From Private Data to Shared Knowledge

T. Hanser[1], L. Johnston[1], J. Marchaland[1], J. Plante[1], R. van Deursen[2], S. Werner[1], and R. Williams[1].

1 Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS
2 Firmenich, Geneva, Switzerland.

Introduction

Artificial Intelligence (AI) has become a powerful research catalyst in science. At the core of modern AI is the ability to automatically extract knowledge from data and build accurate predictive models. To maximize this impact, it is critical to have access to enough good quality data to allow machine learning algorithms to extract relevant knowledge and produce useful models. One of the main challenges in AI is therefore to compile such pivotal datasets, which is particularly difficult in drug discovery due to the confidential nature of the primary information: the chemical structure. Even with the availability of public data, the most valuable knowledge remains embedded and locked in private silos despite the willingness of industry to share non-competitive information (fig.1)

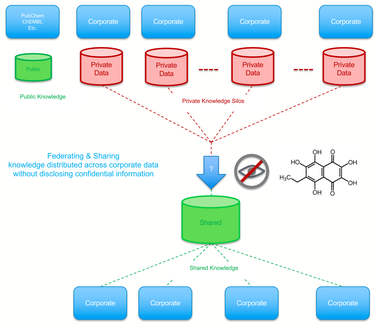

Figure 1: How can we unlock the knowledge embedded in private data silos ?

Methodology

To overcome this obstacle, Lhasa Limited has developed a methodology to enable the transfer of knowledge from corporate data into sharable models whilst preserving the privacy of the original data. The method uses Knowledge Distillation[1] based on the Teacher-Student approach [2,3] adapted to the domain of Molecular Informatics. In this methodology, a private teacher model is trained from the proprietary data and used to label public data (name Cronos data). This public data is subsequently used to train a student model. The private structural information is therefore decoupled from the final student model and is never disclosed to the end user. This new method enables knowledge sharing without a privacy leak. Furthermore we can federate the knowledge from different private teachers into a single shared student model by combining the labels of these teachers for the same structures (fig. 2).

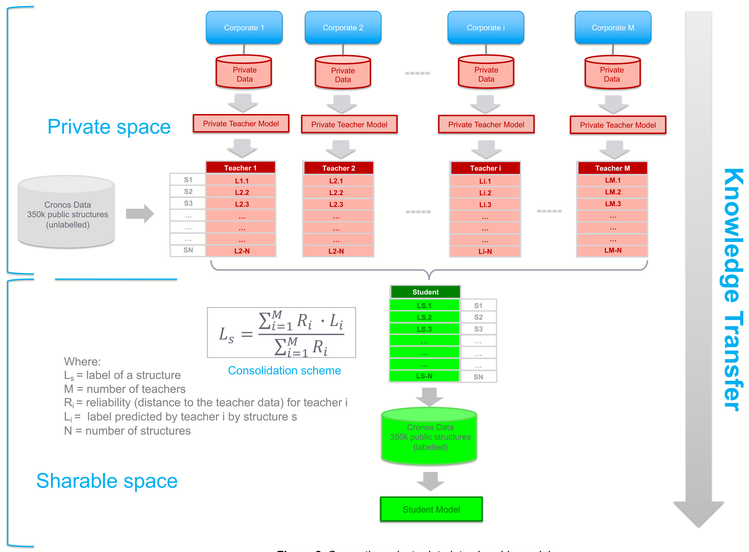

Figure 2: Converting private data into sharable models

We expect the student model to perform better than an average teacher. The underlying hypothesis is that a student that has collected knowledge from several teachers becomes itself more knowledgeable than an average teacher owing to the diversity and complementarity of the information provided by its respective teachers. We have explored and validated this hypothesis in collaboration with a consortium of 8 large pharmaceutical companies. We transferred the knowledge for the hERG endpoint to build a sharable classification model. Models were built using the SOHN algorithm[4] and the Extended Sybyl Atom pair descriptors[5]

Each member of the consortium provided a private dataset for hERG (sizes ranging from 3k to 70k data points) from which we built 8 private teacher models used to label the Cronos data (public structures). For a same structure the labels are consolidated as a weighted average of the labels (class probability distributions) using the distance between, the structure and the teacher private training data (reliability) as a weighting factor (fig. 2).

