Dataset Structure and Active Learning

A Comparative Regression Study

Presentation Structure

- Motivation
- Datasets
- Methods
- Hypotheses
- Results
- Conclusions

Presentation Structure

- **Motivation**
- Datasets
- Methods
- Hypotheses
- Results
- Conclusions

Motivation

- The efficacy of an active learning method depends on the characteristics of the data to which it is applied
- We wanted to study this relationship in the context of regression

Synthetic vs. real data

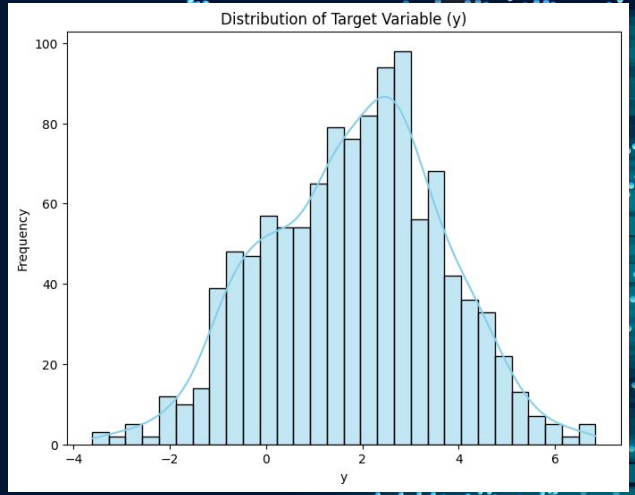
- Synthetic data allows for better isolation of dataset characteristics
- Real data can give us domain-specific insights and – at least to us – is more interesting to work with

Presentation Structure

- Motivation
- **Datasets**
- Methods
- Hypotheses
- Results
- Conclusions

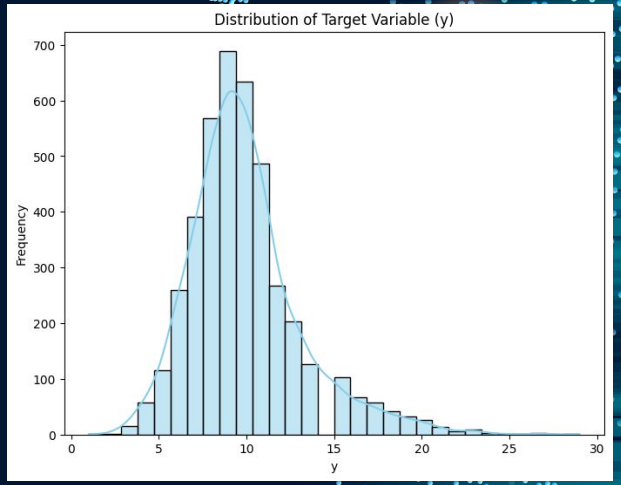
Logd74

- A dataset of 1130 organic compounds and their lipophilicity
- Featurized with Morgan fingerprinting
- Lipophilicity is relevant to drug discovery, as some targets may be within lipid membranes
- (1130 samples, 2048 features)



Abalone Age

- UCI ML repository classic dataset
- Regression Target: Abalone Age
- Features: Sex, Length, Diameter, Height, Weights x6
- Use? Abalone age-identification process is inconvenient
- (4177 samples, 8 features)



