Active learning in chemistry using machine learning

Summer School 2025: Machine Learning and Artificial Intelligence in Synthetic Chemistry

University of Helsinki and the Finnish Society for Synthetic Chemistry

Lucía Morán González

Outline

1. What is active learning (AL)?
2. Why is AL so useful in chemistry?
3. Concept of query strategy
4. Learner model
5. Number of objectives
6. AL workflow
7. Applications
8. Jupyter notebook

What is active learning (AL)?

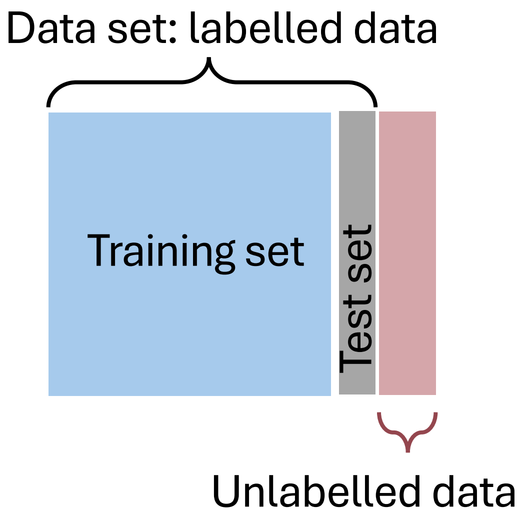
Passive learning

- Train a model → Data set is split into training and test (validation) set(s)
- Training set is *defined* from the beginning

Training set: X, y

Test set X, y

Unlabelled data X

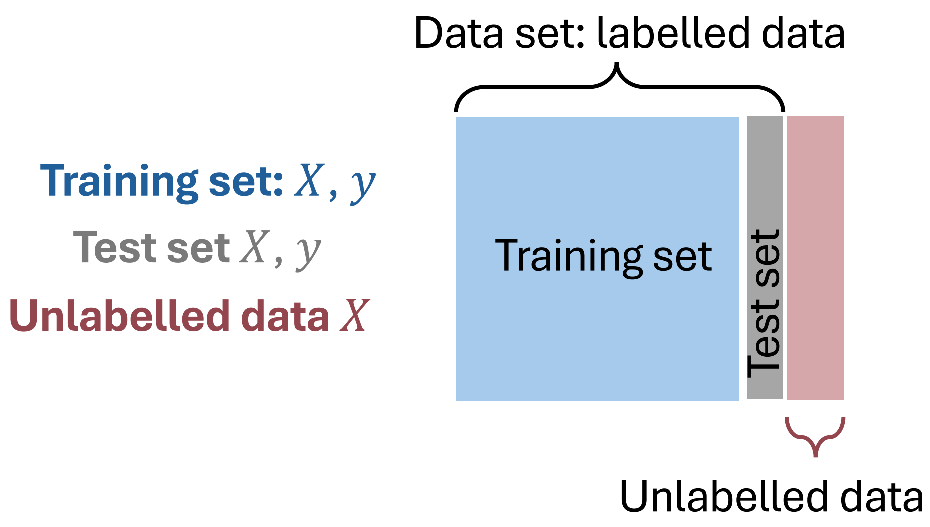


Active learning

What is active learning (AL)?

Passive learning

- Train a model → Data set is split into training and test (validation) set(s)
- Training set is *defined* from the beginning



Active learning

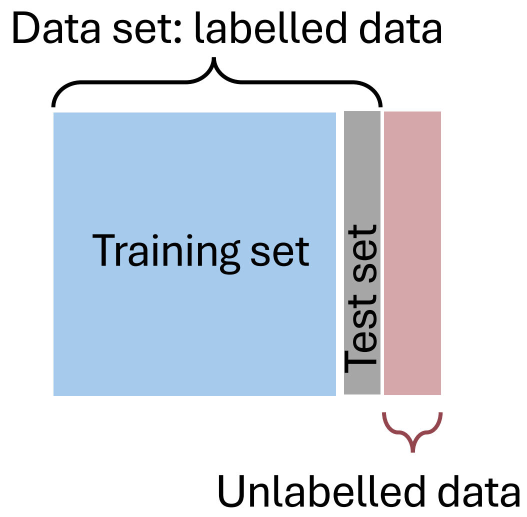
- Train a model → The final training set is not *predefined*

What is active learning (AL)?

Passive learning

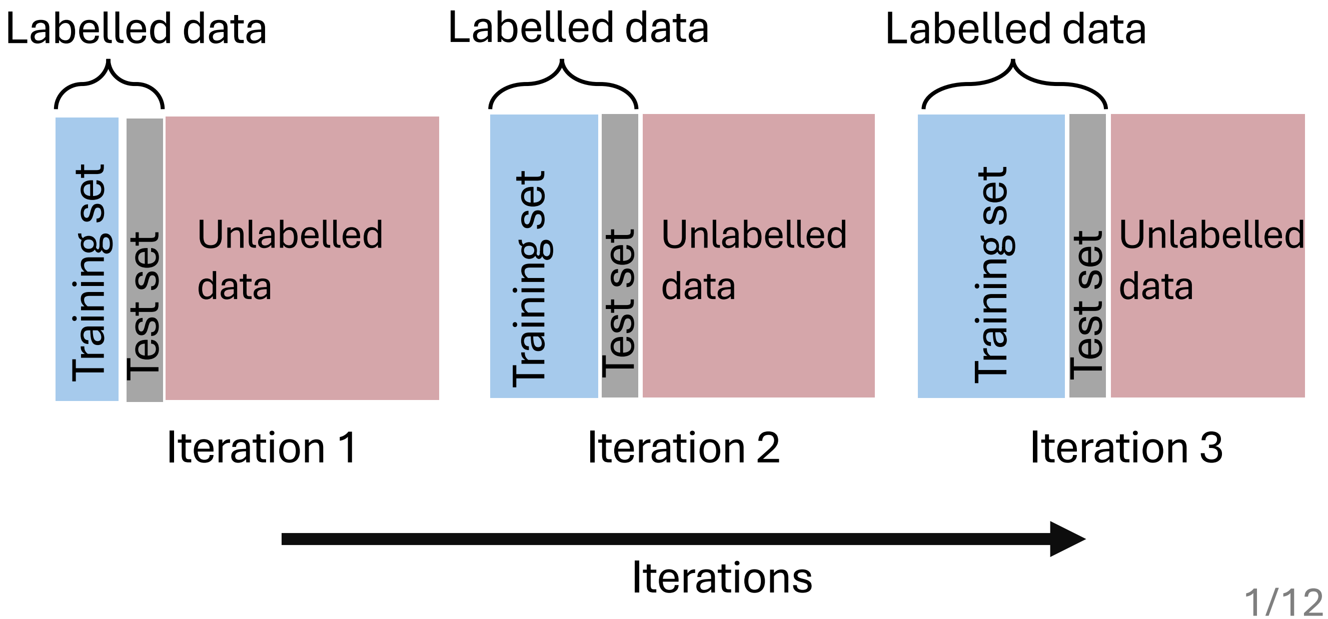
- Train a model → Data set is split into training and test (validation) set(s)
- Training set is *defined* from the beginning

Training set: X, y
Test set X, y
Unlabelled data X



Active learning

- Train a model → The final training set is not *predefined*
- Several **iterations** to increase the training set with the most informative data points (most relevant and representative) while improving model performance



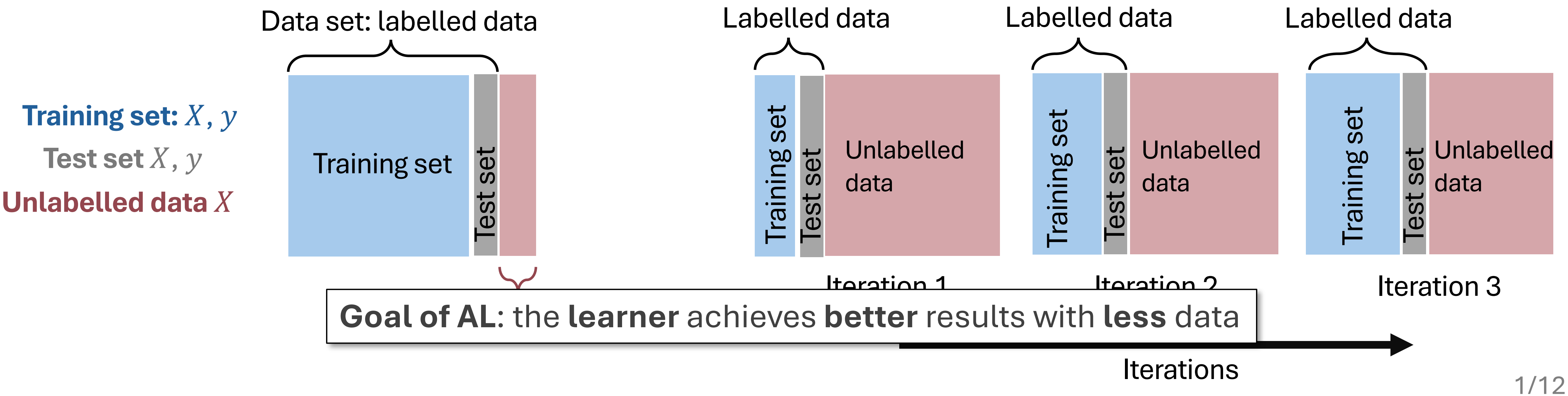
What is active learning (AL)?

Passive learning

- Train a model → Data set is split into training and test (validation) set(s)
- Training set is *defined* from the beginning

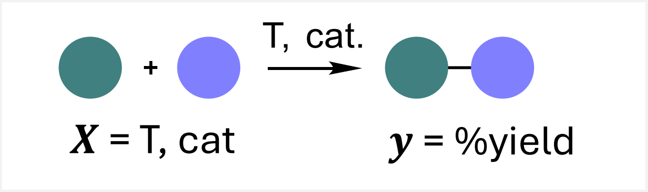
Active learning

- Train a model → The final training set is not *predefined*
- Several **iterations** to increase the training set with the most informative data points (most relevant and representative) while improving model performance



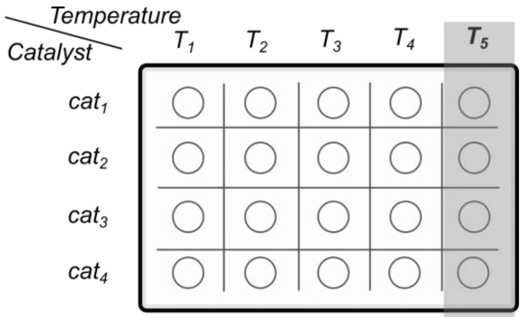
Goal of AL: the **learner** achieves **better** results with **less** data

Why is AL useful? – reaction campaign



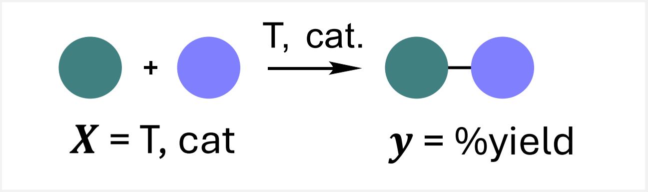
Experimental access to $\boldsymbol{X}$:
catalyst concentration [1mM, 2mM, 3mM, 4mM]
& temperature [25ºC, 30ºC, 50ºC, 100ºC]

Yield prediction



cat = [cat]

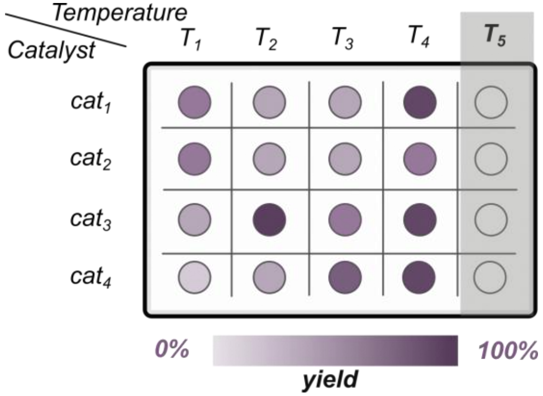
Why is AL useful? – reaction campaign



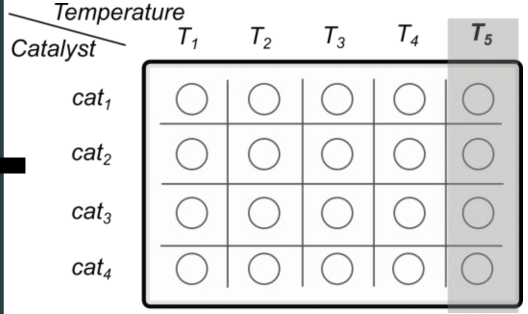
Experimental access to X :
catalyst concentration [1mM, 2mM, 3mM, 4mM]
& temperature [25ºC, 30ºC, 50ºC, 100ºC]

