

Molecule property prediction, federated learning and uncertainty estimation



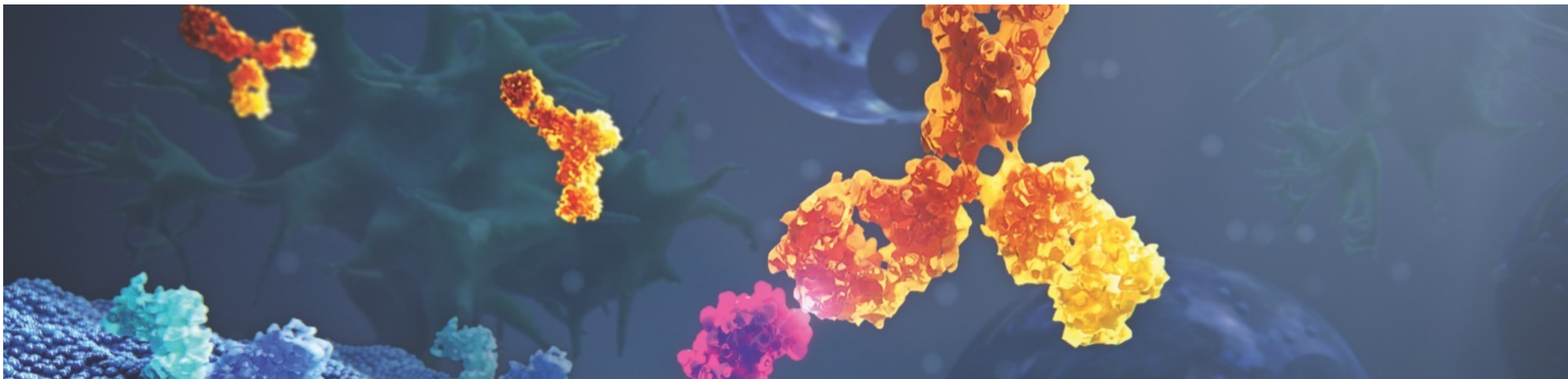
Lewis Mervin

AstraZeneca

Molecular AI, Discovery Sciences, AstraZeneca R&D, Cambridge, UK

AIDD School Leuven

20 October 2022



Contents

- About me
- Molecular AI (MAI) group at AZ
- My general research interests:
 - Molecule prop. prediction & the DMTA cycle
 - How AZ approach *de novo* design
- Current work/research relating to AIDD:
 - How to get the most out of Federated learning (FL)?
 - General Multi-task learning (MTL) outlook
 - How to approach uncertainty quantification going forward?

Who am I?

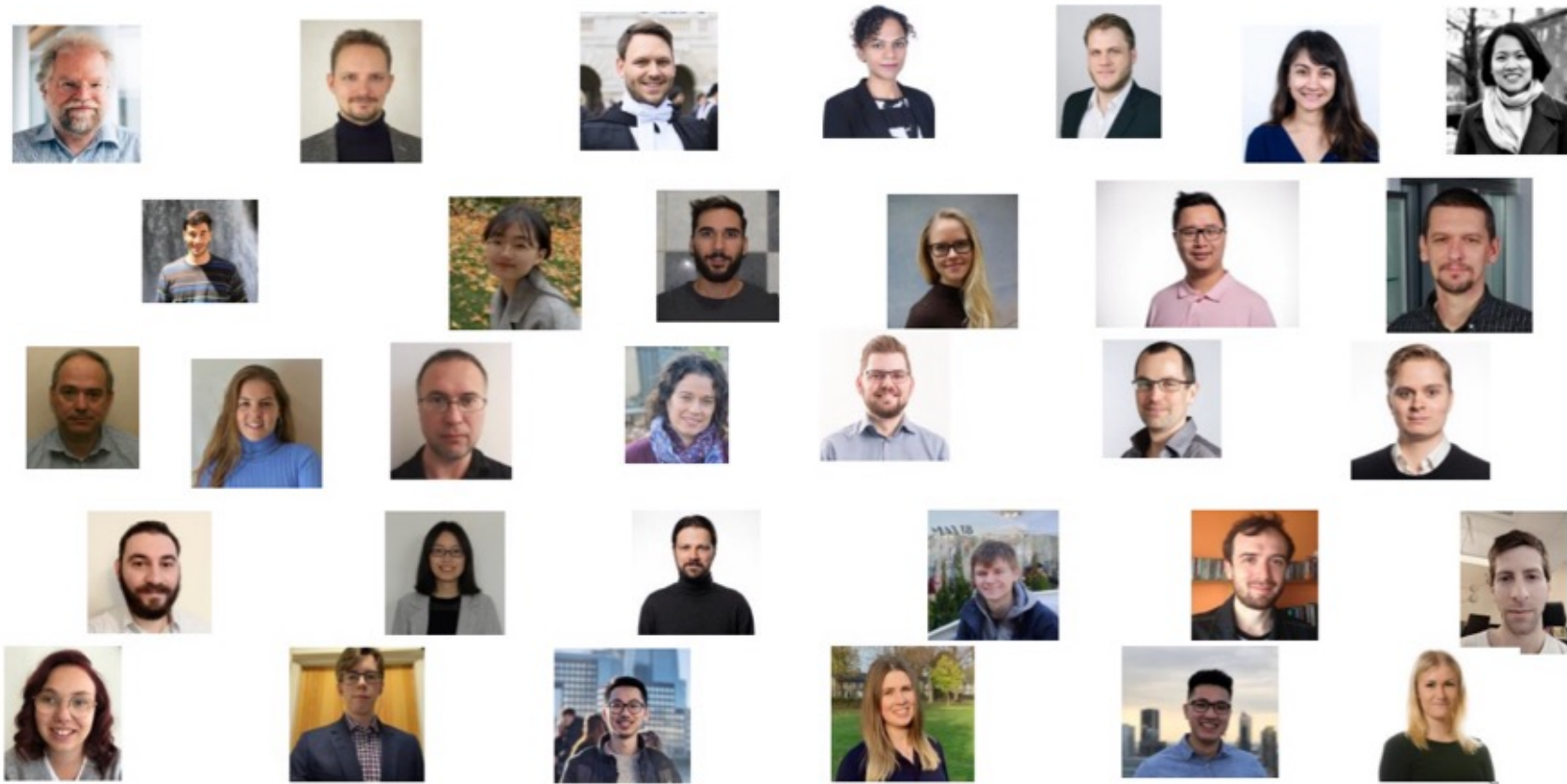
- Data scientist / cheminformatician at AstraZeneca
- Based in UK, part of the Molecular AI Team (mostly based in Sweden)

Career path:

- Industrial CASE Masters/PhD with Uni. Cambridge & AstraZeneca, UK
- EU PostDoc: Systems biology of alcohol addiction (Sybil-AA)
- Re-joined AZ in 2019
- Interests: molecule prop. prediction & comp. methods to improve hit discovery/productivity of drug design



Molecular AI (MAI) group



- Molecular AI group (~20 people)
- General focus: Application of AI to the drug design process
- Broad range of backgrounds, e.g. chemists/biologists/pharmacologists/comp. sci.

Science Molecular AI @AZ

ACS
central
science

Research Article

Cite This: ACS Cent. Sci. 2018, 4, 120–131

Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks

RESEARCH

Molecular De-Novo Design through Deep Reinforcement Learning

Marcus Olivecrona*, Thomas Blaschke†, Ola Engkvist† and Hongming Chen†

RESEARCH ARTICLE Open Access

Exploring the GDB-13 chemical space using deep generative models

Josep Arús-Pous^{1,3*}, Thomas Blaschke^{1,4}, Silas Ulander², Jean-Louis Reymond³, Hongming Chen¹ and Ola Engkvist¹

JCIM
JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

pubs.acs.org/jcim

Application

REINVENT 2.0: An AI Tool for De Novo Drug Design

Thomas Blaschke, Josep Arús-Pous, Hongming Chen, Christian Margreitter, Christian Tyrchan, Ola Engkvist, Kostas Papadopoulos, and Atanas Patronov*

Journal of
Medicinal
Chemistry

This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.

pubs.acs.org/jmc

Article

“Ring Breaker”: Neural Network Driven Synthesis Prediction of the Ring System Chemical Space

Amol Thakkar*, Nidhal Selmi, Jean-Louis Reymond, Ola Engkvist, and Esben Jannik Bjerrum*

Chemical
Science

ROYAL SOCIETY
OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Sci.*, 2021, 12, 3339

All publication charges for this article have been paid for by the Royal Society of Chemistry

Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning†

Amol Thakkar, Veronika Chadimová, Esben Jannik Bjerrum, Ola Engkvist and Jean-Louis Reymond

