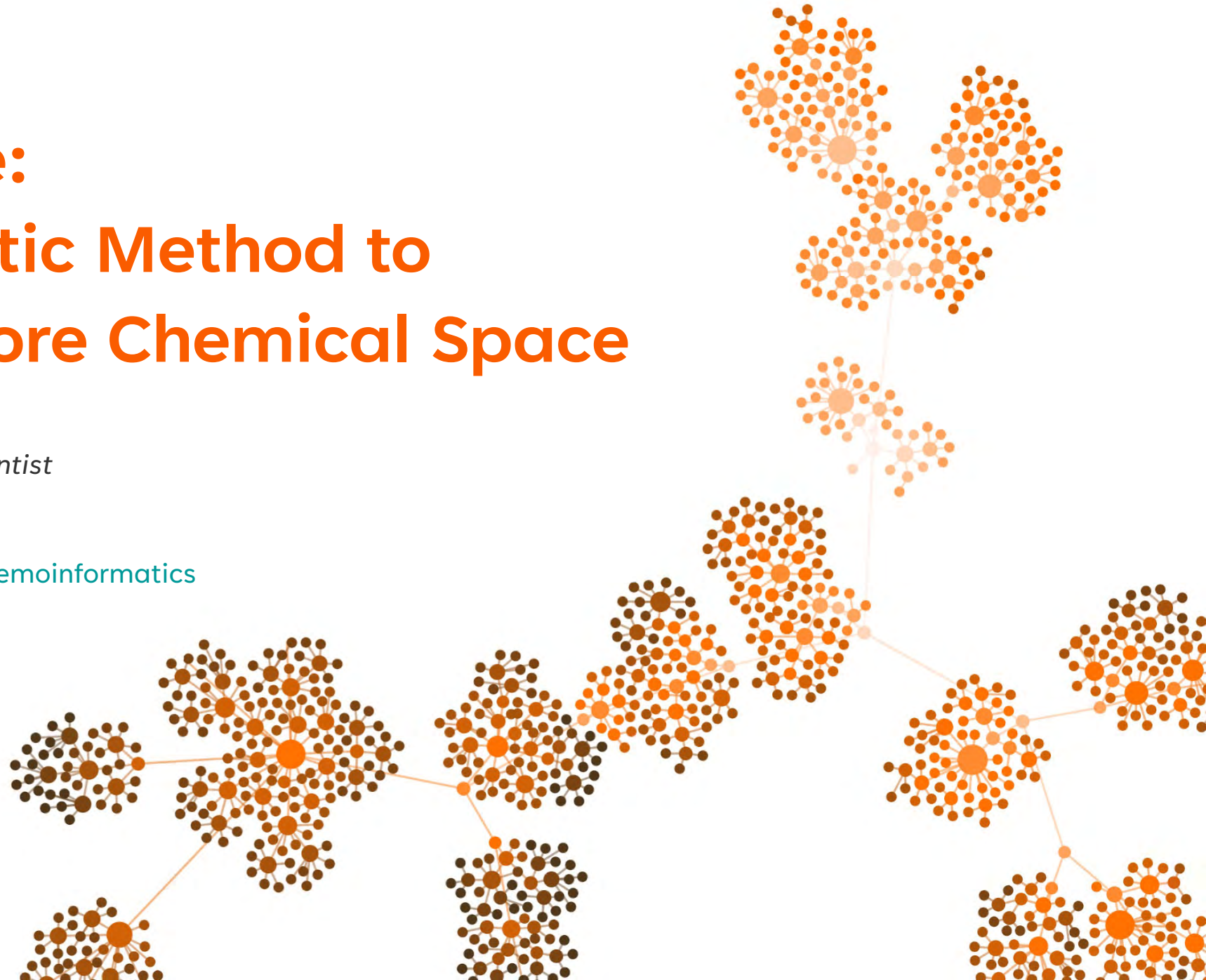
Coverage Score: A Model Agnostic Method to Efficiently Explore Chemical Space

Dan Woodward
Senior Cheminformatics Research Scientist

19 June 2023
Ninth Joint Sheffield Conference on Chemoinformatics

doi/10.1021/acs.jcim.2c00258



Exscientia

Outline

Active learning in drug discovery

- Why is it useful?

Query strategies

- How to select molecules?

Coverage Score

- How does it work?

Validation

- How does Coverage Score perform?

Further work/summary

- Where do we go from here?

Outline

Active learning in drug discovery

- Why is it useful?

Query strategies

- How to select molecules?

Coverage Score

- How does it work?

Validation

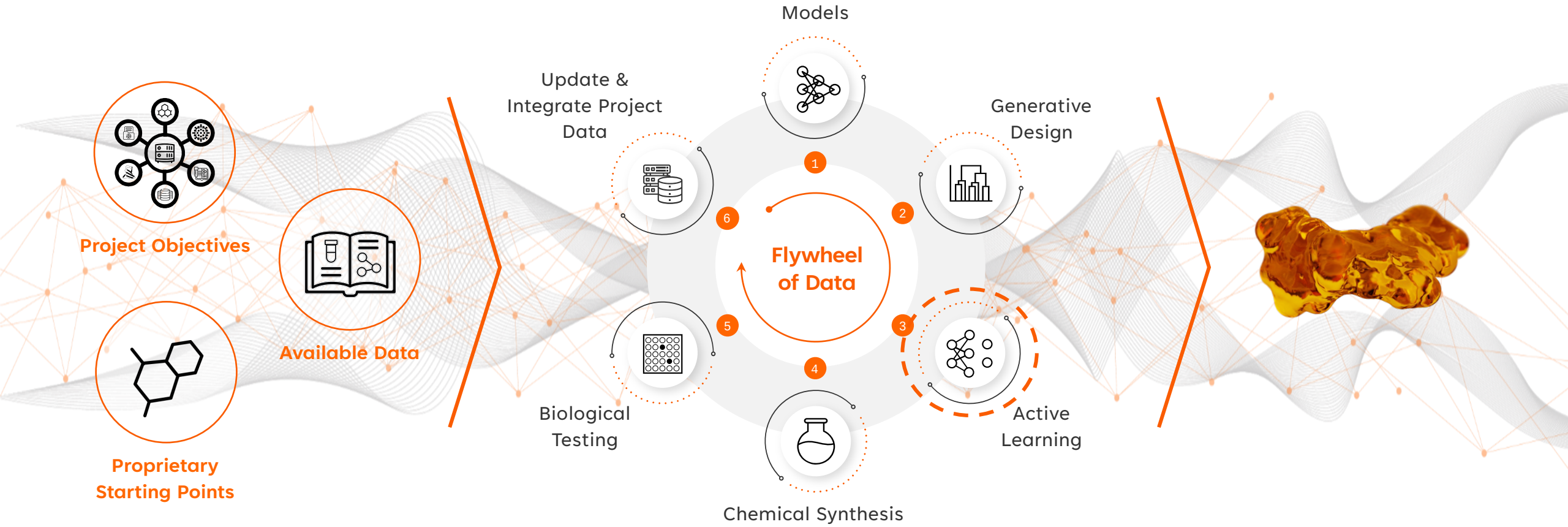
- How does Coverage Score perform?

Further work/summary

- Where do we go from here?

AI-driven design to generate candidate drugs

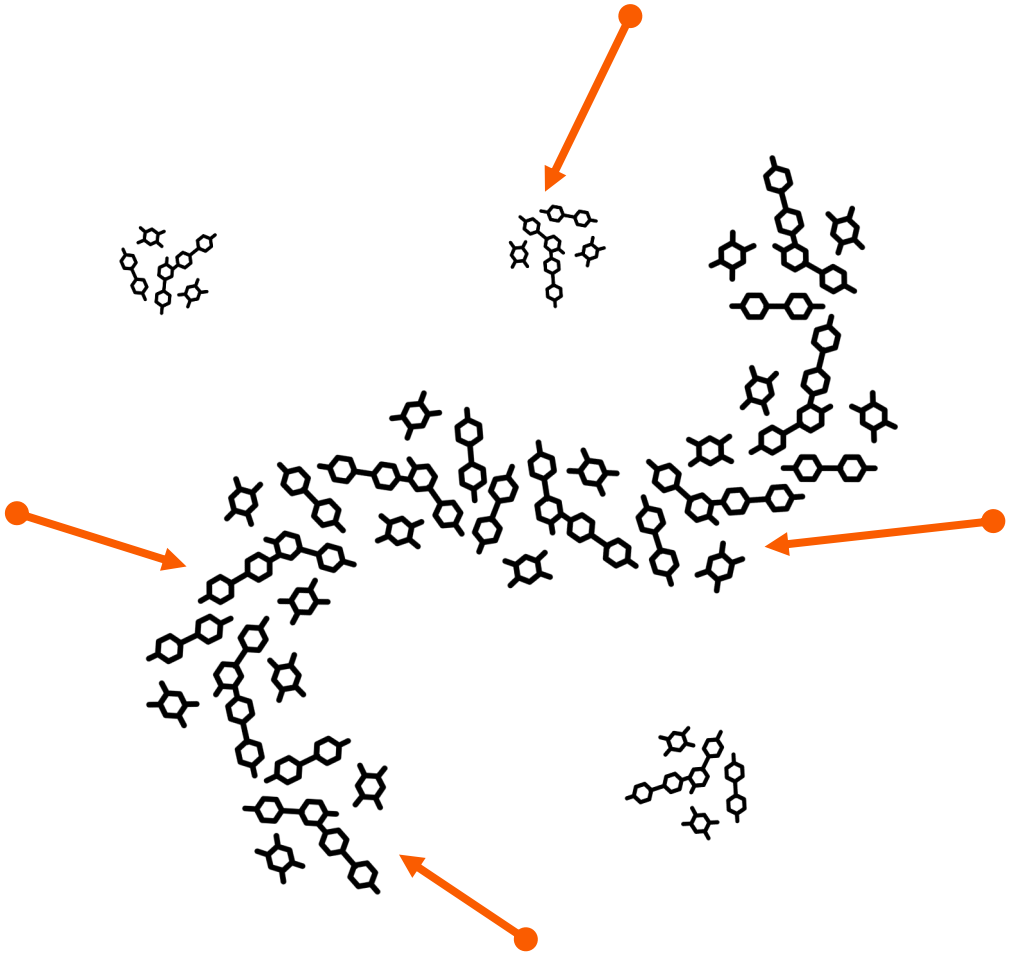
Drug discovery is a learning problem



Small numbers of compounds per design cycle

Why can’t we just screen molecules?

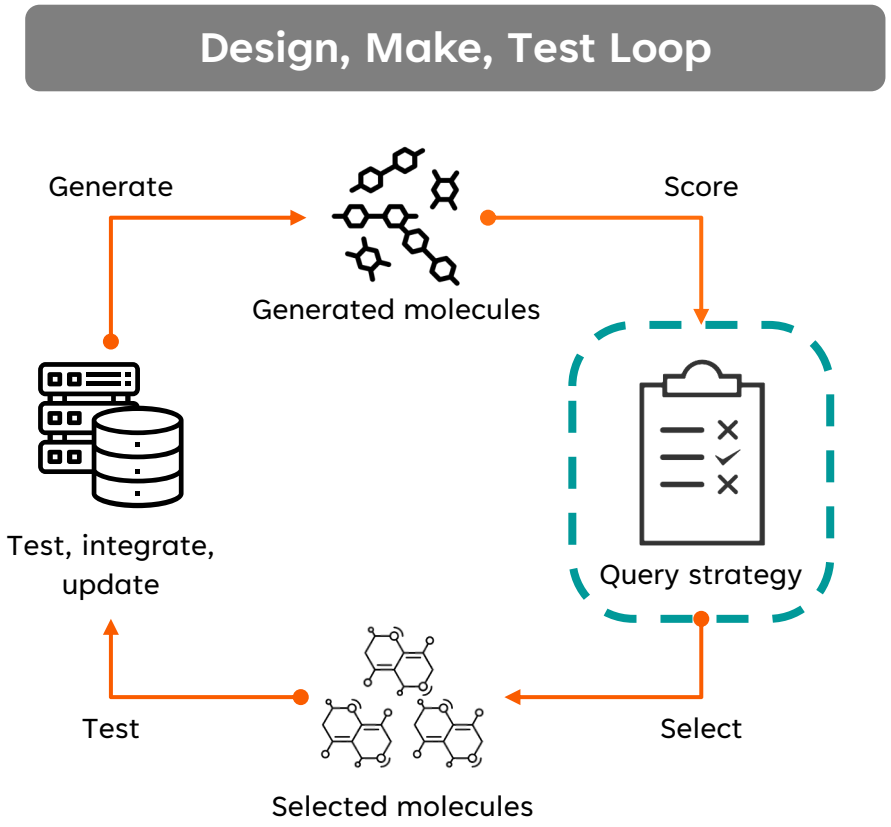
- Druggable chemical space is **huge**! (~10^{60})[1]
- **Slow and costly** to synthesise and assay molecules
- Comparatively **cheap** and fast to run predictive models
- Low data regime, predictive models **less accurate**
- Iteratively decide which molecules are ‘best’ to test



1. Bohacek, R. S., McMartin, C., & Guida, W. C. (1996). The art and practice of structure-based drug design: a molecular modeling perspective. Medicinal research reviews, 16(1), 3–50.

Active learning

- Selecting highly scoring molecules **exploits** what model already knows (minimal information gain)
- Improve model predictions - learn **efficiently**
- More accurate predictions **earlier**, better decisions, **faster** time to candidate
- Query strategies can be data- or model-dependent



Outline

Active learning in drug discovery

- Why is it useful?

Query strategies

- How to select molecules?

Coverage Score

- How does it work?

Validation

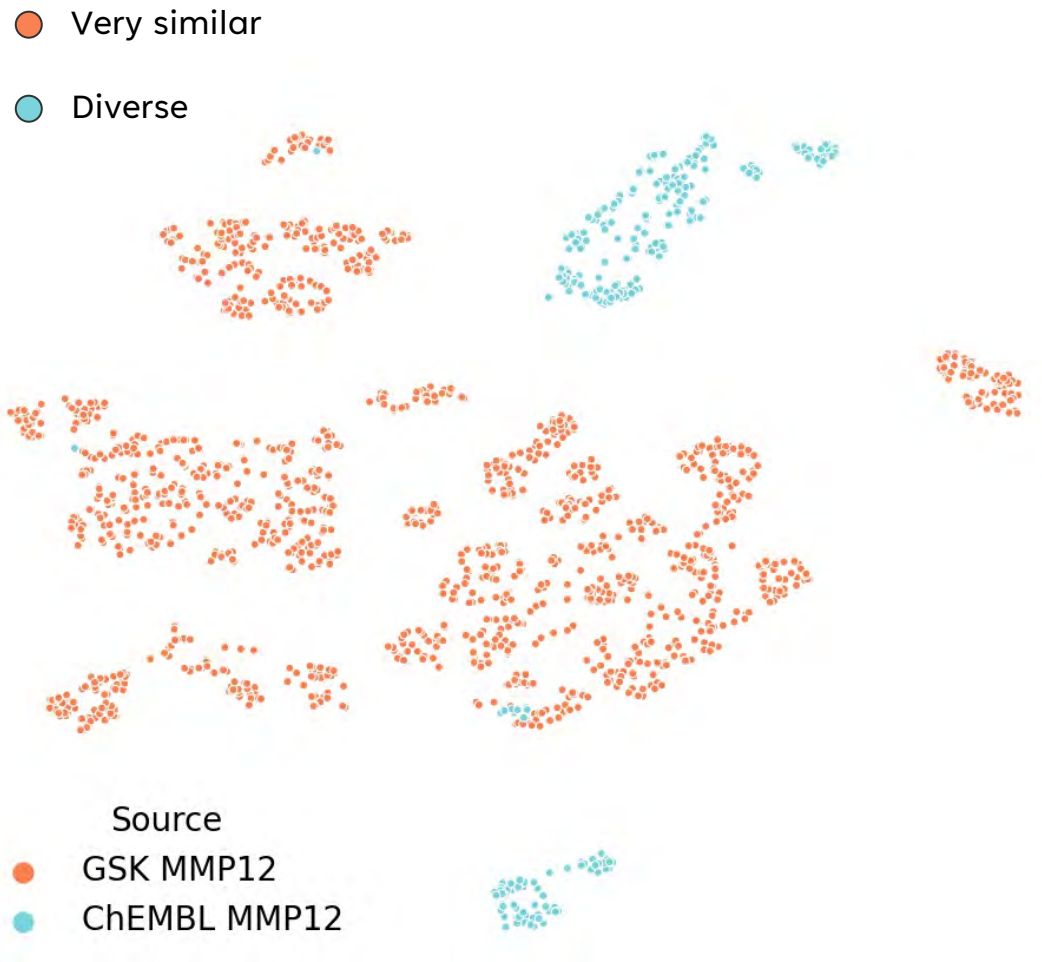
- How does Coverage Score perform?

Further work/summary

- Where do we go from here?

Query strategy comparison

- **Dataset**
  - *x* = molecules from GSK MMP12 set (similar) and ChEMBL (dissimilar)
  - *y* = experimentally determined pIC_{50} values for MMP12
- **Data-dependent**
  - **Diversity**
    maximal dissimilarity
  - **KMeans**
    clustering
  - **Coverage Score**
    Bayesian statistics + information entropy
- **Model-dependent**
  - **Exploitation**
    highest predictive score
  - **Uncertainty**
    highest uncertainty in predictive score



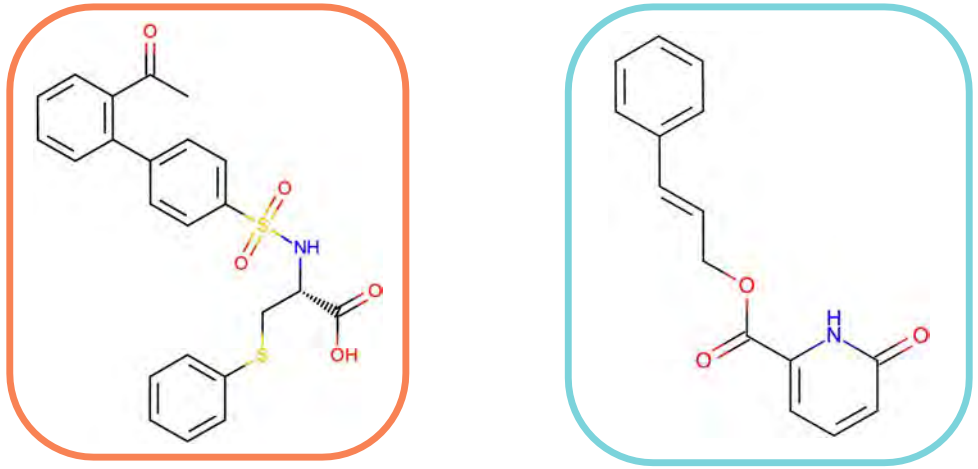
t-SNE plot of D2+ split by D2 (orange) and ChEMBL compounds (pale blue)

Query strategy comparison

- **Dataset**
  - *x* = molecules from GSK MMP12 set (similar) and ChEMBL (dissimilar)
  - *y* = experimentally determined pIC_{50} values for MMP12
- **Data-dependent**
  - **Diversity**
    maximal dissimilarity
  - **KMeans**
    clustering
  - **Coverage Score**
    Bayesian statistics + information entropy
- **Model-dependent**
  - **Exploitation**
    highest predictive score
  - **Uncertainty**
    highest uncertainty in predictive score

Very similar

Diverse



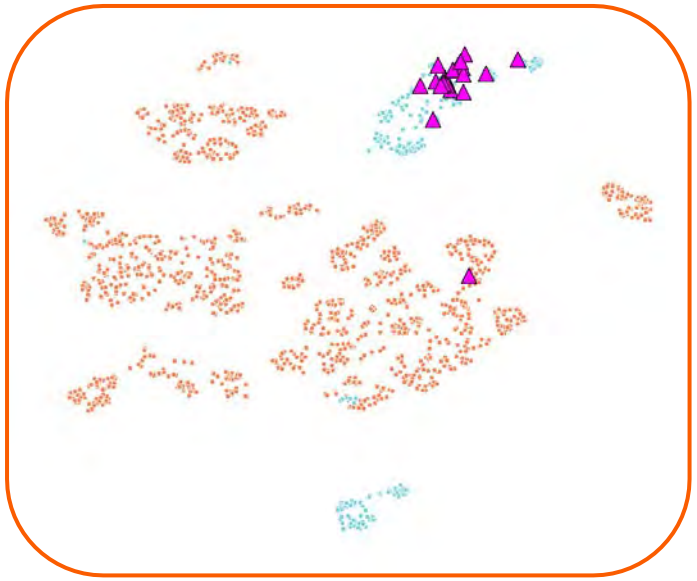
t-SNE plot of D2+ split by D2 (orange) and ChEMBL compounds (pale blue)

Query strategy comparisons

Data-dependent

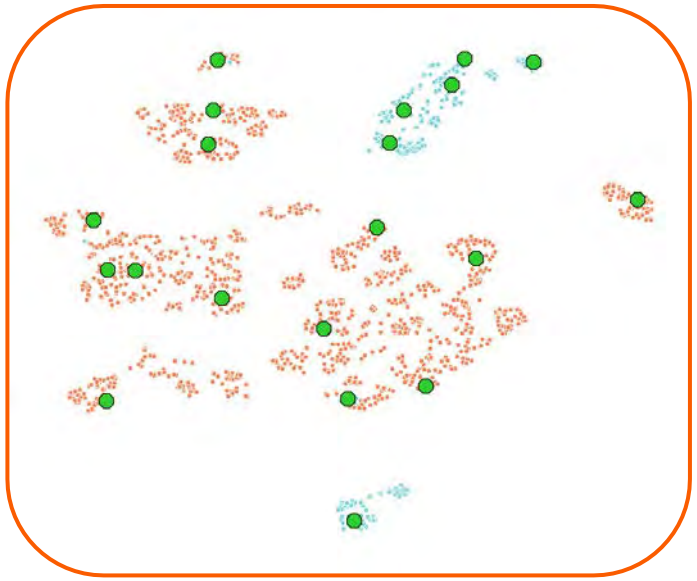
Very similar

Diverse



Diversity
maximal dissimilarity

May select outlier compounds



KMeans
clustering

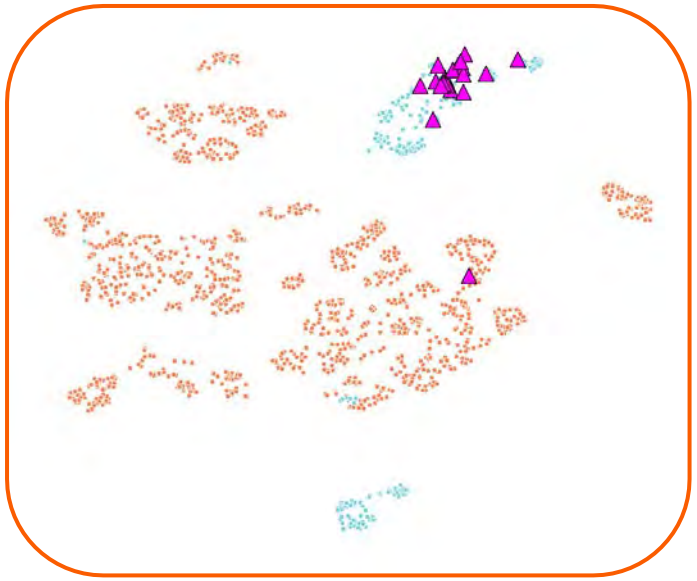
May not provide enough diversity

Query strategy comparisons

Data-dependent

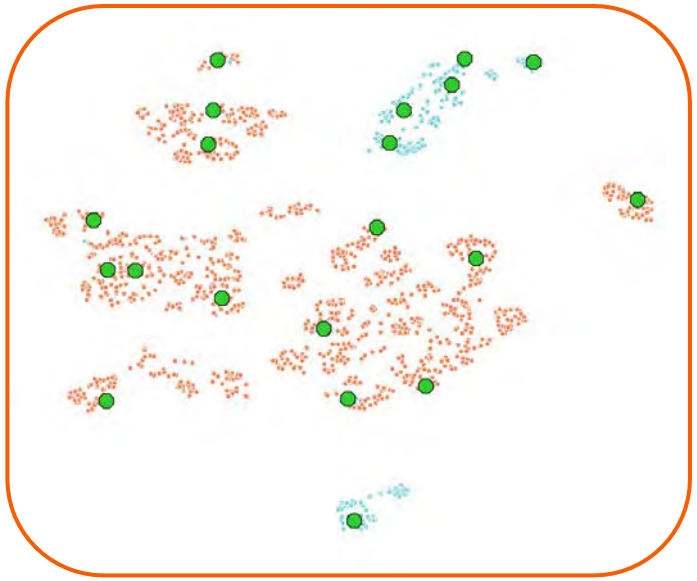
Very similar

Diverse



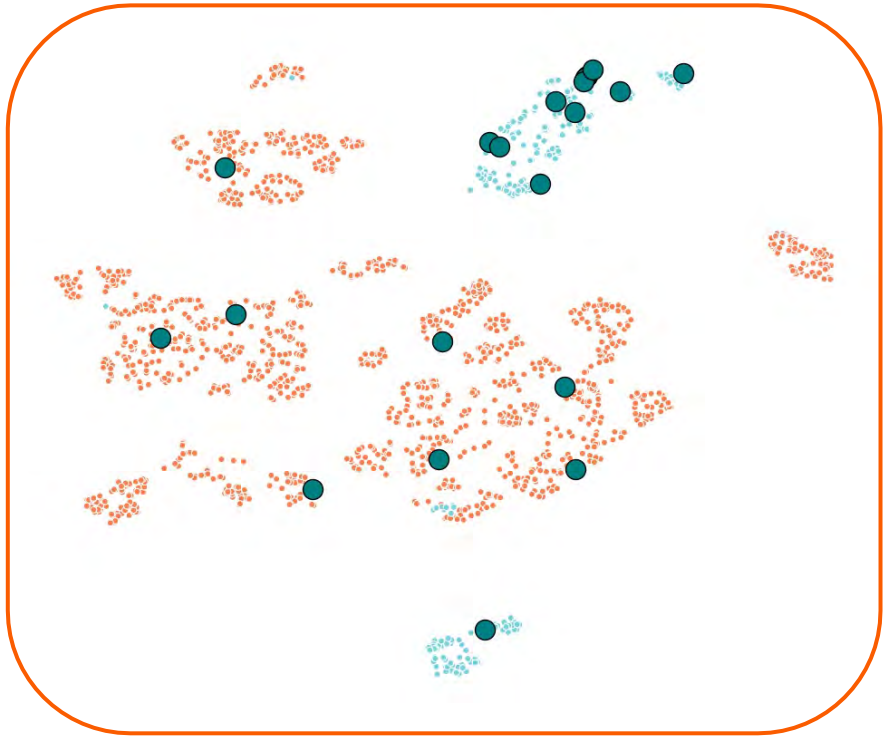
Diversity
maximal dissimilarity

May select outlier compounds



KMeans
clustering

May not provide enough diversity



Coverage Score
Bayesian statistics + information entropy

Useful coverage of chemical space

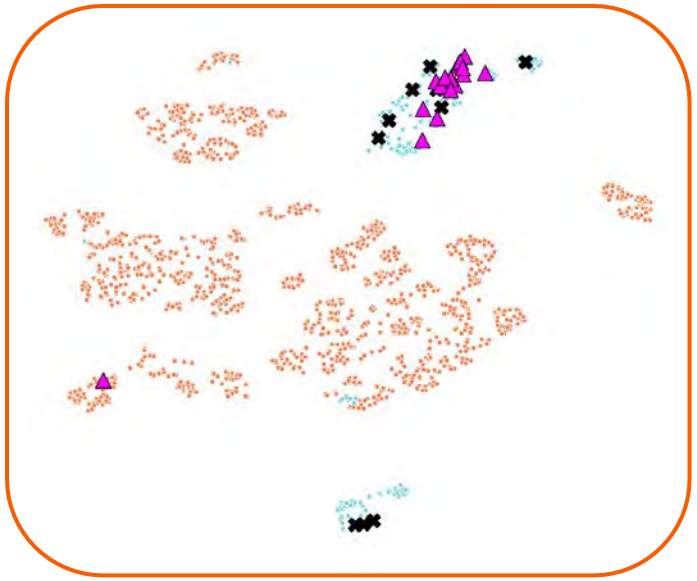
Query strategy comparisons

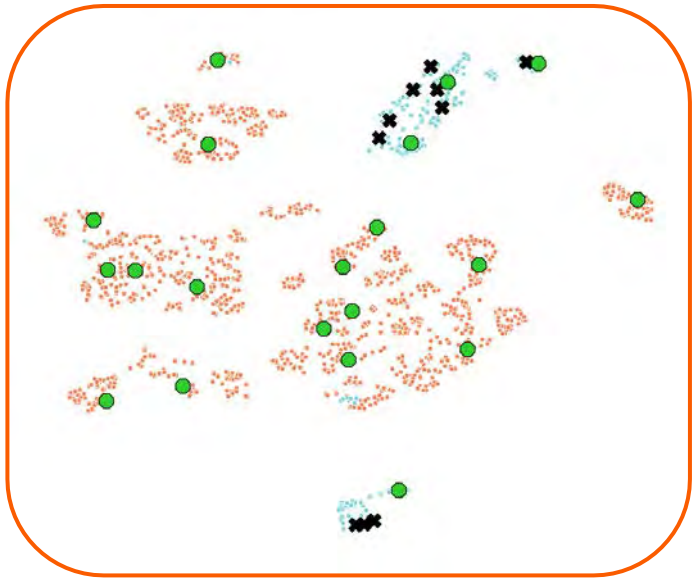
Data-dependent: Dissimilar prior selections

