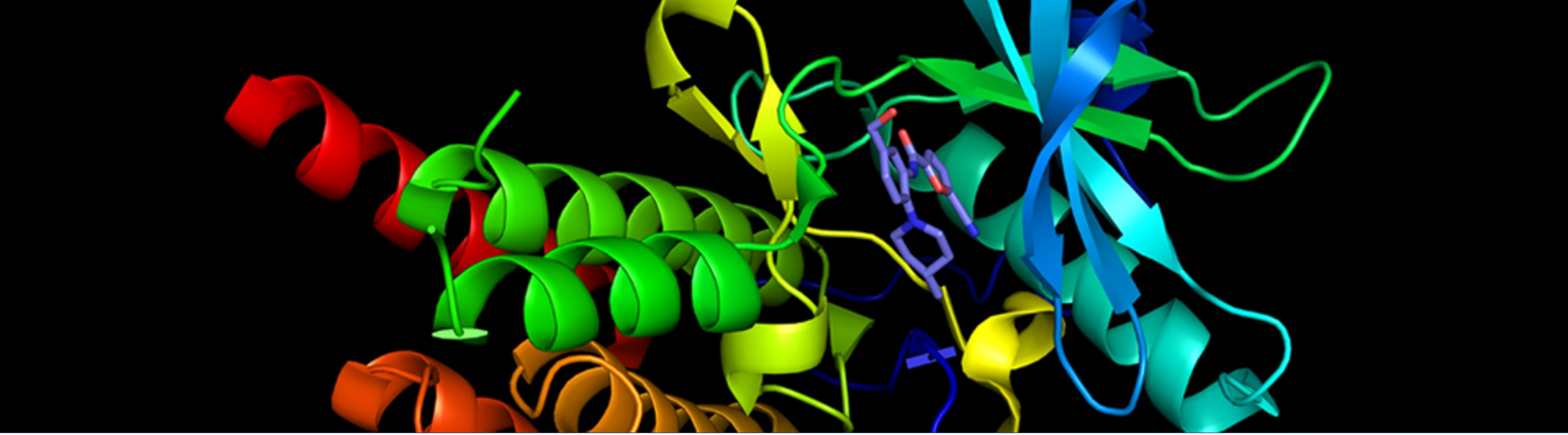


Ligand-based ML for virtual screening of anti-tuberculosis compounds: a viable option?

September 2023

Wouter Heyndrickx, Jorge Esquivias, Christos Varsakelis

Art credit: Discovery Sciences, *Kinase domain of colony-stimulating factor-1 receptor, shown as a rainbow ribbon with a bound inhibitor colored by an atom with purple carbons.*

Ligand-based ML for virtual screening of anti-tuberculosis compounds: a viable option?

Tuberculosis
- 1.5 million deaths annually[1]
- Emerging drug-resistance

How to retrieve novel chemical hit matter?

Virtual screening
- *Approach*: predicting activity from ligand-based ML model
- *Literature*: enrichment factors of at least 10[2]
- *Challenge*: generalization across the vastness of chemical space

How to maximize chances for generalization?

Training **data maximalization**
(>500,000 compound-activity pairs in the public domain from various institutions)

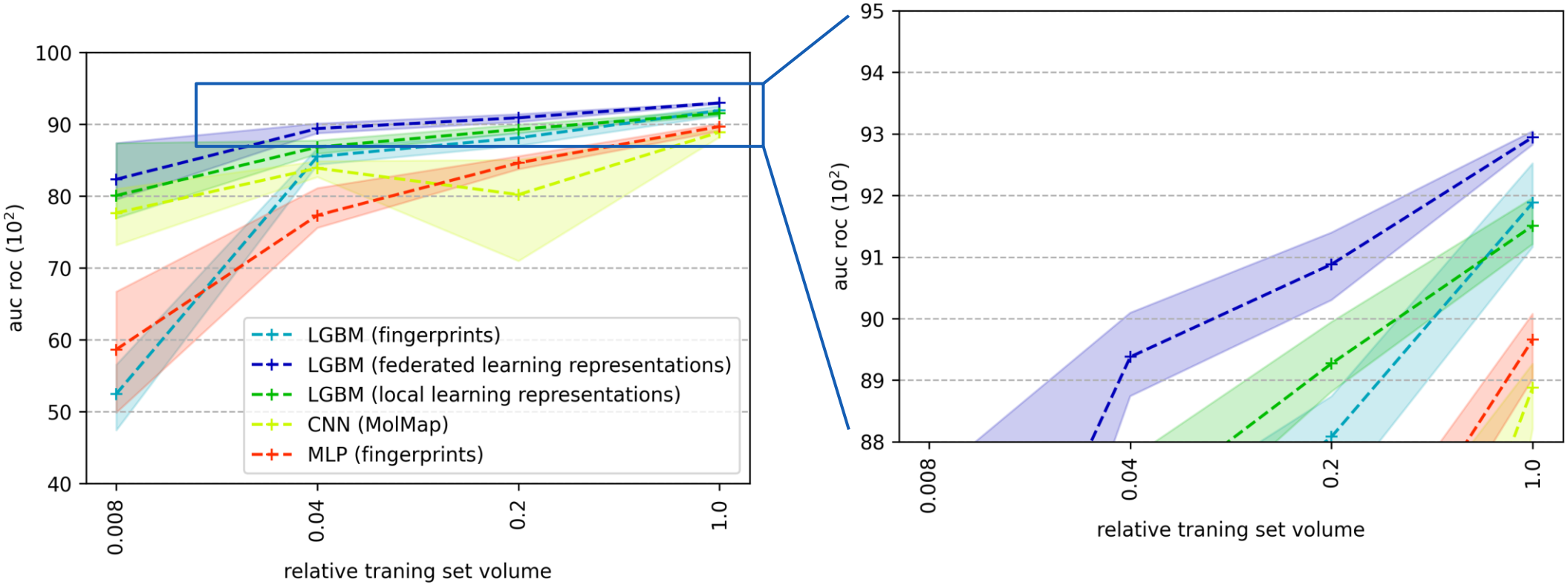
Descriptors derived from **federated learning** (MELLODDY)

