Predicting chemical ecotoxicity by learning latent space chemical representations

Feng Gao\textsuperscript{a}, Wei Zhang\textsuperscript{b}, Andrea A. Baccarelli\textsuperscript{a}, Yike Shen\textsuperscript{a,}\textsuperscript{*}

\textsuperscript{a}Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY 10032, United States

\textsuperscript{b}Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI 48823, United States

Abstract

\textit{In silico} prediction of chemical ecotoxicity (HC\textsubscript{50}) represents an important complement to improve \textit{in vivo} and \textit{in vitro} toxicological assessment of manufactured chemicals. Recent application of machine learning models to predict chemical HC\textsubscript{50} yields variable prediction performance that depends on effectively learning chemical representations from high-dimension data. To improve HC\textsubscript{50} prediction performance, we developed an autoencoder model by learning latent space chemical embeddings. This novel approach achieved state-of-the-art prediction performance of HC\textsubscript{50} with R\textsuperscript{2} of 0.668 ± 0.003 and mean absolute error (MAE) of 0.572 ± 0.001, and outperformed other dimension reduction methods including principal component analysis (PCA) (R\textsuperscript{2} = 0.601 ± 0.031 and MAE = 0.629 ± 0.005), kernel PCA (R\textsuperscript{2} = 0.631 ± 0.008 and MAE = 0.625 ± 0.006), and uniform manifold approximation and projection dimensionality reduction (R\textsuperscript{2} = 0.400 ± 0.008 and MAE = 0.801 ± 0.002). A simple linear layer with chemical embeddings learned from the autoencoder model performed better than random forest (R\textsuperscript{2} = 0.663 ± 0.007 and MAE = 0.591 ± 0.008), fully connected neural network (R\textsuperscript{2} = 0.614 ± 0.016 and MAE = 0.610 ± 0.008), least absolute shrinkage and selection operator (R\textsuperscript{2} = 0.617 ± 0.037 and MAE = 0.619 ± 0.007), and ridge regression (R\textsuperscript{2} = 0.638 ± 0.007 and MAE = 0.613 ± 0.005) using unlearned raw input features. Our results highlighted the usefulness of learning latent chemical representations, and our autoencoder model provides an alternative approach for robust HC\textsubscript{50} prediction.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

\textsuperscript{*}Corresponding author at: Department of Environmental Health Sciences, Columbia University Mailman School of Public Health, P\&S Building, 630 W 168th St Room 16-416, New York, NY 10032, United States. ys3419@cumc.columbia.edu (Y. Shen).

CRediT authorship contribution statement

Feng Gao: Conceptualization, Investigation, Methodology, Formal analysis, Software, Investigation, Resources, Visualization, Writing – original draft, Writing – review & editing. Wei Zhang: Conceptualization, Writing – review & editing. Andrea A. Baccarelli: Funding acquisition, Writing – review & editing. Yike Shen: Conceptualization, Investigation, Formal analysis, Visualization, Resources, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2022.107224.
1. Introduction

Evaluating chemical ecotoxicity (HC$_{50}$) is a valuable approach to identify potential hazardous effects of chemicals on ecosystems and is of great interest in academic and regulatory communities. The traditional gold standard for determining HC$_{50}$ uses in vivo animal models, but such approaches present challenges in the face of rapid increases in the number of chemicals manufactured each year and ethical concerns related to intentional exposure of animals to toxic chemicals (Xia et al., 2008). To circumvent these concerns, the US federal initiative proposed high-throughput in vitro chemical screening to efficiently identify hazardous compounds (Toxicology in the 21st Century, TOX21) (National Research Council, 2007). However, this approach requires laborintensive laboratory procedures that cannot keep pace with the manufacturing of new chemicals. In addition, current in vitro methods cannot capture complex effects such as endocrine effects and have limited ability to predict toxicity in in vivo systems through in vitro to in vivo extrapolation models (Chang et al., 2015; Laue et al., 2020). Hence, selected chemicals must also go through in vivo animal testing to determine HC$_{50}$, highlighting the importance of chemical screening.

One approach to improve screening efficiency and minimize the number of chemicals needed for in vivo animal testing is using in silico models to predict chemical toxicity (e.g., effective concentration 50% [EC$_{50}$] and lethal concentration 50% [LC$_{50}$]). Previous studies predicted chemical acute toxicity based on Abraham parameters and mode of action (MoA) for a diverse set of compounds using quantitative structure-activity relationship (QSAR) models (Barron et al., 2015; Boone and Di Toro, 2019a,b). Barron et al. (2015) used interspecies correlation estimation models to estimate EC/LC$_{50}$ values; Boone and Di Toro (2019a,b) utilized a target site model to predict chemical acute toxicity to aquatic organisms based on polyparameter linear free energy relationship. However, traditional QSAR models are usually linear; such models are not effective in capturing more complex, nonlinear relationships (Cherkasov et al., 2014; Gramatica, 2007; Hou et al., 2020b; Tropsha et al., 2003).