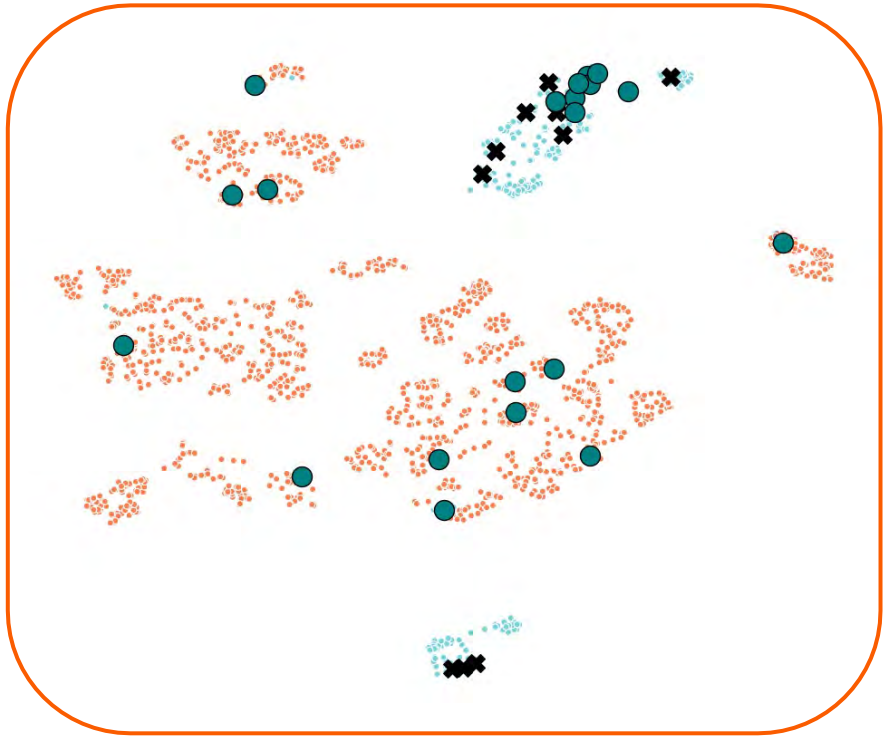
Very similar

Diverse

Prior selections







Diversity
maximal dissimilarity

Misses (almost) all the dense region

KMeans
clustering

Essentially unperturbed by prior selections

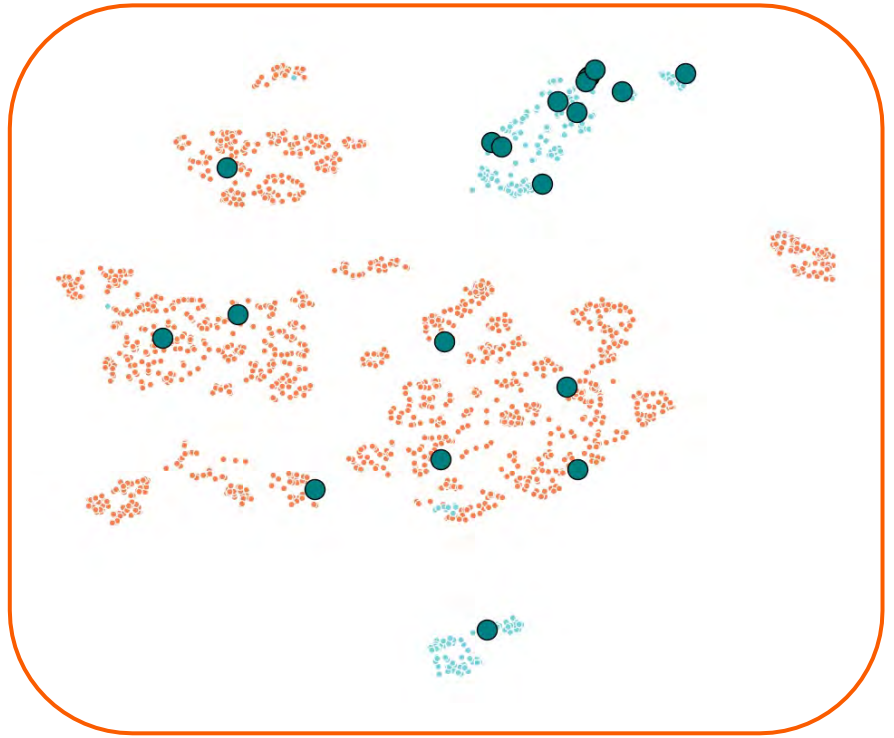
Coverage Score
Bayesian statistics + information entropy

Samples from and based regions with more focus on

Query strategy comparisons

Model-dependent

- Very similar
- Diverse



Exploitation
highest model score

Needs an initial training set (and model)!

Uncertainty
highest uncertainty in score

Needs an initial training set (and model)!

Coverage Score
Bayesian statistics + information entropy

Useful coverage of chemical space

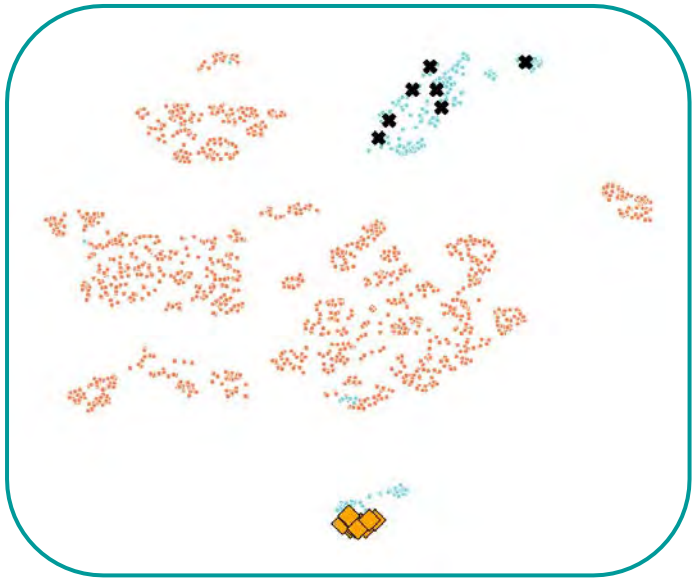
Query strategy comparisons

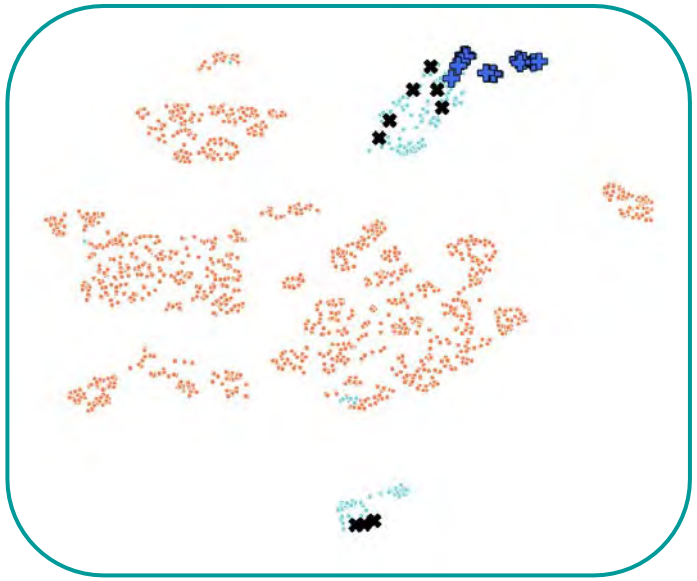
Model-dependent: Dissimilar prior selections

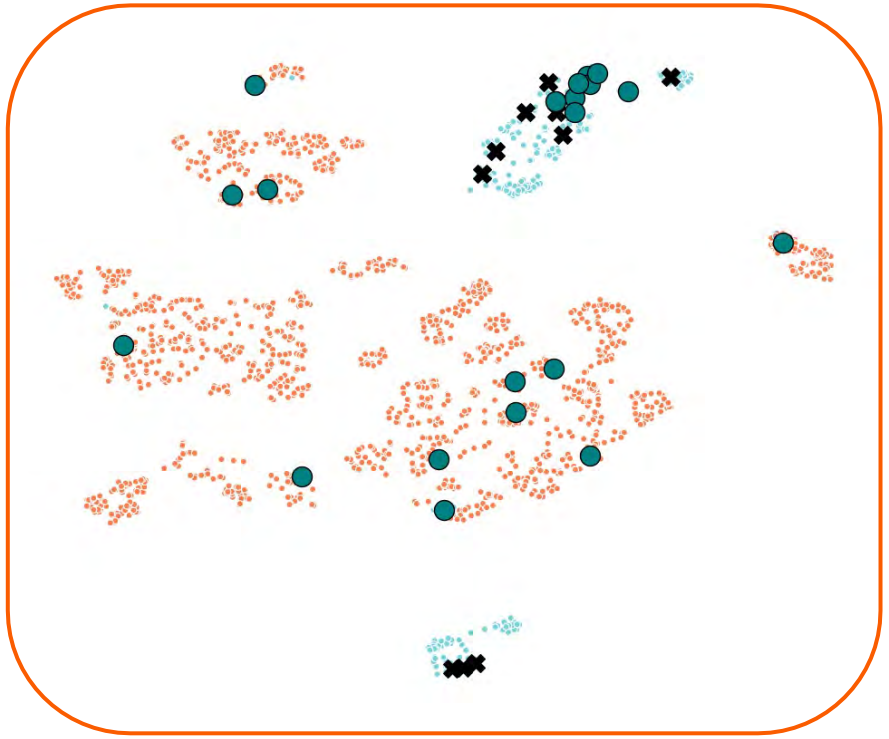
Very similar

Diverse

Prior selections







Exploitation
highest model score

Only thinks compounds are high scoring

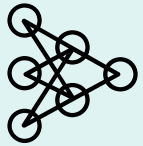
Uncertainty
highest uncertainty in score

Is certain that compounds are low in score

Coverage Score
Bayesian statistics + information entropy

Samples from and based regions with more focus on

Query strategies overview



Model-dependent

Acquisition functions, maximum uncertainty, highest score, expected improvement

Require model and often an uncertainty estimate

If uncertainty is poorly correlated to error in prediction (low data), less useful (and vice versa)

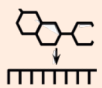
Batch selection may require pseudo-labelled model retraining

Prior molecules can be accounted for via uncertainty metric

Data-dependent

Clustering, maximal dissimilarity, **Coverage Score**

Model-independent



Representation (and/or distance metric) required

Batch selection done greedily or using optimisation

Prior molecules can be accounted for as seed compounds

Outline

Active learning in drug discovery
- Why is it useful?

Query strategies
- How to select molecules?

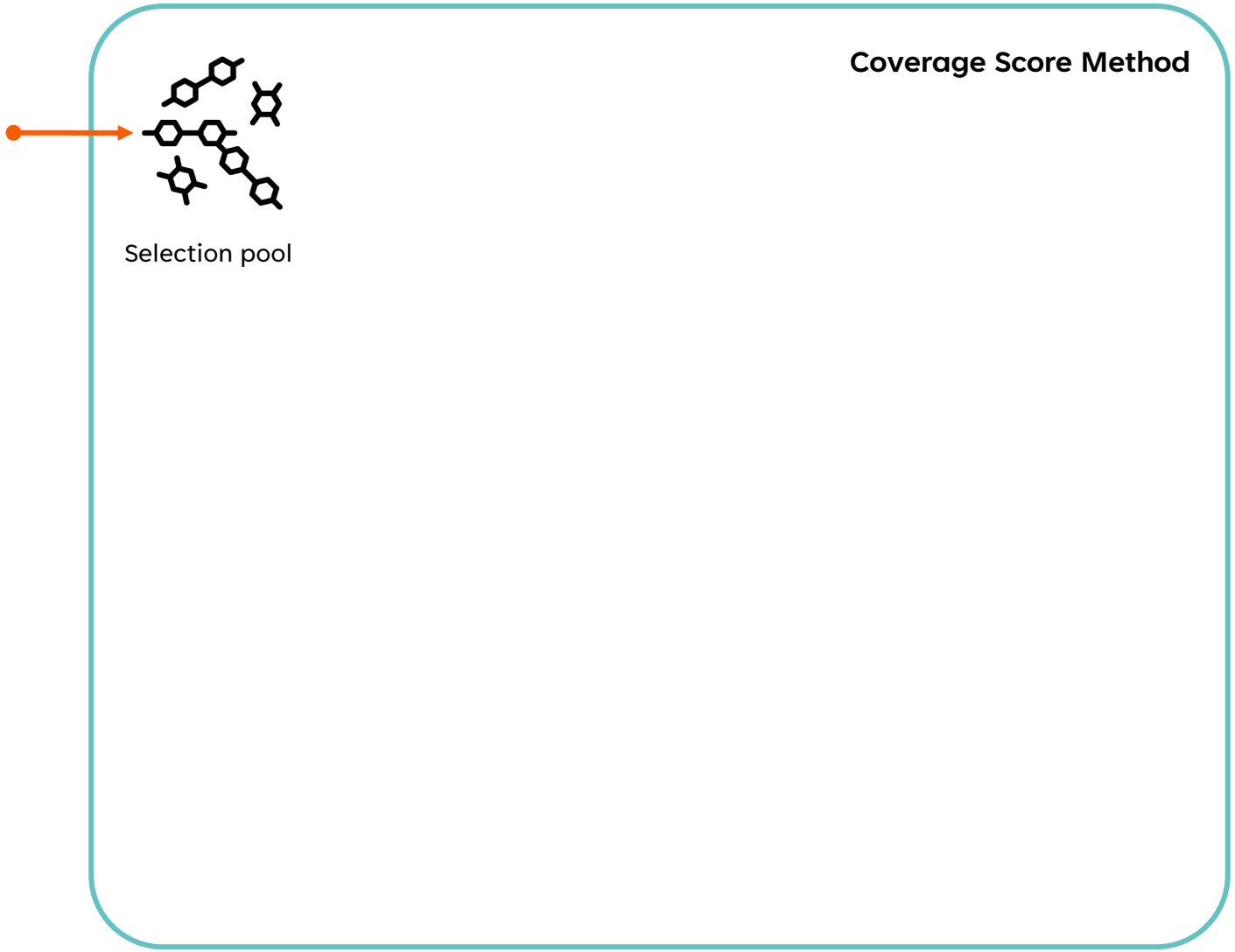
Coverage Score
- How does it work?

Validation
- How does Coverage Score perform?

Further work/summary
- Where do we go from here?

So, what is Coverage Score?

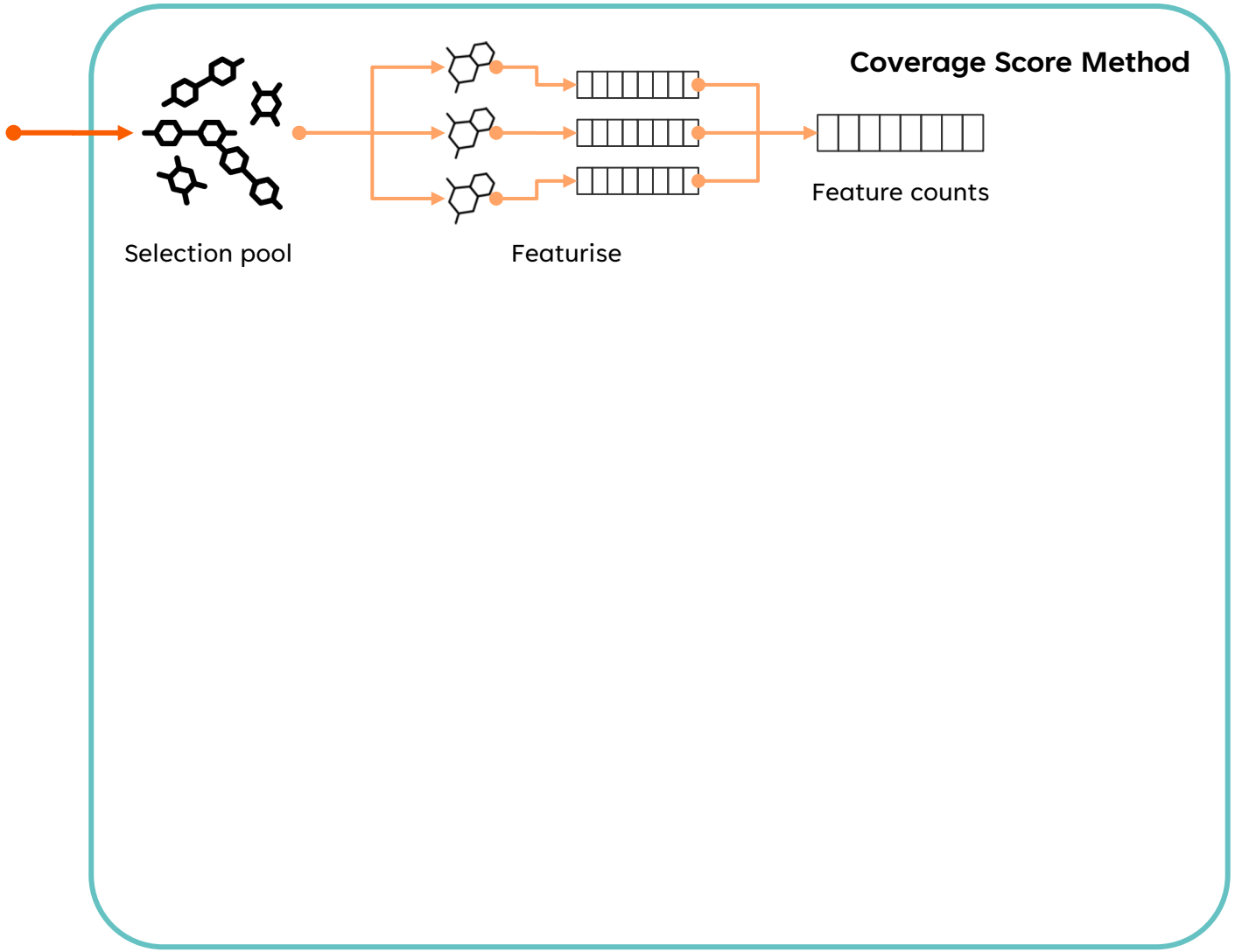
- **Data-dependent** optimisation-based[1] query strategy



1. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE T Evolut. Comput. 2002, 6, 182, DOI: 10.1109/4235.996017

So, what is Coverage Score?

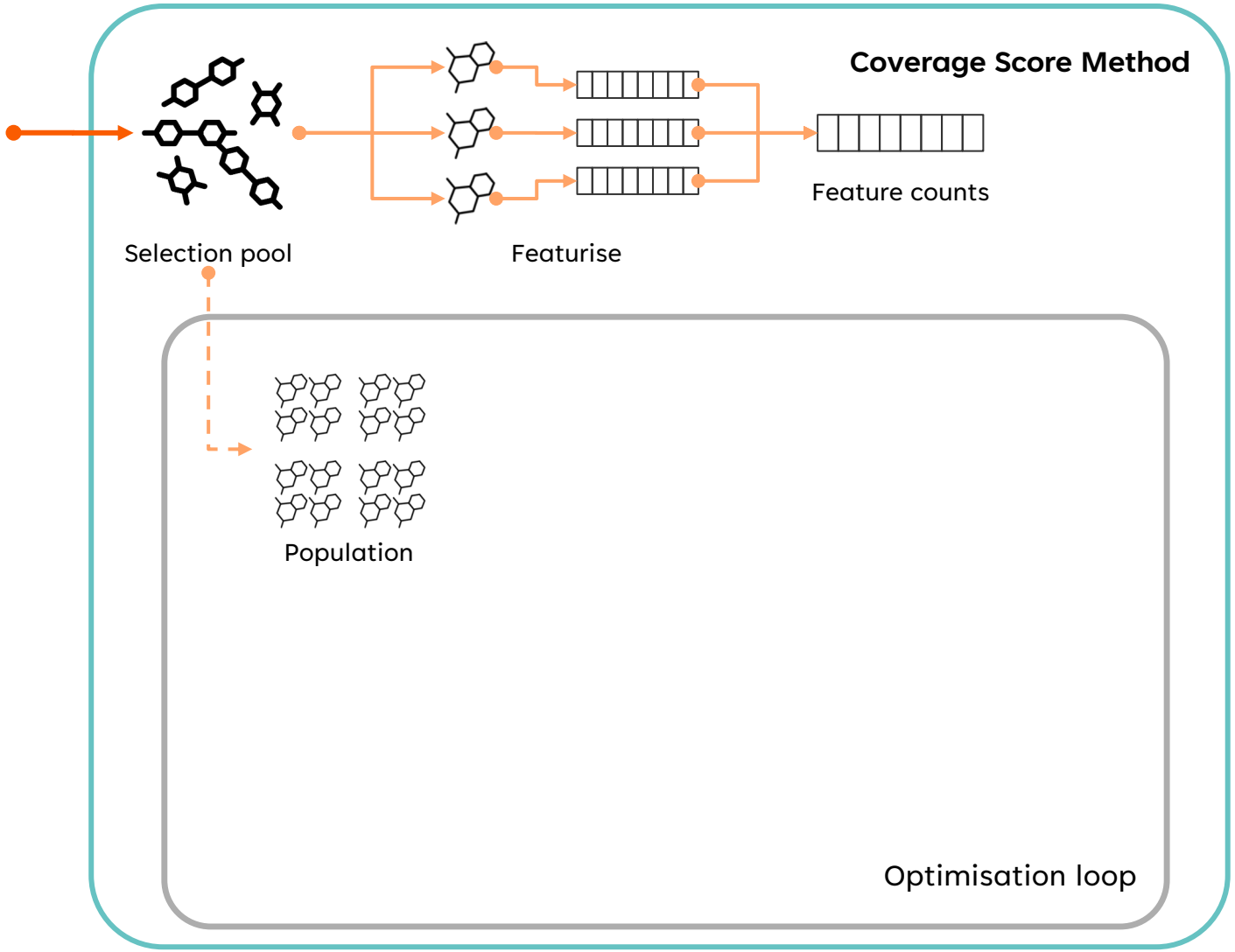
- **Data-dependent** optimisation-based[1] query strategy



1. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE T Evolut. Comput. 2002, 6, 182, DOI: 10.1109/4235.996017

So, what is Coverage Score?