Passive learning

Training set : 16 experiments

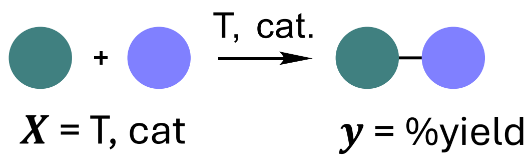


Yield prediction



cat = [cat]

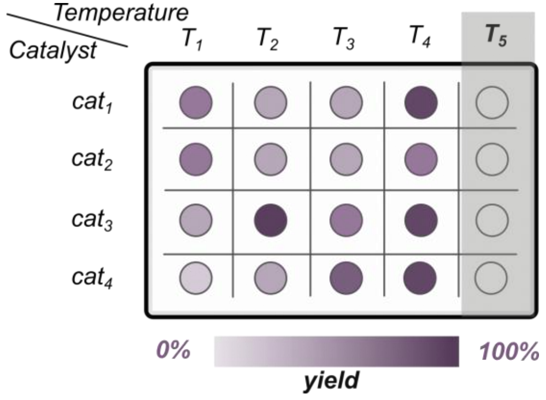
Why is AL useful? – reaction campaign



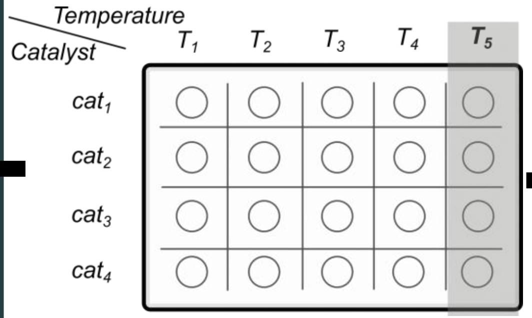
Experimental access to $\boldsymbol{X}$:
catalyst concentration [1mM, 2mM, 3mM, 4mM]
& temperature [25ºC, 30ºC, 50ºC, 100ºC]

Passive learning

Training set : 16 experiments



Yield prediction

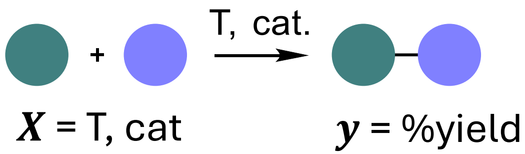


cat = [cat]

Active learning

- **Acquisition function (AF) to query** most informative **unlabelled** data experiments
- **Why ML?** To make the **predictions** and select the most informative points based on different criteria

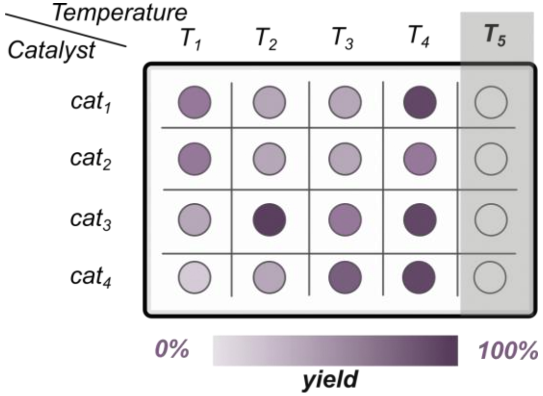
Why is AL useful? – reaction campaign



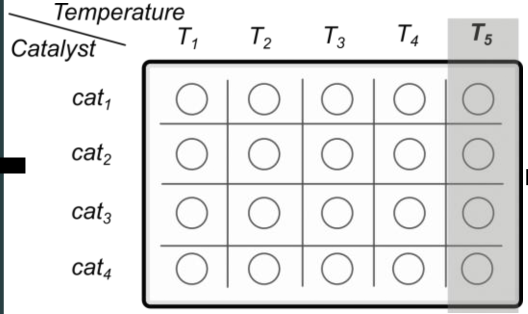
Experimental access to $\boldsymbol{X}$:
catalyst concentration [1mM, 2mM, 3mM, 4mM]
& temperature [25ºC, 30ºC, 50ºC, 100ºC]

Passive learning

Training set : 16 experiments



Yield prediction

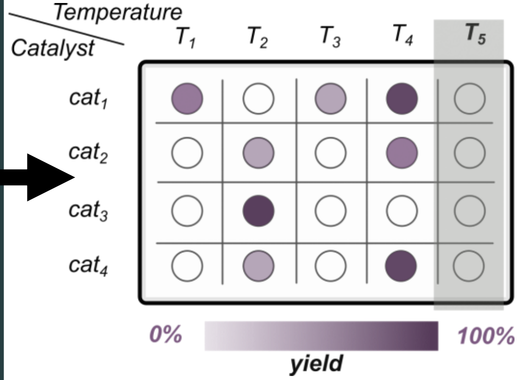


cat = [cat]

Active learning

Training set : 11 experiments

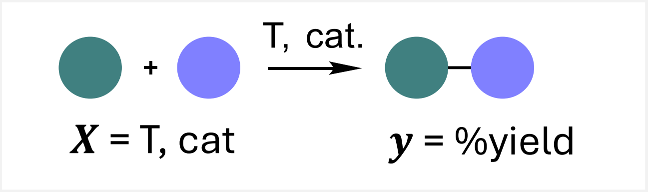
Training set 8 exp.



Iteration 1

- **Acquisition function (AF) to query** most informative **unlabelled** data experiments
- **Why ML?** To make the **predictions** and select the most informative points based on different criteria

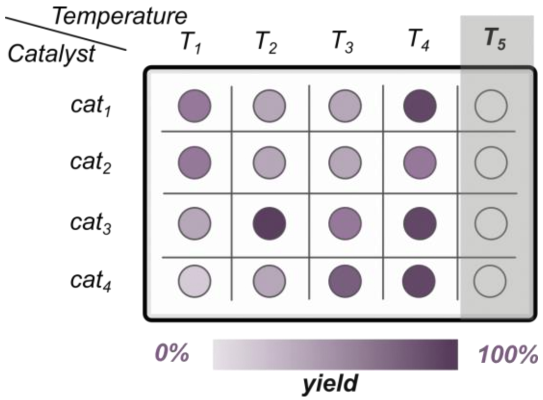
Why is AL useful? – reaction campaign



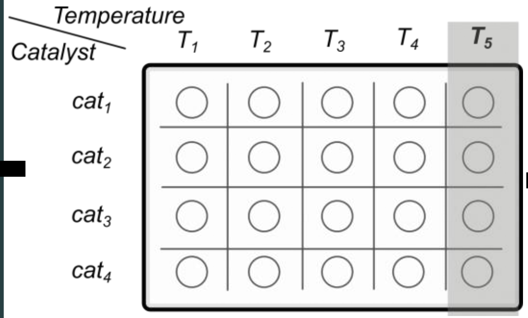
Experimental access to $\boldsymbol{X}$:
catalyst concentration [1mM, 2mM, 3mM, 4mM]
& temperature [25ºC, 30ºC, 50ºC, 100ºC]

Passive learning

Training set : 16 experiments



Yield prediction



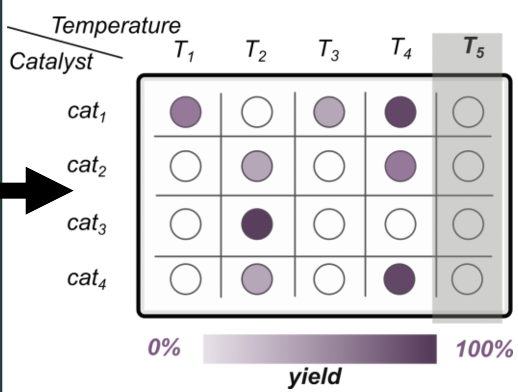
cat = [cat]

Active learning

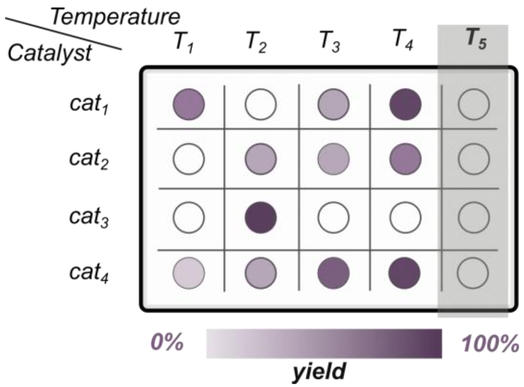
Training set : 11 experiments

Training set 8 exp

+ 3 exp.



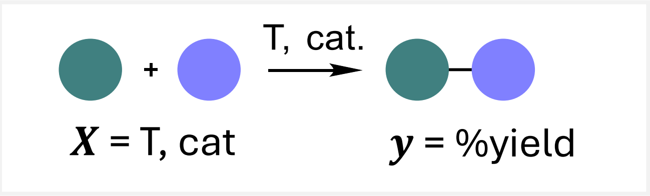
Iteration 1



Iteration 2

- **Acquisition function (AF) to query** most informative **unlabelled** data experiments
- **Why ML?** To make the **predictions** and select the most informative points based on different criteria

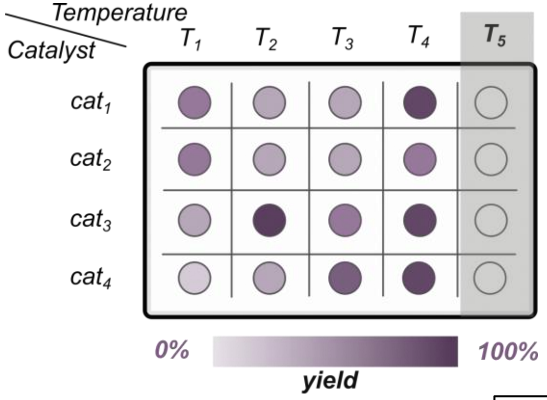
Why is AL useful? – reaction campaign



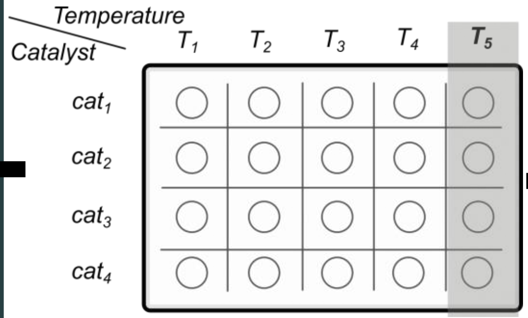
Experimental access to $\boldsymbol{X}$:
catalyst concentration [1mM, 2mM, 3mM, 4mM]
& temperature [25ºC, 30ºC, 50ºC, 100ºC]

Passive learning

Training set : 16 experiments



Yield prediction

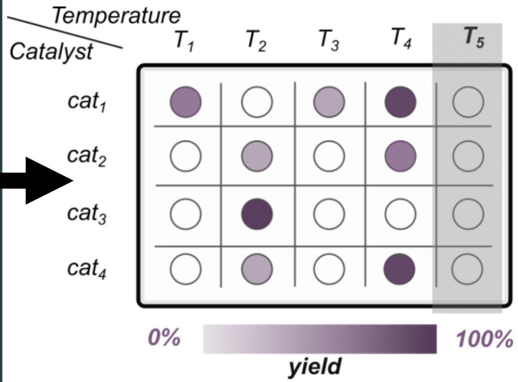


cat = [cat]

Active learning

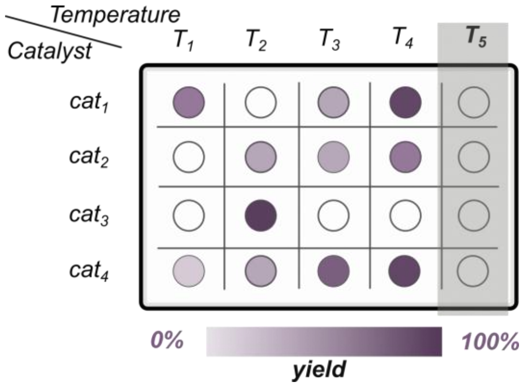
Training set : 11 experiments

Training set 8 exp



Iteration 1

+ 3 exp.

