The background of the slide is a blue gradient. On the left side, there is a vertical column of binary code (0s and 1s) that appears to be receding into the distance. Overlaid on this are several white chemical structures, including various rings, hydroxyl groups, and functional groups, representing molecular chemistry. The overall theme is the intersection of artificial intelligence and pharmaceutical research.

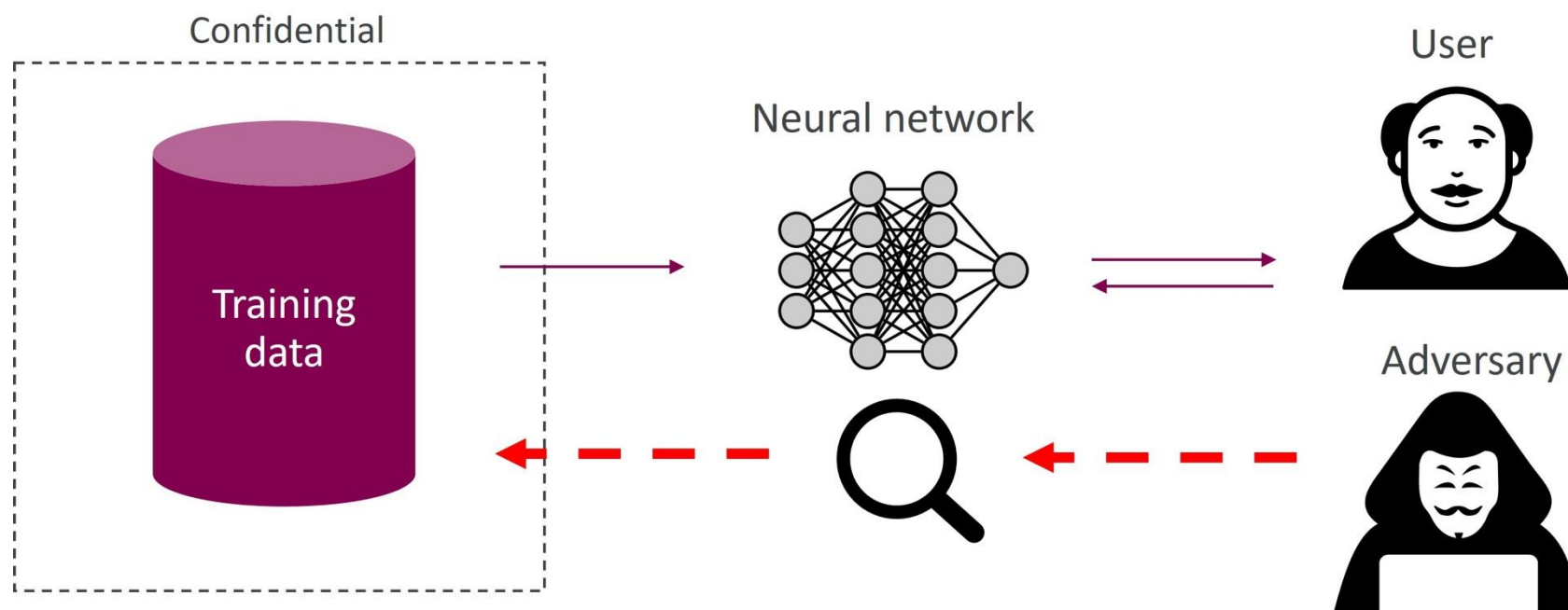
Can Publishing Neural Networks Expose Confidential Training Data? Risks in Drug Discovery

Fabian Krüger, Molecular AI,
BioPharmaceuticals R&D, AstraZeneca,
Gothenburg, Sweden



Problem

- In cheminformatics we often work with confidential data
- Open science in machine learning is important for collaboration and innovation^{1,2}
- Can we still make our trained models publicly available?



