Applying Deep Learning Toxicity Models Across the Chemical Universe

Sunggun Lee

Duke University, Durham, NC

January 19th, 2023

ASCCT Webinar

National Institute of Environmental Health Sciences
Division of the National Toxicology Program

Overview: The Need for Alternative Approaches in Carcinogenicity Assessment

- The ICH Guideline on the Need for Carcinogenicity studies for pharmaceuticals (1995) introduced and outlined the need and study design of carcinogenicity studies
  - Experimental approaches generally requires ~500 rodents and costs around $1.1 million on average
  - Results in flawed extrapolation for carcinogenicity
  - Neglect the 3R's of replacement, reduction, and refinement of animals in toxicology testing


European Medicines Agency

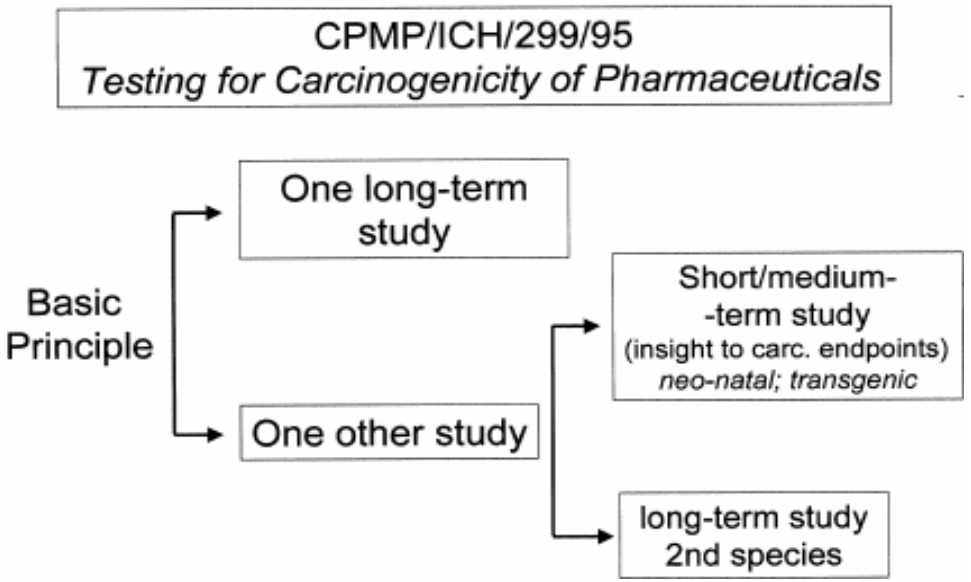
July 1996
CPMP/ICH/140/95

ICH Topic S 1 A
The Need for Carcinogenicity Studies of Pharmaceuticals

Step 5

NOTE FOR GUIDANCE ON THE NEED FOR CARCINOGENICITY STUDIES OF PHARMACEUTICALS
(CPMP/ICH/140/95)

