deepmirror

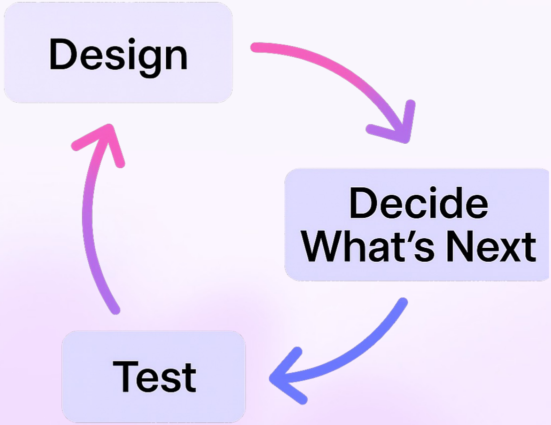
Bio-Assay Derived Fingerprints: Performance Comparison in AutoML Pipelines

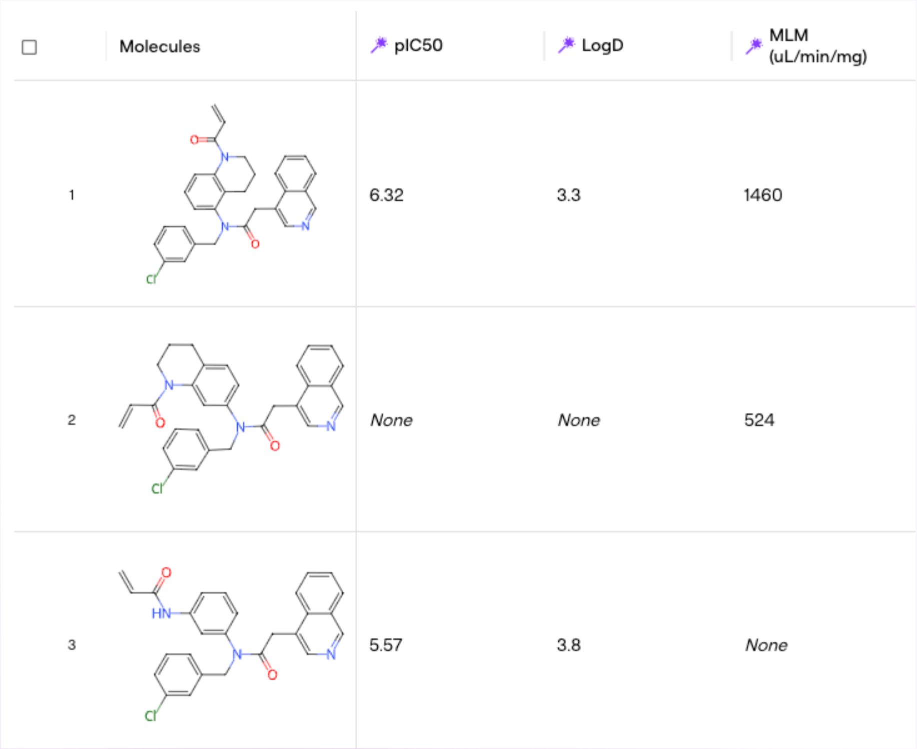
Ryan Greenhalgh | CTO and Co-Founder

deepmirror

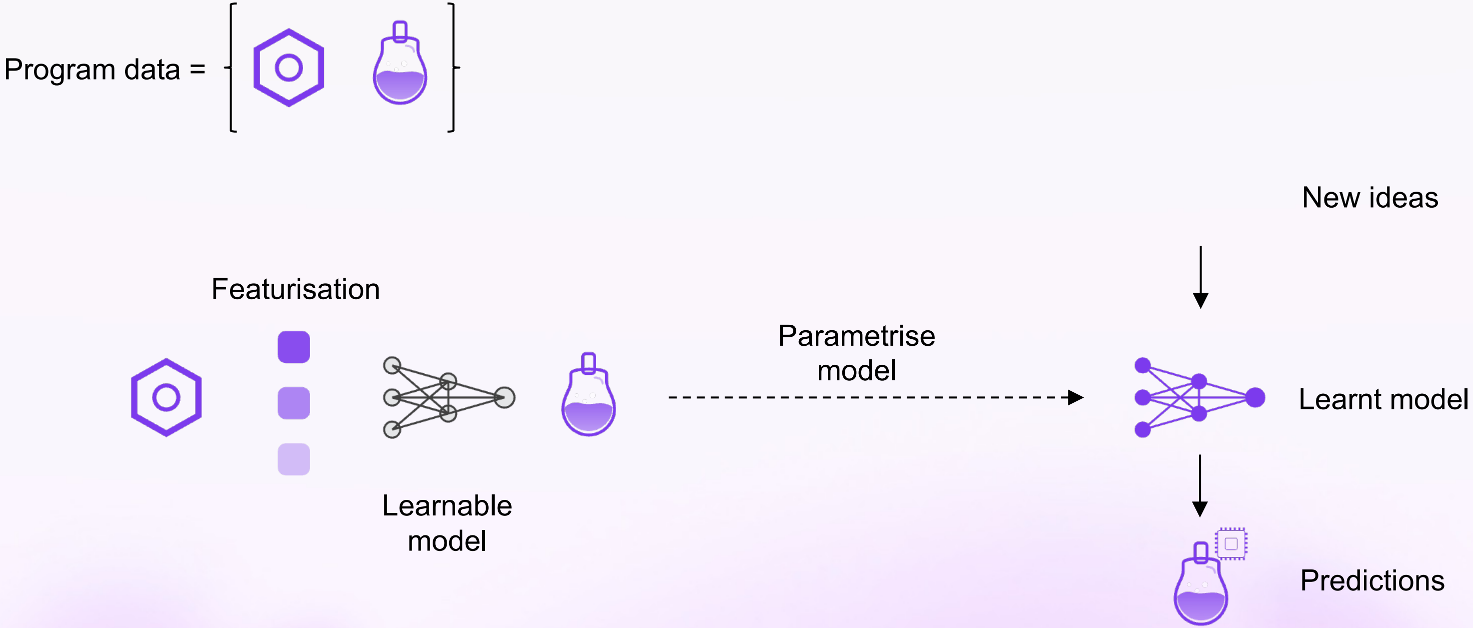
Early Discovery: What to Make Next?

- Early programs <100 datapoints per assay
- Need to **prioritise next compounds**
- Can we build models to help?





