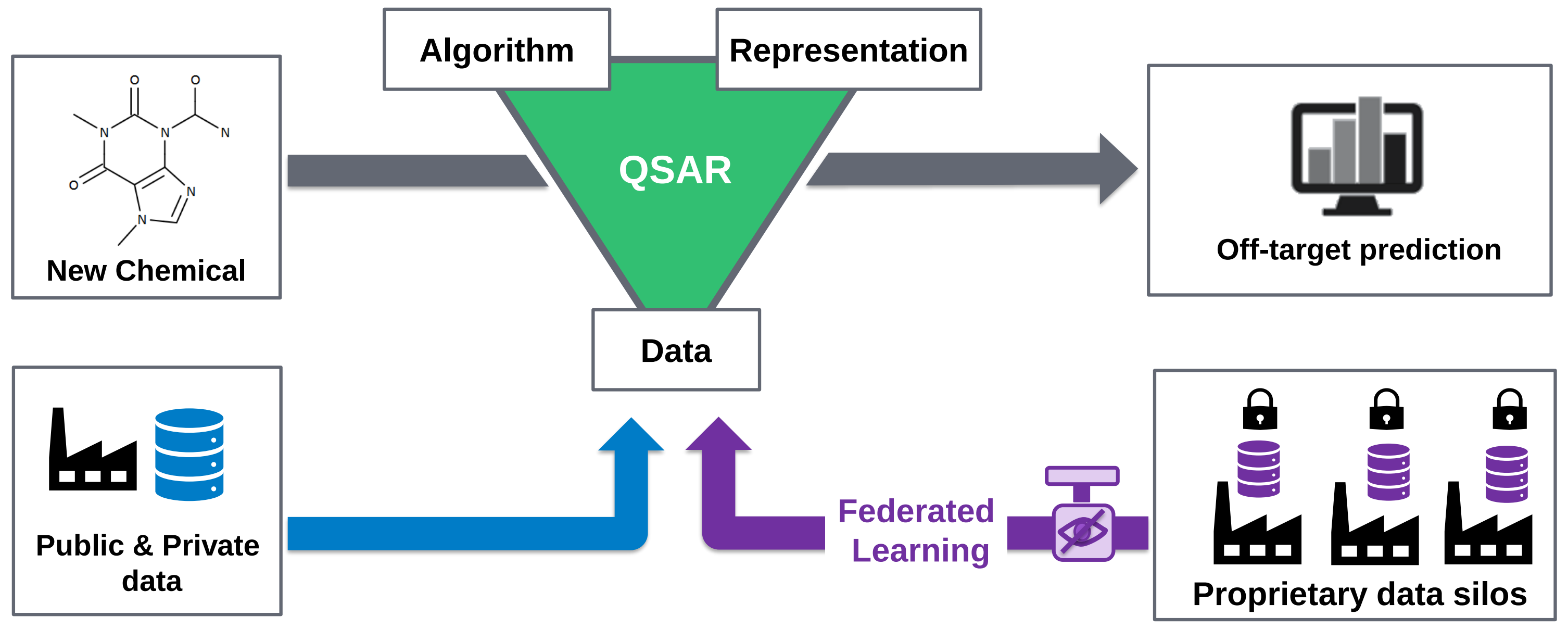
Using privacy-preserving federated learning to enable pre-competitive cross-industry knowledge sharing and improve QSAR models

T. Hanser[1], D. Bastogne[2], A. Basu[3], R. Davies[1], A. Delaunois[2], A. Fowkes[1], A. Harding[1], L. Johnston[1], C. Korlowski[2], E. Kotsampasakou[4], J. Plante[1], L. Rosenbrier-Ribeiro[2], P. Rowell[1], Y. Sabnis[2], A. Sartini[1], A. Sibony[2], S. Werner[1], A. White[4], and T. Yukawa[3].

[1]Lhasa Limited, Leeds, United Kingdom; [2]UCB Biopharma, Braine-l'Alleud, Belgium; [3]Takeda Pharmaceuticals International, Cambridge, MA, USA; and [4]GlaxoSmithKline, Ware, United Kingdom