Recent developments in machine learning models can overcome the limitations of traditional predictive models for chemical ecotoxicity and reactivity prediction (Hou et al., 2020a; Hou et al., 2020b; Mansouri et al., 2021; Raza et al., 2019; Su and Rajan, 2021), which is increasingly valuable in the ever-expanding chemical database. Hou et al. (2020a) applied multiple machine learning models to predict HC$_{50}$ values for 2307 chemicals based on 14 chemical features. Hou et al. (2020b) further developed genetic algorithm optimized neural network models to predict HC$_{50}$ based on 691 calculated chemical features on a dataset of 1815 chemicals. Takata et al. (2020) classified chemicals into four Verhaar schema class categories from the ecological structure–activity relationship (ECOSAR) model and predicted chemical class with ensemble learning methods. Additionally, Raza et al. (2019)
used fully connected neural network (FCNN) models to predict C-F energies in per- and poly-fluoroalkyl substances (PFAS). Yao et al. (2021) developed a variational autoencoder model for the design of nanoporous crystalline reticular materials, which is a generative model that outputs parameters of a pre-defined distribution in the latent space. These studies mostly focused on supervised learning tasks (e.g., classification or regression) and achieved better prediction accuracy than traditional QSAR methods. Additionally, unsupervised models such as t-distributed stochastic neighbor embedding (t-SNE) were used to explore patterns in a PFAS dataset based on chemical properties (Raza et al., 2019; Su and Rajan, 2021). Further developments in machine learning can offer a promising opportunity to develop new predictive models in this field.

In general, the success of machine learning and deep learning algorithms heavily depends on data representations (features) to which they are applied (Bengio et al., 2013; LeCun et al., 2015). Representation learning refers to learning data representations by extracting useful information from input features. Although machine learning models benefit from learning complicated relationships from enormous input features, redundant features may introduce noises and result in poor model performance – known as the curse of dimensionality (Bellman, 2015). Indeed, the curse of dimensionality is a significant challenge in developing machine learning models with high-dimension data, leading to increased errors with increasing number of features (Bellman, 2015).

The curse of dimensionality can be mitigated through dimensionality reduction, which reduces the number of features while maintaining useful information (Bengio et al., 2013). In general, dimensionality reduction methods fall into linear and nonlinear groups. Commonly used linear dimensionality reduction methods include factor analysis and principal component analysis (PCA). Linear methods, however, often do not work well for data with intrinsic nonlinear relationships because it is difficult to successfully embed nonlinear data in the latent space. Nonlinear relationships are instead addressed through nonlinear dimensionality reduction methods such as multidimensional scaling (MDS) (Mead, 1992), t-SNE (Van der Maaten and Hinton, 2008), and uniform manifold approximation and projection (UMAP) (McInnes et al., 2018). However, these nonlinear dimensionality reduction methods are designed to preserve a predefined pairwise distance or local/global structures rather than learning meaningful data representations and are mainly used for visualization purposes.

A novel alternative is to learn low-dimension latent space representations through autoencoders (Bengio et al., 2007; Vincent et al., 2008). Autoencoder is an unsupervised representation learning method that can perform nonlinear dimensionality reduction. Instead of aiming to preserve distances or local structures, it learns latent space embeddings. These embeddings contain representations of data that are more meaningful or comprehensive than predefined distances. An autoencoder consists of a pair of encoder and decoder that can be parameterized by neural networks (Fig. 1). The encoder reduces the high-dimension input features to lower-dimension embeddings, while the decoder reconstructs the input features from the lower-dimension embeddings. This process enables the reduction of dimension by compressing data and noises, while preserving essential information for feature reconstruction. The low-dimensional space embeddings are learned from data and
retrieved by minimizing the reconstruction loss. Then, the learned latent space embeddings can be used for various downstream tasks, e.g., supervised classification or regression tasks and unsupervised clustering. Other machine learning models, such as FCNN, do not have this encoding-decoding process (Figure S1). Thus, an autoencoder model is trained through reconstruction loss, while FCNN is trained through prediction loss.

In this study, we developed an autoencoder model to learn chemical latent space representations and predict HC_{50} from the in vivo USEtox database containing 1815 chemicals and their 691 chemical features. Our autoencoder model can perform nonlinear dimensionality reduction through the encoding-decoding process to account for the intrinsic nonlinearity in data and learn meaningful representation of chemical features, which is different from other linear and non-linear methods such as PCA, MDS and UMAP as well as other machine learning models using raw input features. Specifically, we aimed to utilize this model to learn low-dimension latent space representations of chemicals for dimensionality reduction and denoising to alleviate the curse of dimensionality and to improve the prediction accuracy of HC_{50} by machine learning.

2. Materials and methods

2.1. Data

USEtox is a scientific consensus model for evaluating human and ecotoxicological effects of chemicals. USEtox produces a database of chemical hazard concentrations or chemical ecotoxicity (HC_{50}) from the geometric mean of EC_{50} (effect concentration 50%, i.e., the concentration of a chemical when 50% of its maximal effect is observed) and LC_{50} (lethal concentration 50%, i.e., the concentration of chemical at which 50% of a group died during observation period) (Fantke et al., 2017; Hou et al., 2020b; Rosenbaum et al., 2008). Hou et al. (2020b) calculated input features for 1815 chemicals from USEtox (11 physiochemical properties), U.S. Environmental Protection Agency Toxicity Estimation Software Tool (797 theoretical descriptors), and QikProp (51 physically and pharmaceutically significant properties), and assigned the classification of MoA of chemicals by Verhaar scheme using software ToxTree, resulting in 860 input variables. Hou et al. (2020b) further removed 169 highly uncertain and duplicate variables and obtained 691 final input features. The names and sources of the 691 input features were listed in the Supplementary Material B as a feature glossary. The abundant number of chemical features enables the model to learn chemical embeddings. We therefore followed the same inclusion criteria as Hou et al. (2020b) and collected the HC_{50} value of 1815 chemicals and the corresponding 691 features. Dataset was provided in Supplementary Material B.