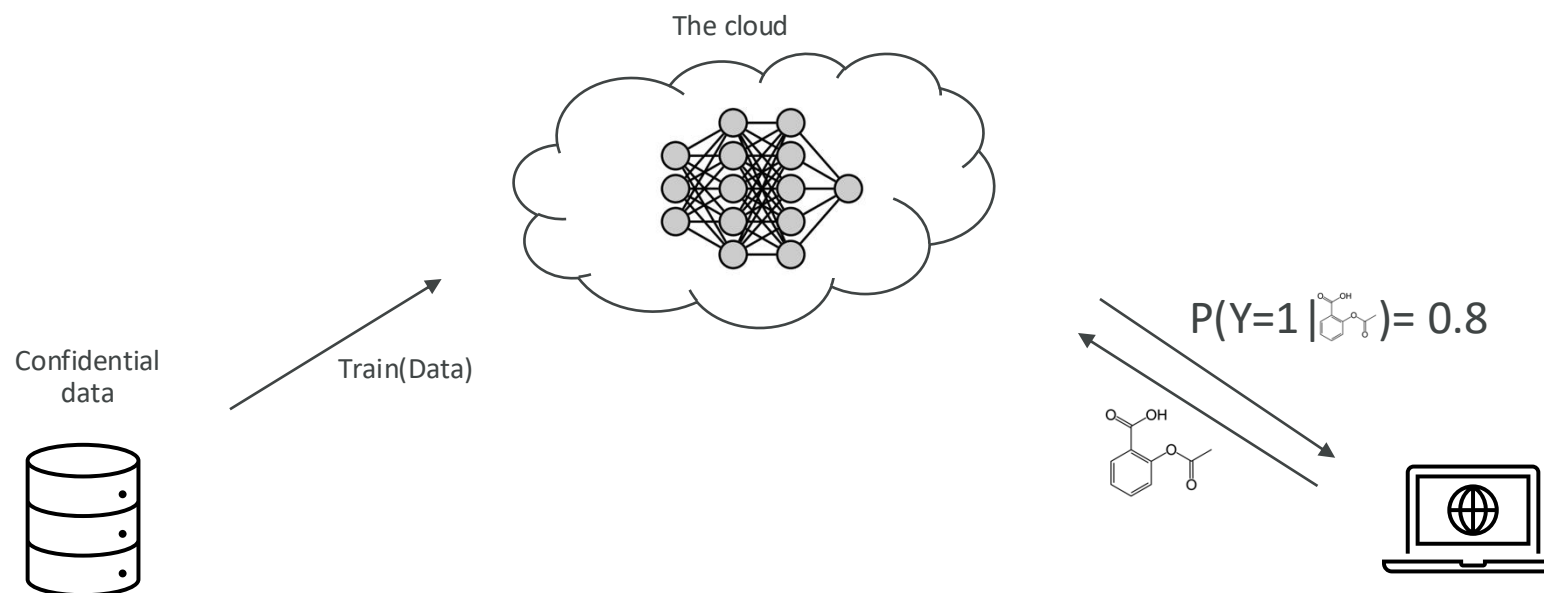
¹Yash Raj Shrestha, Georg von Krogh, and Stefan Feuerriegel. Building open-source ai. Nature Computational Science, 3(11):908–911, 2023.

²Mark Zuckerberg. Open-source ai is the path forward, July 2024. URL <https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/>. Accessed: 2024-09-25.



