high risk of false-negative predictions [19]. Read-across builds on predicting the unknown toxicity of a substance by extrapolation from a set of highly related substances with known toxicological properties. Read across is a commonly used alternative to animal experiments for filling data gaps in REACH dossiers. However, applicability of read-across is restricted to a subset of substances for which sufficient information of related compounds is available [21]. QSARs relate molecular descriptors to toxicity or other properties of a chemical [22, 23]. Local QSARs are either built on a set of related chemicals or on expert knowledge on a shared mode of action of a group of chemicals. Models derived from a diverse set of chemicals are called global QSARs [19].

Descriptors used in QSAR include physicochemical properties and different representations of the molecular structure or properties thereof and more recently also high throughput screening-derived data on biological activity. 2D QSAR uses descriptors derived from the 2D-structure graph of a molecule that are identical for isomorphic graphs (graph invariants) [23]. A frequently used family of descriptors are molecular fingerprints that record the occurrence of local and regional substructure elements in the graph [24]. The 3D QSAR considers the non-covalent interactions between the molecule under consideration and a receptor. In the simplest case, descriptors are based on sampling the force acting on a probe atom that is placed on a grid around the molecule under consideration [23]. Additionally, also 3D-derived fingerprints have been used as descriptors [24].

QSAR approaches utilize multivariate statistical models, like regression or discriminant analysis, to relate molecular descriptors to toxicity. More recently, with the growth in available training data and methodological advancements in artificial intelligence, global QSARs built on machine learning models became available.

Machine learning in 21st century toxicology Toxicology is currently experiencing a paradigm shift from apical endpoints in animal models to more integrated strategies, combined prediction, and high-throughput testing for a range of different, often molecular, endpoints.

Computational methods that involve machine learning have been developed for a while. QSARs, built on machine learning, typically aim at the prediction of a specific and often apical toxicological endpoint. They are therefore powerful, particularly for specific classes of chemicals with a common mode of toxic action. This is of great advantage when making quantitative predictions, e.g. for effect concentrations in a specific domain. However, understanding the effect of structure and activity using QSARs is not straightforward when large amounts of data are generated. The extrapolation outside the range of the training data set is prone to many false predictions. Further, it is known that physicochemical properties are cross-correlated and predictions with different QSARs are often contradictive. Therefore, QSARs are of limited use for the application to a whole universe of chemicals. Further, it is not always straightforward to incorporate new data or knowledge into their models.

Several initiatives and organisations exist worldwide that are dealing with the safety of chemicals regarding human and environmental health. For example, there is the Tox21 program [25], as a collaboration between several U.S. federal agencies, tightly linked to the toxic substances control act (TSCA) [26]. In europe, there is a law that regulates the chemical risk assessment and licensing

– REACH [18], and there is the European Chemicals Agency (ECHA) that is responsible for its implementation within the European Union. The development of new (computational) ways to rapidly test (or predict) whether chemicals have adverse affects or not are part of their programs. For example, the Tox21 Challenge was announced in 2014 with the aim to reveal how well independent researchers can predict the interference between compounds and biochemical pathways given a dataset of chemical structures only. A set of 12 000 chemicals which were measured for 12 different toxic effects was provided for the challenge, and the aim was to develop computational methods that are able to successfully predict those effects. Several computational methods have been developed during this challenge that involved machine learning. The best performing method was the `DeepTox`[27] pipeline. In brief, the chemical representation of the compounds is normalized, a large number of descriptors is computed and deep learning models are trained and evaluated based on the provided test data. Later, these models were successfully validated by the challenge initiators with additional data that has been hold back during the challenge. This pipeline reported AUC values above 85% for the prediction of the association between the provided chemicals and the 12 targets. Further, the authors showed that the concept of deep learning to learn simple objects in the lower layers and combinations of those to more complex objects in higher layers, as shown in image recognition, also holds in toxicology. Hidden neurons in their deep learning models represent known molecular substructures - toxicophores, which have been generated manually by experts for decades.

In the following, we briefly review a selection of further predictive approaches. `eToxPred` [28] was developed to quickly estimate the toxicity of large collections of low molecular weight organic compounds. It employs a Restricted Boltzman Machine classifier and a generative probabilistic model to predict a Tox-score. The reported accuracy is 72%. In a different approach by [29], a support vector machine and random forest were used to build single-label models. Here, the problem of class imbalance has been resolved by methods of under-sampling the data. Four methods were explored to construct multilabel models that can predict the interaction of one chemical against multiple targets simultaneously. The reported accuracy was between 0.74 and 0.81. `TarPred`[30] is a web application for predicting therapeutic and side effect targets of chemical compounds. However, it is not available anymore. The `DeepChem` Project [31] is a Python library that provides datasets, functions, and tutorials for several use cases, with the aim to democratize deep learning for science in general, and for chemistry in particular. It can be used to develop or, as a reference, to compare custom computational approaches in the field.

The main limitation of the above mentioned (and other) ML approaches in toxicology is the lack of applicability to chemicals outside the training data and the availability of sufficient amounts of training data. A specific challenge is that the descriptors need to be fine-grained enough to capture the particularities of molecular initiating events and coarse-grained enough to allow for machine learning. In consequence, the (high) dimensionality of the descriptors/features stands in large contrast to the (low) amount of available labeled training data.

Here, we present a ready-to-use deep learning tool that combines feature reduction with a deep autoencoder and subsequent prediction with a deep feed forward neural network to predict the associations between chemical structures and gene targets that are involved in endocrine disruption (ED). We decreased the discrepancy between large descriptor size and limited amount of labeled training data by i) using a simple representation of chemical structure - the binary fingerprint of moderate size, and ii) by applying feature compression prior to the actual prediction of the association to an effect. Further, we overcame the domain extrapolation problem by training the autoencoder on a huge repertoire of chemical structures and showed that the prediction quality on the subset of labeled training data for ED remained high.

2 Methods and data

Representation, similarity, and visualization of chemical structures. The molecular structure of chemical compounds was encoded in binary topological fingerprints using Python's `RDKit` [32, version 2020.03.2] with the parameters: minimum path size: 1 bond, maximum path size: 7 bonds, fingerprint size: 2048 bits, number of bits set per hash: 2, minimum fingerprint size: 64 bits, target on-bit density 0.0. These fingerprints were generated from chemical structure information provided in the form of InChI or SMILES strings. All training data files containing chemical structure and target association information were serialized using Python's pickle module. Uncompressed (original binary fingerprint) and compressed features of the molecules are visualized in 2D space using the UMAP algorithm [33] and the following metrics. Tanimoto similarity was used to evaluate molecular similarity of the fingerprint representations. Pearson correlation was used to evaluate similarity of compressed features. A k-means clustering with $k \in [2..7]$ was applied to the resulting coordinates of the *uncompressed* features. The assigned clusters were translated to color codes in the visualizations of both uncompressed and compressed features.

The deep learning tasks were implemented using the Python library of the `TensorFlow` framework [34, version 2.2.0], Python's `TensorFlow` wrapper `Keras` [35, version 2.3.1], and the `Scikit-learn` framework [36, version 0.23.1].

A deep regularized autoencoder (AE) was constructed for dimensionality reduction of the binary fingerprints. The AE had a symmetric shape with one-dimensional input and output layers of the size of the input fingerprint L_{FP}. The number of hidden layers N_H and their sizes $S_i, i \in [1..N_H]$ were defined in dependency of L_{FP} and the desired latent space size L_z:

$$N_H = \lfloor \log_2(L_{FP}/L_z) \rfloor \quad (1)$$

$$S_i = L_{FP}/2^i \quad i \in [1..N_H] \quad (2)$$

The ReLU activation function was used towards inner layers, and the Sigmoid activation function was used towards the output layer. The AE model was compiled using the Adam optimizer with a default learning rate of 0.001, a default decay of 0.01, and the binary cross-entropy as a loss function.

