IMPERIAL

Scaffold Splits Overestimate Virtual Screening Performance

Qianrong Guo, Saiveth Hernandez-Hernandez, Pedro Ballester

Department of Bioengineering, Imperial College London, UK

p.ballester@imperial.ac.uk

x.com/pjballester

linkedin.com/in/pedroballester/

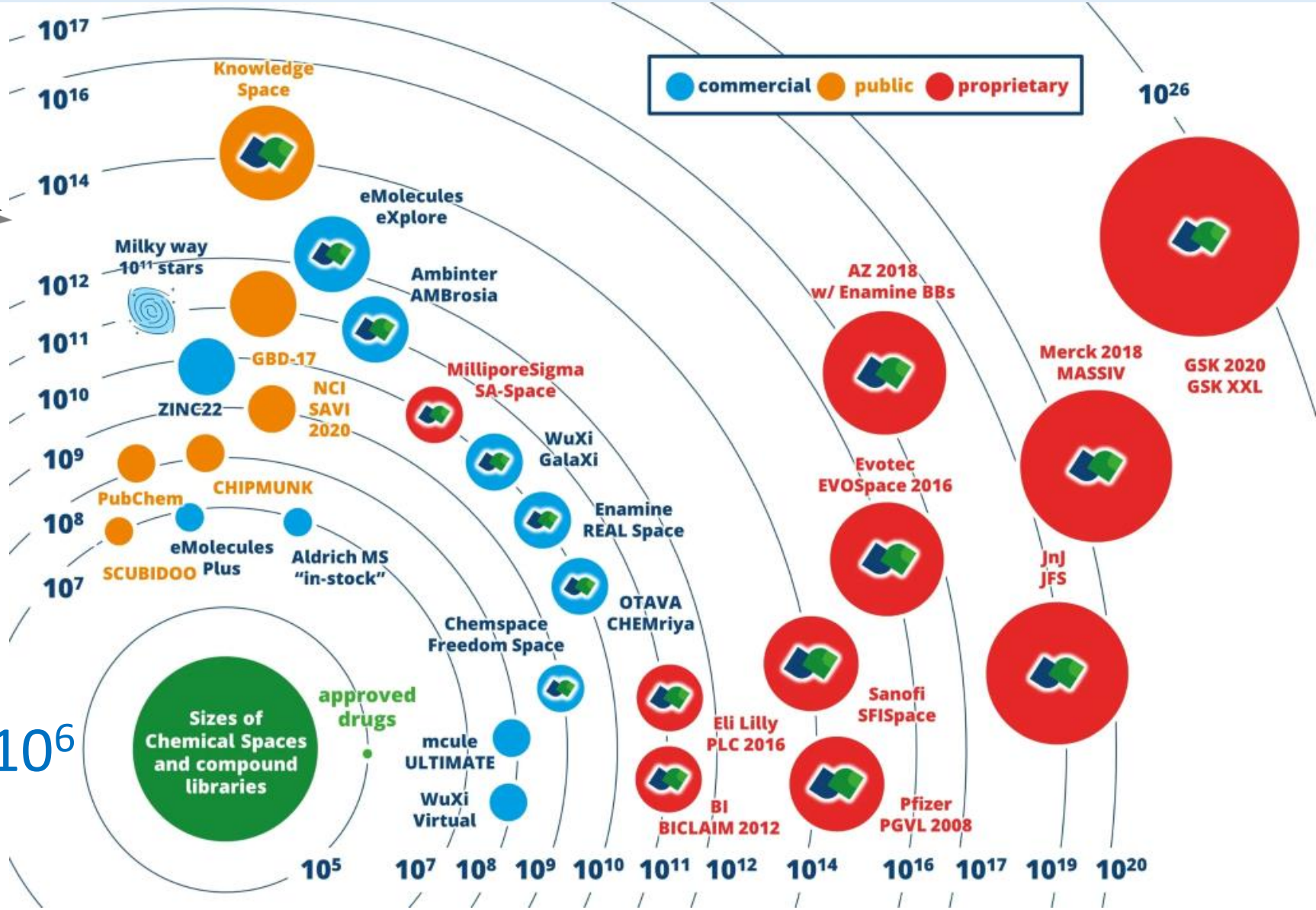
ballestergroup.github.io

ICANN24 (Lugano, Switzerland, Sep 2024)

Virtual Screening (VS): predicting dissimilar molecules

Almost every molecule to predict will be dissimilar to any in training set molecule

activity-labelled molecules that can be used for developing VS methods: at most 10^6



Source: https://www.biosolveit.de/chemical-spaces/

Also needed for other Molecular Property Prediction (MPP)

MPP is a rebranding of ligand-based QSAR/QSPR and structure-based BAP mostly

Their (unverified) claim: MPP models working well on the benchmark will also work well prospectively

MoleculeNet benchmarks

Category	Dataset	Data Type	Task Type	# Tasks	# Compounds	Rec - Split[a]	Rec - Metric[b]
Quantum Mechanics	QM7	SMILES, 3D coordinates	Regression	1	7160	Stratified	MAE
	QM7b	3D coordinates	Regression	14	7210	Random	MAE
	QM8	SMILES, 3D coordinates	Regression	12	21786	Random	MAE
	QM9	SMILES, 3D coordinates	Regression	12	133885	Random	MAE
Physical Chemistry	ESOL	SMILES	Regression	1	1128	Random	RMSE
	FreeSolv	SMILES	Regression	1	642	Random	RMSE
	Lipophilicity	SMILES	Regression	1	4200	Random	RMSE
Biophysics	PCBA	SMILES	Classification	128	437929	Random	PRC-AUC
	MUV	SMILES	Classification	17	93087	Random	PRC-AUC
	HIV	SMILES	Classification	1	41127	Scaffold	ROC-AUC
	PDBbind	SMILES, 3D coordinates	Regression	1	11908	Time	RMSE
	BACE	SMILES	Classification	1	1513	Scaffold	ROC-AUC
Physiology	BBBP	SMILES	Classification	1	2039	Scaffold	ROC-AUC
	Tox21	SMILES	Classification	12	7831	Random	ROC-AUC
	ToxCast	SMILES	Classification	617	8575	Random	ROC-AUC
	SIDER	SMILES	Classification	27	1427	Random	ROC-AUC
	ClinTox	SMILES	Classification	2	1478	Random	ROC-AUC