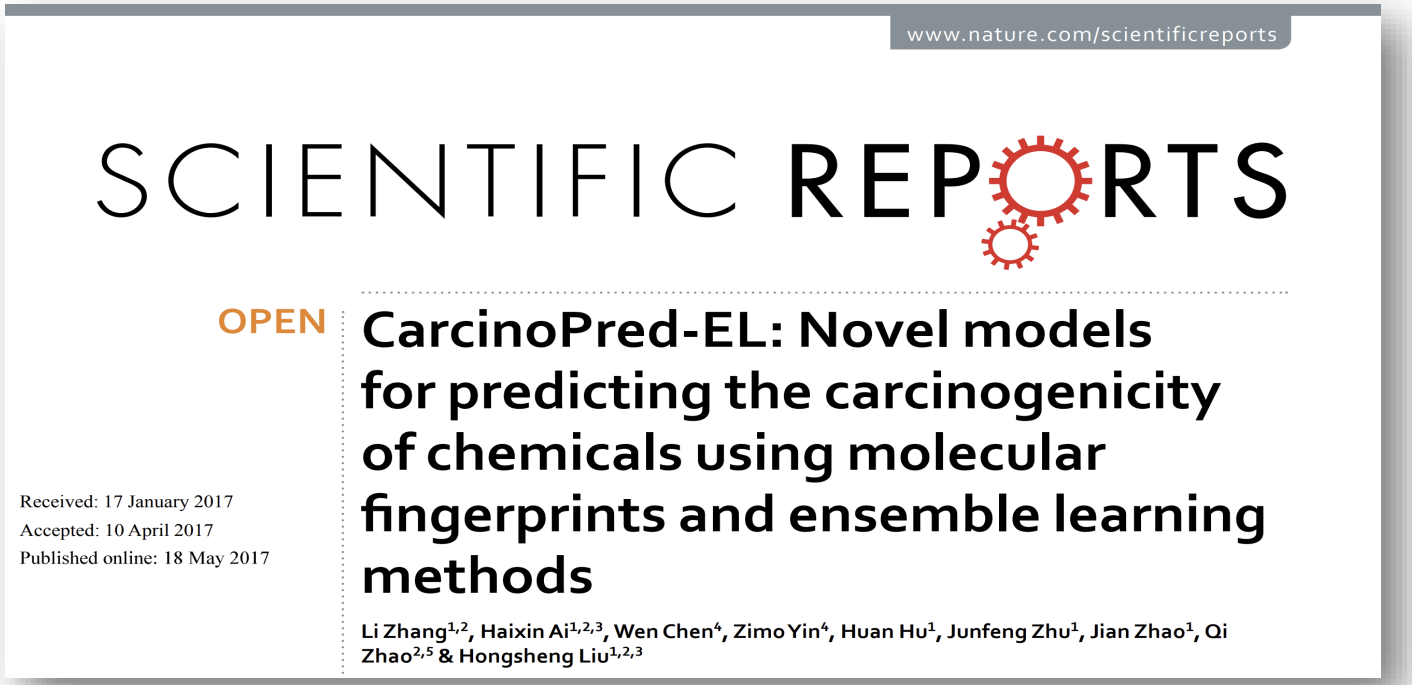
ICH Guideline. 1996.



Silva Lima B, Van der Laan, JW. 2002.

Development of Alternative Approaches for Carcinogenicity Assessment

- New approaches explore in silico methodologies such as QSAR models, adverse outcome pathways, and more
  - ➢ Supported by programs such as Horizon 2020, Tox21, The Seventh Framework Programme 7 (FP7), and other partnerships
- QSAR approaches have been successfully applied for carcinogenicity assessment of certain chemical classes (aromatic amines, food-relevant phytochemicals, etc.)



www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 17 January 2017
Accepted: 10 April 2017
Published online: 18 May 2017

CarcinoPred-EL: Novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods

Li Zhang[1,2], Haixin Ai[1,2,3], Wen Chen[4], Zimo Yin[4], Huan Hu[1], Junfeng Zhu[1], Jian Zhao[1], Qi Zhao[2,5] & Hongsheng Liu[1,2,3]

Zhang L, Ai H, Chen W, Yin Z, Hu H, Zhu J, Zhao J, Zhao Q, Liu H. 2002.

Deep Learning for Carcinogenicity Assessment

- Deep learning models offer a beneficial alternative for toxicity predictions where sufficient training data is available because of their ability to predict complex endpoints
- DeepCarc and DeepDILI predict endpoints for carcinogenicity and drug-induced liver injury, respectively, outperforming conventional QSAR models and state-of-the-art ensemble methods

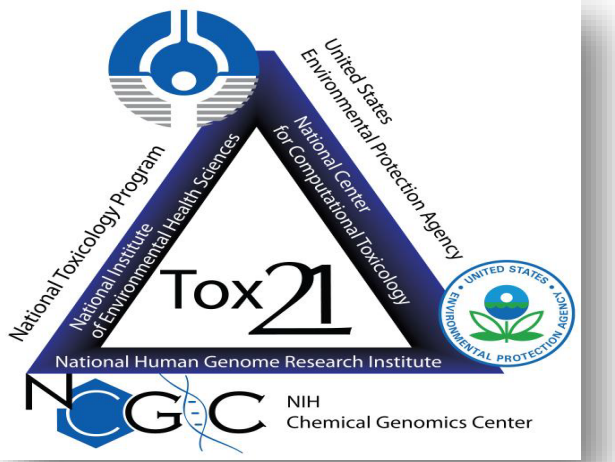


Figure 1: The Tox21 collaboration of NTP, NCGC, and EPA (Shukla et al. 2010)