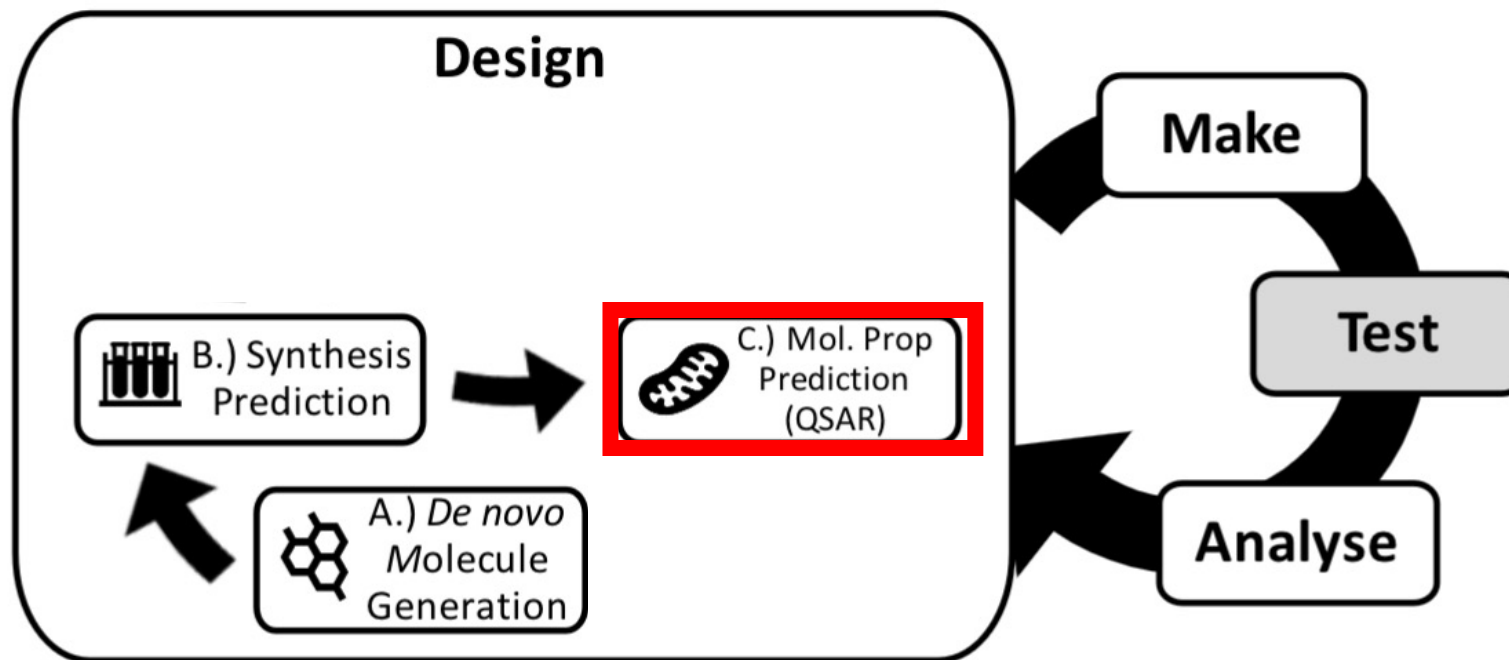
SOFTWARE

Open Access

AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning

Samuel Genheden^{1*}, Amol Thakkar^{1,2}, Veronika Chadimová¹, Jean-Louis Reymond², Ola Engkvist¹ and Esben Bjerrum^{1*}

Where MAI impact the DMTA cycle

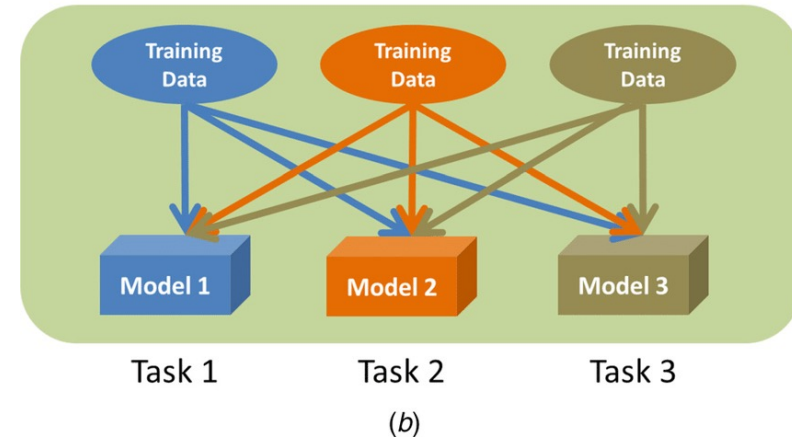
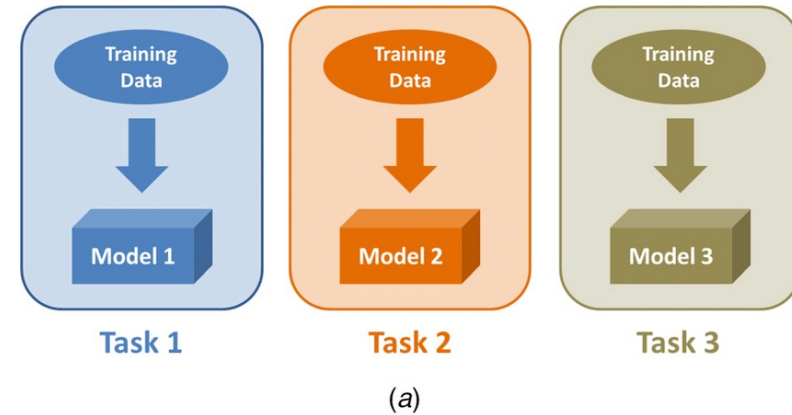


Multi-task learning (MTL) & Federated Learning (FL)

& how to get the most out of them?

What is multi-task learning (MTL)?

- a.) Single-task (ST): one model trained to predict **one task**
 - one model optimised until performance no longer increases
- b.) Multi-task (MT/MTL): training one model to predict **multiple tasks**
 - one model optimising more than one loss function at once
 - enables representations to be shared between trained tasks
 - training signals of related tasks shared between all tasks
- MTL uses the knowledge learnt during training one task to **reduce the loss of other tasks** included in training



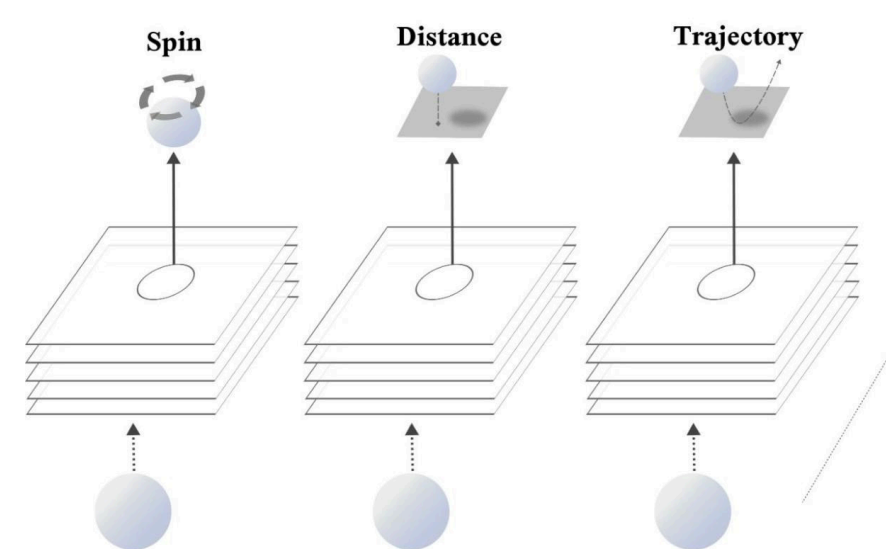
“Even if you are only optimizing one loss as is the typical case, chances are there is an auxiliary task that will help you improve upon your main task” [Caruana, 1998]

Real world examples

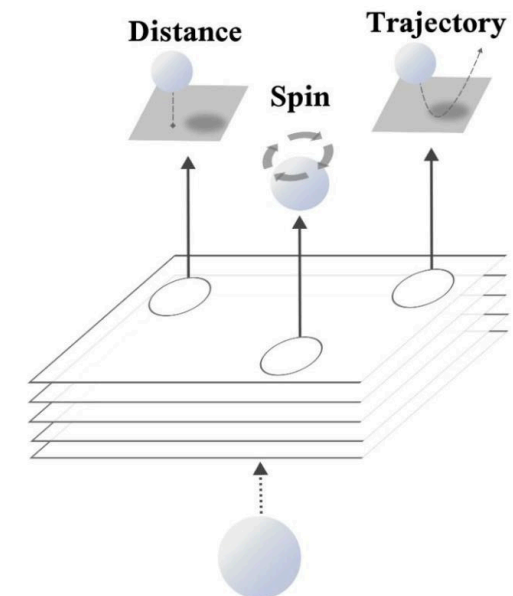
- The Karate Kid (1984)
 - Mr Miyagi teaches the karate kid seemingly unrelated tasks such as **sanding the floor** and **waxing a car**
 - in hindsight, these turn out to equip him with invaluable skills **relevant for karate**
- Predicting ping-pong ball return (right):
 - requires *distance*, *spin*, and *trajectory* of the ping-pong
 - each is unique - predicting *spin* is fundamentally distinct from *location* - but **improving the reasoning of both** will help better prediction of e.g. *trajectory*

Predicting ping-pong ball return

Three Single Task Models



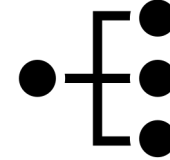
One Multi-Task Model



Why does MTL work?

- **Implicit data augmentation**

- Missing data/sparsity mitigated by augmentation
- Augmented tasks have different noise patterns
- ST models risk overfitting vs. MTL which averages noise patterns



- **Attention focusing**

- Model attention focused to relevant features
- Related tasks give extra evidence for feature [ir]relevance



- **Eavesdropping**

- Some tasks are difficult to learn (complex interactions with features)
- Some features could impede learning certain tasks
- Learn relevant features for difficult tasks via easier tasks



- **Representation bias**

- Biases models to prefer representations many tasks prefer
- Existing well-performing configurations are likely to perform well for novel tasks



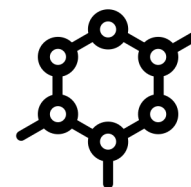
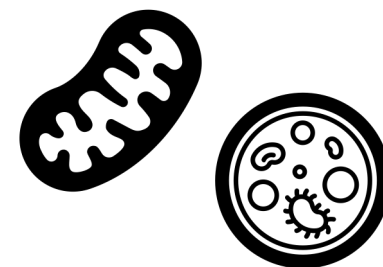
- **Regularization**

- Acts as a regulariser introducing an inductive bias - reduces Rademacher complexity



Why MTL for molecule property prediction?