Scaffold split to evaluate on dissimilar molecules, i.e. to generate two sets with different biases (a.k.a. distribution shift)

0.841

0.737

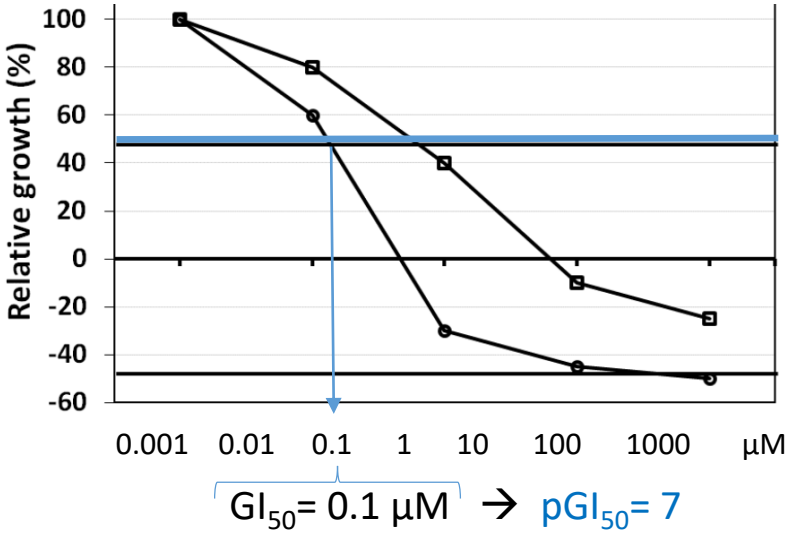
0.988

Near perfect classification!

Source: https://moleculenet.org/

Scaffold splits of the NCI-60 datasets

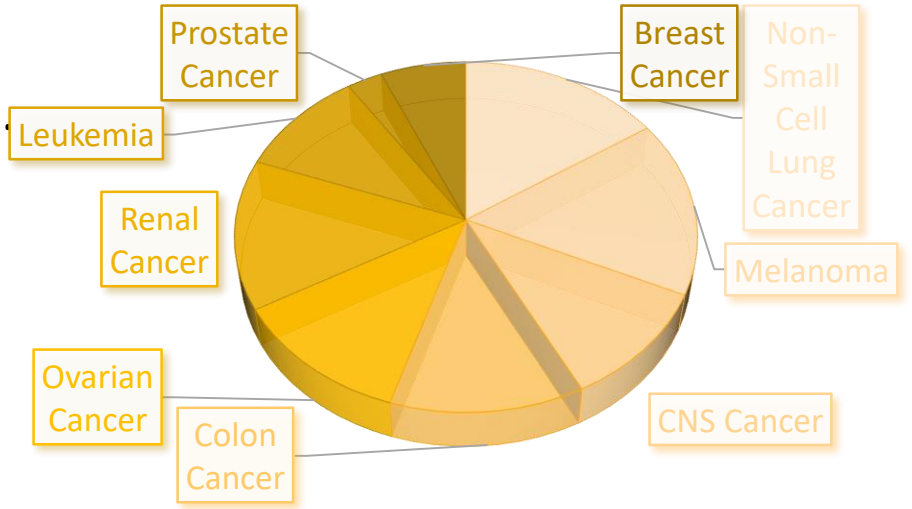
GI_{50}: molecule concentration inducing 50% inhibition of cancer cell line growth.



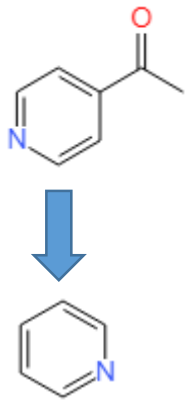
NIH NATIONAL CANCER INSTITUTE

Employed NCI-60 datasets:

- 60 cell lines (9 cancer types).
- 33,118 unique molecules.
- 1,764,938 pGI_{50} measurements (88.8% of this bioactivity matrix)



Bemis-Murcko scaffold: core structure of a molecule by removing its side chain atoms and focusing on its central ring systems and linkers.



- 33,118 molecules
- 14,212 scaffolds

Fold 1: 2031s, 4366m
Fold 2: 2031s, 4405m
Fold 3: 2030s, 5865m
Fold 4: 2030s, 4586m
Fold 5: 2030s, 4993m
Fold 6: 2030s, 4532m
Fold 7: 2030s, 4371m

Training set
Fold 1: 4366m
Fold 2: 4405m
Fold 3: 5865m
Fold 5: 4993m
Fold 6: 4532m
Fold 7: 4371m

Test set
Fold 4: 4586m

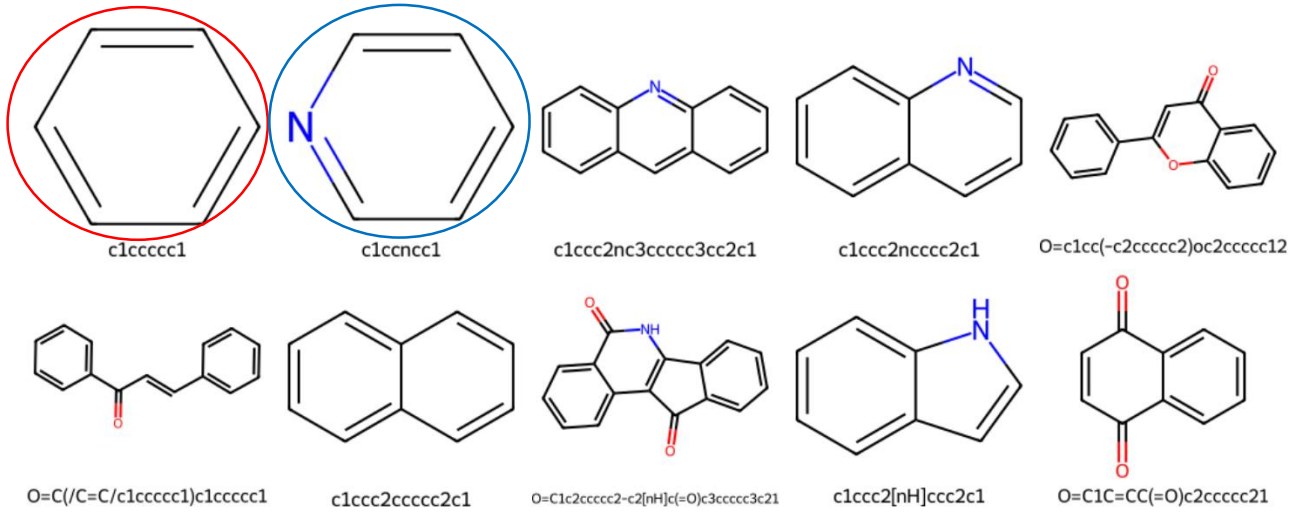
e.g. scaffold split for IGROV1: 27,256 molecules for training, 4,157 molecules for test

