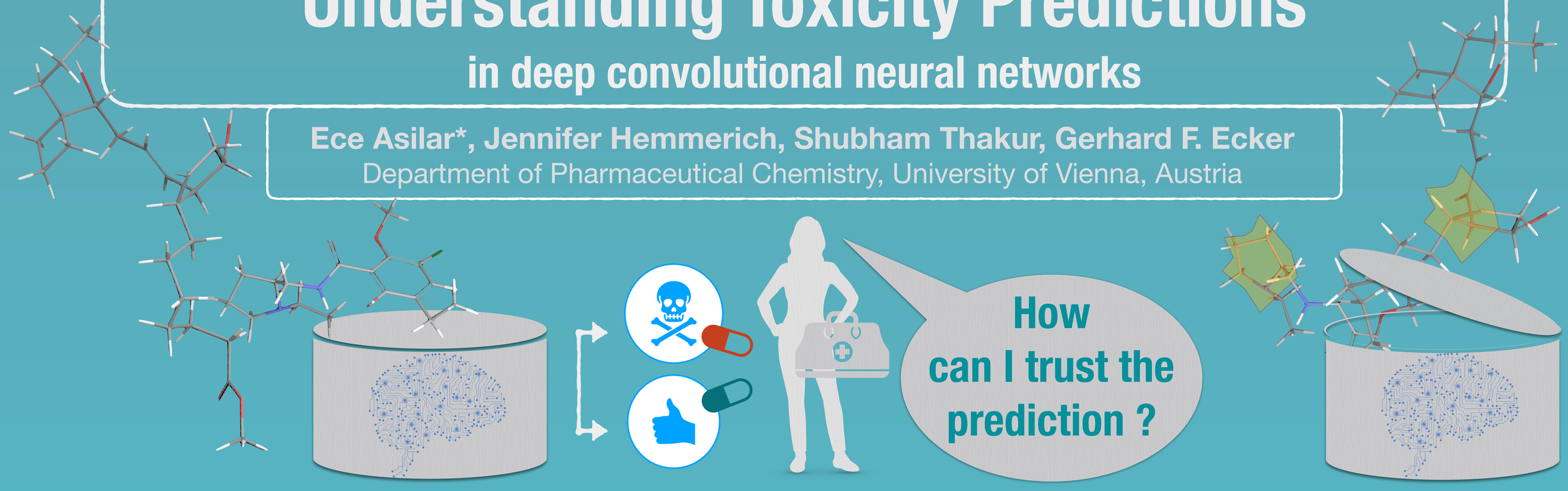
Understanding Toxicity Predictions

in deep convolutional neural networks

Ece Asilar*, Jennifer Hemmerich, Shubham Thakur, Gerhard F. Ecker
Department of Pharmaceutical Chemistry, University of Vienna, Austria



Being one of the emergent technologies in the computer science field, various implementations of **deep learning** have been launched recently in pharmaceutical chemistry research [1]. So far, deep learning architectures exhibit good performance in toxicity prediction tasks [2].

However, these very sophisticated algorithms are not interpretable.

Understanding the reasons behind the predictions is very crucial, especially, when the decision taken by these models has a significant effect on human health.

We employ Local interpretable Model-agnostic Explanations (**LIME**) [3] package to interpret models which are trained to classify compounds according to their toxicity.

Local: explains why a single data point classified as a specific class.

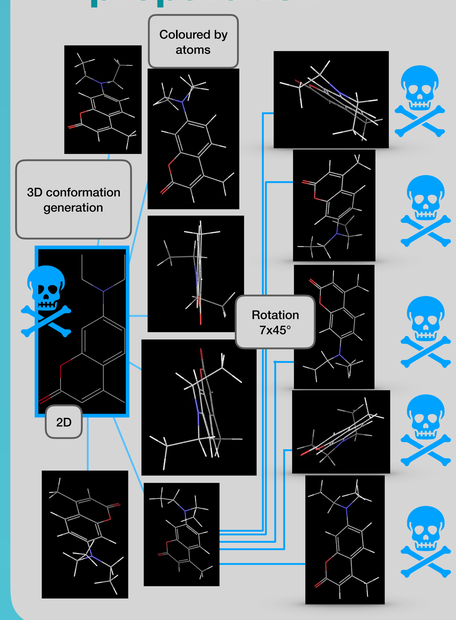
Model-agnostic: The model is treated as a blackbox; does not need to know how it makes the prediction.