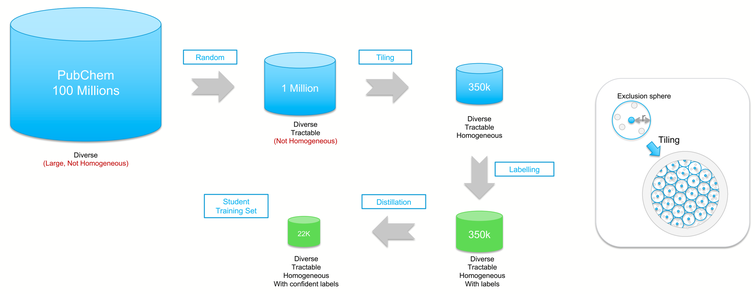

Figure 3: Preparation of the Cronos public structure dataset

Choosing the Cronos set of public structures is an important step as it defines the vessel for the knowledge transfer process. We aimed to design a tractable, diverse and homogeneous representation of a wide chemical space. For that purpose, we randomly sampled 1 million structures from PubChem and applied a tiling process (exclusion sphere based on a Tanimoto similarity) in order to obtain a set of 350k structures. These structures were subsequently labelled using the consolidation across the 8 teacher labels. Finally only the subset of the most confident consolidated labels for each class were retained, producing a final perfectly balanced student training set of 22k structures (fig.3).

Results

To validate and proof the transferability concept we compared the performance of the teachers with the performance of the student using an external benchmark provided by Preissner et al.[6]. The benchmark compiles about 4.5k structures from the public domain. We used MCC (Matthew Correlation Coefficient) as an objective metric for the comparison. In figure 4 and table 1 we can see that the student performs better than the average teacher, validating the hypothesis. Even more promising, the student also outperforms every individual teacher, indicating a positive synergy when combining the teachers. The student model also outperforms the best ranked model (Random-Forest + ECFP4 descriptor) in the Preissner benchmark further demonstrating the potential of this approach (fig.5).

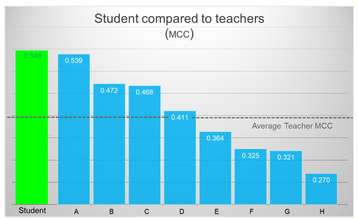

Figure 4: The student model outperforms the average teacher and even outperform every individual the teacher models.

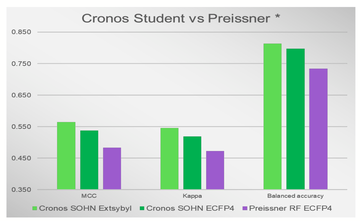

Figure 5: The student model outperforms the best model in the Preissner benchmark even using the same algorithms or descriptors

	A	F	D	H	E	C	G	B	<Teacher>	Student	
MCC	0.539	0.325	0.411	0.270	0.364	0.468	0.321	0.472	0.396	0.546	MCC
Kappa	0.529	0.309	0.389	0.256	0.339	0.468	0.255	0.457	0.375	0.499	Kappa
BAcc	0.793	0.634	0.731	0.610	0.643	0.737	0.683	0.704	0.692	0.813	BAcc
Recall +	0.772	0.353	0.841	0.310	0.350	0.615	0.854	0.473	0.571	0.911	Recall +
Recall -	0.814	0.916	0.722	0.910	0.937	0.858	0.511	0.934	0.825	0.714	Recall -
Precis +	0.582	0.584	0.472	0.535	0.650	0.593	0.370	0.707	0.562	0.517	Precis +
Precis -	0.914	0.808	0.893	0.797	0.811	0.869	0.913	0.841	0.856	0.960	Precis -
Coverage	0.587	0.392	0.454	0.328	0.200	0.442	0.307	0.706	0.427	0.774	Coverage

Table 1: Detailed comparison of the performance of the teachers and the student

These encouraging results show the potential of a new generation of performant predictive models leveraging the privacy preserving access to a huge pool of knowledge distributed across pharmaceutical companies. The potential of these models is further increased by the federation of many knowledge sources with a positive synergistic effect.