□	Molecules	pIC50	LogD	MLM (uL/min/mg)
1		6.32	3.3	1460
2		*None*	*None*	524
3		5.57	3.8	*None*



Program data =

New ideas

Featurisation

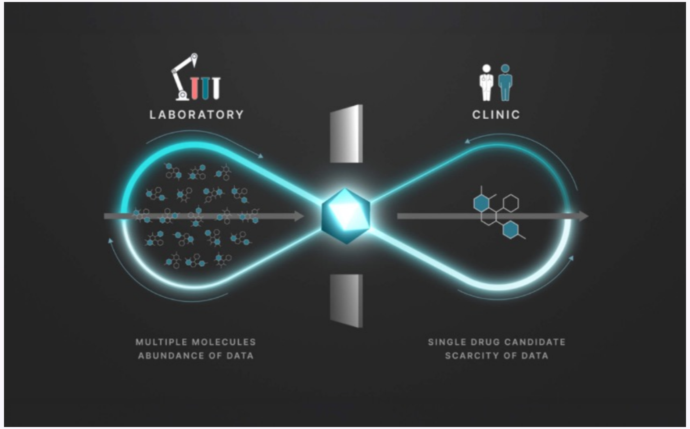
Parametrise model

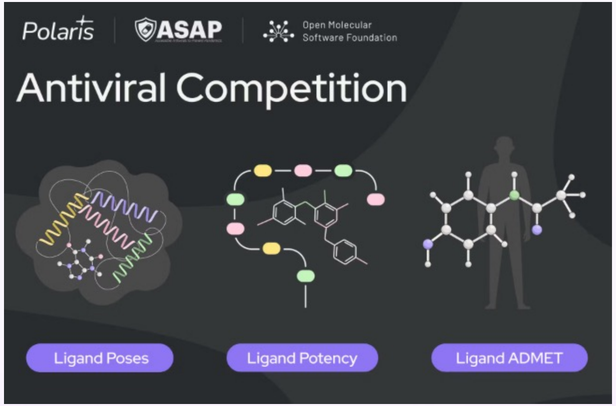
Learnt model

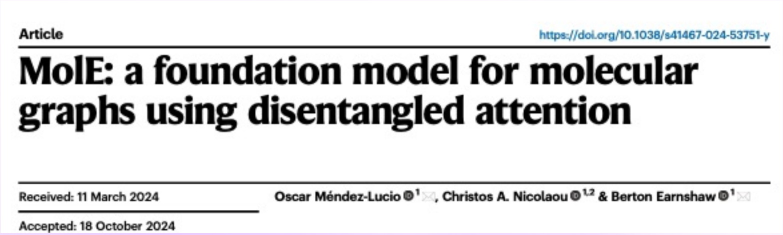
Learnable model

Predictions

deepmirror







Article

https://doi.org/10.1038/s41467-024-53751-y

MolE: a foundation model for molecular graphs using disentangled attention

Received: 11 March 2024

Accepted: 18 October 2024

Oscar Méndez-Lucio [1], Christos A. Nicolaou [1,2] & Berton Earnshaw [1]

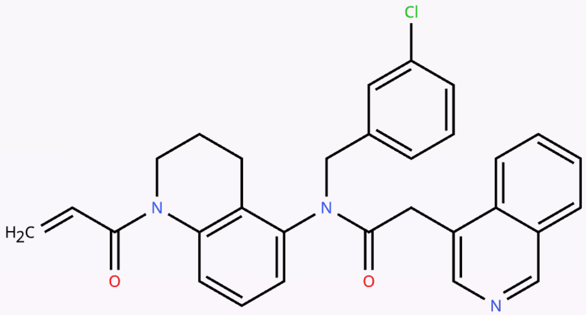
How to learn on fingerprints

Model	Pros	Cons	<100 datapoints	100–10k	>10k
Gaussian Process	Uncertainty-aware, good w/ low data	Slow, tricky to scale	+++	+++	!
Trees / SVM / kNN	Fast, robust	No uncertainty, less expressive	++	+++	!
Deep Learning	Best w/ big data	Overfits on small datasets	xx	!	+++

Green et al., *Current Methods for Drug Property Prediction in the Real World*, arXiv:2309.17161 (2023)

deepmirror

C=CC(=O)N1CCCc2...