A deep neural network was constructed as a sequential feed-forward neural network (FNN) for the classification task. The dimensions of the stacked layers depends on the mode of action of `deepFPlearn`. If feature compression via the autoencoder was enabled, the FNN was used subsequently of the AE and the length of the one-dimensional input layer of the FNN matched the length of the latent space L_z vectors. Otherwise, the number of neurons in the input layer matched the length of the binary fingerprint L_{FP}. The hidden layers were of decreasing sizes, followed by an output layer of size 1. The number of hidden layers $\hat{N}_H$ and their sizes $\hat{S}_j, j \in \left[1..\hat{N}_H\right]$ depended again on the provided input size L_{input} (which is either L_{FP} or L_z). The last four layers with only a few neurons (e.g. less than 32 in the example case of $L_{FP} = 2048$) were not included.

$$\hat{N}_H = \lfloor \log_2(L_{input})/2 \rfloor - 4 \quad (3)$$

$$\hat{S}_j = L_{input}/2^j \quad j \in [1..\hat{N}_H] \quad (4)$$

Dense layers were used with the ReLU activation function towards hidden layers and default kernel regularization of 0.001. The Sigmoid activation function was used towards the output layer. All hidden layers were followed by a dropout layer with a dropout rate of 0.2. The FNN model was compiled using the Adam optimizer with a default learning rate of 0.001, a default decay of 0.01, and the mean squared error as a loss function.

Different datasets were collected from the literature and from public databases. A manually curated dataset S was downloaded from the supplemental material of [29]. It contained chemical-target associations for 7248 chemicals and 6 gene targets. The six considered target genes were: *androgen receptor* (AR), *estrogen receptor* (ER), *gluccochorticoid receptor* (GR), *thyroid receptor* (TR), *PPARγ*, and *Aromatase*. Chemicals that interact with these targets can cause endocrine disruptions (ED) in humans. Originally, this data had been retrieved from bioassay data of the Tox21 program [25], and carefully transformed to binary associations by [29]: Associations were considered as not available (NA) if no bioassay data was available, and as 1 or 0, if an association between chemical and gene target had been confirmed in a bioassay or not, respectively. For details on the data preparation, thresholding and pre-processing of this dataset we refer to the original publication [29]. The dataset S was extended by the artificial target *endocrine disruption* (ED) that combines all existing targets with a logical OR operation. Chemicals in the S dataset were identified by their SMILES representation, and encoded as binary topological fingerprints as described above.

Further, a dataset D was generated from the 719 996 chemicals listed in the CompTox Chemistry Dashboard [37, accessed on 2020/07/13]. Binary topological fingerprints were generated from the provided InChI identifiers as described above.

For benchmarking we downloaded the Tox21 Challenge dataset, and the Side Effect Resource (SIDER) dataset from MoleculeNet [38], a database of benchmarking datasets for classification problems in molecular machine learning.

For training the deep neural networks (AE and FNN models) we enabled early stopping and fall-back mechanisms (model checkpoints). A k-fold cross-validation was applied for training the FNN model, with a default value of $k = 5$.

A hyperparamter tuning (gridsearch) of batch sizes (64, 128 and 256), different optimizers (SGD, Adam), and learning rates (0.01 and 0.001) was applied to each individual target model. The best performing values were selected in the final models for further prediction tasks (see above). Validation loss was monitored. The training procedure was stopped if it did not improve by a certain value min_{delta} (default: 20, can be adjusted through program settings) for a provided number of epochs (patience). The model weights were restored to the respective checkpoint model. Subsequently, a full model containing the complete training set was trained with the same settings. The weights of *all* k models, the *best*, and the *full* model were saved in `.hdf5` format. Training histories were stored in `.csv` format, and a respective visualization showing accuracy, loss, validation accuracy, and validation loss across the training epochs was provided as an `.svg` image. The whole training procedure is logged. Logged metrics were accuracy and validation accuracy per training epoch (FNNs only), loss and validation loss per training epoch, prediction results of test data, area under the curve (ACC), and matthews correlation coefficient (MCC).

Implementation. `deepFPlearn` was implemented as a Python (version 3.7.6) package with three different

usage-modes. The first mode is *convert* which imports the dataset (for training or prediction) and calculates chemical fingerprints for all structures from their respective SMILES or InChi representation. A combined Pandas dataframe is created that contains the SMILES/InChi representation, the calculated fingerprint, and all targets. This dataframe object is then serialized to disc as a Pickle file which has the advantage that importing data is fast for subsequent sessions. The second mode in `deepFPlearn` is *training*. The neural networks can easily be (re-)trained with any dataset that associates chemical structures with an effect. During the training all necessary information is logged to validate and evaluate the trained models. Finally, in its third mode, `deepFPlearn` can be used to *predict* the associations between a provided list of chemical compounds and effects using a selected (pre-trained) model. This model can either be one of our provided models that predicts associations to estrogen, androgen, gluccocorticoid, thyroid receptors, peroxisome proliferator-activated receptor gamma, or aromatase, or an additionally trained model using `deepFPlearn` training mode and a custom dataset.

All settings regarding the neural networks can be globally adjusted, and the individual mode of action as well as respective individual settings can be provided in the form of a JSON configuration file.

Dependencies to external libraries and software are managed using `conda`[1]. The respective `YAML` configuration file for the platform-independent conda environment is part of the source code repository. A singularity container[2] was set up that encapsulates the whole project at the state of publication, including required resources, source code, compiled package, and test data. The `deepFPlearn` software package, conda environment and a singularity container are publicly available, details see section 6.

3 Results

We aimed to develop a stand-alone, ready-to-use deep learning approach to associate chemicals with targets on the gene/pathway level. Further, we aimed to evaluate the potential of feature compression to increase the applicability of this approach to substances beyond the limited amount of available training data. We implemented our ideas in the program `deepFPlearn`, and describe the results of our training procedures and subsequent validation in the following.

Our workflow combined a feature compression step using a deep autoencoder (AE) followed by a classification step using a deep feed forward neural network (FNN), see Figure 1. The research questions at hand were: i) what is the impact of the feature compression prior to the classification, ii) how much structural variety can we capture in this prepended step without a negative impact on the subsequent prediction quality, iii) how is the performance of our models compared to existing approaches, and iv) can we predict feasible results on unknown data? We applied three different training settings based on the selected datasets S and D (details see methods 2 Different datasets), and evaluated the respective performance: First, feature compression was disabled (*no* AE, Figure 1 path from A to C) such that only the FNNs were trained for each individual target of the S dataset using the full-length molecular fingerprints. Second, a *specific* feature compression was applied to the chemical fingerprints of the S dataset followed by training the FNNs using the compressed input features (*specific* AE, Figure 1 from A over B(S) to C(S)). Third, a *generic* feature compression was applied to the chemical fingerprints of the D dataset which is based on the entire repertoire of chemicals stored in the CompTox Database. Subsequently, the FNNs were trained using the compounds together with available target association data from S (*generic* AE, Figure 1 from A over B(D) to C(S)).

The deep feed forward neural networks were trained in a 5-fold cross validation. The weights of the *best*, and the *full* models for each target are saved in `.hdf5` format. These can be restored in the prediction step, and consequently be used for final validation and further application.

Algorithm 1 shows a formal description of the full `deepFPlearn` workflow for training.