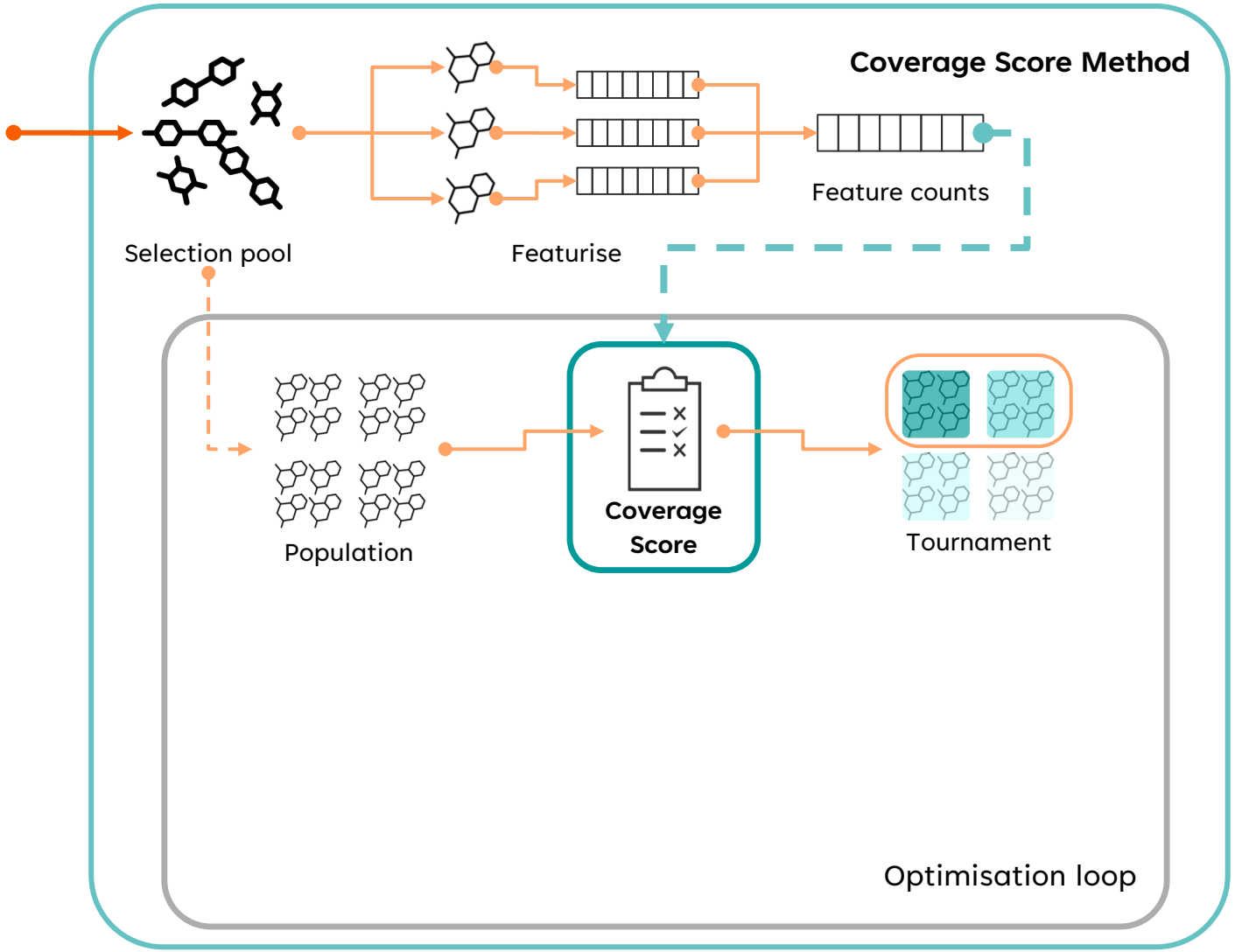
- **Data-dependent** optimisation-based[1] query strategy



1. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE T Evolut. Comput. 2002, 6, 182, DOI: 10.1109/4235.996017

So, what is Coverage Score?

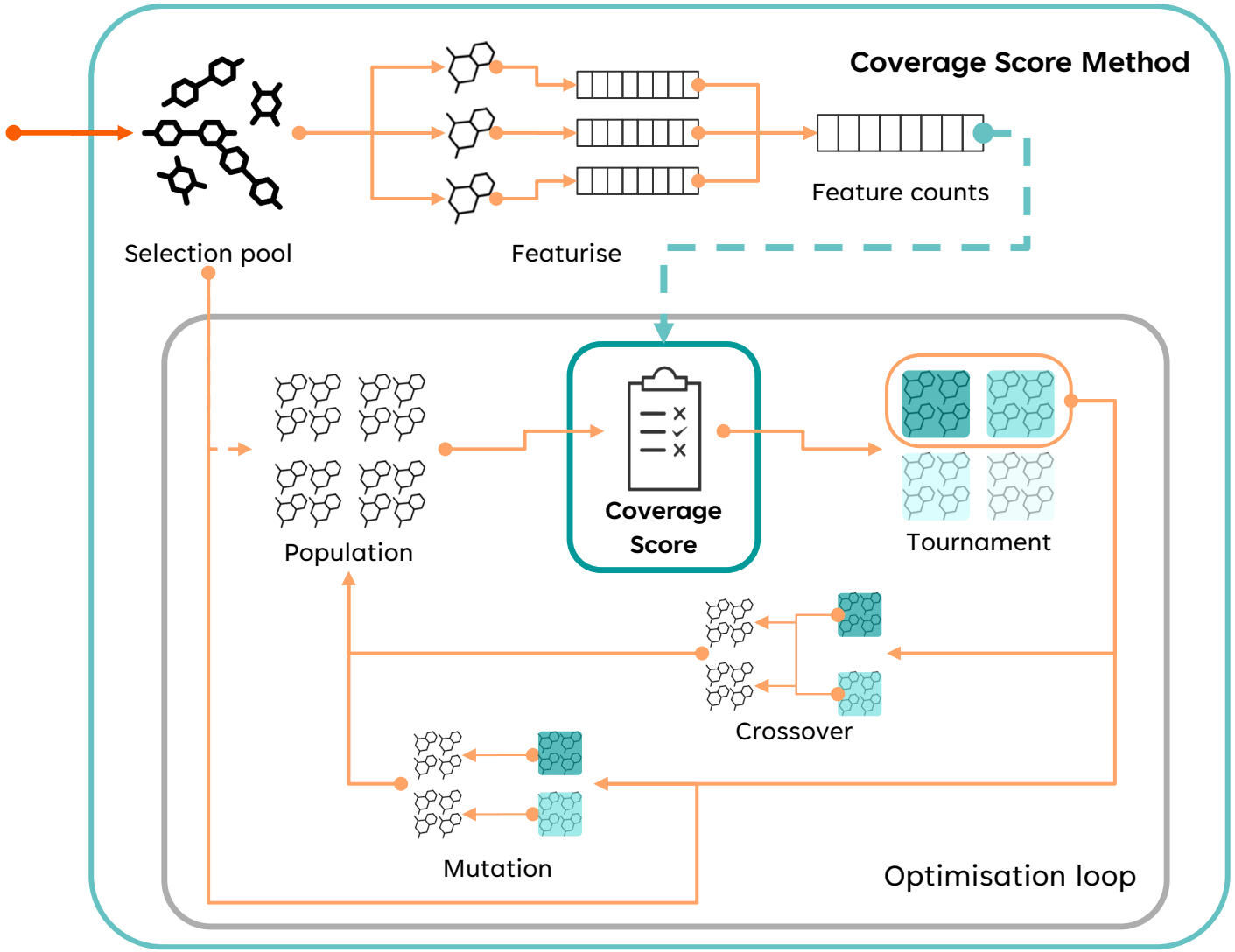
- **Data-dependent** optimisation-based[1] query strategy
- Subset scoring, maximise '**Subset Coverage Score**'



1. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE T Evolut. Comput. 2002, 6, 182, DOI: 10.1109/4235.996017

So, what is Coverage Score?

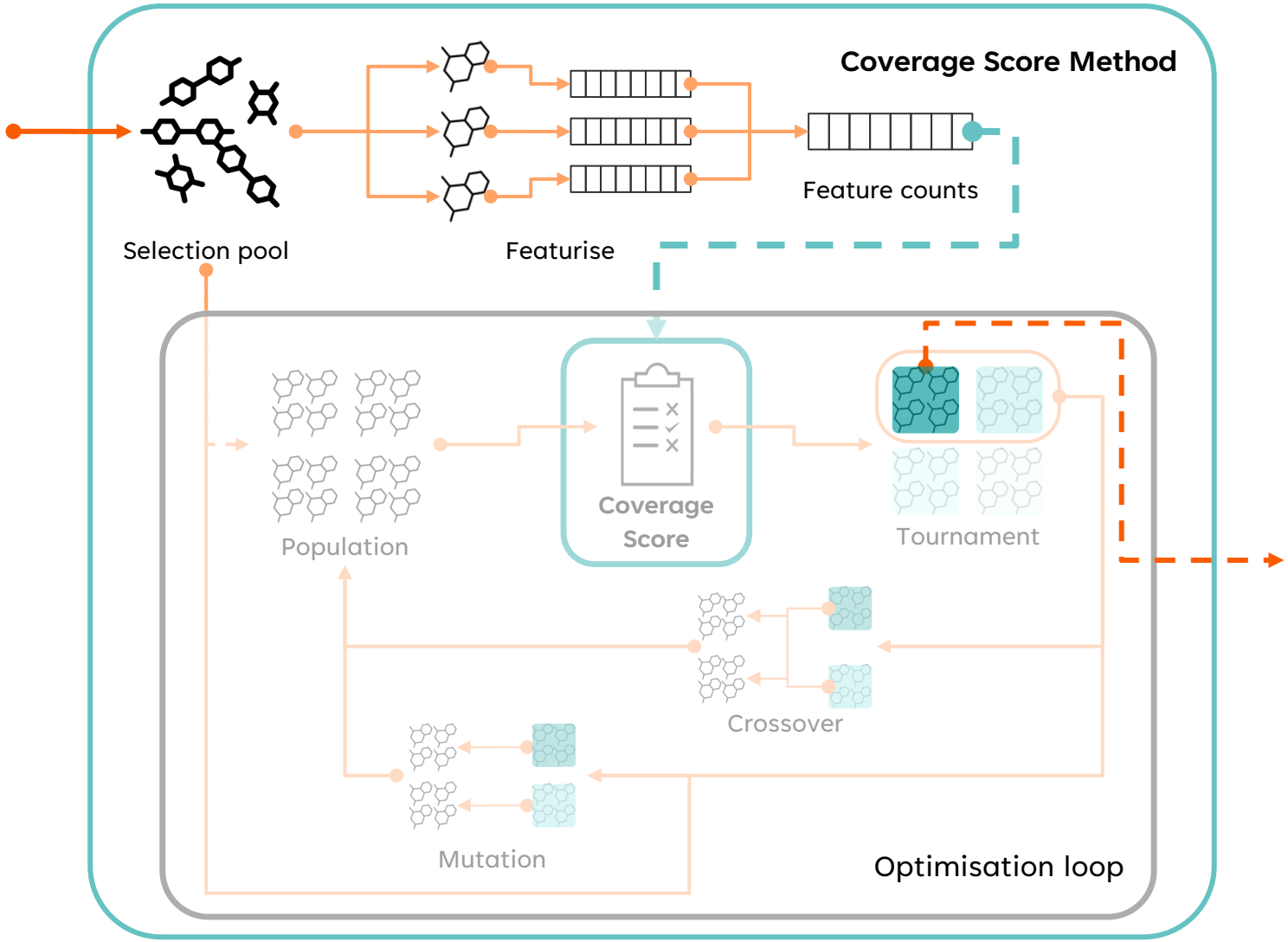
- **Data-dependent** optimisation-based[1] query strategy
- Subset scoring, maximise '**Subset Coverage Score**'



1. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE T Evolut. Comput. 2002, 6, 182, DOI: 10.1109/4235.996017

So, what is Coverage Score?

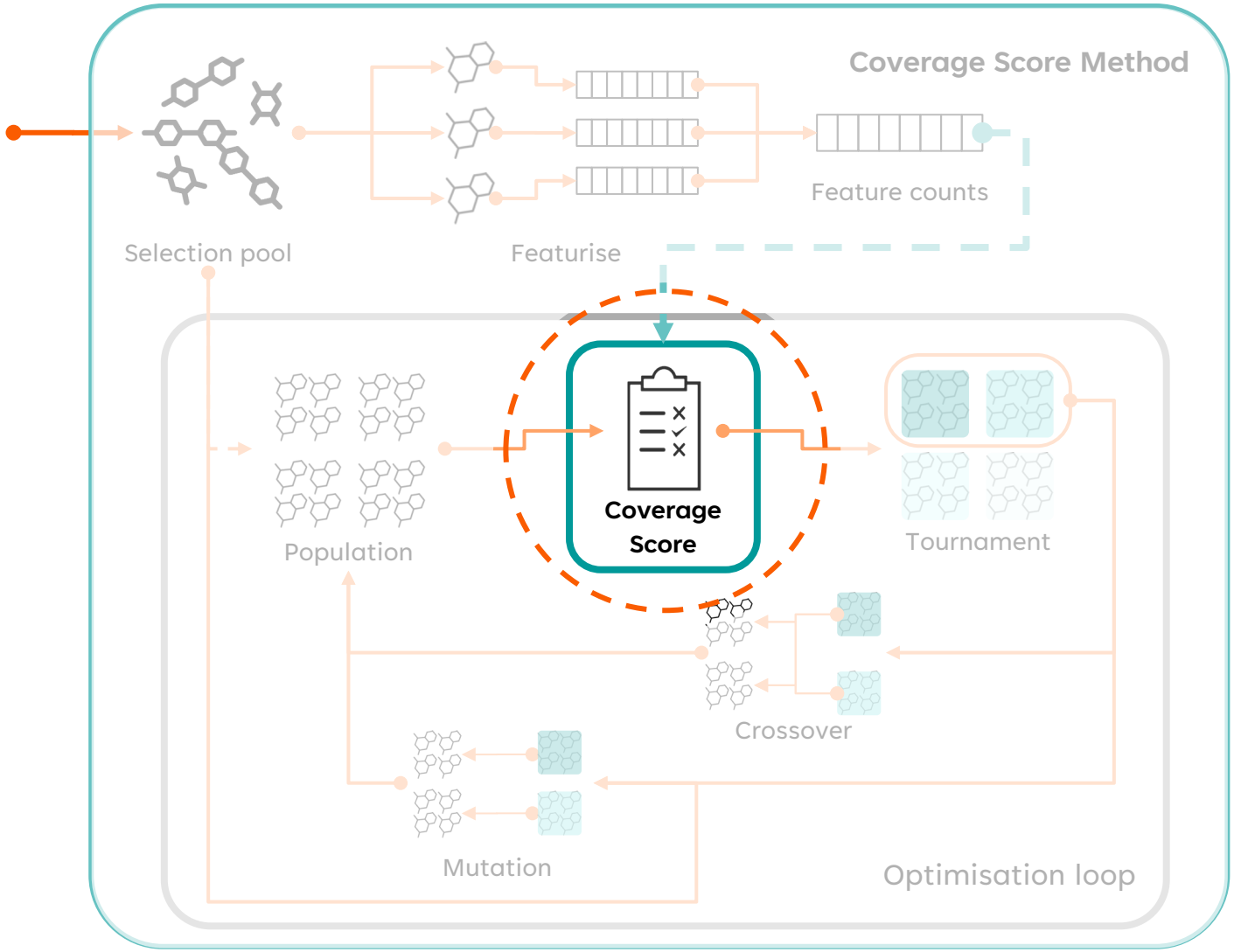
- **Data-dependent** optimisation-based[1] query strategy
- Subset scoring, maximise '**Subset Coverage Score**'
- **Optimisation,** evaluation of each unique subset of 10 out of 100, per ns would take ~200 years!



1. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE T Evolut. Comput. 2002, 6, 182, DOI: 10.1109/4235.996017

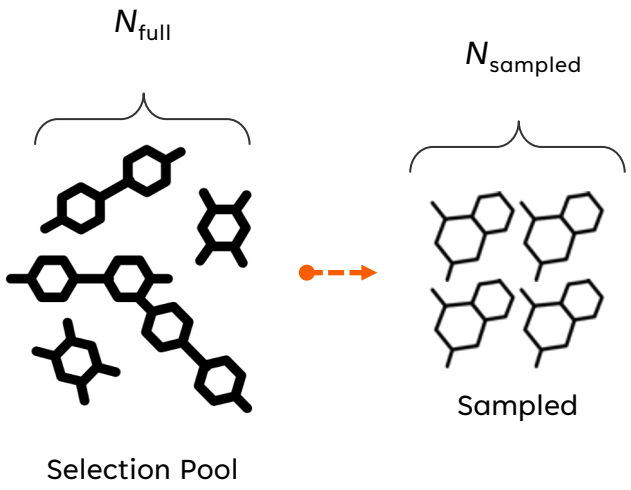
So, what is Coverage Score?

- **Data-dependent** optimisation-based[1] query strategy
- Subset scoring, maximise '**Subset Coverage Score**'
- **Optimisation,** evaluation of each unique subset of 10 out of 100, per ns would take ~200 years!

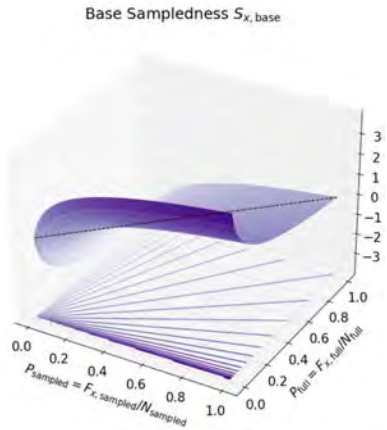


1. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE T Evolut. Comput. 2002, 6, 182, DOI: 10.1109/4235.996017

Calculating Coverage Score

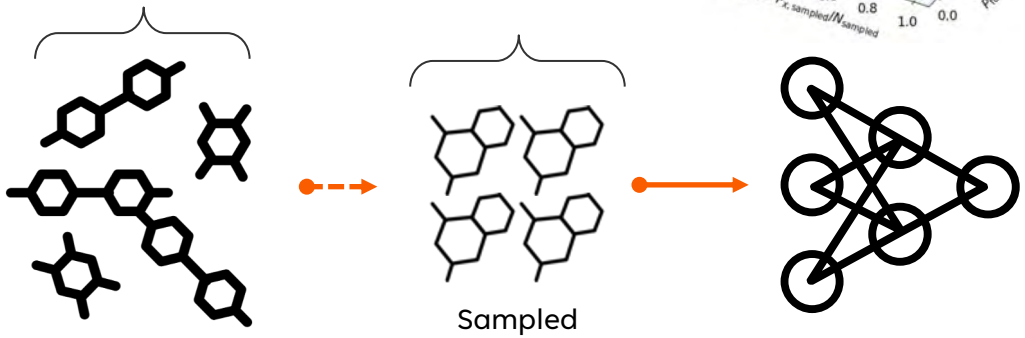


Calculating Coverage Score



N_{full}

$N_{sampled}$



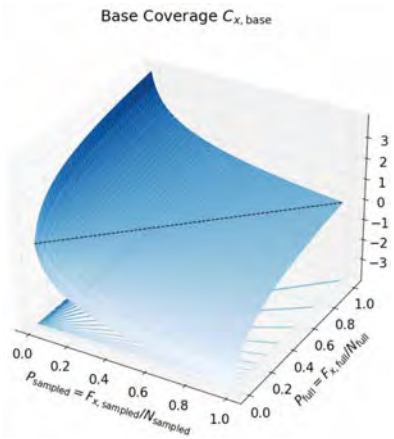
Selection Pool

Sampled

Bayesian Classifier (sampledness)

Features associated with sampledness, **positive** score

Calculating Coverage Score



N_{full}

$N_{sampled}$

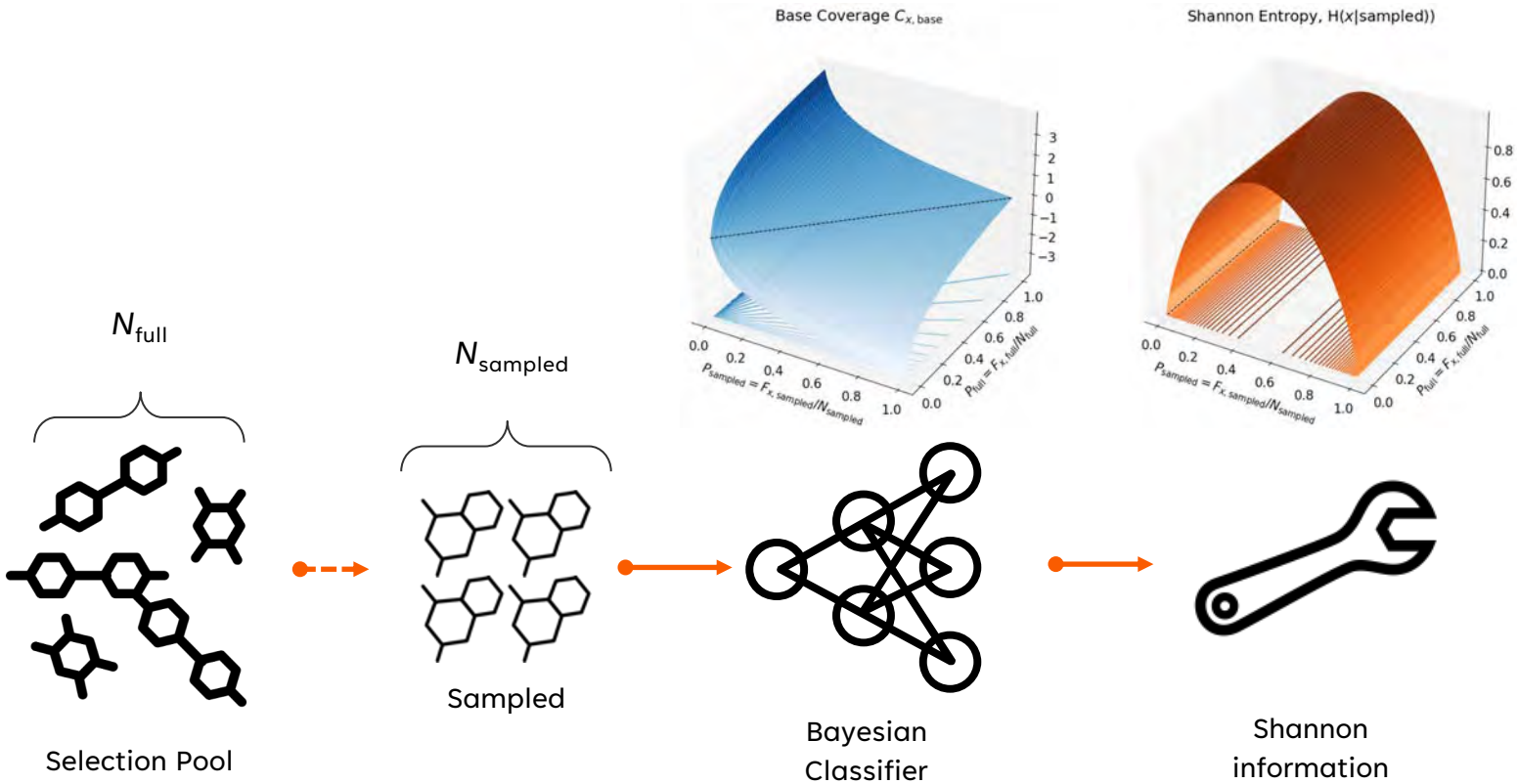
Selection Pool

Sampled

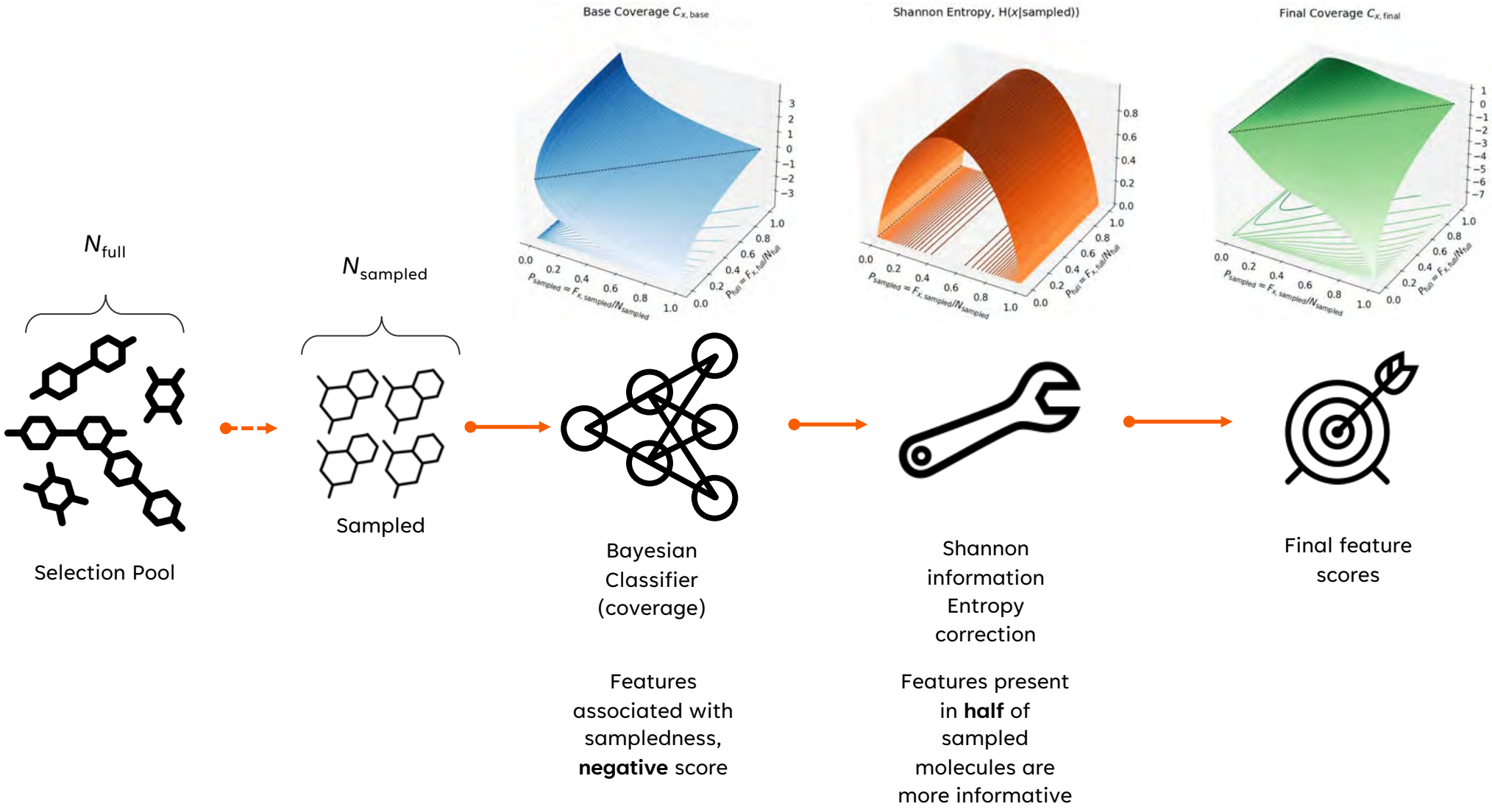
Bayesian Classifier (coverage)