2.2. Autoencoder

Autoencoder is an unsupervised deep learning model that can be used to compress input data and learn low-dimension latent space embeddings. Our autoencoder model takes 691 features as input, produces the learned chemical embeddings, and predicts the HC_{50} values. The model learns latent space representation of the input features by compressing them into low-dimensional space, and the learned latent space embeddings are then used to reconstruct the input features. This encoding-decoding process keeps
essential information within the low-dimensional space while other less important information is compressed. Specifically, our autoencoder comprises two parts: an encoder that learns a lower dimension embedding of each chemical based on their 691 input features, and a decoder that reconstructs the input features (Fig. 1). For the selection of hidden units, number of layers and the activation function, we tested multiple architectures with different activation functions including ReLU, Sigmoid, and Tanh. The encoder gradually reduces the input dimensions, so we gradually reduced the number of neurons in each encoder layer. The architectures that were evaluated included 691-128-691, 691-256-128-256-691, 691-128-64-128-691, 691-512-256-128-256-512-691, 691-512-256-128-64-128-256-512-691, and 691-512-n-512-691 (the number of neurons, n = 2, 16, 32, 64, 128, or 256). Detailed prediction performances were summarized in Table S1 of Supplementary Material A. Based on the numerical results of these model architectures, the best-performing current architecture (i.e., 691-512-128-512-691) was selected. ReLU is a widely used activation function that is useful to solve the gradient vanishing problem of Sigmoid and Tanh functions (Agarap, 2018; Glorot et al., 2011; Yarotsky, 2017). Indeed, the ReLU activation function had the best performance (Table S1). In summary, our encoder consists of three fully connected layers, with an input layer of 691 neurons, a hidden layer of 512 neurons and an embedding layer of 128 neurons that is capable of reducing the feature dimension to 128. The decoder contains two fully connected layers with one hidden layer of 512 neurons and an output layer of 691 neurons. By forcing the encoder to reduce the dimensions of input features and to learn the embeddings of each sample, we used the learned embeddings to predict HC$\text{50}$. To validate the effectiveness of the learned embeddings, we input the learned embeddings into a simple linear layer for the prediction of HC$\text{50}$ (Fig. 1). Although a more complex neural network with multiple hidden layers and nonlinear activations can be used, a simple linear layer for prediction can better reflect the representative power of learned chemical embeddings. Thus, our autoencoder consists of two processes, i.e., the encoding-decoding process that learns chemical embeddings, and the prediction process using learned embeddings to predict HC$\text{50}$. These two processes were trained jointly. For the encoding-decoding (reconstruction) process, we utilized the reconstruction loss defined as $\text{loss}_{\text{reconstruction}} = \frac{1}{n} \sum_{i=1}^{n} (\text{Feature}_{i,\text{input}} - \text{Feature}_{i,\text{reconstructed}})^2$, which measured the differences between the reconstructed and original features. For the prediction process, we introduced the prediction loss term quantified as $\text{loss}_{\text{prediction}} = \frac{1}{n} \sum_{i=1}^{n} (\text{HC50}_{i,\text{predicted}} - \text{HC50}_{i,\text{measured}})^2$. The prediction loss measured the differences between the predicted HC$\text{50}$ value and the measured HC$\text{50}$ value. Hence the total loss was defined as $\text{loss}_{\text{total}} = \text{loss}_{\text{reconstruction}} + \text{loss}_{\text{prediction}}$. The autoencoder was trained using the Adam optimizer, with L2 regularization weight 0.0005. Evaluation of the trained model on the validation set was performed every 200 epochs and was only tested on the test set if the validation results were improved, which was used to avoid overfitting. We set a maximum of 1500 epochs for training (the optimal epochs were usually observed between 200 and 800 epochs), and a learning rate of 1E-3.
2.3. **Chemical embedding visualization**

To interpret and evaluate the quality of learned embeddings, we first visualized the learned latent space embeddings under their t-SNE coordinates (Van der Maaten and Hinton, 2008). T-SNE provides a tool to visualize high-dimensional data (Van der Maaten and Hinton, 2008). It embeds the local structure of the data and tends to extract clustered local groups of samples. This method converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data, and thus can be used to discover patterns in the data without knowing any HC$_{50}$ values in advance. In the t-SNE plot, chemicals with similar embeddings were near each other, and the HC$_{50}$ value of each chemical was colored with brighter color for greater HC$_{50}$ value and with darker color for lower HC$_{50}$ values (Fig. 2). We also colored the embeddings under the t-SNE coordinates with MoA since it is an important determinant of chemical toxicity (Barron et al., 2015; Boone and Di Toro, 2019a).

2.4. **Other dimension reduction methods and models**

For comparison, we tested multiple popular dimensionality reduction methods including PCA, UMAP, kernel PCA, factor analysis, and MDS. We also applied several popular machine learning methods (LASSO, Ridge regression, RF and FCNN) using original 691 features and compared their performances with that of our autoencoder model.

2.4.1. **Principle component analysis (PCA) and kernel PCA**—PCA is a linear dimensionality reduction method that projects the input features onto a lower dimensional space with a set of orthogonal components and captures the majority of variance (Wold et al., 1987). Kernel PCA is an extension of PCA that achieves nonlinear dimensionality reduction through the use of kernels (Mika et al., 1998; Schölkopf et al., 1997). Here, we used radial basis function (RBF) kernel. We reduced the input feature dimensions to 128 using PCA and kernel PCA.