To explore the impact of the number of teachers on the performance of the student, we ran the experiment with an incremental number of teachers (stochastically sampled on the full labelling matrix to avoid permutation bias). We can see in figure 6 that after just 3 teachers the student outperforms the average teacher with an MCC > 0.4. Beyond 4 teachers, the student performance begins to plateau.

To make sure that the performance of the student is not due to a single teacher, we ran the experiment leaving one teacher out at each round. Figure 7 shows that removal of any given teacher does not substantially impact the performance of the student, demonstrating that the model is truly collaborative.

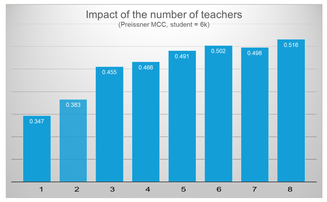

Figure 6: With 3 or more teachers, the student model already outperforms the average teacher (MCC=0.4)

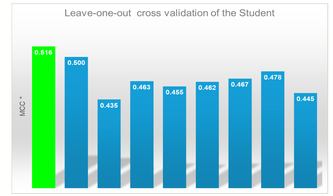

Figure 7: The model is truly collaborative as no teacher demonstrates a substantial drop when removed

Finally we wanted to test the potential of this approach in a prospective context. For this purpose we combined the Cronos labelled data with curated public data and private data from three of the members of the consortium. From these three sources of data we built a model called 'hybrid model' that represents a model with the highest predictive potential. We then compared the performance of this model with a model built only with the private data against a prospective test set composed of structures developed since the Cronos consortium data were collected (time split).

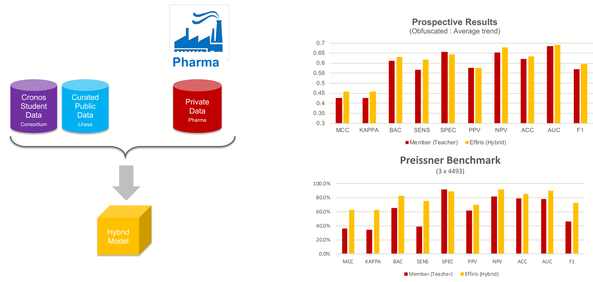

Figure 8: Prospective validation experiment the hybrid model outperforms the private model

The hybrid model outperforms the private model (fig. 8) in the prospective validation with a substantial gain in sensitivity. The hybrid model performs much better than the private model against the Preissner benchmark.

Conclusion

Using the teacher-student approach, Lhasa Limited has introduced a new privacy-preserving method to transfer knowledge from private data into sharable models. We have validate the concept in collaboration with 8 large pharmaceutical companies. The resulting student model outperforms private teacher models it learned from, indicating a positive synergistic effect. The student model also outperforms the best ranked model in the Preissner benchmark context. These encouraging results open an avenue for a new generation of highly performant models in the domain of drug discovery and development. The approach is however domain agnostic and can be applied in many other contexts. A first application of this methodology drives the knowledge transfer for secondary pharmacology endpoints within the Effiris consortium [7]

References

[1] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
[2] Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., & Talwar, K. (2016). Semi-supervised knowledge transfer for deep learning from private training data. arXiv preprint arXiv:1610.05755.
[3] Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., & Erlingsson, Ú. (2018). Scalable private learning with pate. arXiv preprint arXiv:1802.08908.
[4] Hanser, T., Barber, C., Rosser, E., Vessey, J. D., Webb, S. J., & Werner, S. (2014). Self organising hypothesis networks: a new approach for representing and structuring SAR knowledge. Journal of cheminformatics, 6(1), 21.
[5] Hanser, T., Steinmetz, F. P., Plante, J., Rippmann, F., & Krier, M. (2019). Avoiding hERG-liability in drug design via synergetic combinations of different (Q) SAR methodologies and data sources: a case study in an industrial setting. Journal of cheminformatics, 11(1), 9.
[6] Siramshetty, V. B., Chen, Q., Devarakonda, P., & Preissner, R. (2018). The Catch-22 of predicting hERG blockade using publicly accessible bioactivity data. Journal of chemical information and modeling, 58(6), 1224-1233.
[7] https://www.lhasalimited.org/products/Effiris.htm