Modified from: https://pub.iapchem.org/ojs/index.php/admet/article/view/496
https://datagrok.ai/help/datagrok/solutions/domains/chem/scripts/murcko-scaffolds

https://dtp.cancer.gov/discovery_development/nci-60/cell_list.htm

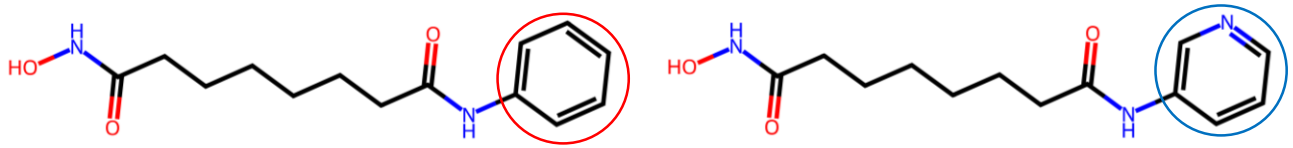
Scaffold split: unrealistically high train-test similarities!

Top 10 most-frequent scaffolds among molecules tested on TK-10 (a renal cancer cell line)



Scaffold split will often permit high similarities between training and test molecules (scaffolds different in a single atom, one scaffold containing the other) that rarely occur prospectively (massive diversity of screening libraries used as real-world test set)

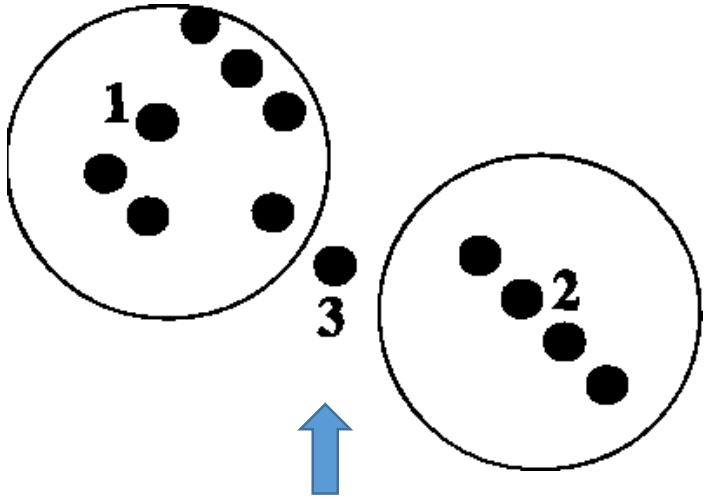
Scaffold split can place the molecule on the left in the training set and that on the right in the test set!



C(CCCCCC(=O)NO)C(=O)Nc1ccccc1 c1cc(cnc1)NC(=O)CCCCCCC(=O)NO

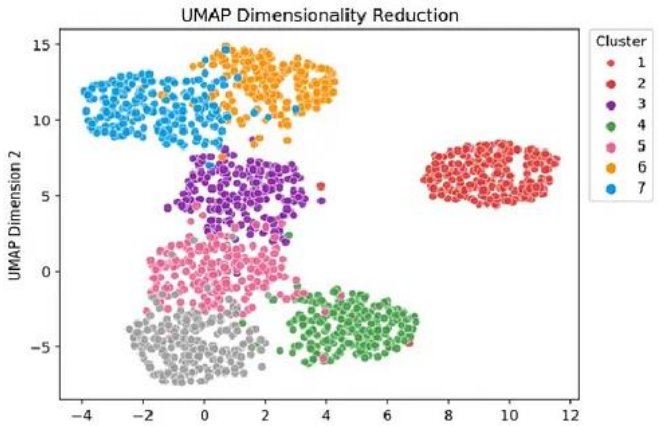
Butina and UMAP clustering splits

Butina clustering: centroids are selected as the molecules with more neighbours. Then each cluster is formed with molecules with similarity > cutoff=0.9 (found optimal) to its centroid.



33,118 molecules x 263 features

UMAP clustering: UMAP learns the manifold structure of the data in a topology-preserving manner assuming k clusters. Here outputs a two-dimensional embedding. K= 7 was optimal.



Butina clustering split:

- 7 folds as UMAP and scaffold.
- Butina clusters distributed across folds by their decreasing size (same-size folds)

UMAP clustering split:

- 7 folds, fold = UMAP cluster

Butina: https://pubs.acs.org/doi/full/10.1021/ci9803381

UMAP: https://www.mdpi.com/2218-273X/13/3/498

Linear Regression (LR) and Random Forest (RF)

Features

263 pre-calculated features **X** per molecule:

- 256 binary (MorganFpt,256 bits,radius 2)
- 7 real-valued (physico-chemical)

Package	Function
AllChem.GetMorganFingerprintAsBitVect	Generate the Morgan Fingerprints [9] for the molecules.
rdMolDescriptors.CalcTPSA	Calculate the area of the total polar surface.
rdMolDescriptors.CalcExactMolWt	Calculate the molecular weight.
rdMolDescriptors.CalcCrippenDescriptors	Calculate the Crippen-Wildman partition coefficient ($logP$) parameters [10].
rdMolDescriptors.CalcNumAliphaticRings	The number of aliphatic rings.
rdMolDescriptors.CalcNumAromaticRings	The number of aromatic rings.
rdMolDescriptors.CalcNumHBA	The number of hydrogen bond acceptors.
rdMolDescriptors.CalcNumHBD	The number of hydrongen bond doner.