- Data being modelled is **heavily biased**^[1]
 - Due to the amount, degree of diversity and distribution of data points
 - ST-models often incapable of arriving at realistic probability estimates across the many tasks
- Behaviour/characteristics of **biological/assay task space**
 - Protein activity profiles often co-correlated
 - Protein family, homologous/orthologous protein, common off-targets
 - Biological properties linked with ADME, PK/PD, physiochemical properties
 - e.g. lysosomotropism linked with lipophilicity
 - e.g. thermodynamic solubility and kinetic water solubility
 - Overlap between primary/orthogonal/artefact assays & screening cascades
 - Overlap of experimental machinery/protocols/procedures
- Behaviour/characteristics of **chemical space**
 - Overlap of compound decks/screening libraries
 - Information transfer from standardised compound sets

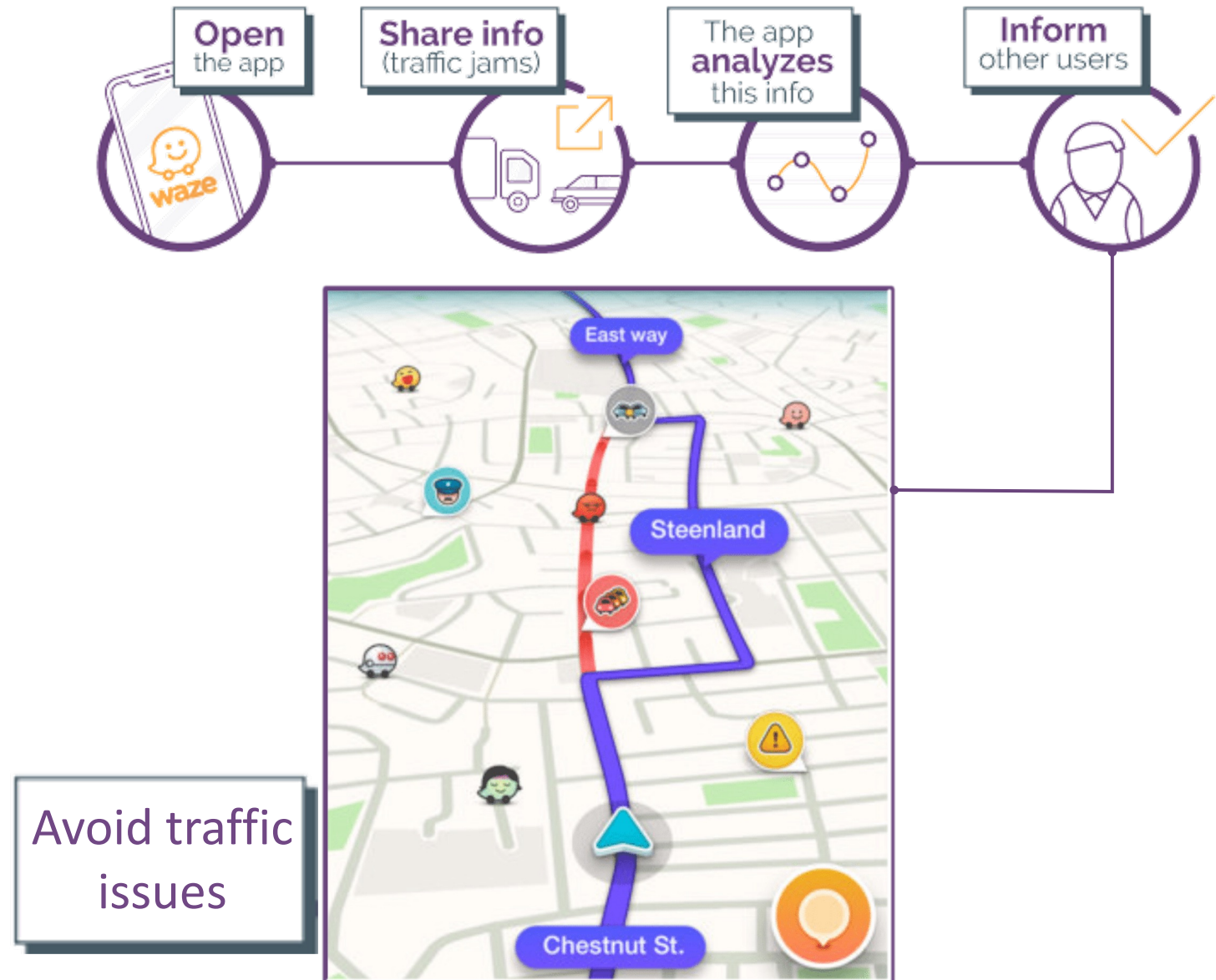


[1] Lewis H.Mervin *et al*, Uncertainty quantification in drug design (2021), Drug Discovery Today

Application toward Federated Learning (FL)

What is Federated Learning (FL)?

HOW DOES WAZE WORK



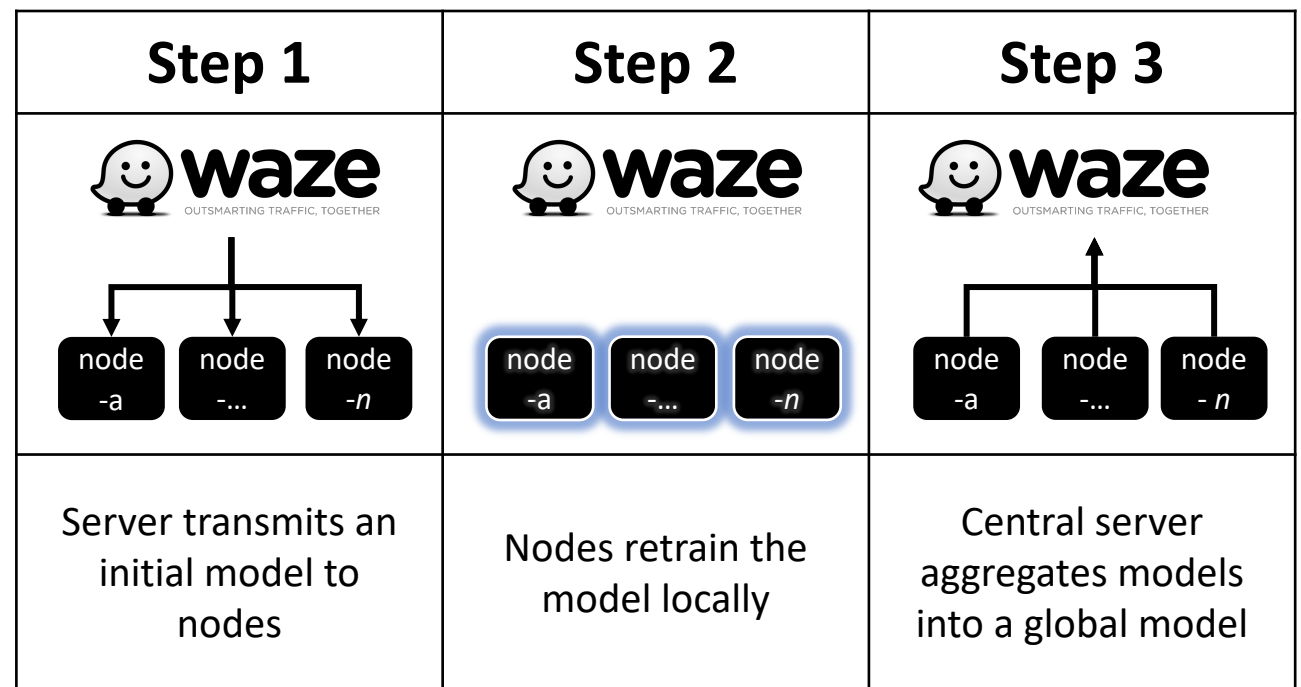
Collaboratively learn
a **shared prediction
model**

What is *privacy preserving* FL?

Two examples of federated architectures shown (not exhaustive)

