Journal of Cheminformatics

RESEARCH

Open Access

Publishing neural networks in drug discovery might compromise training data privacy

Fabian P. Krüger[1,2,3*], Johan Östman[4], Lewis Mervin[5], Igor V. Tetko[3] and Ola Engkvist[1,6]

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. Combining multiple attacks increases these risks. Molecules from minority classes, often the most valuable in drug discovery, are particularly vulnerable. We also found that representing molecules as graphs and using message-passing neural networks may mitigate these risks. We provide a framework to assess privacy risks of classification models and molecular representations, available at https://github.com/FabianKruger/molprivacy. Our findings highlight the need for careful consideration when sharing neural networks trained on proprietary chemical structures, informing organisations and researchers about the trade-offs between data confidentiality and model openness.

Scientific contribution

This study presents the first systematic assessment of the privacy risks associated with the sharing of neural networks trained to predict molecular properties. We are the first to develop a comprehensive framework for assessing these privacy risks in the context of cheminformatics, enabling the evaluation of vulnerabilities across different molecular representations and model architectures. Our work bridges the gap between privacy research and cheminformatics, providing a foundation for safer data sharing practices in drug discovery.

Keywords Membership inference attack, Privacy, Drug discovery, Cheminformatics, QSAR, Machine learning

*Correspondence:
Fabian P. Krüger
fabian.krueger@tum.de
[1] Discovery Sciences, Molecular AI, AstraZeneca R&D, Mölndal 431 83, Sweden
[2] TUM School of Computation, Information and Technology, Department of Mathematics, Technical University of Munich, Munich 80333, Germany
[3] Molecular Targets and Therapeutics Center, Institute of Structural Biology, Helmholtz Munich - Deutsches Forschungszentrum Für Gesundheit Und Umwelt (GmbH), Neuherberg 85764, Germany
[4] AI Sweden, Gothenburg 41756, Sweden
[5] Discovery Sciences, Molecular AI, AstraZeneca R&D, Cambridge CB2 0AA, UK
[6] Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg 412 96, Sweden

Introduction

The use of neural networks has gained significant traction in early drug discovery, with organisations increasingly relying on these models for a range of important modelling tasks [1]. One of the most common applications is the prediction of molecular properties [2, 3]. The performance of these models is heavily dependent on the quality and quantity of available datasets [2]. However, generating these datasets in drug discovery is an expensive and resource-intensive process, often requiring significant investment in both time and money [4]. As a result, organisations are highly protective of their data, as they have invested significant resources in generating the proprietary datasets and are accordingly reluctant to make this information publicly available.

While organisations are interested in keeping their proprietary datasets private due to the significant investments involved, they still recognise the value of engaging with the broader drug discovery community and artificial intelligence (AI) communities [5]. In the AI research field, it is common practice to share models through open-source platforms or alternatively to offer them as secure web services, fostering collaboration and innovation [6]. This interaction is mutually beneficial, as it allows for the refinement and validation of models while also advancing the field as a whole [7]. However, this type of collaboration inevitably raises concerns about data security, an issue of growing importance in AI research [8]. As organisations seek to balance the advantages of community engagement with the need to protect valuable data, the issue of privacy is becoming increasingly important.

In this work, we adopt an interdisciplinary approach that bridges the fields of drug discovery and data privacy research. This bridge has largely been missing and we firmly believe that there are great opportunities for scientific progress by bringing the two fields closer to each other. To empirically evaluate the privacy of machine learning models, membership inference attacks have become the most widely used method [9–11]. These attacks can be conceptualized as a privacy game, where the adversary seeks to determine whether a specific sample was part of the model's training data (Algorithm 1). There are various levels of information the adversary might have access to regarding the model [12]. In our study, we focus on the so-called black-box scenario, where the adversary is provided with the output logits of the trained model, rather than the model's weights, which would correspond to a white-box scenario. This black-box scenario is similar to making machine learning models available as web services.

Algorithm 1 Membership Inference Attack. This algorithm formalizes the membership inference attack game we use to evaluate the privacy of our neural networks. The attack assumes knowledge about the underlying data distribution (chemical space) Π from which the training dataset is sampled. Given an adversary A, a training algorithm T, and the data distribution Π, the process involves sampling points from the data distribution, training a model on these samples, and then using the adversary to infer whether a specific data point (chemical structure) was part of the training set or not. The algorithm tests the adversary's ability to distinguish between data points sampled from the training set and those not included, thereby evaluating potential information leakage from the model.

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: $\tilde{b} \leftarrow A(T, \Pi, z, f_\theta(z))$

Building on the growing body of research on membership inference attacks, Hu et al. conducted an extensive survey, highlighting that they have been studied in the domains of image data, text data, tabular data, as well as node classification in graph data [13]. Among the different implementations of attacks, likelihood ratio attacks (LiRA) and robust membership inference attacks (RMIA) have been shown to be the most effective in identifying training data samples, setting state-of-the-art performance benchmarks for the most commonly used benchmark datasets [11, 14]. Despite the growing interest in membership inference attacks, their application to molecular property prediction in drug discovery remains largely unexplored. To the best of our knowledge, Pejo et al. conducted the only study about membership inference attacks in the context of molecular property prediction, but they focused on federated learning scenarios using attacks tailored to this approach [15]. The broader implications and potential risks of membership inference attacks in molecular property prediction, particularly in traditional centralised machine learning models, still require investigation.

In this study, we provide the first comprehensive analysis of membership inference attacks against neural networks trained to predict molecular properties. We thereby highlight the risk that releasing machine learning models may expose proprietary chemical structures to the public, a challenge that organisations, for instance, must consider. To our knowledge this is the first study

to investigate how different molecular representations affect the privacy of the resulting models. Additionally, we create a framework where the privacy risks of classification model architectures and representation algorithms can be assessed and compared. A scheme of our workflow is described in Fig. 1. Our study also explores whether different membership inference attacks can be used together, and we present some characteristics of the identified chemical structures that provide insights into the specific privacy risks. The approaches and findings of this study have relevance beyond the pharmaceutical sector, offering applicability to any field that relies on predictions of molecular properties, such as materials science or toxicology. Our framework also allows for the systematic assessment of privacy threats associated with predictive models in these fields.

Results

In this section, we present the results of membership inference attacks on different neural networks trained on different datasets for specific tasks: Blood-Brain Barrier crossing (BBB) to predict the ability of molecules to cross the blood-brain barrier [17], Ames mutagenicity prediction (Ames) to assess potential mutagenicity [18, 19], DNA Encoded Library enrichment (DEL) to analyse enrichment [20], and inhibition of the potassium ion channel encoded by the human ether-à-go-go-related gene (hERG) to assess cardiac toxicity risks [21]. The datasets differ in size with BBB and Ames being relatively small (859 and 3,264 training data molecules) and DEL and hERG being relatively large (48,837 and 137,853 training data molecules). We explore the potential of combining different attacks to identify additional molecules contained in the training data. We also investigate whether the identified molecules have distinct properties that distinguish them from the rest of the training data. Finally, we provide a detailed example of a specific attack to illustrate our findings.

Membership inference attacks

We wanted to see if we could identify whether a molecule was part of the training data from querying a neural network and analysing its outputs. To achieve this, we used two different membership inference attacks: likelihood ratio attacks (LiRA) and robust membership inference attacks (RMIA) [11, 14]. We evaluated their ability to distinguish between molecules in the training data and those outside it by measuring the true positive rate (TPR) at a false positive rate (FPR) of 0. In this context, we refer to molecules that were part of the training data as positives. Evaluating membership inference attacks at low FPRs was recommended by Carlini et al. [11]. Here we examine the TPR at an FPR of 0, which is the most conservative approach. For models trained on smaller datasets, we observed significantly higher TPRs than would be observed when randomly guessing if the chemical structure was part of the training dataset (Fig. 2). For example, in the blood-brain barrier crossing dataset, median TPRs were between 0.01 and 0.03 for most representations, corresponding to the identification of between 9 and 26 of the 859 training molecules. The baseline in our experimental setup for identifying molecules by chance is identifying 2 molecules of the training data (See Supplementary Information for a comprehensive derivation of this baseline). Models trained on larger datasets also showed significantly high TPRs, but only for