LR

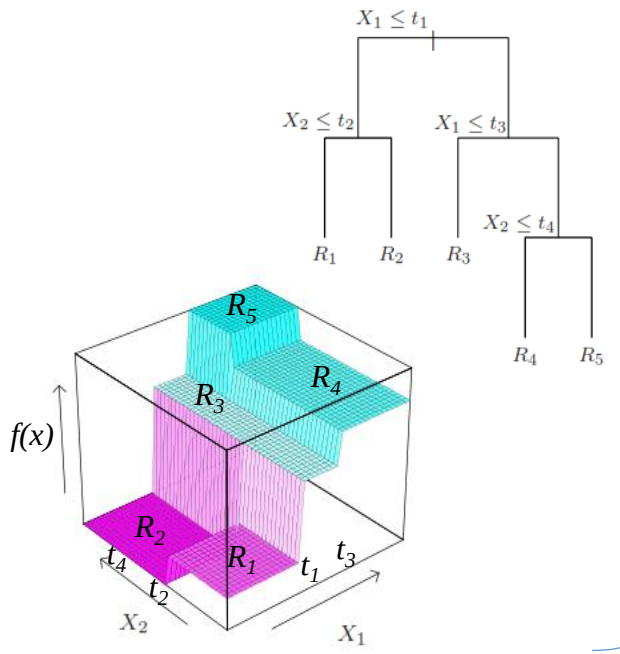
$$y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

RF

One trained regression tree

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

$$\hat{c}_m = \text{ave}(y_i | x_i \in R_m)$$



Random Forest of regression trees

Algorithm 15.1 *Random Forest for Regression or Classification.*

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

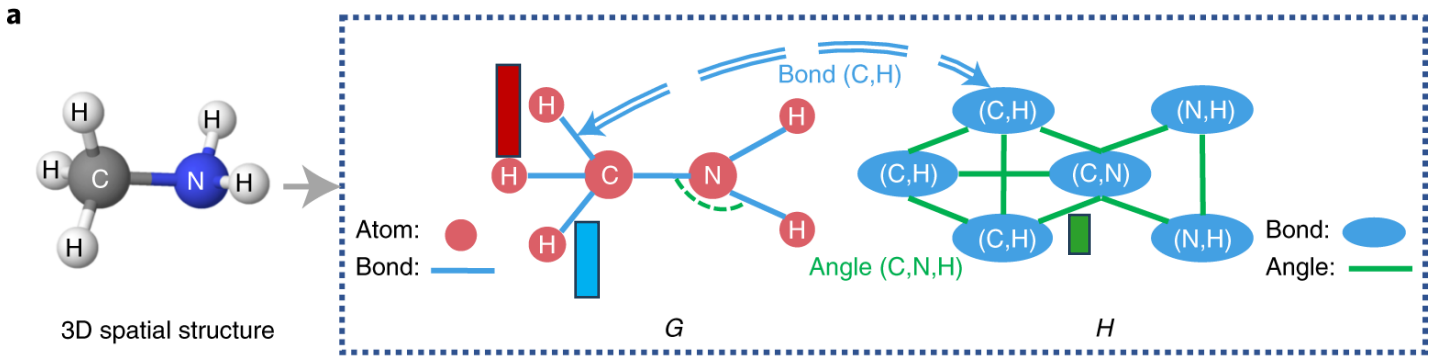
To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\text{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

Source: https://www.statlearning.com/

Geometry-Enhanced molecular representation learning Method (GEM)



Each molecule, two node-edge graphs: G (atom-bond) and H (bond–angle)

Feature type	Feature	Description	Size
atom	atom type	type of atom (e.g., C, N, O), by atomic number (one-hot)	119
	aromaticity	whether the atom is part of an aromatic system (one-hot)	2
	formal charge	electrical charge (one-hot)	16
	chirality tag	CW, CCW, unspecified or other (ont-hot)	4
	degree	number of covalent bonds (one-hot)	11
	number of hydrogens	number of bonded hydrogen atoms (one-hot)	9
	hybridization	sp, sp^2, sp^3, sp^3d, or sp^3d^2 (one-hot)	5
bond	bond dir	begin dash, begin wedge, etc. (one-hot)	7
	bond type	single, double, triple or aromatic (one-hot)	4
	in ring	whether the bond is part of a ring (one-hot)	2
	bond length	bond length (float)	-
bond angle	bond angle	bond angle (float)	-

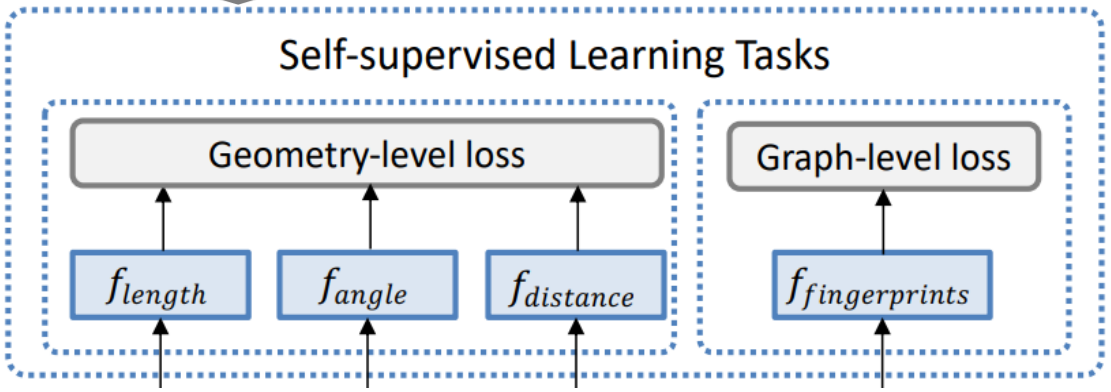
Input features for atoms, bonds and bond angles

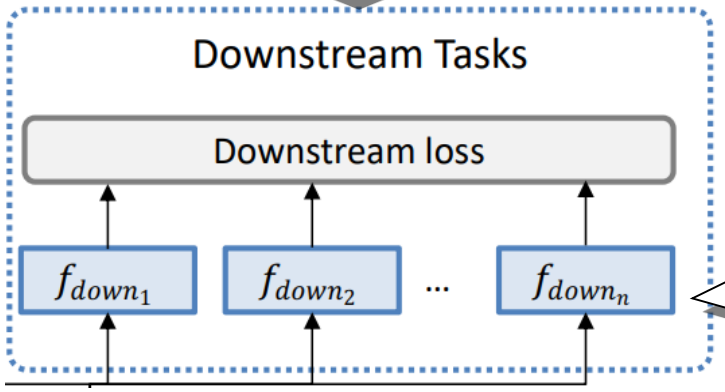
Figure Source: Fang X, Liu L, Lei J, et al. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 2022, 4(2): 127-134.

