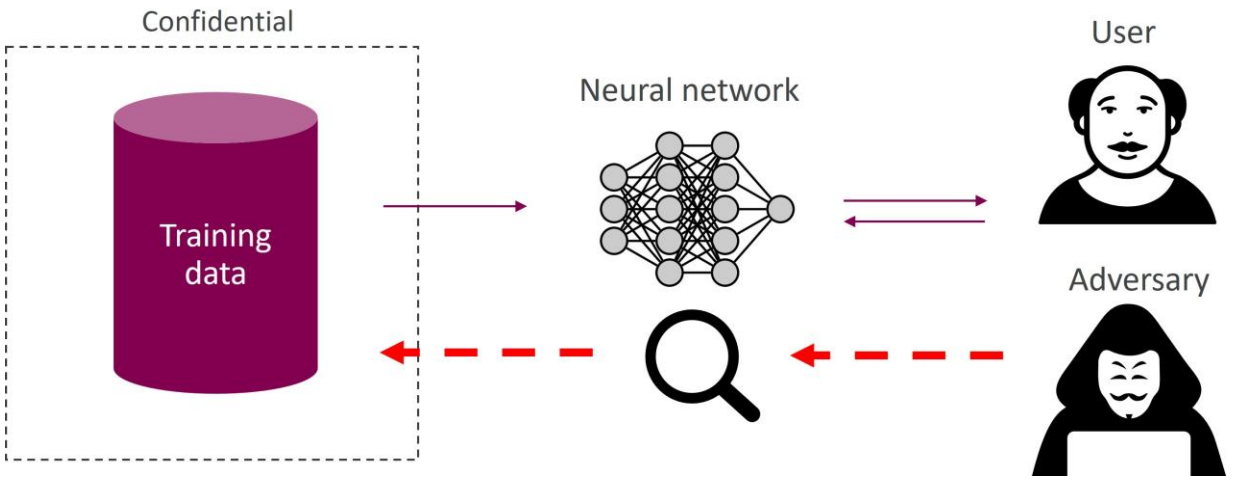
Publishing Neural Networks in Drug Discovery Might Compromise Training Data Privacy

Fabian Krüger[1,2,3], Johan Östman[4], Lewis Mervin[1], Igor Tetko[3], Ola Engkvist[1,5]

AstraZeneca R&D[1], Technical University of Munich[2], Helmholtz Munich[3], AI Sweden[4], Chalmers University of Technology[5]

Abstract

This study investigates the risks of exposing confidential chemical structures when machine learning models trained on these structures are made publicly available. We use membership inference attacks, a common method to assess privacy that is largely unexplored in the context of drug discovery, to examine neural networks for molecular property prediction in a black-box setting. Our results reveal significant privacy risks across all evaluated datasets and neural network architectures.

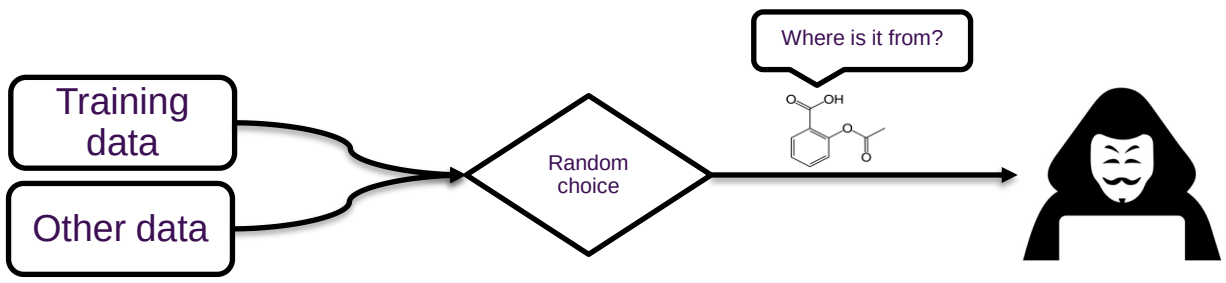


Introduction

- Datasets in drug discovery are expensive to generate. Leaking information about proprietary data can severely harm an organization.
- Organizations need to balance benefits from open science and collaboration with the scientific community with their privacy concerns.
- There is a lack of studies on how much training data information an be inferred from neural networks in a drug discovery context.

Methods

- Developed and evaluated neural networks trained on diverse molecular representations (fingerprints, graphs, SMILES) across four drug discovery datasets.
- Applied state-of-the-art membership inference attacks (LiRA[1] and RMIA[2]) in a black-box setting to measure how well attackers can identify molecules from training data.