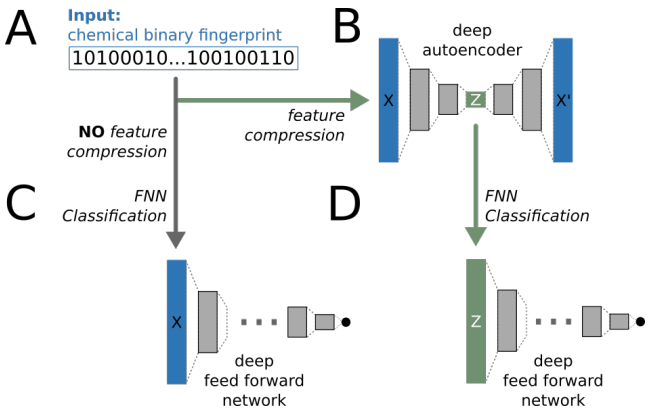


Figure 1: The `deepFPlearn` workflow. A) The binary chemical fingerprints serve as input for the neural networks. B) A deep autoencoder (AE) is used to compress the binary fingerprints. C) A deep feed forward network (FNN) is used for direct classification of the input. D) A deep feed forward network (FNN) is used for classification of the compressed input. Sizes of layers, activation and loss functions are different for each network and depend on the input size, see methods section.

Prediction *without* feature compression performed worst. The training stopped early before 300 training epochs were reached, and the models were perfectly able to predict the 0's — meaning no association between chemical and effect — in the validation

[1] `https://www.anaconda.com`
[2] `https://sylabs.io/`

Algorithm 1: Simplified workflow of the training procedure: Training data is provided and optionally an autoencoder (AE) is trained or its weights are loaded. When an AE is used, features are compressed prior to classification with the feed-forward network (FNN).

Data: X feature matrix and Y target matrix
Result: Train `deepFPlearn` in 5-fold cross validation, return AE models, best and full models of FNNs

```
if compressFeatures then
    if trainAE then
        encoder = trainAutoencoder(X)
    end
    else
        encoder = loadAutoencoder(options.encoder-weights-file)
    end
    trainFeatures = compressFeatures(X, encoder)
end
else
    trainFeatures = X
end
foreach target t do
    (X_t, Y_t) = drop_na(trainFeatures, Y, t)
    scores = []
    models = []
    while (train, test) in Kfold(k = [1 : 5], X_t, Y_t) do
        // model of current fold
        models[k] = trainFNN(X_t[train], Y_t[train])
        scores[k] = model.evaluate(X_t[test], Y_t[test])
    end
    // best model
    save(models[scores.maxIdx(scores[MCC])])
    // train and save full model
    modelFull = trainFNN(X_t, Y_t)
    save(modelFull)
end
```

data, when no feature compression was applied. Training accuracy and loss values almost reached a value of 1, and 0 respectively, see Figure 3 A - top panel (uncompressed). The good performance on the test data right after the first training epoch can be explained by the over-represented fraction of 0's in the validation and training sets. All these issues indicated that the models learned the training data by rote and that the imbalance of 0 and 1 associations had a substantial effect on the final performances of the models. To confirm that we applied a down-sampling procedure. We sampled from the 0-associated compounds such that the ratio between 1's and 0's in the training set is $0.5, 0.6, \ldots, 1.0$. We observed that we could reduce the previously observed issues by downsampling the 0's at the expense of validation accuracy and loss (data not shown).

Feature compression improved training performance and preserved relevant structural information. Two different deep auto encoders (AE) were trained: one using the *specific* dataset S, and the other using the entire repertoire of $\sim 770k$ chemical structures that can be downloaded from the CompTox Database as a more *generic* dataset D. Topological fingerprints of size 2048bit served as input and latent space dimensions of $L_z \in [1024, 512, 256]$ were tested. Best training performances could be reached for $L_z = 256$ in both cases. The *specific* AE trained for 130 epochs, while the *generic* AE trained for 160 epochs. Both reached a validation loss of 0.215. See Figure 2 A for the training histories of the AEs. We could further show, that relevant (structural) information is preserved during feature compression. See Figure 2 B for the respective UMAP visualization of uncompressed and compressed features and subsequent k-means clustering. Compounds from the S dataset that clustered in the uncompressed feature space, also clustered (in the same clusters) in the latent space of the AE.

FNN training with *specific* feature compression yielded good prediction performance. The second variant of training the FNNs used feature compression. In particular, features were compressed using the encoder of the *specific* autoencoder. Subsequently, the deep feed forward networks were trained individually for each target with a 5-fold cross validation. We observed that, with around 650 epochs, the FNN training with compressed feature inputs took much longer compared to our first training variant where we did not use feature compression, see Figure 3 A – top panel. Validation accuracy and loss values were above 0.8 and below 0.2, respectively (on average across the five folds of the cross validation, see Figure 3 A – middle panel (specific AE)). The areas under the curve reached values above 0.82. Specificity values were above 93 % for all gene targets, and we reached up to 61 % sensitivity.

FNN training with *generic* feature compression preserved good prediction performance. In the third setting features were compressed using the encoder of the *generic* autoencoder. We observed that the number of training epochs for training the FNN models increased slightly up to 720, see Figure 3 B. The training and prediction quality was comparably high with a slight drop of only 1% on average in terms of ROC-AUC, compared to the feature compression using the *specific* AE, see Figure 3 B. The ratio between positive (1) and neg-

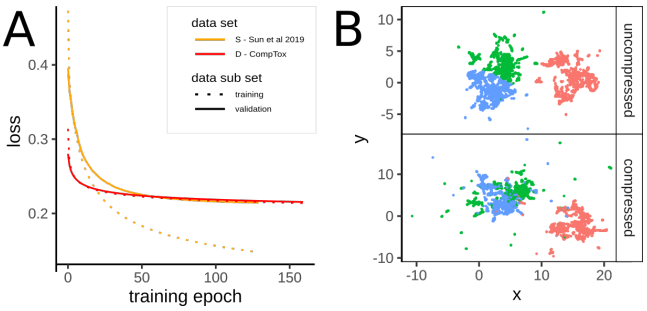


Figure 2: A) Training and validation loss during training of the *generic* (D – CompTox) and the *specific* (S – Sun et al 2019) autoencoder. B) UMAP visualizations of uncompressed and compressed representations of all compounds from S dataset; color indicates cluster assignment of a k-means clustering with $k = 3$ on the uncompressed features.

ative (0) associations between chemical and gene target differs substantially between the individual targets, see Figure 3 C. As expected models with a more balanced class distribution perform better, compare Figures 3 B and C The ED model is an exception here. It can be considered as better balanced on the one hand, but as a more generic predictor on the other hand.

Training and prediction of benchmarking datasets confirmed our concept.

As a proof of our concept regarding the feature compression using the AE in general, and the potential to extrapolate to a larger set of compounds using the *generic* AE in particular, we tested our approach on a selection of benchmarking datasets for molecular machine learning provided by MoleculeNet [38]. We selected the Tox21 Challenge dataset which also associates chemicals and gene targets, and the Side Effect Resource (SIDER) database that associates drugs with grouped adverse drug reactions. These datasets contain 7831 and 1427 compounds, and 12 and 27 targets, respectively, and comprised binarized associations between those compounds and the targets. The same issues as for our initial training dataset also hold for these datasets, i.e. many associations were not available (NA values), and there is a high imbalance between the number of class associations, i.e. we observed an 1/0 ratio of 0.05 (Tox21) and 1.9 (SIDER) (on average across all targets), respectively. We followed the recommended [38] metric and splitting patterns to generate training data from these datasets. Our trained models performed slightly better than described in the MoleculeNet and as implemented in the `deepchem` library. In particular, we reached a maximum of ROC-AUC of 0.86 in the Tox21 case, and 0.68 in the SIDER case, see Figure 4 A and B, where MoleculeNet reported 0.83 and 0.67 for the deep learning approaches implemented in `deepchem`. Additionally, as expected we observed only a small drop of performance loss on the benchmarking datasets when we used our *generic* AE for feature compression, see Figure 4 C.