Features associated with sampledness, **negative** score

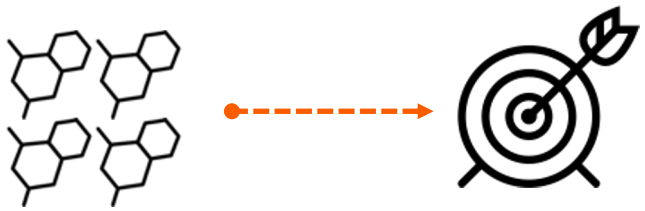
Calculating Coverage Score



Calculating Coverage Score



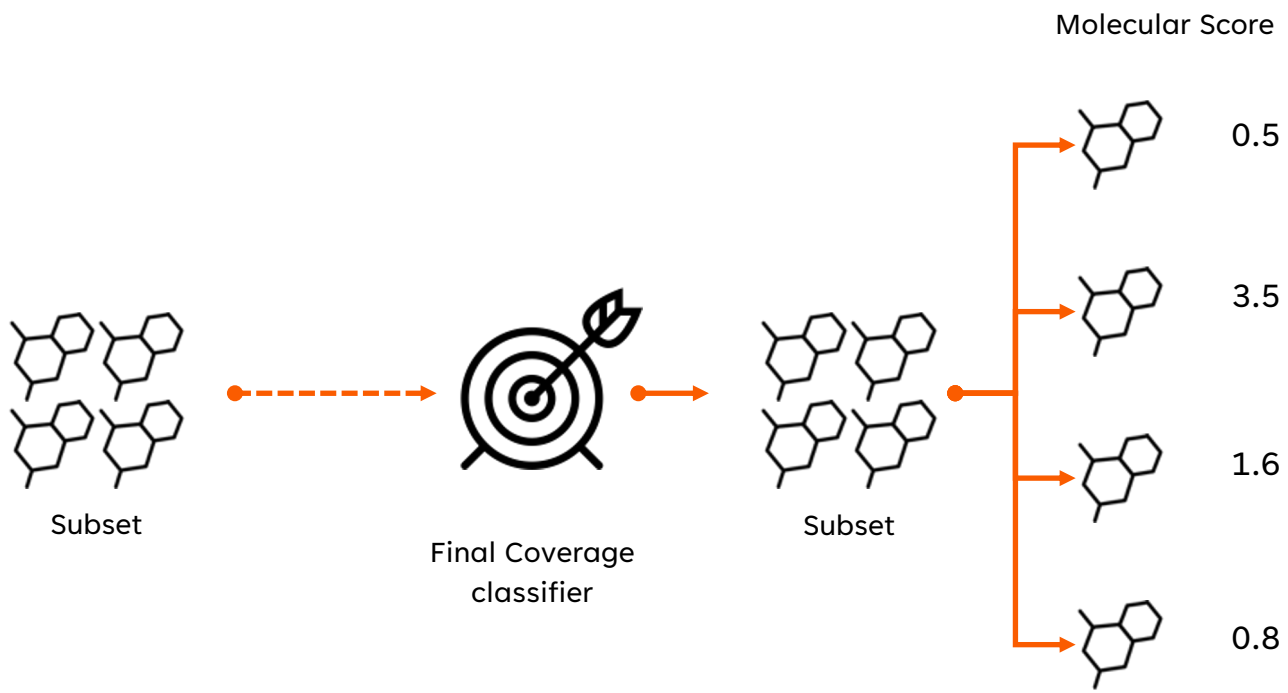
Calculating Coverage Score



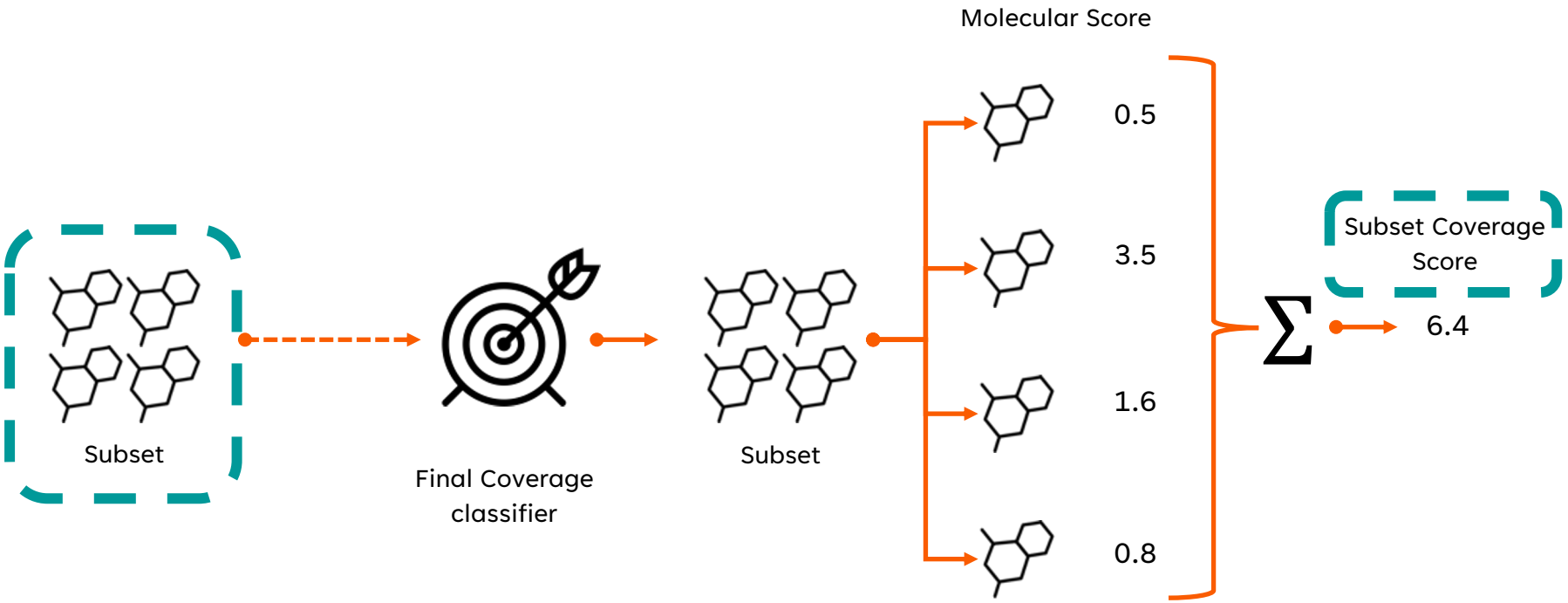
Subset

Final Coverage classifier

Calculating Coverage Score

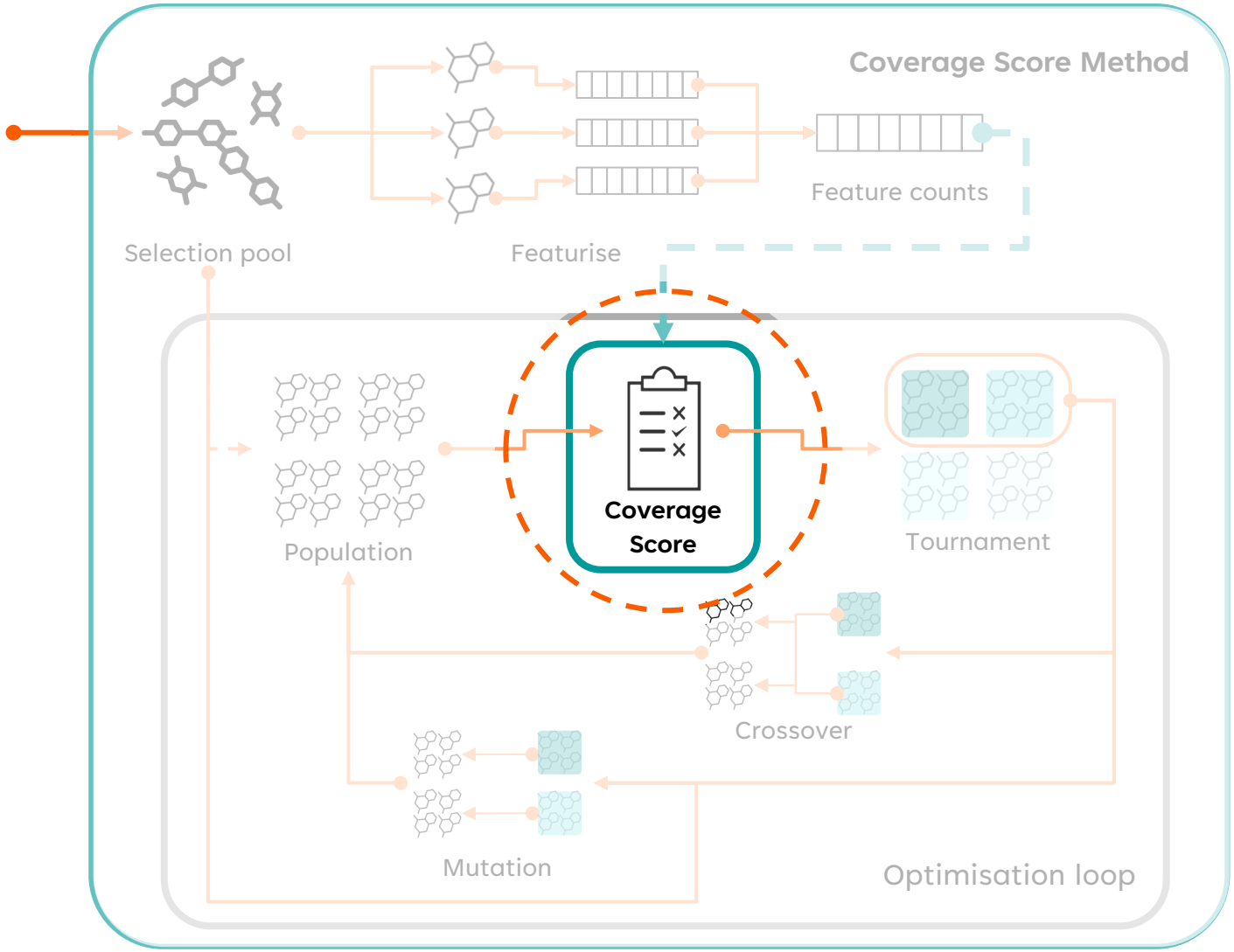


Calculating Coverage Score



So, what is Coverage Score?

- **Data-dependent** optimisation-based[1] query strategy
- Subset scoring, maximise '**Subset Coverage Score**'
- **Optimisation,** evaluation of each unique subset of 10 out of 100, per ns would take ~200 years!



1. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE T Evolut. Comput. 2002, 6, 182, DOI: 10.1109/4235.996017

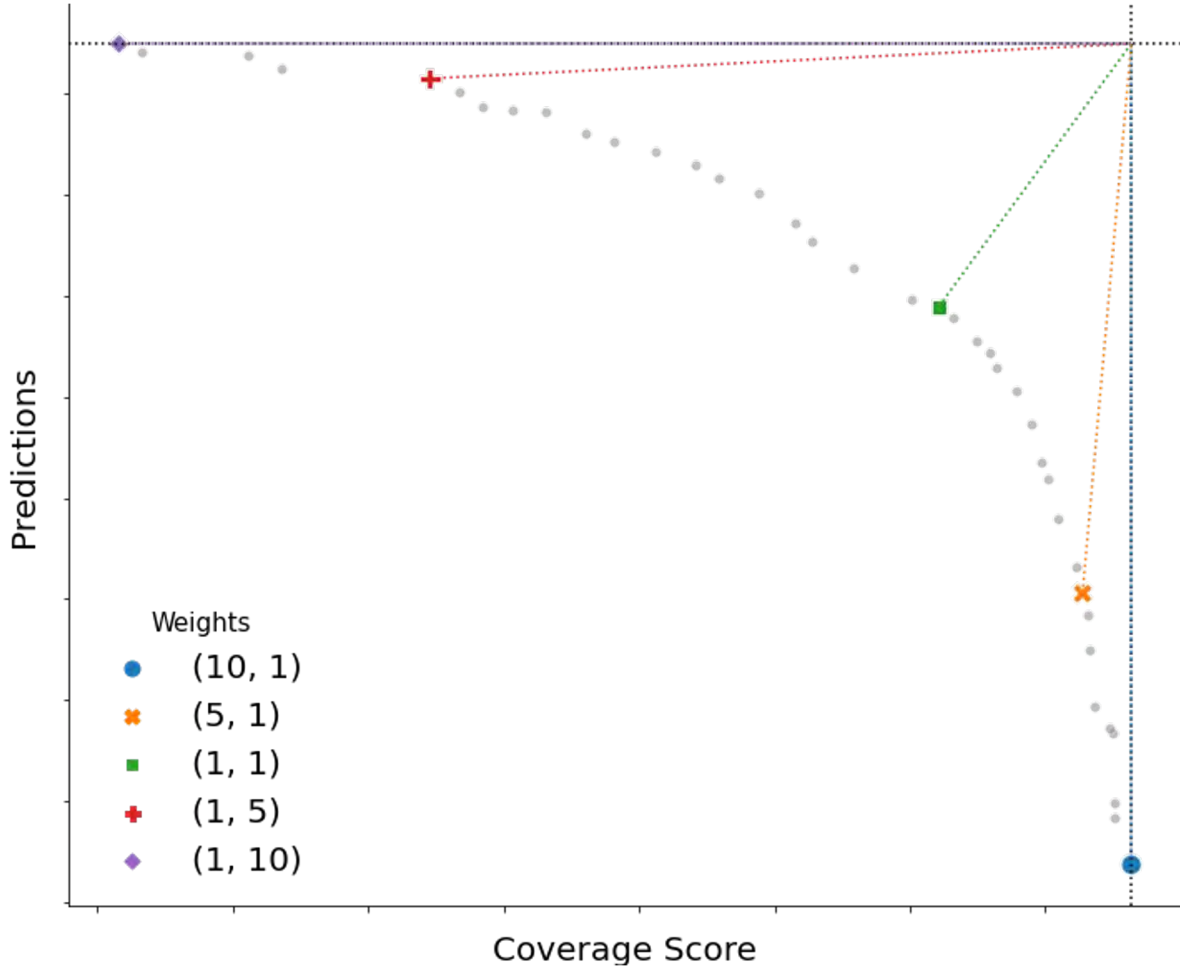
Optimisation of additional properties

- **Genetic algorithm** can optimise for multiple properties
- Balancing **exploration** (subset coverage score) and **exploitation** (molecule scores/properties)
- **Additional subset scores** defined by:

$$p_S = \sum_{\text{mol} \in S} p_{\text{mol}}$$

- Final subset selected through normalised weighted selection

$$S^* = \arg\max_S \sum_p w_p \hat{p}_S \quad , w_p \in \mathbb{R}, \hat{p}_S \in [0,1]$$



Outline

Active learning in drug discovery

- Why is it useful?

Query strategies

- How to select molecules?

Coverage Score

- How does it work?

Validation

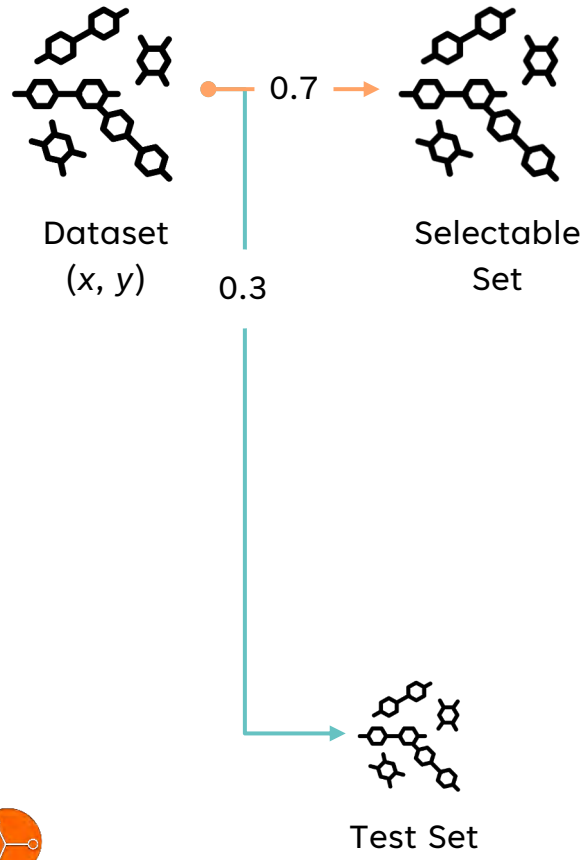
- How does Coverage Score perform?

Further work/summary

- Where do we go from here?