[1] World Health Organization - Tuberculosis. World Health Organization - Tuberculosis. https://www.who.int/health-topics/tuberculosis#tab.
[2] Ekins, S.; Freundlich, J. S.; Hobrath, J. V.; Lucile White, E.; Reynolds, R. C. Combining Computational Methods for Hit to Lead Optimization in Mycobacterium Tuberculosis Drug Discovery. Pharm. Res. 2014. https://doi.org/10.1007/s11095-013-1172-7.

Can model quality be boosted by
- Increased data quantity?
- Descriptors derived from massive scale federated learning (MELLODDY)?

Binary classification
Phenotypic activity
@10uM
>500k datapoints

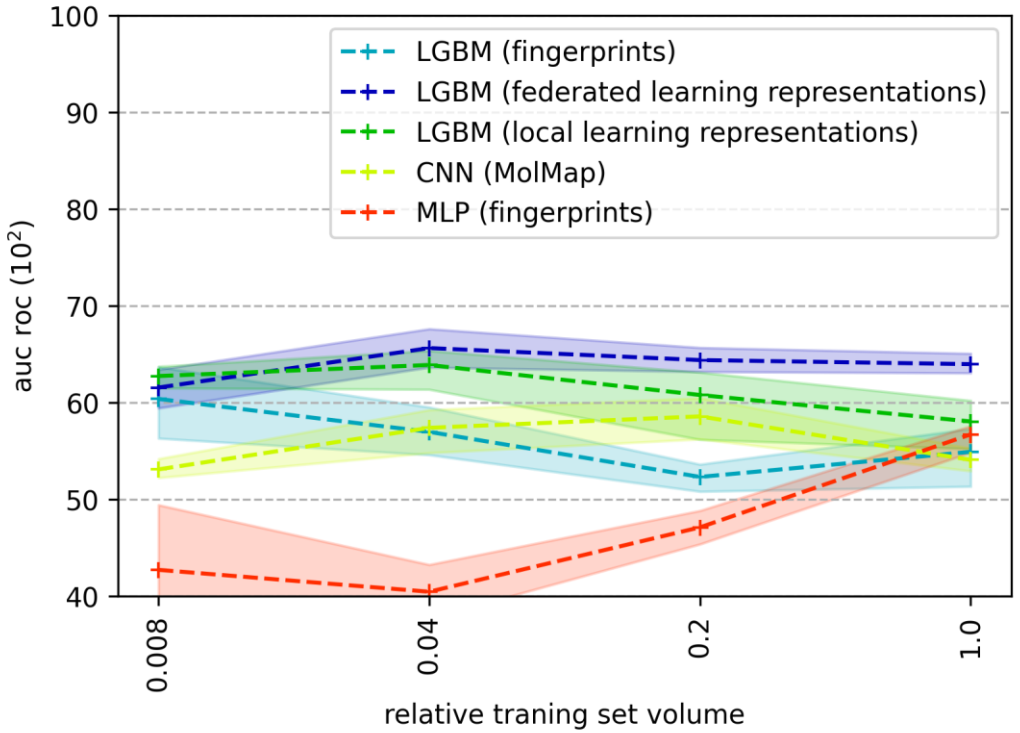
Performance on 'internal' scaffold-network-split test set
(Janssen + Public)