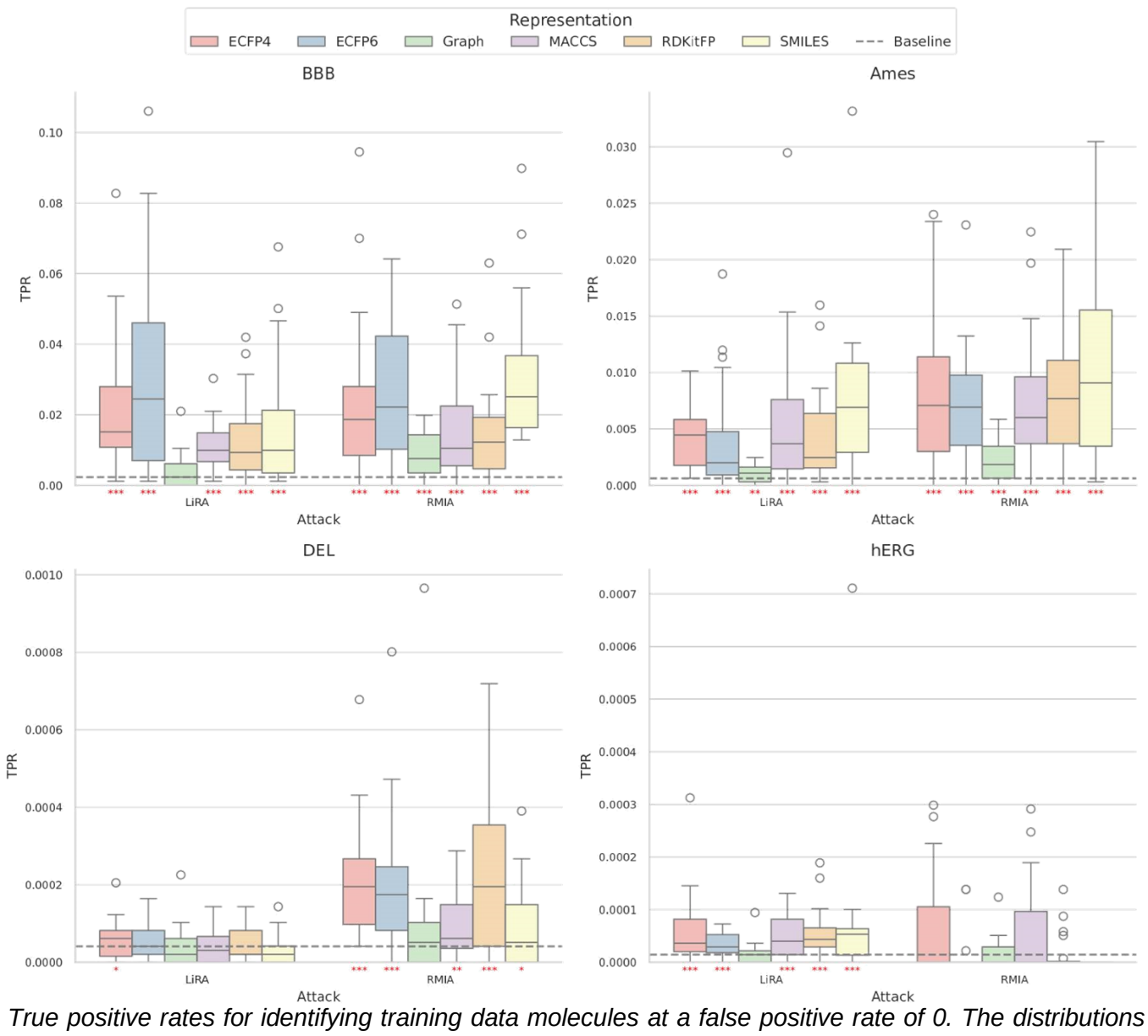
Scheme of how membership inference attacks are evaluated.

Conclusions

- It is consistently possible to identify parts of the training data, even at false positive rates as low as 0 (under some assumptions).
- Combining both attacks allows getting even more information about the training data.
- Minority class molecules are easier to identify.
- Message passing neural network has the least information leakage.

Paper:

Results



True positive rates for identifying training data molecules at a false positive rate of 0. The distributions of 20 experimental repetitions are shown for each representation and dataset, for both the likelihood ratio attack (LiRA[1]) and the robust membership inference attack (RMIA[2]). Distributions with significantly higher true positive rates (information leakage) than the baseline (random guessing) are indicated by red stars. Training dataset sizes are: 859 molecules for the blood-brain barrier permeability dataset [3]; 3,264 for the Ames mutagenicity prediction dataset [4]; 48,837 for the DNA-encoded library enrichment dataset [5]; and 137,853 for the hERG channel inhibition dataset [6].

References

1. Membership inference attacks from first principles, Carlini et al., IEEE, 2022.
2. Low-cost high-power membership inference attacks, Zarifzadeh et al., ICML, 2024
3. A bayesian approach to in silico blood-brain barrier penetration modeling, Martins et al., JCIM, 2012
4. Benchmark data set for in silico prediction of ames mutagenicity, Hansen et al., JCIM, 2009
5. Machine learning on dna-encoded library count data using an uncertainty-aware probabilistic loss function, Lim et al., JCIM 2022
6. Hergcentral: a large database to store, retrieve, and analyze compound-human ether-a-go-go related gene channel interactions to facilitate cardiotoxicity assessment in drug development, Du et al., Assay and drug development technologies, 2011

Acknowledgements

This study was partially funded by the Horizon Europe funding programme under the Marie Skłodowska-Curie Actions Doctoral Networks grant agreement “Explainable AI for Molecules - AiChemist”, no. 101120466. The work of Johan Östman was funded by Vinnova, the Swedish innovation agency, under grant 2023-03000.