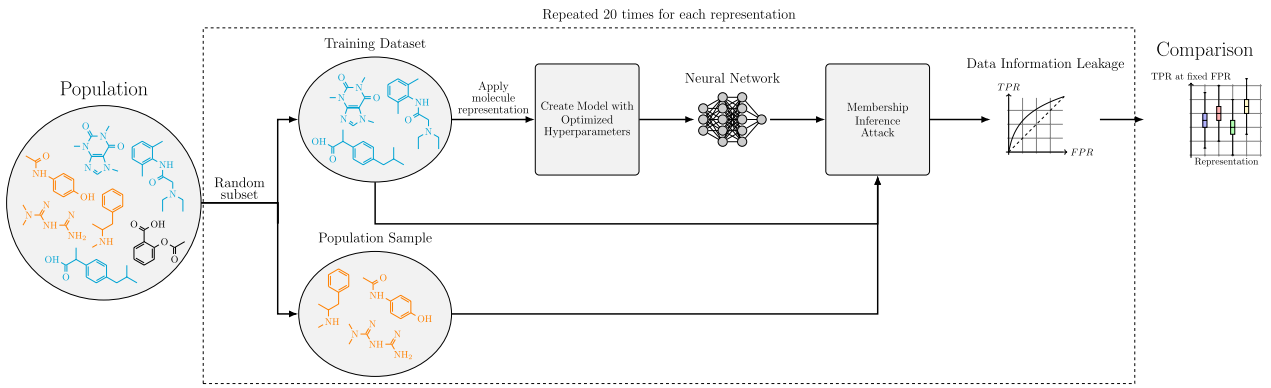

Fig. 1 Overview of our workflow to evaluate privacy risks of neural network for molecular property prediction. Two random, non-overlapping subsets are created from each dataset. One subset is transformed into the desired molecular representation and used to train a neural network, optimised through Bayesian hyperparameter tuning [16]. We then apply membership inference attacks (Algorithm 1) to determine if chemical structures in the training data can be distinguished from those in the other subset. We evaluate this using two different attack implementations. This process is repeated 20 times for each dataset and molecular representation. We assess the results by analyzing true positive rates at fixed false positive rates, comparing them to random guessing, and examining the impact of the molecular representations

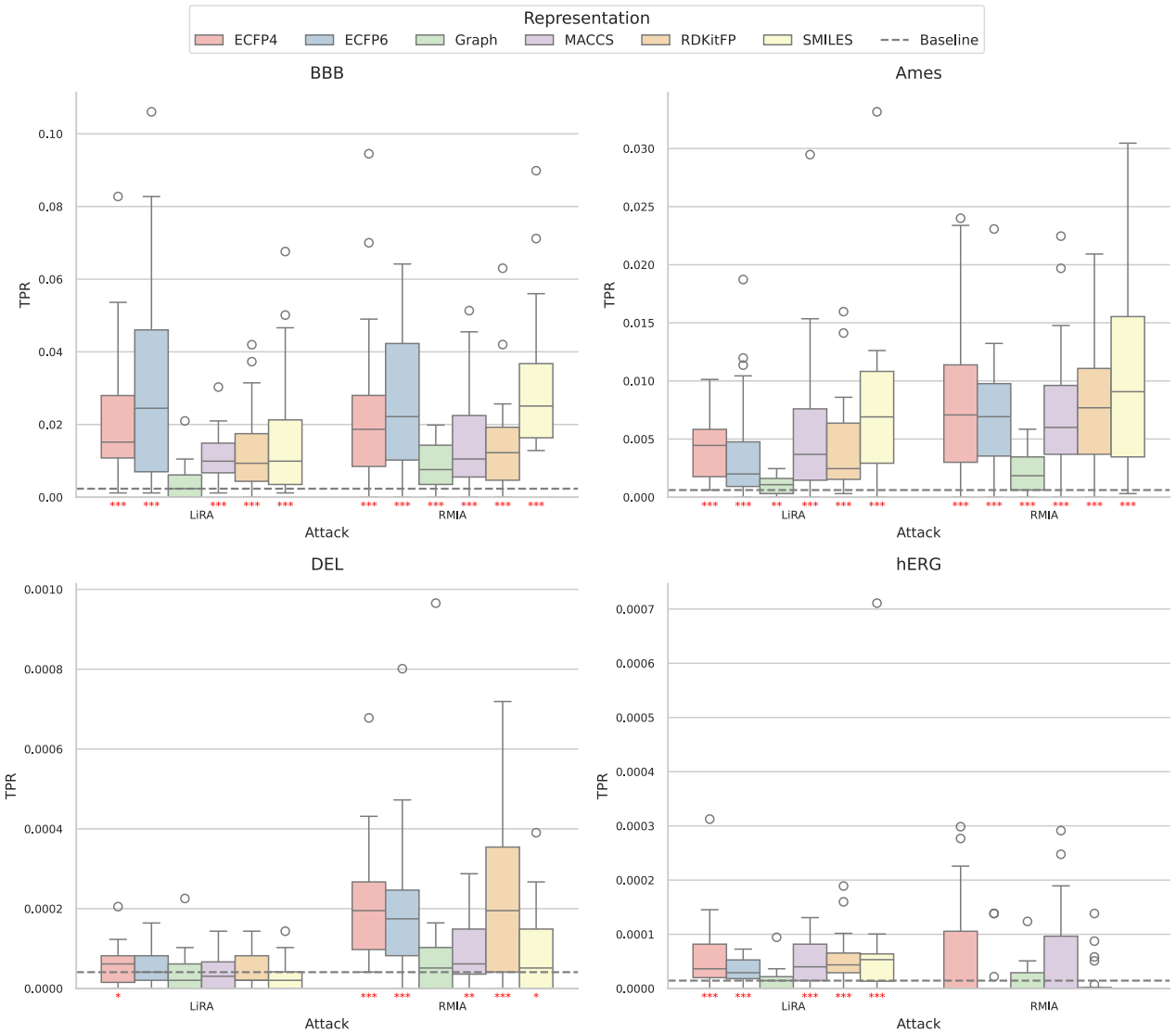

Fig. 2 True positive rates for identifying training data molecules at a false positive rate of 0. The distributions of 20 experimental repetitions are shown for each representation and dataset, for both the likelihood ratio attack (LiRA) and the robust membership inference attack (RMIA). Distributions with significantly higher true positive rates than the baseline are indicated by red stars. A single star represents a p-value less than 0.05, two stars represent a p-value less than 0.01, and three stars represent a p-value less than 0.001. Training dataset sizes (total amount of positives) are: 859 molecules for the blood-brain barrier permeability dataset; 3,264 for the Ames mutagenicity prediction dataset; 48,837 for the DNA-encoded library enrichment dataset; and 137,853 for the hERG channel inhibition dataset

one of the attacks, which varied between datasets (Fig. 2). The observed TPRs decreased with increasing dataset size.

To verify the consistency of our trends, we repeated our analysis of the TPR at an FPR of 10^{-3}, as shown in Supplementary Fig. 1. We observed similar trends at this FPR. One notable difference was that RMIA always performed at least as well as LiRA across every dataset and representation. Specifically, RMIA was significantly better in half of the cases. For the other half, no significant difference was observed. In addition, even for the larger datasets, RMIA consistently provided higher TPRs than the baseline. We also investigated the corresponding ROC curves for all datasets and representations, which show our trends are consistent even for larger FPRs (Supplementary Fig. 2). The high TPRs across all four datasets at both FPRs indicate significant information leakage, showing that chemical structures from the

training data can be identified. The amount of information leakage seems to be higher for models trained on smaller datasets.

When comparing different molecular representations for neural networks, we found that models trained on graph representations showed the least information leakage across all datasets (Fig. 2). The graph representation consistently had the lowest TPRs across all datasets and attacks, with a median TPR that was on average 66% ± 6% lower than median TPRs of the other representations at an FPR of 0. In fact, for our larger datasets (DEL enrichment and hERG channel inhibition), models trained on graph representations were the only ones for which it was not possible to identify more training data molecules than by random guessing (Fig. 2). We observed the same trend for an FPR of 10^{-3}, where the graph representation consistently had the lowest TPRs (Supplementary Fig. 2). We tested whether this was due to differences in model performance (Fig. 3), but found no clear correlation between model performance and information leakage. For the small datasets, most of the models trained on different representations performed similarly. For the larger datasets, there were some outliers in model performances. In the DNA encoded library enrichment dataset, this included models trained on MACCS keys, which performed significantly worse than the other representations. In the hERG channel inhibition dataset this included models trained on graph and SMILES representations, which performed significantly better than the other representations. Our findings suggest that graph representations combined with message passing neural networks may offer the safest architecture in terms of data privacy, without sacrificing model performance.