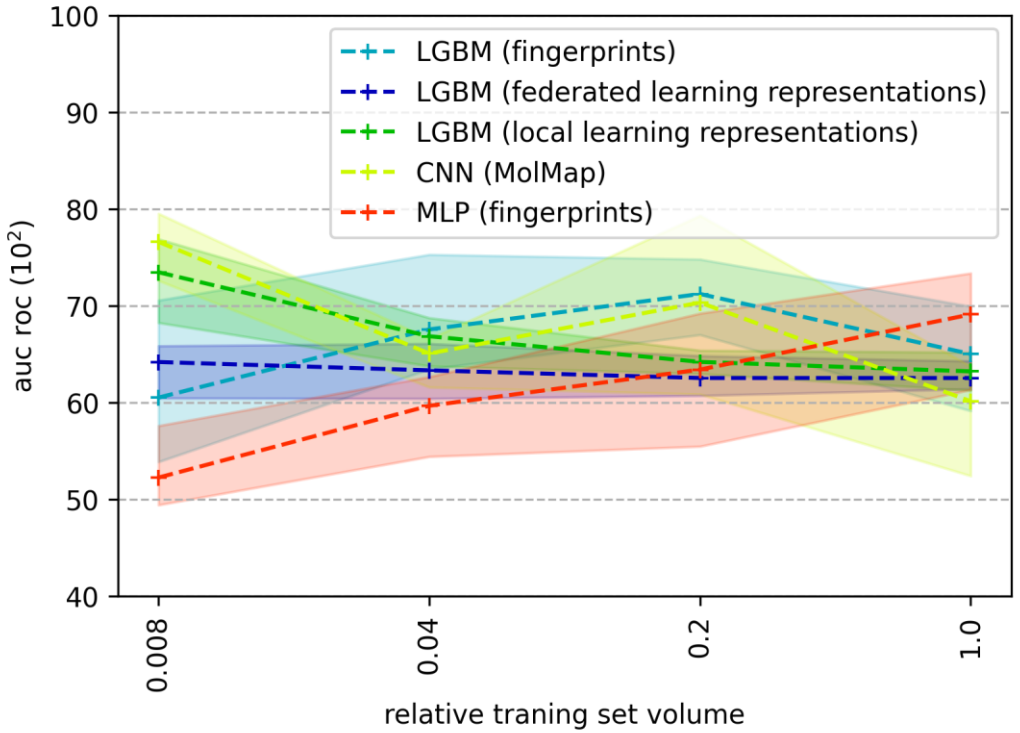
Both data quantity and federated learning boost model performance

Can model quality be boosted by
- Increased data quantity?
- Descriptors derived from massive scale federated learning (MELLODDY)?

Performance on external test set (industry[1])
Tanimoto distance >= 0.4



Performance on external test set (consortium[2])
Tanimoto distance >= 0.4

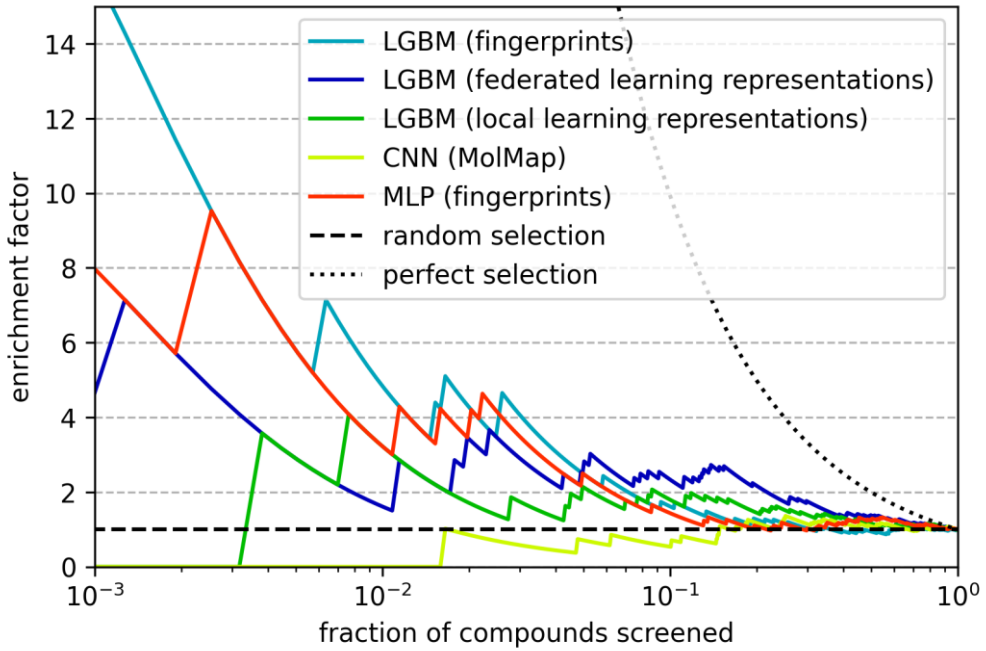


On a distant, external test set, no consistent effect from the data quantity or algorithm