Model Training

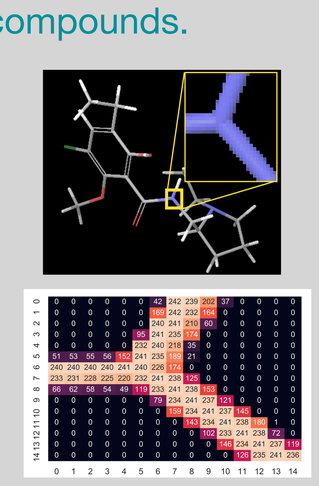
We utilised a supervised deep learning algorithm for classifying the compounds according to their toxicity.

These sophisticated algorithms are prone to overfitting in the case of small and imbalanced dataset. To avoid bias towards to the majority class, we balanced the dataset by generating 3D conformers. The images of the 3D conformers are then used as an input.

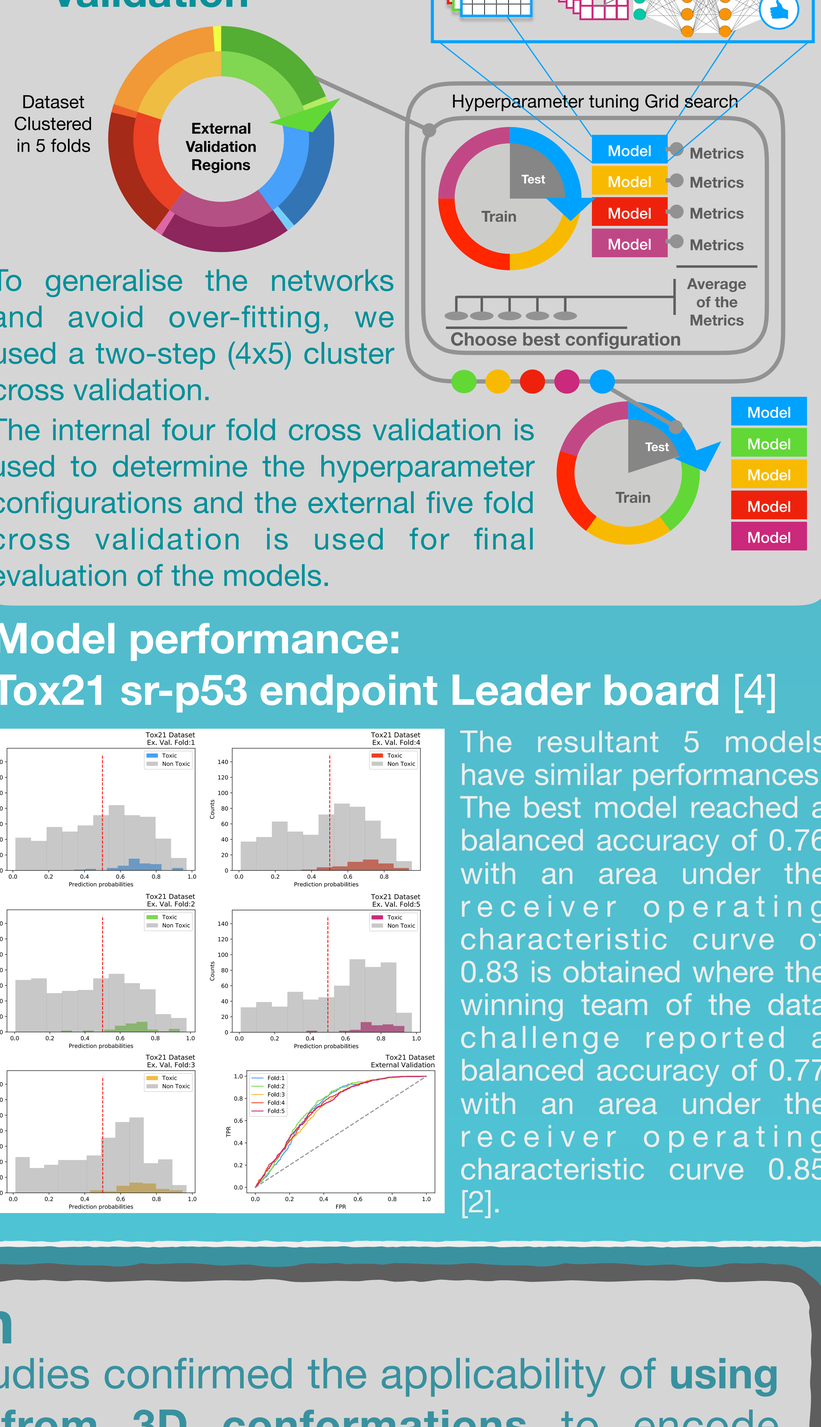
Image data preparation



Using images allows **connecting** various chemical features to the **geometry** of the compounds.