Cheminformatics

Mordred

RDKit

Morgan

MACCS

fast, proven, interpretable

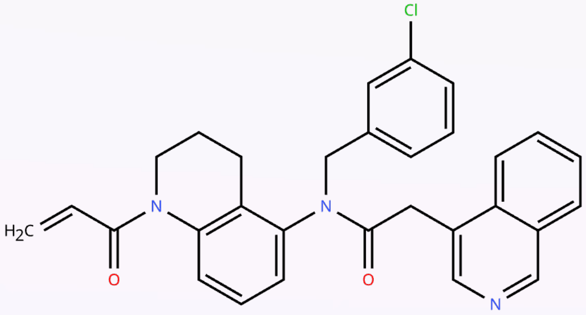
Machine Learning

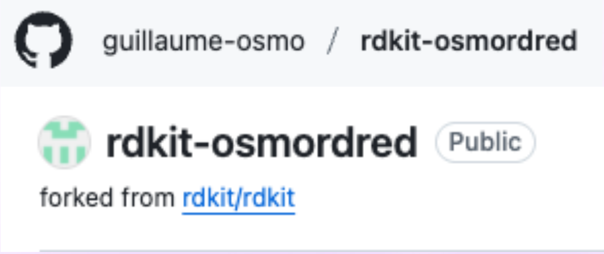
CheMeleon

ChemBERTa

Mol2Vec

RDKit / Mordred / MACCS

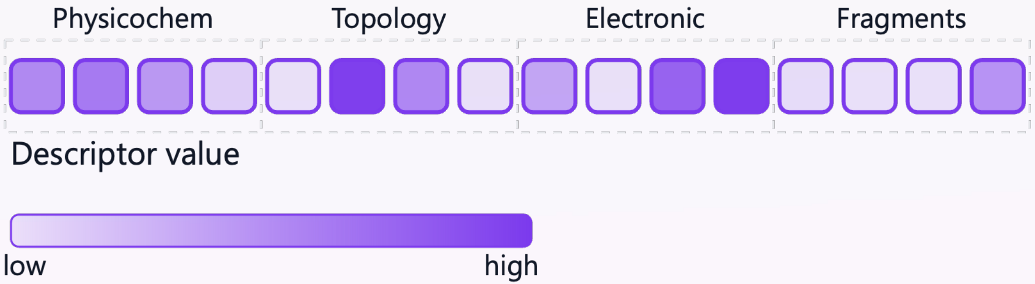




guillaume-osmo / rdkit-osmordred

rdkit-osmordred Public

forked from rdkit/rdkit



Physicochem

Topology

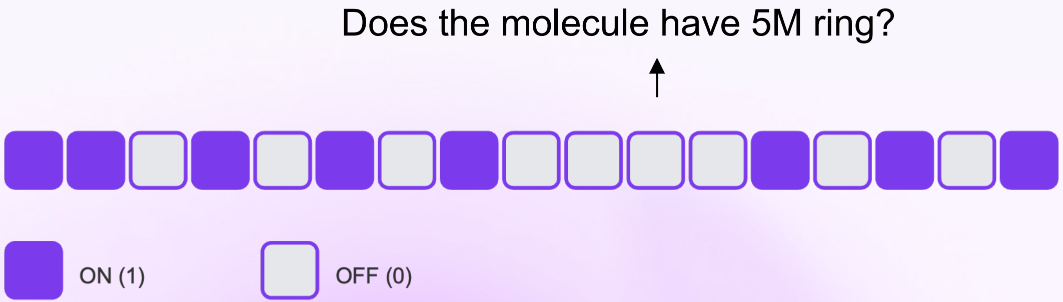
Electronic

Fragments

Descriptor value

low

high



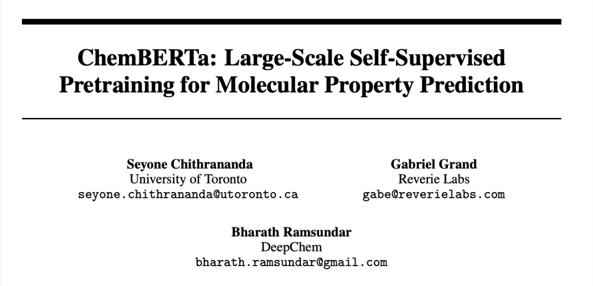
Does the molecule have 5M ring?

■ ON (1)

□ OFF (0)

Mol2vec/ChemBERTa

Leveraging a CBOW/transformer trained on large datasets i.e. PubChem dataset (77M compounds).

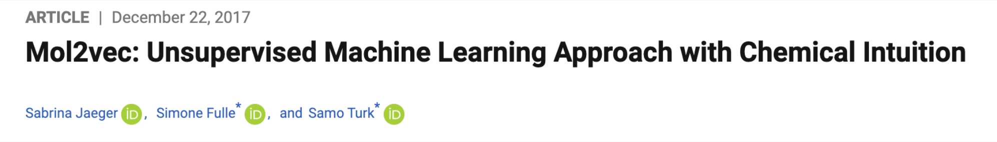


ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction

Seyone Chithrananda
University of Toronto
seyone.chithrananda@utoronto.ca

Gabriel Grand
Reverie Labs
gabe@reverielabs.com

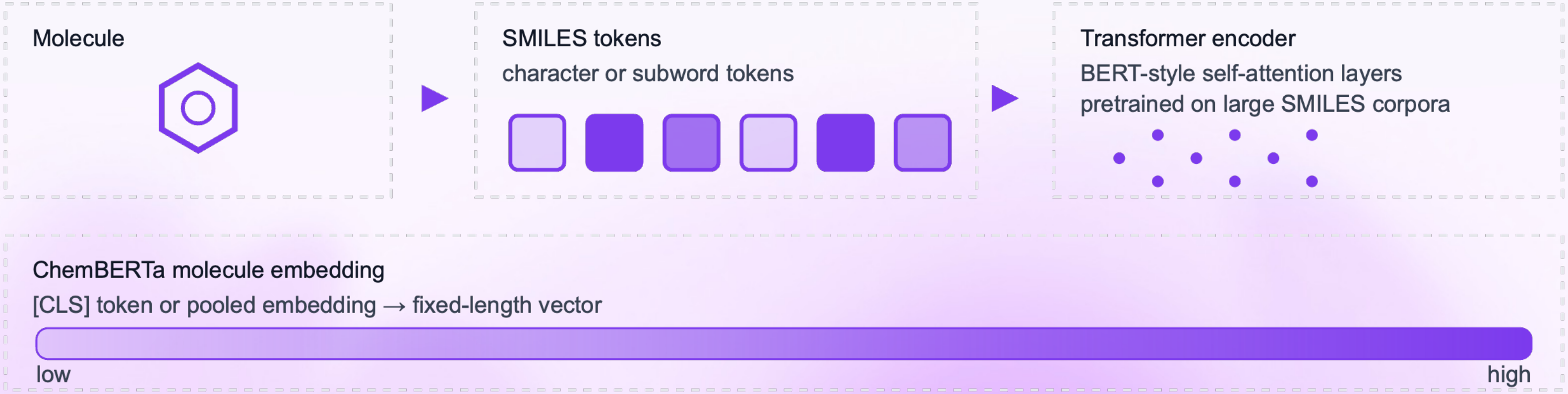
Bharath Ramsundar
DeepChem
bharath.ramsundar@gmail.com



ARTICLE | December 22, 2017

Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition

Sabrina Jaeger, Simone Fulle*, and Samo Turk*



Molecule

SMILES tokens
character or subword tokens

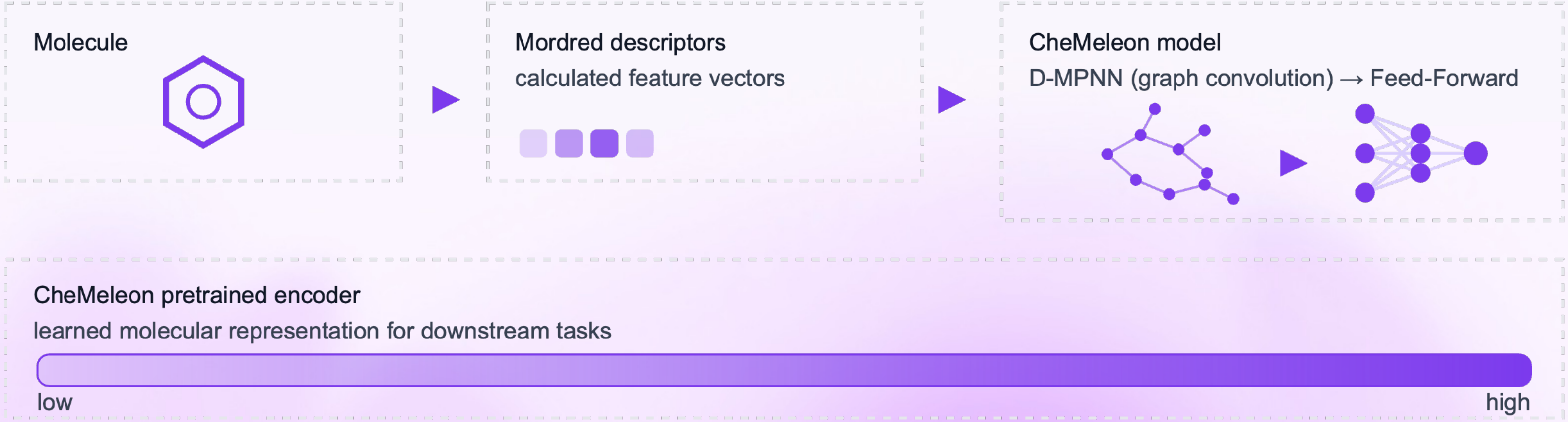
Transformer encoder
BERT-style self-attention layers pretrained on large SMILES corpora