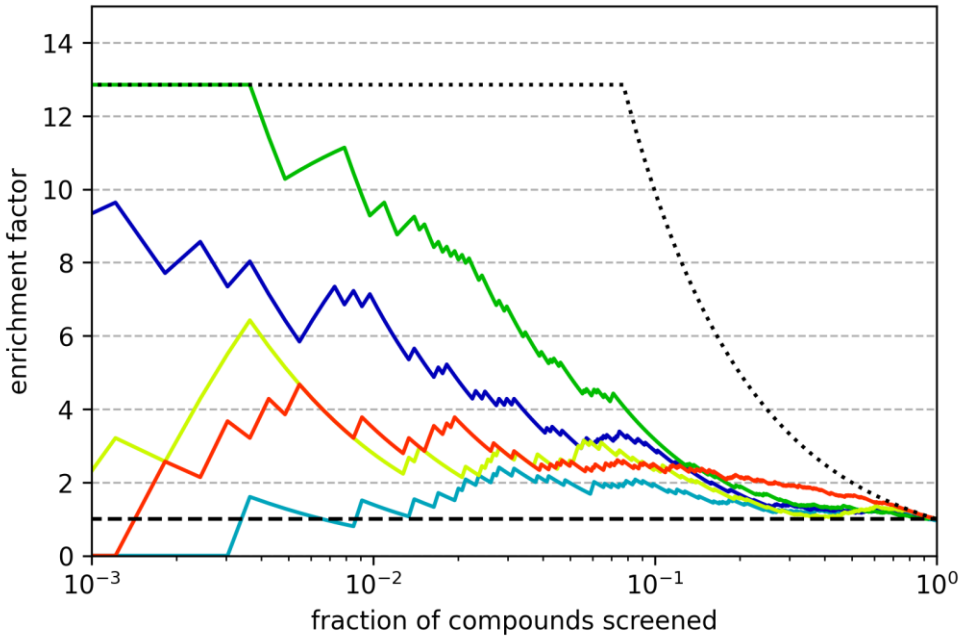
[1] 1 L. Ballell, R. H. Bates, R. J. Young, D. Alvarez-Gomez, E. Alvarez-Ruiz, V. Barroso, D. Blanco, B. Crespo, J. Escribano, R. González, S. Lozano, S. Huss, …, N. Cammack, ChemMedChem, 2013, 8, 313–321.
[2] T. R. Lane, F. Urbina, L. Rank, J. Gerlach, O. Riabova, A. Lepioshkin, E. Kazakova, A. Vocat, V. Tkachenko, S. Cole, V. Makarov and S. Ekins, Mol. Pharm., 2022, 19, 674–689.

Which enrichment factors can be expected in external virtual screening?

Performance on external test set (industry[1])
Tanimoto distance >= 0.4



Performance on external test set (consortium[2])
Tanimoto distance >= 0.4



For full dataset

2-7 fold enrichment

[1] 1 L. Ballell, R. H. Bates, R. J. Young, D. Alvarez-Gomez, E. Alvarez-Ruiz, V. Barroso, D. Blanco, B. Crespo, J. Escribano, R. González, S. Lozano, S. Huss, …, N. Cammack, ChemMedChem, 2013, 8, 313–321.
[2] T. R. Lane, F. Urbina, L. Rank, J. Gerlach, O. Riabova, A. Lepioshkin, E. Kazakova, A. Vocat, V. Tkachenko, S. Cole, V. Makarov and S. Ekins, Mol. Pharm., 2022, 19, 674–689.

Conclusions

- Models leveraging federated learning tend to outperform others
- Model predictive performance based on (scaffold-split) folds,
  is not representative of performance in external screening
- Models can be useful to increase the experimental hit rate in screening with factor 2-7

Powered by RDKit
- Standardization (Chem.MolStandardize)
- Scaffold network (Chem.Scaffolds.rdScaffoldNetwork)
- Similarity (DataStructs.cDataStructs.BulkTanimotoSimilarity)
- Fingerprints (Chem.AllChem.GetMorganFingerprint)
- Mw (Chem.Descriptors.ExactMolWt())
- Atom counting (mol.GetNumAtoms())
- Chem.RemoveStereochemistry()