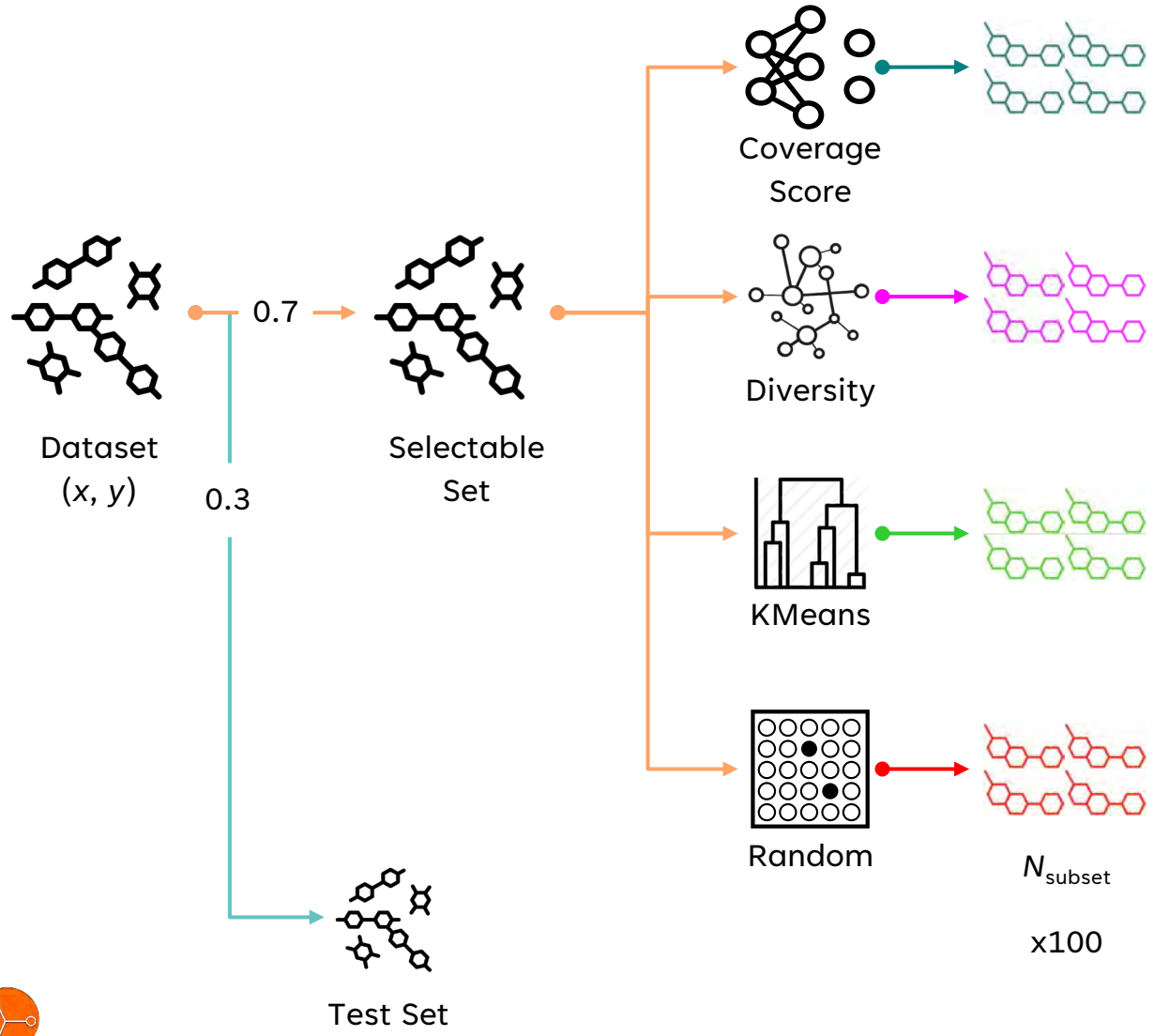
Validating selection methods

Model performance



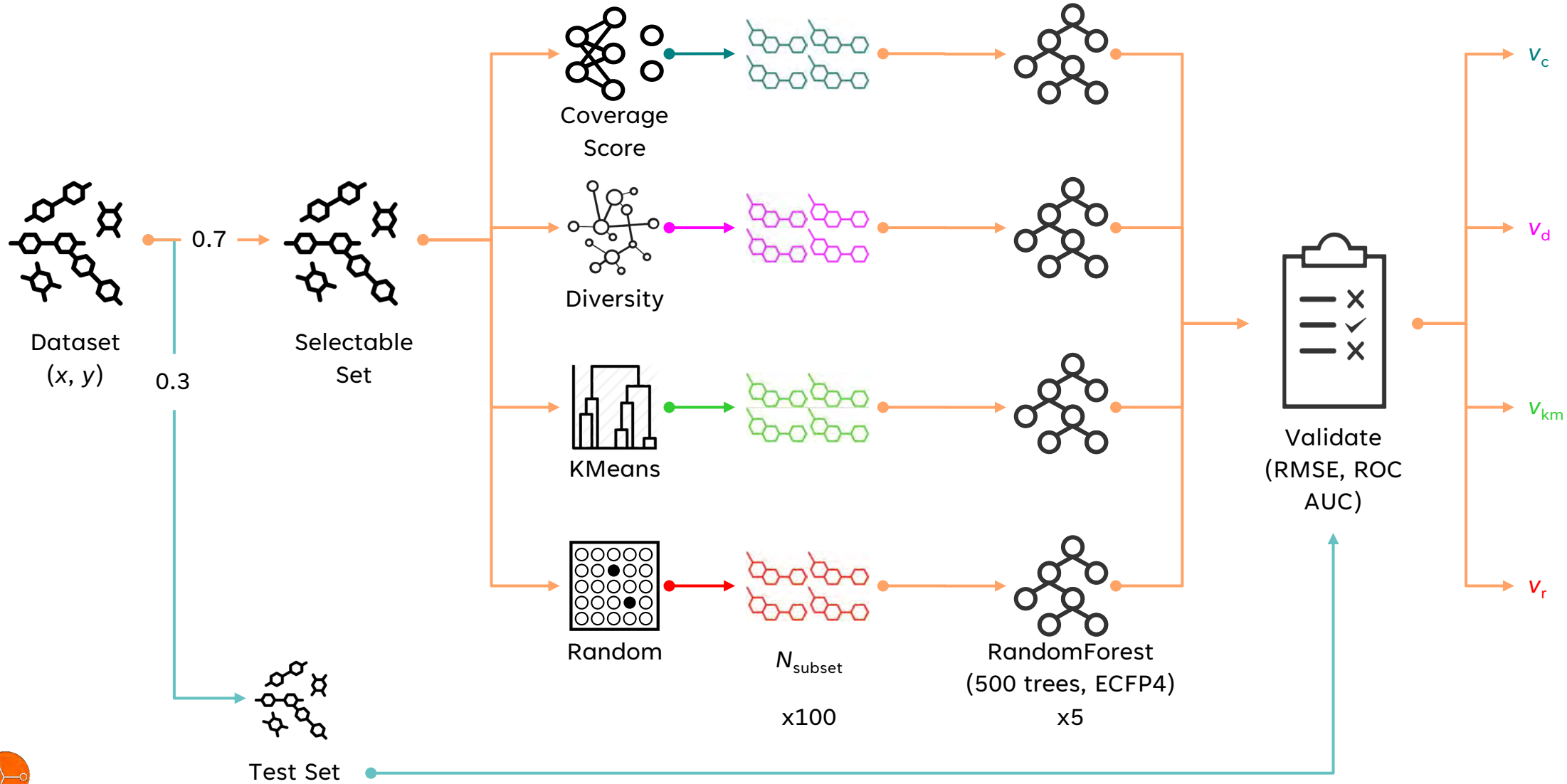
Validating selection methods

Model performance



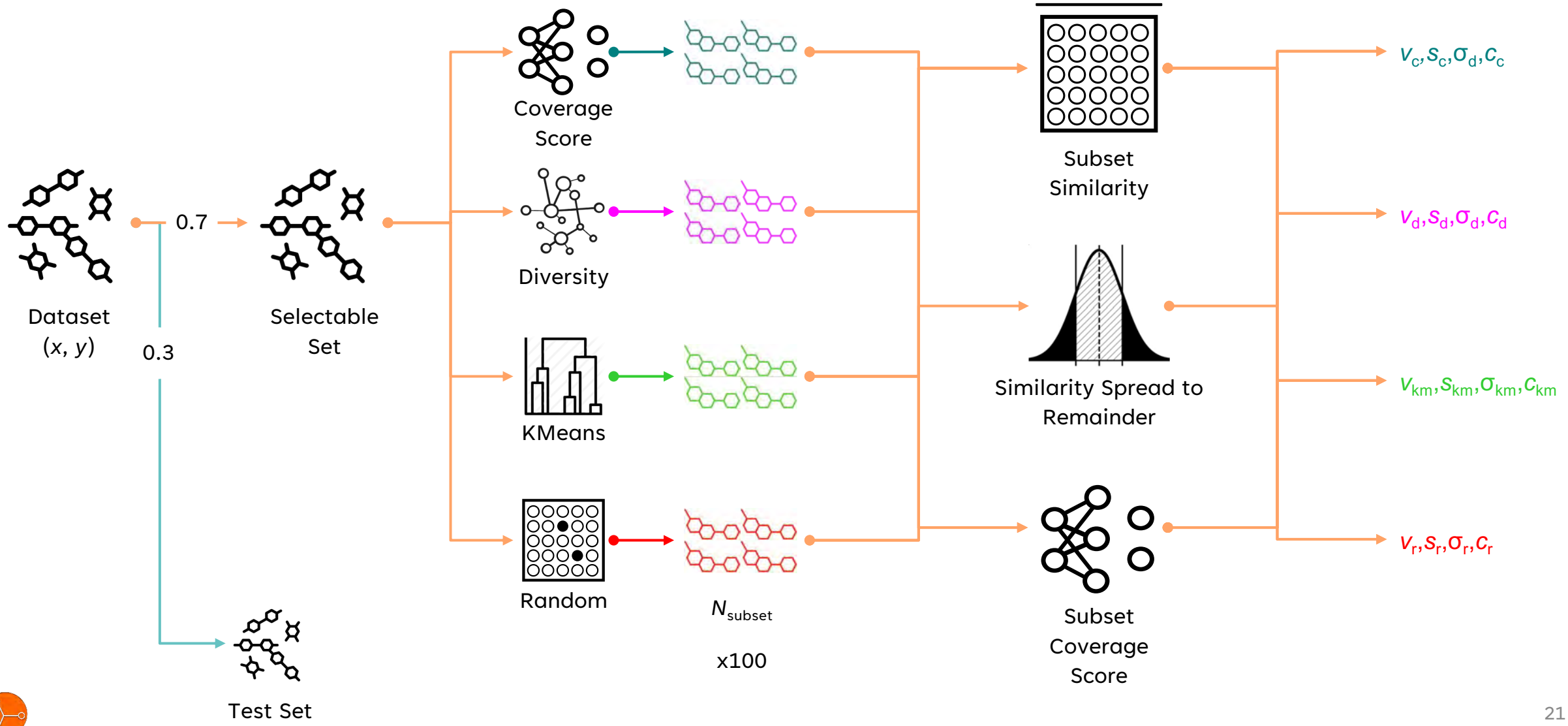
Validating selection methods

Model performance



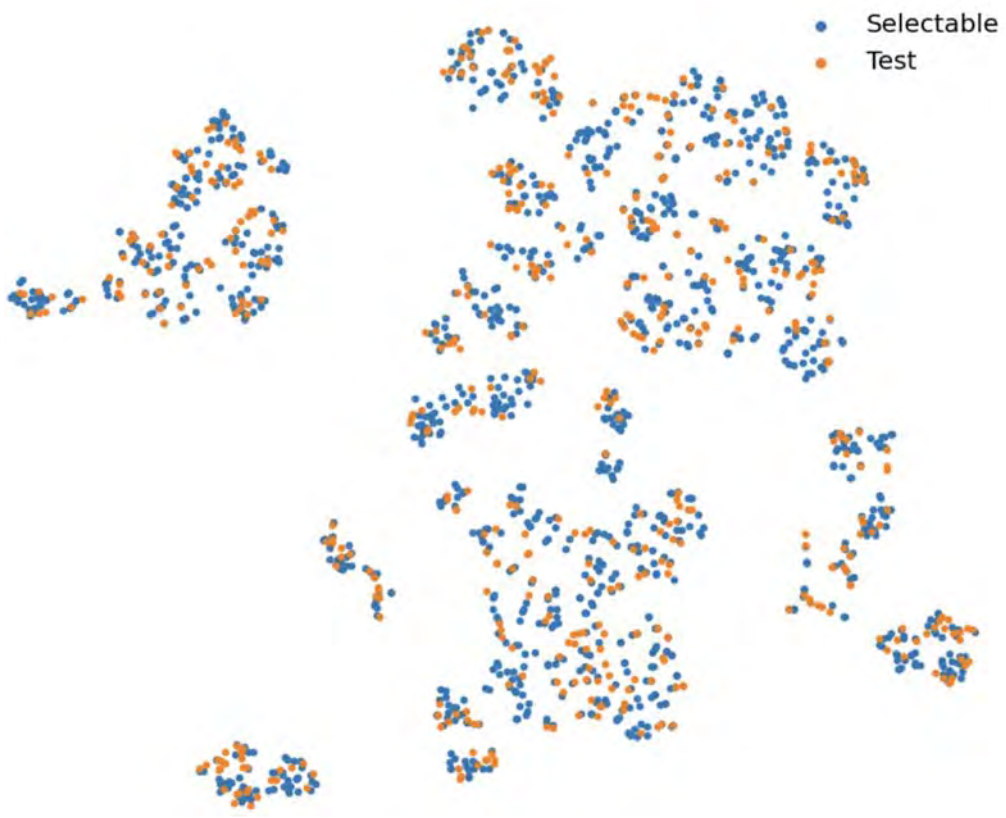
Validating selection methods

Additional metrics



Datasets

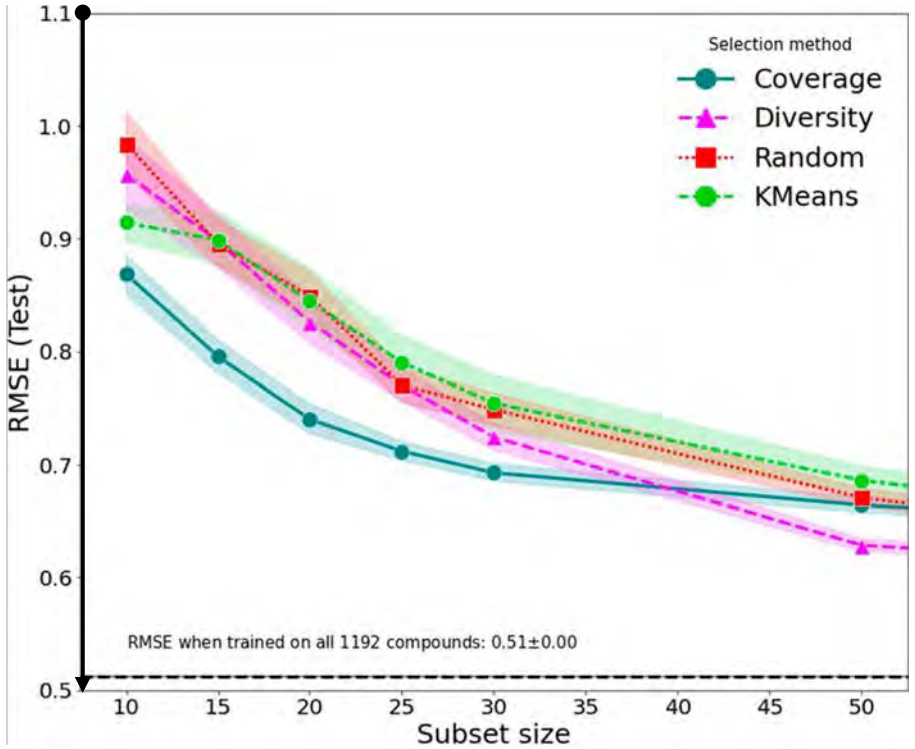
- **Five** different datasets tested
- Regression (RMSE) and classification (ROC AUC) tasks
- **D2**
  - *x* = GSK set of molecules (1704)
  - *y* = experimentally determined pIC_{50} values for MMP12

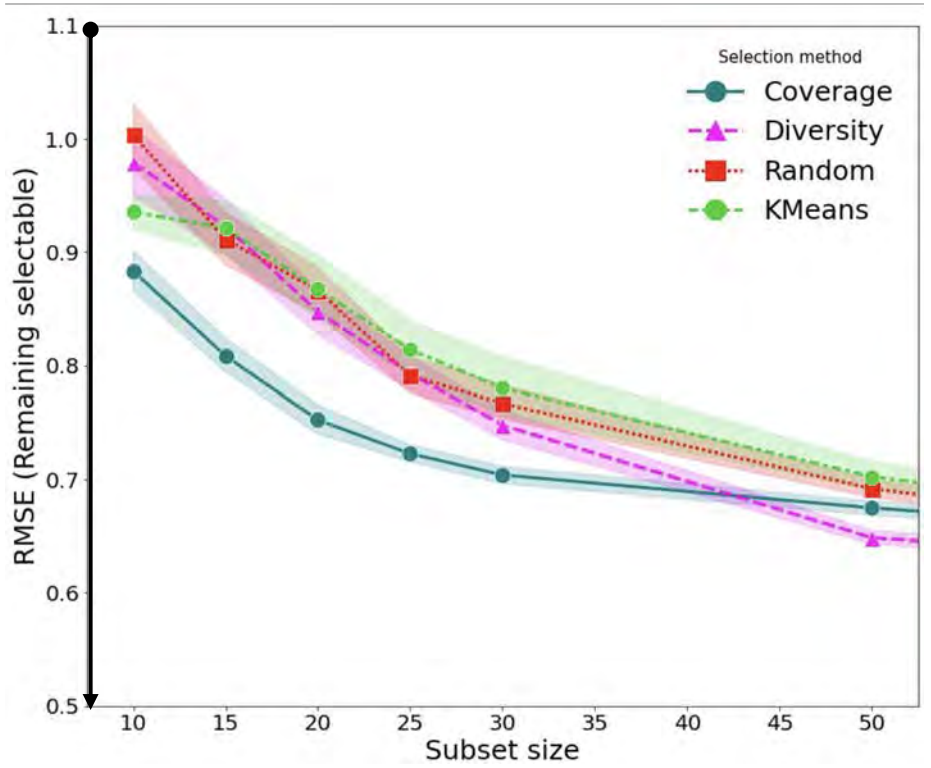


t-SNE plot of D2 split by selectable (0.7) and test (0.3) sets

D2 selections

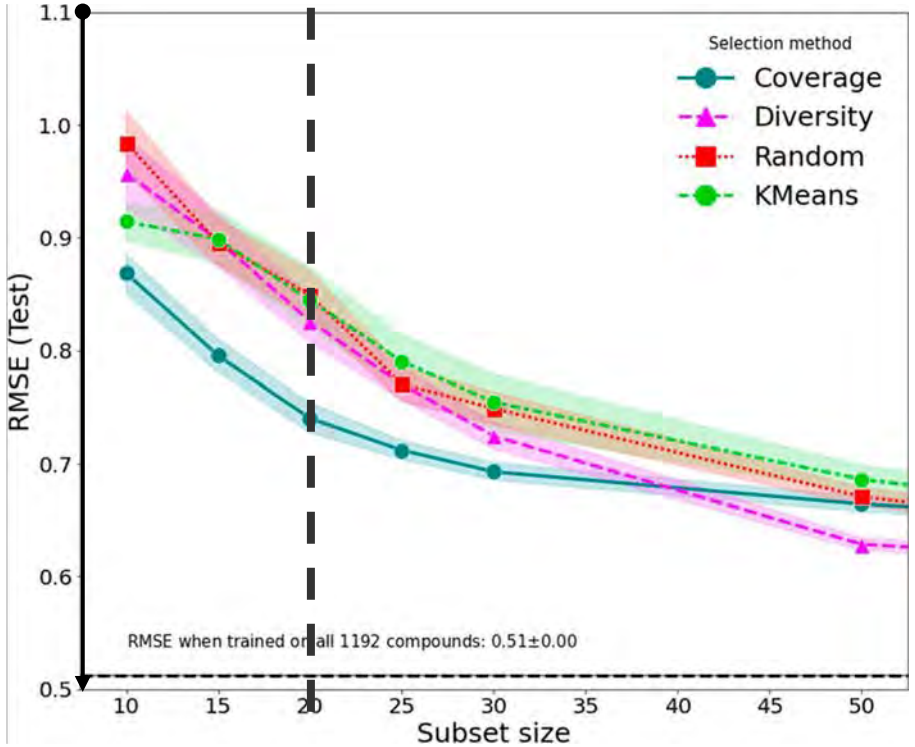
GSK set pIC_{50} for MMP12

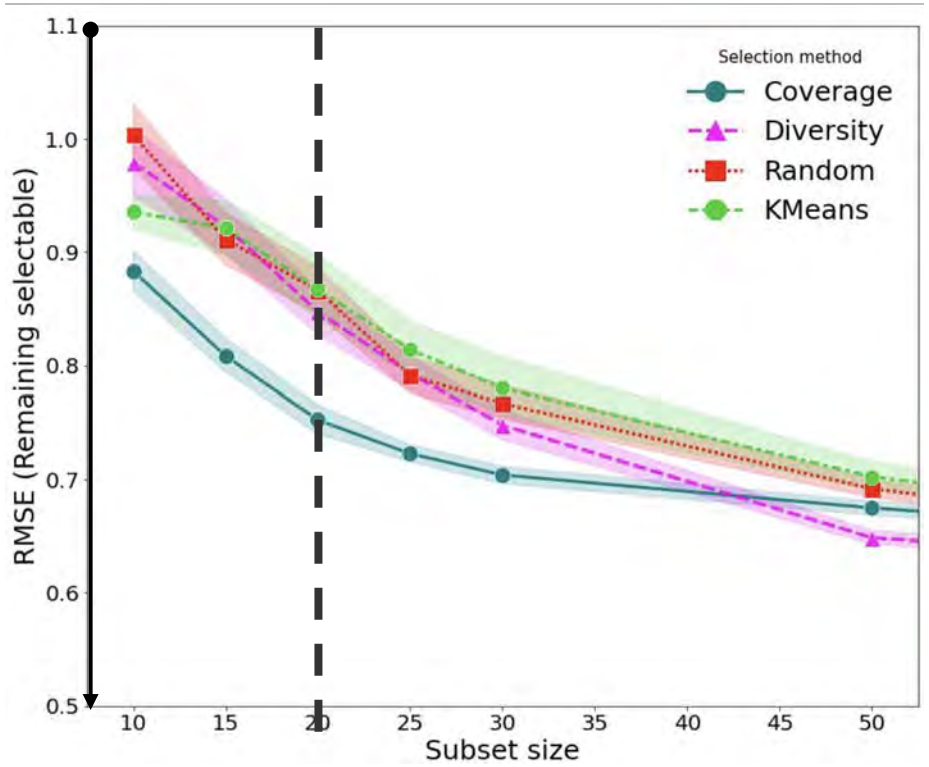




D2 selections

GSK set pIC_{50} for MMP12

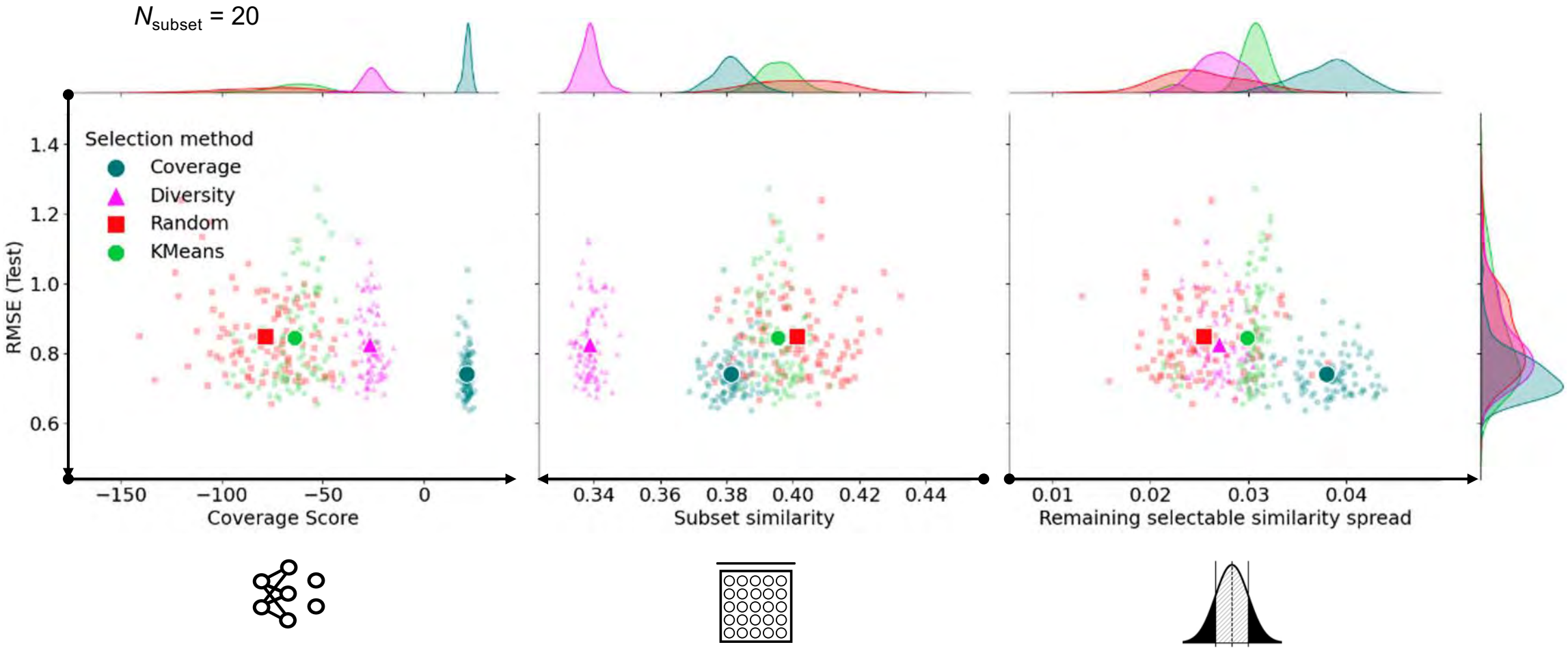




D2 selections

GSK set pIC_{50} for MMP12

Each dot = 1 selected subset
Large markers = average over all 100 selections



Datasets

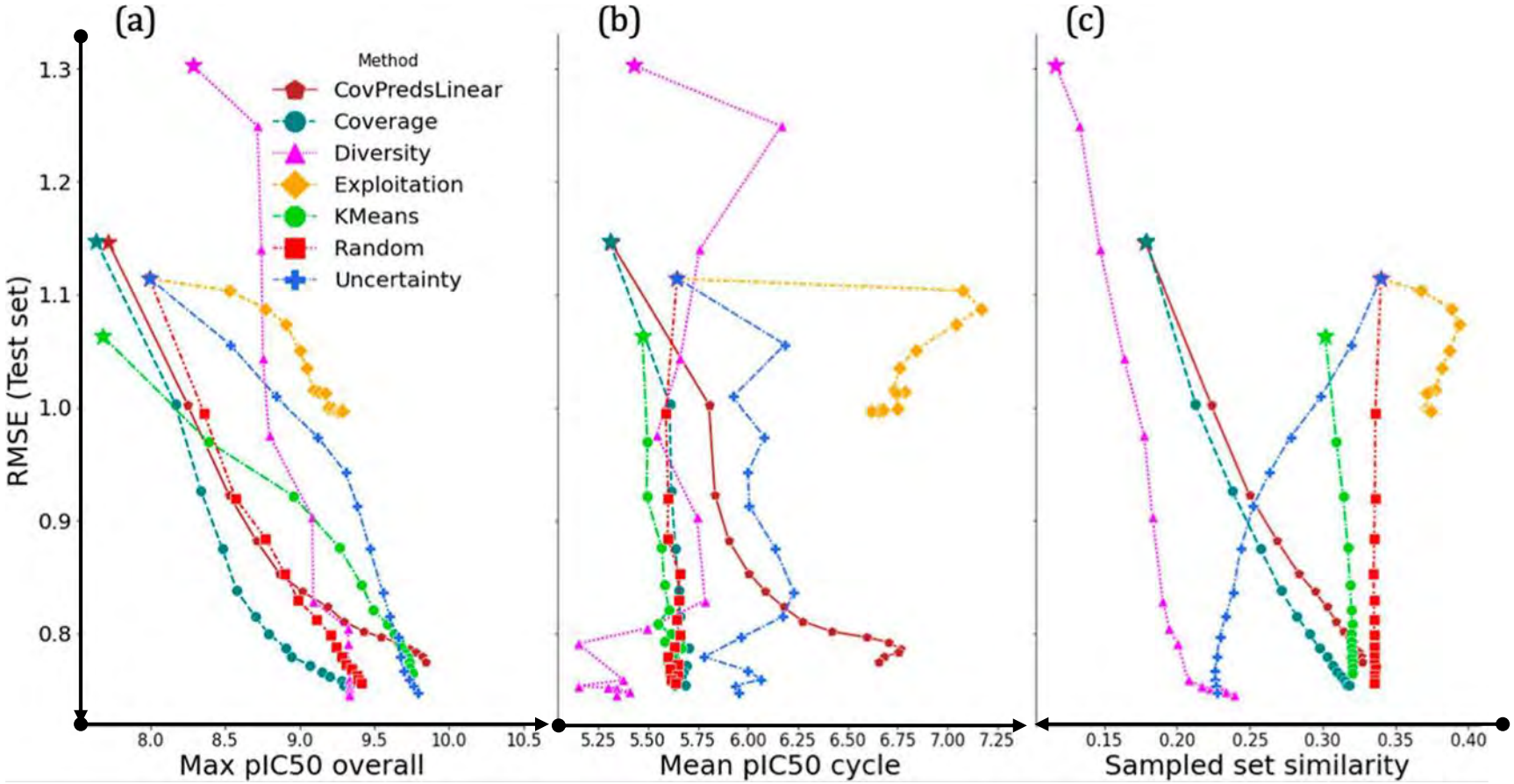
- **Five** different datasets tested
- Regression (RMSE) and classification (ROC AUC) tasks
- D2
  - *x* = GSK set of molecules (1704)
  - *y* = experimentally determined pIC_{50} values for MMP12
- **D2+**
  - *x* = D2 + molecules from ChEMBL (2076)
  - *y* = experimentally determined pIC_{50} values for MMP12
  - Simulated subsequent 15 cycles of selection N_{subset} = 20

t-SNE plot of D2+ split by D2 (orange) and ChEMBL compounds (pale blue)

D2+ selections

MMP12 pIC_{50} compounds

- 15 cycles of selection (N_{subset} = 20)
- ★ markers = initial cycle, subsequent cycles connected
- Additional query strategies included:
  - **Exploitation** → highest predictive score
  - **Uncertainty** → highest uncertainty in score
  - **CovPredsLinear** → Coverage Score with predictions, linear increments in weights (CS, P), each cycle (50 → 1, 1 → 50), initial solely Subset Coverage Score based



Outline

Active learning in drug discovery
- Why is it useful?

Query strategies
- How to select molecules?

Coverage Score
- How does it work?

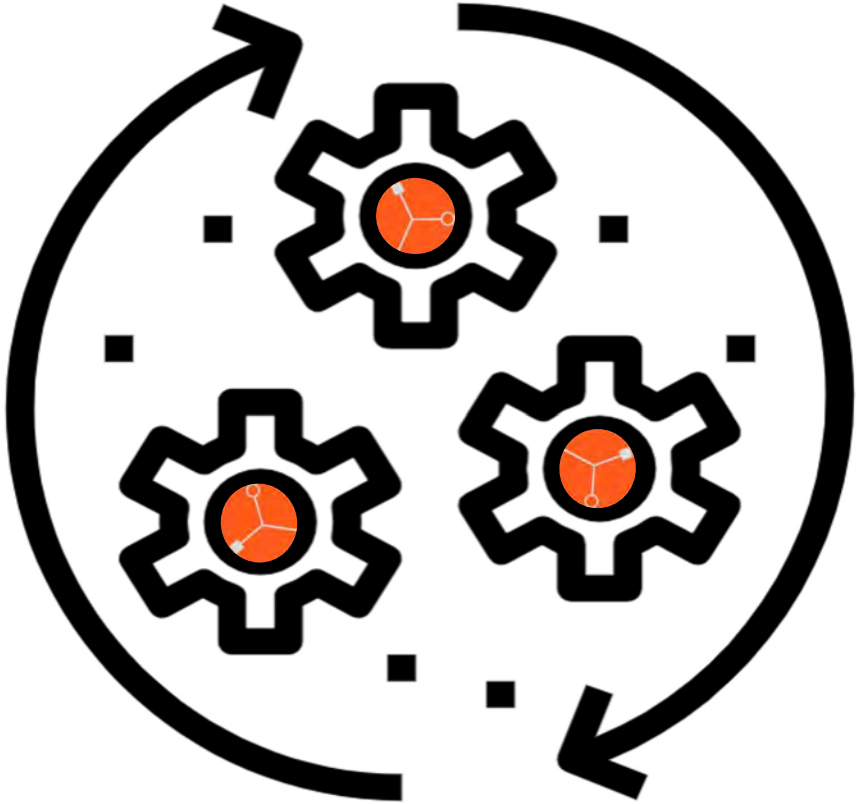
Validation
- How does Coverage Score perform?

Further work/summary
- Where do we go from here?

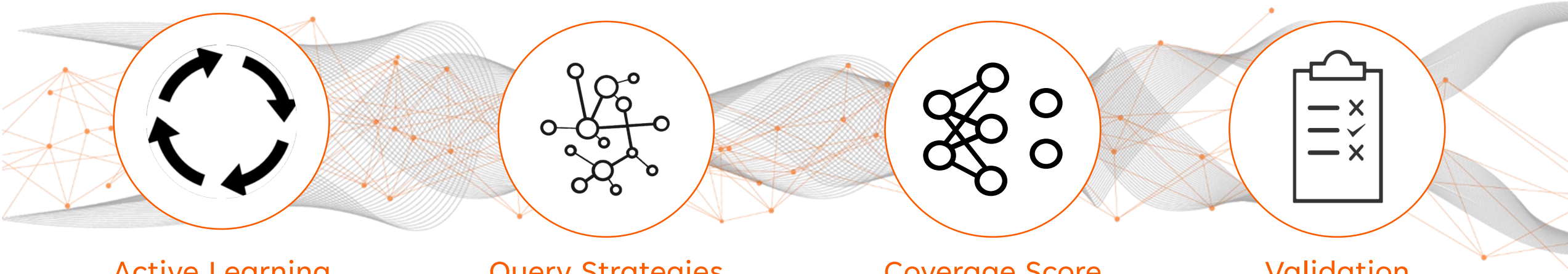
Future work

Where to go from here?