Unlocking additional value for *in silico* secondary pharmacology profiling

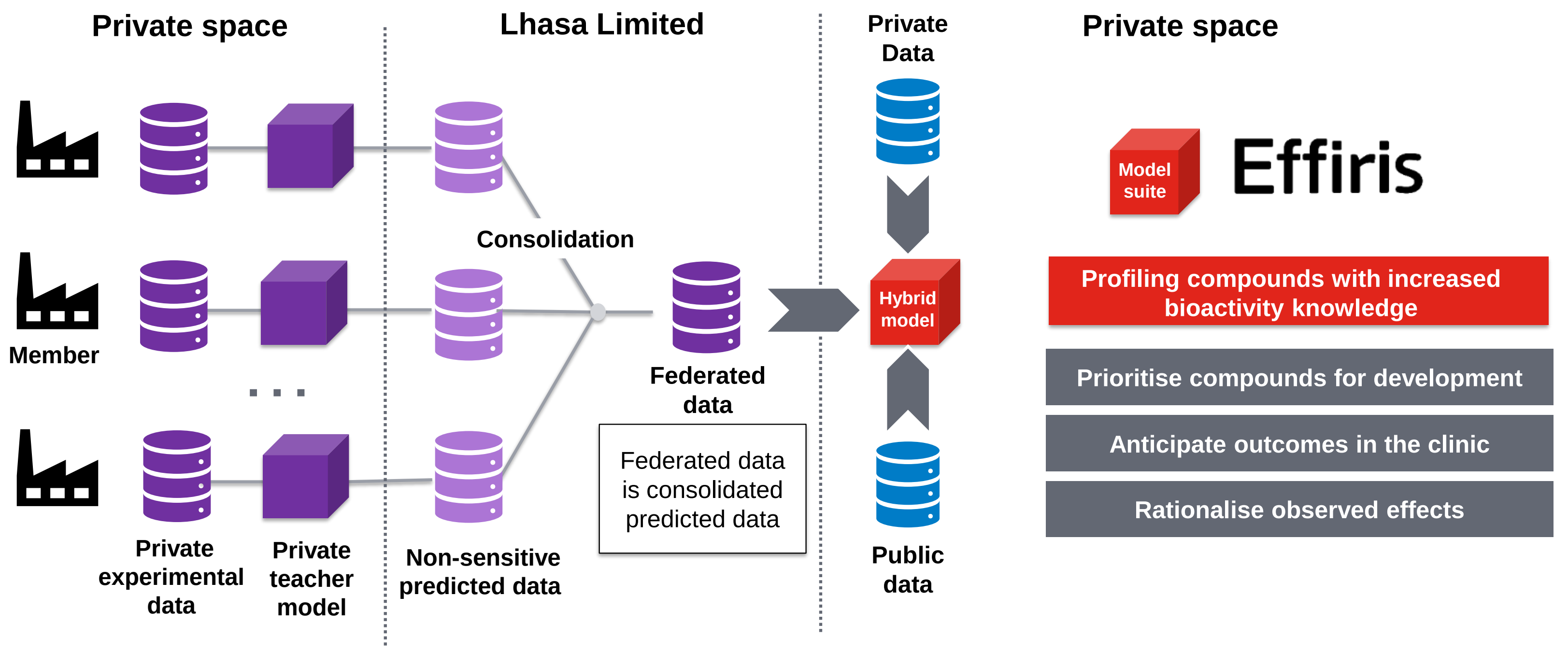
Secondary pharmacology profiling using quantitative structure-activity relationship (QSAR) models offer an efficient way to provide insight into biological properties during compound optimisation and prioritisation [1]. The performance of these models is highly dependent on the quality and the quantity of the data available to train them, particularly when investigating new areas of chemical space.



A wealth of knowledge is locked in high-quality proprietary data silos and allowing QSAR models to access this knowledge would lead to unprecedented performances and decision support. Federated learning can overcome the confidentiality through a privacy-preserving approach to extract knowledge from proprietary data and facilitate pre-competitive collaboration [2].