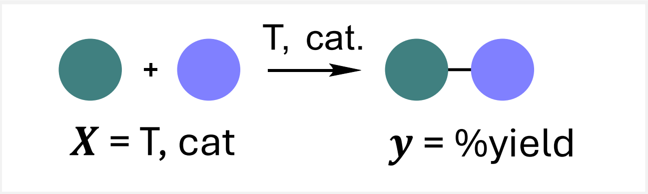


Iteration 2

Less data to get a **general** prediction model

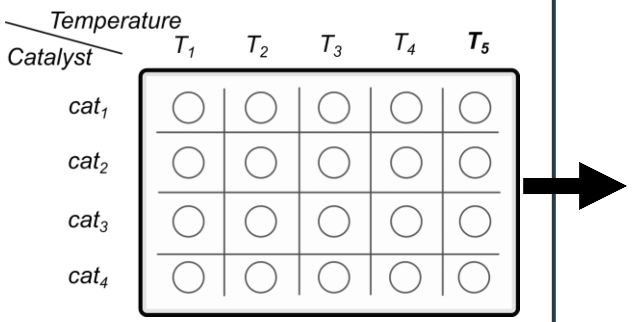
- **Acquisition function (AF) to query** most informative **unlabelled** data experiments
- **Why ML?** To make the **predictions** and select the most informative points based on different criteria

Why is AL useful? – reaction campaign



Experimental access to $\boldsymbol{X}$:
catalyst concentration [1mM, 2mM, 3mM, 4mM]
& temperature [25ºC, 30ºC, 50ºC, 100ºC]

Active learning

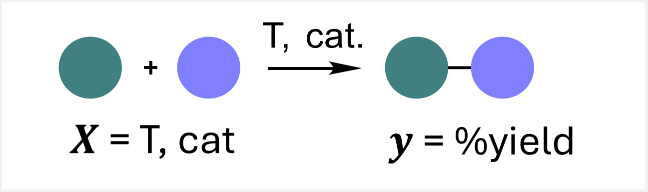


cat = [cat]

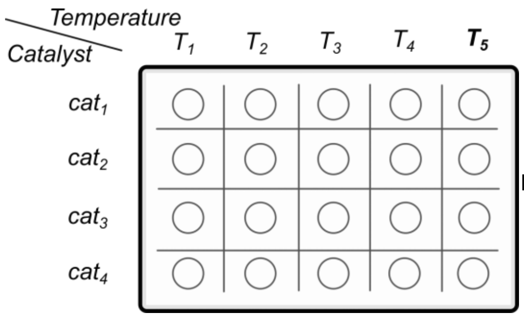
Globally **optimized** reaction conditions

- **Acquisition function (AF) to query** most informative **unlabelled** data experiments
- **Why ML?** To make the **predictions** and select the most informative points based on different criteria

Why is AL useful? – reaction campaign



Experimental access to X :
catalyst concentration [1mM, 2mM, 3mM, 4mM]
& temperature [25ºC, 30ºC, 50ºC, 100ºC]



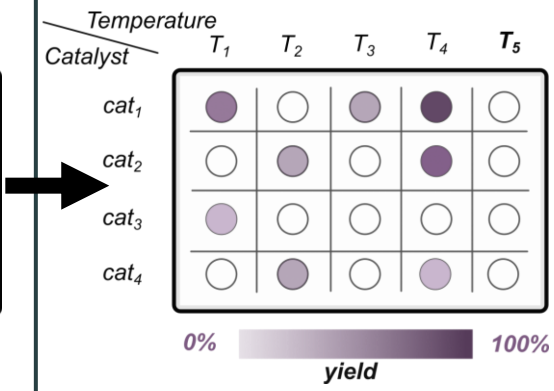
cat = [cat]

Globally **optimized** reaction conditions

Active learning

Training set : 11 experiments

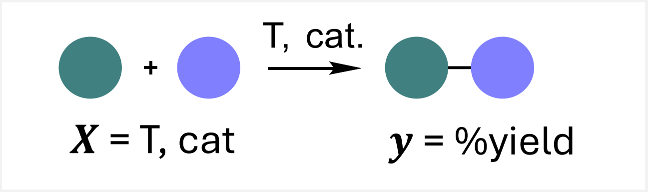
Training set 8 exp



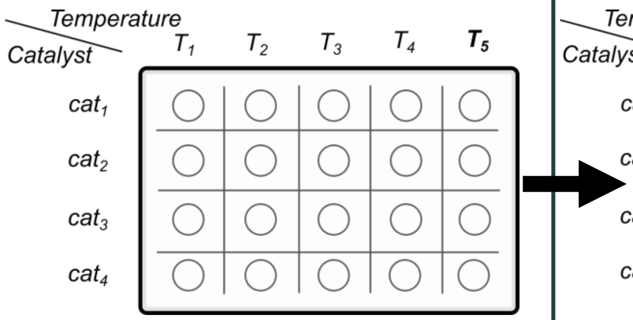
Iteration 1

- **Acquisition function (AF) to query** most informative **unlabelled** data experiments
- **Why ML?** To make the **predictions** and select the most informative points based on different criteria

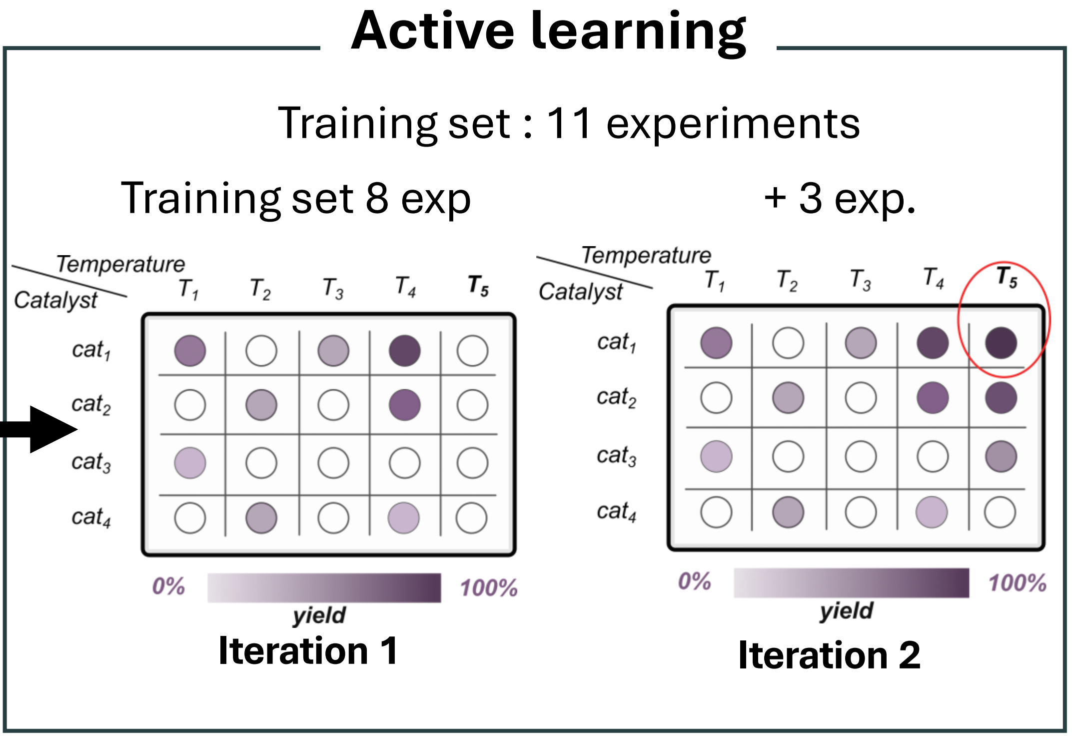
Why is AL useful? – reaction campaign



Experimental access to $\boldsymbol{X}$:
catalyst concentration [1mM, 2mM, 3mM, 4mM]
& temperature [25ºC, 30ºC, 50ºC, 100ºC]



cat = [cat]

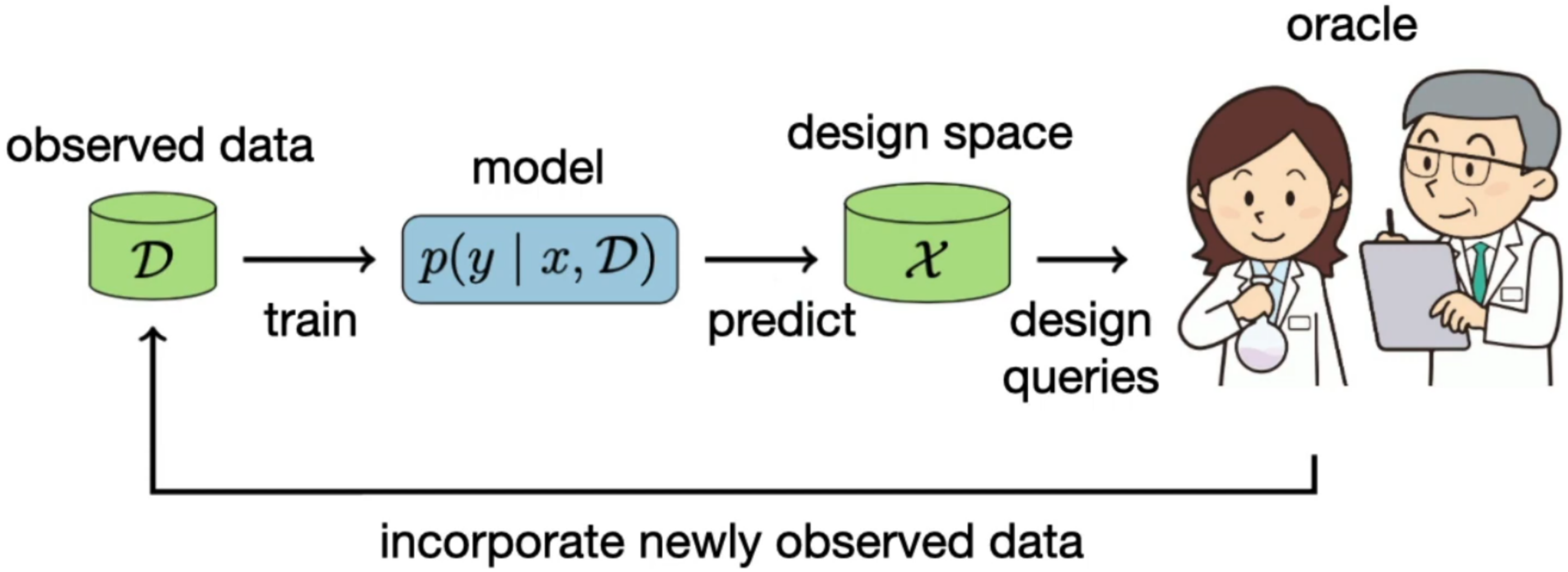


Globally **optimized** reaction conditions

- **Acquisition function (AF) to query** most informative **unlabelled** data experiments
- **Why ML?** To make the **predictions** and select the most informative points based on different criteria

Workflow

Bayesian Optimization experiment



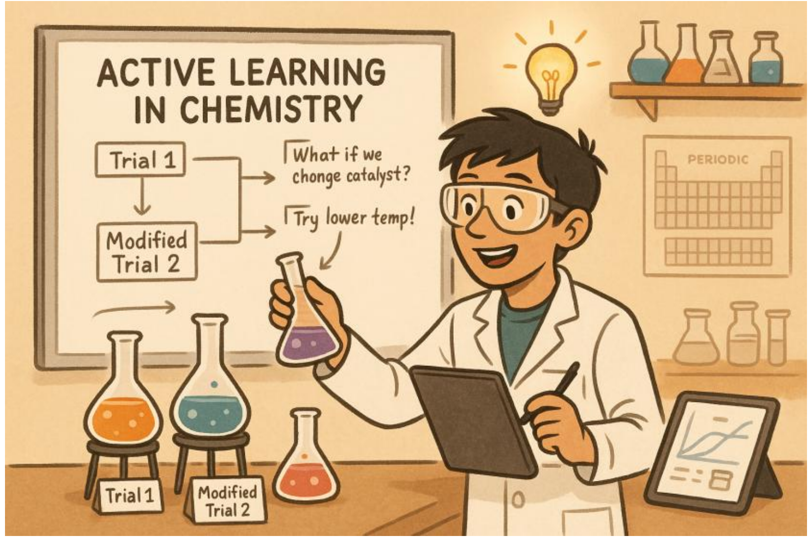
Why is AL useful in chemistry?