Combining membership inference attacks

After confirming that both membership inference attacks could identify molecules from the training data, we investigated whether they identified the same molecules or whether they could be used together to gain more information about the training data. To do this, we calculated the percentage of maximum possible overlap between the sets of molecules identified by each attack (Fig. 4). For our small datasets, we observed significantly higher overlap than would have been observed by chance if the attacks were completely uncorrelated. However, the overlap was

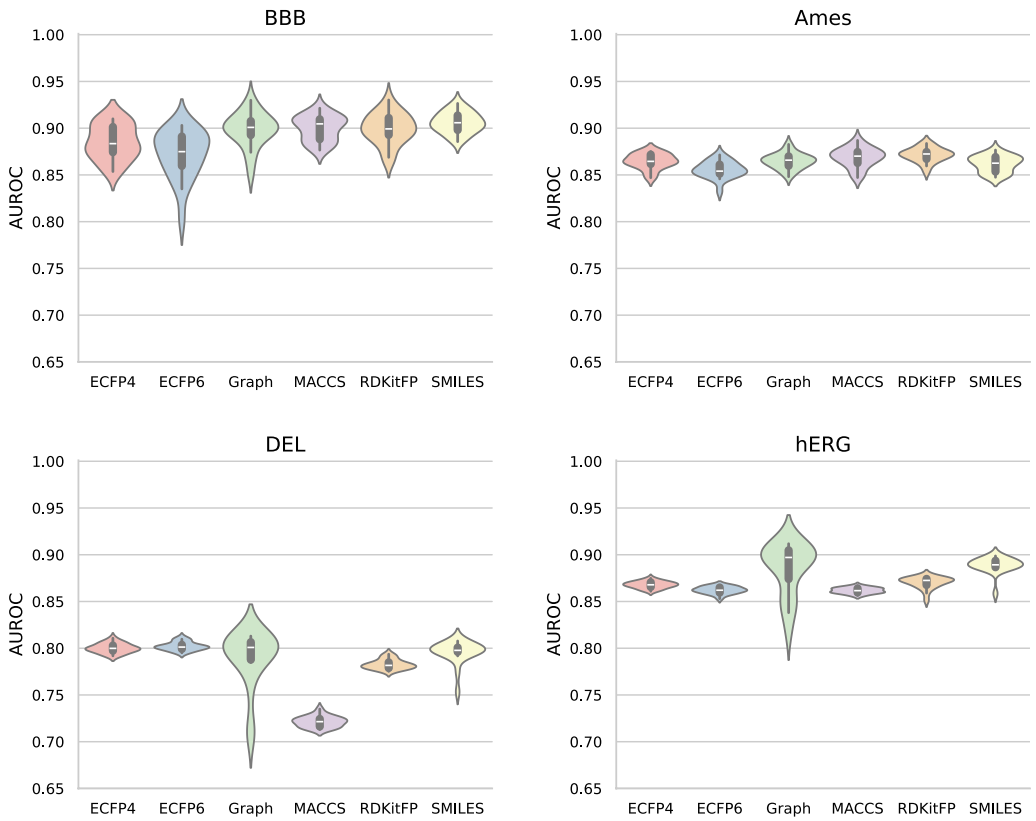

Fig. 3 Classification performance of neural networks trained on different molecular representations in molecular property prediction tasks. The performance is measured as the area under the receiver operating characteristic curve (AUROC). The performance is displayed as the distribution over 20 experiment repetitions

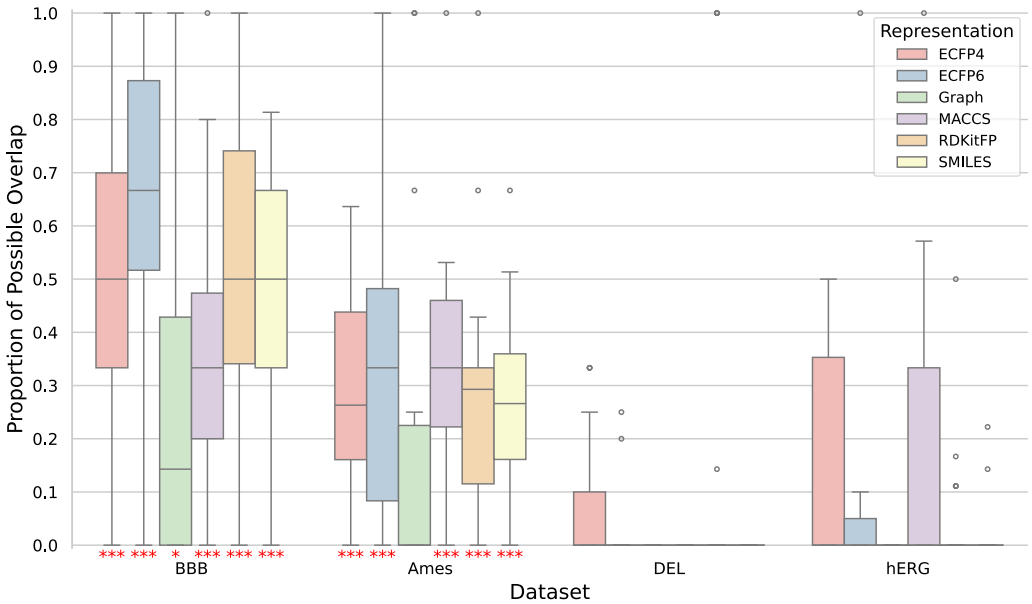

Fig. 4 Overlap between the sets of molecules identified by the likelihood ratio attack (LiRA) and the robust membership inference attack (RMIA). The percentage of possible overlap is defined as the proportion of molecules from the smaller set that are also present in the larger set. The less overlap exists between the attacks, the more information is gained when combining them. Overlap that was significantly higher than observed when randomly drawing two uncorrelated subsets is indicated by red stars. A single star represents a p-value less than 0.05, two stars represent a p-value less than 0.01, and three stars represent a p-value less than 0.001

still well below 100%, indicating that using both attacks can identify a wider range of molecules in the training data. For our larger datasets (DEL enrichment and hERG inhibition), there was no significant overlap, which is reasonable given our earlier findings that only one of the attacks significantly outperformed random guessing in each dataset. How much the observed overlap deviated from overlap occurring due to chance is shown in Supplementary Fig. 3. Our results suggest that using multiple different membership inference attacks is advantageous and allows the identification of more molecules from the training data.

We also investigated the overlap of identified molecules in models trained on different representations. We found a consistently large overlap between models trained on ECFP4 and ECFP6. For other representations, the overlap varied depending on the dataset and the attacks used. Detailed results can be found in Supplementary Fig. 4.

Analysing the identified training data molecules

Next, we wanted to see if the molecules identified from the training data shared any common characteristics. To do this, we analysed whether they differed in their distributions of property labels and molecular sizes compared to the overall training data. For the property labels, we found that the identified molecules had a significantly higher proportion of minority class molecules compared to the overall dataset (Table 1). The minority class refers to the less frequently occurring label category within a dataset, such as active compounds in a screening assay where the majority are inactive. This significant difference in label distribution was observed in all our imbalanced datasets and held true for both small datasets (blood-brain barrier crossing) and larger ones (DNA encoded library enrichment, hERG channel inhibition) across both membership inference attacks. We confirmed this finding by examining the TPRs of minority class molecules and discovered that their TPRs were consistently higher than the overall TPRs (Supplementary Fig. 5). Specifically, the median TPR of the minority class was approximately three times greater for all representations of the blood-brain barrier crossing dataset and up to 20 times greater for some representations of the DNA encoded library enrichment and hERG channel inhibition datasets. Detailed TPR distributions for all datasets and representations can be found in Supplementary Fig. 5. Regarding molecular sizes, we only found differences between identified and not identified structures in models trained on ECFP representations (Supplementary Figure 6). For models trained on other representations, we did not find any significant differences. While the identified structures do not seem to show a clear trend

Table 1 Property label distributions of the identified molecules and the overall datasets.

		LiRA		RMIA	
Dataset	**Representation**	**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	***

The amount of positive compounds in each dataset is written in parentheses in the 'Dataset' column. The numbers in the 'Mean' columns refer to the percentage of positive compounds in the identified molecules. Stars indicate significant differences in the property label distribution of the identified molecules compared to the property label distribution in the training data. A single star represents a p-value less than 0.05, two stars represent a p-value less than 0.01, and three stars represent a p-value less than 0.001

regarding their molecular size, our findings do indicate that it is easier to identify molecules from the minority class.