- The Tox21 dataset is comprised of ~10,000 compounds including food-additives, household cleaning products, medicines, and environmental chemical hazards
- DeepCarc and DeepDILI were applied to predict toxicity for 7176 compounds in the Tox21 dataset

https://github.com/TingLi2016/DeepCarc
https://github.com/TingLi2016/DeepDILI

Deep Learning Approach and Implementation

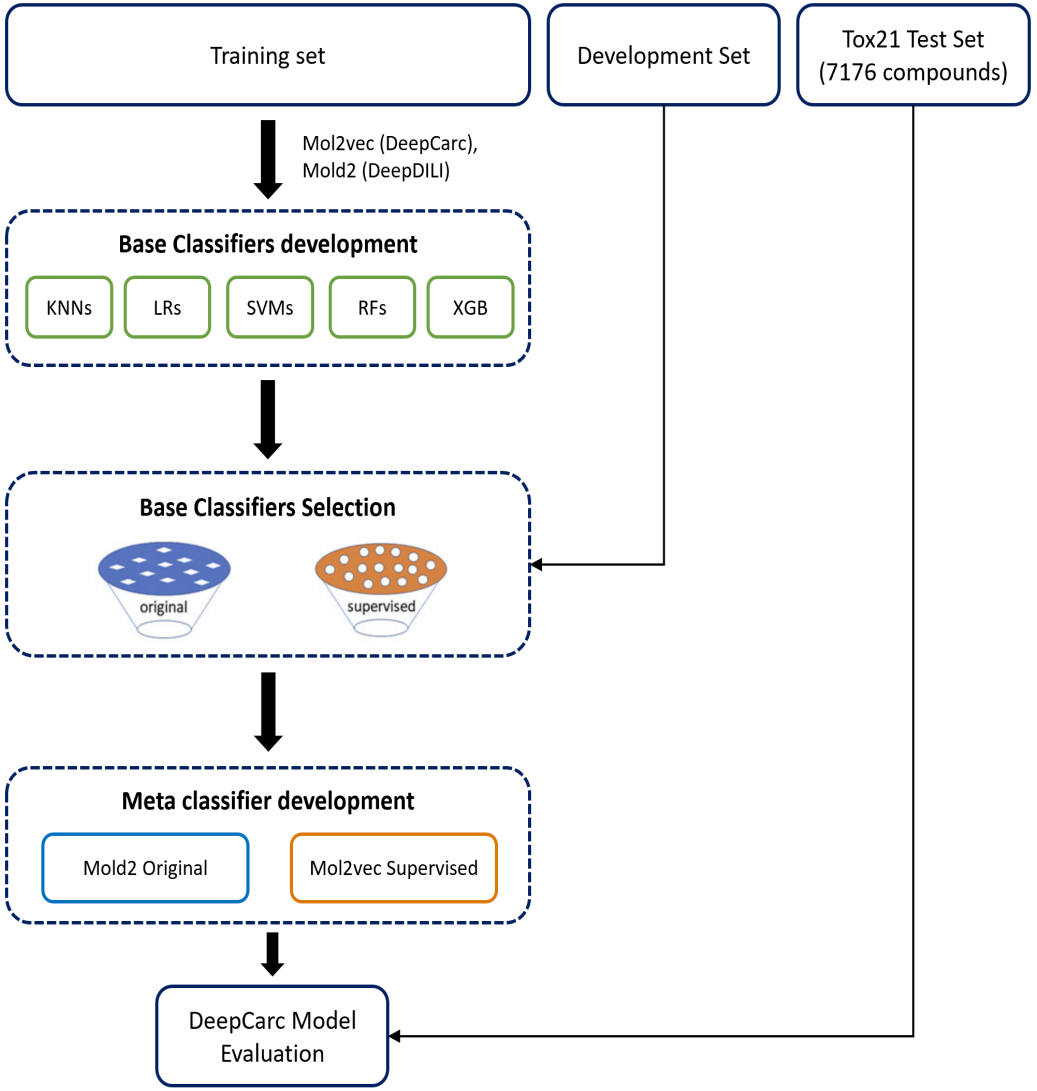


Figure 2: General workflow of the DeepCarc and DeepDILI models

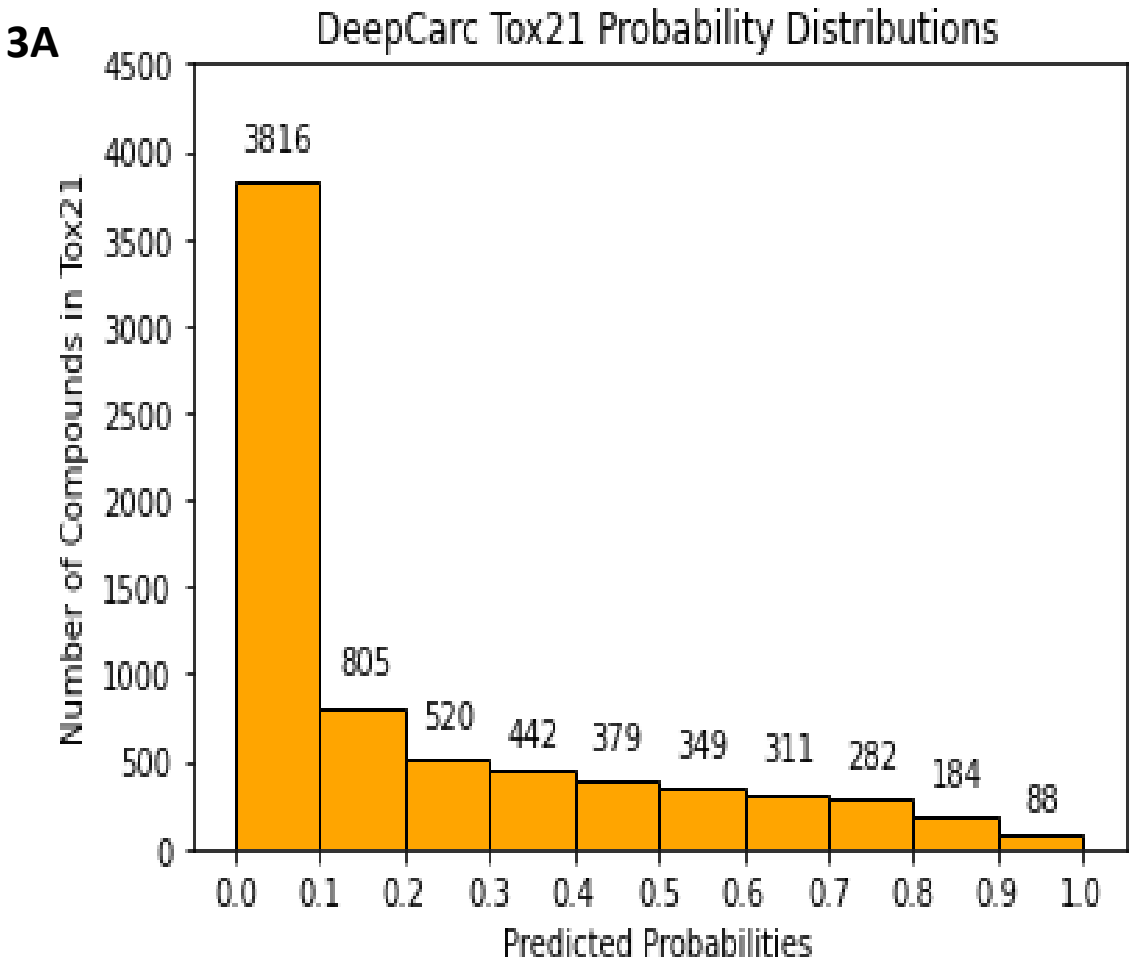
Base Classifier Selection Strategies

- Original Strategy (DeepDILI): 100 classifiers generated by each of the five algorithms with the best hyperparameters were rank-ordered based on MCC values
- Supervised Strategy (DeepCarc): The base classifiers with MCC values higher than the average MCC of both the training and development set were selected

Chemical Descriptors

- Mold2 (DeepDILI): 777 chemical-physical descriptors are calculated from chemical structure
- Mol2vec (DeepCarc): Compounds are vectors of 300 dimensions made up of individual substructures