AL helps chemists refine their reactions/designs by learning through hands-on experimentation

- Labelling data points (training set) is often expensive **€€€**. Unlabelled data (test set) is often cheap. Save time and resources. Adjust the chemical space domain
  E.g. Run reactions/calculations (training set) €€€

- **Not** all labels are equally **useful**. Noise, outliers, missing values...
  E.g. Reactions involving structurally similar substrates exhibit comparable selectivity profiles

- We want to collect the **best data** at minimal cost.
  Most informative to be able to predict



Why is AL useful in chemistry?

AL helps chemists refine their reactions/designs by learning through hands-on experimentation

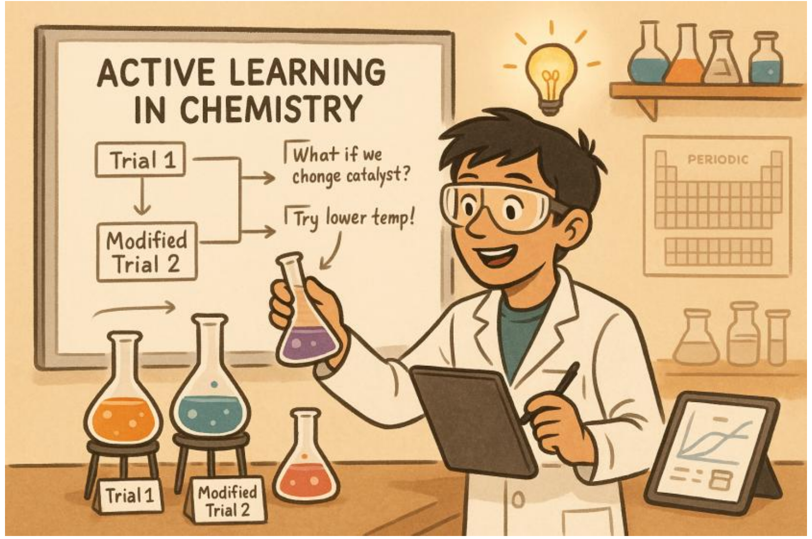
- Labelling data points (training set) is often expensive **€€€**. Unlabelled data (test set) is often cheap. Save time and resources. Adjust the chemical space domain
  E.g. Run reactions/calculations (training set) €€€

- **Not** all labels are equally **useful**. Noise, outliers, missing values...
  E.g. Reactions involving structurally similar substrates exhibit comparable selectivity profiles

- We want to collect the **best data** at minimal cost.
  Most informative to be able to predict

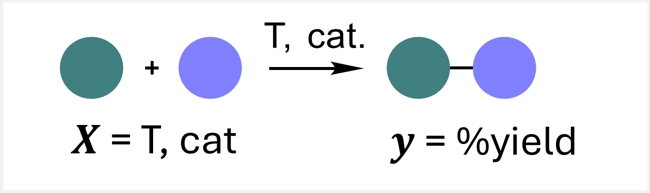
How to choose the data points to be trained?

Learner + Criteria to query



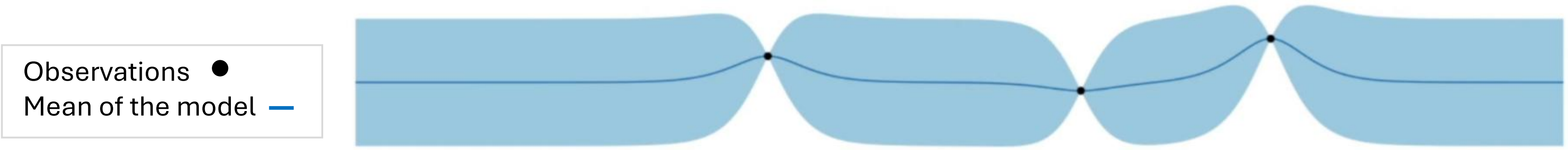
Learner: ML model

- Provide **predictions** and **uncertainties** for the unlabelled data points



$\hat{y} \pm \sigma^2(\mathbf{x})$

- ML models: Gaussian Processes (GP), Neural networks, Random Forests



Criteria to query points to label

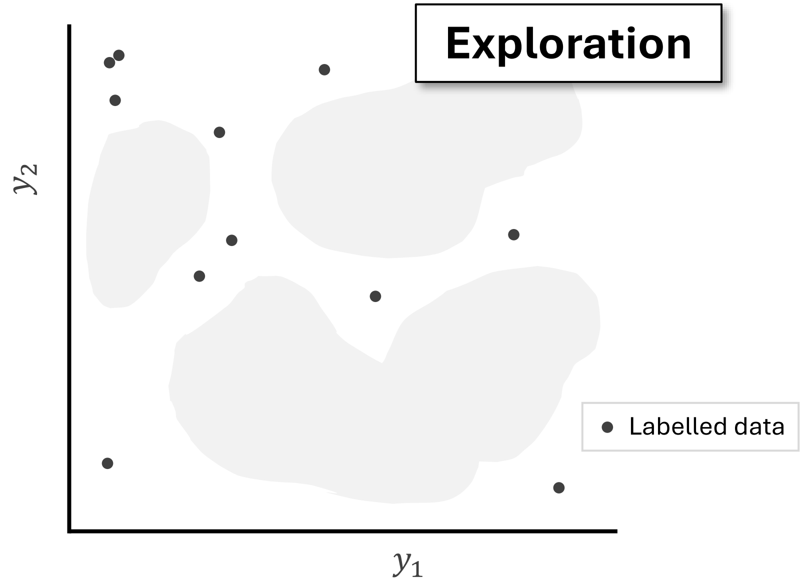
- **Exploration** implies selecting data points that are **diverse** and cover different regions of the feature space – long-term benefit

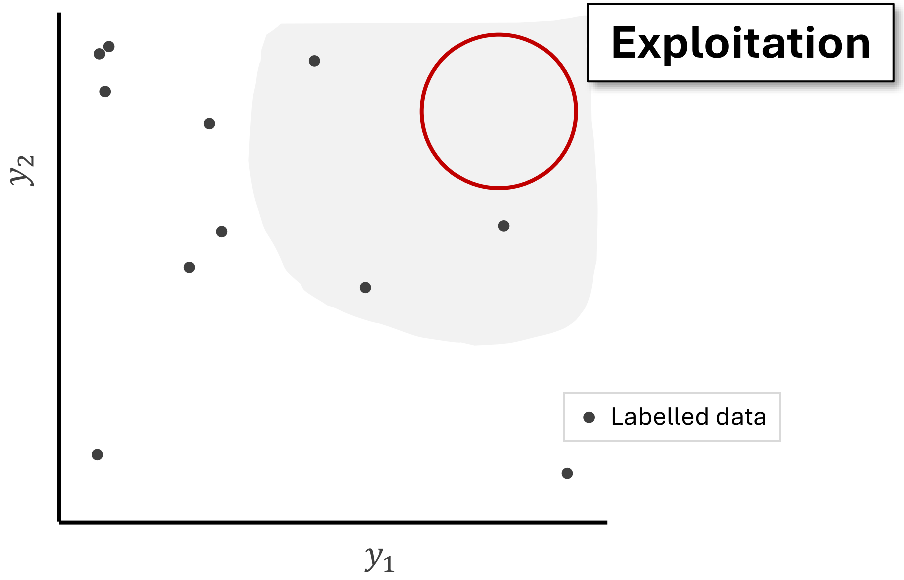
  e.g. solubility of compounds

- **Exploitation** means selecting data points that are informative and **reduce** the model **uncertainty**. Maximize, minimize or specific areas of the target property – most reward

  e.g. maximize the selectivity

- Optimal: balance the trade-off between them and find the optimal query strategy.





Criteria to query points to label

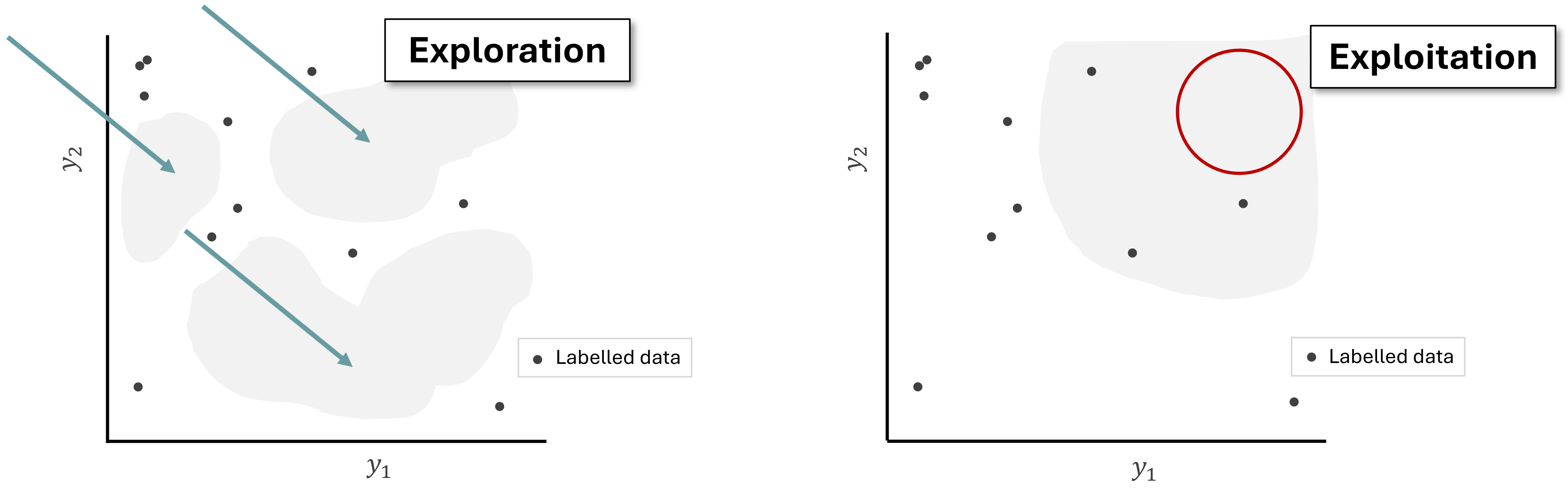
- **Exploration** implies selecting data points that are **diverse** and cover different regions of the feature space – long-term benefit

  e.g. solubility of compounds

- **Exploitation** means selecting data points that are informative and **reduce** the model **uncertainty**. Maximize, minimize or specific areas of the target property – most reward

  e.g. maximize the selectivity

- Optimal: balance the trade-off between them and find the optimal query strategy.



Criteria to query points to label

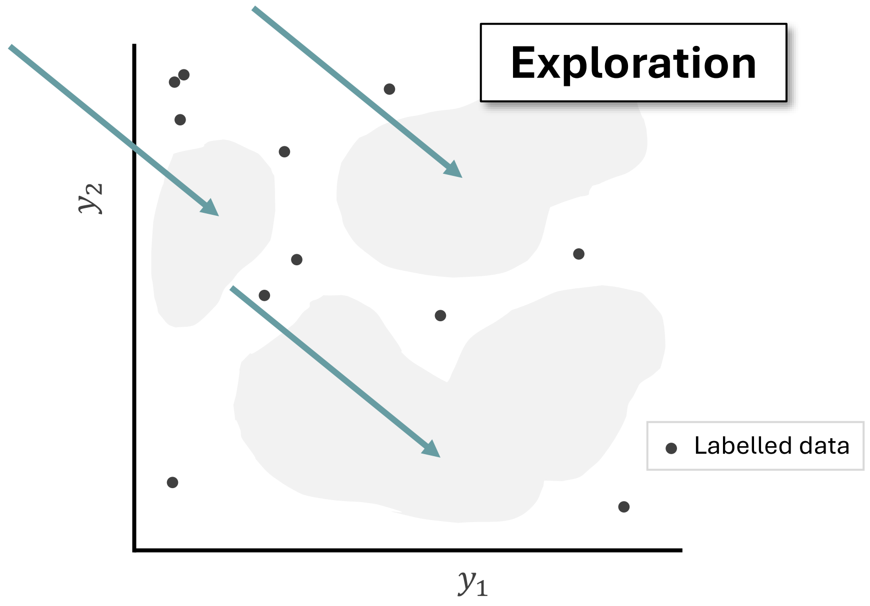
- **Exploration** implies selecting data points that are **diverse** and cover different regions of the feature space – long-term benefit

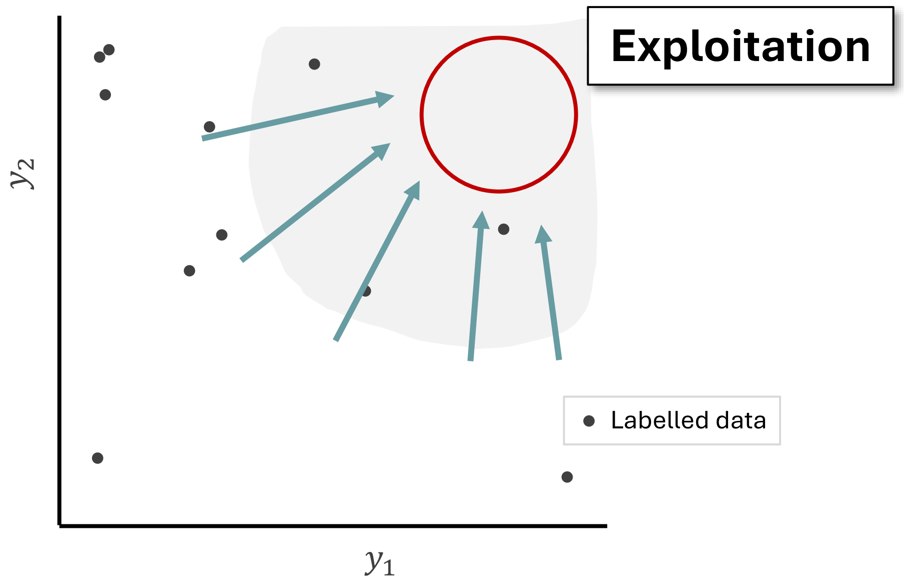
  e.g. solubility of compounds

- **Exploitation** means selecting data points that are informative and **reduce** the model **uncertainty**. Maximize, minimize or specific areas of the target property – most reward

  e.g. maximize the selectivity

- Optimal: balance the trade-off between them and find the optimal query strategy.