We also investigated whether molecules with uncommon structural features are easier to identify. Uncommon structures were defined based on both their highest (nearest neighbour) and average Tanimoto similarity to the rest of the training data, and identification was assessed at an FPR of 0. For the highest Tanimoto similarity, Mann–Whitney U tests revealed that in more than 80% of dataset-representation combinations, the identified molecules had significantly lower similarity to their nearest neighbour in the training set compared to non-identified molecules. In addition, we examined whether the fraction of identified molecules varied systematically with Tanimoto similarity-following trends such as linear or exponential relationships-but no consistent pattern emerged across all combinations of datasets and representations (Supplementary Figures 7 and 8). Similar results were observed for average Tanimoto similarity. In 75% of cases, identified molecules had a significantly lower average similarity to the rest of the training data compared to non-identified molecules. However, when analysing the fraction of identified molecules across different similarity values, we again did not observe a consistent relationship between Tanimoto similarity and identification rates (Supplementary Figures 9 and 10). Overall, these results show that molecules with lower structural similarity to the training data tend to be easier to identify, but their identification rates do not follow a simple, systematic trend based on similarity alone.

Case study

To illustrate our results, we present a specific example of attacking one neural network model trained to predict whether molecules can pass the blood-brain barrier. Molecules are represented by ECFP4s, a common representation in many related applications. This particular model was chosen because it is representative of the 20 experimental repetitions we conducted, with its TPR falling within the interquartile range of our results. Figure 5 shows the chemical structures identified using LiRA on this model under the most stringent conditions (an FPR of 0). It was possible to identify 23 of the 859 structures from the training data (Fig. 5). The baseline for random guessing in that case is identifying 2 of 859 structures (See Supplementary information). 21 of the 23 identified structures are from the minority class (Fig. 5). When we relaxed the FPR to 1.1×10^{-2} (allowing for 10 false positives among the identified structures), we were able to identify 100 structures from the training data (baseline for random guessing is 10 structures in that case). This illustrates the rapid increase in identified structures as the restrictions on the FPR are relaxed. Additionally, when combining both LiRA and RMIA, we identified 53 structures at an FPR of 0. We hope that this concrete illustration shows the potential risks that membership inference attacks pose to neural network models used in drug discovery.

Discussion

We investigated if it is possible to identify molecules from the training data only using the output of trained neural networks, a so-called black-box attack scenario.

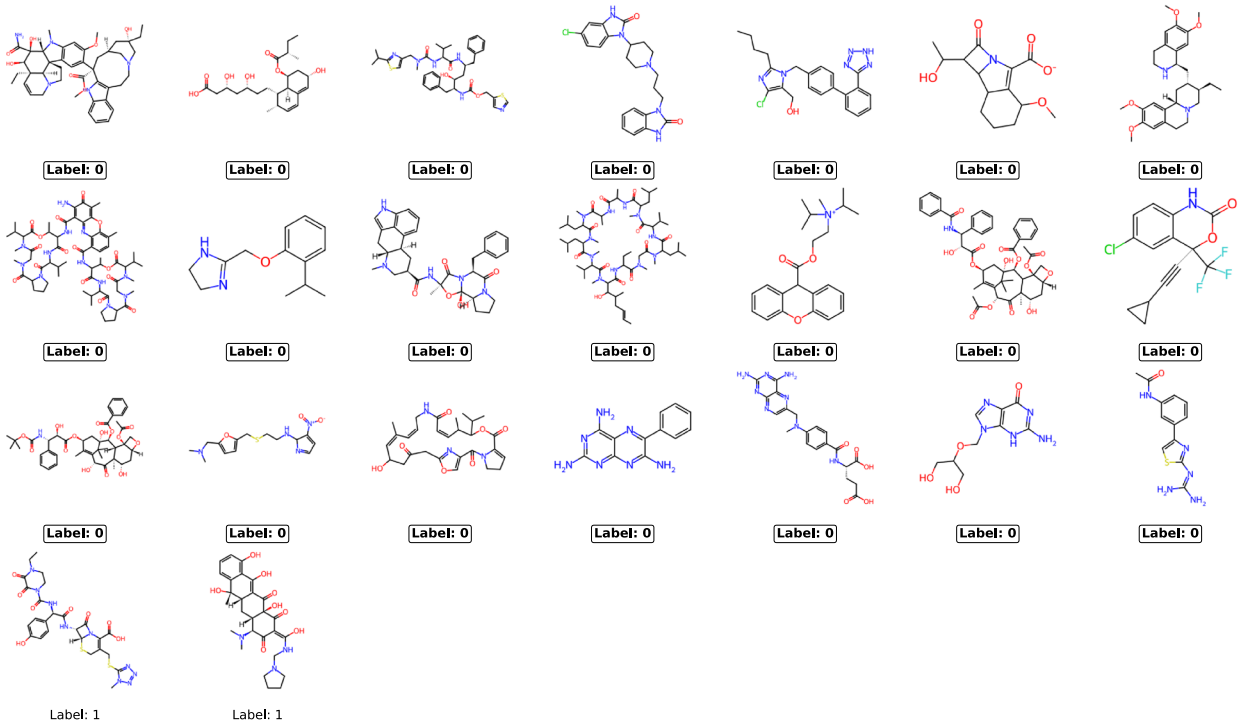

Fig. 5 Chemical structures identified using the likelihood ratio attack (LiRA) against a neural network model trained to predict whether molecules pass the blood-brain barrier. Molecules were represented using ECFP4s in this model. Structures that are from the minority class have the label 0 and are surrounded by a solid line. These structures correspond to molecules that cannot pass the blood-brain barrier. It was possible to identify 23 of the 859 training structures at an FPR of 0

To investigate this question, we have applied state-of-the-art membership inference attacks to neural networks trained on different machine learning tasks for molecular property prediction. We showed that it is possible to confidently identify a subset of the training data. We also showed that combining multiple different membership inference attacks allows us to identify even more molecules since each attack identifies different molecules. Furthermore, we investigated the identified molecules and found that they contain a much higher proportion of molecules from the minority class. Thus the investigation presents evidence that there can be significant information leakage of chemical structures from the training data when publishing a trained neural network model, which we will discuss in the following section.

It is important to note that our results focus on membership inference attacks against neural networks trained on classification tasks. This investigation does not cover regression tasks. Further research is needed to explore this area.

A limitation of the membership inference attacks we used is that they require the adversary to have data similar to the training data of the target model. While in some real-world scenarios it may hold — for instance, many organizations do have comparable internal datasets or can leverage publicly available datasets [22] — this does not fully capture the complexity of real-world applications. While assuming that the adversary has data from a similar distribution is a useful starting point for exploring privacy vulnerabilities, in drug discovery, private datasets often contain novel chemistries or rare scaffolds that lie outside common public libraries, potentially degrading the efficacy of these attacks when the adversary's data distribution diverges from that of the target model. Although Shokri et al. [9] showed that synthetic data generated by the target model can still be used to perform attacks requiring shadow models, the feasibility of this approach for molecular data, where the gap between known and unexplored chemical space can be substantial, requires further investigation. Future work should address how these attacks perform under distribution shifts in order to better assess their applicability under these conditions.

In a real world scenario, this might often be the case. For many tasks in drug discovery, there are some small publicly available datasets [22]. Additionally, many organisations have their own internal datasets for these tasks. Furthermore, Shokri et al. showed that even when similar data is not available, synthetic data generated from the target model can be used to successfully perform attacks that require shadow models [9].

We also want to emphasize that membership inference attacks assess whether it is possible to identify samples from the training data, not whether it is possible to reconstruct the training data from the model. These attacks are commonly used to assess information leakage in privacy assessments and viewed as a building block towards other attacks, e.g. reconstruction attacks [12]. In the context of drug discovery, they may have even more practical applications. For example, if an organisation offers neural network based molecular property predictions as an online service, membership inference attacks could determine whether specific molecules were part of the model's training data. Since the presence of a molecule in the training data suggests that it is being actively researched, a competitor could use this information to gain valuable insights that could give them a strategic advantage.

Our study shows that neural networks trained for molecular property prediction in drug discovery can leak training data information, as demonstrated through membership inference attacks. However, message-passing neural networks using graph representations of molecules showed significantly reduced vulnerability to these attacks. We argue that this shows that these models are the safest architecture in terms of privacy conservancy of the training data in our setting. An alternative interpretation could be that message-passing neural networks are not inherently safer, but rather that the specific membership inference attacks we used were less effective against this particular combination of model and representation. However, we think this is very unlikely, as the results are held across two different attacks, both of which rely only on model outputs rather than architecture-specific features. The only way the attacks are influenced by the specific architecture is through the training of shadow models that share the architecture of the target model. Notably, LiRA and RMIA are robust to mismatches in shadow model architectures, as shown by Carlini et al. and Zarifzadeh et al. [11, 14], meaning that variations in shadow model architectures do not significantly affect the success of the attacks. This supports our claim that graph representations of molecules with message-passing neural networks are the safest architecture in terms of protecting training data privacy in drug discovery.