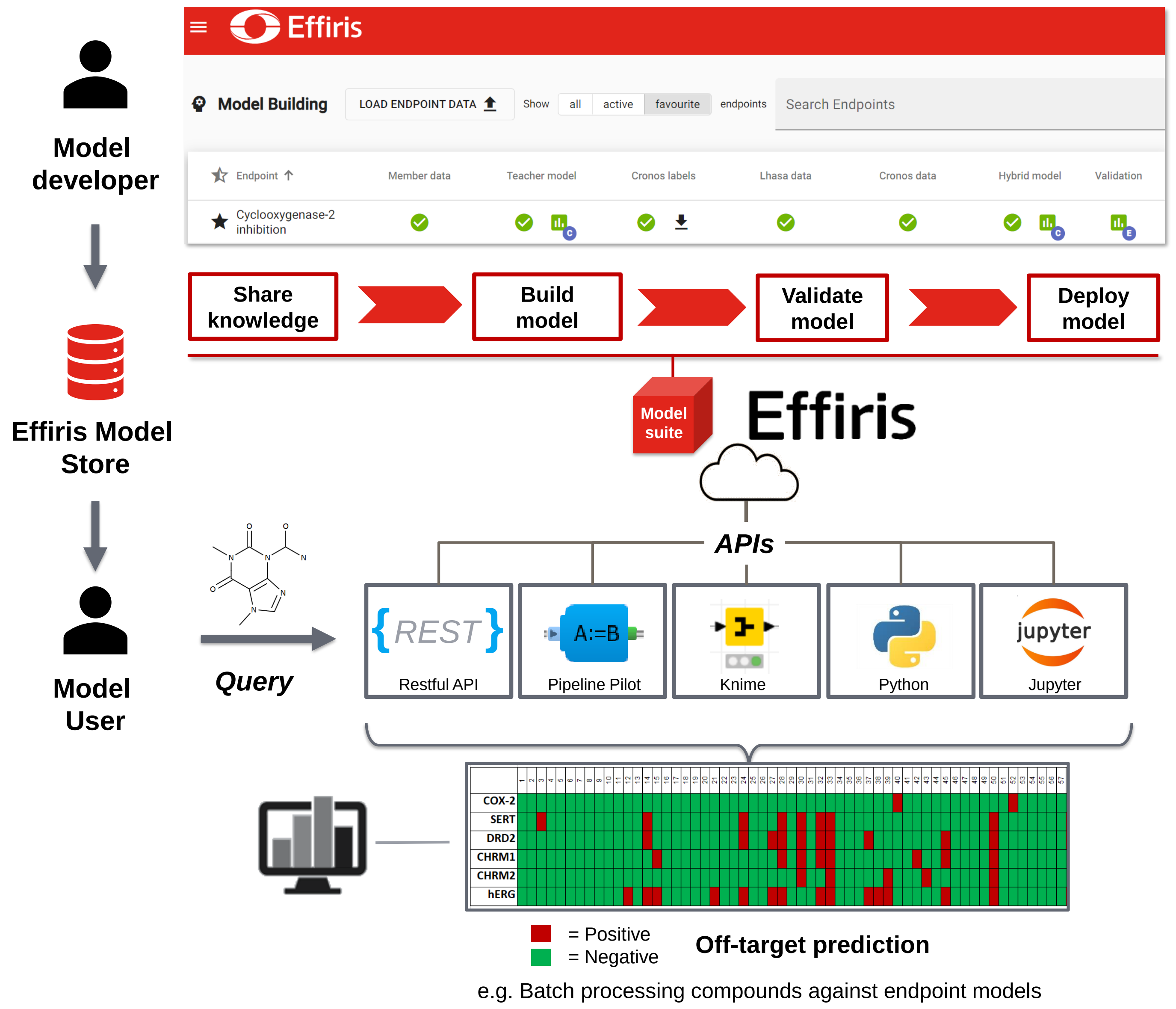
Federated data enables pre-competitive sharing of knowledge between organisations

Lhasa Limited has developed a federated learning platform called Effiris, which enables extraction of knowledge from multiple proprietary data silos without loss of confidential information. The approach labels a common set of public structures with endpoint predictions from teacher models trained on proprietary data. Acting as an honest broker, Lhasa Limited consolidates the predicted data from all the partners into a single robust dataset, whilst also performing quality checks. The consolidated data is then shared with members, who are then able to combine this new data with in-house and public data to generate hybrid models, which have learnt from multiple sources of knowledge.


Private space
Lhasa Limited
Private Data
Private space
Member
Consolidation
Federated data
Federated data is consolidated predicted data
Private experimental data
Private teacher model
Non-sensitive predicted data
Hybrid model
Public data
Model suite
Effiris
Profiling compounds with increased bioactivity knowledge
Prioritise compounds for development
Anticipate outcomes in the clinic
Rationalise observed effects