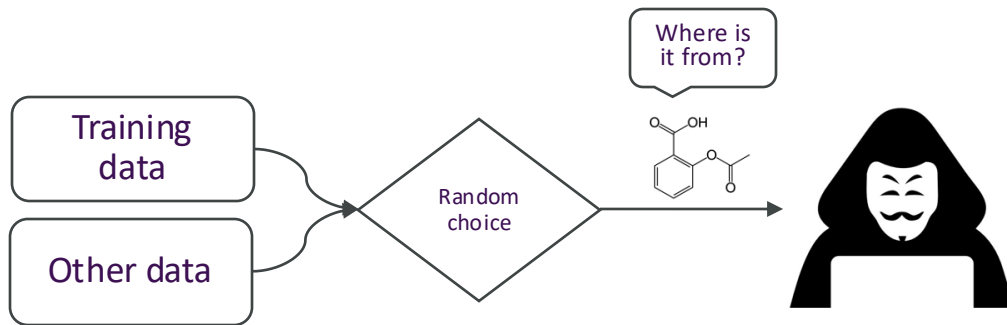
Research question

- How much information about the training data can be identified when you make a model public?
- Here we look at neural networks trained on different tasks for drug discovery
- Black-box setting (no access to weights)



Approach

- To see how much of the training data could be identified, we used membership inference attacks (MIA), which are a widely used method for privacy assessment^{1,2,3}



-
- Input:** Adversary A , Training Algorithm T , Data distribution Π
 - Sample n points from Π : $D \sim \Pi^n$
 - Train model using T on D : $f_\theta \leftarrow T(D)$
 - Flip a coin: $b \sim \{0, 1\}$
 - if** $b = 0$ **then**
 - Sample $z \sim D$
 - else**
 - Sample $z \sim \Pi(\cdot \mid z \notin D)$
 - end if**
 - Let A guess b : $\tilde{b} \leftarrow A(T, \Pi, z, f_\theta(z))$
-

¹Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP), pages 3–18. IEEE, 2017.

²Sasi Kumar Murakonda and Reza Shokri. MI privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. arXiv preprint arXiv:2007.09339, 2020.

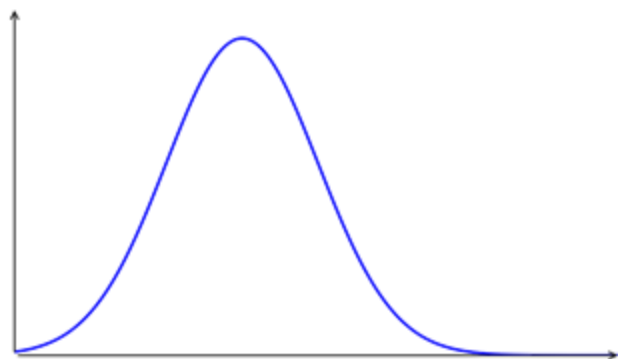
³Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In 2022 IEEE Symposium on Security and Privacy (SP), pages 1897–1914. IEEE, 2022



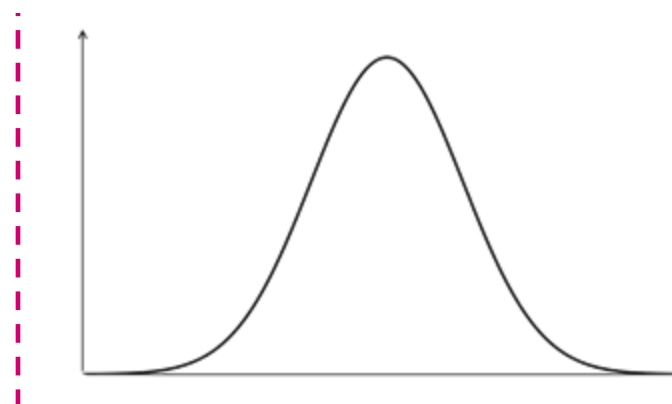
Approach

- We used two different state-of-the-art attacks (LiRA¹ and RMIA²)
- They rely on having data from a similar distribution to train so-called shadow models

Shadow model predictions for CC(=O)Oc1ccc(O)cc1
from shadow models trained on D_i



Distribution of predictions for CC(=O)Oc1ccc(O)cc1
from shadow models trained on $D_i \cup \{\text{target}\}$



Target model prediction of CC(=O)Oc1ccc(O)cc1

5 ¹Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In 2022 IEEE Symposium on Security and Privacy (SP), pages 1897–1914. IEEE, 2022.

²Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In Forty-first International Conference on Machine Learning, 2024.



Datasets

- 4 datasets (all binary classification tasks)

Dataset	Size [# molecules]	Class imbalance [% Positives]
Ability to cross blood-brain barrier (BBB) ¹	1,909	76
Mutagenicity prediction (Ames) ²	7,255	54
DEL enrichment for carbonic anhydrase IX binding (DEL) ³	108,528	4.9
hERG inhibition (hERG) ⁴	306,341	4.5

¹Ines Filipa Martins, Ana L Teixeira, Luis Pinheiro, and Andre O Falcao. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling*, 52(6):1686–1697, 2012.

²Katja Hansen, Sebastian Mika, Timon Schroeter, Andreas Sutter, Antonius Ter Laak, Thomas Steger-Hartmann, Nikolaus Heinrich, and Klaus-Robert Muller. Benchmark data set for in silico prediction of ames mutagenicity. *Journal of chemical information and modeling*, 49(9):2077–2081, 2009.

³Katherine S Lim, Andrew G Reidenbach, Bruce K Hua, Jeremy W Mason, Christopher J Gerry, Paul A Clemons, and Connor W Coley. Machine learning on dna-encoded library count data using an uncertainty-aware probabilistic loss function. *Journal of chemical information and modeling*, 62(10):2316–2331, 2022.

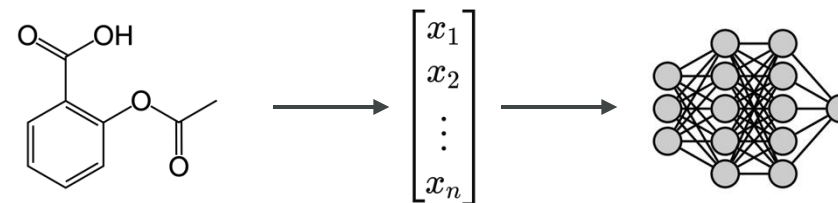
⁴Fang Du, Haibo Yu, Beiyuan Zou, Joseph Babcock, Shunyou Long, and Min Li. hergcentral: a large database to store, retrieve, and analyze compound-human ether-a-go-go related gene channel interactions to facilitate cardiotoxicity assessment in drug development. *Assay and drug development technologies*, 9(6):580–588, 2011



Models

- MLPs on different molecular representations

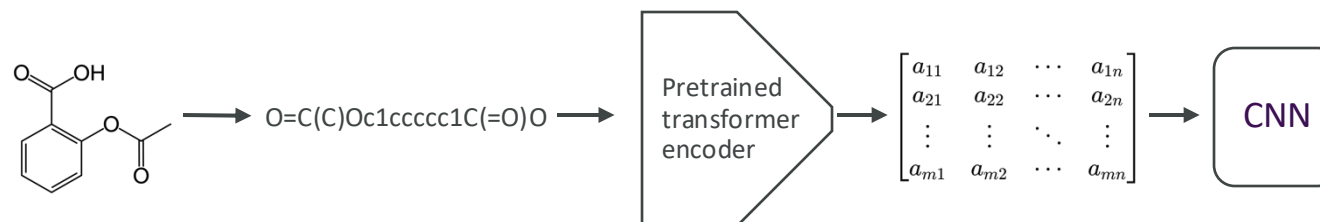
- ECFPs
- MACCS keys
- RDKitFP



- Message passing neural networks (graph)