`deepFPlearn` is ready to be applied to huge datasets.

We used `deepFPlearn` with generic feature compression and a selection of the trained models (ER, AR, and ED) to identify potential further associations among the huge amount of compounds in the CompTox D dataset. For the majority of those compounds the probability to act as endocrine disruptors at all was not known. From the investigated 719 997 compounds, 51 467 (AR), and 61 336 (ER) were returned to be associated with at least one of these two targets with a probability $P > 0.75$.

Recently, Escher et al. [39] classified the association between selected compounds in six of the high throughput Tox21 assays, including AR and ER, whether the observed effect was likely due to specific action via the MoA under consideration or rather due to unspecific consequences of cytotoxicity.

Recently, Escher et al. [39, 40] classified the association between selected compounds in six of the high throughput Tox21 assays, including AR and ER. The aim has been to evaluate whether the observed effect was likely due to specific action via the mode of action (MoA) under consideration or rather due to unspecific consequences of cytotoxicity. We selected the 141 and 103 compounds which had been investigated and classified for androgen and estrogren receptor, and for which we have a deepFPlearn prediction probability in the D dataset as described above. `deepFPlearn` was able to significantly (according to a Kolmogorow-Smirnow test of the respective probability distributions) classify the selective compounds (specificity ratio in relation to experimental cytotoxicity > 10) best. For all other compounds (with specificity rations < 10) `deepFPlearn` returned unsecure probabilities (< 0.5), see Figure 4 D.

Further, we manually selected some predicted associations (ED model), that were not part of the S (train) data set, among the top 200 (most probable positive predictions) and bottom 200 (most probable negative predictions) and empirically investigated their biological feasibility. In our positive predictions we found compounds like Estriol, 17alpha-Ethinylestradiol, 17beta-Ethinylestradiol, or Mestranol, and Prednisolone, Dexamethasone, Betamehtasone and respective derivates that are well known to interact with the human estrogen receptors and pathways, or to interact with the glucocorticoid pathway. Interestingly, some of those were also identified to selectively interact with the androgen receptor in cell assay screenings [39]. Among the candidates with the highest probabilities for the association to any of the endocrine disruption targets we found compounds like Ezlopitant dihydrate dihydrochloride, 5-Bromo-2,2-diethyl-5-nitro-1,3-dioxane, or Schinifoline, a metabolite of the Japanese Pepper plant Zanthoxylum schinifolium. Those have not been tested in bioassays so far. Among the candidates with the lowest probablities ($P < 0.01$) we identified derivates of carbamic, acetic, or amino acids, that have not yet been discussed to be connected to steroid hormone related endocrine disruptions.

4 Discussion

There is a great need for systematic prediction of chemical-effect associations in the field of toxicology. In particular for the prioritization of chemicals for experimental screening, a smart selection of chemicals for monitoring, and for the design of future chemicals. Although a number of approaches and implementations exist that, in part, address these issues, there is no actual tool that can be readily applied on a large scale. With `deepFPlearn` we present a ready-to-use application that can be used to investigate any set of chemicals for their potential associations to gene targets that are involved in endocrine

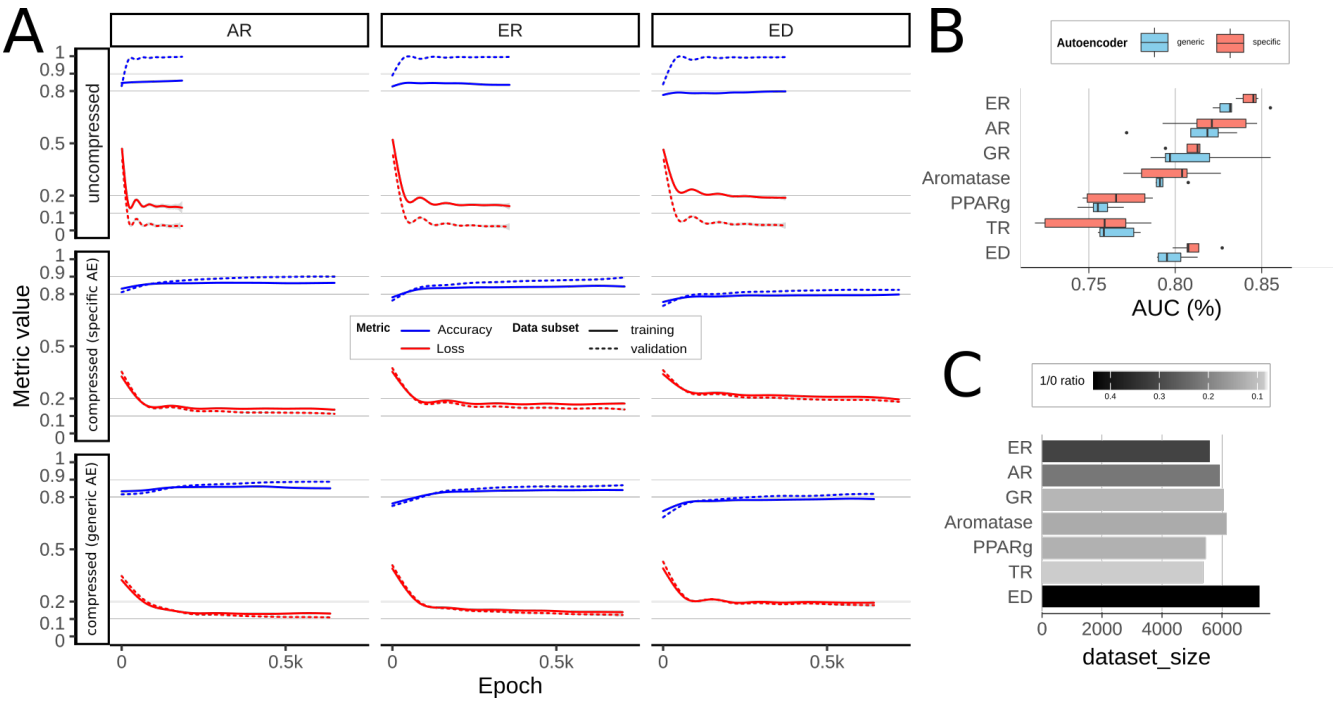


Figure 3: A) Training histories of the feed forward neural networks stratified by the selected targets/models for androgen (AR) and estrogen (ER) receptors, and endocrine disruption (ED), and the degree of feature compression (uncompressed, specific, and generic); shown metrics are accuracy (dotted, blue), validation accuracy (straight, blue), loss (dotted, red) and validation loss (solid, red) during training. The variance in metric values per epoch results from different folds of the 5-fold cross validation. B) Comparison of the AUC values of individual models using the specific and generic AE. C) Sizes of the individual data sets per target, stratified by the ratio of 1 and 0 associations.