Tox21 Probability Distributions



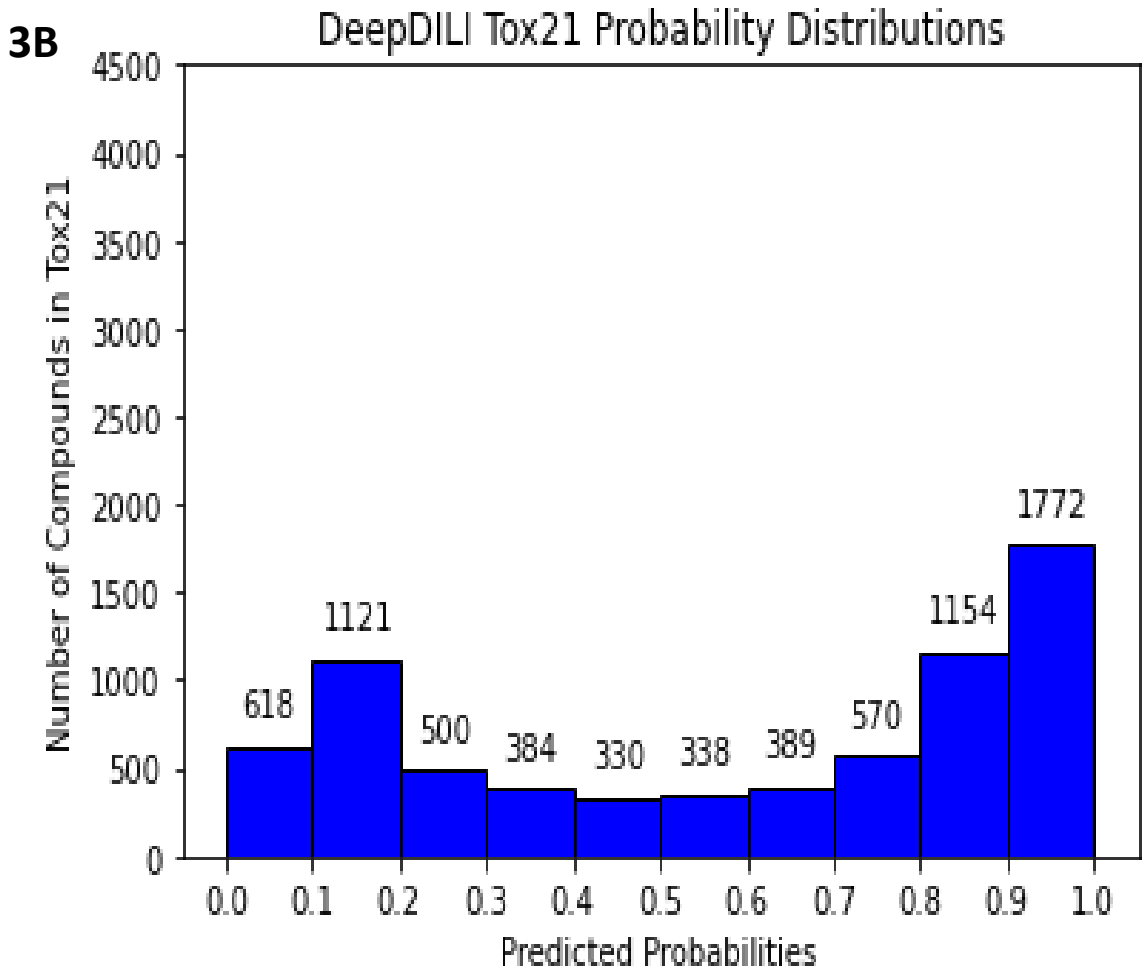


Figure 3: Tox21 Probability Distribution. A, DeepCarc. **B,** DeepDILI.