We are confident that our results would be similar even if attacks tailored to graph classification neural networks, such as those proposed by Wu et al. [23], were used. Our conclusion is supported by Zarifzadeh et al. [14], who showed both theoretically and empirically that the Attack-P method of Ye et al. [24] — which is essentially identical to the threshold-based attack of Wu et al. — is less effective than both LiRA and RMIA. Therefore, we focused on the use of RMIA and LiRA, as they are widely recognised as state-of-the-art techniques in the field and can be applied to any model architecture.

Our findings align with those of Zarifzadeh et al., who investigated membership inference attacks in the domains of computer vision (using CIFAR-10, CIFAR-100, and CINIC-10 datasets) and tabular data (using the Purchase-100 dataset) [14]. At a false positive rate (FPR) of 0, they reported true positive rates ranging from 0.0082 to 0.0778, which is in the same range as our results. This indicates that the findings of attacks on neural networks in other deep learning fields translate into the field of molecular property prediction. Another finding of Zarifzadeh et. al was that RMIA consistently outperformed LiRA [14]. They derived this both theoretically and empirically. Our results generally support this, with one exception: for attacks on the hERG channel inhibition dataset, LiRA outperformed RMIA at an FPR of 0. However, at an FPR of 10^{-3}, this was not the case. At an FPR of 10^{-3}, our results completely agreed with the findings of Zarifzadeh et al. [14]. The small discrepancy at an FPR of 0 may be due to the computational constraints we faced with the hERG dataset, which was the largest in our study. Due to its large size, we had to use a small amount of samples Z from the underlying distribution to do the likelihood ratio test against for RMIA. This limitation arose because comparing all data points against many points Z across our models was computationally prohibitive. In contrast, LiRA does not have these constraints, which may explain its better performance compared to RMIA in this case. While RMIA generally outperforms LiRA, the latter remains a valuable approach, as it identifies different molecules, making it a complementary method, which we will discuss in a later paragraph.

Our results also show that membership inference attacks are most effective on smaller datasets. This is consistent with the findings of Shokri et al. [9], who link the success of attacks to the generalisability of the model and the diversity of the training data — both of which

improve with larger datasets. It is important to note that our neural networks are by no means designed in a way that makes them vulnerable to attack. On the contrary, we have implemented robust regularisation techniques that have been shown to make neural networks more resilient to membership inference attacks and improve privacy guarantees. In particular, our models use early stopping, dropout, and L2 weight regularisation. For the latter two, it has been specifically shown to reduce the efficiency of membership inference attacks [9, 25].

In practice, it could even be possible to increase the effectiveness of the attacks further by augmenting the attack query with some similar data as was shown by Zarifzadeh et al. [14]. We did not explore this due to computational limitations and given the broader scope of our study.

Another way to further increase the effectiveness of privacy attacks could be to incorporate scaffold-based inference strategies. Shifting from identifying complete molecular structures to detecting the presence or absence of specific molecular scaffolds within the training data could be an easier task that still provides information about sensitive intellectual property. Future research in this direction could potentially uncover additional privacy vulnerabilities in molecular property prediction.

We found that by applying multiple membership inference attacks, we were able to identify more molecules within the training data. This is consistent with previous work by Ye et al. [24], which demonstrated that some data points are only identified by certain attacks. We extended this by investigating the current state-of-the-art methods, LiRA and RMIA, and explicitly quantifying the overlap between these attacks across different datasets. From a practical point of view, using both attacks makes sense because it is possible to reuse the same shadow models between attacks, allowing more training data to be identified with limited computational overhead. In addition, the attacks remain feasible even when minimal computational resources are available. For example, RMIA has been shown to perform effectively with as few as two shadow models [14]. In such cases, the attacks can be run on any device capable of training neural networks with architectures similar to the target model.

Our finding that molecules in the minority class are more likely to be identified could be explained by the lower diversity in the training data for these compounds, as discussed above. This observation has important implications for drug discovery. In many datasets, the pharmacologically relevant compounds often belong to the minority class. For example, in high-throughput screening assays such as DNA-encoded library enrichment, researchers focus on the few molecules that bind to the target protein, while the majority that do not bind are of less interest [26]. This pattern is also seen in various cell-based screening assays, such as phenotypic assays aimed at identifying molecules that inhibit cancer cell proliferation [27]. In these scenarios, the minority class contains the compounds of greatest interest, making their identification far more valuable.

Our findings also indicate that molecules with low similarity to the rest of the training data are easier to identify, which has potential implications for drug discovery. Unique molecular structures that differ from established library compounds may correspond to proprietary lead compounds or novel scaffolds under development. Our results suggest that models might be more likely to memorize and subsequently reveal information about such structures. This observation aligns with previous research indicating that models tend to memorize outliers [28], which is suspected to contribute to the easier identification of these molecules [11].

To address these privacy concerns, we have developed a Python package to assess the privacy of training data for molecular property prediction.[1] This package allows users to evaluate their own data by applying our workflow to determine the extent to which training data can be identified when using different molecular representation methods. In addition, the package supports the testing of new representation methods by providing insight into their training data privacy and model performance on both user-provided and pre-supplied datasets. We hope that this tool will help researchers assess privacy risks before publishing their models.

Our research shows the potential dangers of information leakage from training data when publishing a trained neural network for drug discovery tasks. This risk exists even when the weights of the neural network are not published, and the model is offered as a supposedly safe web service. This has significant implications for organisations, which must constantly balance the need to make scientific discoveries openly available with the imperative to protect confidential data. We have shown that information leakage is consistently observed, but it can be mitigated by representing molecules as graphs and using message-passing neural networks, which also proved to be among the best performing models on our datasets. However, when planning to publish a model, it is crucial to consider not only performance but also the

[1] https://github.com/FabianKruger/molprivacy

privacy implications of different model architectures. Our findings also open up new research questions, such as how to adapt reconstruction attacks to the domain of molecules and how to develop models that are safer in terms of training data privacy in this field. The baseline for developing safer models might be to represent molecules as graphs and use message-passing neural networks for predictions. Our research highlights the essential balance between publicly available innovation and privacy, a balance that will impact the future of AI-driven drug discovery.

Methods

In this section, we first describe how we trained neural networks on biological datasets to predict molecular properties. Then, we outline the membership inference attacks used to evaluate the vulnerabilities of the models. Finally, we explain the methods used to compare and analyse the molecules leaked by these attacks. A high-level overview of our workflow is presented in Fig. 1. The code for our models and membership inference attacks, along with the datasets used in this study, are available on GitHub.

https://github.com/FabianKruger/molprivacy

Datasets

We used four different datasets to predict pharmacologically relevant molecular properties. The datasets differ in size, task, and class imbalance. The first dataset is used for mutagenicity prediction [18, 19]. It contains Ames test results for 7,255 drugs. Of these, 54% show positive results. The second dataset assesses blood-brain barrier permeability [17]. It contains 1,909 molecules, with 76% able to penetrate the barrier. The third dataset provides information on the inhibition of the potassium ion channel encoded by the human Ether-à-go-go-Related Gene (hERG) [21]. Inhibition is defined as a half-maximal inhibitory concentration of less than 10 μM. This dataset contains 306,341 compounds, with 4.5% being inhibitors. These three datasets were obtained from Therapeutics Data Commons [22]. The fourth dataset contains information on whether a molecule is enriched in a DNA-encoded library (DEL) for binding to carbonic anhydrase IX [20]. Positive enrichment is defined as the top 5% of enrichment scores. This dataset includes 108,528 molecules, with 4.9% showing enrichment after cleaning the data.

We pre-processed all datasets to remove ambiguities and incorrect compounds. Molecules were standardised for correct bonding, aromaticity, and hybridisation. Salts were removed to isolate the primary compound and Simplified Molecular Input Line Entry System (SMILES) [29] strings were converted to their canonical forms. Duplicate molecules and those with conflicting labels were removed. Molecules with canonical SMILES strings longer than 200 characters were also excluded. These steps were performed using the RDKit package version 2024.03.1. The reported dataset sizes are after cleaning. The cleaned datasets were randomly divided into a training set (45%), a validation set (10%), and a population subset (45%). The population subset was used for membership inference attacks, while the training and validation sets were used for model training and hyperparameter optimisation.

Model architectures

To capture the variety in molecular representation approaches, we trained neural networks on a range of commonly used representations. Our study included extended-connectivity fingerprints (ECFPs) [30], molecular access system (MACCS) keys [31], graph representations, RDKit fingerprints (RDKitFPs) [32], and SMILES [29] representations. We chose these representations to cover various conceptually different approaches to molecular representation. For ECFPs, we investigated fingerprints with radii of 2 and 3, both mapped to 2048-bit vectors. MACCS keys were represented as binary vectors, indicating the presence or absence of 166 structural patterns. RDKitFPs identified all subgraphs in the molecule up to a length of 7, hashed into 2048-bit vectors. These three representations were generated using RDKit [32]. The graph representation was generated using Chemprop version 1.6.1 [33].