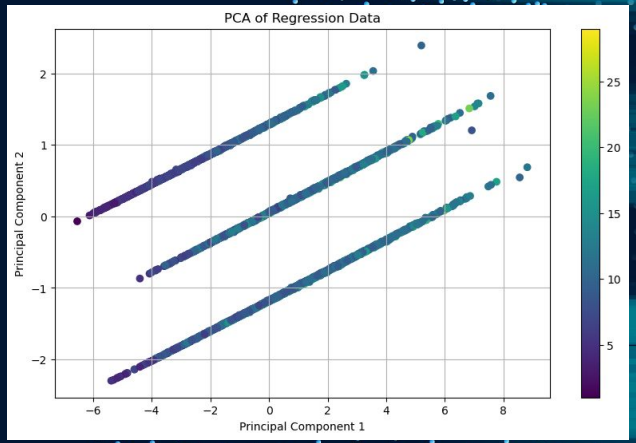
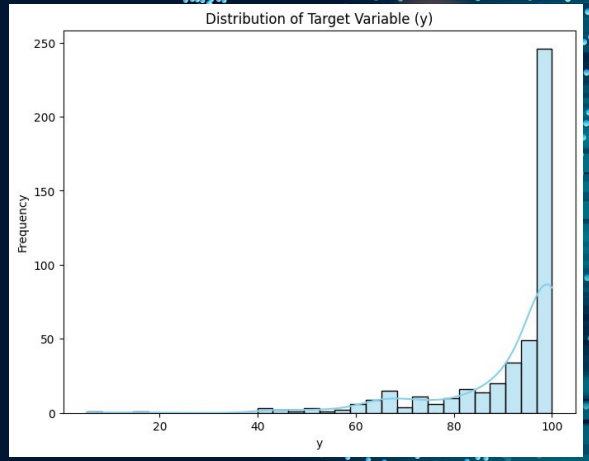
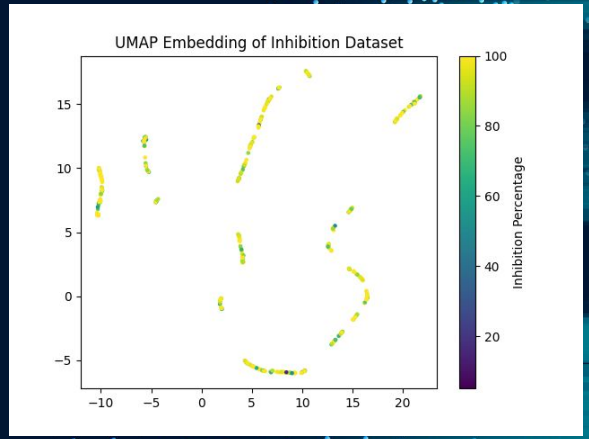


image credits:
https://shopoysters.hogislandoysters.com/products/red-abalone

Kinase Inhibition Data

- A dataset of 569 proteins and their inhibition levels by a small molecule drug
- Featurized with protein physicochemical properties using biopython
- The goal of using this dataset is to capture which chemical properties cause a certain small molecule inhibitor to be most effective
- (569 samples, 28 features)