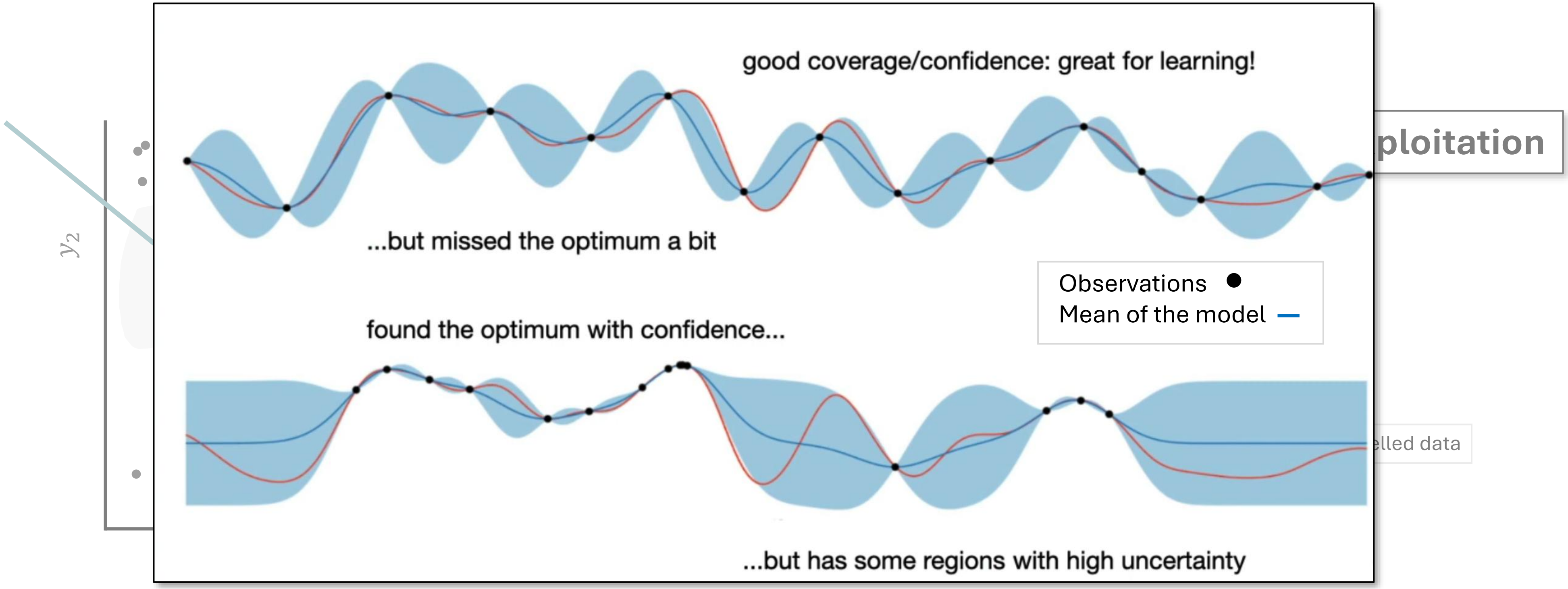
Criteria to query points to label

- **Exploration** implies selecting data points that are **diverse** and cover different regions of the feature space – long-term benefit

  e.g. solubility of compounds

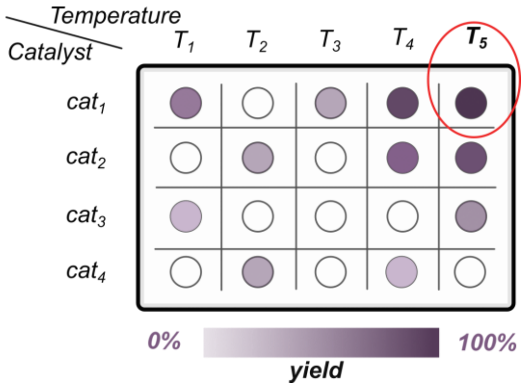
- **Exploitation** means selecting data points that are informative and **reduce** the model **uncertainty**. Maximize, minimize or specific areas of the target property – most reward
  e.g. maximize the selectivity

- Optimal: balance the trade-off between them and find the optimal query strategy.

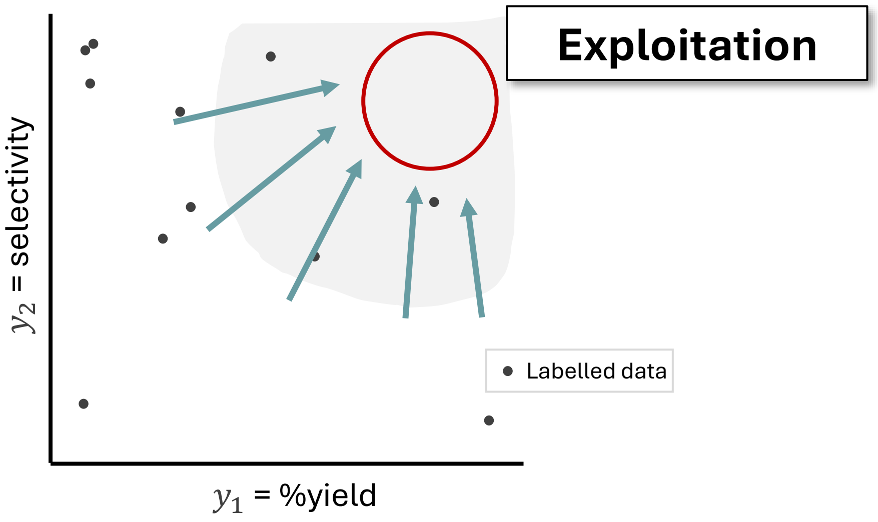


Number of objectives in chemistry

- Single objective problem y_1 = %yield



- Multiobjective problem (y_1, y_2) = (selectivity, %yield); (HOMO-LUMO gap, polarizability)



AL workflow

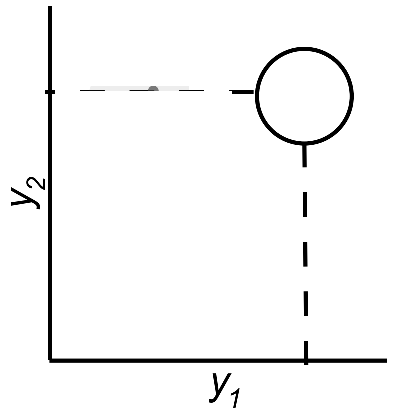
- Two target properties: y_1 (%yield) and y_2 (selectivity)