The type of neural network we used varied depending on the specific molecular representation. We used multi-layer perceptrons (MLPs) for ECFPs, MACCS keys, and RDKitFPs. We employed message passing neural networks implemented in Chemprop for the graph representation. For the SMILES representation, we used a pre-trained transformer encoder combined with a convolutional neural network based on Karpov et al. [34]. All our models were implemented in Pytorch version 2.2.2 [35]. We pre-trained the transformer encoder to convert non-canonical SMILES strings to their canonical counterparts for 20 epochs using the ChEMBL_V29

dataset from Therapeutics Data Commons [36]. We randomly split this dataset into 90% training data and 10% validation data. For the transformer encoder, we used the same hyperparameters as in the original publication but increased the context length of the transformer from 110 to 202 tokens in order to also generate encodings for larger molecules. We determined the hyperparameters for the MLPs, message passing neural networks, and convolutional neural networks using a Bayesian optimisation method, which we will describe in the next paragraph. All our models had one output node to predict the logits for our binary classification problems.

Hyperparameter optimization

To avoid introducing subjective bias into our models, we decided to automatically optimise the hyperparameters of the neural networks using a tree structured Parzen estimator [16]. This was done using Optuna version 3.6.0. [37]. We optimised dropout rate, number and dimension of hidden layers, learning rate, and weight decay for MLPs. For message passing neural networks, we optimised message passing steps, dropout, encoder hidden dimension, bias addition in the encoder, aggregation function, number and dimension of classifier hidden layers, learning rate, and weight decay. For convolutional neural networks, we kept the filter sizes from the original publication and optimised dropout, learning rate, and weight decay. Detailed ranges for the hyperparameter search spaces are shown in Supplementary Table 1. We optimised each neural network architecture for three hours on an NVIDIA Volta V100 GPU. During this time, we evaluated the validation cross-entropy loss for different hyperparameter combinations. Each training run was performed for a maximum of 20 epochs. We stopped runs early if the validation loss did not improve for two consecutive epochs or if, after 15 epochs, the validation loss was below the median value for that epoch.

Model training

After finding the optimised hyperparameters, we trained the final models until their performance converged on the validation set. We used early stopping with a patience of 10 epochs and saved the model weight of the epoch with the lowest validation loss. For all our models, we used a weighted binary cross-entropy loss as a loss function. The weights accounted for the class imbalance and were inversely proportional to the frequency of the classes. We used the adaptive moment estimation with decoupled weight decay regularization (AdamW) optimiser for MLPs and message passing neural networks [38]. For convolutional neural networks, we used the original adaptive moment estimation (Adam) optimiser to remain consistent with the original implementation [39]. Training was done in batch sizes of 64 samples. We repeated our experiment 20 times for each dataset and representation to capture the marginal distribution of all randomness in the experiment, including dataset splitting, hyperparameter optimisation, and model weight initialisation. We examined the performance of each model on the population sample dataset (Fig. 1), as it was not used in any way for training or hyperparameter optimisation, and testing the performance of the model is independent of the membership inference attacks.

Membership inference attacks

To determine whether an adversary can discriminate between molecules that are in the training data and those that are not, we applied two state-of-the-art membership inference attacks: likelihood ratio attacks (LiRA) [11] and robust membership inference attacks (RMIA) [14]. Both methods assign a score to each sample, indicating the confidence that it was part of the training dataset. LiRA performs a likelihood ratio test by comparing the likelihood of the model output when the sample is included in the training dataset against when it is not (Algorithm 2). To approximate these likelihoods, so-called shadow models are trained on data from the same distribution. Some shadow models include the target sample in their training data, while others do not. We used random subsets containing 50% of our population subset dataset to train 10 shadow models for each target model. Each target model training data sample was included in some shadow models and excluded from others. The shadow models had the same hyperparameters as the target model and were trained for 15 epochs. For each target sample, two Gaussian distributions of the rescaled output logits are modelled: one for shadow models that included the target sample in their training data and one for those that did not. The likelihood of observing the rescaled output logits of the target model is then calculated for each distribution. The ratio between these likelihoods represents the likelihood ratio that the target sample was in the training data. For more details on LiRA, we refer the reader to the original publication [11].

Algorithm 2 Likelihood Ratio Attack (LiRA) tests whether a specific target data point *m* - in our case, a molecular structure *x* with the corresponding label *y* - was part of the training data for a target neural network model f_θ. In this attack, shadow models $s_i,\ i = 1, \ldots, N$ are trained on data drawn from a distribution similar to that of f_θ's training data (in our case, a similar chemical space). Some shadow models include *m* in their training data, while others do not. The re-scaled confidence of each shadow model when predicting *m* is then calculated. These confidences are modeled as two Gaussian distributions: one for the shadow models that included *m*, and one for those that did not. Finally, we determine whether the confidence of the target model f_θ is more likely to belong to the distribution of models that included *m* or the distribution of models that did not. The likelihood ratio between these distributions, combined with a decision threshold *t*, determines whether *m* is predicted to have been part of f_θ's training data.

Input: Target model f_θ, Target data point $m = (x, y)$, Data distribution Π, Number of shadow models N, Decision threshold t

Define re-scaling function: $\Phi(p) = \log\left(\frac{p}{1-p}\right)$
$O_{in} \leftarrow \emptyset$
$O_{out} \leftarrow \emptyset$
for $i = 1$ to N **do**
    Sample dataset from Π: $D_i \sim \Pi^n$
    Flip a fair coin: $c_i \sim \{0, 1\}$
    if $c_i = 1$ **then**
        Include m in D_i
    else
        Exclude m from D_i
    end if
    Train shadow model s_i on D_i
    Calculate re-scaled shadow model confidence for m: $o_i \leftarrow \Phi(s_i(x)_y)$
    if $c_i = 1$ **then**
        $O_{in} \leftarrow O_{in} \cup \{o_i\}$
    else
        $O_{out} \leftarrow O_{out} \cup \{o_i\}$
    end if
end for

} **Train Shadow Models**

Calculate μ_{in}, σ^2_{in} from O_{in}
Calculate $\mu_{out}, \sigma^2_{out}$ from O_{out}

} **Fit Gaussian Distributions**

Calculate re-scaled target model confidence for m: $o_{target} \leftarrow \Phi(f_\theta(x)_y)$
Calculate likelihoods:
    $L_{in} \leftarrow N(o_{target} \mid \mu_{in}, \sigma^2_{in})$
    $L_{out} \leftarrow N(o_{target} \mid \mu_{out}, \sigma^2_{out})$
Calculate likelihood ratio: $LR \leftarrow \frac{L_{in}}{L_{out}}$

} **Calculate Likelihood Ratio**

if $LR > t$ **then**
    Predict that data point m was in f_θ's training data
else
    Predict that data point m was not in f_θ's training data
end if

} **Predict Membership**

RMIA compares the likelihood ratio of observing the target model f_θ after applying the training algorithm T with two different conditions: first, when the target sample m is included in the training dataset D, and second, when a random, different, sample z is included instead. This process is repeated with many different random samples. RMIA then attempts to calculate the probability that these likelihood ratios exceed a threshold gamma

$$\text{Score}(m, f_\theta) \approx P_{z\sim\Pi}\left(\frac{P(F_\Theta = f_\theta | f_\theta = T(D \cup \{m\})}{P(F_\Theta = f_\theta | f_\theta = T(D \cup \{z\})} \geq \gamma\right).$$

In our experiments, we chose a gamma value of 2. It was shown that the attack is robust to different values for gamma [14]. Each likelihood of the ratio is calculated using Bayes' rule (for brevity we abbreviate the conditions on both probabilities with *m* and *z* here)

$$\frac{P(f_\theta|m)}{P(f_\theta|z)} = \left(\frac{P(m|f_\theta)}{P(m)}\right) \cdot \left(\frac{P(z|f_\theta)}{P(z)}\right)^{-1}.$$

The probability $P(m|f_\theta)$ is approximated by the probability of the correct class prediction and the probability $P(m)$ is approximated as the empirical mean of this over

all shadow models (Algorithm 3). The probabilities for the random points Z are computed similarly. The complete implementation of this attack is shown in Algorithm 3. For more details on RMIA, we refer readers to the original publication [14]. For this attack, we reused the shadow models from LiRA and used the 50% of the population sample dataset not included in their training as random sample points Z in the attack. We based our implementation of LiRA and RMIA on the implementation in the LeakPro repository of AI Sweden.[2]

