Bioinformatics, 2023, 1–4
doi: DOI HERE
Advance Access Publication Date: Day Month Year
Application note

`deepFPlearn`$^{+}$: Enhancing Toxicity Prediction Across the Chemical Universe Using Graph Neural Networks

Kyriakos Soulios ,[1,3,*] Patrick Scheibe ,[2] Matthias Bernt ,[1] Jörg Hackermüller [1,3] and Jana Schor [1]

[1] Department of Computation Biology, Helmholtz Centre for Environmental Research - UFZ, Permosterstraße 15,04318 Leipzig, Germany, [2] Department of Neurophysics, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1a, 04103 Leipzig, Saxony, Germany and [3] Faculty of Mathematics and Computer Science, University of Leipzig, Augustusplatz 10, 04109 Leipzig,Germany

*Corresponding author. jana.schor@ufz.de

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Summary: Sophisticated approaches for the *in-silico* prediction of toxicity are required to support the risk assessment of chemicals. The number of chemicals on the global chemical market and the speed of chemical innovation stand in massive contrast to the capacity for regularizing chemical use. We recently proved our ready-to-use application deepFPlearn as a suitable approach for this task. Here, we present its extension `deepFPlearn`$^{+}$ incorporating i) a graph neural network to feed our AI with a more sophisticated molecular structure representation and ii) alternative train-test splitting strategies that involve scaffold structures and the molecular weights of chemicals. We show that the GNNs outperform the previous model substantially and that our models can generalize on unseen data even with a more robust and challenging test set. Therefore, we highly recommend the application of `deepFPlearn`$^{+}$ on the chemical inventory to prioritize chemicals for experimental testing or any chemical subset of interest in monitoring studies.
Availability and Implementation: The software is compatible with python 3.6 or higher, and the source code can be found on our GitHub repository: `https://github.com/yigbt/deepFPlearn`. A complete data and models archive is also available on Zenodo: `https://zenodo.org/record/8146252`. Detailed installation guides via Docker, Singularity, and Conda are provided within the repository for operability across all operating systems.
Contact: jana.schor@ufz.de
Supplementary information: The supplementary material is provided in the submission.

Key words: deep learning, graph neural network, scaffold split, chemical structure, chemical-effect associations

Key Messages

- Graph neural networks are superior in providing molecular structure information to downstream tasks.
- Scaffold splitting ensures models generalize better to unseen molecular structures providing more realistic performance estimates.
- `deepFPlearn`$^{+}$ enhances the predictive power and feasibility of deep learning in toxicology for smart chemical prioritization and sustainable design.

Introduction

Recently, we developed the ready-to-use and stand-alone program deepFPlearn that predicts the association between chemical structures and effects on the gene/pathway level using a combined deep learning approach [Schor et al., 2022]. We achieved high accuracy using a deep autoencoder (AE) to reduce features and a feed-forward neural network to predict whether the input chemical interacts with the nuclear receptors involved in endocrine disruption. Our pre-training strategy allowed for capturing a vast