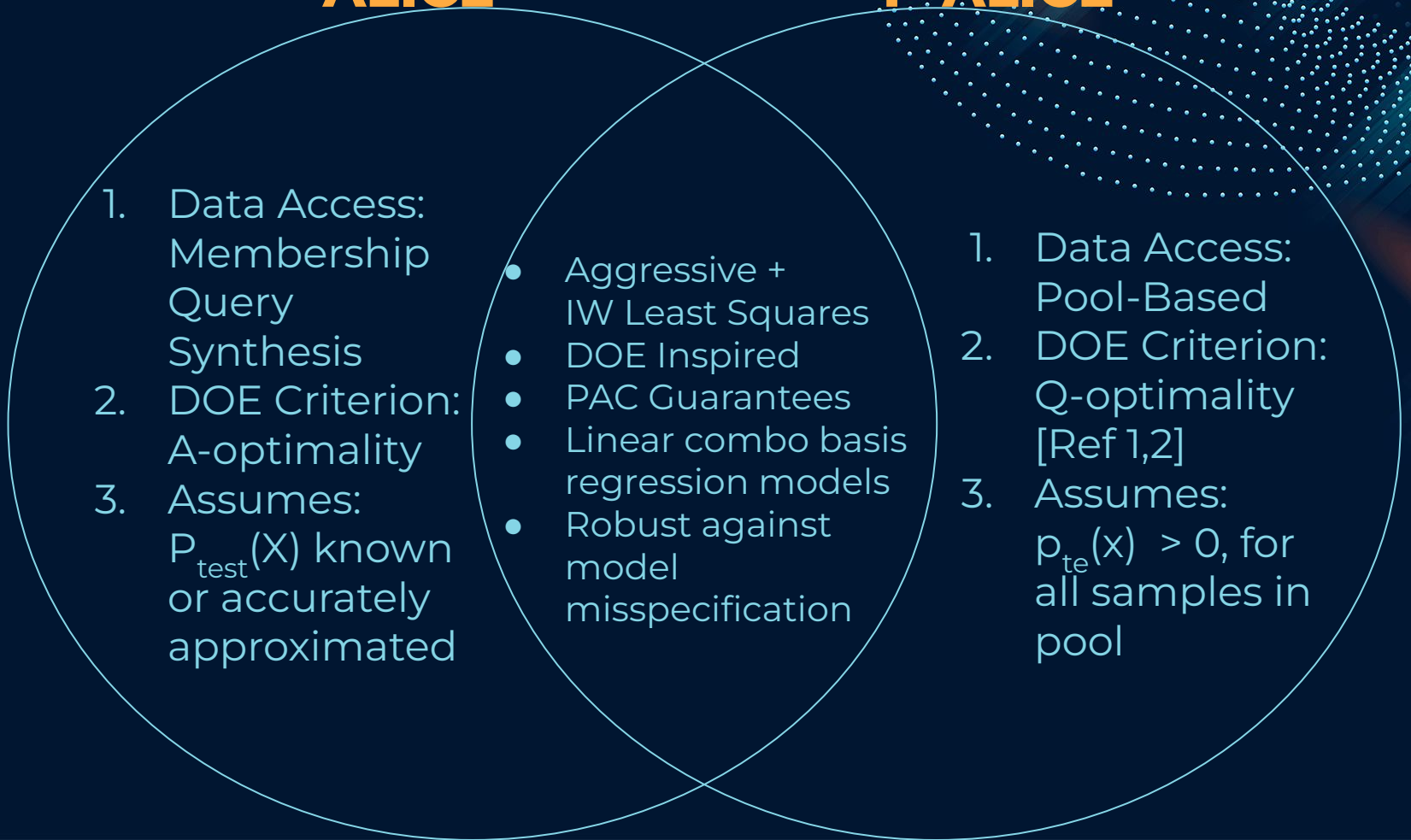


Presentation Structure

- Motivation
- Datasets
- **Methods**
- Hypotheses
- Results
- Conclusions

Random forest batch selection

- Random forest base learner
- Batch query methods:
  - Random
  - Uncertainty
  - K-means diversity
  - Hierarchical diversity


ALICE
P-ALICE
1. Data Access: Membership Query Synthesis
2. DOE Criterion: A-optimality
3. Assumes: $P_{test}(X)$ known or accurately approximated
• Aggressive + IW Least Squares
• DOE Inspired
• PAC Guarantees
• Linear combo basis regression models
• Robust against model misspecification
1. Data Access: Pool-Based
2. DOE Criterion: Q-optimality [Ref 1,2]
3. Assumes: $p_{te}(x) > 0$, for all samples in pool

P-ALICE

P-ALICE/ALICE Base Learner Formula

$$\widehat{f}(\boldsymbol{x}) = \sum_{\ell=1}^{t} \theta_\ell \varphi_\ell(\boldsymbol{x})$$

Eqn 3

Input: Test input points $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ and basis functions $\{\varphi_\ell(\boldsymbol{x})\}_{\ell=1}^{t}$.

Output: Learned parameter $\widehat{\boldsymbol{\theta}}_{\mathrm{W}}$

Compute the $t \times t$ matrix $\widehat{\boldsymbol{U}}$ with $\widehat{U}_{\ell,\ell'} = \frac{1}{n_{\mathrm{te}}} \sum_{j=1}^{n_{\mathrm{te}}} \varphi_\ell(\boldsymbol{x}_j^{\mathrm{te}}) \varphi_{\ell'}(\boldsymbol{x}_j^{\mathrm{te}})$;
For several different values of λ (possibly around $\lambda = 1/2$)
  Compute $\{b_\lambda(\boldsymbol{x}_j^{\mathrm{te}})\}_{j=1}^{n_{\mathrm{te}}}$ with $b_\lambda(\boldsymbol{x}) = (\sum_{\ell,\ell'=1}^{t} [\widehat{\boldsymbol{U}}^{-1}]_{\ell,\ell'} \varphi_\ell(\boldsymbol{x}) \varphi_{\ell'}(\boldsymbol{x}))^\lambda$;
  Choose $\mathcal{X}_\lambda^{\mathrm{tr}} = \{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ from $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ with probability proportional to $\{b_\lambda(\boldsymbol{x}_j^{\mathrm{te}})\}_{j=1}^{n_{\mathrm{te}}}$;
  Compute the $n_{\mathrm{tr}} \times t$ matrix $\boldsymbol{X}_\lambda$ with $[X_\lambda]_{i,\ell} = \varphi_\ell(\boldsymbol{x}_i^{\mathrm{tr}})$;
  Compute the $n_{\mathrm{tr}} \times n_{\mathrm{tr}}$ diagonal matrix $\boldsymbol{W}_\lambda$ with $[W_\lambda]_{i,i} = (b_\lambda(\boldsymbol{x}_i^{\mathrm{tr}}))^{-1}$;
  Compute $\boldsymbol{L}_\lambda = (\boldsymbol{X}_\lambda^\top \boldsymbol{W}_\lambda \boldsymbol{X}_\lambda)^{-1} \boldsymbol{X}_\lambda^\top \boldsymbol{W}_\lambda$;
  Compute P-ALICE$(\lambda) = \mathrm{tr}(\widehat{\boldsymbol{U}} \boldsymbol{L}_\lambda \boldsymbol{L}_\lambda^\top)$;
End
Compute $\widehat{\lambda} = \mathrm{argmin}_\lambda$ P-ALICE(λ);
Gather training output values $\boldsymbol{y}^{\mathrm{tr}} = (y_1^{\mathrm{tr}}, y_2^{\mathrm{tr}}, \ldots, y_{n_{\mathrm{tr}}}^{\mathrm{tr}})^\top$ at $\mathcal{X}_{\widehat{\lambda}}^{\mathrm{tr}}$;
Compute $\widehat{\boldsymbol{\theta}}_{\mathrm{W}} = \boldsymbol{L}_{\widehat{\lambda}} \boldsymbol{y}^{\mathrm{tr}}$;