2.4.2. **Uniform manifold approximation and projection (UMAP)**—UMAP is a dimension reduction technique that can be used for visualization similar to t-SNE, but also for general nonlinear dimension reduction. UMAP constructs a high-dimensional graph representation of the data, and then optimizes a low-dimensional graph to be as structurally similar as possible. UMAP ensures that local structure is preserved in balance with global structure. McInnes et al. (2018) described the mathematics of UMAP. We reduced the input feature dimensions to 128 using UMAP.

2.4.3. **Factor analysis**—Factor analysis assumes that the observed variables are linear transformations of lower-dimensional latent factors with additional Gaussian noises. This method yields a maximum likelihood estimate of the loading matrix, through which latent variables can be transformed to the observed variables using singular value decomposition based approach (Barber, 2012; Kim and Mueller, 1978). We reduced the feature dimension to 128 using factor analysis.
2.4.4. Multidimensional scaling (MDS)—MDS achieves dimensionality reduction aimed to approximate the distances in the original high-dimensional space. General application of MDS is to analyze similarity or dissimilarity in data modeled as distances in a geometric space (Mead, 1992). We reduced the input features to 128 dimensions using MDS.

2.4.5. Least absolute shrinkage and selection operator (LASSO) and ridge regression—Mathematically, LASSO consists of a linear regression model with an added L1 regularization term. The objective function to minimize is:

\[
\min_{w} \left| \left| Xw - y \right| \right|_2^2 + \alpha \left| \left| w \right| \right|_1
\]

Similarly, ridge regression adds an L2 regularization term to the common ordinary least square objective function:

\[
\min_{w} \left| \left| Xw - y \right| \right|_2^2 + \alpha \left| \left| w \right| \right|_2^2
\]

Here, \( w \) are the weights of the linear regression model, and \( \alpha \) controls the regularization strength.

2.4.6. Random forest (RF)—RF is an ensemble learning method that fits multiple decision trees on various subset of the dataset and averages results from each individual tree to improve the predictive accuracy and control over-fitting. RF previously performed best among six machine learning models and five traditional models for predicting HC\(_{50}\) based on 14 chemical properties (Hou et al., 2020a). We built a RF model and carefully tuned hyperparameters including max features (suggested by scikit-learn package: ‘auto’, ‘sqrt’, ‘log2’) and number of trees (100, 200, 300, 500, 1000).

2.4.7. Fully connected neural network (FCNN)—FCNN consists of an input layer that takes the input features (i.e., chemicals), multiple hidden layers, and finally an output layer that makes final predictions of HC\(_{50}\). Nonlinear activation functions are usually used between hidden layers. Each layer contains multiple neurons, with each neuron calculated as:

\[
y = f(b + \sum_{i=1}^{n} w_i x_i)
\]

where \( f \) is the activation function, \( w_i \) is the weight, \( b \) is bias and \( x_i \) is the \( i \)-th input to this neuron. FCNN can learn complex non linear relationships between input features and predicted targets. Here, we developed an FCNN model with one hidden layer. ReLU activation was used between layers. Parameters in the model were updated through optimizing the loss function \( \text{loss}_{\text{prediction}} = \frac{1}{n} \sum_{i=1}^{n} (HC_{50, \text{predicted}} - HC_{50, \text{measured}})^2 \). The Adam optimizer was used to update the weights through back propagation. For the FCNN
model, we used one hidden layer with 512 hidden units and ReLU activation to compare with our HC50 prediction using a linear layer without activation functions based on the learned embeddings from the autoencoder.

### 2.4.8. Ecological structure activity relationships (ECOSAR) predictive model evaluation

The ECOSAR software can be downloaded from the website of US EPA (USEPA, 2022). The software runs on Microsoft Windows systems. We input the CAS ID as batches to the ECOSAR model and extracted the calculated EC50 values across different species. We then obtained HC50 values by calculating the geometric mean of EC50 across multiple species. All chemicals in our dataset that have predictions in ECOSAR were used (N = 1728) to evaluate the performance of ECOSAR model.

### 2.5. Applicability domain

The applicability domain of our trained model can be decided by calculating a distance $$d = \sqrt{\sum_{i=1}^{n} (\text{property}_{i,j} - \overline{\text{property}})^2}$$, where $$\overline{\text{property}} = \frac{1}{n} \sum_{i=1}^{n} \text{property}_{i,j}$$. Hence, d measures the distance between a new chemical and the center of the training dataset based on their physiochemical properties and theoretical molecular descriptors. This distance can be calculated when our model is applied to a new chemical to decide if it is in or outside of the application domain.