y_1 = %yield

y_2 = selectivity

labelled

predicted area



AL workflow

y_1 = %yield

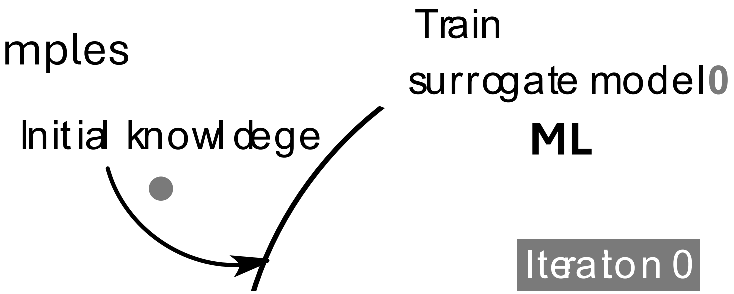
y_2 = selectivity

labelled

predicted area

Learner (surrogate model):
ML model

Most representative samples
e.g. clustering



AL workflow

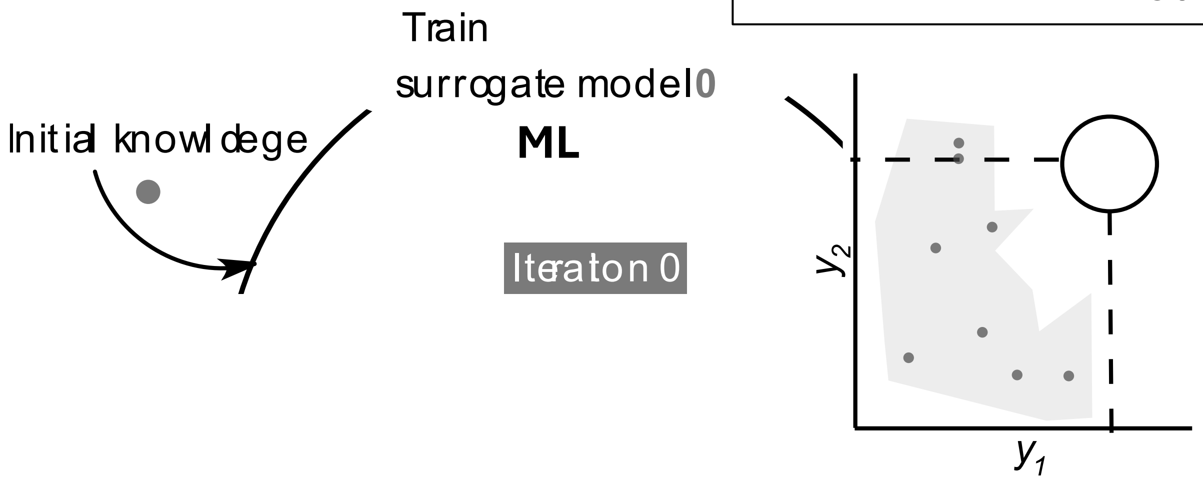
y_1 = %yield

y_2 = selectivity

labelled

predicted area

Learner (surrogate model):
ML model



Provide **predictions** and **uncertainties** for the unlabelled data points

AL workflow

y_1 = %yield

y_2 = selectivity

labelled

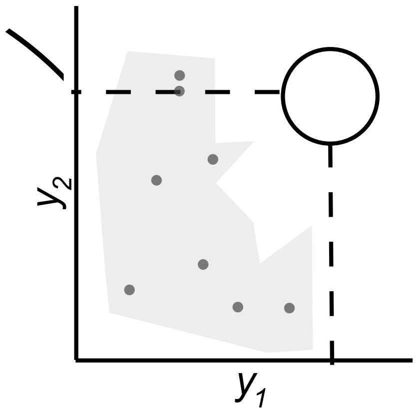
predicted area

Learner (surrogate model):
ML model

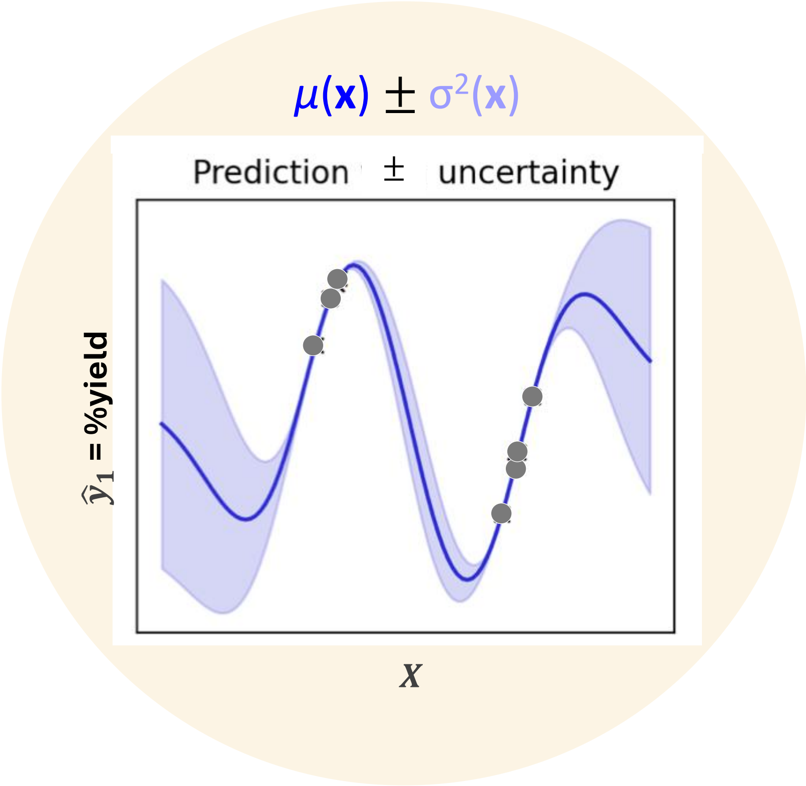
Initial knowldege

Train
surrogate model0
ML

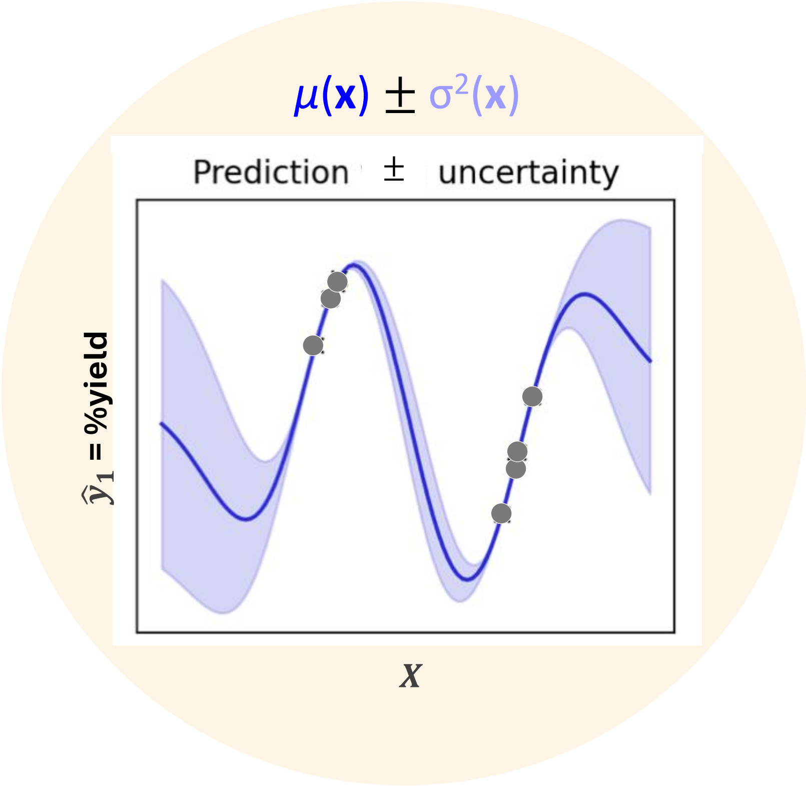
Iteraton 0

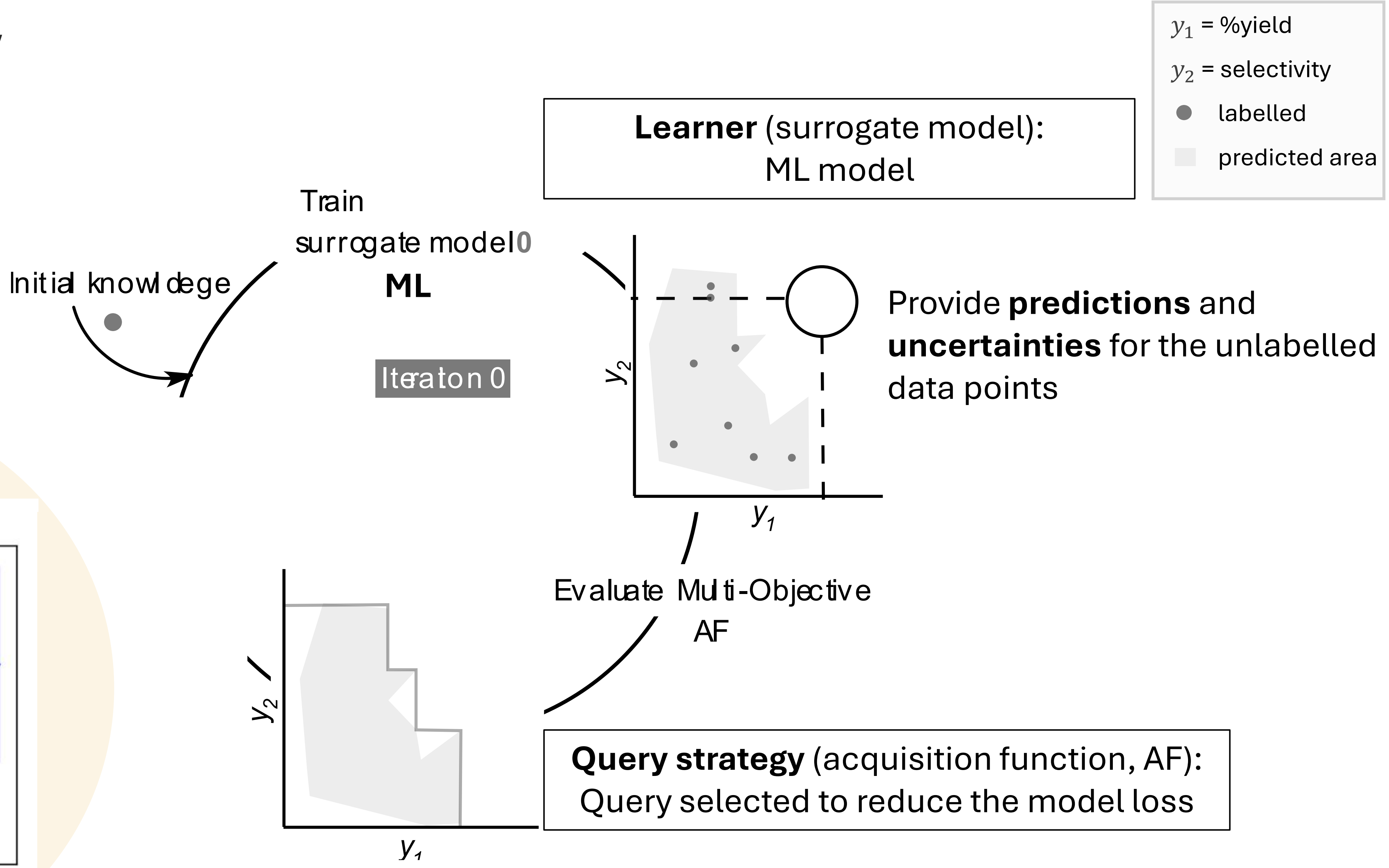


Provide **predictions** and **uncertainties** for the unlabelled data points



AL workflow





Learner ranks and select the unlabelled data points by their potential value

AL workflow

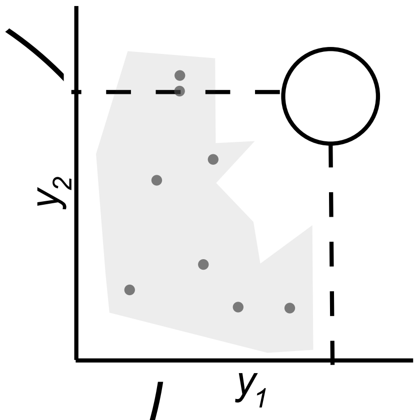
y_1 = %yield
y_2 = selectivity
labelled
predicted area

Learner (surrogate model):
ML model

Train
surrogate model0
ML

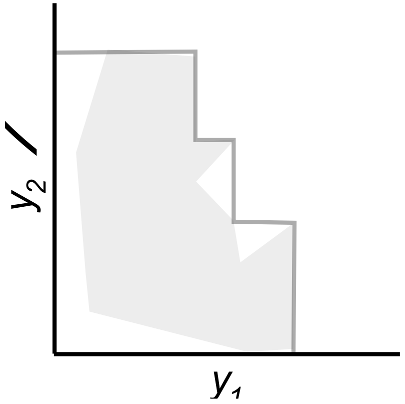
Initial knowldege

Iteraton 0

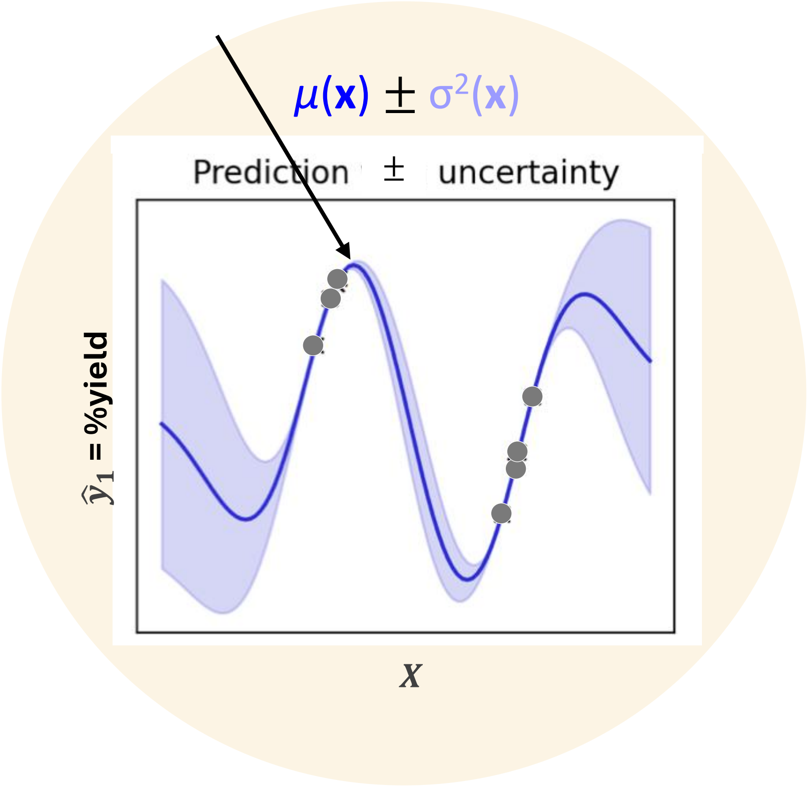


Provide **predictions** and **uncertainties** for the unlabelled data points

Evaluate Multi-Objective
AF



Query strategy (acquisition function, AF):
Query selected to reduce the model loss