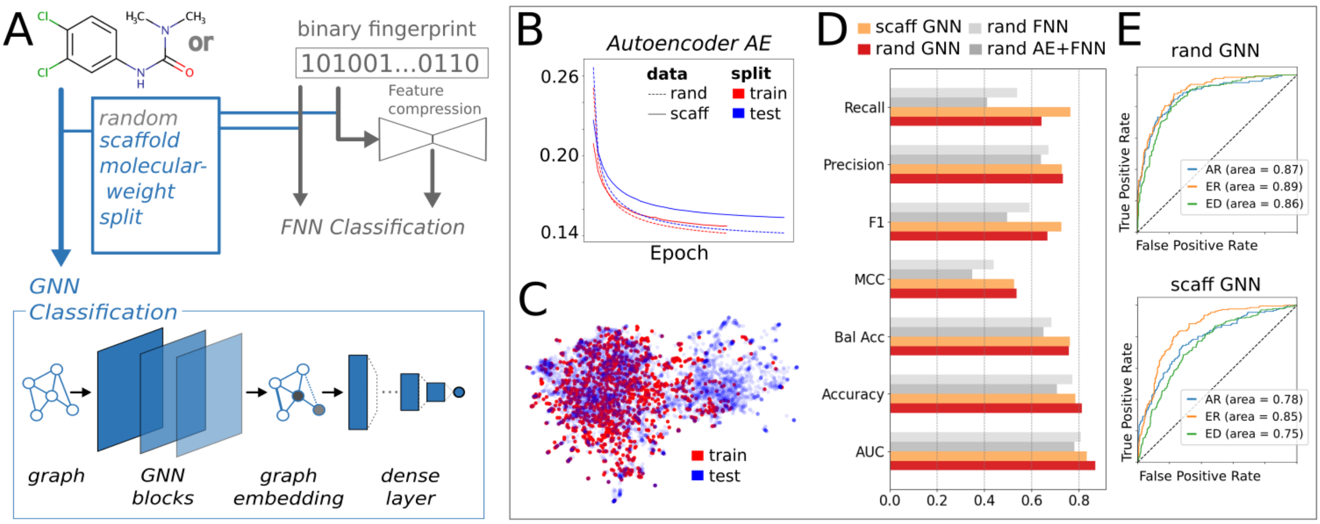

Fig. 1: A) DL framework with deepFPlearn$^+$ extensions indicated in blue and existing parts in gray. B) History of AE mean squared error for random and scaffold split of the train data. C) Uniform Manifold Approximation and Projection (UMAP) embedding of the scaffold-split with a noticeable distribution-shift between train and test data. D) Comparison of all collected metrics across all model architectures for the ED (endocrine disruptor) target using the best fold for each model. E) ROC-AUC of random and scaffold split of GNN for three targets, AR (androgen receptor), ER (estrogen receptor-alpha) and ED (endocrine disruptor).

range of molecular structures. Additionally, deepFPlearn classifies chemicals quickly and can be customized. Our tool significantly aids in the systematic in silico study of chemicals, and can handle the vast, constantly growing chemical universe.

deepFPlearn uses binary topological fingerprints to encode the molecular structure of chemicals, which show the absence or presence of certain substructures or atomic configurations by ones and zeros respectively, in a fixed-size binary vector. deepFPlearn$^+$ now can also use molecular graphs to encode the structure of chemicals preserving the full connectivity information which is not fully captured in binary fingerprints, at the expense of lost chirality information. Graph neural networks can process such graph-based information.

A graph neural network (GNN) is an artificial neural network that operates on graph data Scarselli et al. [2009]. A graph $G = (V, E)$ comprises a set of nodes V and a list of edges E that encodes the nodes' relations. In molecular structure graphs, nodes represent atoms, and covalent chemical bonds form the edges. More attributes, like the type of atom or direction/type of the bond, can be stored in V and E. A GNN then forms an optimizable transformation on all graph attributes that preserve graph symmetries. One approach for training a GNN is using message passing (MP), which involves disseminating information stored in nodes and edges as a cumulative message throughout the network to arrive at a prediction. In a D-MPNN, the message is associated with directed edges to avoid unnecessary loops in the message-passing trajectory. See [Yang et al., 2019] for the detailed algorithm.

Substructures form functional entities - scaffolds - and different scaffold combinations result in different properties and modes of interactions with other molecules. A *chemical scaffold* is a structure that is shared among a group of molecules, which are then likely to have a similar relationship with targets Katritzky et al. [2000]. The cheminformatic software package RDKit Landrum [2021] implements algorithms from [Bemis and Murcko, 1996] that generate chemical scaffolds from their 2D representation. Typically, machine learning methods require datasets to be split into train and test sets. While trained on the first, the model is evaluated on the latter to investigate how well the trained model can generalize on unseen data. A random split creates a similar distribution of classes in the train and test set. This is only sometimes best for chemical structures since both sets contain similar structures, leading to an overestimation of the model's prediction performance. Scaffold splitting separates structurally different molecules into different subsets, providing a more significant challenge for learning algorithms than the random split. Splitting based on molecular weight divides datasets into training and test sets by considering the size of the molecules. This approach aims to ensure that molecules of different sizes are represented in both sets, reflecting the diversity of molecular scales.