### 2.6. Hyperparameter tuning and model validation

We used five-fold cross-validation for hyperparameter tuning and model validation. Five-fold cross-validation enables the maximum efficiency in model selection and learning time (Hawkins et al., 2003; King et al., 2021). The dataset was randomly shuffled into five equivalent subsets. Each subset was used as an independent test set and taken out during the training process. Models were then trained using the remaining four subsets of data. To fine-tune the model and select the best parameter combinations, 12.5% of the training data (10% of the full dataset) were used as a validation subset; the remaining 87.5% of the training data (70% of full dataset) were used to train the model. The best parameter sets were retrieved by examining the performance of trained models on the validation set. The best combination of hyperparameters was then used to build the model that was tested on the held-out test subset. Each subset was therefore tested once. The final prediction results were averaged based on the performances on the five test sets. Finally, our model performance was evaluated by three criteria: $$R^2$$, mean absolute error (MAE), and root mean squared error (RMSE). $$R^2$$ is defined as $$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$, where $$\hat{y}$$ is the predicted HC50 value of i-th sample, $$y_i$$ is the corresponding true value, and $$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$. MAE is defined as $$\text{MAE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$ and RMSE is defined as $$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$. We also computed the standard deviation (±) of each model by repeating the above process five times.
3. Results and discussion

3.1. Comparison of embedding sizes on HC$_{50}$ prediction

Autoencoder model can compress input feature dimensions without losing essential information. The embedding size must be optimized to avoid redundant features or losing essential information (Hinton and Salakhutdinov, 2006). Therefore, we compared different embedding sizes (i.e., 2, 16, 32, 64, 128, and 256) for optimal model selection. An embedding size of 128 achieved the best performance, with $R^2$ of 0.668 ± 0.003, MAE of 0.572 ± 0.001, and RMSE of 0.781 ± 0.005 (Fig. 2). Using a larger embedding size of 256 resulted in a slightly worse performance with $R^2$ of 0.667 ± 0.007, MAE of 0.571 ± 0.005, and RMSE of 0.778 ± 0.009 (Fig. 2). Yet, the performance also declined with decreasing embedding size from 128 to 2 (embedding size 64: $R^2 = 0.657 ± 0.009$, MAE = 0.576 ± 0.006, RMSE = 0.790 ± 0.013; embedding size 32: $R^2 = 0.650 ± 0.007$, MAE = 0.581 ± 0.009, and RMSE = 0.800 ± 0.008; embedding size 16: $R^2 = 0.634 ± 0.006$, MAE = 0.595 ± 0.007, and RMSE = 0.816 ± 0.008) (Fig. 2). At an embedding size of 2, the $R^2$ was 0.632 ± 0.007, with MAE of 0.599 ± 0.005 and RMSE of 0.820 ± 0.008 (Fig. 2). While smaller embedding sizes provided highly concentrated information of chemical features, additional features allowed the autoencoder model to achieve better performance. Therefore, we used an embedding size of 128 in this study.

3.2. Representation of chemicals in low-dimensional latent space

Our autoencoder model effectively learned useful chemical embeddings in low-dimensional latent space to predict HC$_{50}$. To examine the quality of the generated embeddings, we visualized the embeddings under their t-SNE coordinates colored by HC$_{50}$ values (Fig. 3a) and MoA (Fig. 3b). The embeddings themselves reflected the patterns of HC$_{50}$ values, with each point representing one of the chemicals in the dataset (Fig. 3a). Chemicals with similar embeddings should be near each other and thus share similar HC$_{50}$ values. There was a clear pattern of change from smaller HC$_{50}$ values (darker color) to larger HC$_{50}$ values (brighter color). The ability to reflect HC$_{50}$ values based on the learned chemical embeddings suggests that our autoencoder method is effective in learning low-dimension latent space representations of chemical HC$_{50}$. In Fig. 3b, chemicals acting by a specific mechanism (Class 4) were clustered in blue in the right region, corresponding to lower HC$_{50}$ values in Fig. 3a. On the other hand, inert chemicals (Class 1) had higher HC$_{50}$ values and were mostly clustered in green in the left region, separated from Class 4 chemicals. As expected, unclassifiable chemicals (Class 5) had a scattered distribution.

3.3. Autoencoder dimension reduction compared to PCA, kernel PCA, UMAP, factor analysis, and MDS

The effectiveness of the learned chemical embeddings from the autoencoder in terms of dimension reduction and denoising was demonstrated by predicting HC$_{50}$. For comparison, we predicted HC$_{50}$ values using a simple linear regression model in the autoencoder,
factor analysis, PCA, kernel PCA, UMAP, and MDS (Fig. 4 and Figure S2). The average prediction accuracy of the autoencoder method had an $R^2$ of 0.668 ± 0.003. For other methods, the $R^2$ values were 0.601 ± 0.031 and 0.631 ± 0.008 for the 128 PCA and kernel PCA features, respectively, and 0.400 ± 0.008 for the 128 UMAP features (Fig. 4). In addition, the $R^2$ values for factor analysis and MDS were 0.577 ± 0.003 and 0.505 ± 0.007, respectively (Figure S2). The MAE of the autoencoder method was smaller (0.572 ± 0.001) than those of PCA (0.629 ± 0.005) and kernel PCA (0.625 ± 0.006) as well as that of UMAP (0.801 ± 0.002), factor analysis (0.680 ± 0.003), and MDS (0.731 ± 0.003). Similarly, the RMSE ranked as follows: autoencoder (0.781 ± 0.005), kernel PCA128 (0.827 ± 0.008), PCA128 (0.860 ± 0.036), factor analysis (0.886 ± 0.003), MDS (0.957 ± 0.006), and UMAP128 (1.055 ± 0.007). Thus, the autoencoder method performed better than other methods because it benefitted from learning the latent space representation of chemicals from the data. We further colored the results according to chemical types (i.e., acid, base, neutral, or amphoter) and calculated the MAE for each chemical type in the autoencoder model. The MAE was similar for neutral (green), base (orange), and acid (brown) compounds (i.e., 0.56, 0.57, and 0.59, respectively). However, the MAE (0.69) was slightly worse for amphoter (blue) compounds, probably due to more complex mechanisms of amphoters. Additionally, the autoencoder results were colored by MoA (Fig. 5), and the MAE was calculated for each MoA. The MAE was lower for Class 1 chemicals (0.41) and Class 2 chemicals (0.43). However, prediction was more challenging for more toxic chemicals, with larger MAE for Class 3 chemicals (0.60) and Class 4 chemicals (0.72). Finally, the MAE for Class 5 (unclassified chemicals) was 0.63 (Fig. 5).