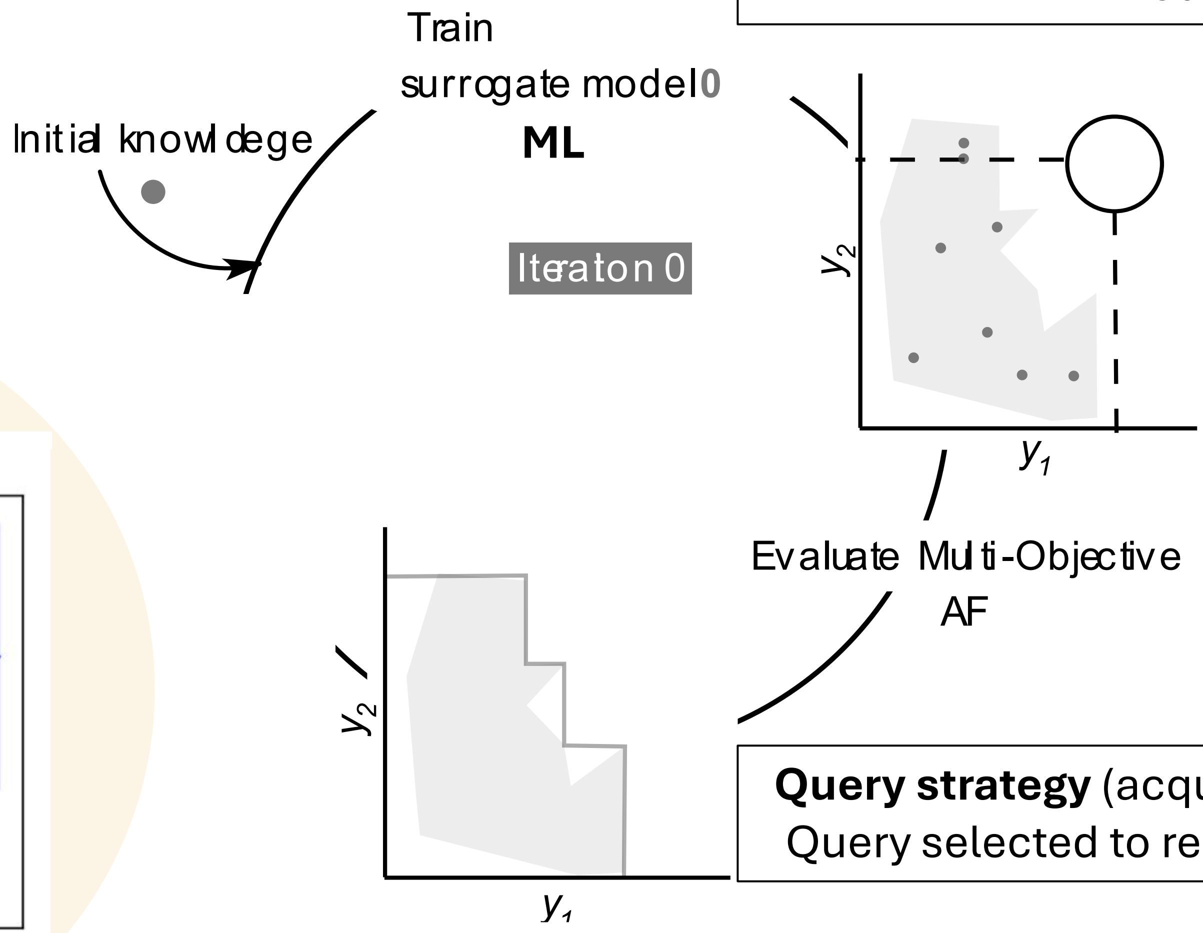


Learner ranks and select the unlabelled data points by their potential value

AL workflow

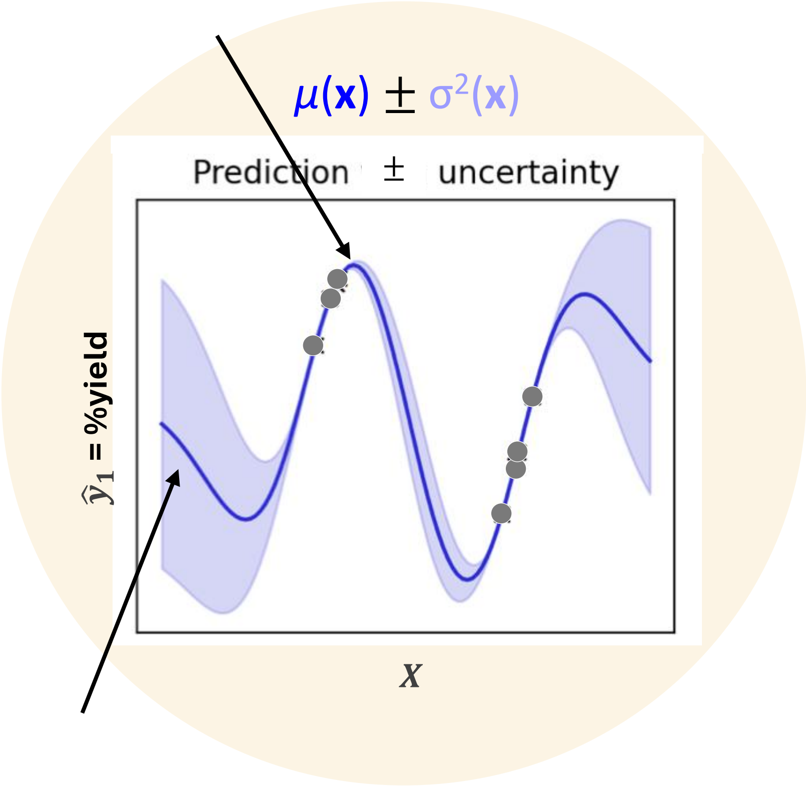
y_1 = %yield
y_2 = selectivity
labelled
predicted area

Learner (surrogate model):
ML model



Provide **predictions** and **uncertainties** for the unlabelled data points

Query strategy (acquisition function, AF):
Query selected to reduce the model loss



Learner ranks and select the unlabelled data points by their potential value

AL workflow

y_1 = %yield

y_2 = selectivity

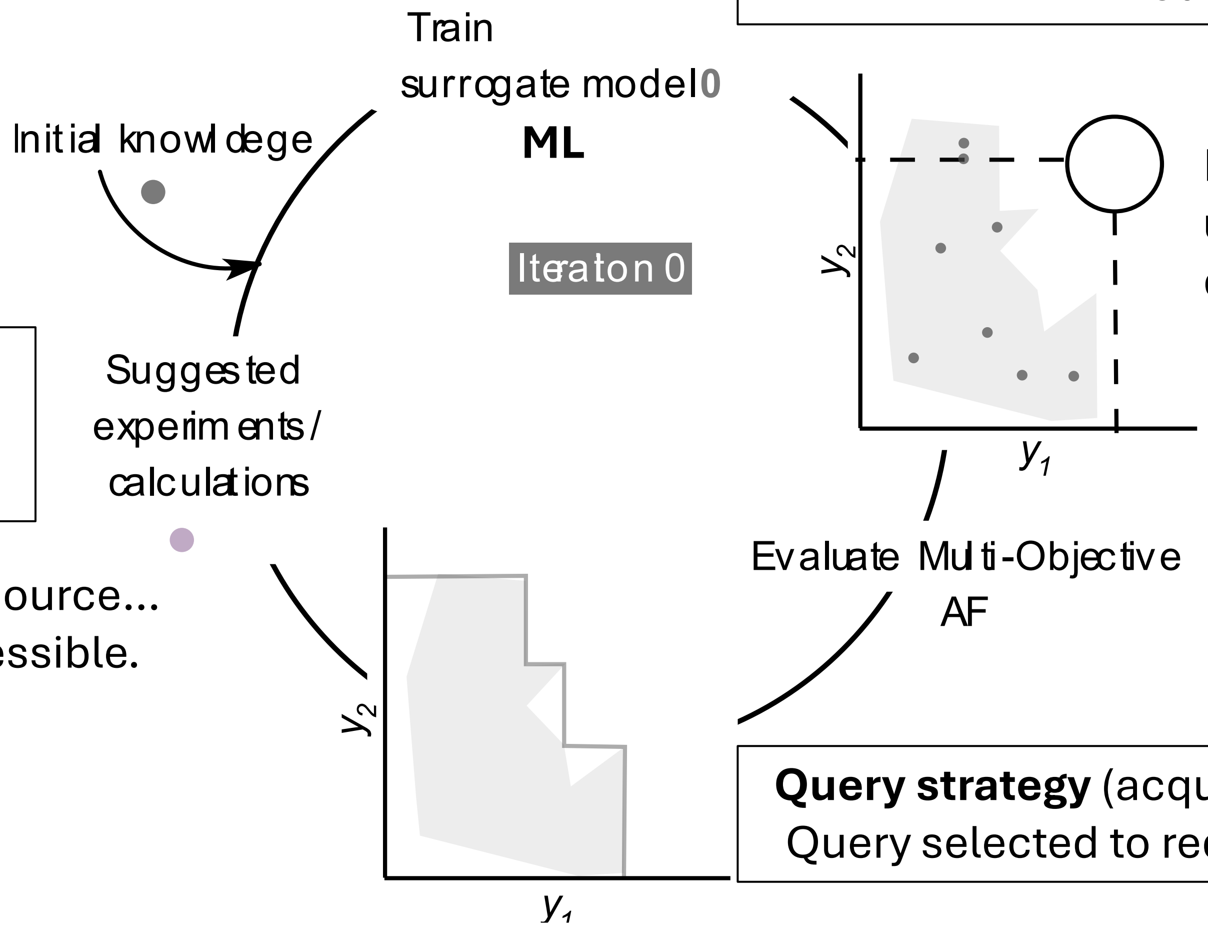
labelled

predicted area

Learner (surrogate model):
ML model

Oracle: Source of labels for the data points that are queried by the learner.

Human expert, ground truth source…
Reliable, consistent and accessible.

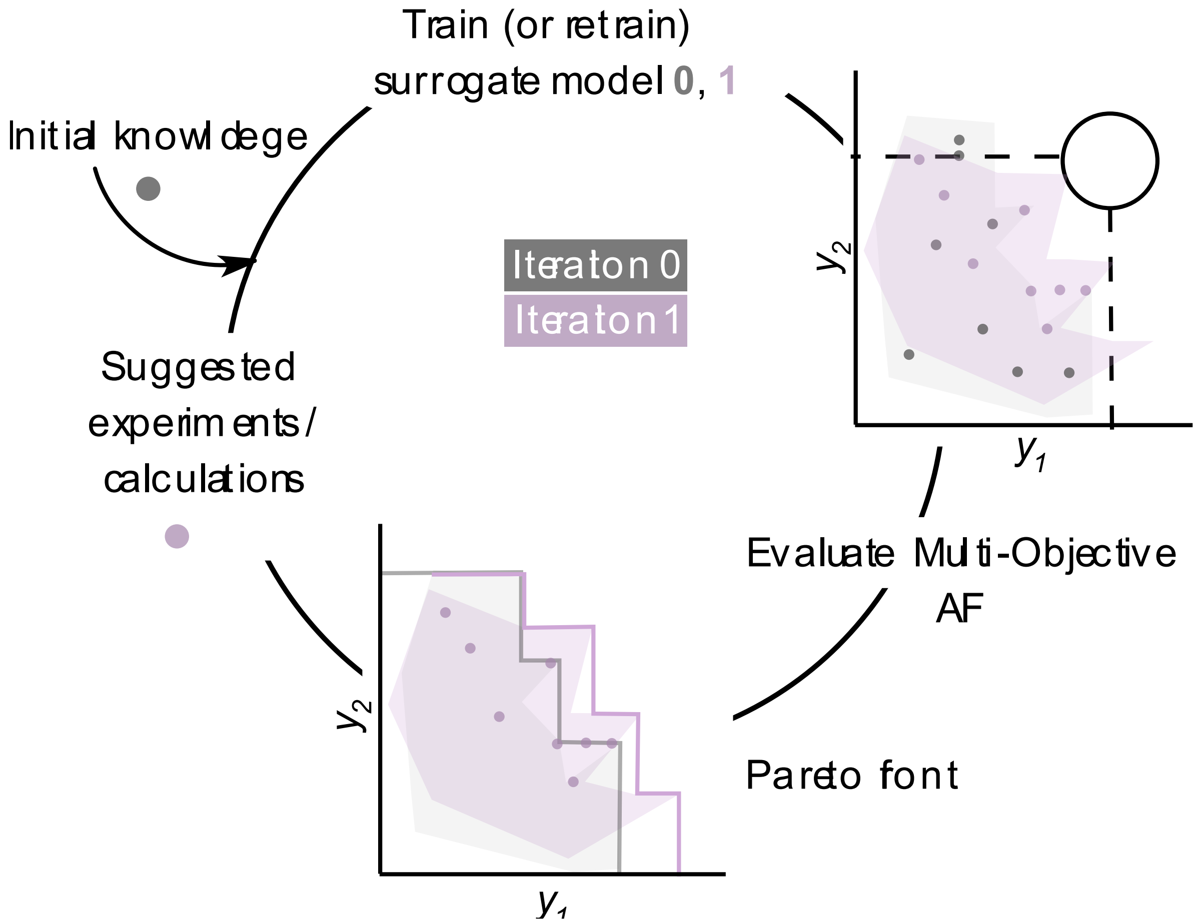


Provide **predictions** and **uncertainties** for the unlabelled data points

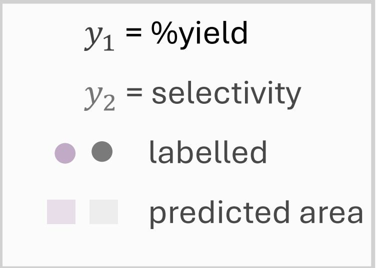
Query strategy (acquisition function, AF):
Query selected to reduce the model loss