- **Automatic** exploration / exploitation balancing
- Investigation into **optimal** feature coverage surface
- **Representation** analysis, including **3D descriptors** (PLIFs)
- Model **confidence** and **domain of applicability** as a validation metric



Summary



Active Learning

- Vital to learn effectively by selecting informative molecules
- Useful in low-data regime

Query Strategies

- Query strategies can be model- or data-dependent
- Pros and cons to multiple approaches

Coverage Score

- Genetic optimisation-based method
- Finds subset that maximises a ‘subset coverage score’
- Can optimise for additional properties

Validation

- Subsets contain dissimilar compounds
- Subsets can better training sets
- Balance of exploitation and exploration

Acknowledgements

Willem van Hoorn

Cedric Bouysset

Alice Cappechi

Anthony Bradley

Rob Smith

Ilenia Giangreco

Exscientia

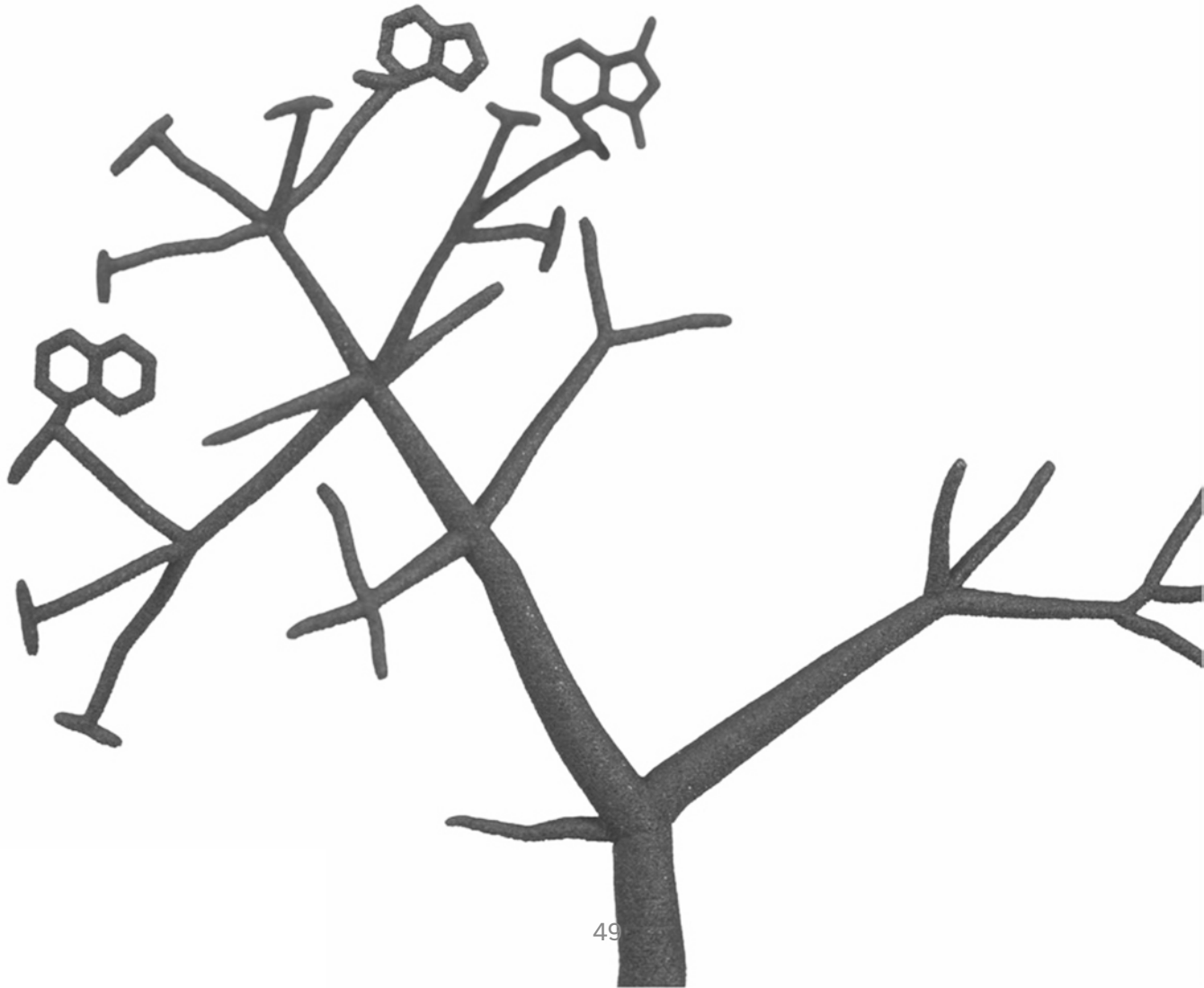
Exscientia plc

OXFORD HEADQUARTERS
The Schrödinger Building
Oxford Science Park
Oxford OX4 4GE

dwoodward@exscientia.ai

Registered address: The Schrödinger Building,
Oxford Science Park, Oxford, OX4 4GE, United Kingdom

Registered number: 13483814



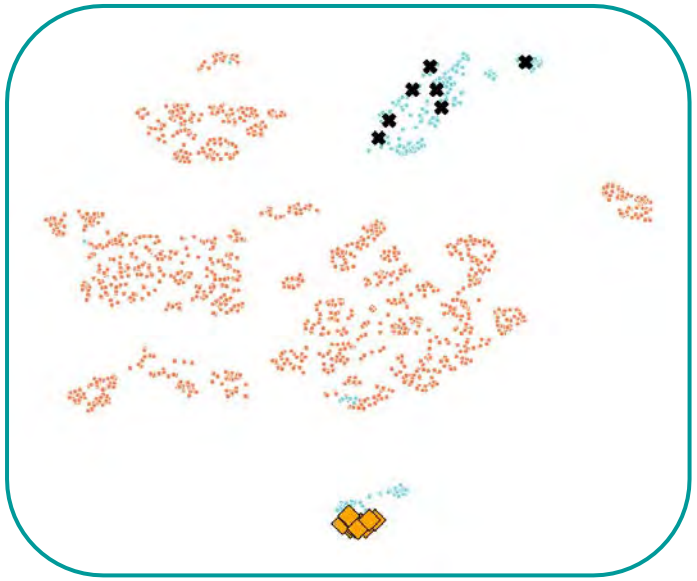
Query strategy comparisons

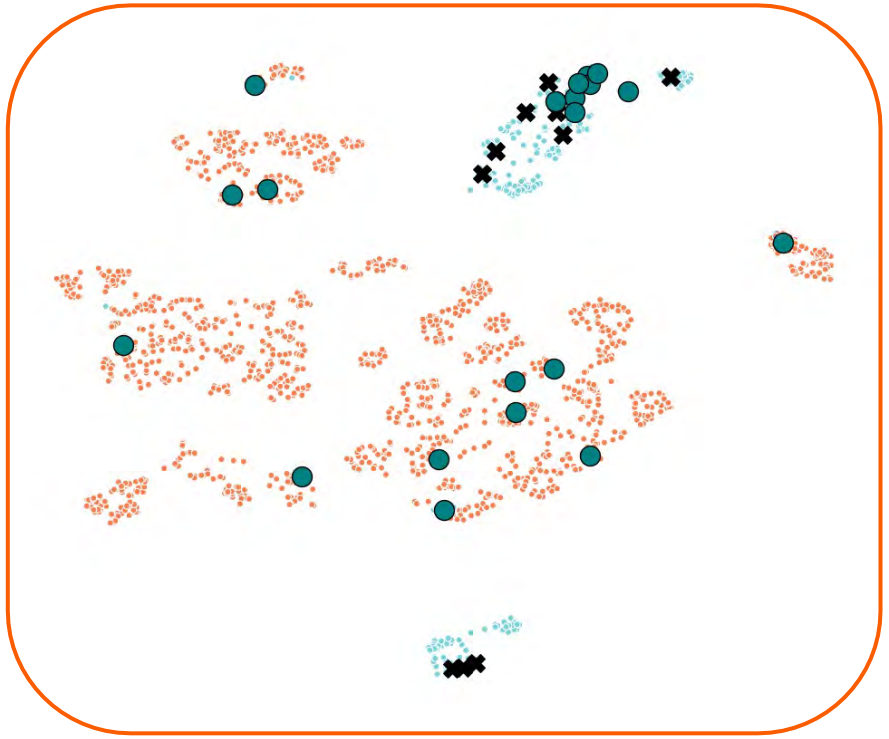
Model-dependent: Dissimilar prior selections

Very similar

Diverse

Prior selections





Exploitation
highest model score

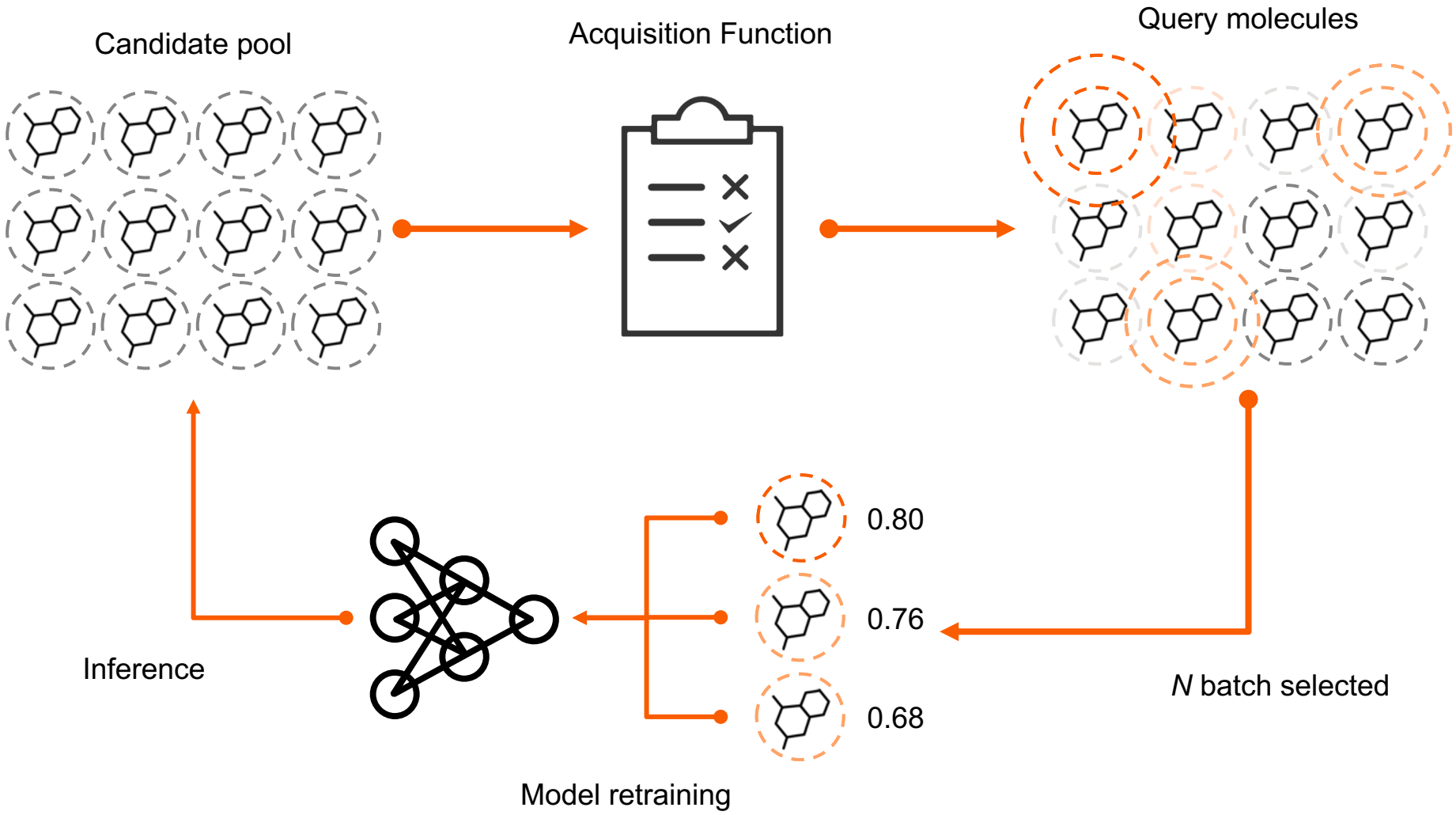
Only thinks compounds are high scoring

Uncertainty-batch
highest uncertainty in score, batches of 5, **retrained** on predictions

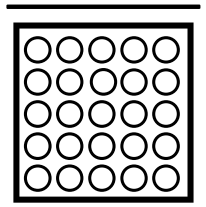
Coverage Score
optimisation-based

Samples from and based regions with more focus on

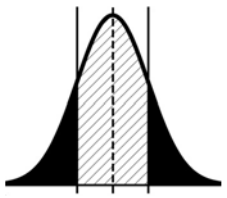
Model-dependent Loop



Validation

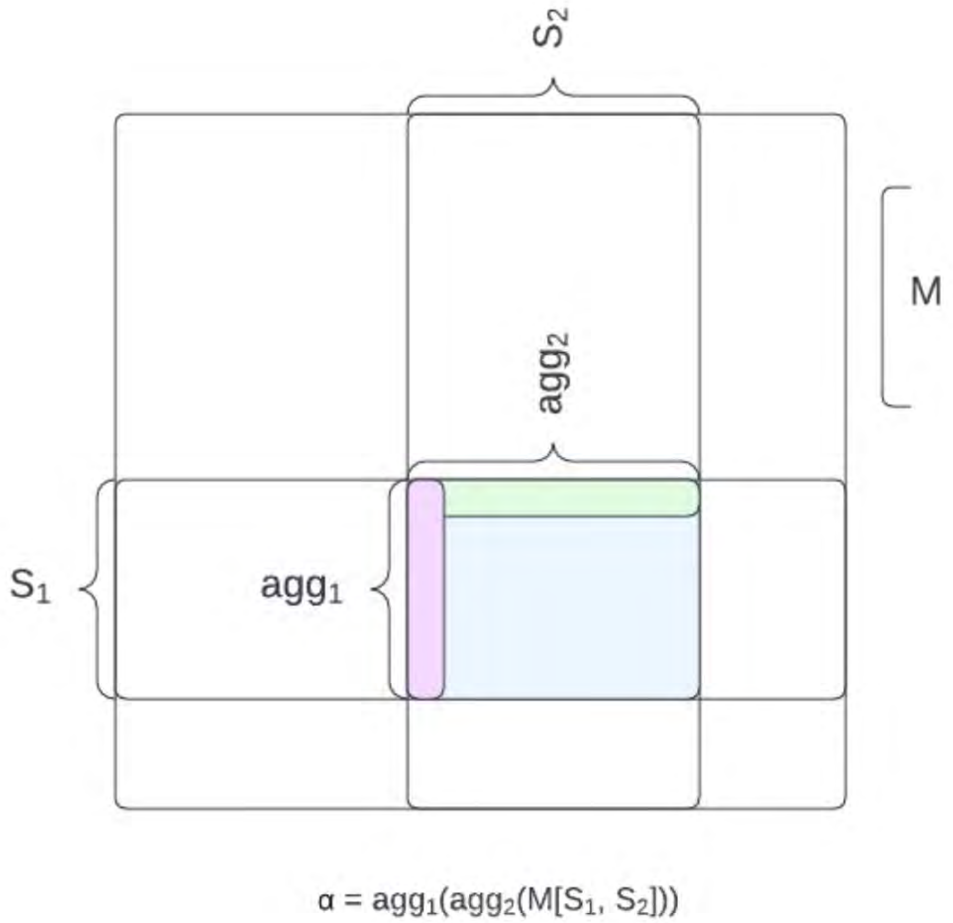


- Subset similarity:
  - $S_1 = S_2 = S$
  - agg_1 = mean
  - agg_2 = mean



- Remaining selectable similarity spread:
  - $S_1 = S_{subset}$
  - $S_2 = S_{full} - S_{subset}$
  - agg_1 = std dev
  - agg_2 = mean

$M = N_{full}$ x N_{full} similarity matrix

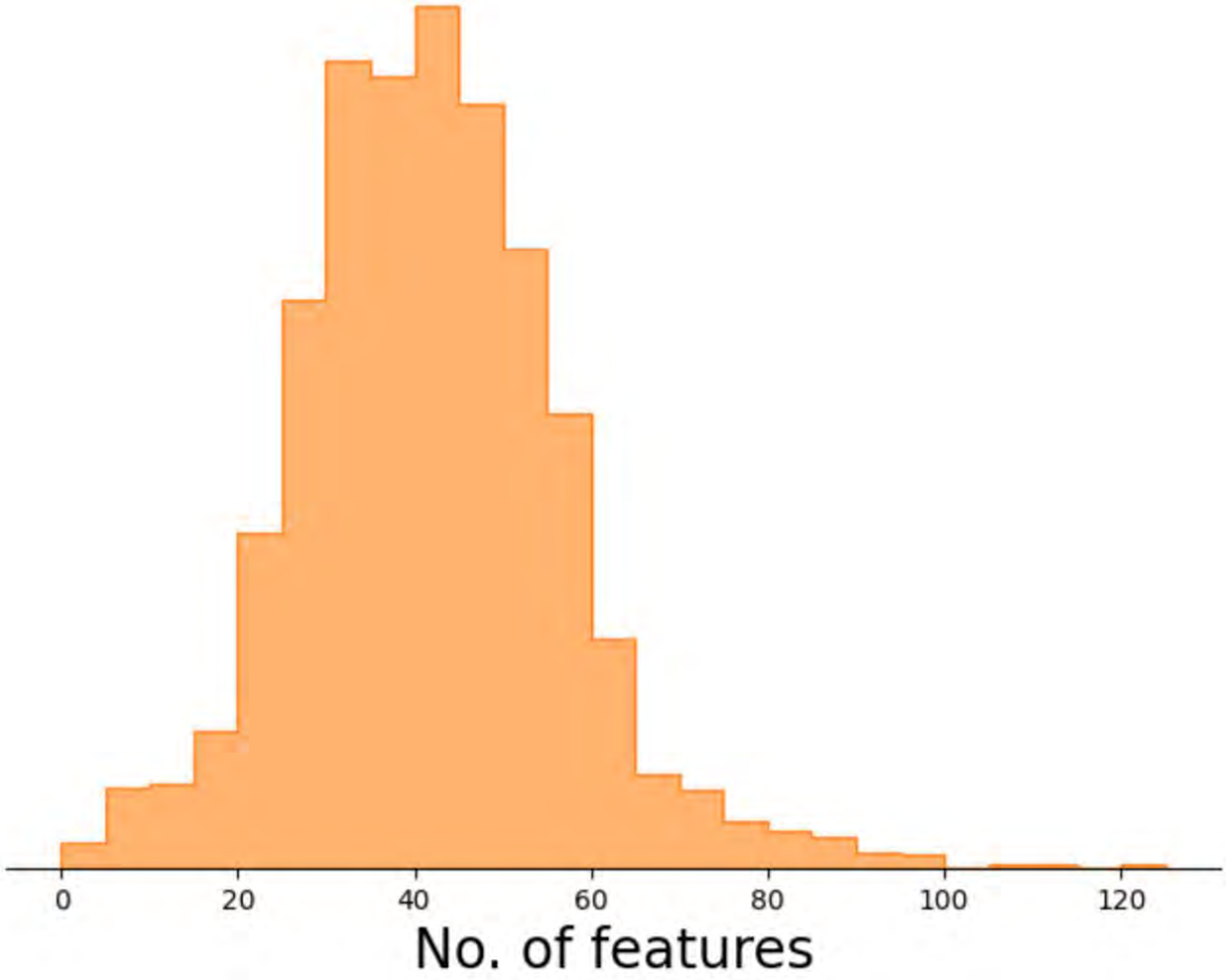


$\alpha = agg_1(agg_2(M[S_1, S_2]))$

D3 & D3F

Blood Brain Barrier Penetrance

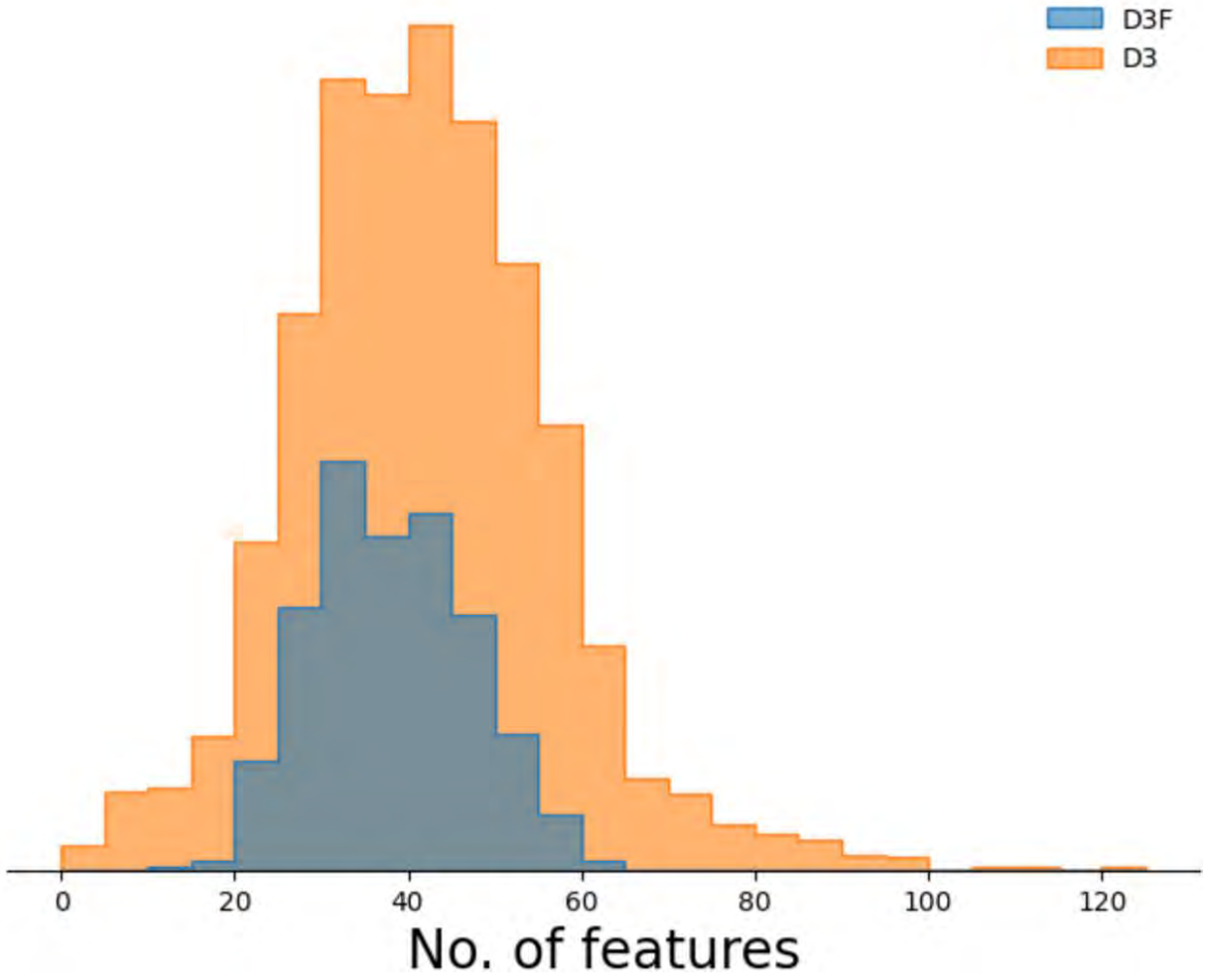
- **D3**
  - *x* = Desalted, deduplicated molecules from MoleculeNet Blood Brain Barrier Penetrance dataset.
  - *y* = {0, 1} classification of brain penetrant (1) or not (0).



D3 & D3F

Blood Brain Barrier Penetrance

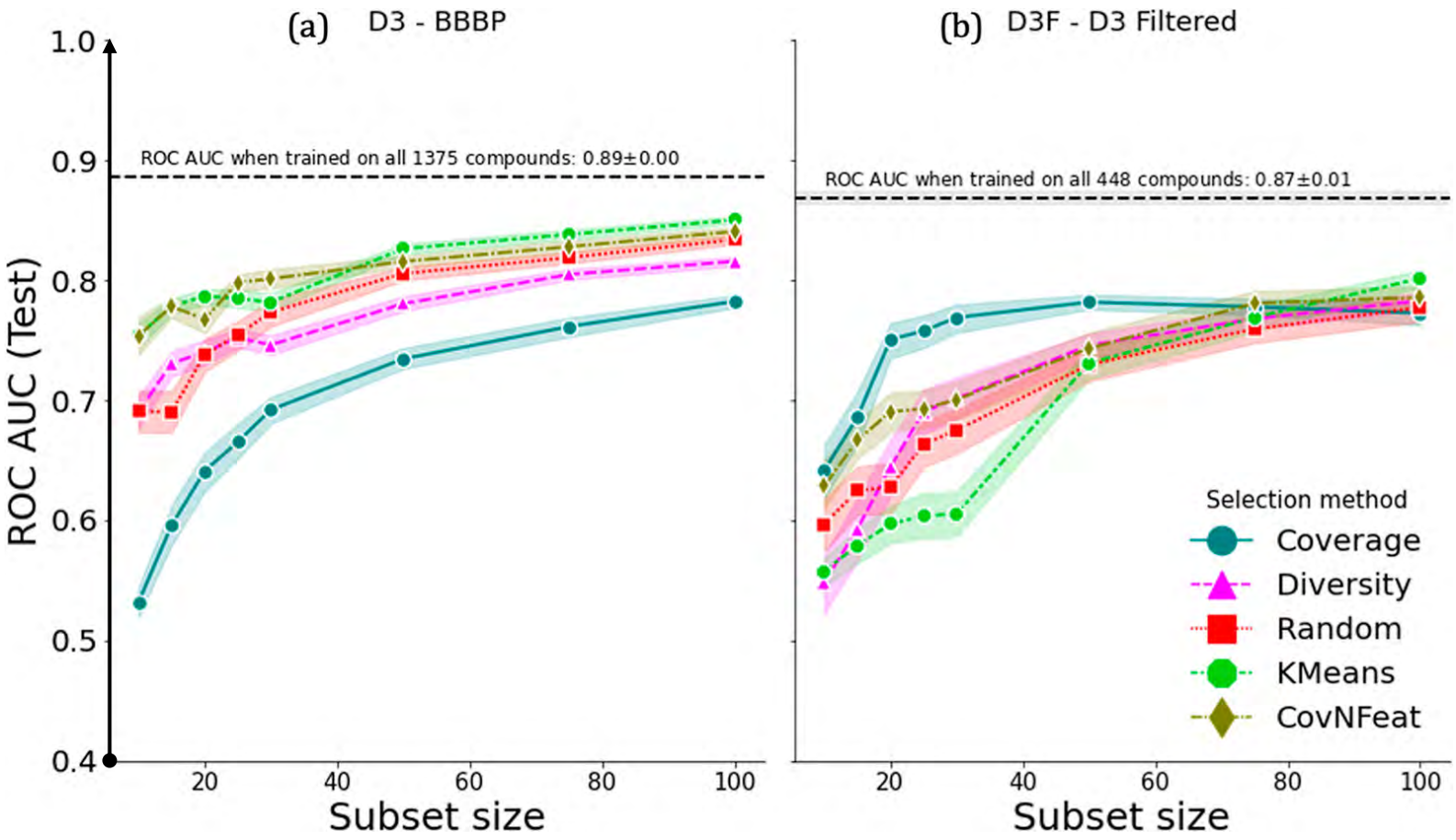
- **D3**
  - *x* = Desalted, deduplicated molecules from MoleculeNet Blood Brain Barrier Penetrance dataset.
  - *y* = {0, 1} classification of brain penetrant (1) or not (0).
- **D3F**
  - D3 filtered for drug-like molecules.