National Institute of Environmental Health Sciences
Division of the National Toxicology Program

Chemical Property Visualization of High-Risk Compounds

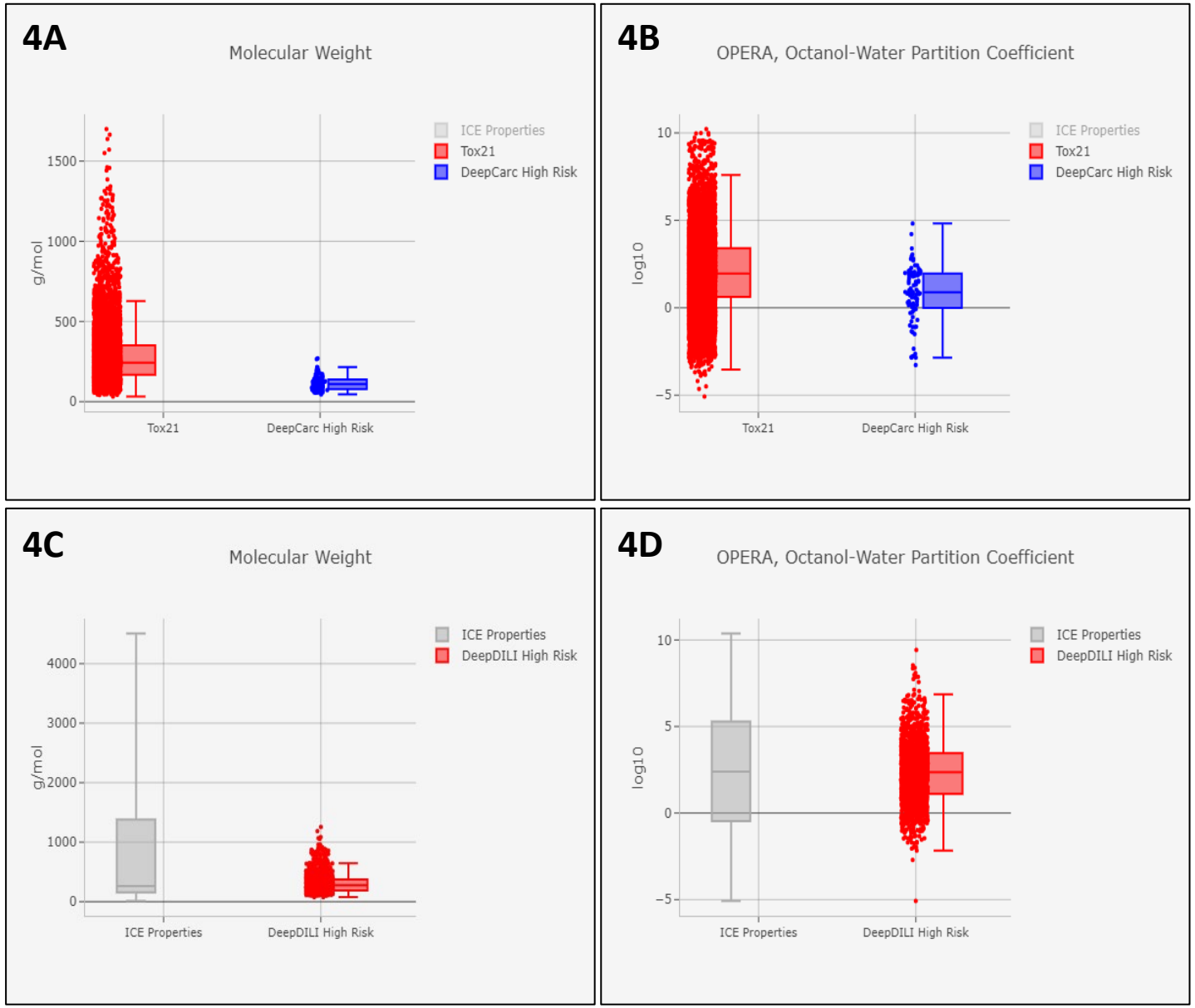


Figure 4: Visualization of Chemical Properties.

A, DeepCarc High Risk MW. **B,** DeepCarc High Risk log *P*. **C,** DeepDILI High Risk MW. **D,** DeepDILI High Risk log *P*

Integrated Chemical Environment

National Toxicology Program. 2021.