ChemBERTa molecule embedding
[CLS] token or pooled embedding → fixed-length vector

CheMeleon

chemeleon Public

Similar to ChemBERTa, but applied to physical fingerprints and implemented with GNNs.



Molecule

Mordred descriptors

calculated feature vectors

CheMeleon model

D-MPNN (graph convolution) → Feed-Forward

CheMeleon pretrained encoder

learned molecular representation for downstream tasks

CheMeleon

ChemBERTa

Mordred

Mol2Vec

Morgan

RDKit

MACCS

Traditional fingerprints rely on physico-chemical properties, structural chemistry, or information-theoretic descriptors of the molecule

ChEMBL

PubChem

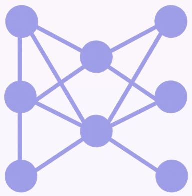
Does training on **assay outcomes (ADMET, potency, selectivity)** make learned fingerprints more predictive for real-world drug discovery?

Combining heterogeneous data

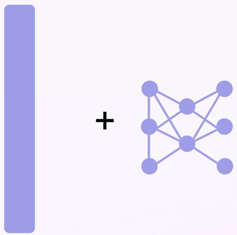
Molecule	assay-unit-source	assay-unit-source	HLM-hr-labA	HLM-hr-labB	P1-pic50-orgA	P2-pic50-orgB
C=CC(=O…		7				12
C(=O..				8	180	
C=CC(=O…	1.2					
CCN…			143			

Combining IC50 or Ki Values from Different Sources Is a Source of Significant Noise. *J. Chem. Inf. Model.* **64**(5), 1560–1567 (2024). G.A. Landrum & S. Riniker. DOI: 10.1021/acs.jcim.4c00049

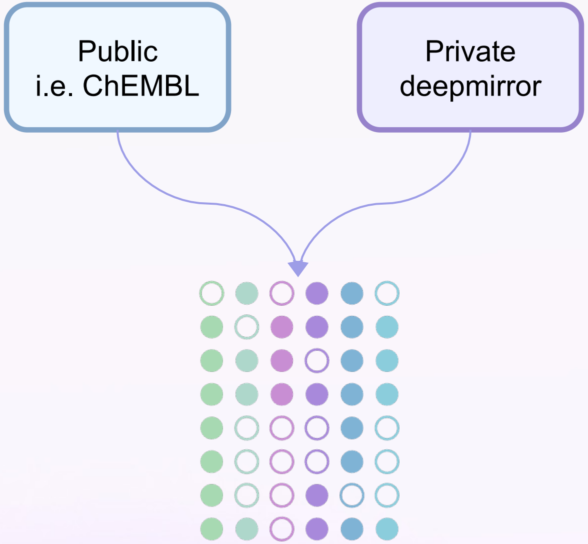
Pre-training ADME/Potency FPs



Message passing graph neural network



Aggregation layer + feed-forward MLP

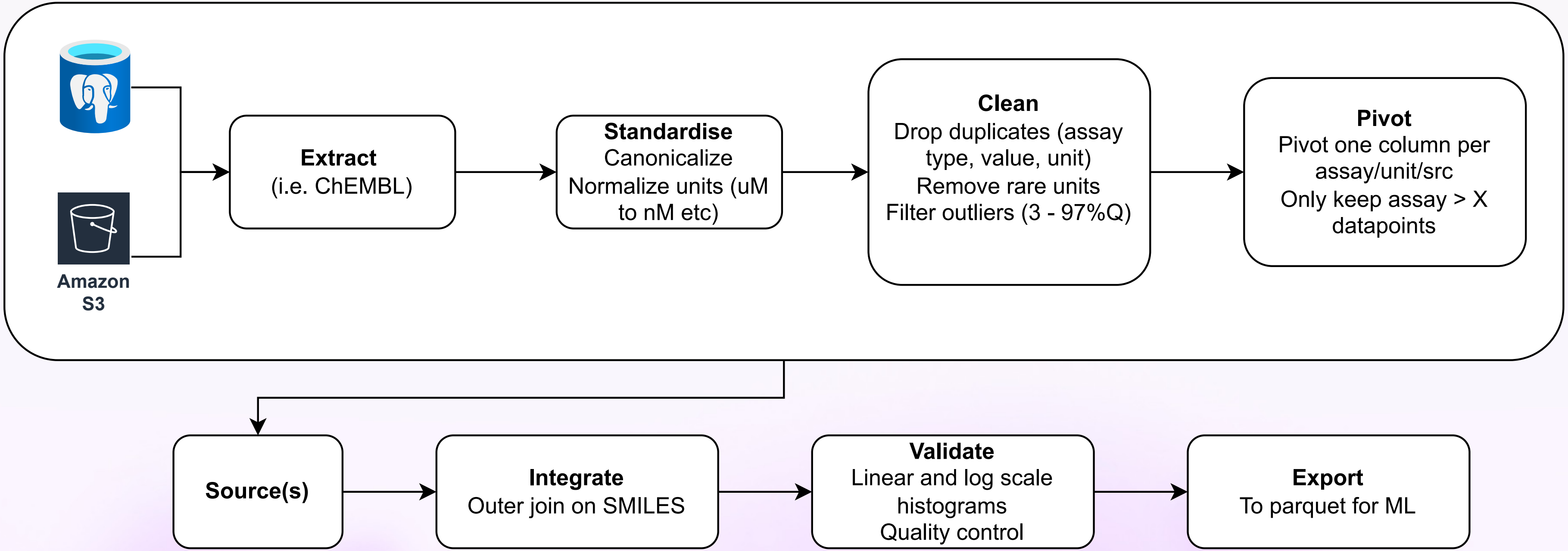


Public i.e. ChEMBL

Private deepmirror

Standardize + harmonized assay data

deepmirror



Amazon S3

Extract
(i.e. ChEMBL)