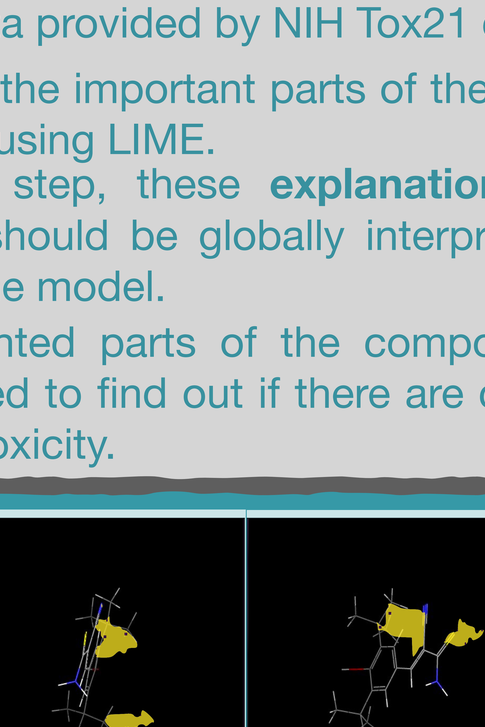
4x5 Cross validation



To generalise the networks and avoid over-fitting, we used a two-step (4x5) cluster cross validation.

The internal four fold cross validation is used to determine the hyperparameter configurations and the external five fold cross validation is used for final evaluation of the models.

Model performance: Tox21 sr-p53 endpoint Leader board [4]



The resultant 5 models have similar performances. The best model reached a balanced accuracy of 0.76 with an area under the receiver operating characteristic curve of 0.83 is obtained where the winning team of the data challenge reported a balanced accuracy of 0.77 with an area under the receiver operating characteristic curve 0.85 [2].

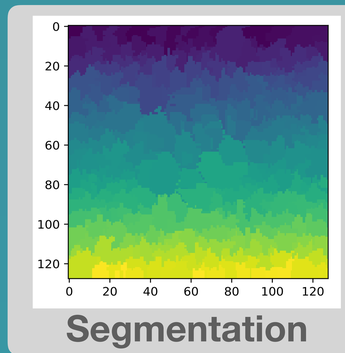
Prediction interpretation using LIME

For a given prediction, LIME fits a linear model on a set of artificially generated local samples in the feature space.

Then it picks the most important features yield a model which is faithful to the original prediction.

An interpretable representation for image classification is a presence/absence of a patch of similar pixels, i.e super-pixels.

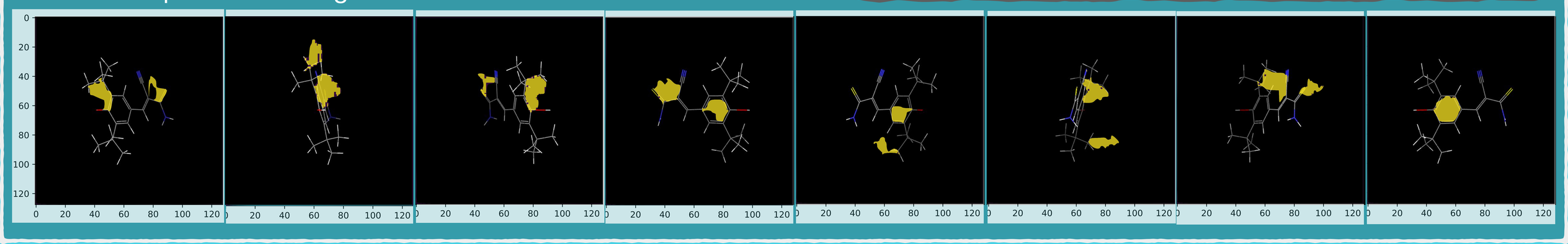
The candidate super-pixels are designed with a segmentation function.



Segmentation

Example: A toxic compound which is predicted as toxic with a probability of 0.84 using our models and predicted as toxic with LIME's local linear model as toxic with a probability of 0.87.

The most important two segments are shown.



Conclusion

- Our recent studies confirmed the applicability of **using the images from 3D conformations** to encode molecules using **deep convolutional neural networks** [3] on the data provided by NIH Tox21 challenge.
- We derived the important parts of the images in giving the decision using LIME.
- As a next step, these **explanations** on individual predictions should be globally interpreted to measure the trust in the model.
- The highlighted parts of the compounds should be further studied to find out if there are common features leading the toxicity.

universität wien

eTRANSAFE

[1] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018) The rise of deep learning in drug discovery. Drug Discov. Today 23, 1241 – 1250, https://doi.org/10.1016/j.drudis. 2018.01.039
[2] Mayr, A. et al. (2016). DeepTox: Toxicity Prediction using Deep Learning. Front. Environ. Sci. doi: 10.3389/fenvs.2015.0008
[3] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016).
[4] https://tripod.nih.gov/tox21/challenge/