Developing Federated QSAR Models for Secondary Pharmacology

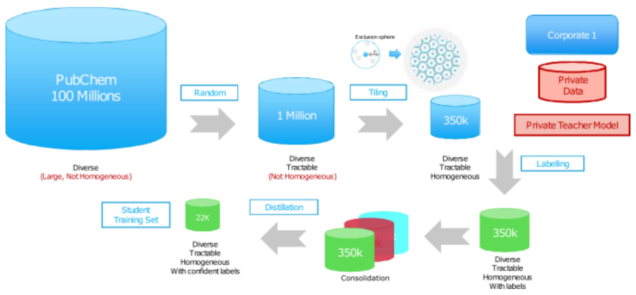
Thierry Hanser, Jean-Francois Marchaland, Jeffrey Plante, Stephane Werner

Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS

Introduction

One difficult problem in the development of pharmaceutical compounds is the tendency for compounds to have secondary pharmacological, or off target, effects. This occurs when a compound has enough affinity for an unintended target to cause a disruption, thereby leading to a toxicological concern. A paper by Bowes[1] has detailed a number of secondary pharmacological targets and their associated toxicity. Modelling these complex endpoints is difficult, because much of the data resides within individual pharmaceutical companies and they are generally unable to disclose their proprietary information and potentially lose their competitive advantage. At Lhasa Limited, we have developed a means of extracting the toxicity knowledge held within the company to enable collaboration between companies without disclosing any private information. This allows companies to build more effective *in silico* models for the prediction of these off target effects and thus highlight compounds that could have a toxic liability further down the line, thereby speeding up the pace of pharmaceutical research