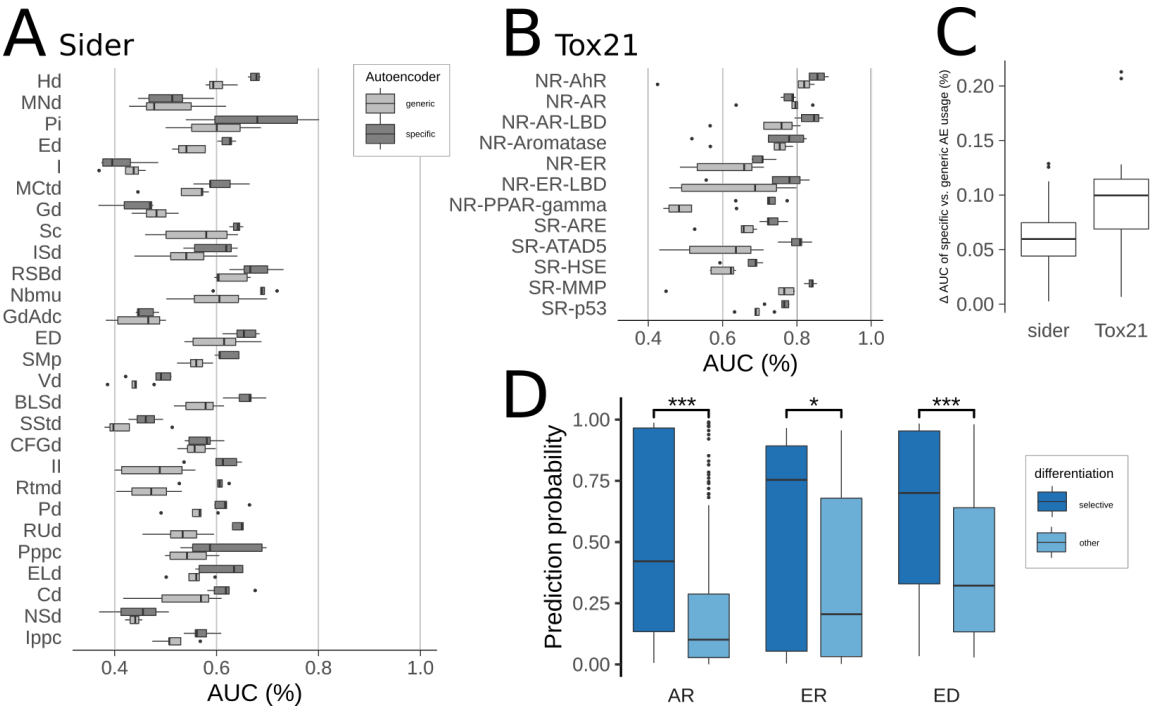


Figure 4: Comparison of the AUC values of individually trained models of the selected benchmarking data – A) Sider, and B) Tox21 – using the specific and generic AE. The abbreviations of the gene/pathway targets are extracted from the original publications. Their interpretation is not required in the scope of this study. C) Comparison of the AUC difference between specific and generic AE of the selected benchmarking data sets. D) `deepFPlearn` prediction probability using AR, ER, and ED model of compounds that have been experimentally measured for quantified target association and respectively differentiated into selective and non-specifically acting compounds. Probability distributions are compared using the Kolmogorow-Smirnow test, and significance levels for rejecting the null hypotheses that both distributions are similar were: ***, **, and * for p_values below 0.001 and 0.01, and 0.05, respectively.

disruption. It is a deep learning approach, and we provide the possibility to retrain the integrated models to any kind of further associations of interest.

The small number of available labeled training data stands in large contrast to the high number of features that are necessary to describe the molecular structure of a chemical. We decided to use the topological fingerprint of size 2048bit since it had been shown that this variant of the fingerprint returns the most promising results in different machine learning approaches [29]. The natural interaction of chemicals and biomolecules is also biased towards "no interaction" (label of 0) such that the data suffers from a substantial imbalance between 1 and 0 labels. For only a handful of chemicals – compared to the whole universe of chemicals – the association to a gene target has been measured and classified in low or high throughput experimental tests. There is a lack of available labeled training data since the methods to measure associations between chemicals and biomolecules are still lab-based and are therefore very slow, even if applying high-throughput technologies, and cannot cope with the high speed in which new chemicals are released. To make things worse, many positive associations (label of 1) are potentially wrong due to mistakes during the interpretation of the screening results, e.g. due to unclear effect tresholds, high variability, and respective limitations in statistics, or in the experimental setups, e.g. due to volatility or solubility of compounds in the respective test systems [41]. We showed that the deep learning models overfit due to the above issues, and that the correction of the imbalance decreased this observation at the expense of training performance. We were able to overcome the discrepancy between large descriptor size (fingerprint) and the limited amount of training data by introducing feature compression using a deep autoencoder on the simple binary fingerprint representation prior to the actual prediction of the association to an effect. Further, we decreased the problem of domain extrapolation by training the autoencoder on a huge repertoire of chemicals structures without reducing the predictive power of the subsequent classification with a deep feed forward network.

We tested different training situations, i) without using feature compression, ii) training an autoencoder with the chemicals of a specific subset of chemicals, and iii) training the autoencoder on a huge set of chemical structures. We could prevent our subsequent classification models from overfitting by using the specific or the generic autoencoder. Further, we reached good training performances with ROC-AUC values above 80%, with satisfying sensitivity up to 63%, and specificity up to 97%. These values drop by only 1% with the integration of the generic autoencoder, which allows to capture a much wider range of chemical structures without reducing the prediction quality of our classification models.

The application to the large CompTox dataset returned a high number of chemicals classified as associated or not to selected gene targets. Among the positive predictions of the large CompTox data set we found compounds that are well known to interact with human estrogen receptors and pathways, or the glucocorticoid pathway but have not been part of our training set. Examples are Estriol, 17alpha-Ethinylestradiol, 17beta-Ethinylestradiol, or Mestranol, and Prednisolone, Dexamethasone, Betamehtasone and respective derivates. Our results also provide lists for candidates that have not yet been tested in bioassays but might be cross-checked or prioritized for testing regarding their involvment in endocrine disruption. Examples are Ezlopitant dihydrate dihydrochloride, 5-Bromo-2,2-diethyl-5-nitro-1,3-dioxane, or Schinifoline, which were the three candidates with the highest probabilities for an ED potential in our analysis. Likewise, on the bottom of the lists we identified substances with probapilities < 0.01, such as derivates of carbamic, acetic, or amino acids, that have not yet been discussed to be connected to steroid hormone related endocrine disruptions. This confirms our prediction quality, and enhances the trust in our models. Our high values for specificity also suggest a potential application of `deepFPlearn` to exclude undesirable secondary effects during drug design. We suggest a more detailed analysis of the predicted associations and respective experimental validation in up-coming studies to confirm or decline potential secondary effects in endocrine disruption.

5 Conclusion

With `deepFPlearn` we were able to draw the associations between chemical structures and effects on the gene/pathway level with our deep learning approach. In contrast to existing approaches and implementations, `deepFPlearn` is a ready-to-use tool. It comes as a stand-alone Python software package, and (additionally) wrapped in a Singularity Container to overcome the dependency on the operating system and required software. `deepFPlearn` is able to capture a much wider range of chemical structures than those contained in the training data of the classification network, and it can be applied to classify hundreds of thousands of chemicals in seconds. With its different application modes we provide the flexibility to train custom models with any meaningful dataset that associates chemicals with an effect. We are convinced that `deepFPlearn` substantially contributes to the systematic *in silico* investigation of chemicals, e.g. for their associations to certain effects that have not been considered so far. With `deepFPlearn` we can cope with the large, constantly and rapidly growing chemical universe and support a prioritization of chemicals for experimental testing, assist in the smart selection of chemicals for monitoring and contribute in the sustainable design of future chemicals.

6 Declarations

Availabilty of data and materials

We provide the `deepFPlearn` Python package on GitHub `https://github.com/yigbt/deepFPlearn`. The software is distributed under the GNU General Public Licence (Vers. 3, 29th June 2007). We further provide a ready-to-use Singularity Container image containing all required tools and libraries as well as the `deepFPlearn` package in the first release state, which built the basis of this manuscript at `https://cloud.sylabs.io/library/mai00fti/default/dfpl.sif`.
All datasets used for training, and the respectively trained models are publicly available as release-asset in the package repository on GitHub.

Competing interests

The authors declare that they have no competing interests.

Funding

Author's contribution

JS and JH planned the study; JS and CL defined the neural network architectures; JS pre-processed all data; JS, PS and MB implemented the software package; All authors analyzed the results and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

References

[1] *CAS Registry.* `https://www.cas.org/support/documentation/chemical-substances`. Accessed: 2020-11-13.