Standardise
Canonicalize
Normalize units (uM to nM etc)

Clean
Drop duplicates (assay type, value, unit)
Remove rare units
Filter outliers (3 - 97%Q)

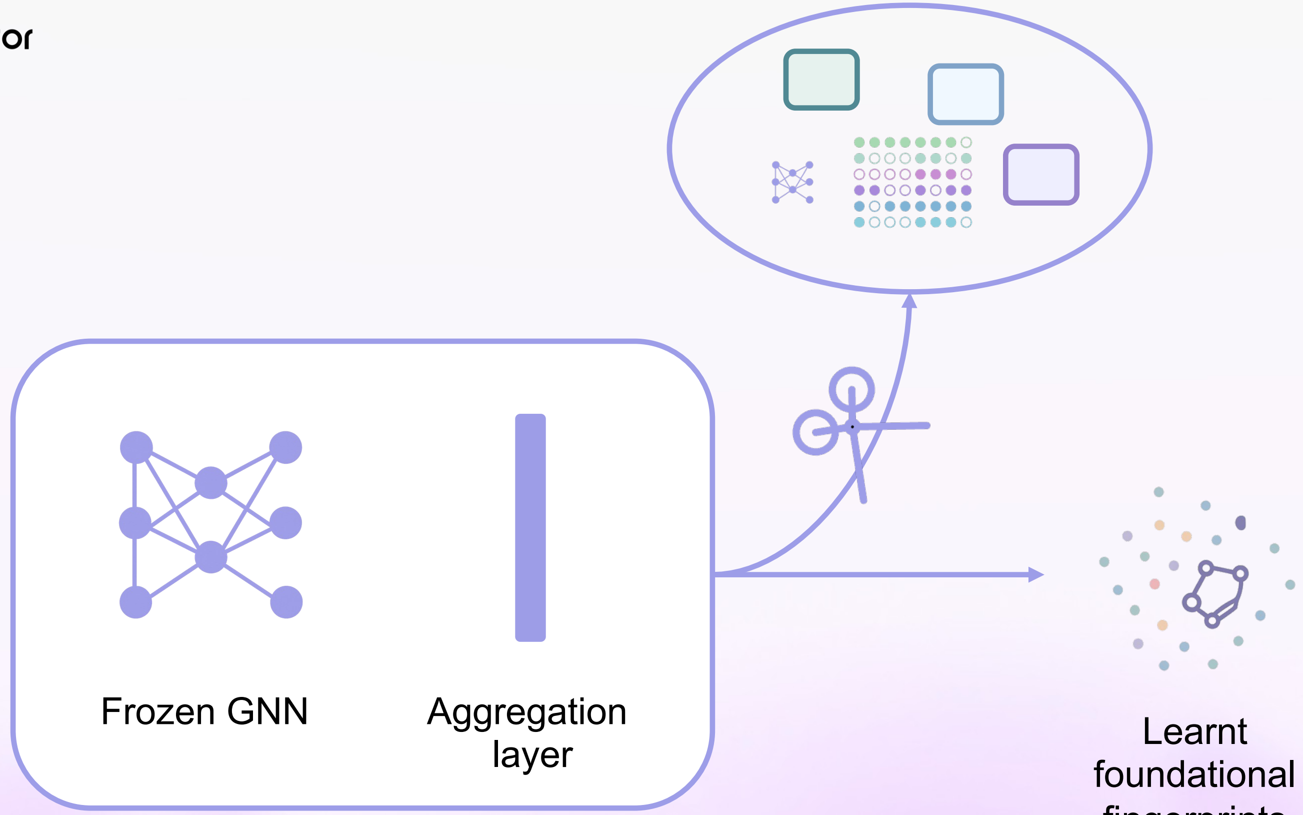
Pivot
Pivot one column per assay/unit/src
Only keep assay > X datapoints

Source(s)