- Transformer with CNN (SMILES)²



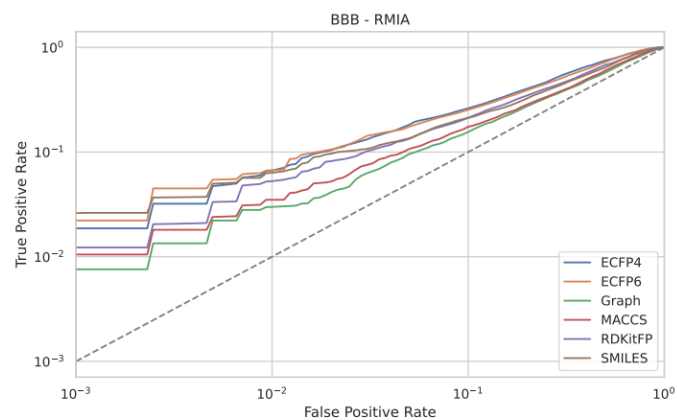
¹Heid, Esther, et al. "Chemprop: a machine learning package for chemical property prediction." *Journal of Chemical Information and Modeling* 64.1 (2023): 9-17.

²Pavel Karpov, Guillaume Godin, and Igor V Tetko. Transformer-cnn: Swiss knife for qsar modeling and interpretation. *Journal of cheminformatics*, 12:1–12, 2020.

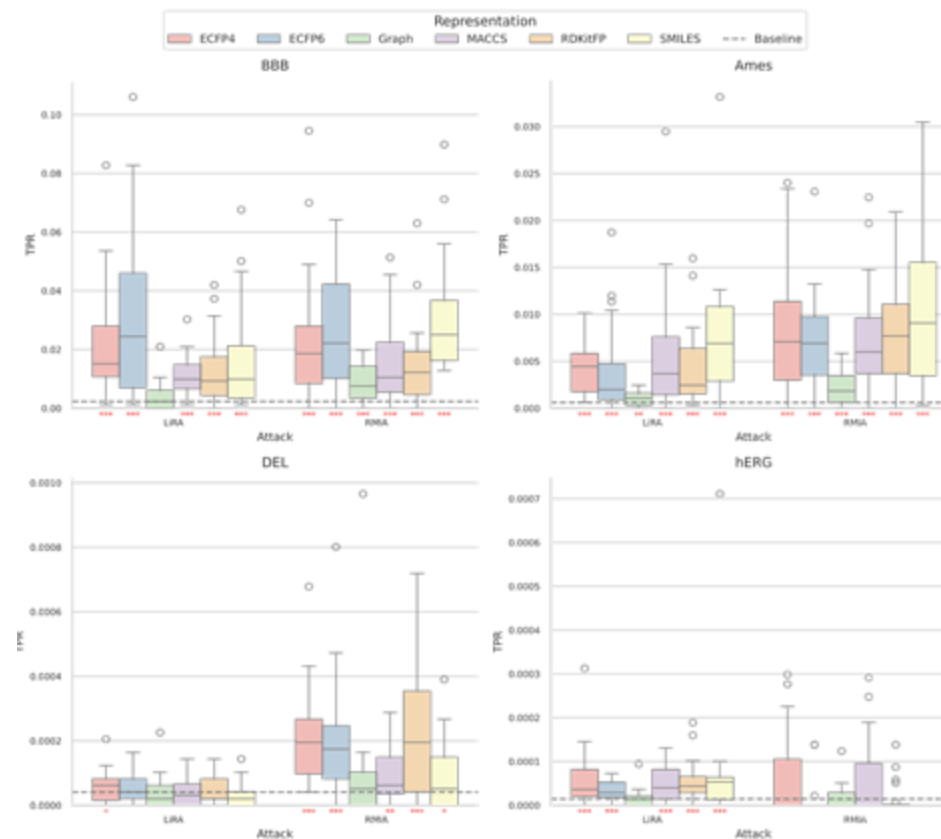


Results

- Low false positive rates (FPRs) for identifying training data members are most relevant from a privacy perspective¹
- Identifying training data is consistently possible