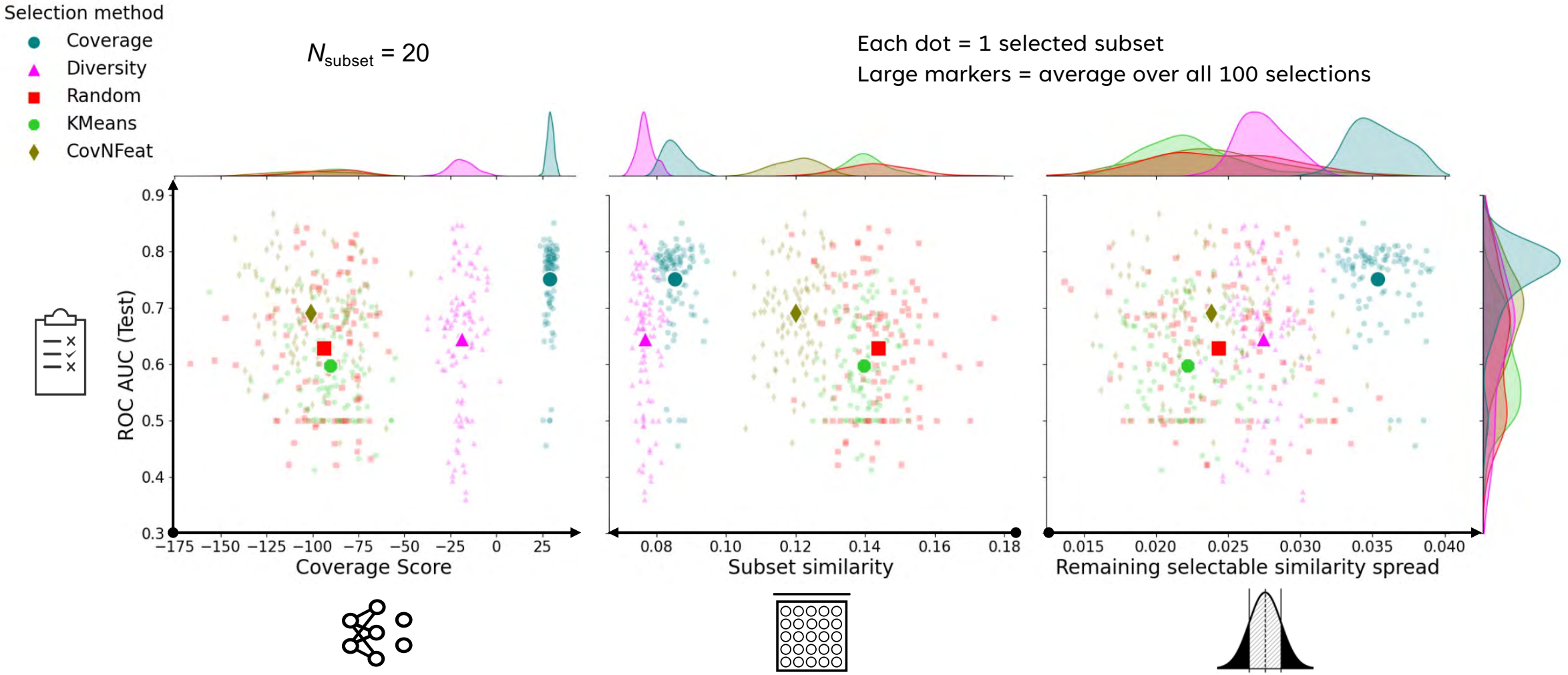
D3 & D3F selections

Blood Brain Barrier Penetrance

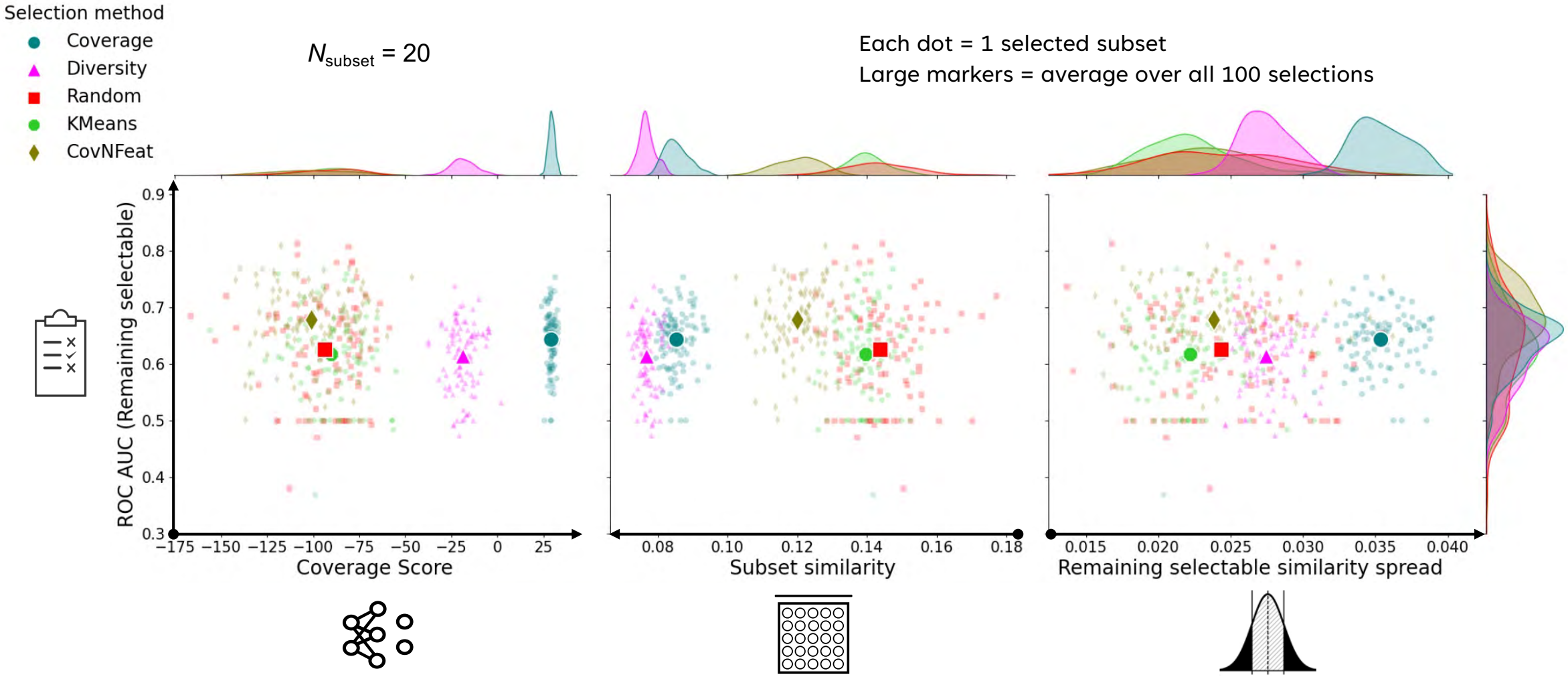
- D3 (left):
  - Coverage performs poorly, optimising for molecules with a larger number of features as well (**CovNFeat**) does better.
- D3F (right):
  - Coverage performs much better.



D3F selections

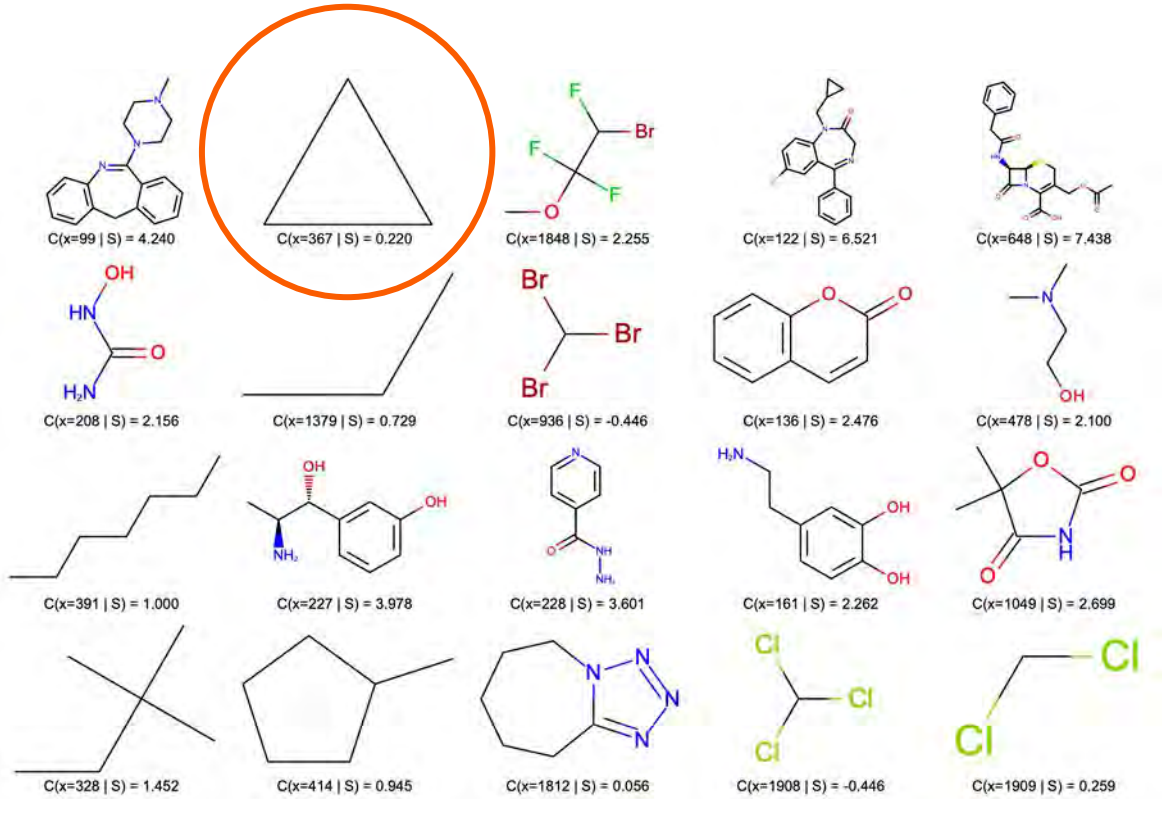


D3F selections

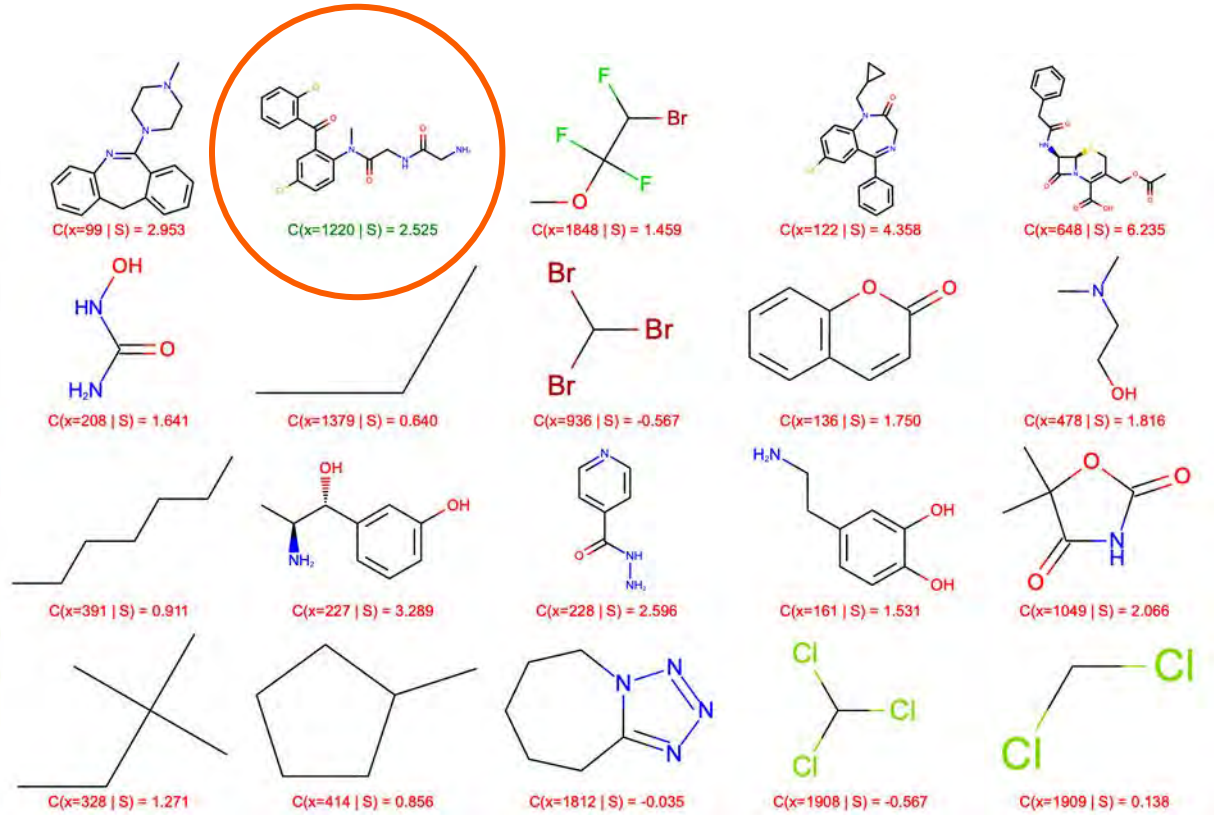


D3

Coverage Score selection, N_{subset} = 20



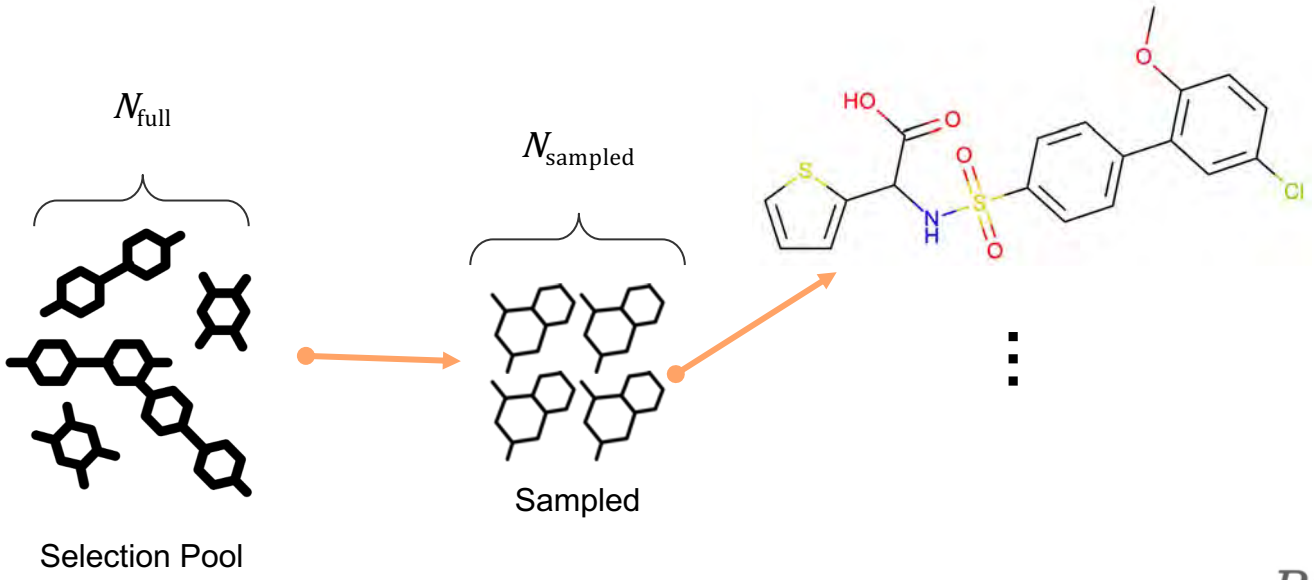
Subset Coverage Score = 43.49

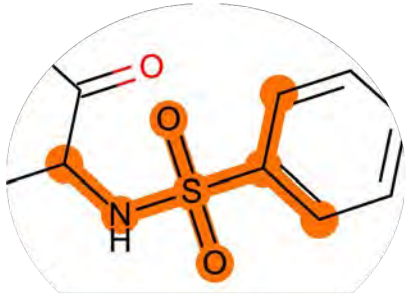


Subset Coverage Score = 34.87

Calculating Coverage Score

Feature Counts



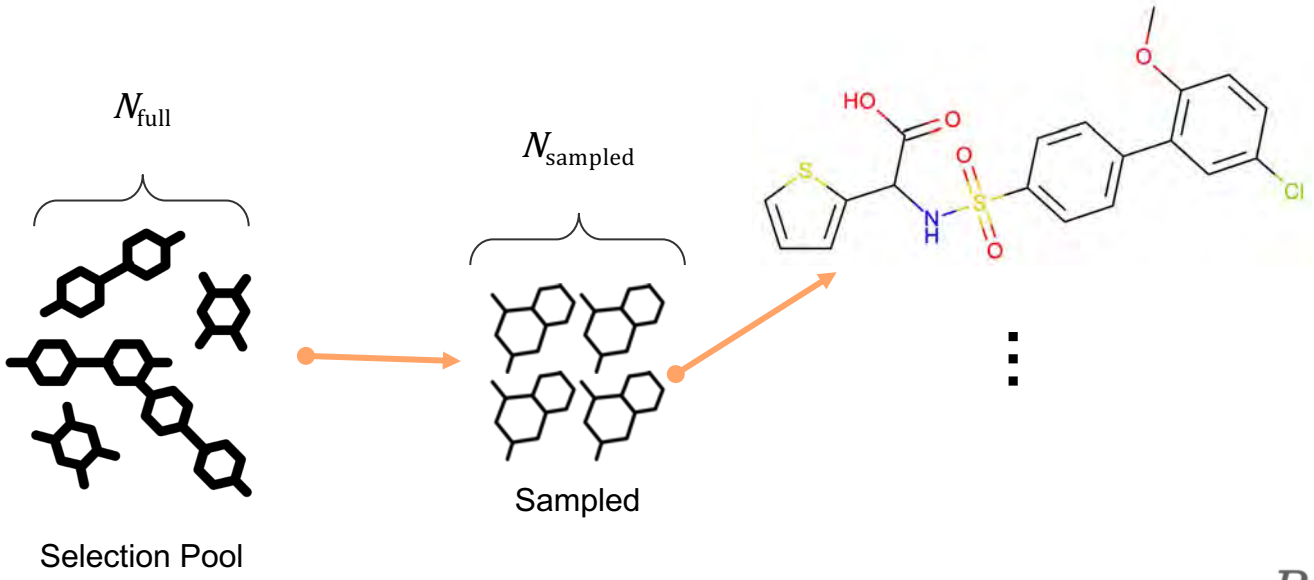


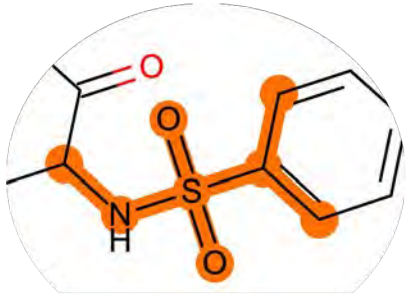
$$P(\text{sampled}) = \frac{N_{\text{sampled}}}{N_{\text{full}}} = P_{\text{s}}$$

$$P(\text{sampled}|x) = P_{\text{s,x}} = \ ?$$

Calculating Coverage Score

Feature Counts





$$P(\text{sampled}) = \frac{N_{\text{sampled}}}{N_{\text{full}}} = P_{\text{s}}$$

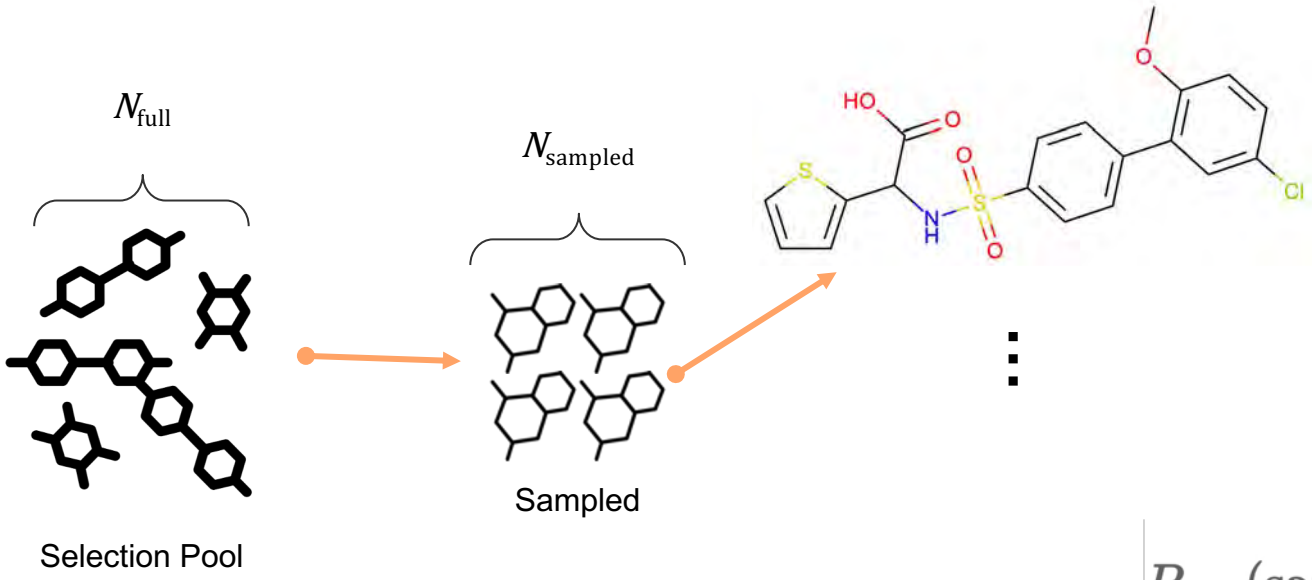
$$P(\text{sampled}|x) = P_{\text{s,x}} = \frac{F_{x,\text{sampled}}}{F_{x,\text{full}}} \quad ?$$

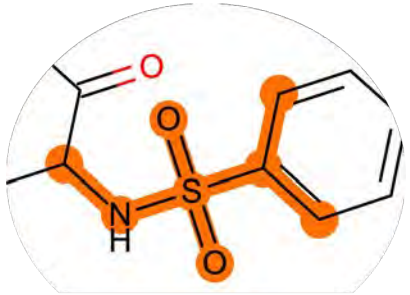
$$F_{x,\text{sampled}} \leq F_{x,\text{full}} \cap F_{x,\text{sampled}} \in \mathbb{Z}_0^+$$

$$F_{x,\text{full}} = 1 \rightarrow P_{s,x} \in \{0, 1\}$$

Calculating Coverage Score

Feature Counts





$$P(\text{sampled}) = \frac{N_{\text{sampled}}}{N_{\text{full}}} = P_{\text{s}}$$

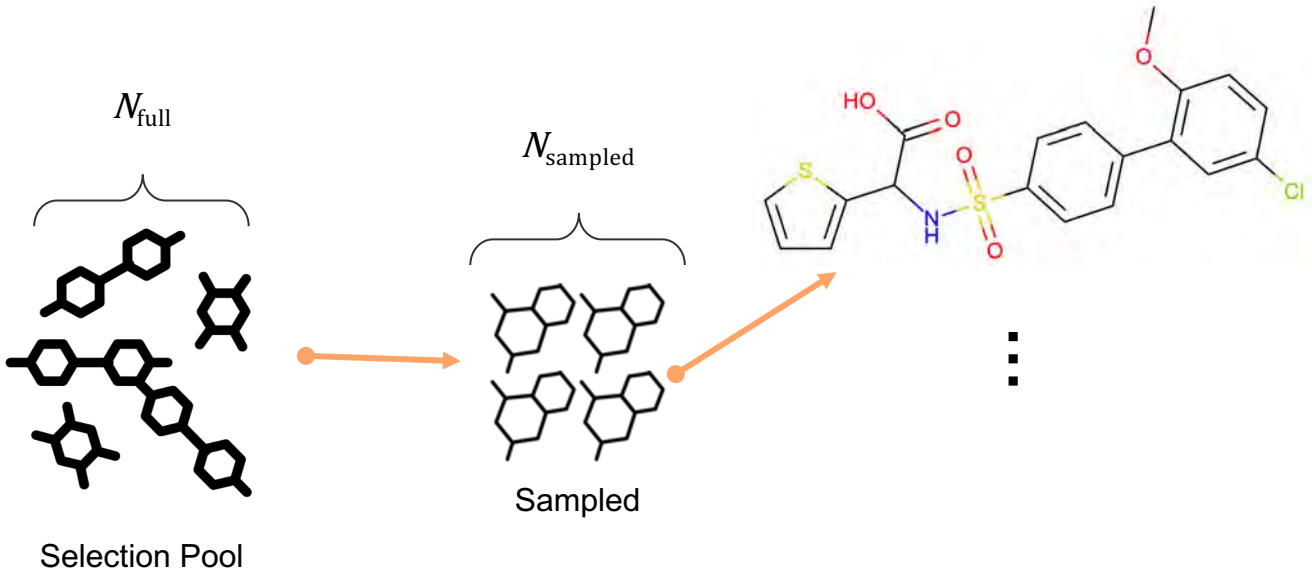
$$P_{\text{corr}}(\text{sampled}|x) = P_{\text{corr,s},x} = \frac{F_{x,\text{sampled}} + 1}{F_{x,\text{full}} + \frac{1}{P_{\text{s}}}}$$

✓

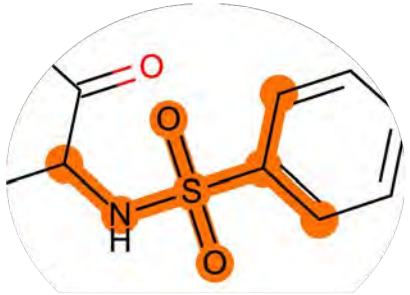
$$F_{x,\text{full}} = 0 \rightarrow P_{\text{corr}}(\text{sampled}|x) = P(\text{sampled})$$

Calculating Coverage Score

Feature Counts



$$P(\text{sampled}) = \frac{N_{\text{sampled}}}{N_{\text{full}}} = P_{\text{s}}$$