GEM pretraining and training

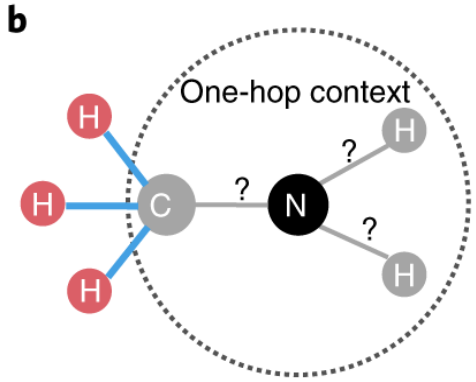
GEM pretraining by the authors using the 3D conformers of 20 million unlabelled molecules from ZINC15

Pretrained GEM fine-tuning by us using the same labelled training sets as LR or RF

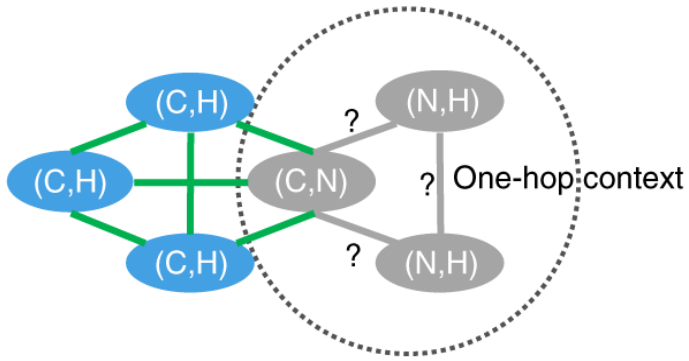




n=1: only predicting pGI_{50}



Predict bond lengths



Predict bond angles

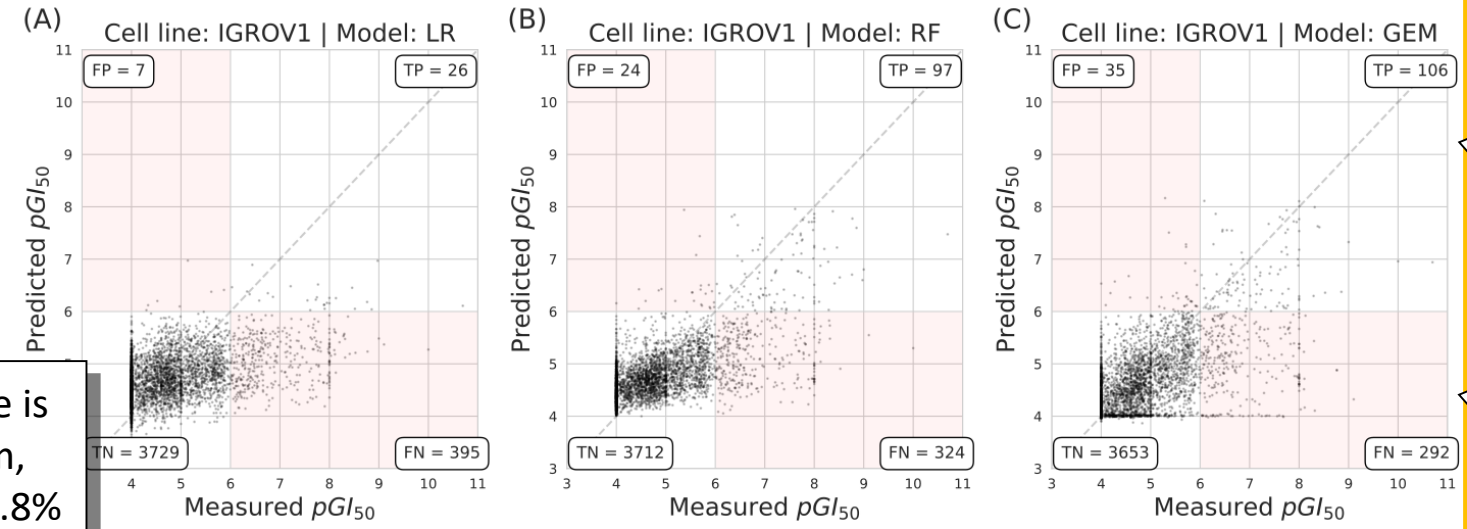
	C	H	H	H	N	H	H
C		?					
H							
H				?			
H							
N						?	
H		?					
H							

Predict atomic distances

Figure Source: Fang X, Liu L, Lei J, et al. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 2022, 4(2): 127-134.