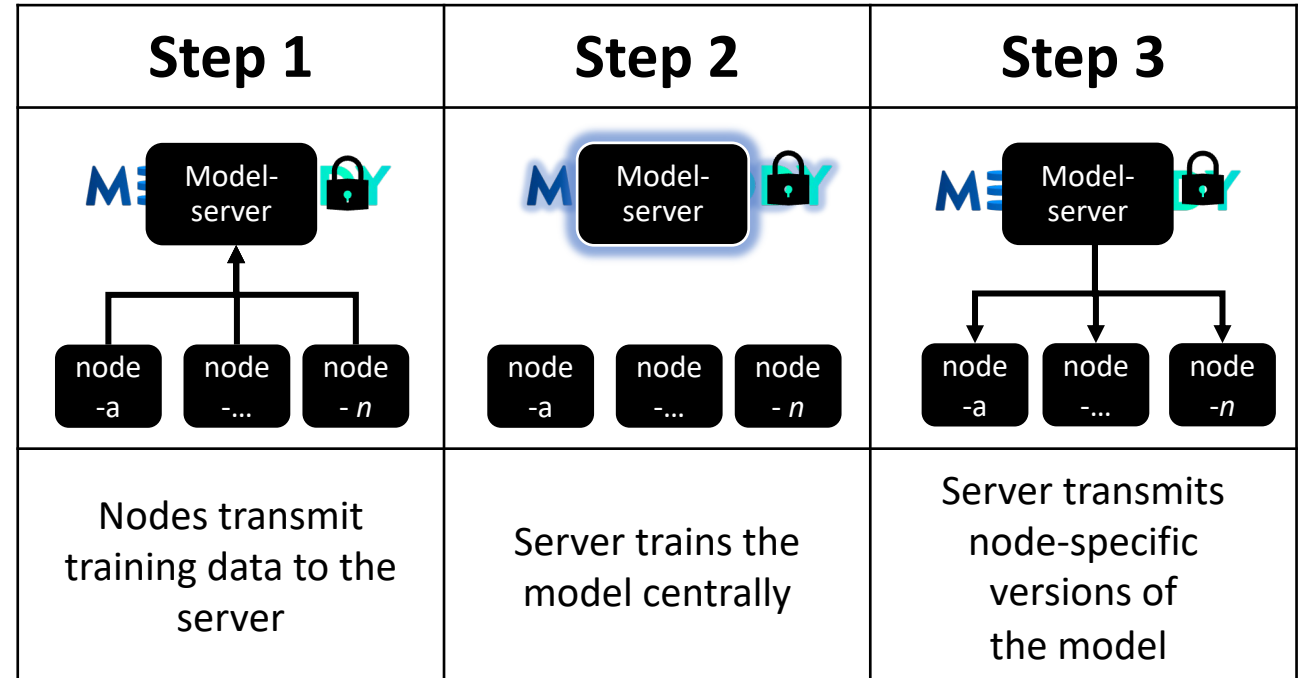
Node-centric

Some level of security

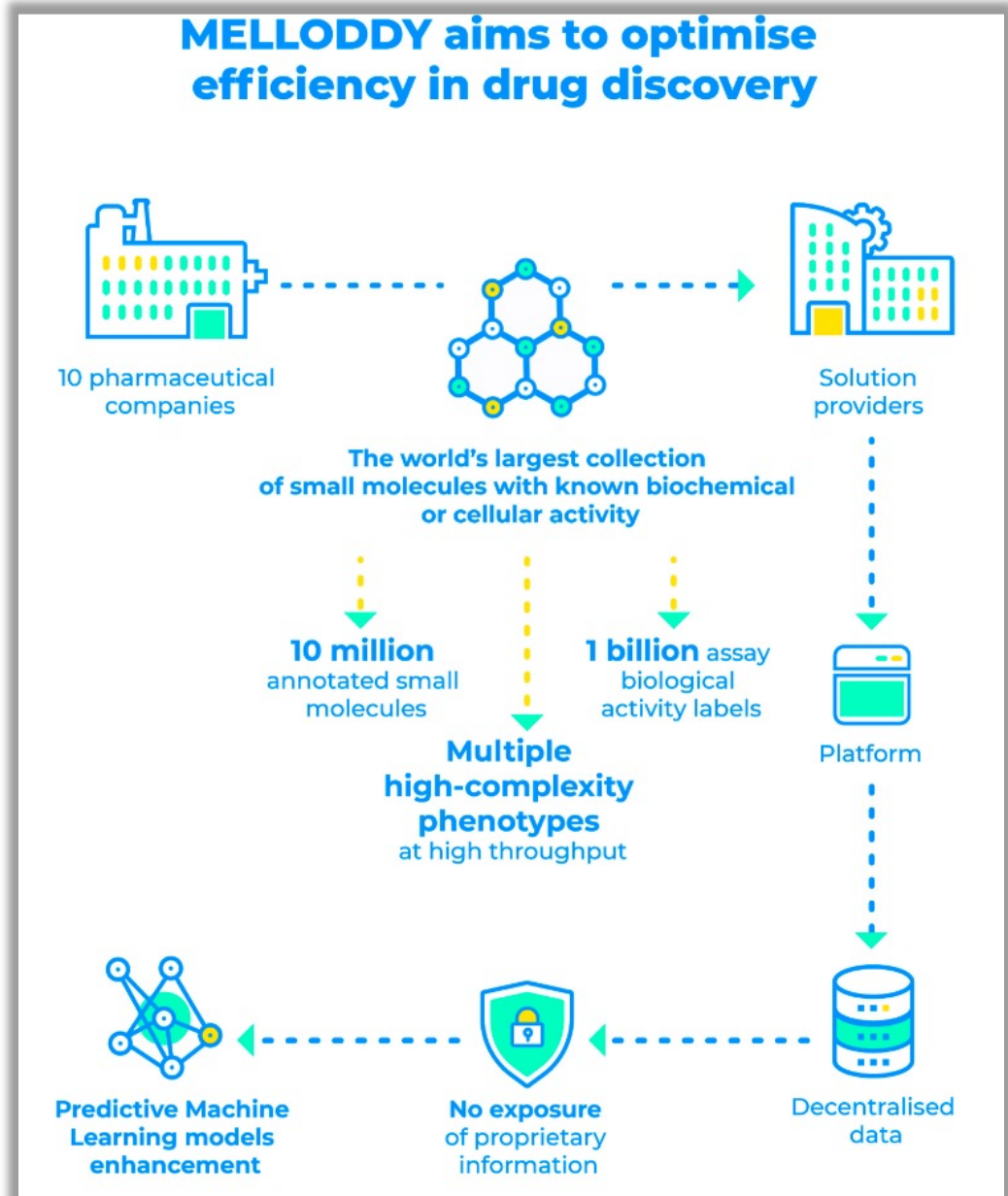


Server-centric

Even higher levels of security

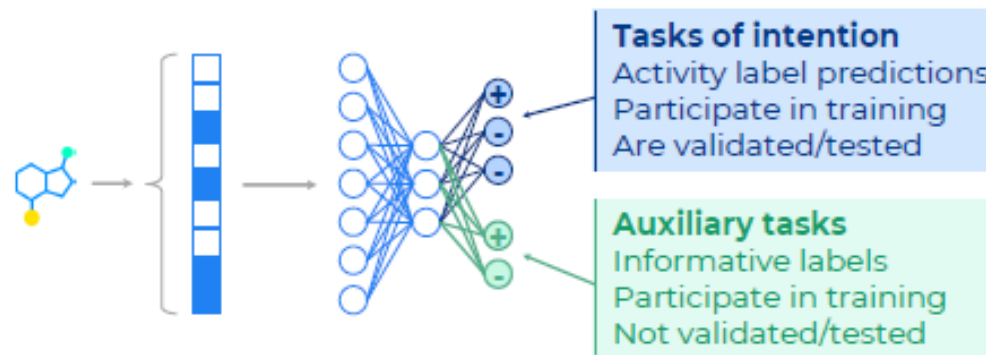


show the benefits of multi-task modelling across pharma partners at the largest achievable scale



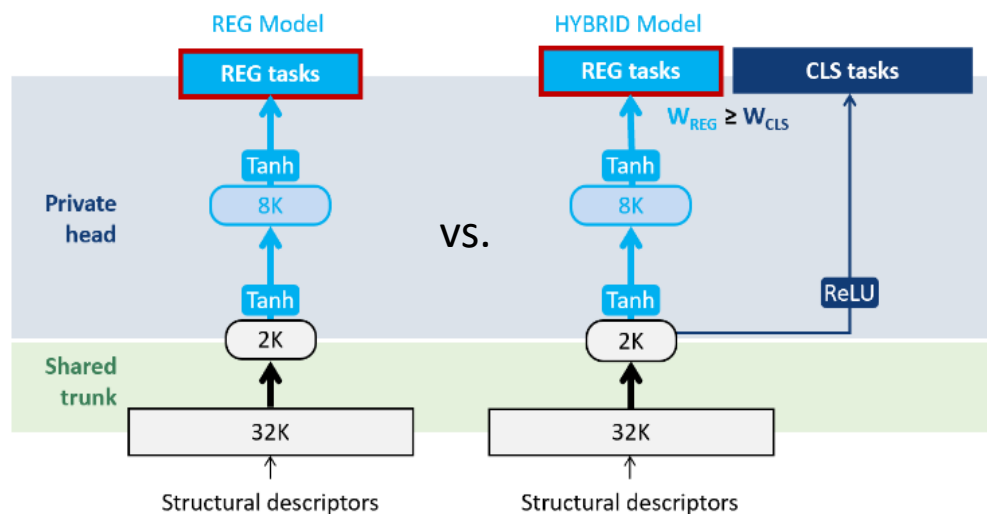
Privacy preservation of data and federated models is paramount