[2] Fischer S, 2017. *KEMI Market List.* `http://doi.org/10.5281/zenodo.3959394`. Version NORMAN-SLE-S17.0.1.4.

[3] European Environment Agency, 2019. *The European environment — state and outlook 2020.* Tech. rep. ISBN 9789294800909.

[4] Barupal D & Fiehn O, 2019. *Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach. Environ Health Perspect*, 127(9). doi:doi:10.1289/EHP4713.

[5] Schwarzenbach R, Escher B, Fenner K, Hofstetter T, Johnson C, von U Gunten, & Wehrli B, 2006. *The Challenge of Micropollutants in Aquatic Systems. Science*, 313(5790):1072–1077. doi:10.1126/science.1127291.

[6] Busch W, Schmidt S, Kühne R, Schulze T, Krauss M, & Altenburger R, 2016. *Micropollutants in European rivers: A mode of action survey to support the development of effect-based tools for water monitoring. Environmental Toxicology and Chemistry*, 35(8):1887–1899. doi:10.1002/etc.3460.

[7] Norman Network, 2020. *EMPODAT Database.* `https://www.norman-network.com/nds/empodat/chemicalStatistics.php`.

[8] Rappaport SM, 2016. *Genetic Factors Are Not the Major Causes of Chronic Diseases. Plos One*, 11(4):e0154387. doi:10.1371/journal.pone.0154387.

[9] Lim S, Vos T, Flaxman A, Danaei G, Shibuya K, Adair-Rohani H, Amann M, Anderson H, Andrews K, Aryee M, *et al.*, 2012. *A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet*, 380(9859):2224–2260. doi:10.1016/S0140-6736(12)61766-8.

[10] Landrigan P, Fuller R, Acosta N, Adeyi O, Arnold R, Basu N, Baldé A, Bertollini R, Bose-O'Reilly S, Boufford J, *et al.*, 2018. *The Lancet Commission on pollution and health. Lancet*, 391(10119):462–512. doi:10.1016/S0140-6736(17)32345-0.

[11] Köhler HR & Triebskorn R, 2013. *Wildlife ecotoxicology of pesticides: Can we track effects to the population level and beyond? Science*, 341(6147):759–765. doi:10.1126/science.1237591.

[12] Hallmann CA, Foppen RPB, Van Turnhout CAM, De Kroon H, & Jongejans E, 2014. *Declines in insectivorous birds are associated with high neonicotinoid concentrations. Nature*, 511(7509):341–343. doi:10.1038/nature13531.

[13] Desforges JP, Hall A, McConnell B, Rosing-Asvid A, Barber JL, Brownlow A, De Guise S, Eulaers I, Jepson PD, Letcher RJ, Levin M, Ross PS, Samarra F, Víkingson G, Sonne C, & Dietz R, 2018. *Predicting global killer whale population collapse from PCB pollution. Science*, 361(6409):1373–1376. doi:10.1126/science.aat1953.

[14] Lemm J, M, Globevnik L, Stefanidis K, Panagopoulos Y, van Gils J, Posthuma L, Kristensen P, Feld C, Mahnkopf J, Hering D, & Birk S, 2021. *Multiple stressors determine river ecological status at the European scale: Towards an integrated understanding of river status deterioration. Glob Chang Biol*, 27(9):1962–1975. doi:doi:10.1111/gcb.15504.

[15] Posthuma L, van Gils J, Zijp MC, van de Meent D, & de Zwartd D, 2019. *Species sensitivity distributions for use in environmental protection and assessment and and management of aquatic ecosystems for 12 386 chemicals. Environmental Toxicology and Chemistry*, 38(4):703–711. doi:10.1002/etc.4373.

[16] European Commission, 2019. *COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT and THE EUROPEAN COUNCIL and THE COUNCIL and THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS. The European Green Deal.* Tech. Rep. COM(2019) 640 final.

[17] European Chemials Agency, 2019. *Universe of registered substances.* `https://echa.europa.eu/de/universe-of-registered-substances`. Accessed 2020-11-12.

[18] European Chemicals Agency, 2018. *Evaluation under REACH: progress report 2017 — 10 years of experience.* Tech. rep., Helsinki.

[19] Raies AB & Bajic VB, 2016. *In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdisciplinary Reviews: Computational Molecular Science*, 6(April):147–172. doi:10.1002/wcms.1240.

[20] Lepailleur A, Poezevara G, & Bureau R, 2013. *Automated detection of structural alerts (chemical fragments) in (eco)toxicology. Computational and Structural Biotechnology Journal*, 5(6):e201302013. doi:10.5936/csbj.201302013.

[21] Vink SR, Mikkers J, Bouwman T, Marquart H, & Kroese ED, 2010. *Use of read-across and tiered exposure assessment in risk assessment under REACH–a case study on a phase-in substance. Regulatory toxicology and pharmacology : RTP*, 58(1):64–71. doi:10.1016/j.yrtph.2010.04.004.

[22] Perkins R, Fang H, Tong W, & Welsh WJ, 2003. *Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. Environmental Toxicology and Chemistry*, 22(8):1666–1679. doi:10.1897/01-171.

[23] Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'Min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, & Tropsha A, 2014. *QSAR modeling: Where have you been? Where are you going to? Journal of Medicinal Chemistry*, 57(12):4977–5010. doi:10.1021/jm4004285.

[24] Lo YC, Rensi SE, Torng W, & Altman RB, 2018. *Machine learning in chemoinformatics and drug discovery. Drug Discovery Today*, 23(8):1538–1546. doi:10.1016/j.drudis.2018.05.010.

[25] Thomas RT, Paules RS, Simeonov A, Fitzpatrick S, Crofton K, Casey W, & Mendrick D, 2018. *The US Federal Tox21 Program: A Strategic and Operational Plan for Continued Leadership. ALTEX.* doi:doi:10.14573/altex.1803011.

[26] Schmidt CW, 2016. *TSCA 2.0: A New Era in Chemical Risk Management. Environmental health perspectives*, 124(10):A182–A186. doi:doi:10.1289/ehp.124-A182.

[27] Mayr A, Klambauer G, Unterthiner T, & Hochreiter S, 2015. *DeepTox: Toxicity Prediction Using Deep Learning. Frontiers in Environmental Science*, 3. doi:10.3389/fenvs.2015.00080.

[28] Pu L, Naderi M, Liu T, Wu HC, Mukhopadhyay S, & Brylinski M, 2019. *eToxPred: a machine learning-based approach to estimate the toxicity of drug candidates. BMC Pharmacology and Toxicology*, 20(1). doi:10.1186/s40360-018-0282-6.

[29] Sun L, Yang H, Cai Y, Li W, Liu G, & Tang Y, 2019. *In Silico Prediction of Endocrine Disrupting Chemicals Using Single-Label and Multilabel Models. Journal of Chemical Information and Modeling*, 59. doi:10.1021/acs.jcim.8b00551.

[30] Liu X, Gao Y, Peng J, Xu Y, Wang Y, Zhou N, Xing J, Luo X, Jiang H, & Zheng M, 2015. *TarPred: A web application for predicting therapeutic and side effect targets of chemical compounds. Bioinformatics*, 31(12):2049–2051. doi:10.1093/bioinformatics/btv099.

[31] Ramsundar B, Eastman P, Walters P, & Pande V, 2019. *Deep Learning for the Life Sciences.* O'Reilly Media and Inc. ISBN 9781492039839.

[32] Landrum G, 2006. *RDKit: Open-source Cheminformatics.* doi:10.2307/3592822.

[33] McInnes L, Healy J, Saul N, & Großberger L, 2018. *UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software*, 3(29):861. doi:10.21105/joss.00861.