We evaluated the success of our attacks by determining the true positive rates (TPRs) for identifying training data molecules at different false positive rates (FPRs). We focused our evaluation on low FPRs, as was recommended by Carlini et. al [11] and is discussed in their paper in more detail. In both of our attacks, we give the adversary a training data sample with a probability of 0.67 and a non-training data sample with a probability of 0.33. The reason for this is that we did not want the training datasets for the models to become too small, while

Algorithm 3 Robust Membership Inference Attack (RMIA) tests whether a specific target data point *m*-in our case, a molecular structure *x* with the corresponding label *y*-was part of the training data for a target neural network model f_θ. In this attack, shadow models s_i, $i = 1, \ldots, N$ are trained on data drawn from a distribution similar to that of f_θ's training data (in our case, a similar chemical space). Some shadow models include *m* in their training data, while others do not. The probability of *m* is approximated by averaging the correct class probability over all shadow models. Similarly, the probability of *m* given f_θ is approximated as the probability of the correct class assignment by model f_θ. The ratio between these probabilities is then calculated and compared to the ratios obtained for other points *z*. The final score is the proportion of points *z* for which the ratio is at least γ times higher for data point *m*. This score, combined with a decision threshold *t*, determines whether *m* is predicted to have been part of f_θ's training data.

Input: Target model f_θ, Target data point $m = (x, y)$, Data distribution Π, Samples from data distribution $Z \sim \Pi^k$, Number of shadow models N, Likelihood ratio threshold γ, Decision threshold t

for $i = 1$ to N **do**
  Sample dataset D_i from Π such that D_i and Z are disjoint: $D_i \sim \Pi^n(\cdot \mid D \cap Z = \emptyset)$
  Flip a fair coin: $c_i \sim \{0, 1\}$
  if $c_i = 1$ **then**
    Include m in D_i
  else
    Exclude m from D_i
  end if
  Train shadow model s_i on D_i
end for

} **Train Shadow Models**

$P(m) \approx \frac{1}{N} \sum_i s_i(x)_y$
$P(m|f_\theta) \approx f_\theta(x)_y$
$\text{Ratio}_m \leftarrow \frac{P(m|f_\theta)}{P(m)}$

} **Approximate Probabilities**

Counter $C \leftarrow 0$
for each $z = (x', y') \in Z$ **do**
  $P(z) \approx \frac{\sum_{i=1}^{N} \mathbb{I}[c_i=0] \cdot s_i(x')_{y'}}{\sum_{i=1}^{N} \mathbb{I}[c_i=0]}$
  $P(z|f_\theta) \approx f_\theta(x')_{y'}$
  $\text{Ratio}_z \leftarrow \frac{P(z|f_\theta)}{P(z)}$
  if $\frac{\text{Ratio}_m}{\text{Ratio}_z} \geq \gamma$ **then**
    $C \leftarrow C + 1$
  end if
end for

} **Calculate Individual Likelihood Ratios**

$\text{Score}(m, f_\theta) \leftarrow \frac{C}{|Z|}$

} **Calculate Final Score**

if Score $> t$ **then**
  Predict that data point m was in f_θ's training data
else
  Predict that data point m was not in f_θ's training data
end if

} **Predict Membership**

[2] https://github.com/aidotse/LeakPro

still using all the data points for the attack. This approach allowed us to use 45% of the dataset size as training data for the target model. With a membership probability of 0.67, the baseline TPR at an FPR of 0 is $\frac{2}{N}$, where N is the size of the training dataset. A detailed derivation of this baseline is provided in the Supplementary information. To determine if the attacks leak training data information, we compared the TPRs of our attacks to the baseline TPR of $\frac{2}{N}$. We tested for significance using Wilcoxon signed-rank tests over the 20 repetitions of each experiment. We repeated this experiment with the TPR at an FPR of 10^{-3} to see if we observe similar trends. We also investigated the ROC curves for identifying training data molecules to see the trends at all possible FPRs.

Leaked molecule analysis

We investigated whether our two membership inference attacks identify the same molecules or can be used complementarily to gain more information about the training data. To do this, we analysed the overlap between the identified molecules from each attack. In our setting, we have the training dataset Ω, from which we identify two subsets, $A \subseteq \Omega$ and $B \subseteq \Omega$, each corresponding to one attack. These subsets can have different sizes and can overlap. We define the percentage of the maximum possible overlap as

$$f(A,B) = \frac{|A \cap B|}{\min(|A|, |B|)}.$$

This scalar value ranges from 0 to 1, where 1 indicates that the larger subset contains all molecules of the smaller subset. We examined the percentage of maximum possible overlap between the two membership inference attacks for every dataset and representation. For each combination, we plotted the distribution of the 20 experiment repetitions. To determine whether the overlap is significantly different from what would occur by chance when drawing two uncorrelated subsets, we calculate the difference between the expected overlap by chance and the observed overlap. This is done for our 20 experiment repetitions. We use a Wilcoxon signed-rank test to assess if the difference between the observed and random overlap is significantly different from 0. The overlap by chance can be thought of as a random variable following a hypergeometric distribution, because when we independently draw the smaller subset, we draw without replacement from the training dataset Ω, which contains the larger subset as possible overlap successes

$$|A \cap B| \sim \text{Hypergeometric}(N = |\Omega|, K = \max(|A|, |B|), n = \min(|A|, |B|)).$$

The mean of the hypergeometric distribution is defined as $\mathbb{E}[|A \cap B|] = \frac{nK}{N}$, which is the overlap that is expected to be observed by chance. We also calculated the overlap between the identified molecules from models trained on different molecular representations of the same training data.

In addition, we investigated characteristics of the molecules that could be identified. To do this, we compared the distributions of property labels of identified molecules with the underlying property label distribution. For each of the 20 experiment repetitions, we calculated the percentage of positive compounds for both identified and not identified molecules. We then used Mann–Whitney U tests to analyse whether the two distributions differed significantly. We also calculated the TPRs of the minority class. We did this similarly as before but only considered the training data molecules of the minority class in this case. We also assessed whether the identified molecules differed in size compared to the rest of the training dataset. To do this, we calculated the number of atoms in each molecule and pooled the amounts across all 20 experiment repetitions for both identified and not identified molecules. We compared the distributions of molecule sizes and determined significance using Mann–Whitney U tests.

To investigate whether molecules with low similarity to the rest of the training data are easier to identify, we computed pairwise Tanimoto similarity scores for all molecules in the training set. Specifically, we generated ECFP fingerprints with a radius of 2 and a size of 2048, using RDKit for both fingerprint computation and Tanimoto similarity calculations [32]. For each molecule, we determined the highest similarity to any other molecule in the training set (nearest neighbour similarity), as well as the average similarity across all training molecules. This approach allowed us to account for individual outliers as well as local clusters distant from the majority of the training data. To analyse the highest similarity values, we divided the range from 0 to 1 into ten equal-width bins, each covering a 0.1 increment, and then calculated the fraction of molecules identified at an FPR of 0 within each bin. In contrast, for average similarity, where values were more closely distributed, we used quantile binning to divide the data into deciles, ensuring that each bin contained an equal number of samples. We then calculated the fraction of molecules identified at an FPR of 0 within each bin. In addition, we performed one-tailed Mann–Whitney U tests to determine whether the similarity values of identified molecules were significantly lower than those of non-identified molecules, considering both the highest and average similarity metrics. Our analysis was performed on all molecular representations for the BBB and Ames datasets. However, due to computational

constraints, we were unable to perform the same analysis for the DEL and hERG datasets, as calculating all pairwise similarity values would have been infeasible due to the exponential increase in computational complexity with dataset size.

Abbreviations

AI	Artificial intelligence
LiRA	Likelihood ratio attack
RMIA	Robust membership inference attack
BBB	Blood-Brain barrier
DEL	DNA-Encoded library
hERG	Human ether-à-go-go-related gene
AUROC	Area under the receiver operating characteristic curve
TPR	True positive rate
FPR	False positive rate
SMILES	Simplified molecular input line entry system
ECFP	Extended-connectivity fingerprints
MACCS	Molecular access system
RDKitFP	RDKit fingerprints
MLP	Multi-Layer perceptron
Adam	Adaptive moment estimation
AdamW	Adaptive moment estimation with decoupled weight decay regularization

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-00982-w.