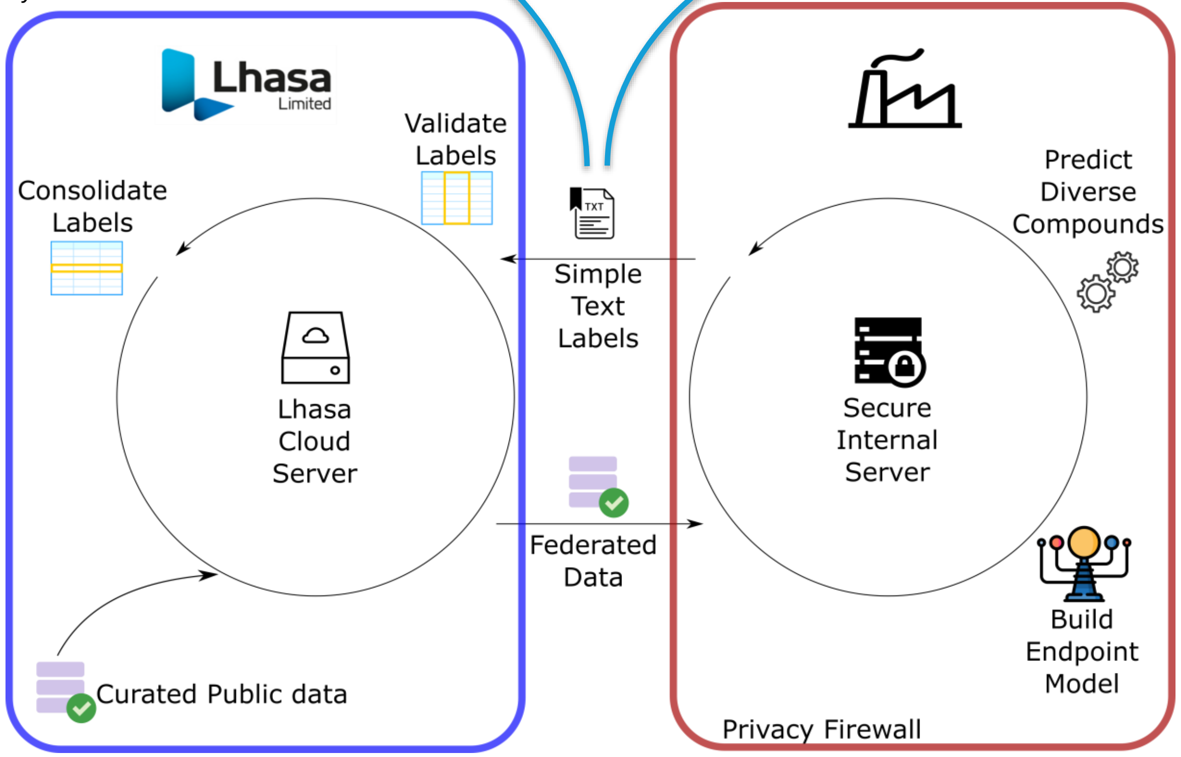
Methodology



We have generated a diverse dataset that covers all known chemical space by randomly sampling PubChem and then using a tiling approach to ensure maximum structural diversity. This dataset serves as the substrate for the private transfer of knowledge from inside the private space of the company to a public space consisting solely of public structures. These compounds are labelled using the private model and these labels are collected and consolidated using a weighted averaging approach to generate an overall label for each compound. The 350k compound dataset is then subsampled to select only those compounds where the confidence in the label is the highest, as those compounds contain the most knowledge. The consolidated dataset is returned to each company allowing them to build a hybrid model using their chemical space, along with the knowledge across all companies.

A simple text file securely transfers the knowledge

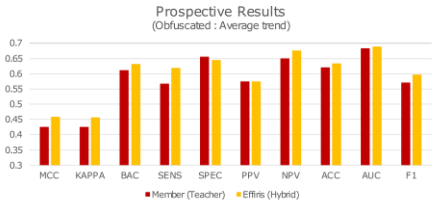
```
244466→[D=0.89, 0.11 R=0.38 m=11] →NORMAL
244467→[D=0.33, 0.67 R=0.47 m=11] →NORMAL
244468→[D=0.48, 0.52 R=0.39 m=10] →NORMAL
244469→[D=0.94, 0.06 R=0.39 m=10] →NORMAL
244470→[D=0.77, 0.23 R=0.45 m=10] →OUT_OF_APPLICABILITY
```

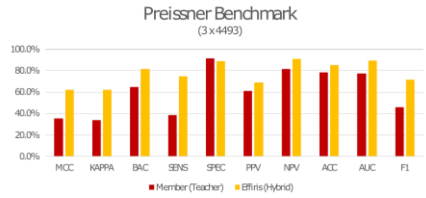


The generalised workflow is shown above. Initially Lhasa curates public data for an endpoint, and also finds the best conditions for modelling the endpoint. The endpoint is then activated for a company to build a model. This model labels our diverse dataset and the member sends a simple text file to Lhasa. In this manner the model is able to act as a teacher transferring the knowledge but not private information[2]. This labelled data is validated to ensure fidelity and these labels are consolidated into a single label per compound. This federated data is returned to the company along with the public data to enable them to build a hybrid model.

Endpoint	Lhasa Curated Data	Private Harvesting	Federated Data	Prospective Validation
hERG	✓	✓	✓	✓
D2	✓	⚙	⌛	⌛
COX-2	✓	⚙	⌛	⌛
A2a	✓	⚙	⌛	⌛
SERT	✓	⚙	⌛	⌛
others	✓	⌛	⌛	⌛

Results





Shown above are averaged results for predicting hERG activity with a 10 µM threshold. The first graph shows a prospective validation, showing what the real world improvement would be if the hybrid model were used over the teacher model. The second validation is the performance of both models against our internal test set compiled from a recent paper by Preissner[3]. The hybrid model brings statistically significant improvement even within the chemical space of each individual company. On average the MCC will increase by 0.04 reflecting an increase in sensitivity without much loss in specificity. It has a much more major impact on the predictivity outside of each company's focussed chemical space. The averaged MCC increases by 0.25 reflecting a greatly improved sensitivity with a stable specificity. The hybrid model is using information about general hERG activity and performs better than the internal model when validating with external data. This shows the generalisability of the model and allows each company to have better predictions outside their area of chemical space when starting a new project.

Conclusion

We have demonstrated a means of securely leveraging the private toxicity knowledge contained within a company to enable others to have better performing models across a wider area of chemical space. This method doesn't leak any proprietary information out into the public sphere as the knowledge is carried on the substrate of a diverse set of public compounds. We have validated this process using the hERG endpoint and seen improvements, both within the chemical space of the company, as well as in a more general chemical space, where previously the performance was lacking. We are currently harvesting data for many other endpoints and will be doing a thorough prospective evaluation on those in the future.

References

1. Bowes, Joanne, et al. "Reducing safety-related drug attrition: the use of in vitro pharmacological profiling." *Nature reviews Drug discovery* 11.12 (2012): 909-922.
2. Papernot, Nicholas, et al "Semi-supervised knowledge transfer for deep learning from private training data" arXiv preprint arXiv:1610.05755 (2016)
3. Siramshetty, Vishal B., et al. "The Catch-22 of predicting hERG blockade using publicly accessible bioactivity data." *Journal of Chemical Information and Modeling* 58.6 (2018): 1224-1233.