[34] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, *et al.*, 2015. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.* Software available from tensorflow.org.

[35] Chollet F, 2015. *keras.* `https://github.com/fchollet/keras`.

[36] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss, R, Dubourg, V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, & Duchesnay E, 2011. *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research*, 12:2825–2830.

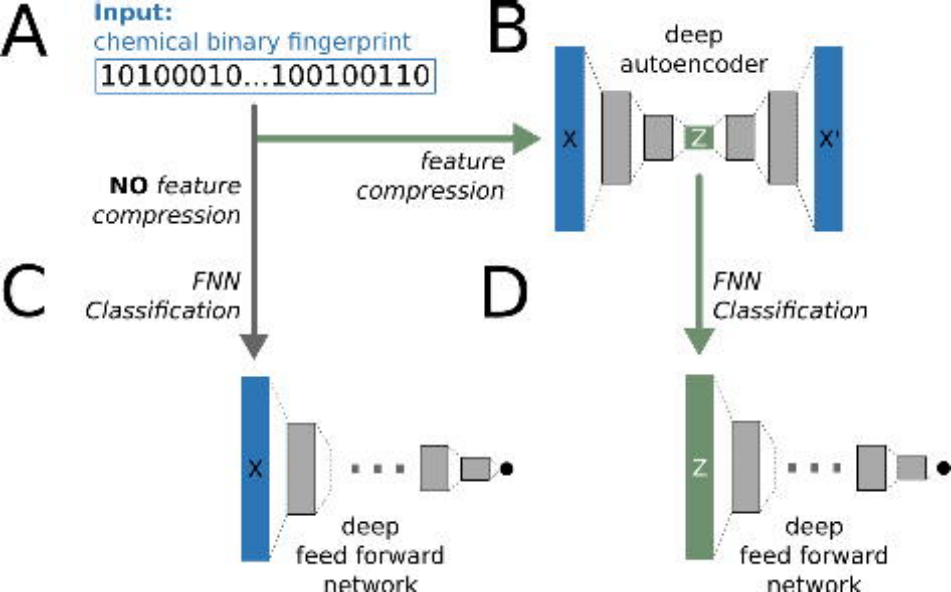
[37] Williams AJ, Grulke CM, Edwards J, Mceachran AD, Mansouri K, Baker NC, Patlewicz G, Shah I, Wambaugh JF, Judson RS, & Richard AM, 2017. *The CompTox Chemistry Dashboard: a community data resource for environmental chemistry Open Access. Journal of Cheminformatics*, 9:61. doi:10.1186/s13321-017-0247-6.

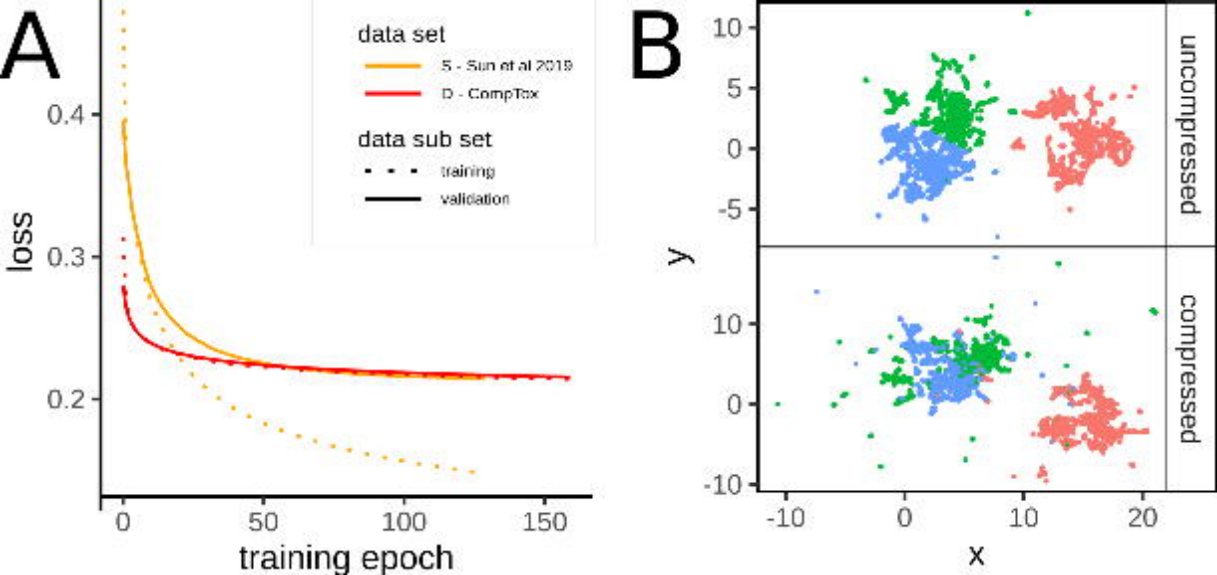
[38] Zhenqin W, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, & Pande V, 2018. *MoleculeNet: A Benchmark for Molecular Machine Learning. Chemical Science*, 2:513–530. doi:10.1039/C7SC02664A.

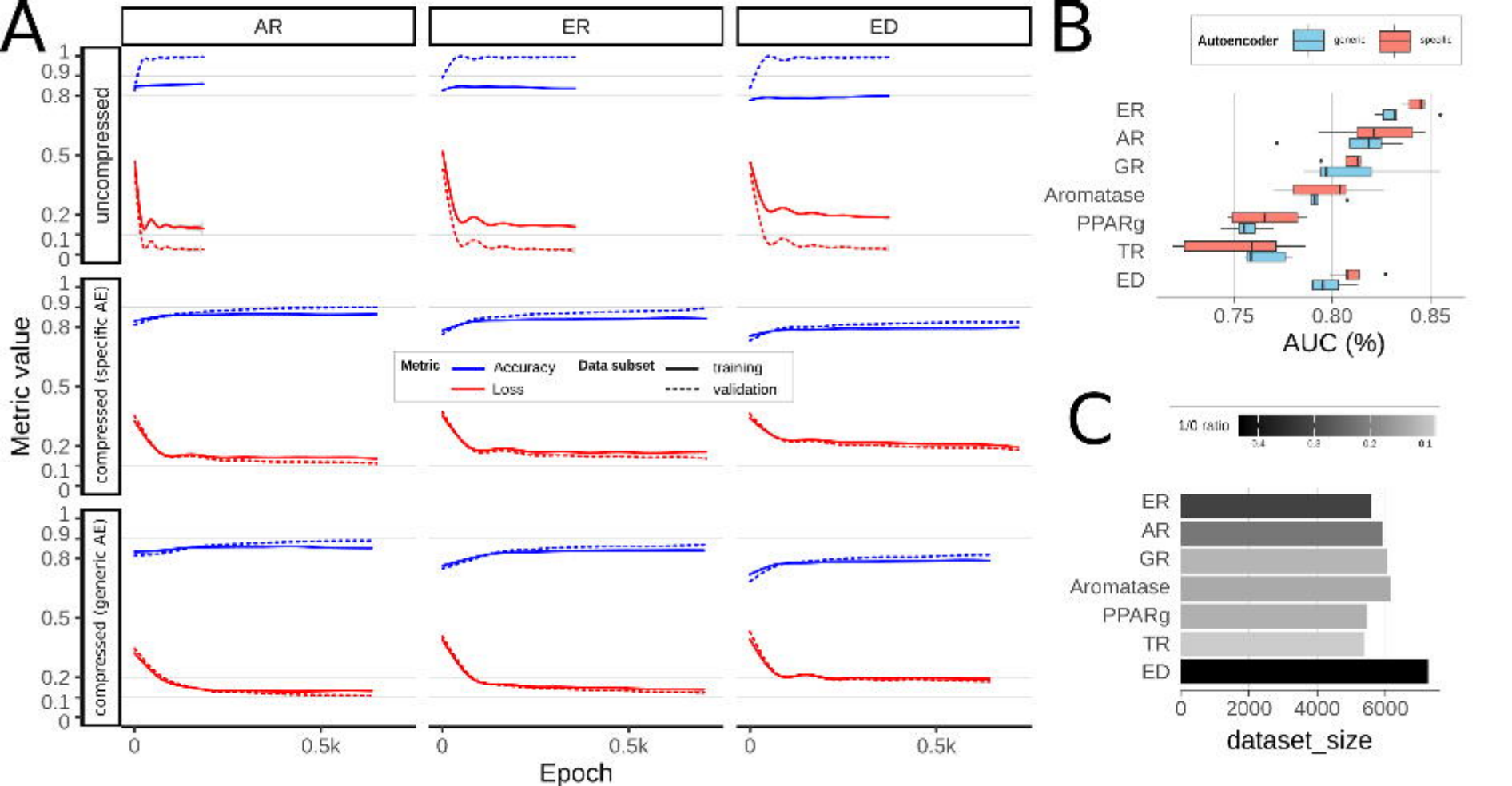
[39] Escher BI, Henneberger L, König M, Schlichting R, & Fischer FC, 2020. *Cytotoxicity burst? Differentiating specific from nonspecific effects in tox21 in vitro reporter gene assays. Environmental Health Perspectives*, 128(7):1–10. doi:10.1289/EHP6664.