National Institute of Environmental Health Sciences
Division of the National Toxicology Program

Consumer Risk Distribution of High-Risk Compounds

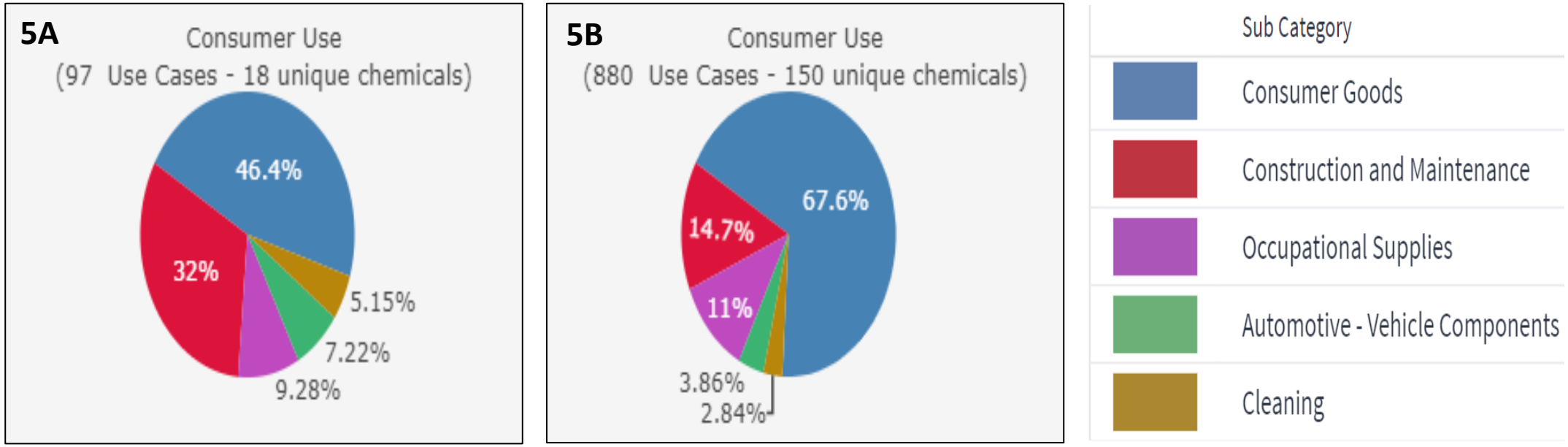


Figure 5: Consumer use distribution. A, DeepCarc high risk chemicals. **B,** DeepDILI high risk chemicals.

- Many of the high-risk compounds were not included in the US EPA's Chemical and Products Database
  - 79.5% for DeepCarc
  - 91.5% for DeepDILI

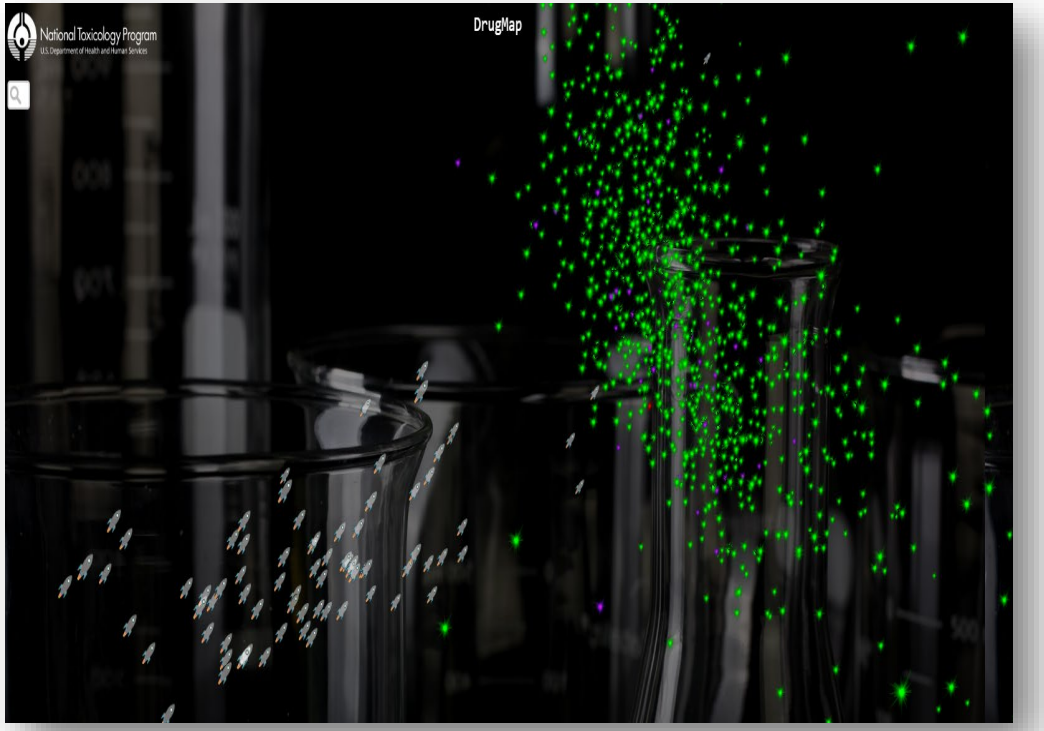
Integrated Chemical Environment

National Toxicology Program. 2021.