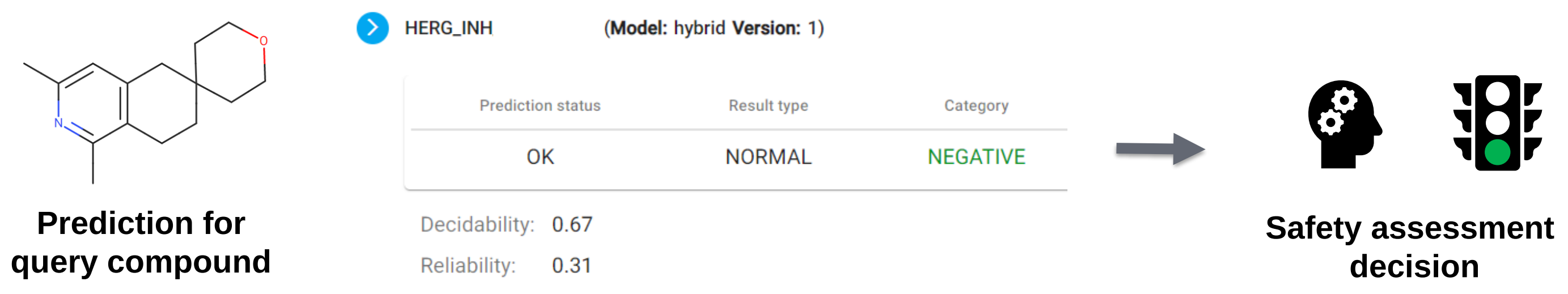
User-friendly federated learning platform and workflow integration tools

To facilitate the adoption of federated QSAR models across the organisations in the Effiris consortium, a user-friendly platform was developed to support model building, validation and deployment. In addition, the platform allows users to control when knowledge is shared. Effiris models are automatically deployed to a web service and can be easily consumed through ready-to-use APIs. The platform facilitates integration of Effiris predictions with existing environments, for example, PipeLine Pilot protocols, Knime workflows, Jupyter notebooks or directly in existing web portals and thus enabling user to readily consume the knowledge within the federated models.



e.g. Batch processing compounds against endpoint models

Effiris provides decision support through its applicability domain framework [3], where the off-target prediction is presented alongside a decidability and reliability score, which reflect the concordance and the density of the supporting information, respectively.



HERG_INH (**Model:** hybrid **Version:** 1)

Prediction status	Result type	Category
OK	NORMAL	NEGATIVE

Decidability: 0.67

Reliability: 0.31

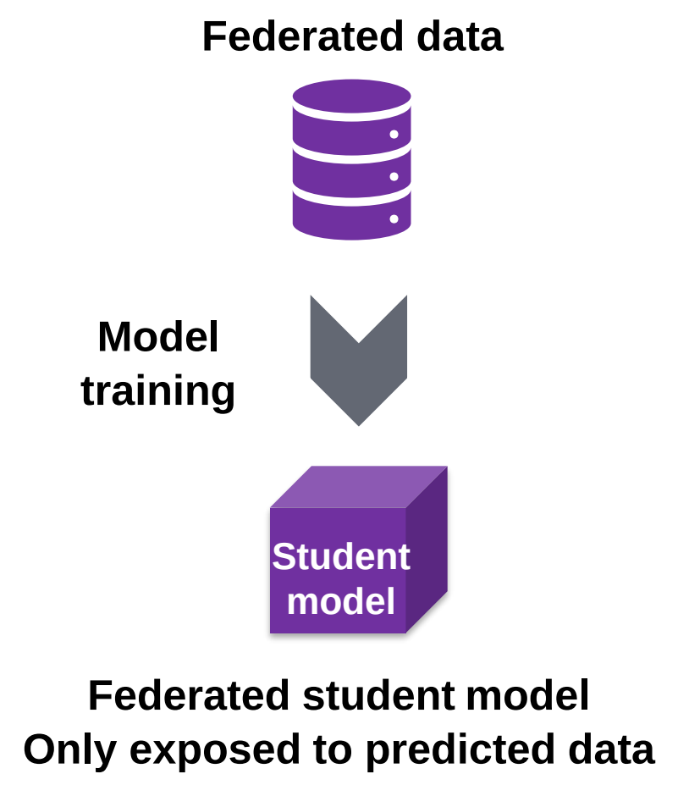
References & Disclosures

[1] Jenkinson S *et al*., J. Pharmacol. Toxicol. Methods, 2020, 105, 106869, doi:10.1016/j.vascn.2020.106869
[2] Papernot N *et al*., Conference paper at ICLR 2017, arXiv:1610.05755
[3] Hanser T *et al*., Advances in Computational Toxicology pp 215-232, doi: 10.1007/978-3-030-16443-0_11

- The initiative is sponsored by the Effiris consortium, which is comprised of the companies in the author list
- There are no conflicts of interest to declare