3.4. Autoencoder compared with LASSO, ridge regression, random forest, and FCNN

Next, we compared our autoencoder model with other common machine learning models including LASSO, ridge regression, RF, and FCNN with the cross-validation procedure identical to that of the autoencoder method. The performance of RF ($R^2 = 0.663 ± 0.007$, MAE = 0.591 ± 0.008, and RMSE = 0.790 ± 0.009) was slightly poorer than that of the autoencoder model ($R^2 = 0.668 ± 0.003$, MAE = 0.572 ± 0.001, and RMSE = 0.781 ± 0.005) (Fig. 4). Although the $R^2$ value of the RF model was comparable to that of the autoencoder model, its MAE and RMSE were worse. Notably, the prediction of the autoencoder method was solely based on one simple linear layer in contrast with a RF model. The performance of the FCNN with one hidden layer ($R^2 = 0.614 ± 0.016$, MAE = 0.610 ± 0.008, and RMSE = 0.839 ± 0.016) was inferior to that of the autoencoder model. Our autoencoder model also outperformed popular linear models LASSO ($R^2 = 0.617 ± 0.037$, MAE = 0.619 ± 0.007, RMSE = 0.842 ± 0.039) and ridge regression ($R^2 = 0.638 ± 0.007$, MAE = 0.613 ± 0.005, RMSE = 0.819 ± 0.009) (Fig. 4). These results further highlighted the benefit of HC50 prediction based on learning chemical embeddings rather than on raw chemical properties.

3.5. Autoencoder compared with other methods from the literature

Hou et al. (2020b) used genetic algorithm optimized neural network models to predict HC50 and achieved an $R^2$ of 0.63 (Table 1). However, Hou et al. (2020b) randomly sampled test data instead of using cross-validation for the whole dataset to test every data point. Notably, fair comparison of different approaches should use the same dataset and validation methods. Our dataset matched that of Hou et al. (2020b), so we compared our results to theirs by
matching validation procedures. Using the same validation procedures as Hou et al. (2020b), our autoencoder model yielded an $R^2$ of 0.67 (almost identical to that of the five-fold cross validation procedure) (Table 1). Thus, our approach was stable for different validation methods. Another study by Hou et al. (2020a) used a different input dataset with only 14 input features, precluding direct comparison of their results with the results of this study. We also directly predicted the HC$_{50}$ values using ECOSAR software ($R^2 = 0.23$), which had a much poorer performance than other tested models (Table 1).

### 3.6. Management implications

Effective use of machine learning approaches in management of chemical risks requires transparent and accessible communication of model development and results interpretation to stakeholders. Since machine learning methods are complex and not readily accessible to all stakeholders, such communication can be very challenging, but could be aided by flow diagrams, comparison with benchmark models (e.g., QSAR and other models), and selective use as screening tools. Additionally, with the ability to learn from complex high-dimensional data, the autoencoder approach may be useful to predict chronic and more subtle endpoints that drive management and market decisions. For example, the lowest observed effect concentration may be predicted for chemicals used in products to assist the Safer Choice Program in EPA Ecolabeling.

### 4. Conclusion

Overcoming challenges and limitations in in silico prediction of chemical ecotoxicity could provide valuable new approaches to identifying chemical hazards. This study developed an autoencoder model for predicting the HC$_{50}$ values. This autoencoder model can perform nonlinear dimensionality reduction and learn informative latent space chemical representations from input data, providing a unique way of processing redundant, high-dimension input data that is different from other methods such as PCA, MDS and UMAP. These learned chemical representations can achieve significant dimension reduction, supporting their use for more accurately predicting chemical ecotoxicity compared to methods directly using raw input features.

### Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

### Acknowledgements

This study was supported by Grants R35ES031688 and P30ES009089 from the National Institute of Environmental Health Sciences. We would like to thank Bu Zhao, Dr. Ming Xu, and Dr. Ping Hou for discussion on their collected dataset and Dr. Sheila M. Cherry from Fresh Eyes Editing for polishing the writing of the manuscript.

### Abbreviations:

- **HC$_{50}$**: hazardous concentration 50%
- **EC$_{50}$**: effective concentration 50%
| Abbreviation | Description |
|--------------|-------------|
| LC<sub>50</sub> | lethal concentration 50% |
| MoA | mode of action |
| QSAR | quantitative structure-activity |
| PCA | principal component analysis |
| MDS | multidimensional scaling |
| UMAP | uniform manifold approximation and projection |
| t-SNE | t-distributed stochastic neighbor embedding |
| LASSO | least absolute shrinkage and selection operator |
| RF | random forest |
| FCNN | fully connected neural network |
| ECOSAR | ecological structure-activity relationship |
| MAE | mean absolute error |
| RMSE | root mean squared error |

**References**

Agarap AF, 2018. Deep learning using rectified linear units (relu). arXiv:1803.08375 2018.

Barber D, 2012. Bayesian reasoning and machine learning. Cambridge University Press.