The idea to reach out beyond random split is that the increased differences between training and test sets form a more robust test of the model's generalizability Wu et al..

Here, we present deepFPlearn$^+$ - an extension of deepFPlearn incorporating graph neural networks to predict the association of an entire graph (one chemical) with selected nuclear receptors involved in endocrine disruption and alternative train-test data splitting strategies.

deepFPlearn$^+$

We extended our original deepFPlearn deep learning framework by a D-MPNN graph neural network and two alternative train test data splitting strategies, based on structural scaffolds and the molecular weight of the chemicals, see Fig. 1A. We trained deepFPlearn on the same data as in [Schor et al., 2022] and compared the model performances for different setups. In particular, we performed a 5-fold cross-validation for training, employed Weights & Biases (www.wandb.com) for hyperparameter tuning, and used the small but labeled dataset S (extracted from the supplemental material of Sun et al. [2019]) containing 7248 compounds for classification and the huge unlabeled dataset D (extracted from the CompTox Chemistry Dashboard [Williams et al., 2017, accessed on 2020/07/13]) containing 719 996 compounds for pre-training and feature compression. When we enabled scaffold split for pre-training, the average validation loss of the autoencoders increased slightly from 0.1386 (random split) to 0.1547 (Fig. 1B), which we expected since the molecular structures differ substantially in the train and the test data (Fig. 1C). With enabled molecular weight split for pre-training, the average validation loss of the autoencoders increased substantially to 0.2266 (Supplementary Fig. 1). See supplementary Fig. 1 for the training performances and UMAP visualizations of the training data for all three cases. Further, the validation loss diverges considerably from the training loss. We expected this even more since this approach ensures that the differences between molecules of the train and test datasets are much higher. Consequently, the potential of over- and underfitting specific molecular weight ranges grows dramatically, and the model's generalizability declines. The GNN (with random split) substantially outperforms the fingerprint classification models [Schor et al., 2022]. We reached binary accuracy values of 0.87, 0.84 and 0.81 and ROC-AUC values of 0.87, 0.89 and 0.86 for androgen receptor (AR), estrogen receptor (ER), and endocrine disruption (ED), respectively. Enabling scaffold split slightly reduced the model's performance. And still, we reached higher values with the GNN than we did when using the molecular fingerprint as input and the traditional feed-forward neural network. See figure 1D and E. See the supplementary material for detailed performance comparisons showing all combinations of split strategies, model architectures, and classification targets.

We used the trained GNN model for ER to predict the association between all substances from the D dataset, which have no association with ER in our training data. Exemplary, we found Hexylresorcinol, Quinoxyfen, and Clofoctol among the top positive predictions, with a probability higher than 80%. The CompTox dashboard provides bioassay information for all three substances, indicating an association with ER. Hexylresorcinol is a substituted dihydroxybenzene with antiseptic, anthelmintic, and local anesthetic properties, contained in topical applications for minor skin infections and oral solutions or throat lozenges for pain relief and first-aid antiseptic Chaudhuri and Chaudhuri, Matthews et al.. Quinoxyfen is a fungicide used mainly to control Erysiphe graminis - powdery mildew in cereals. The EFSA regulated the maximum amount of this substance based on bioactivity and toxicity studies in December 2020 EFS. Clofoctol is a bacteriostatic antibiotic for treating upper and lower respiratory tract infections. The CompTox Chemistry Dashboard reports bioactivity with ER in the ToxCast summary reports.