Learner ranks and select the unlabelled data points by their potential value

AL workflow

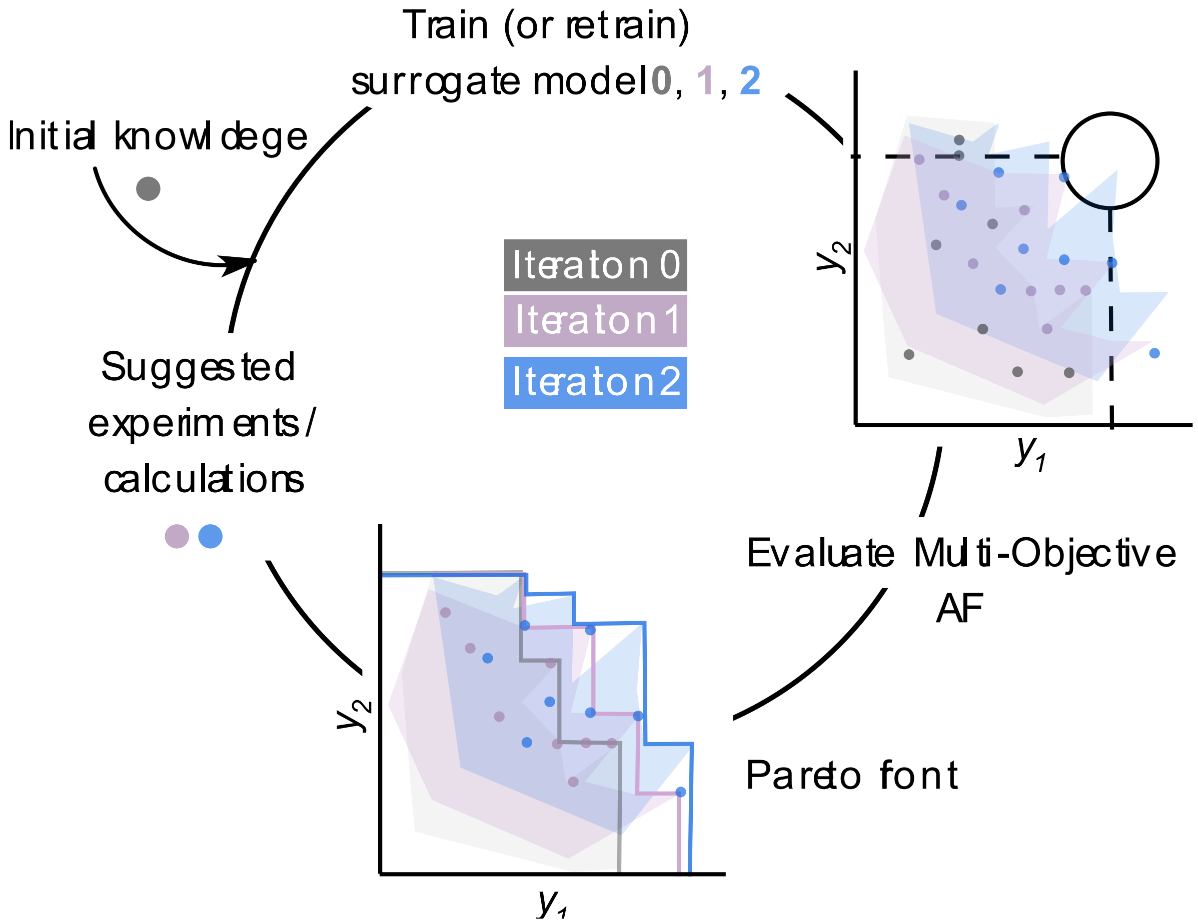


Train (or retrain)
surrogate model 0, 1
Initial knowldege
Iteraton 0
Iteraton 1
Suggested
experiments /
calculations
y_2
y_1
Evaluate Multi-Objective
AF
Pareto font
y_2
y_1

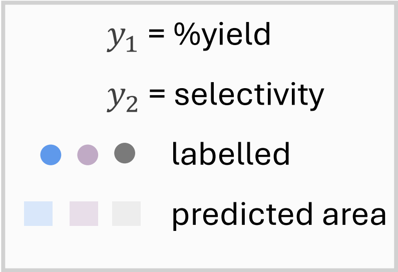


y_1 = %yield
y_2 = selectivity
labelled
predicted area

AL workflow

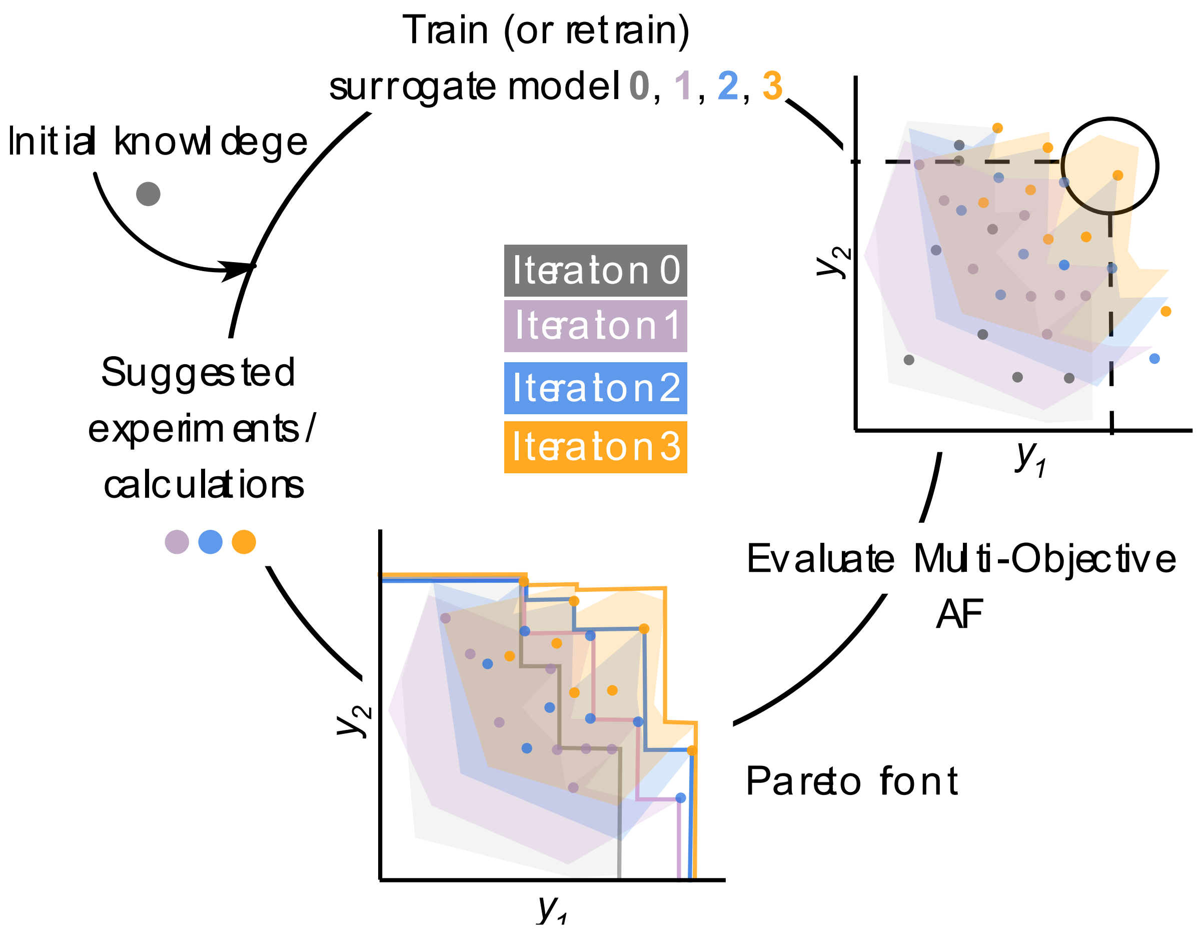


Train (or retrain)
surrogate model 0, 1, 2
Initial knowldege
Suggested experiments / calculations
Iteraton 0
Iteraton 1
Iteraton 2
y_2
y_1
Evaluate Multi-Objective AF
Pareto font
y_2
y_1



y_1 = %yield
y_2 = selectivity
labelled
predicted area

AL workflow



y_1 = %yield
y_2 = selectivity
labelled
predicted area

AL components

Parameter	Tools	Comments
Learner	Gaussian Processes	Gold standard: $\mu(\mathbf{x}) \pm \sigma^2(\mathbf{x})$
	Neural Network	Based on latent space, Euclidean distance
Acquisition function	Expected Improvement (EI); Probability of improvement (PI)	Highest probability of improvement the current best solution
	Lower/Upper Confidence Bound (UCB)	Balance exploration - exploitation
Oracle	DFT calculations, experiments	

Applications of AL in chemistry

- **Reaction optimization**

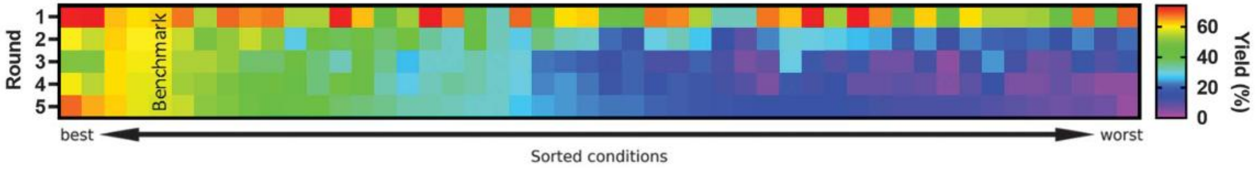
Find the best conditions (temperature, solvent, catalyst, etc.) to maximize yield/selectivity

RESEARCH

ORGANIC CHEMISTRY

Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling

Nicholas H. Angello[1,2]†, Vandana Rathore[1,2]†, Wiktor Beker[3], Agnieszka Wołos[3,4], Edward R. Jira[2,5], Rafał Roszak[3,4], Tony C. Wu[6,7], Charles M. Schroeder[1,2,5,8], Alán Aspuru-Guzik[6,7,9,10,11,12], Bartosz A. Grzybowski[3,4,13,14]*, Martin D. Burke[1,2,15,16,17]*



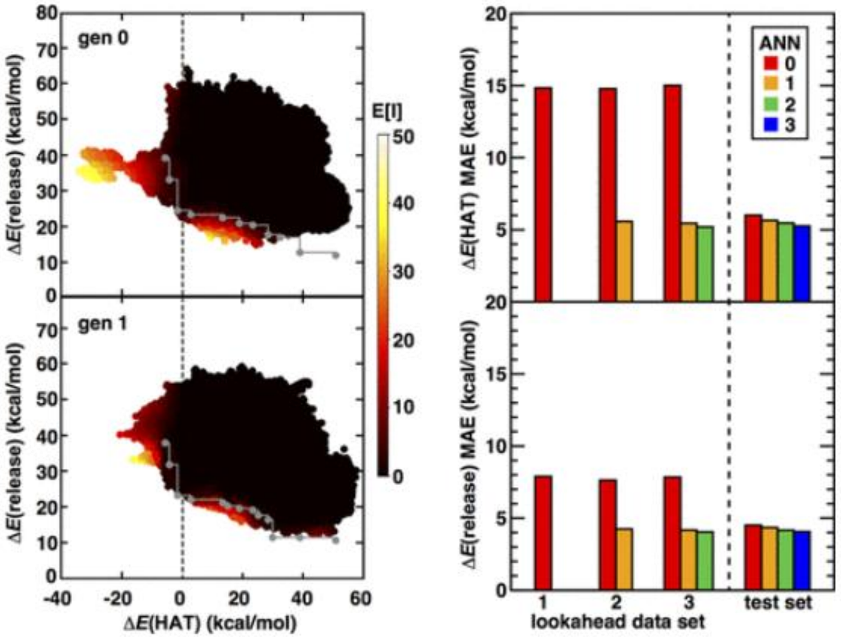
- **Catalyst design**

Millions of possible metal/ligand combinations

ARTICLE | April 27, 2022

New Strategies for Direct Methane-to-Methanol Conversion from Active Learning Exploration of 16 Million Catalysts

Aditya Nandy, Chenru Duan, Conrad Goffinet, and Heather J. Kulik*



EDBO+: Bayesian Optimization platform

- Graphical User Interface: https://edboplus.org/
- Interactive platform to guide chemists to optimize single and multiple objective in a reaction
- https://www.youtube.com/watch?v=Fo_ZplPyLZo

JACS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS Article

A Multi-Objective Active Learning Platform and Web App for Reaction Optimization

Jose Antonio Garrido Torres,[#] Sii Hong Lau,[#] Pranay Anchuri,[#] Jason M. Stevens, Jose E. Tabora, Jun Li, Alina Borovika, Ryan P. Adams, and Abigail G. Doyle[*]