Barron M, Lilavoi C, Martin T, 2015. Moatox: A comprehensive mode of action and acute aquatic toxicity database for predictive model development. Aquat. Toxicol 161, 102–107. [PubMed: 25700118]

Bellman RE, 2015. Adaptive control processes: A guided tour. Princeton University Press.

Bengio Y, Courville A, Vincent P, 2013. Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell 35, 1798–1828. [PubMed: 23787338]

Bengio Y, Lamblin P, Popovici D, Larochelle H, 2007. Greedy layer-wise training of deep networks. Adv. Neural Inf. Process. Syst

Boone KS, Di Toro DM, 2019a. Target site model: Application of the polyparameter target lipid model to predict aquatic organism acute toxicity for various modes of action. Environ. Toxicol. Chem 38 (1), 222–239. [PubMed: 30255636]

Boone KS, Di Toro DM, 2019b. Target site model: Predicting mode of action and aquatic organism acute toxicity using abraham parameters and feature-weighted k-nearest neighbors classification. Environ. Toxicol. Chem 38 (2), 375–386. [PubMed: 30506854]

Chang X, Kleinstreuer N, Ceger P, Hsieh J-H, Allen D, Casey W, 2015. Application of reverse dosimetry to compare in vitro and in vivo estrogen receptor activity. Appl. In Vitro Toxicol 1 (1), 33–44.

Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, 2014. Qsar modeling: Where have you been? Where are you going to? J. Med. Chem 57, 4977–5010. [PubMed: 24351051]

Fantke P, Bijster M, Guignard C, Hausschild M, Huijbregts M, Jolliet O, Kounina A, Magaud V, Margni M, McKone T, Posthuma L, Rosenbaum RK, van de Meent D, Zelm R.v., 2017. Usetox® 2.0 documentation (version 1.1).

Glorot X, Bordes A, Bengio Y, 2011. Deep sparse rectifier neural networks. Proceedings of the 14th International Conference on Artificial Intelligence and Statistics.
Gramatica P, 2007. Principles of qsar models validation: Internal and external. QSAR Comb. Sci 26, 694–701.

Hawkins DM, Basak SC, Mills D, 2003. Assessing model fit by cross-validation. J. Chem. Inf. Comput. Sci 43, 579–586. [PubMed: 12653524]

Hinton GE, Salakhutdinov RR, 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507. [PubMed: 16873662]

Hou P, Jolliet O, Zhu J, Xu M, 2020. Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. Environ. Int, 135, 105393. [PubMed: 31862642]

Hou P, Zhao B.u., Jolliet O, Zhu J.i., Wang P, Xu M, 2020b. Rapid prediction of chemical ecotoxicity through genetic algorithm optimized neural network models. ACS Sustain. Chem. Eng 8 (32), 12168–12176.

Kim J-O, Mueller CW, 1978. Factor analysis: Statistical methods and practical issues. Sage.

King RD, Orhobor OI, Taylor CC, 2021. Cross-validation is safe to use. Nat. Mach. Intell 3, 276–276.

Laue H, Hostettler L.u., Badertscher RP, Jenner KJ, Sanders G, Arnolt JA, Natsch A, 2020. Examining uncertainty in in vitro–in vivo extrapolation applied in fish bioconcentration models. Environ. Sci. Technol 54 (15), 9483–9494. [PubMed: 3263948]

LeCun Y, Bengio Y, Hinton G, 2015. Deep learning. Nature 521 (7553), 436–444. [PubMed: 26017442]

Mansouri K, Karmaus AL, Fitzpatrick J, Patleweicz G, Pradeep P, Alberga D, Alepee N, Allen TEH, Allen D, Alves VM, Andrade CH, Auernhammer TR, Ballabio D, Bell S, Benfenati E, Bhattacharya S, Bastos JV, Boyd S, Brown JB, Capuzzi SJ, Chushak Y, Ciallella H, Clark AM, Consonni V, Daga PR, Ekins S, Farag S, Fedorov M, Fourches D, Gadaleta D, Gao F, Gearhart JM, Goh G, Goodman JD, Grisoni F, Grulke CM, Hartung T, Hirn M, Karpov P, Korotcov A, Lavado GJ, Lawless M, Li X, Luechetfeld T, Lunghini F, Mangiordi GF, Marcou G, Marsh D, Martin T, Mauri A, Muratov EN, Myatt GJ, Nguyen D-T, Nicolotti O, Note R, Pande P, Parks AK, Peryea T, Polash AH, Rallo R, Roncaglioni A, Rowlands C, Ruiz P, Russo DP, Sayed A, Sayre R, Sheils T, Siegel C, Silva AC, Simeonov A, Sosnin S, Southall N, Strickland J, Tang Y, Teppen B, Tetko IV, Thomas D, Tkachenko V, Todeschini R, Toma C, Tripodi I, Trisciuoggi D, Tropsha A, Varnek A, Vukovic K, Wang Z, Wang L, Waters KM, Wedlake AJ, Wijeyesakere SJ, Wilson D, Xiao Z, Yang H, Zahoranszky-Kohalmi G, Zakharov AV, Zhang FF, Zhang Z, Zhao T, Zhu H, Zorn KM, Case W, Kleinstreuer NC, 2021. Catmos: Collaborative acute toxicity modeling suite. Environ. Health Perspect. 129 (4), 047013.

McInnes L, Healy J, Melville J, 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426.