Fig. 3 Pseudo code of proposed pool-based active learning algorithm

Ref [1]

P-ALICE Base Learners

1. Linear Regression
2. EDA Spearman Significant Features for Squared, Cubic, and First-Order Interactions
3. -1/+1 map for 0-1 Binary Features

Used for Abalone Age & Kinase Inhibition

Used for Logd74

$$\varphi_\ell(\vec{x}) : \{x_l, x_l^2, x_l^3, sin(x_l), ReLu(x_l), \mathbb{1}(x_l), \cdots\}$$

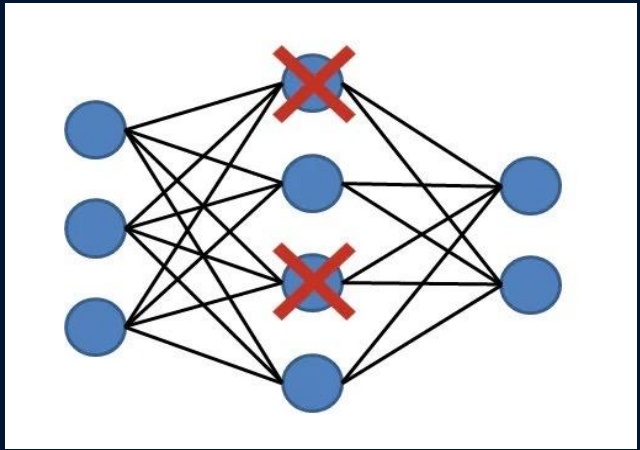
$$x_l : \{x_i, x_i x_j, x_i x_j x_k, \cdots\}$$

Neural network batch selection

- Base learner: a multi-layer perceptron (MLP)
- Batch query methods:
  - Random
  - Uncertainty
  - COVDROP

Neural network uncertainty estimation

- To estimate uncertainty, perform forward passes with dropout repeatedly and calculate variance of predictions



COVDROP

- Premise: the k most uncertain instances may be highly correlated
- Instead, choose points that maximize joint entropy
  - Equivalent to choosing points maximizing determinant of covariance matrix

Determinant Maximization Sampling

Input: $\mathcal{S} = \{x_0, \ldots, x_N\}$, sample pool
Input: $\mathcal{M}$, trained model
Input: K batch size; M # of starts
Data: $\mathrm{Cov} \in \mathbb{R}^{N \times N}$, covariance for $\mathcal{M}$ on $\mathcal{S}$, (computation given elsewhere)
begin
  Extract variances Var from Cov ;
  μ probability measure on $\mathcal{S}$, proportional to $\mathrm{Quantile}(\mathrm{Var})$;
  Choose $B_m^0 = \left(x_{m1}^0, \ldots, x_{mK}^0\right)$, $0 \le m < M$ random starting batches ;
  $\mathrm{Cov}(B_m^0)$, principal submatrices of Cov;
  Compute scores $S_m^0 := \log\det \mathrm{Cov}(B_m^0)$ via Cholesky decomposition;
  Optional *Keep only $M' < M$ highest scoring starts*;
  $i \leftarrow 0$, $i \in \mathbb{Z}/K\mathbb{Z}$;
  while *not converged* **do**
    for $m = 1$ **to** M **do**
      for $j = 1$ **to** N **do**
        Let $B_{m,j}^i$ be B_m^i with x_j substituted at the ith position;
        Compute score $S_{m,j}^i := \log\det \mathrm{Cov}(B_{m,j}^i)$ via rank-1 Cholesky update;
      $B_m^{i+1} \leftarrow B_{m,J}^i$ where $J = \arg\max_j S_{m,j}^i$;
    $i \leftarrow i + 1$;
  Result: $B = B_b^i$ where $b = \arg\max_m B_m^i$