National Institute of Environmental Health Sciences
Division of the National Toxicology Program

DrugBankMap Visualization of High-Risk Compounds

6A



6B

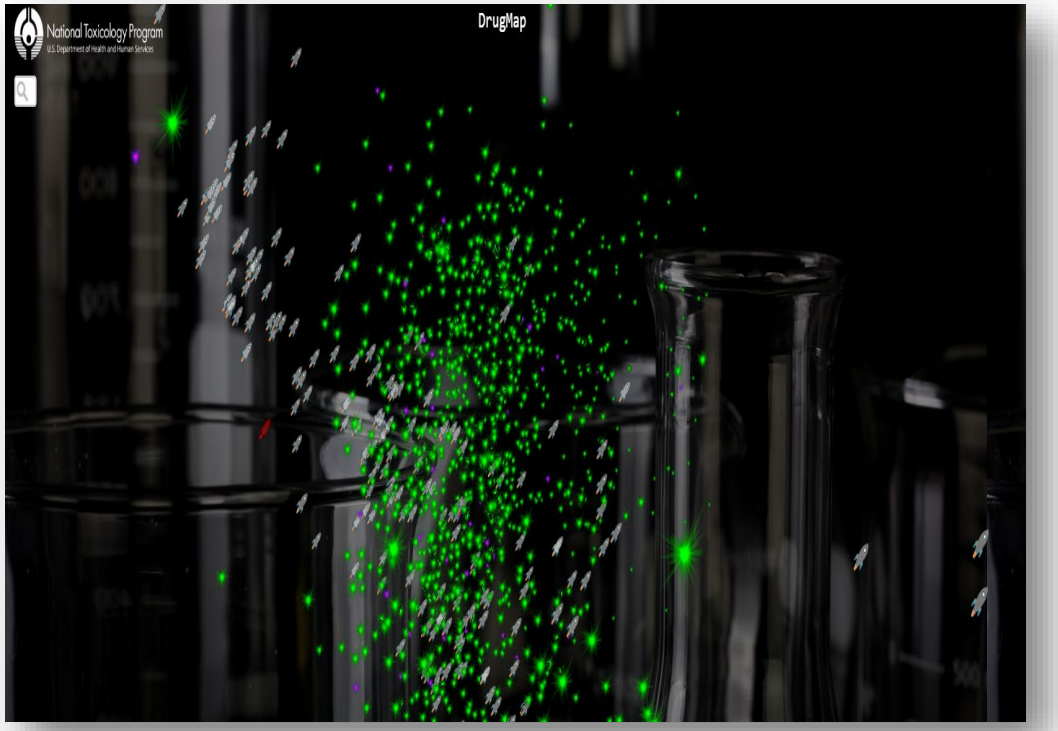


Figure 6: ChemMaps[5] Visualization on the DrugBankMap. A, DeepCarc high risk chemicals. **B,** DeepDILI top 200 high risk chemicals. (Approved drugs in green, withdrawn drugs in purple, and high-risk compounds as gray rockets)

A. Borrel, N. C. Kleinstreuer, & D. Fourches. (2018) Exploring drug space with ChemMaps.com. Bioinformatics, (1), 1–3.

Conclusions and Moving Forward

- With our increasingly data driven world and the availability of large datasets such as the Tox21 dataset, in silico methodologies for carcinogenicity assessment are emerging as powerful tools to supplement toxicity testing
- Results show that further research needs to be conducted on some of the high-risk chemicals that are widely used in the consumer space
- The DeepCarc and DeepDILI models will be applied to other datasets such as the EPA's DSSTox which contains around 1 million compounds
- DeepCarc and DeepDILI models will be further refined by exploring its performance with different chemical descriptors and by utilizing a different neural network