Data augmentation through auxiliary tasks



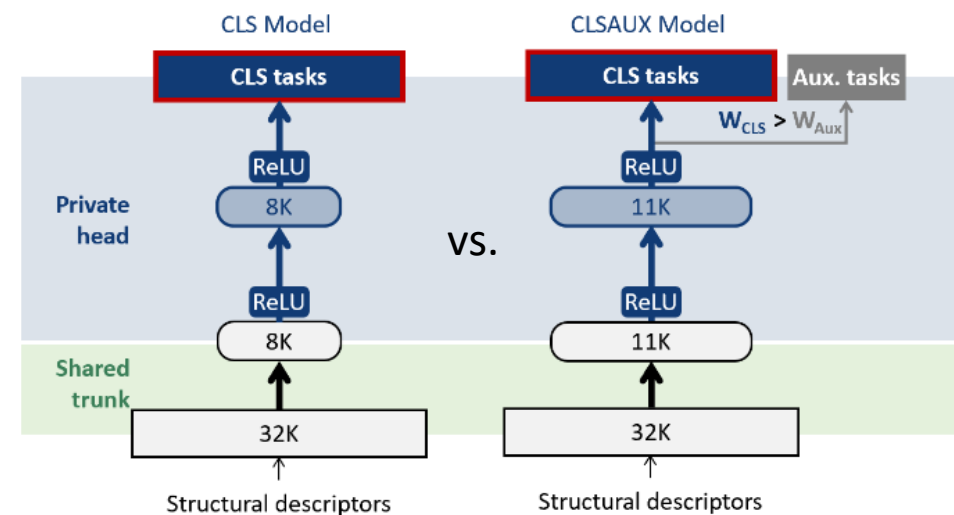
Regression Setup

- Auxiliary data: Classification tasks
- Hybrid model approach

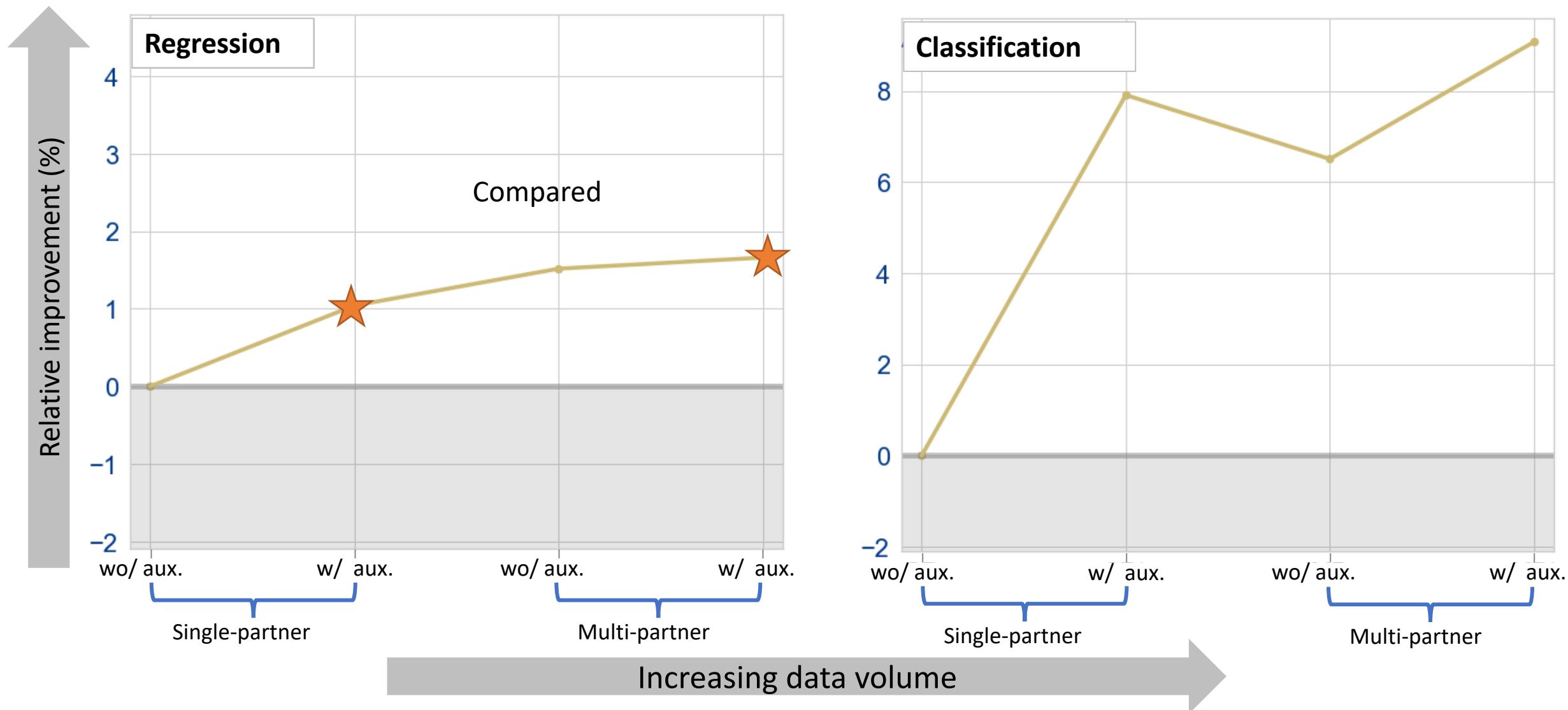


Classification Setup

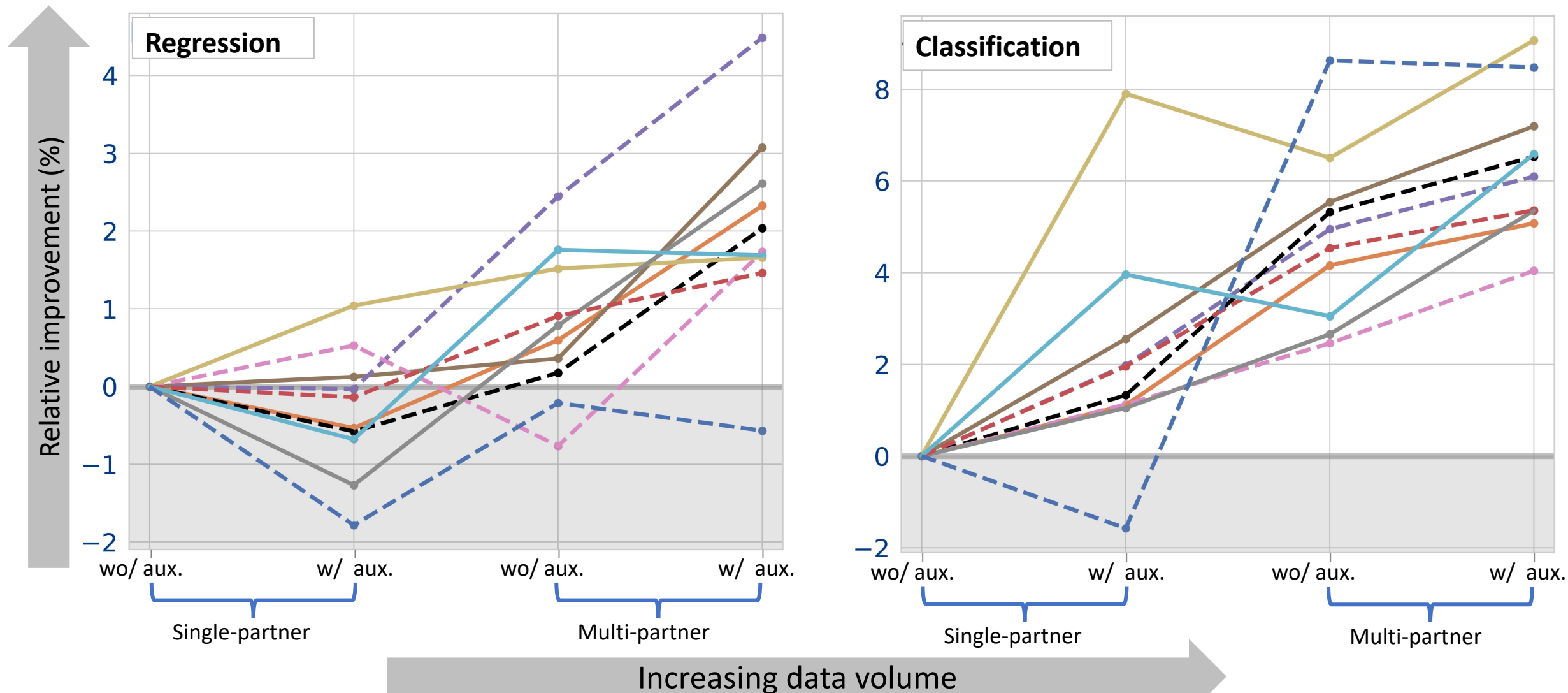
- Auxiliary data: HTS data
- Data volume increase **by 10-100x**



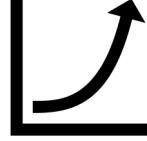



Increasing data volume boosts performance, with saturation



Increasing data volume boosts performance, with saturation



Some questions* raised from MELLODDY

- How to **assess performance increase**? 
 - Applicability domain – evaluate on unlabeled space?
- How to assess improvement of **uncertainty quantification**? 
- How to **get value** from models? 
- How to do MT learning in the future? 

Future outlook for MTL & FL

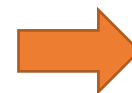
Multi-task learning: future outlook

- Research to explore **best practice**/ how to **benefit** from side information & auxiliary tasks

User data		Side information		
Compounds	Main pXC50	Side info task 1 pXC50	...	Side info task n pXC50
C1	4.8	3.4	...	4.3
C2	5.2	5.5	...	6.5
...



multi-task learning of the commonalities/correlations of compound activity in related side information tasks **benefits main task** predictions



Superior predictions using side info & multi-task learning



Hyperparameter

How to perform MTL in the future?

Q: What type of side information should be used?

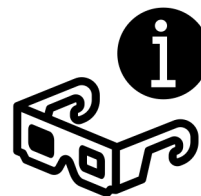
- Fingerprints
 - High throughput screening (HTS) fingerprints
 - Cell Painting
 - Morphology features
 - Pseudolabels



- [Predicted] properties, e.g.:
 - Physchem
 - QSAR models [+MELLODDY]
 - Physics-based methods?



- Protein space side information?
 - Sequence/graph/voxel/homology metadata
 - 3D/Structural descriptors

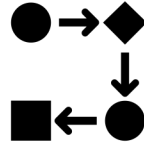


(not exhaustive)

How to represent MTL side information?

- **Pre-processing**

- Scaling (max/min)
- Variance filtering
- (Recursive) Feature selection
- Pseudolabels



- **Binary vs. continuous** tasks as side information

- MLDY conclusions:

- Regression benefitted from binary classification tasks
- Classification benefitted from binary auto-thresholding HTS screens

IOIO
IOIO

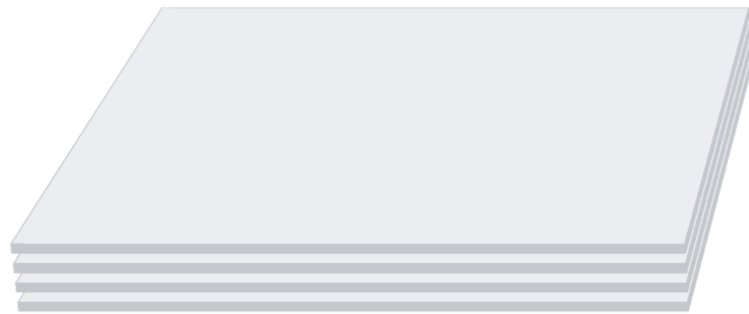
- Should be a **hyper-parameter choice**:

- which tasks benefit a primary task of interest?

How to identify which tasks should be learnt together?

- Intractable to search all task combinations for MTL
- Task sets may change throughout a model lifetime anyway
- **Draw inspiration from Meta-learning**, e.g.:
 - Learn representation minimizing loss for the weights **after 1+ steps** of training vs. the current set of weights
 - **Optimizes for the future**, not the present
- Task Affinity Groupings (TAG):
 - **Updates** model parameters with respect to 1 task
 - **Evaluate** the change on the other tasks
 - **Undoes** the update
 - **Process repeated** for all tasks to gather information how every task interacts

Task Affinity Groupings (TAG)^[1]



Train all tasks

- **Network selection algorithm** analyses task interaction data
- Groups tasks together that maximize **inter-task affinity**
- Outlines which tasks are **beneficial / antagonistic**

<https://ai.googleblog.com/2021/10/deciding-which-tasks-should-train.html>

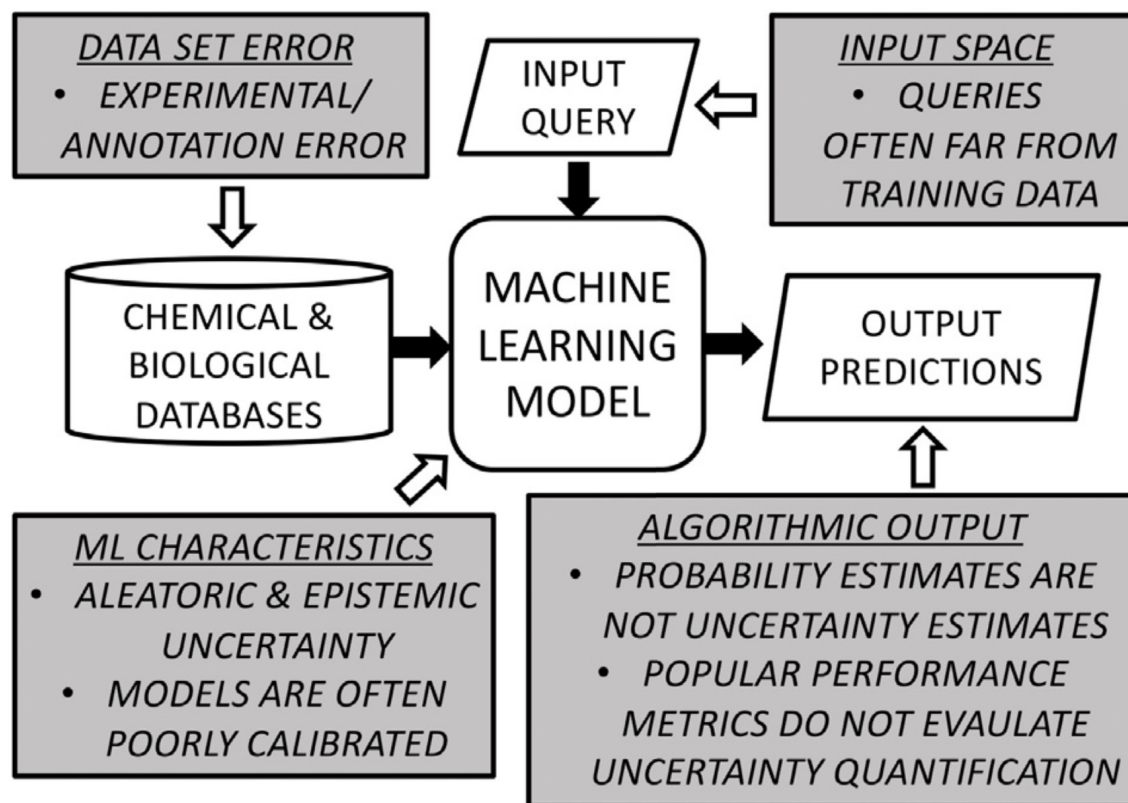
[1] Efficiently Identifying Task Groupings in Multi-Task Learning (NeurIPS 2021), Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, Chelsea Finn

How can we get better estimates
of uncertainty?