Integrate
Outer join on SMILES

Validate
Linear and log scale histograms
Quality control

Export
To parquet for ML



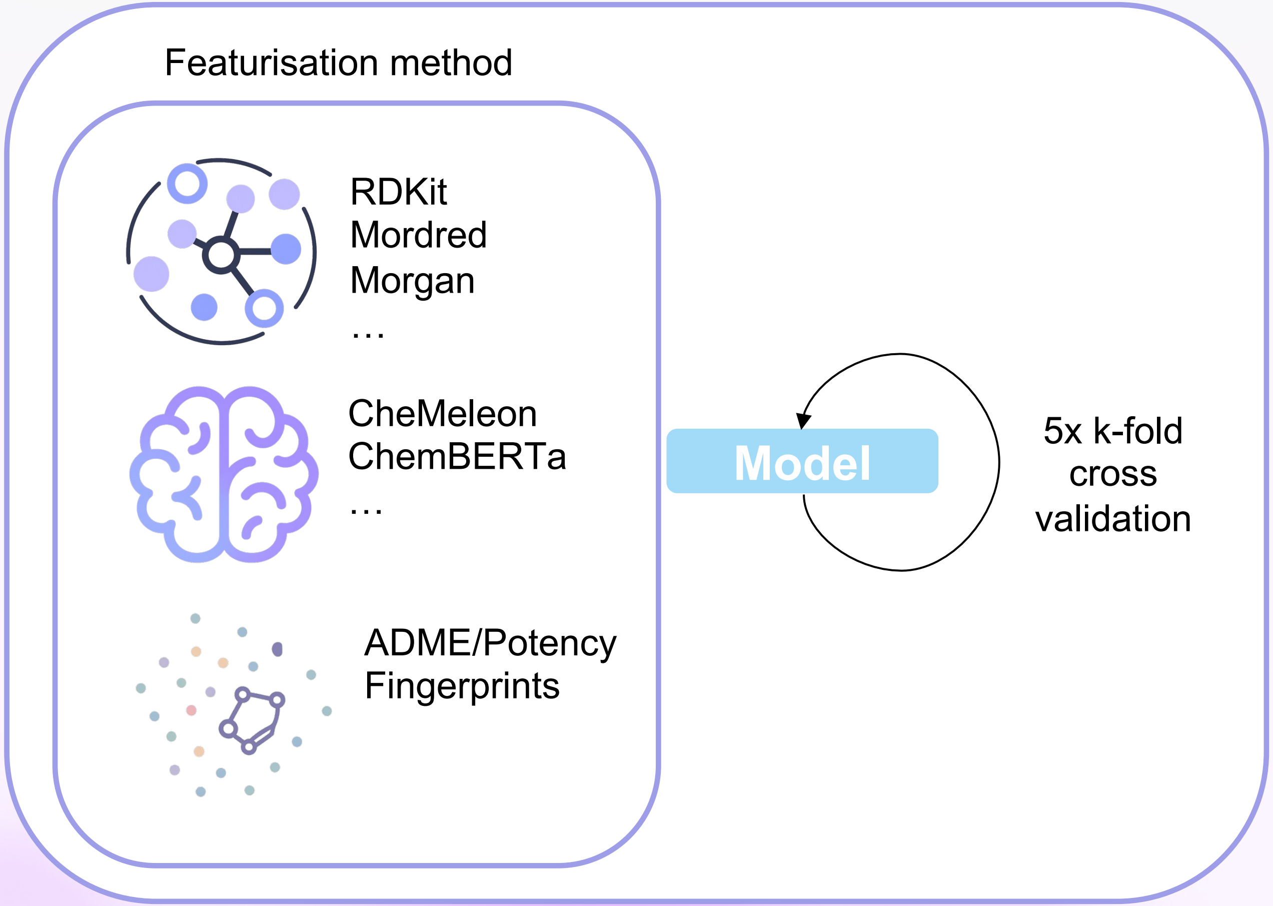
Frozen GNN

Aggregation layer

Learnt foundational fingerprints

Private programs

100s anonymized user-provided data.