Results: 1 left-out fold x 1 CL x 1 seed x 3 algorithms

Scaffold Split



Hit Rate	MCC	ROC AUC	RMSE
78.8 %	0.204	0.530	0.849

Hit Rate	MCC	ROC AUC	RMSE
80.2 %	0.402	0.612	0.729

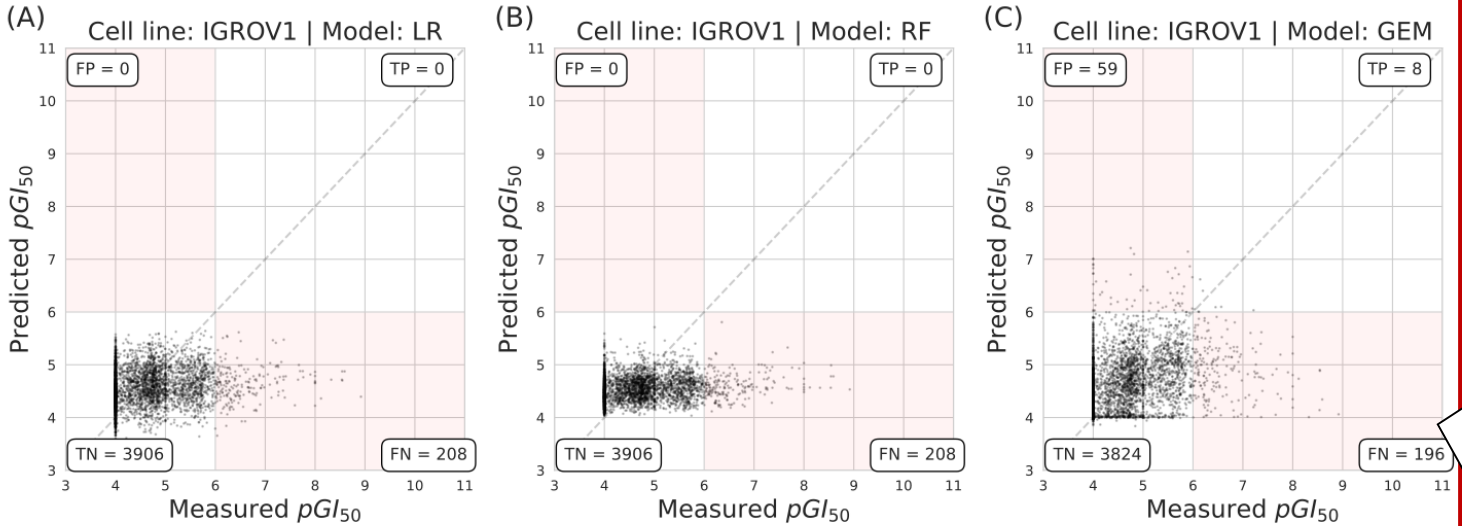
Hit Rate	MCC	ROC AUC	RMSE
75.2 %	0.417	0.628	0.738

ROC AUC value is almost random, but hit rate 78.8%

regression-classification evaluation: active if pGI_{50}>6

Highest hit rate 80.2% → RF selected for prospective use

UMAP Split



Hit Rate	MCC	ROC AUC	RMSE
0.0 %	0.000	0.500	0.766

Hit Rate	MCC	ROC AUC	RMSE
0.0 %	0.000	0.500	0.737

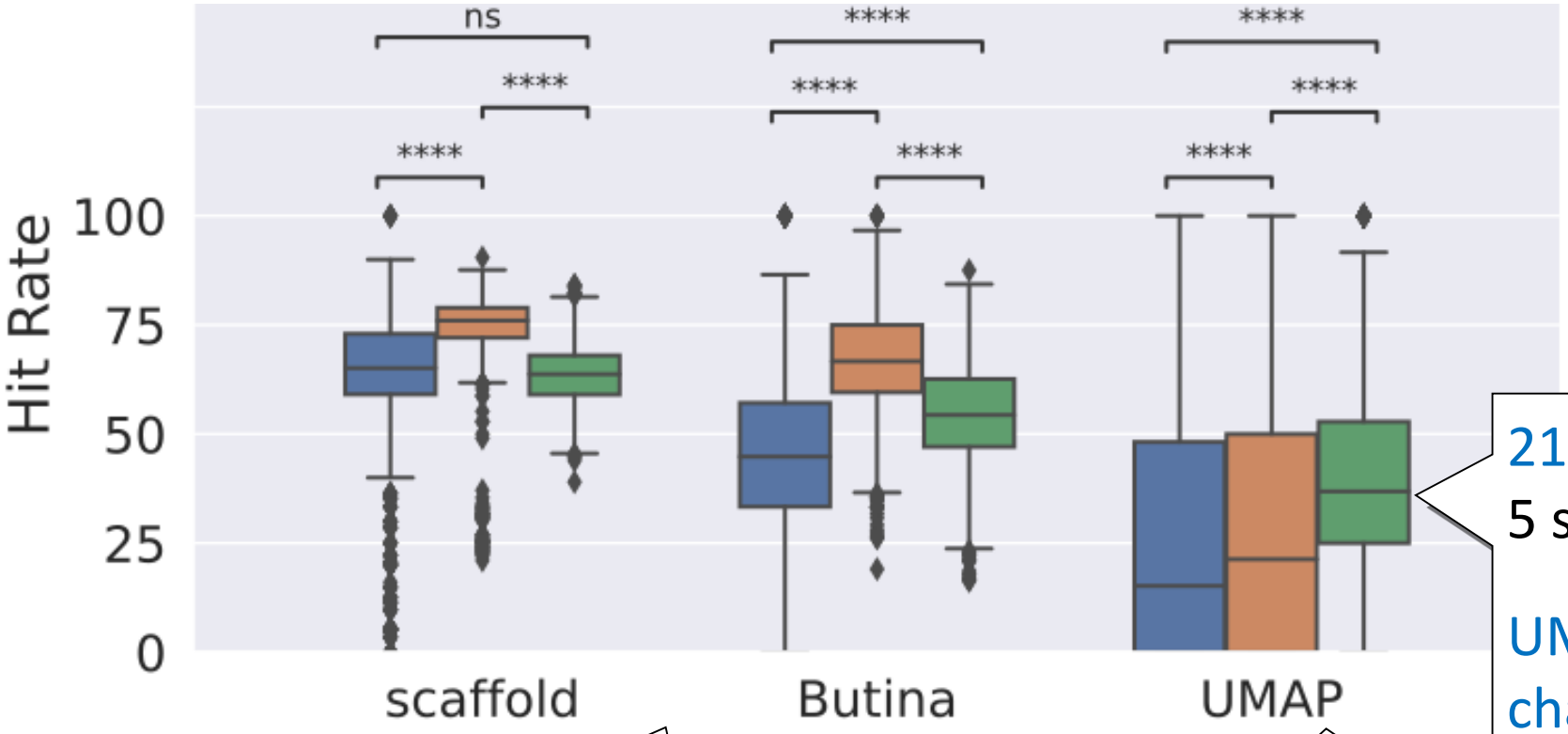
Hit Rate	MCC	ROC AUC	RMSE
11.9 %	0.041	0.512	0.729

RMSE is not helpful either: e.g.
LR SS (0.849) vs UMAP (0.766)
but hit rate
LR SS (78.8%) vs UMAP (0%)