Why consider uncertainty estimation in drug design?

- Predictions without uncertainty are **difficult to interpret** & not always actionable
- Estimation recognised as a **principal shortcoming** of current approaches
- Better communication of uncertainty to **aid the adoption of ML**
- Key for autonomous decision making & integrating ML with chemistry automation to create an **autonomous DMTA cycle**
- Locating regions of chemical space with high uncertainty helps to **prioritise experiments** to expand future applicability domains (e.g., by active learning)

Factors to consider...



Drug Discovery Today

Lewis H.Mervin *et al*, Uncertainty quantification in drug design (2021), *Drug Discovery Today*

Various methods to model uncertainty

- Empirical
- Frequentist / Bayesian
- Ensemble-based



Drug Discovery Today
Volume 26, Issue 2, February 2021, Pages 474-489

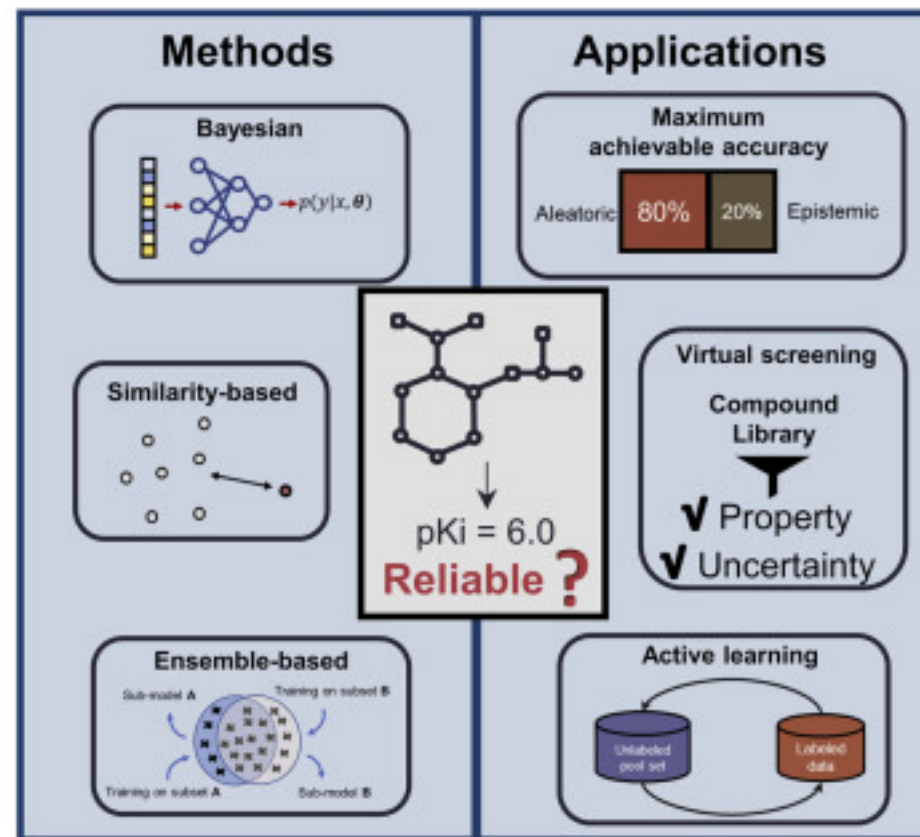


Review
Keynote

Uncertainty quantification in drug design

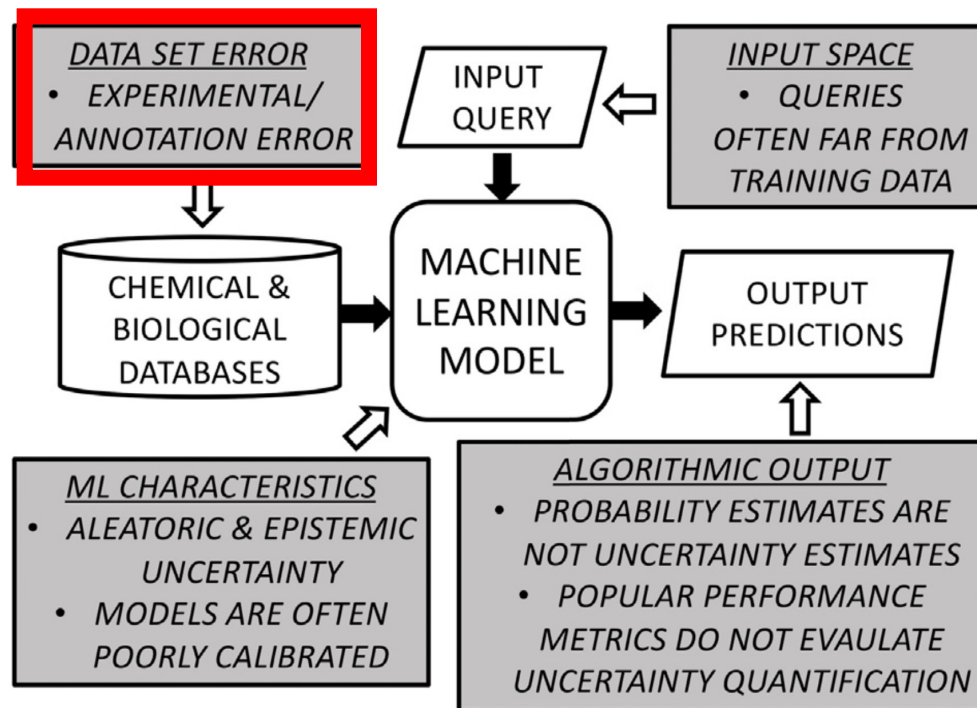
Lewis H. Mervin¹ , Simon Johansson^{2,3}, Elizaveta Semenova⁴, Kathryn A. Giblin⁵, Ola Engkvist³

Lewis H.Mervin *et al*, Uncertainty quantification in drug design (2021), Drug Discovery Today



Jie Yu, Uncertainty quantification: Can we trust artificial intelligence in drug discovery? (2022) iScience

Uncertainty estimation has historically focused on behavioural characteristics of base estimators, not the underlying (biological) data



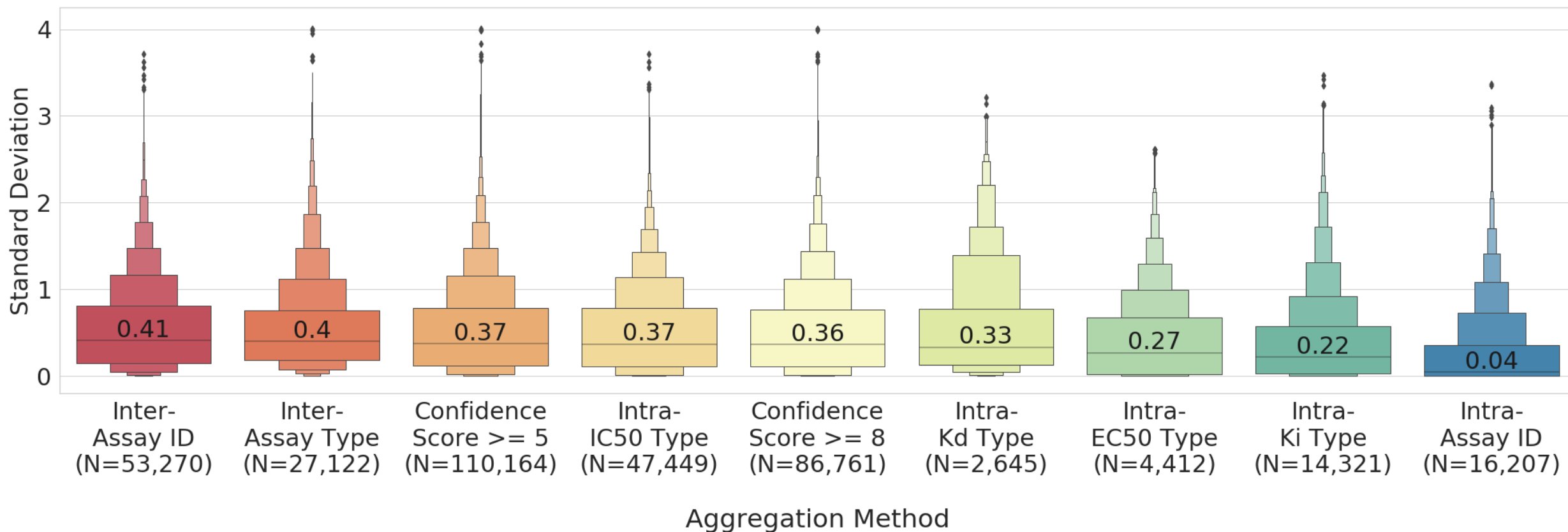
Drug Discovery Today

Lewis H.Mervin *et al*, Uncertainty quantification in drug design (2021), *Drug Discovery Today*