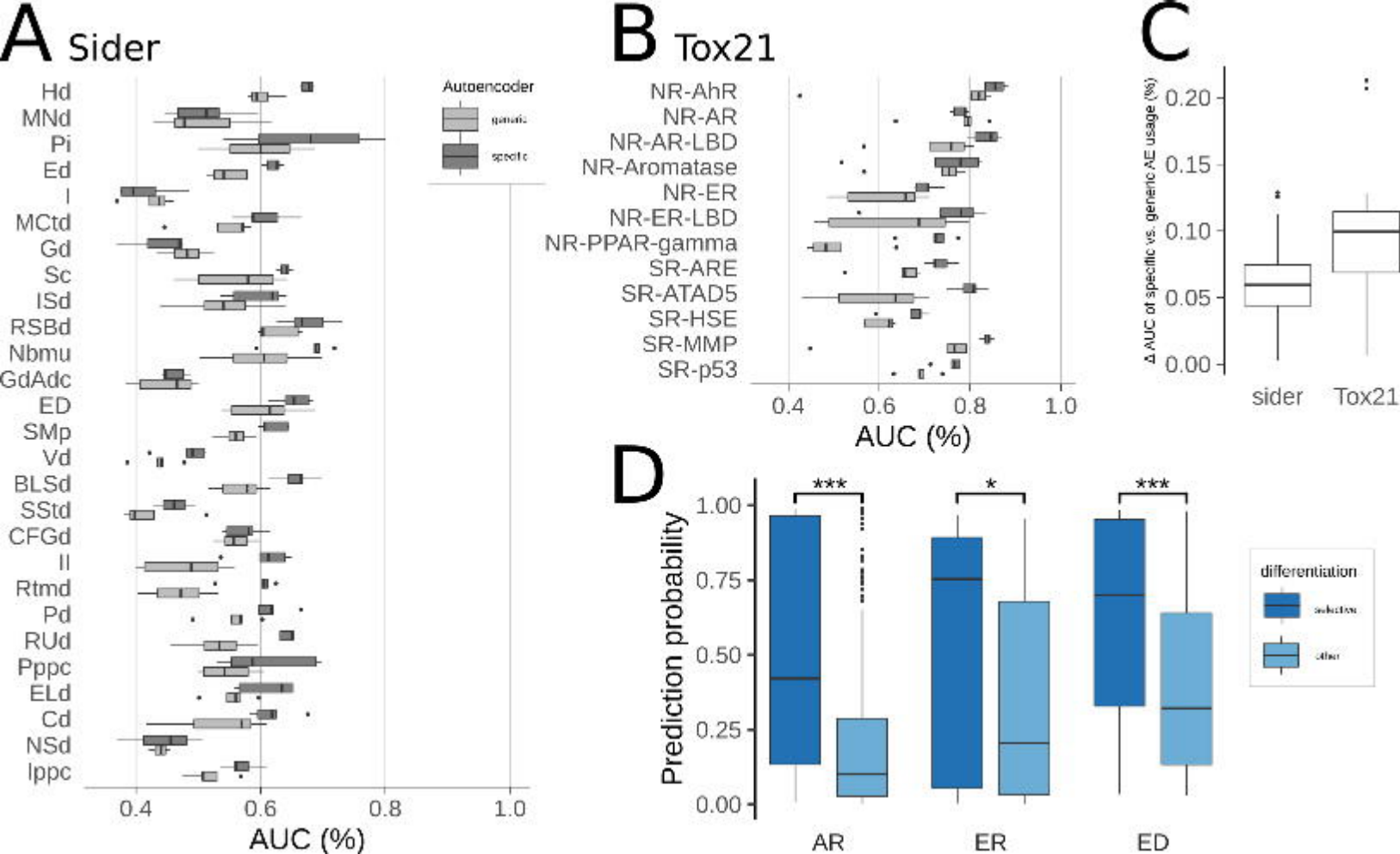
[40] Escher B & Neale P, 2021. *Effect-Based Trigger Values for Mixtures of Chemicals in Surface Water Detected with In Vitro Bioassays. Environ Toxicol Chem*, 40(2):487–499. doi:10.1002/etc.4944.

[41] Huang R, Xia M, Sakamuru S, Zhao J, Shahane S, Attene-Ramos M, Zhao T, Austin C, & Simeonov A, 2016. *Modelling the Tox21 10 K chemical profiles for in vivo toxicity prediction and mechanism characterization. Nat Commun*, 7:10425. doi:doi:10.1038/ncomms10425.


A
Input:
chemical binary fingerprint
10100010...100100110
B
deep
autoencoder
x
z
x'
feature
compression
NO feature
compression
C
FNN
Classification
D
FNN
Classification
x
deep
feed forward
network
z
deep
feed forward
network


A
data set
S - Sun et al 2019
D - CompTox
data sub set
training
validation
loss
0.4
0.3
0.2
0
50
100
150
training epoch
B
uncompressed
compressed
y
x
10
5
0
-5
-10
20


A
AR
ER
ED
uncompressed
compressed (specific AE)
compressed (generic AE)
Metric value
Epoch
Metric
Accuracy
Loss
Data subset
training
validation
B
Autoencoder
generic
specific
ER
AR
GR
Aromatase
PPARg
TR
ED
AUC (%)
C
1/0 ratio
dataset_size


A Sider
Autoencoder
generic
specific
Hd
MNd
Pi
Ed
I
MCtd
Gd
Sc
ISd
RSBd
Nbmu
GdAdc
ED
SMp
Vd
BLSd
SStd
CFGd
II
Rtmd
Pd
RUd
Pppc
ELd
Cd
NSd
Ippc
0.4
0.6
0.8
1.0
AUC (%)
B Tox21
NR-AhR
NR-AR
NR-AR-LBD
NR-Aromatase
NR-ER
NR-ER-LBD
NR-PPAR-gamma
SR-ARE
SR-ATAD5
SR-HSE
SR-MMP
SR-p53
0.4
0.6
0.8
1.0
AUC (%)
C
Δ AUC of specific vs. generic AE usage (%)
0.20
0.15
0.10
0.05
0.00
sider
Tox21
D
Prediction probability
1.00
0.75
0.50
0.25
0.00
***
*
***
AR
ER
ED
differentiation
other