Conclusion

In the deepFPlearn architecture, FFNNs had established a sophisticated baseline for performance, whereas autoencoders facilitated a pre-trained model attuned to a vast array of chemical structures from unlabeled datasets. Although these methods performed good enough, we substantially improved the performance of deepFPlearn by tuning the input data representation and, subsequently, the model's architecture towards a graph neural network. To address the generalizibility assessment of GNNs that lack pretraining, we introduced more challenging train-test splitting strategies of scaffold and molecular weight split and compared the model performances to the original random split approach. Scaffold split was most feasible since it provides a more realistic test set regarding the application domain of deepFPlearn$^+$. It tests the model's generalizability by separating structurally different molecules into separate subsets.

With deepFPlearn$^+$ we increased the predictive power and the feasibility of the overall beneficial deep-learning approach in predictive toxicology. We could confirm our predictions with current toxicity knowledge from the latest ToxCast release, suggesting that there are more exciting chemicals among them that should be considered in subsequent wet lab and *in silico* analyses. In ongoing data analyses, we dive deeper into the biological backgrounds of our predictions and integrate explainable AI into our framework to better understand and explain our models' decisions. We promote deepFPlearn$^+$ś application on the chemical inventory and custom subsets of substances to prioritize chemicals for experimental testing, assist in the smart selection of chemicals for monitoring and contribute to the sustainable design of future chemicals.

Competing interests

No competing interest is declared.

Author contributions statement

K.S., P.S., and J.S. implemented all extensions in deepFPlearn$^+$. M.B. merged new with existing code and updated the GitHub CI. All authors wrote and reviewed the manuscript.

Acknowledgments

This work was supported by the framework of the European Partnership for the Assessment of Risks from Chemicals (PARC) and has received funding from the European Union's Horizon Europe research and innovation program under grant agreement 101057014. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Secondly, this work was also supported in part by the Helmholtz association. The AI project 'XAI-graph' funded by Helmholtz AI which aims to investigate the explainability and the quantification of uncertainty of deep learning methods.

References

Review of the existing maximum residue levels for quinoxyfen according to article 12 of regulation (ec) no 396/2005. 18. ISSN 18314732, 18314732.

G. W. Bemis and M. A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39:2887–2893, 1 1996. ISSN 0022-2623. doi: 10.1021/jm9602928.

R. Chaudhuri and R. K. Chaudhuri. Hexylresorcinol: Providing skin benefits by modulating multiple molecular targets photostability view project skin hydration and barrier building view project hexylresorcinol: Providing skin benefits by modulating multiple molecular targets.

A. R. Katritzky, J. S. Kiely, N. Hébert, and C. Chassaing. Definition of templates within combinatorial libraries. *Journal of Combinatorial Chemistry*, 2:2–5, 1 2000. ISSN 1520-4766. doi: 10.1021/cc990047k.

G. Landrum. Rdkit: Open-source cheminformatics software, 2021. ISSN 00028282.

D. Matthews, O. Adegoke, and A. Shephard. Bactericidal activity of hexylresorcinol lozenges against oropharyngeal organisms associated with acute sore throat. *BMC Research Notes*, 13:1–4, 2 . ISSN 17560500. doi: 10.1186/S13104-020-04954-1/FIGURES/1.

F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20, 2009. ISSN 10459227. doi: 10.1109/TNN.2008.2005605.

J. Schor, P. Scheibe, M. Bernt, W. Busch, C. Lai, and J. Hackermüller. Ai for predicting chemical-effect associations at the chemical universe level - deepfplearn. *Briefings in Bioinformatics*, 23, 2022. ISSN 14774054. doi: 10.1093/bib/bbac257.

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

A. J. Williams, C. M. Grulke, J. Edwards, A. D. Mceachran, K. Mansouri, N. C. Baker, G. Patlewicz, I. Shah, J. F. Wambaugh, R. S. Judson, A. M. Richard, and W. A. Gov. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry Open Access. *Journal of Cheminformatics*, 9:61, 2017. doi: 10.1186/s13321-017-0247-6.

Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande. Moleculenet: A benchmark for molecular machine learning. *Chemical Science*, 9:513–530, 3 .

K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, and R. Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59, 2019. ISSN 1549960X. doi: 10.1021/acs.jcim.9b00237.