- Aleatoric uncertainty cannot be reduced, only identified and quantified
- Epistemic uncertainty can be reduced through more comprehensive study
- UQ intends to work towards reducing epistemic uncertainties to aleatoric uncertainties where possible

- **Max achievable accuracy/confidence of models = quality of experimental data**
 - i.e when **models approximate experimental error**

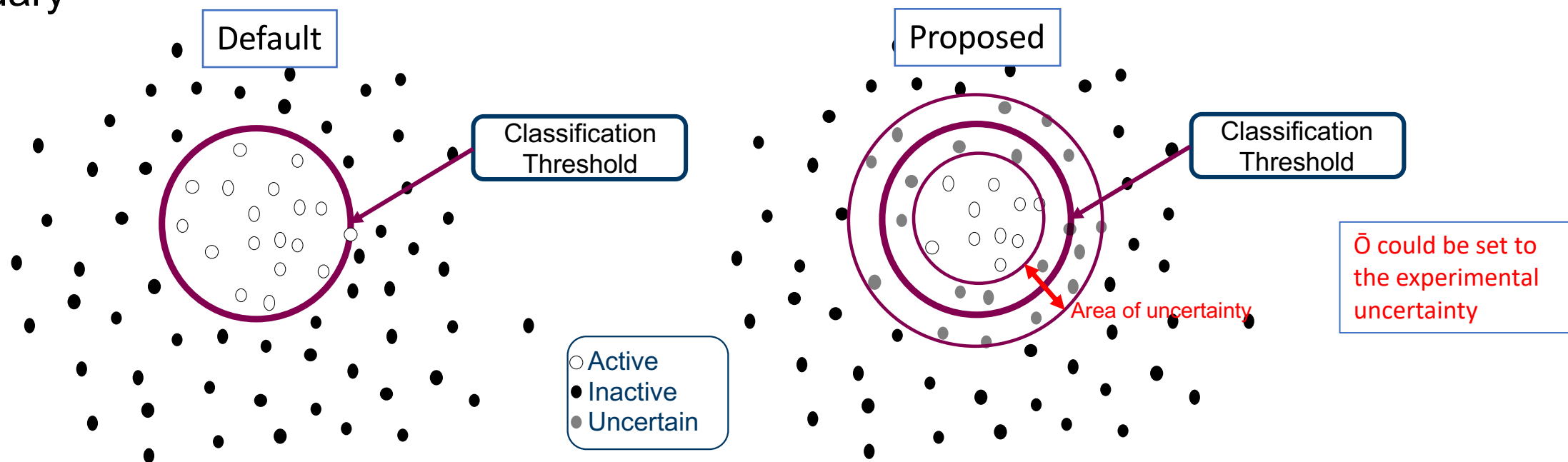
Experimental error of literature bioactivity data depends on consistency of experimental setup



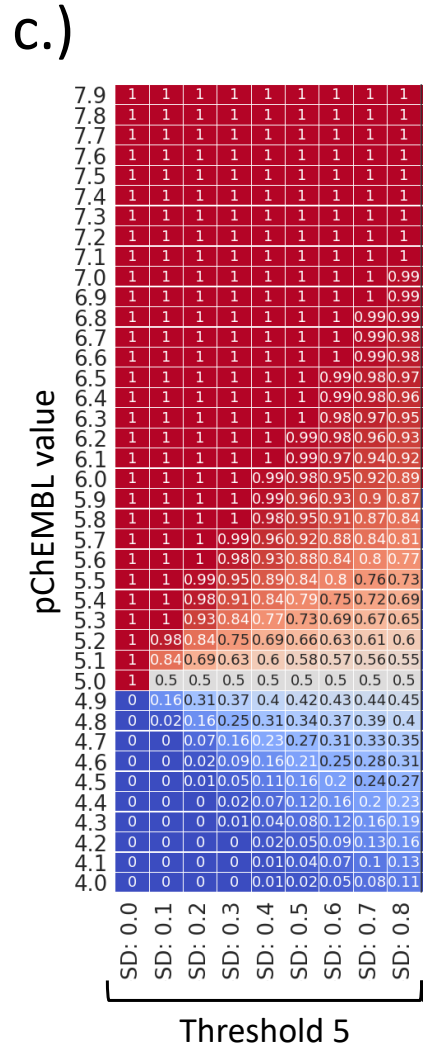
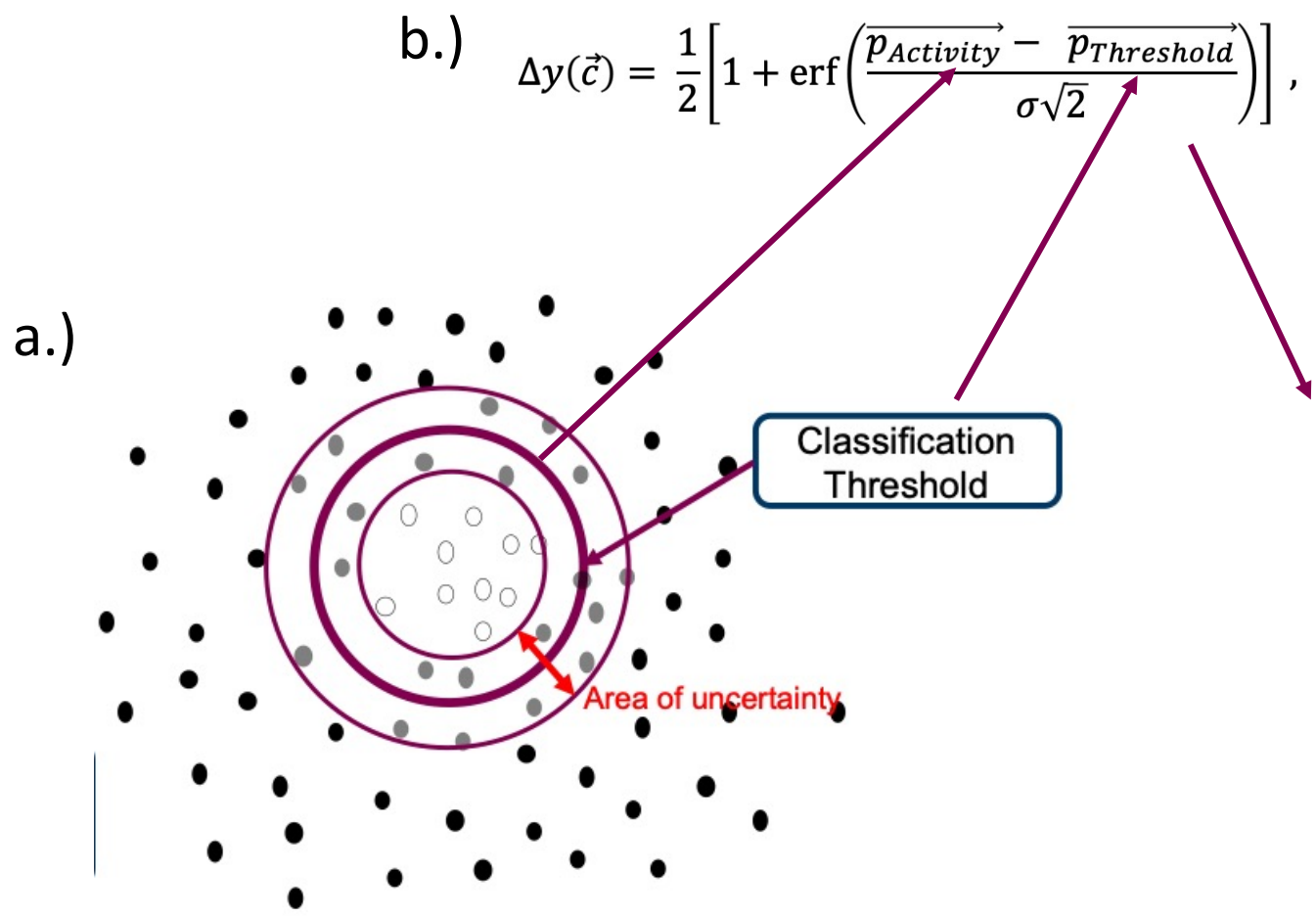
- **Overall SD** = 0.22–0.41 range depending bioactivity aggregation & different grouping schemes
- **Smallest SD** = intra- [when experimental results from same experiment (replicates)]

We should account for the uncertainty at label assignment

- A common approach to predict molecular properties is to use binary classifiers
- Biochemical experiments have associated reproducibility limits due to experimental error and classification threshold is usually arbitrary
- It is important to account for the uncertainty of activity label assignment at the decision boundary