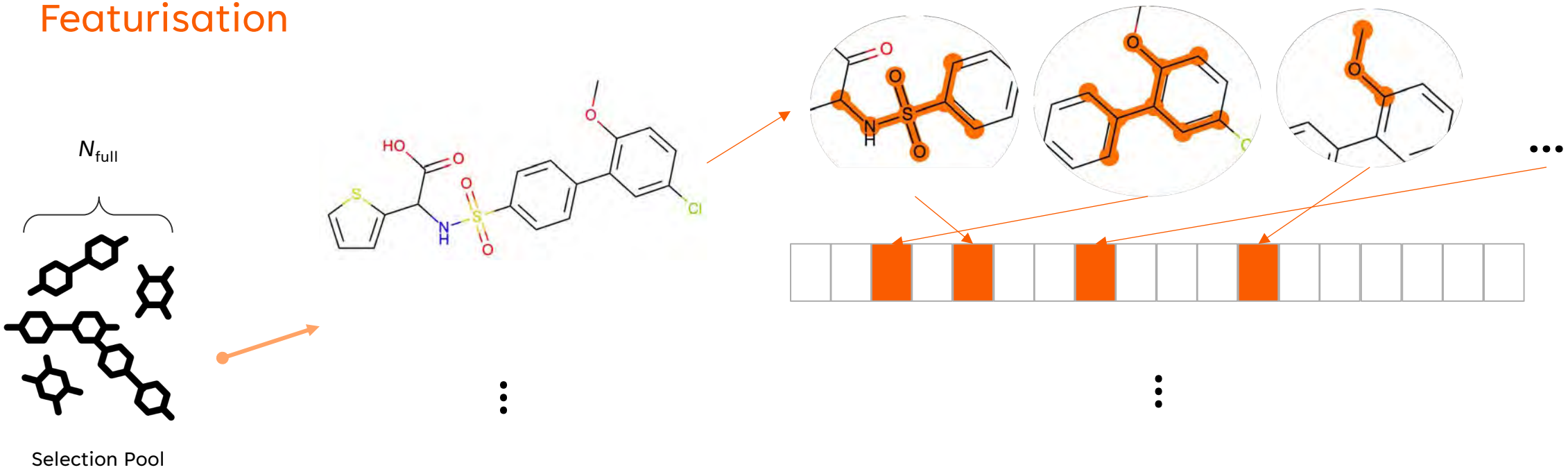


$$C_{x,\text{base}} = -\ln\left(\frac{P_{\text{corr,s},x}}{P_{\text{s}}}\right)$$

$$P_{\text{corr,s},x} > P_{\text{s}} \rightarrow C_{x,\text{base}} < 0$$

Calculating Coverage Score

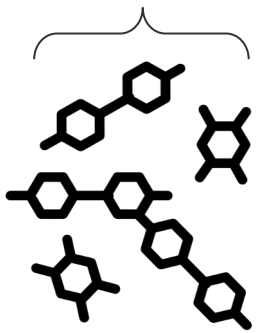
Featurisation



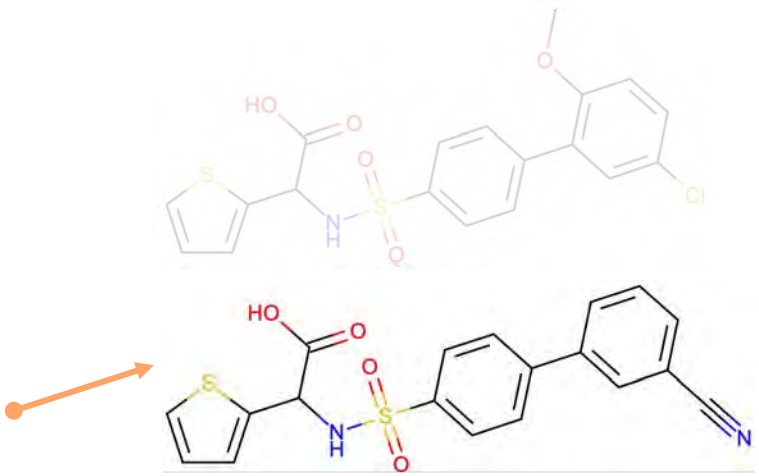
Calculating Coverage Score

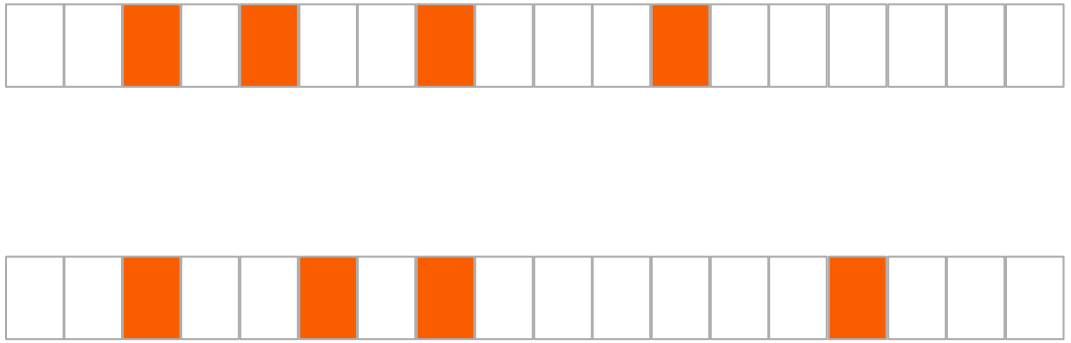
Featurisation

N_{full}



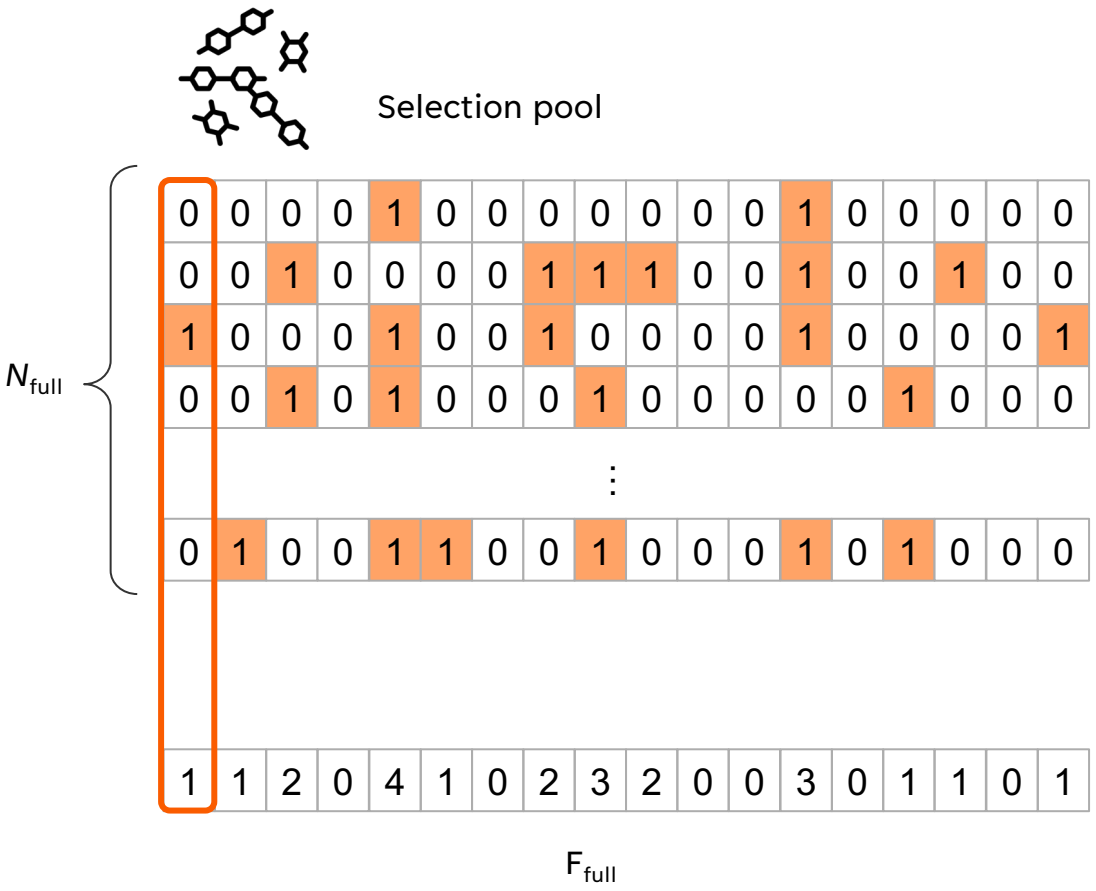
Selection Pool





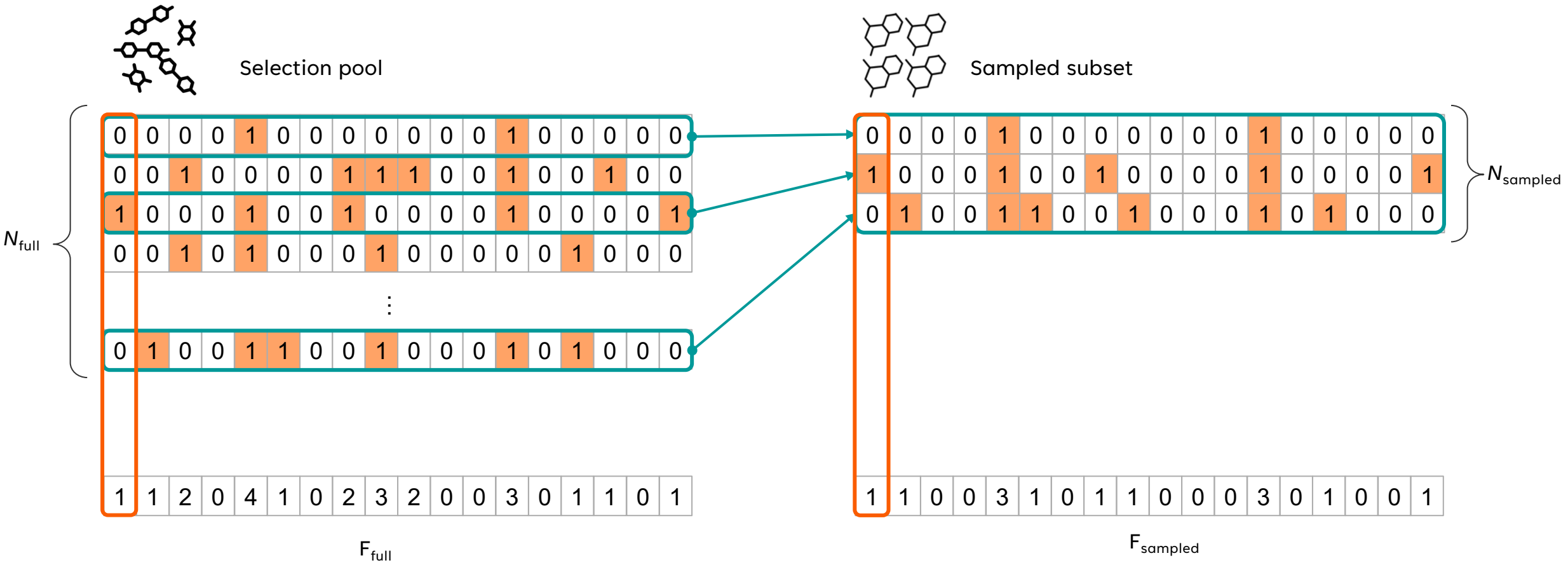
Calculating Coverage Score

Feature counts



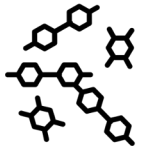
Calculating Coverage Score

Feature counts



Calculating Coverage Score

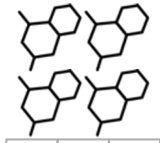
Feature Coverage Score



N_{full} = 10

$\mathbf{F}_{full}$

1	1	2	0	4	1	0	2	3	2	0	0	3	0	1	1	0	1



$N_{sampled}$ = 4

$\mathbf{F}_{sampled}$

1	1	1	0	4	1	0	1	2	0	0	0	3	0	2	0	0	1

$$\mathbf{P}_{full} = \frac{\mathbf{F}_{full}}{N_{full}}$$

$$\mathbf{P}_{sampled} = \frac{\mathbf{F}_{sampled}}{N_{sampled}}$$

$\mathbf{C}_{base}$

$\mathbf{H}$

$$\mathbf{C}_{base} = \ln\left(\frac{\mathbf{P}_{full}N_{sampled} + 1}{\mathbf{P}_{sampled}N_{sampled} + 1}\right)$$

Base Coverage Score
(Bayesian statistics)

$$\mathbf{H} = -\frac{1}{\ln(2)}\left(\mathbf{P}_{sampled}\ln(\mathbf{P}_{sampled}) + (1 - \mathbf{P}_{sampled})\ln((1 - \mathbf{P}_{sampled}))\right)$$

Shannon Information Entropy

$\mathbf{C}_{final}$

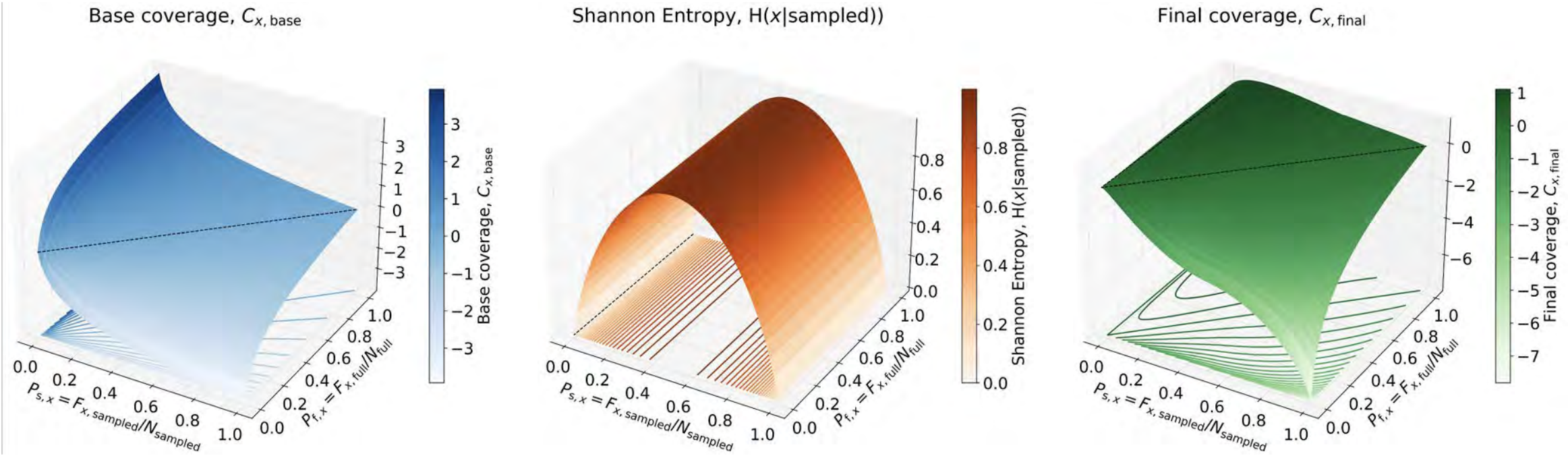
$$\mathbf{C}_{final} = \begin{cases} \mathbf{C}_{base}(2 - \mathbf{H}) & \text{if } \mathbf{C}_{base} < 0 \cap \mathbf{P}_{sampled} > 0.5 \\ \mathbf{C}_{base}\mathbf{H} & \text{otherwise} \end{cases}$$

Final Coverage Score

Calculating Coverage Score

Feature Coverage Score surfaces

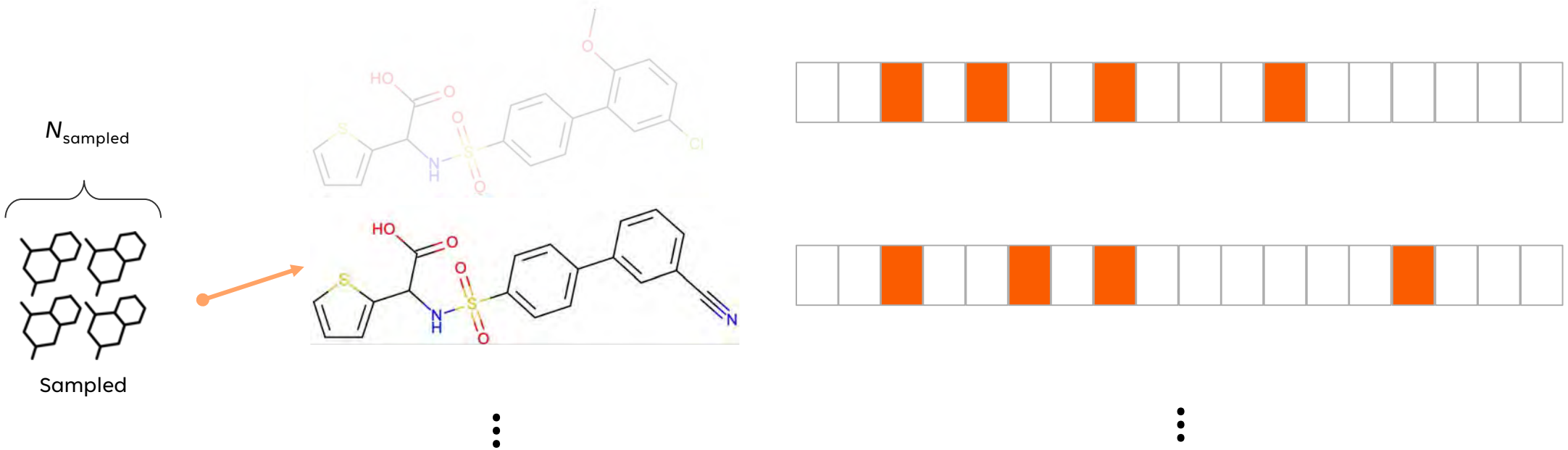
$N_{\text{sampled}} = 50$



- **Higher** the feature count in the **sampled set**, the **lower the score**
- **Higher** the feature count in the **full set**, the **higher the score**
- Score is **penalising** when $\frac{\mathbf{F}_{\text{full}}}{N_{\text{full}}} < \frac{\mathbf{F}_{\text{sampled}}}{N_{\text{sampled}}}$, i.e., when proportion sampled > proportion full

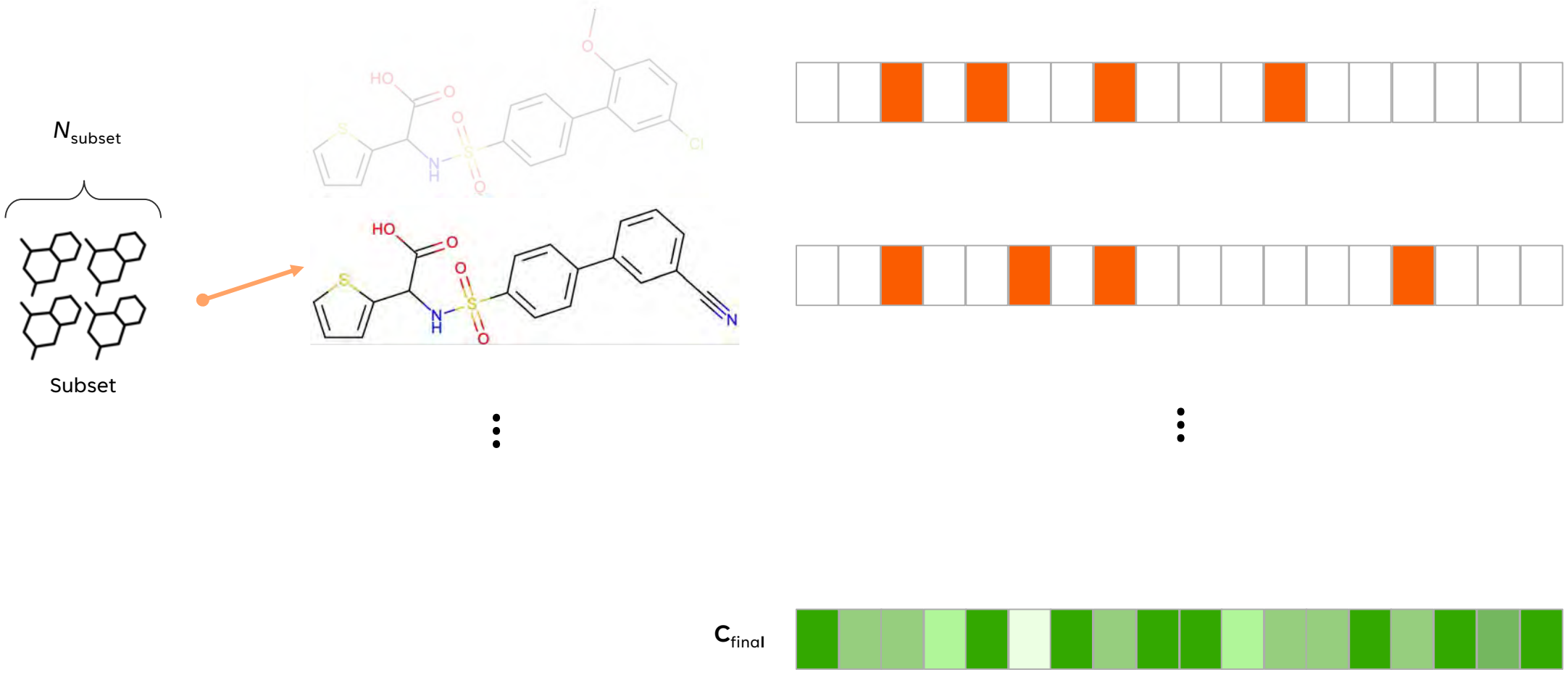
Calculating Coverage Score

Molecular Coverage Score



Calculating Coverage Score

Molecular Coverage Score

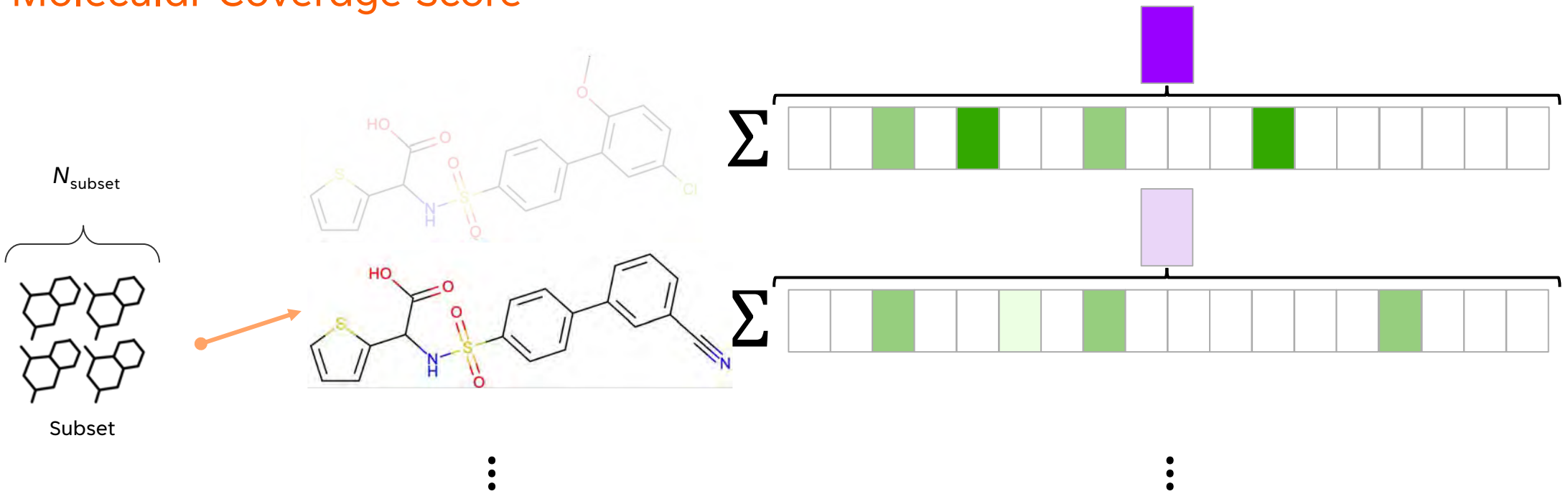


Calculating Coverage Score

Molecular Coverage Score

Calculating Coverage Score

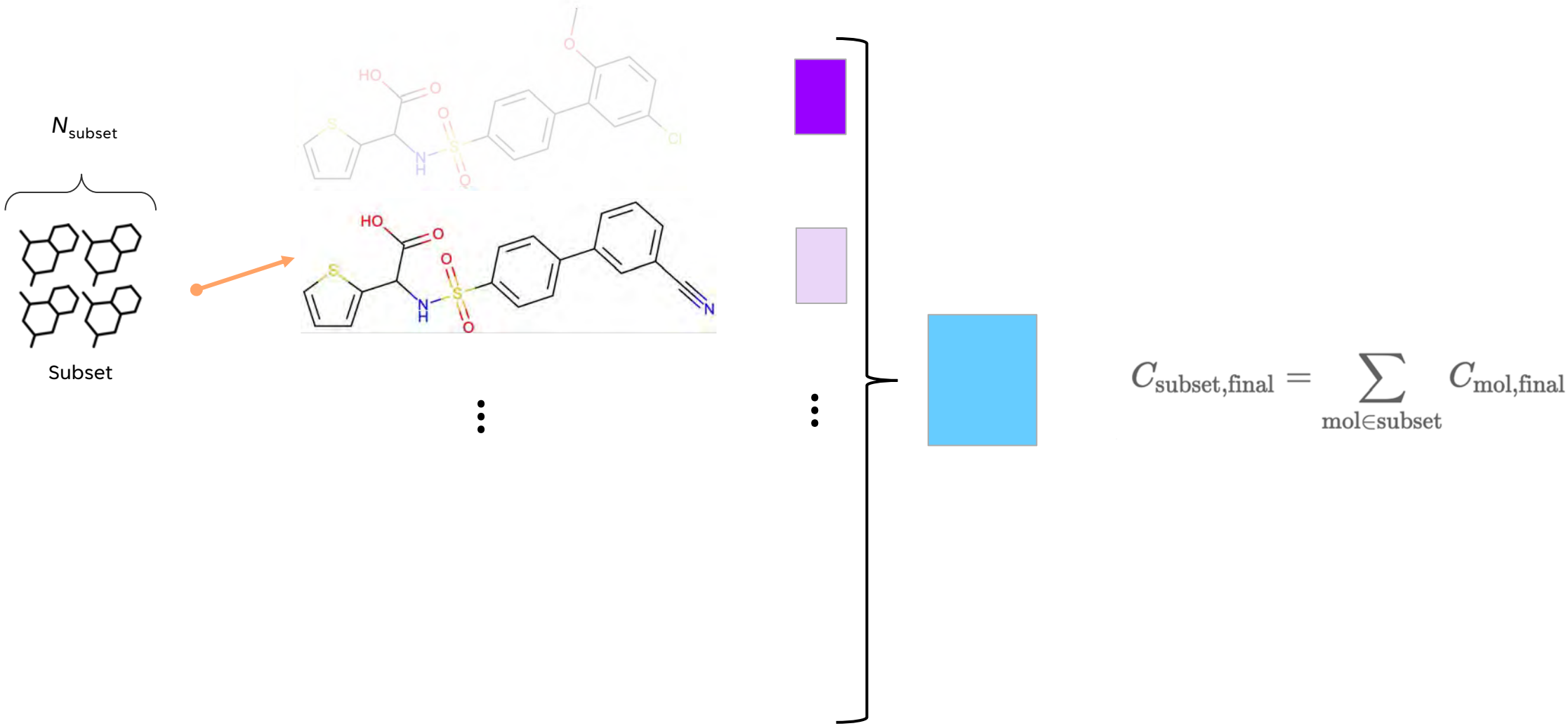
Molecular Coverage Score



$$C_{\text{mol,final}} = \sum_{x \in \text{mol}} C_{x,\text{final}}$$

Calculating Coverage Score

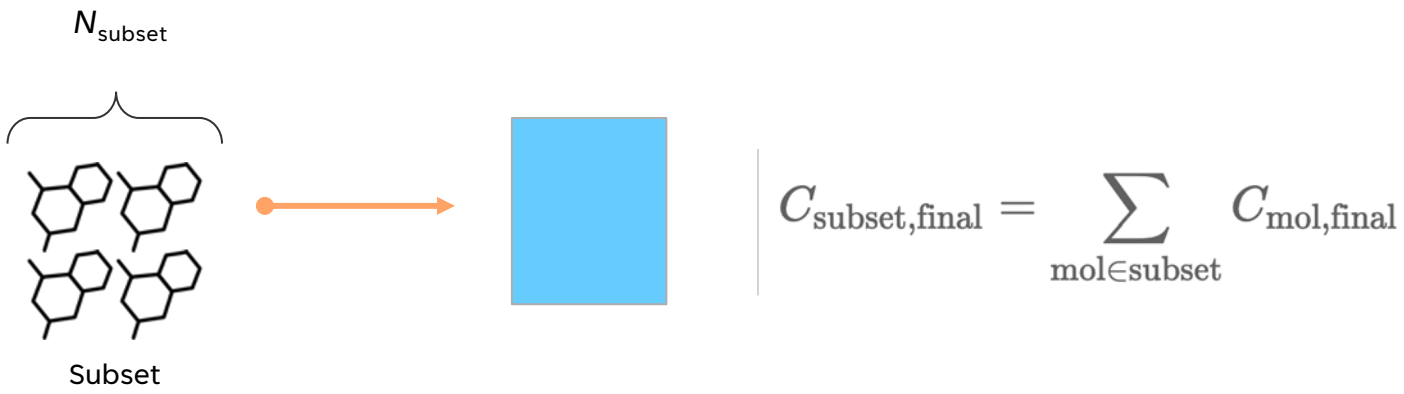
Subset Coverage Score



$$C_{\text{subset,final}} = \sum_{\text{mol} \in \text{subset}} C_{\text{mol,final}}$$

Calculating Coverage Score

Subset Coverage Score



$$C_{\text{subset,final}} = \sum_{\text{mol} \in \text{subset}} C_{\text{mol,final}}$$