Featurisation method

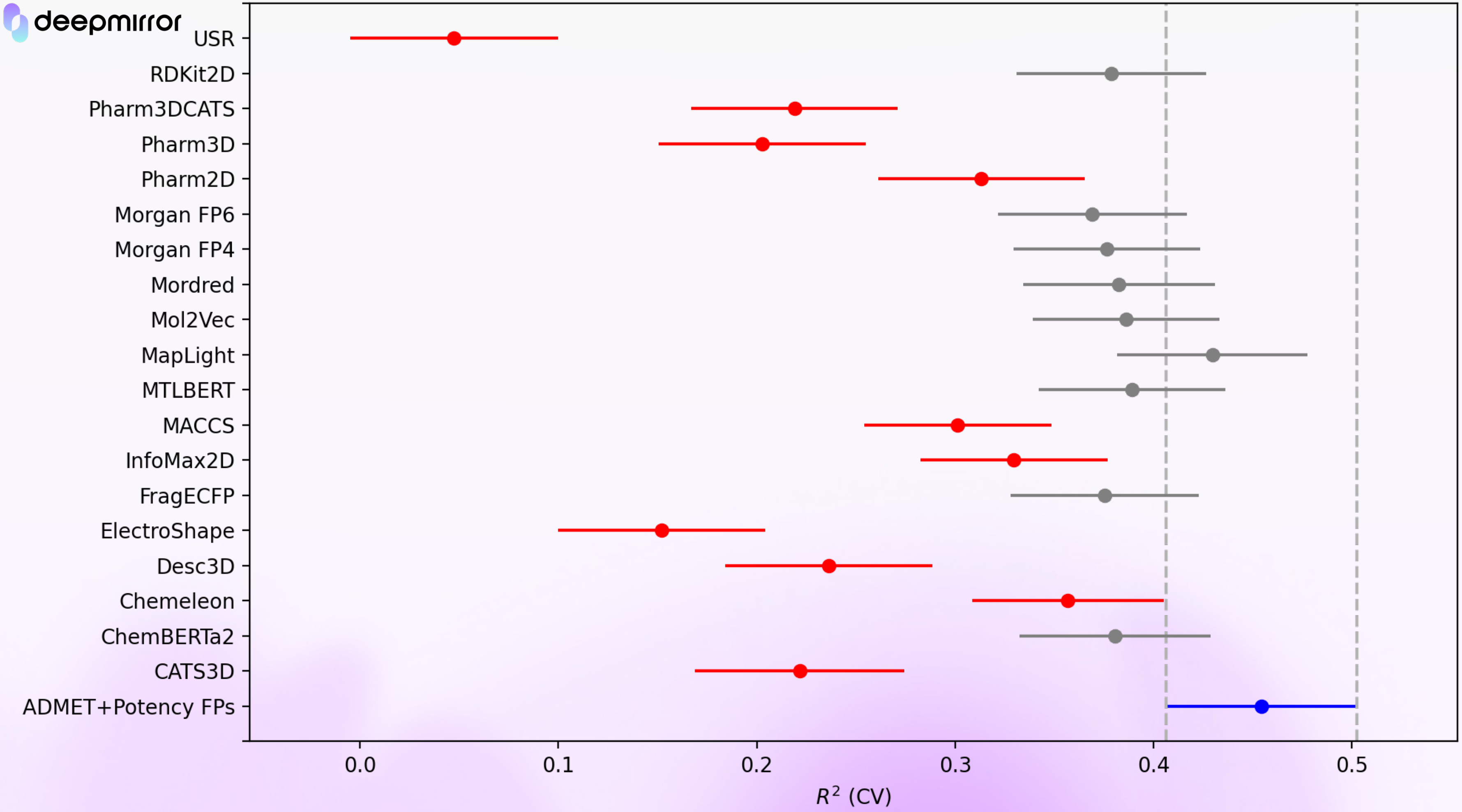
RDKit
Mordred
Morgan
…

CheMeleon
ChemBERTa
…

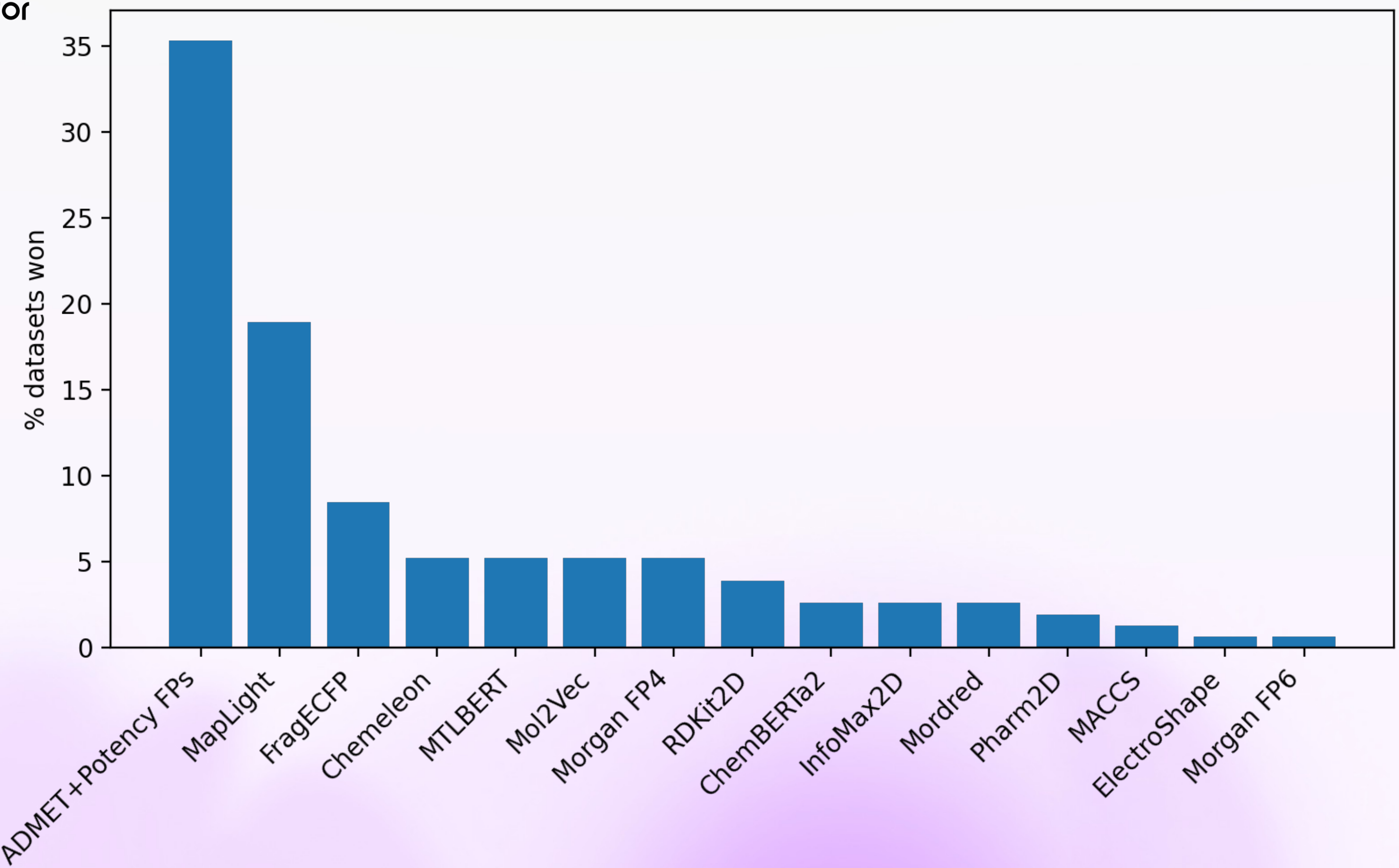
ADME/Potency
Fingerprints

Model

5x k-fold
cross
validation

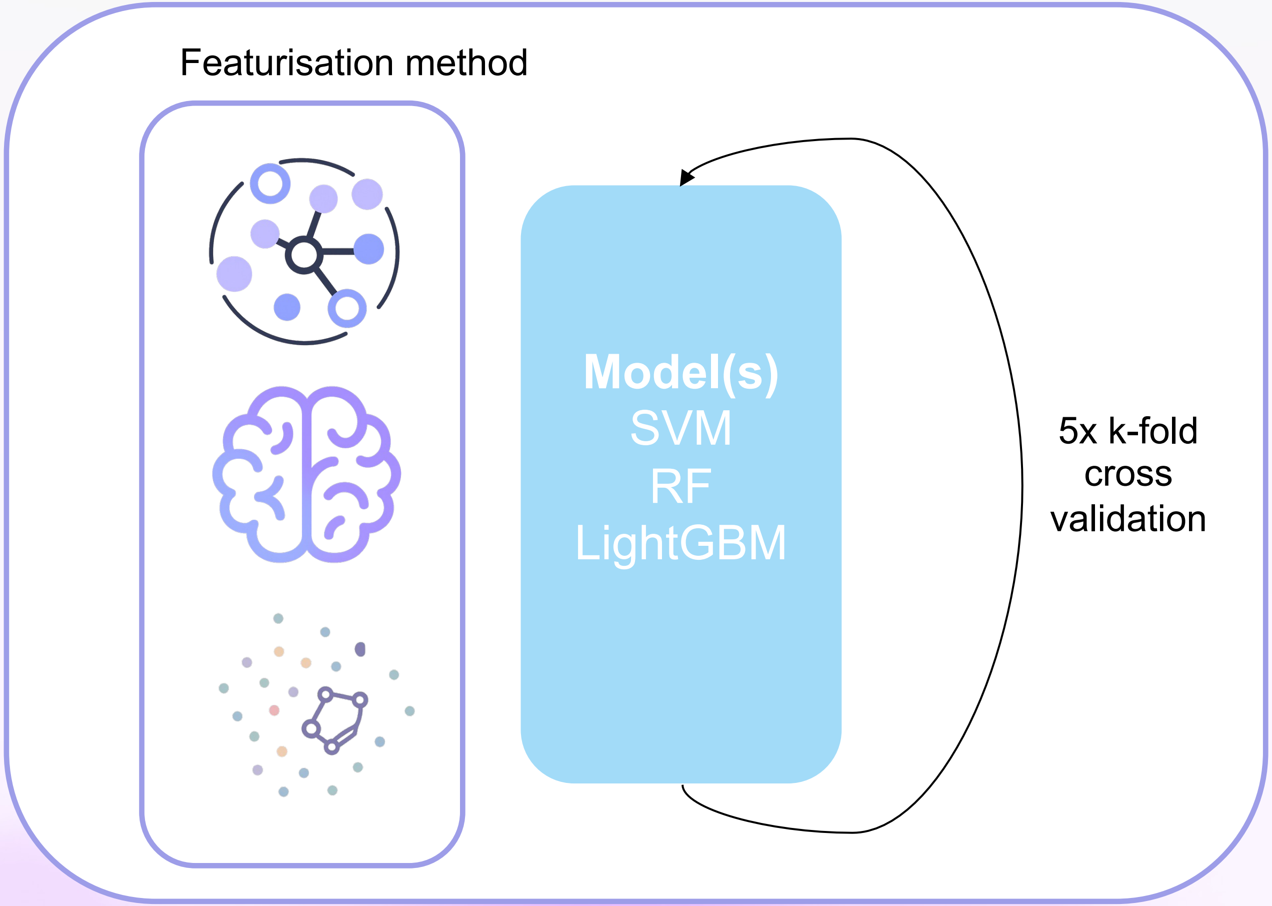

USR
RDKit2D
Pharm3DCATS
Pharm3D
Pharm2D
Morgan FP6
Morgan FP4
Mordred
Mol2Vec
MapLight
MTLBERT
MACCS
InfoMax2D
FragECFP
ElectroShape
Desc3D
Chemeleon
ChemBERTa2
CATS3D
ADMET+Potency FPs
0.0
0.1
0.2
0.3
0.4
0.5
R^2 (CV)

deepmirror


% datasets won
0
5
10
15
20
25
30
35
ADMET+Potency FPs
MapLight
FragECFP
Chemeleon
MTLBERT
Mol2Vec
Morgan FP4
RDKit2D
ChemBERTa2
InfoMax2D
Mordred
Pharm2D
MACCS
ElectroShape
Morgan FP6

Private programs

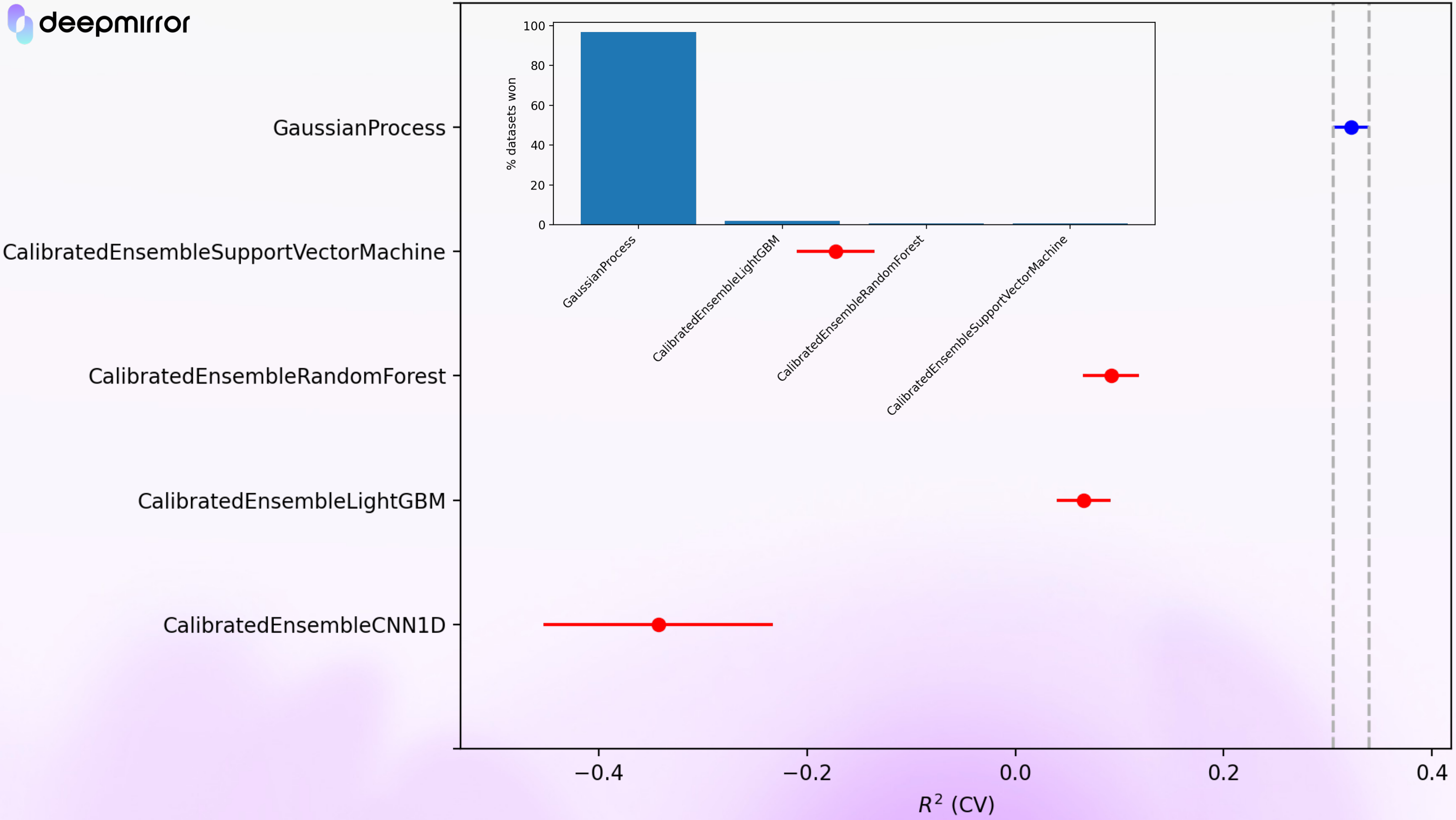
100s anonymized user-provided data.



Featurisation method

Model(s)
SVM
RF
LightGBM

5x k-fold cross validation


GaussianProcess
CalibratedEnsembleSupportVectorMachine
CalibratedEnsembleRandomForest
CalibratedEnsembleLightGBM
CalibratedEnsembleCNN1D
−0.4
−0.2
0.0
0.2
0.4
R^2 (CV)
% datasets won
0
20
40
60
80
100
GaussianProcess
CalibratedEnsembleLightGBM
CalibratedEnsembleRandomForest
CalibratedEnsembleSupportVectorMachine

deepmirror

- Use ADME and potency data to boost performance on programs.
- Across many programs, the best features depend on the specific context, there isn’t a single one that works best for all.
- ADME-based fingerprints are often selected.
- Classical physics-based methods can perform just as well.

deepmirror

Reach out: ryan@deepmirror.ai

Thank you:

Daniel Crusius
+ rest of team

Sign up

Thank you

AMPHISTA THERAPEUTICS

FERRARIAE UNIVERSITAS · EX LABORE FRUCTUS · 13 91

HES PHARMA

…and more

Guillaume Godin &
Jackson Burns