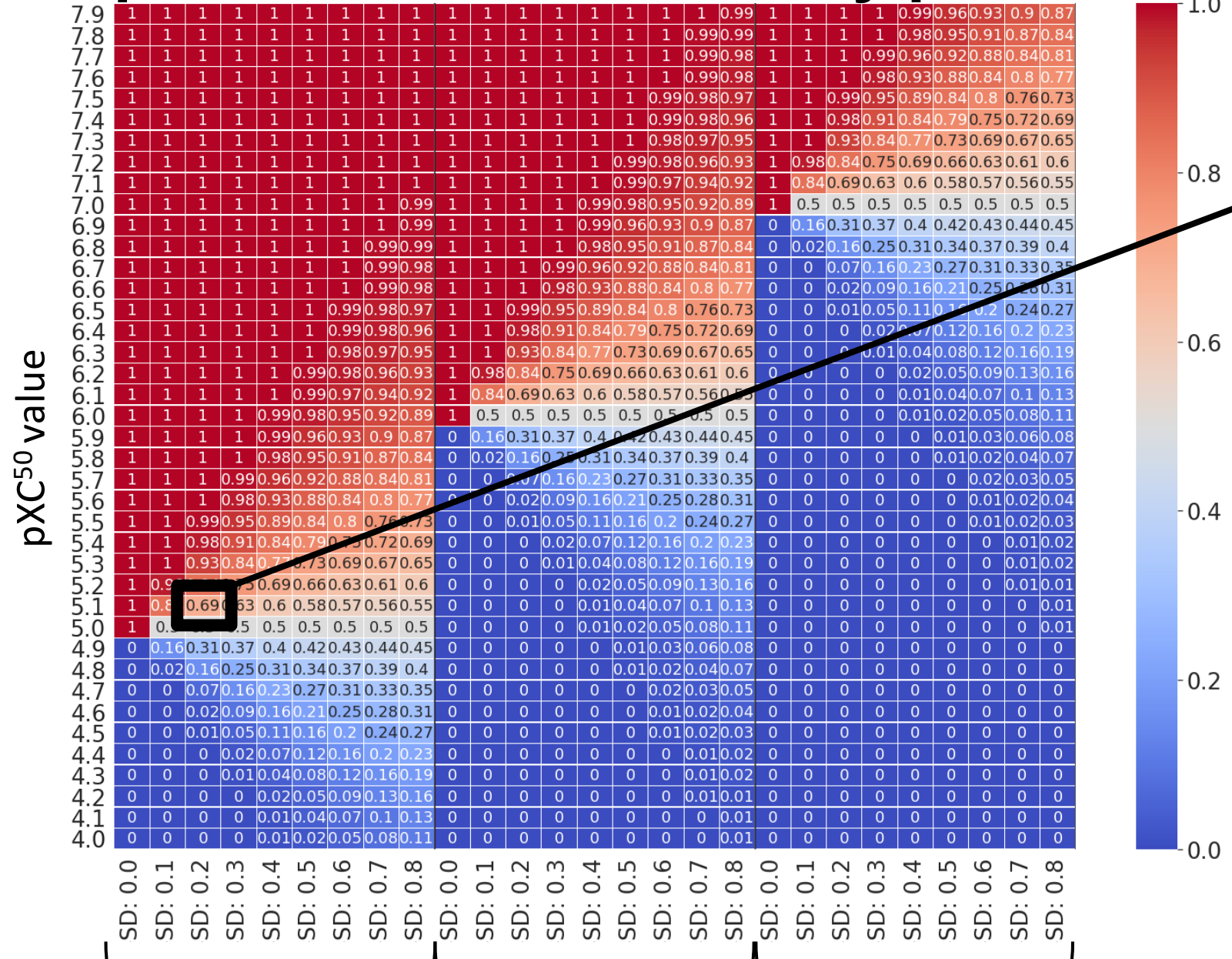
How to convert activity labels into probabilities (with cumulative distribution function)



somewhere between classification & regression

a.) define classification threshold & SD
 B.) apply CDF given input comp activity and the THR & SD
 c.) represents p(activity) on continuous scale

Lookup table for the bioactivity probabilities



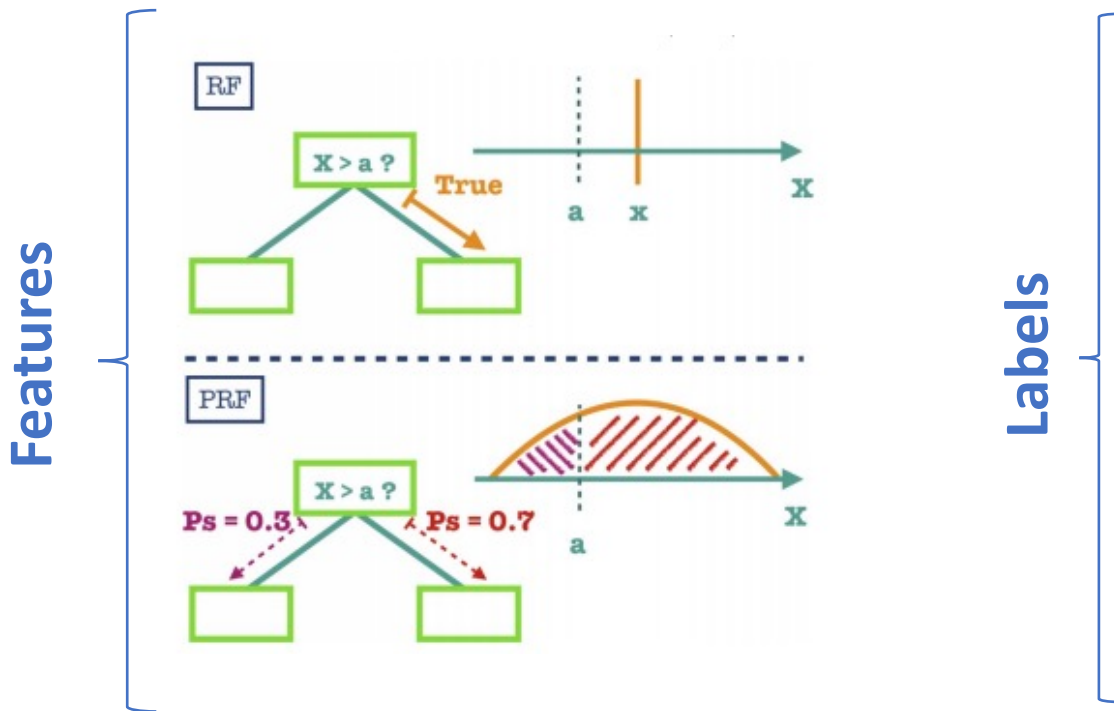
e.g., A compound with a **pChEMBL=5.1** (8 μ M) would be assigned a **new Δy of ~0.63** for an activity **threshold of 5.0** and a user-defined $\sigma = 0.3$

=>63% chance to belong to the active class compared to classic RF classifier which assumes that it is 100% active.

Threshold 5 Threshold 6 Threshold 7

Training a Probabilistic Random Forest (PRF)

- PRF: modification to the long-established Random Forest (RF) algorithm and takes into account uncertainties in *features* and/or *labels*



	Compound	pXC ⁵⁰	Activity label
RF	C ¹	4.8	0
	C ²	5.2	1

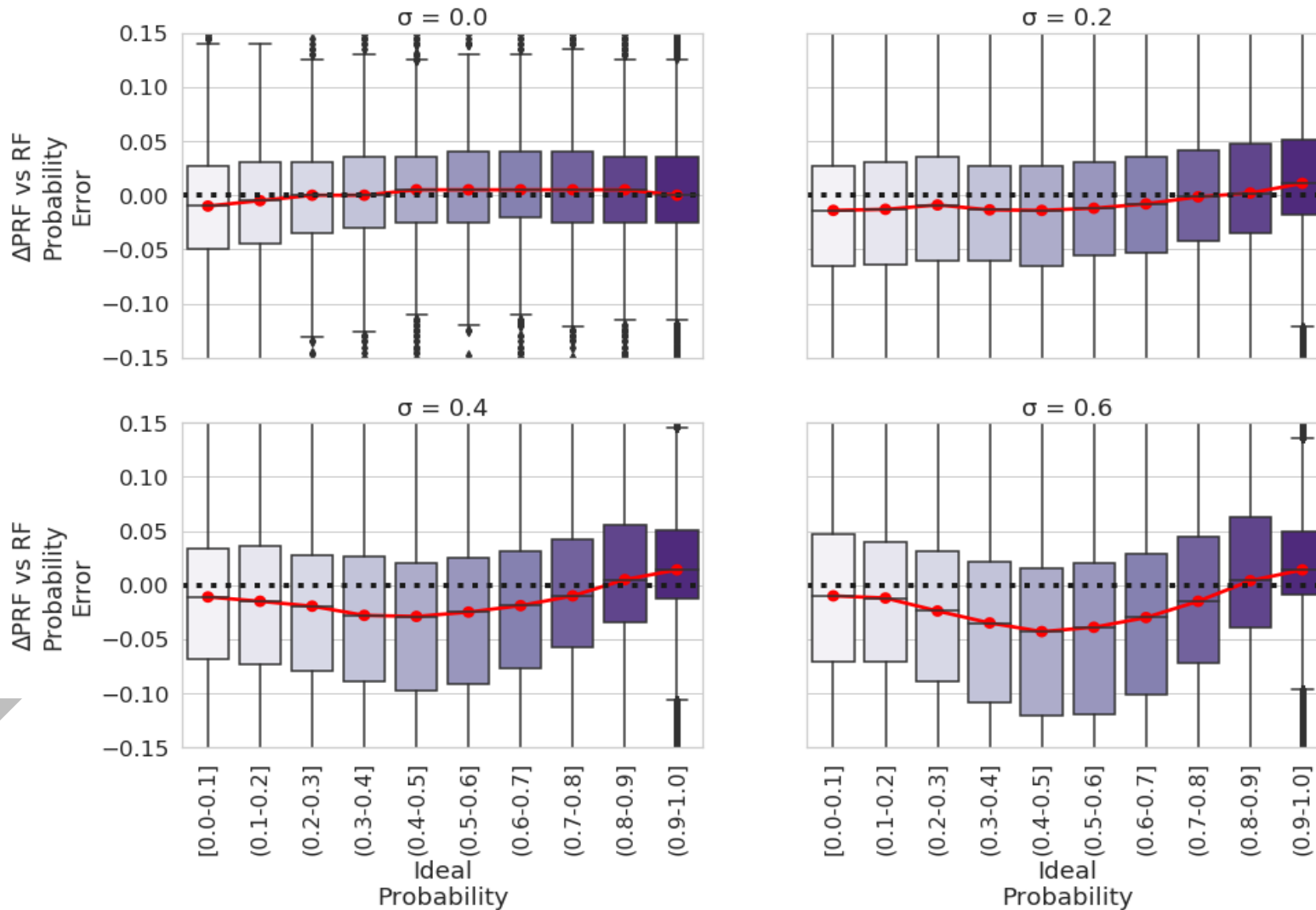
	C ⁿ	4.6	0
PRF	Compounds	pXC ⁵⁰	Probability to be active at a given threshold (e.g. 5)
	C ¹	4.8	0.39
	C ²	5.2	0.64

	C ⁿ	4.6	0.15

- RF uses **discrete variables** for the activity label (threshold applied to bioactivity data)
- PRF treats labels as **probability distribution functions** (rather than deterministic quantities)

PRF outperforms RF near the decision boundary

Benefit of PRF



- PRF > RF when there is a degree of uncertainty in the data (i.e., $\sigma \geq 0.2$)
- PRF has largest benefit over RF toward the midpoint of the probability scale
- This is because the RF weights the marginal cases equivalent in distinguishing between activity classes

Summary

- Overview of MAI & DMTA cycle provided
- Molecule property prediction forms a key part of the *de novo* design platform in-house
- Future work will evaluate how to benefit from MELLODDY models & how to do MTL going forward
- Various uncertainty quantification methods available, most focus on behavioural characteristics of base estimators
- We should consider uncertainty in experimental data - CDF/PRF can do this

Acknowledgments