RF now a 0% hit rate! (LR too) vs GEM stills finding actives

NB: GEM ↑TP in each split

Hit rate in left-out fold: 3 algorithms x 60 cell lines



Algorithms

- LR
- RF
- GEM

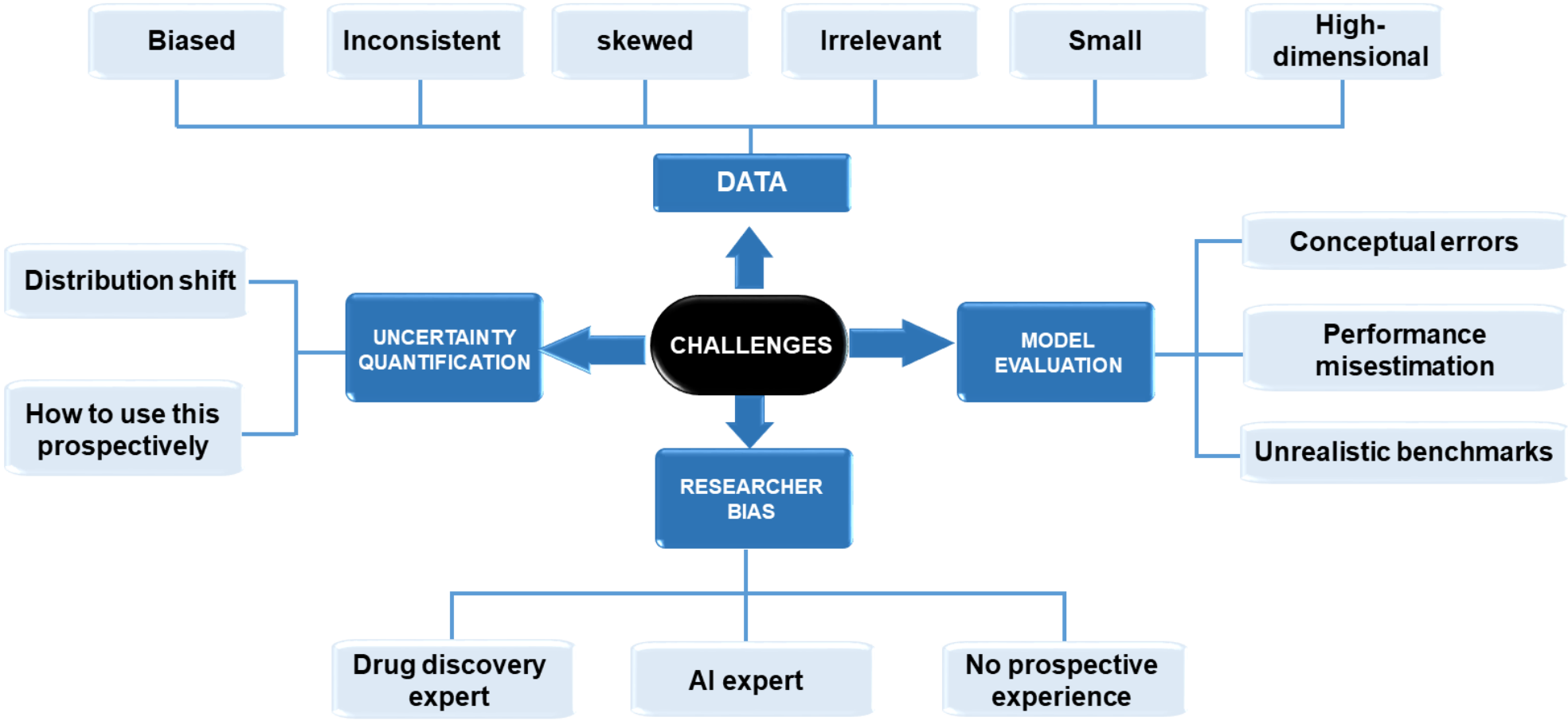
2100 hit rates per boxplot:
5 seeds x 7 folds x 60 cell lines

UMAP clustering split is most challenging for all algorithms

If we only used any of these splits, RF would be selected for prospective use

Using the more realistic UMAP split, GEM would be selected instead

Biased datasets: far from being the only MPP challenge



Ghislat et al. (2024) "Data-centric challenges with the application and adoption of artificial intelligence for drug discovery" *Expert Opinion on Drug Discovery*. https://arxiv.org/abs/2407.05150