Mead A, 1992. Review of the development of multidimensional scaling methods. J. Royal Statistical Soc.: Ser. D (The Statistician) 41, 27–39.

Mika S, Schölkopf B, Smola A, Müller K-R, Scholz M, Rätsch G, 1998. Kernel pca and de-noising in feature spaces. Adv. Neural Inf. Process. Syst 11.

National Research Council, 2007. Toxicity testing in the 21st century: A vision and a strategy: National Academies Press.

Raza A, Bardhan S, Xu L, Yamijala SS, Lian C, Kwon H, Wong BM, 2019. A machine learning approach for predicting defluorination of per-and polyfluoroalkyl substances (pfas) for their efficient treatment and removal. Environ. Sci. Technol. Lett 6, 624–629.

Rosenbaum RK, Bachmann TM, Gold LS, Huijbregts MAJ, Jolliet O, Juraske R, Koehler A, Larsen HF, MacLeod M, Margni M, McKone TE, Payet J, Schuhmacher M, van de Meent D, Hauschild MZ, 2008. Usetox—the unep-setac toxicity model: Recommended characterisation factors for human toxicity and freshwater ecotoxicity in life cycle impact assessment. Int. J. Life Cycle Assess 13 (7), 532–546.

Schölkopf B, Smola A, Müller K-R, 1997. Kernel principal component analysis. Artif. Neural. Netw. ICANN.

Su A, Rajan K, 2021. A database framework for rapid screening of structure-function relationships in pfas chemistry. Scientific Data 8, 1–10. [PubMed: 33414438]

Environ Int. Author manuscript; available in PMC 2022 May 01.
Takata M, Lin B-L, Xue M, Zushi Y, Terada A, Hosomi M, 2020. Predicting the acute ecotoxicity of chemical substances by machine learning using graph theory. Chemosphere 238, 124604. [PubMed: 31450113]

Tropsha A, Gramatica P, Gombar VK, 2003. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSAR models. QSAR Comb. Sci 22, 69–77.

US EPA. Ecological structure activity relationships (ecosar) predictive model. United States Environmental Protection Agency; 2022.

Van der Maaten L, Hinton G, 2008. Visualizing data using t-sne. J. Mach. Learn. Res 9.

Vincent P, Larochelle H, Bengio Y, Manzagol P-A, 2008. Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th International Conference on Machine Learning.

Wold S, Esbensen K, Geladi P, 1987. Principal component analysis. Chemometrics Intellig. Lab. Syst 2, 37–52.

Xia M, Huang R, Witt KL, Southall N, Fostel J, Cho M-H, Jadhav A, Smith CS, Inglese J, Portier CJ, Tice RR, Austin CP, 2008. Compound cytotoxicity profiling using quantitative high-throughput screening. Environ. Health Perspect 116 (3), 284–291. [PubMed: 18335092]

Yao Z, Sánchez-Lengeling B, Bobbitt NS, Bucior BJ, Kumar SGH, Collins SP, Burns T, Woo TK, Farha OK, Snurr RQ, 2021. Inverse design of nanoporous crystalline reticular materials with deep generative models. Nat. Mach. Intell 3, 76–86.

Yarotsky D, 2017. Error bounds for approximations with deep relu networks. Neural Networks 94, 103–114. [PubMed: 28756334]
Fig. 1. Schematic of the autoencoder model. Dark blue nodes are input features, green nodes are generated embeddings, and light blue nodes are output layers. HC50 is predicted using one linear layer.
Fig. 2. Embedding size comparison of the autoencoder models. Embedding layer size of a) 2; b) 16; c) 32; d) 64; e) 128; and f) 256 neurons.
Fig. 3. Visualization of chemical embeddings in t-SNE coordinates.
a) t-SNE colored on HC50 (brighter color represents greater HC50 values); b) t-SNE colored on Mode of Action (Class 1 = inert chemicals, Class 2 = less inert chemicals, Class 3 = reactive chemicals, Class 4 = chemicals acting by a specific mechanism, and Class 5 = unclassifiable chemicals).
Fig. 4. Comparison of different dimensionality reduction methods and machine learning models.
a) Autoencoder; b) PCA based on top 128 features; c) kernel PCA based on top 128 features; d) UMAP based on top 128 features; e) random forest (RF) model; f) fully connected neural network (FCNN) model; g) LASSO regression; h) ridge regression. Red lines are 1:1 plots (y = x) indicating predicted HC$_{50}$ values equal to measured HC$_{50}$ values. Green = neutral chemicals; orange = base chemicals; brown = acid chemicals; blue = amphoter chemicals.
Fig. 5. Autoencoder model results colored by Mode of Action
(Class 1 = inert chemicals; Class 2 = less inert chemicals; Class 3 = reactive chemicals; Class 4 = chemicals acting by a specific mechanism; and Class 5 = unclassifiable chemicals).
Table 1

The coefficient of determination ($R^2$) for the HC$_{50}$ prediction using other methods from the literature.

| Method                                           | Chemical numbers | Feature numbers | $R^2$ |
|--------------------------------------------------|------------------|-----------------|-------|
| ECOSAR                                           | 1728             | NA              | 0.23  |
| Random forest model (Hou et al., 2020a)          | 2307             | 14              | 0.63  |
| Genetic algorithm optimized neural network model (Hou et al., 2020b) | 1815             | 691             | 0.63  |
| Autoencoder model with random sampling           | 1815             | 691             | 0.67  |

* Results that can be compared. These results used the same dataset and validation method (random sampling).