Additional file

Supplementary file 1 (PDF 413 KB)

Acknowledgements
Not applicable

Author contributions
F.K. conducted the primary research, including conceptualization, experimentation, and analysis, and drafted the manuscript. J.Ö., L.M., I.T., and O.E. contributed to the conceptualization of the study and provided critical feedback on the manuscript. J.Ö. additionally supplied code for the membership inference attacks, while L.M. provided data for the study.

Funding
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.

Data availability
All data and code are hosted on GitHub at https://github.com/FabianKruger/molprivacy and can be installed via pip.

Declarations

Competing interests
The authors declare no competing interests.

Received: 10 December 2024 Accepted: 4 March 2025
Published online: 26 March 2025

References

1. Chen Hongming, Engkvist Ola, Wang Yinhai, Olivecrona Marcus, Blaschke Thomas (2018) The rise of deep learning in drug discovery. Drug Dis Today 23(6):1241–1250
2. Muratov Eugene N, Bajorath Jürgen, Sheridan Robert P, Tetko Igor V, Filimonov Dmitry, Poroikov Vladimir, Oprea Tudor I, Baskin Igor I, Varnek Alexandre, Roitberg Adrian et al (2020) Qsar without borders. Cheml Soc Rev 49(11):3525–3564
3. Dara Suresh, Dhamercherla Swetha, Jadav Surender Singh, Madhu Babu CH, Ahsan Mohamed Jawed (2022) Machine learning in drug discovery: a review. Artif Intell Rev 55(3):1947–1999
4. Vamathevan Jessica, Clark Dominic, Czodrowski Paul, Dunham Ian, Ferran Edgardo, Lee George, Li Bin, Madabhushi Anant, Shah Parantu, Spitzer Michaela et al (2019) Applications of machine learning in drug discovery and development. Nat Rev Drug Dis 18(6):463–477
5. Oldenhof Martijn, Ács Gergely, Pejó Balázs, Schuffenhauer Ansgar, Holway Nicholas, Sturm Noé, Dieckmann Arne, Fortmeier Oliver, Boniface Eric, Mayer Clément et al (2023) Industry-scale orchestrated federated learning for drug discovery. Proc AAAI Conf Artif Intell 37:15576–15584
6. Zuckerberg Mark (2024) 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: 25-09-2025
7. Yash Raj Shrestha (2023) Georg von Krogh, and Stefan Feuerriegel Building open-source ai. Nat Comput Sci 3(11):908–911
8. Murdoch Blake (2021) Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 22:1–5
9. Shokri Reza, Stronati Marco, Song Congzheng, Shmatikov Vitaly (2017) Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE
10. Murakonda Sasi Kumar, Shokri Reza (2020) Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. arXiv preprint arXiv:2007.09339
11. Carlini Nicholas, Chien Steve, Nasr Milad, Song Shuang, Terzis Andreas, Tramer Florian (2022) Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE
12. Salem Ahmed, Cherubin Giovanni, Evans David, Köpf Boris, Paverd Andrew, Suri Anshuman, Tople Shruti, Zanella-Béguelin Santiago (2023) Sok: Let the privacy games begin! a unified treatment of data inference privacy in machine learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 327–345. IEEE
13. Hongsheng Hu, Zoran Salcic, Lichao Sun, Dobbie Gillian Yu, Philip S, Xuyun Zhang (2022) Membership inference attacks on machine learning: A survey. ACM Comput Surv (CSUR) 54(11):1–37
14. Zarifzadeh Sajjad, Liu Philippe, Shokri Reza (2024) Low-cost high-power membership inference attacks. In Forty-first International Conference on Machine Learning
15. Pejo Balazs, Remeli Mina, Arany Adam, Galtier Mathieu, Acs Gergely (2022) Collaborative drug discovery: Inference-level data protection perspective. arXiv preprint arXiv:2205.06506
16. Bergstra James, Bardenet Rémi, Bengio Yoshua, Kégl Balázs (2011) Algorithms for hyper-parameter optimization. Advances in neural information processing systems 24
17. Martins Ines Filipa, Teixeira Ana L, Pinheiro Luis, Falcao Andre O (2012) A bayesian approach to in silico blood-brain barrier penetration modeling. J Chem Inf Model 52(6):1686–1697
18. Hansen Katja, Mika Sebastian, Schroeter Timon, Sutter Andreas, Ter Laak Antonius, Steger-Hartmann Thomas, Heinrich Nikolaus, Muller Klaus-Robert (2009) Benchmark data set for in silico prediction of ames mutagenicity. J Chem Inf Model 49(9):2077–2081
19. Congying Xu, Cheng Feixiong, Chen Lei, Zheng Du, Li Weihua, Liu Guixia, Lee Philip W, Tang Yun (2012) In silico prediction of chemical ames mutagenicity. J Chem Inf Model 52(11):2840–2847
20. Lim Katherine S, Reidenbach Andrew G, Hua Bruce K, Mason Jeremy W, Gerry Christopher J, Clemons Paul A, Coley Connor W (2022) Machine learning on dna-encoded library count data using an uncertainty-aware probabilistic loss function. J Chem Inf Model 62(10):2316–2331
21. Fang Du, Haibo Yu, Zou Beiyan, Babcock Joseph, Long Shunyou, Li Min (2011) 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 Drug Dev Technol 9(6):580–588
22. Huang Kexin, Tianfan Fu, Gao Wenhao, Zhao Yue, Roohani Yusuf, Leskovec Jure, Coley Connor W, Xiao Cao, Sun Jimeng, Zitnik Marinka (2021) Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. arXiv preprint arXiv:2102.09548
23. Wu Bang,Yang Xiangwen, Pan Shirui, Yuan Xingliang (2021) Adapting membership inference attacks to gnn for graph classification: Approaches and implications. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1421–1426. IEEE
24. Ye Jiayuan, Maddi Aadyaa, Murakonda Sasi Kumar, Bindschaedler Vincent, Shokri Reza (2022) Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3093–3106,
25. Jain Prateek, Kulkarni Vivek, Thakurta Abhradeep, Williams Oliver (2015) To drop or not to drop: Robustness, consistency and differential privacy properties of dropout. arXiv preprint arXiv:1503.02031
26. Satz Alexander L, Brunschweiger Andreas, Flanagan Mark E, Gloger Andreas, Hansen Nils JV, Kuai Letian, Kunig Verena BK, Xiaojie Lu, Madsen Daniel, Marcaurelle Lisa A et al (2022) DNA-encoded chemical libraries. Nature Rev Methods Primers 2(1):3
27. Wei Zheng, Natasha Thorne, McKew John C (2013) Phenotypic screens as a renewed approach for drug discovery. Drug Disc Today 18(21):1067–1073
28. Feldman Vitaly, Zhang Chiyuan (2020) What neural networks memorize and why: discovering the long tail via influence estimation. Adv Neural Inf Process Syst 33:2881–2891
29. David Weininger (1988) Smiles, a chemical language and information system 1 introduction to methodology and encoding rules. J Chem Inf Comput Sci 28(1):31–36
30. Rogers David, Hahn Mathew (2010) Extended-connectivity fingerprints. J Chem Inf Model 50(5):742–754
31. Durant Joseph L, Leland Burton A, Henry Douglas R, Nourse James G (2002) Reoptimization of mdl keys for use in drug discovery. J Chem Inf Comput Sci 42(6):1273–1280
32. Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello et al (2024) rdkit/rdkit: 2024_09_1 (q3 2024) release. https://www.rdkit.org
33. Yang Kevin, Swanson Kyle, Jin Wengong, Coley Connor, Eiden Philipp, Gao Hua, Guzman-Perez Angel, Hopper Timothy, Kelley Brian, Mathea Miriam et al (2019) Analyzing learned molecular representations for property prediction. J Chem Inf Model 59(8):3370–3388
34. Karpov Pavel, Godin Guillaume, Tetko Igor V (2020) Swiss knife for QSAR modeling and interpretation Transformer-CNN. J Chem 12:1–12
35. Paszke Adam, Gross Sam, Massa Francisco, Lerer Adam, Bradbury James, Chanan Gregory, Killeen Trevor, Lin Zeming, Gimelshein Natalia, Antiga Luca et al (2019) Pytorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32:87654
36. Barbara Zdrazil, Eloy Felix, Fiona Hunter, Manners Emma J, James Blackshaw, Sybilla Corbett, de Veij Marleen, Ioannidis Harris, Lopez David Mendez, Mosquera Juan F, et al (2024) The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res 52(1):1180-D1192
37. Akiba Takuya, Sano Shotaro, Yanase Toshihiko, Ohta Takeru, Koyama Masanori (2019) Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631
38. Loshchilov Ilya, Hutter Frank, et al (2017) Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101, 5
39. Kingma Diederik P (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

Publisher's Note