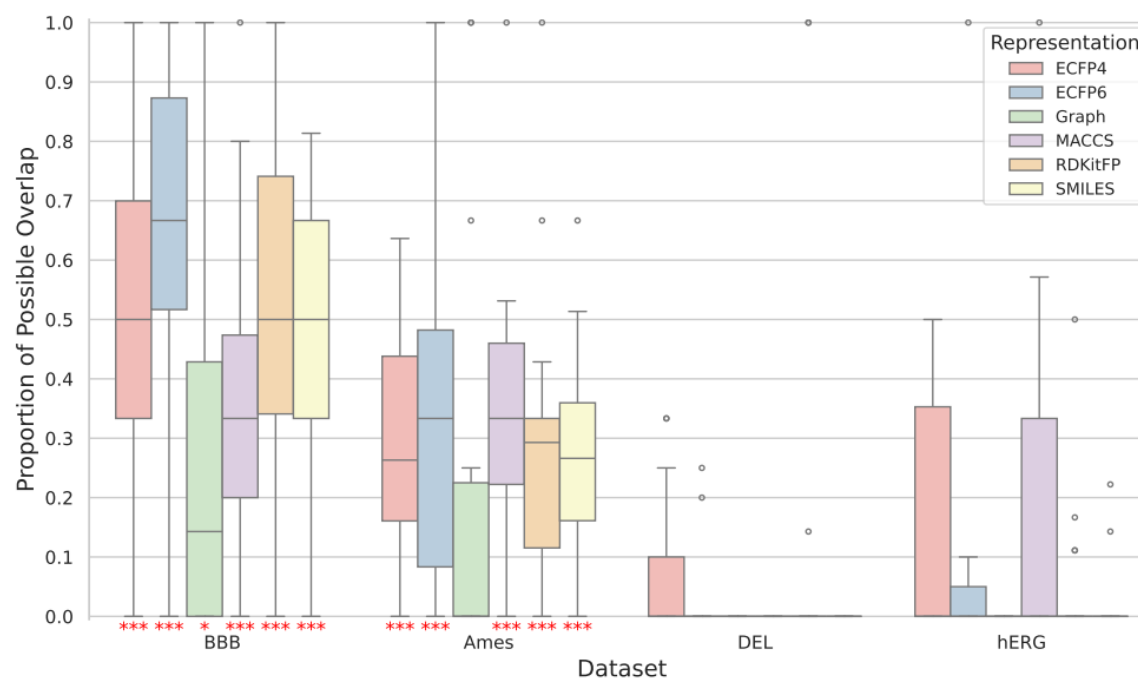


FPR=0:



Results

- Could you combine the attacks to get even more information?
- Overlap between the attacks:



Results

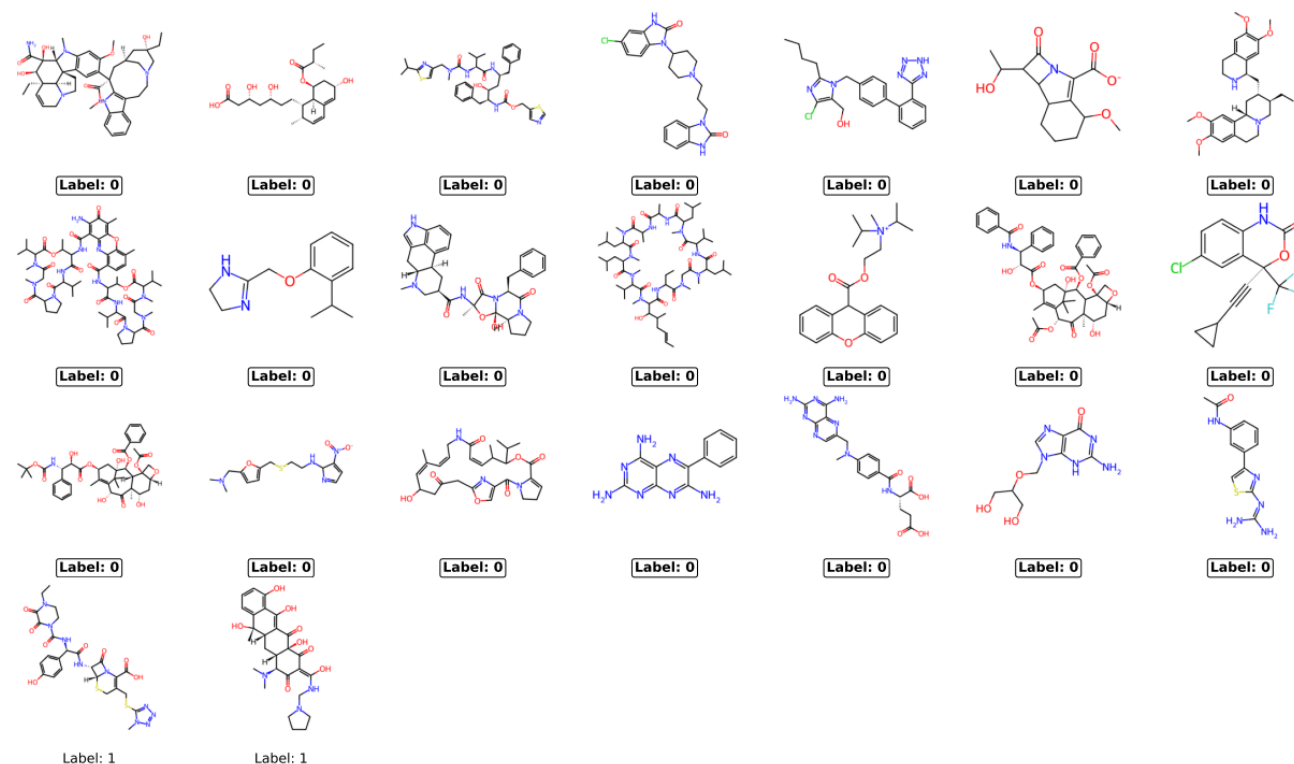
- Minority class is identified more
- Often most important structures

Dataset	Representation	LiRA		RMIA	
		Mean	Significance	Mean	Significance
BBB (0.76)	ECFP4	0.16	***	0.25	***
	ECFP6	0.07	***	0.17	***
	Graph	0.40	**	0.42	**
	MACCS	0.31	***	0.37	**
	RDKitFP	0.19	***	0.31	***
	SMILES	0.21	***	0.20	***
Ames (0.54)	ECFP4	0.54		0.45	*
	ECFP6	0.49		0.45	
	Graph	0.51		0.77	**
	MACCS	0.50		0.53	
	RDKitFP	0.60		0.47	
	SMILES	0.44		0.44	*
Del (0.05)	ECFP4	0.16		0.78	***
	ECFP6	0.12		0.82	***
	Graph	0.00	***	0.43	
	MACCS	0.23		0.69	***
	RDKitFP	0.14	**	0.62	***
	SMILES	0.05	**	1.00	***
hERG (0.04)	ECFP4	0.80	***	0.55	
	ECFP6	0.44		0.47	
	Graph	0.29		0.53	
	MACCS	0.75	***	0.78	***
	RDKitFP	0.66	***	0.76	***
	SMILES	0.72	***	1.00	***



Example case study

- Model trained on ECFP4 for predicting blood-brain barrier crossing
- 23 of 859 training structures identified at FPR=0 with LiRA
- 21 of the 23 structures were from the minority class
- Combining it with RMIA allowed identifying 53 structures at FPR=0








Conclusion

- It is consistently possible to identify parts of the training data, even at FPRs 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
- More information: <https://doi.org/10.48550/arXiv.2410.16975>

PUBLISHING NEURAL NETWORKS IN DRUG DISCOVERY MIGHT
COMPROMISE TRAINING DATA PRIVACY

A PREPRINT

 Fabian P. Krüger^{1,2,3}  Johan Östman⁴  Lewis Mervin⁵  Igor V. Tetko³  Ola Engkvist^{1,6}



Thank you.



Model classification performance