Conclusions

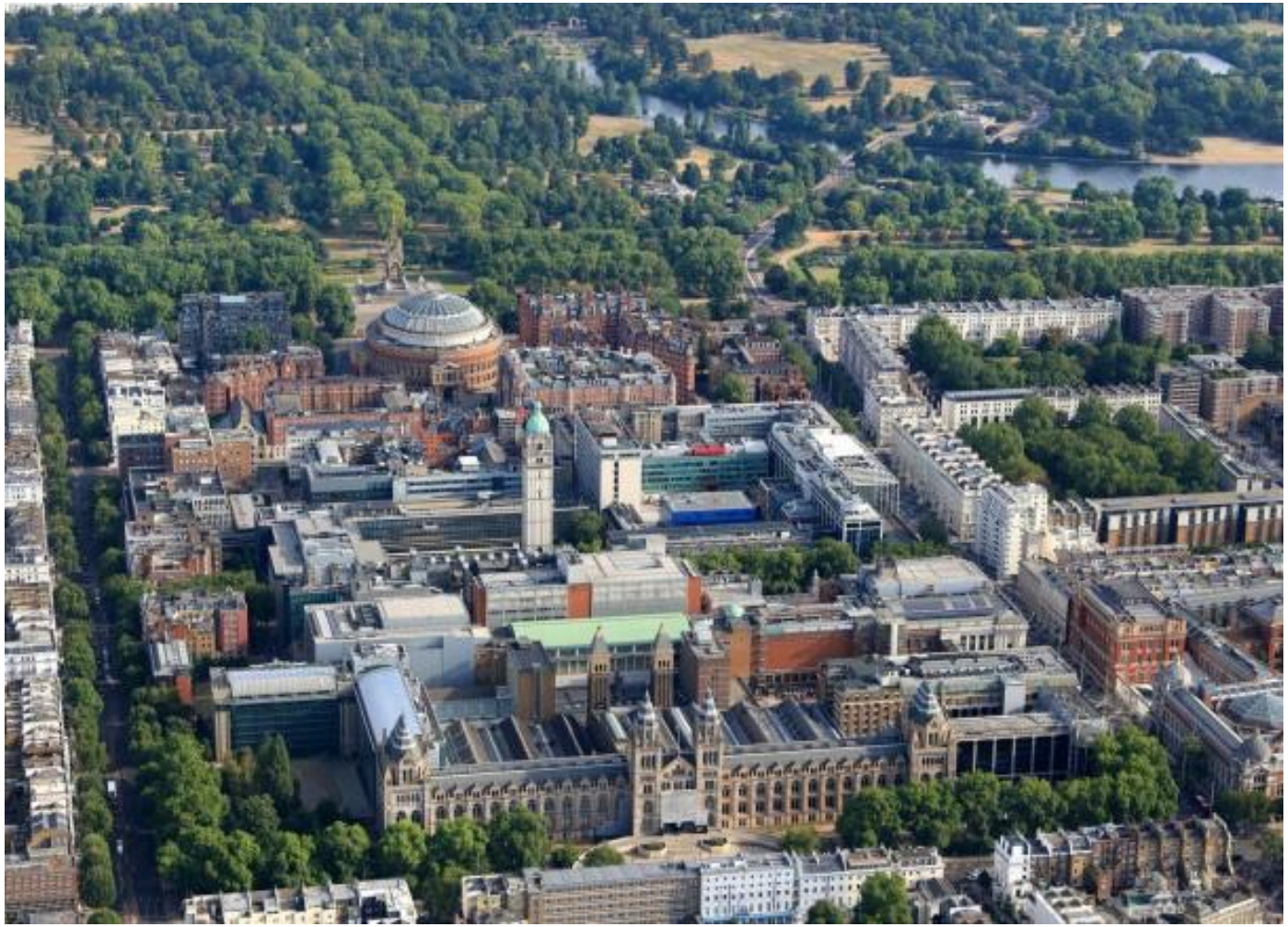
1. Scaffold splits do not generate realistic distribution shifts because similar molecules often have different scaffolds
2. Clustering splits ensure lower similarities between training and test molecules → more challenging than scaffold splits
3. UMAP clustering splits are substantially harder than Butina clustering splits for all the supervised learning algorithms
4. As training-test similarities do not depend on the label to predict, scaffold splits are also likely to distort model selection in similar molecular property prediction problems

Do you know anyone looking for a postdoc in this area?

Postdoc1 on AI for structure-based virtual screening
Postdoc2 on generative AI for de novo drug design

If interested, please email me
p.ballester@imperial.ac.uk
with a CV with publications
and a motivation letter.

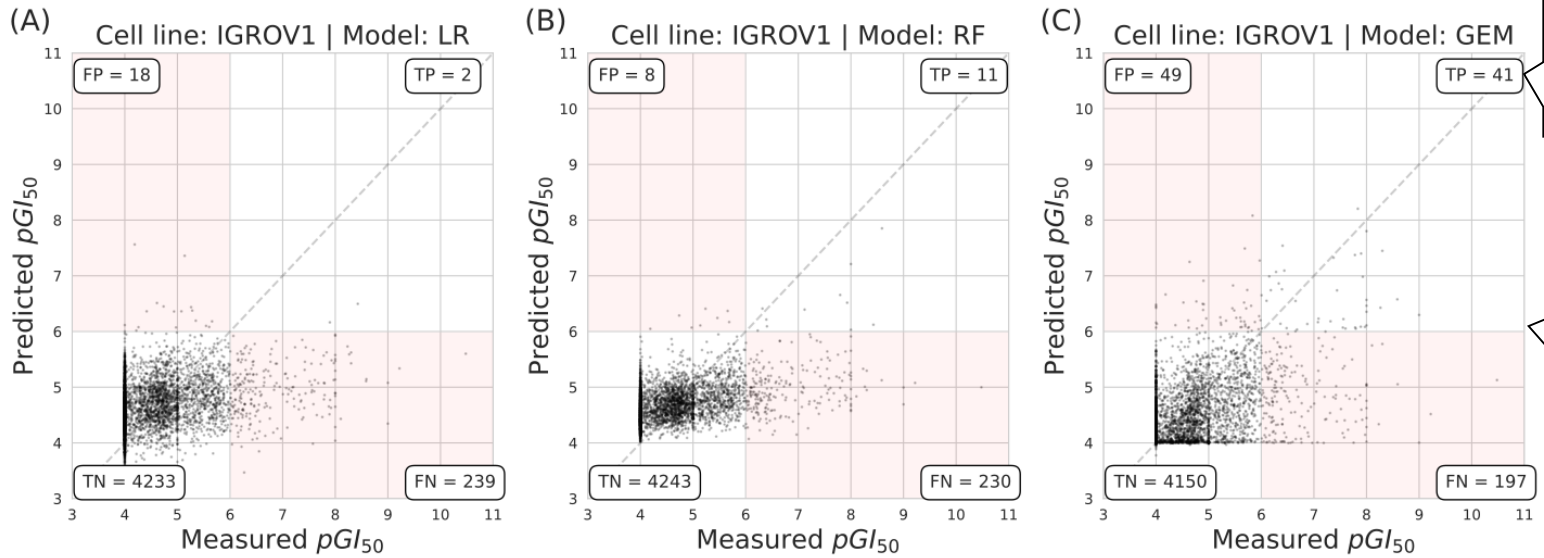
Q & A



Results: 1 left-out fold x 1 CL x 1 seed x 3 algorithms

Highest hit rate 57.9% → RF selected for prospective use

Butina split



Hit Rate	MCC	ROC AUC	RMSE
10.0 %	0.014	0.502	0.780

Hit Rate	MCC	ROC AUC	RMSE
57.9 %	0.152	0.522	0.705

Hit Rate	MCC	ROC AUC	RMSE
45.6 %	0.257	0.580	0.695

NB: GEM highest TP in each split

regression-classification evaluation: active if $pGI_{50}>6$

RMSE also useless: e.g.
RF SS (0.849) vs Butina (0.780)
but hit rate
RF SS (78.8%) vs Butina (10%)

ROC AUC useless:
almost random,
but hit rate 57.9%