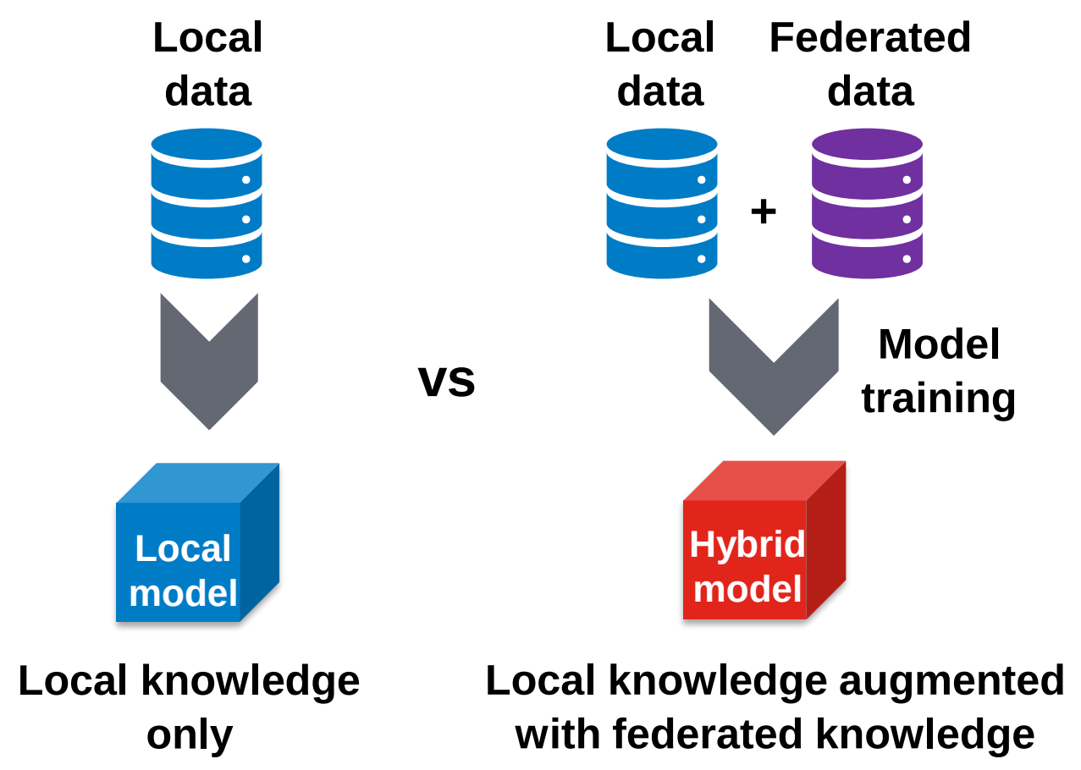
Shared bioactivity knowledge can be used to improve QSAR models

To examine if knowledge has been successfully extracted and transferred across the consortium, a classification model called the student was trained on the federated data only. This model was validated against a test set and a minimum predictivity corresponding to a Matthews correlation coefficient (MCC) value of 0.2 was required to conclude if knowledge had been transferred. From the eight prioritised endpoints, Effiris was able to transfer knowledge for six targets. The main factor that hampered knowledge transfer for the remaining two endpoints was a strong bias of the original private data across all the contributing members.



Endpoint	Federated student model (MCC)	Knowledge transferred? (MCC > 0.2)
COX-1	0.06	NO
COX-2	0.31	YES
GABA-A	0.19	NO
SERT	0.21	YES
DRD2	0.37	YES
CHRM1	0.42	YES
CHRM2	0.56	YES
hERG	0.36	YES

To examine if the shared knowledge for the six endpoints could be used to improve QSAR models, the federated data was combined with local public data to create hybrid classification models, which were then validated using a challenging cluster cross validation. The validation demonstrated that the approach resulted in improved QSAR models, which were able to make more correct predictions over a wider area of chemical space. These models will enable assessors to identify and act upon toxicity liabilities earlier in their workflows.



Average performance of models for the six endpoints where the federated data has demonstrated successful knowledge transfer

	Local QSARs	Hybrid QSARs	QSARs benefit from shared knowledge?
Performance (MCC)	0.50	0.54	YES
Coverage (%)	50	71	YES

Conclusions and Future work

- Federated learning enables the sharing of knowledge contained in proprietary data silos without loss of confidential information
- Effiris is able to transfer bioactivity knowledge between organisations using federated data and improve the accuracy and coverage of QSAR models
- The Effiris consortium will continue to share knowledge and work towards developing a suite of federated QSAR models to support early-stage hazard identification
- Deep learning and novel representation approaches will be examined to improve knowledge transfer and support potency predictions using federated models.

shared **knowledge** ● shared **progress**