Presentation Structure

- Motivation
- Datasets
- Methods
- **Hypotheses**
- Results
- Conclusions

Hypotheses (Batch sampling)

- We expected batch methods that consider the makeup of a batch to outperform methods that do not
- Thus, both methods below should outperform selecting the k most uncertain instances:
  - Random forest cluster-based batch sampling
  - Neural network COVDROP

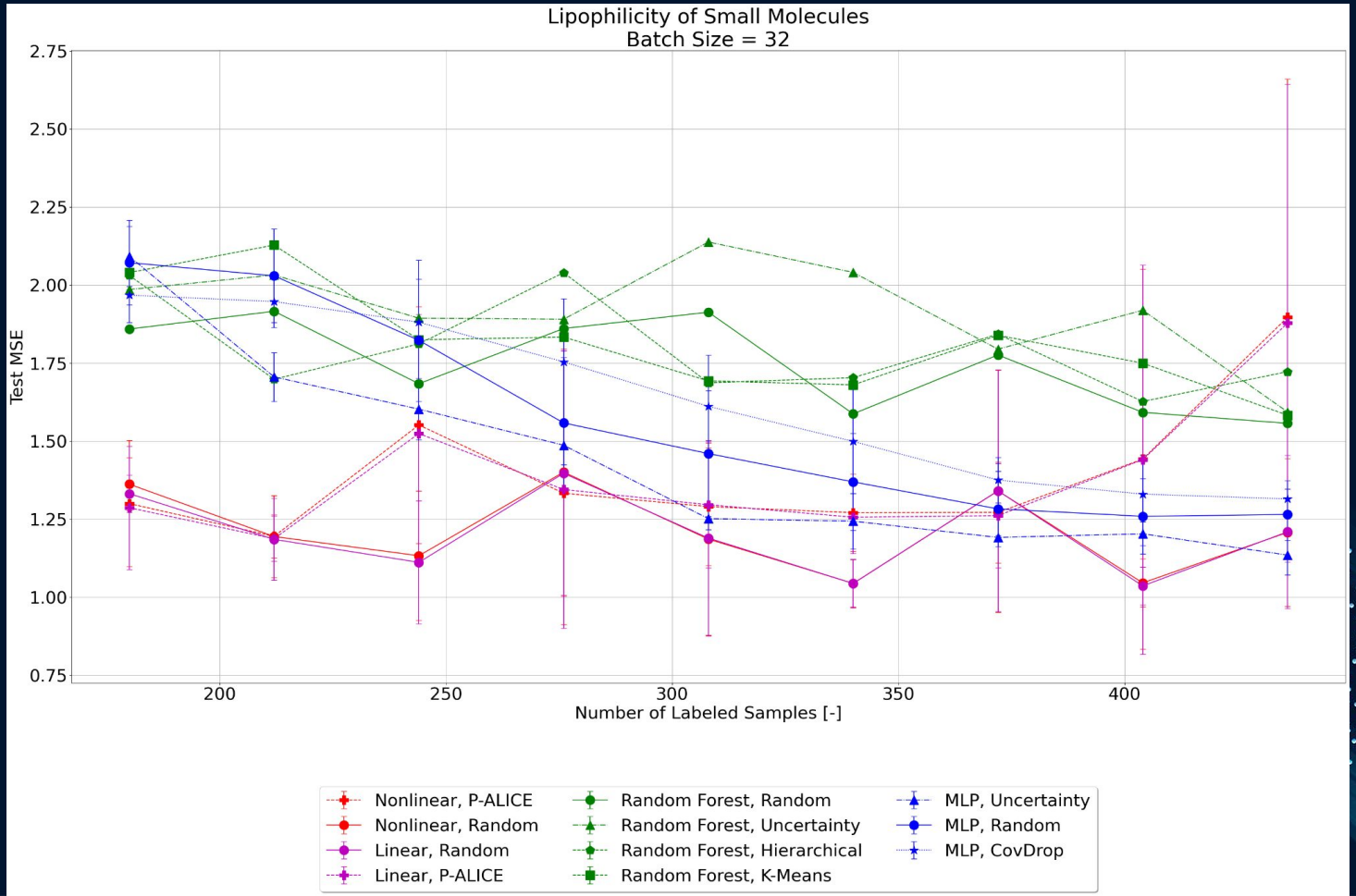
Hypotheses (Dataset structure)

- We expected batch methods that consider the makeup a batch to have a greater performance advantage when the dataset is highly heterogeneous
  - E.g. logd74
- We expected P-ALICE active sampling to be less effective on noisy and high-dim. datasets
  - E.g. logd74

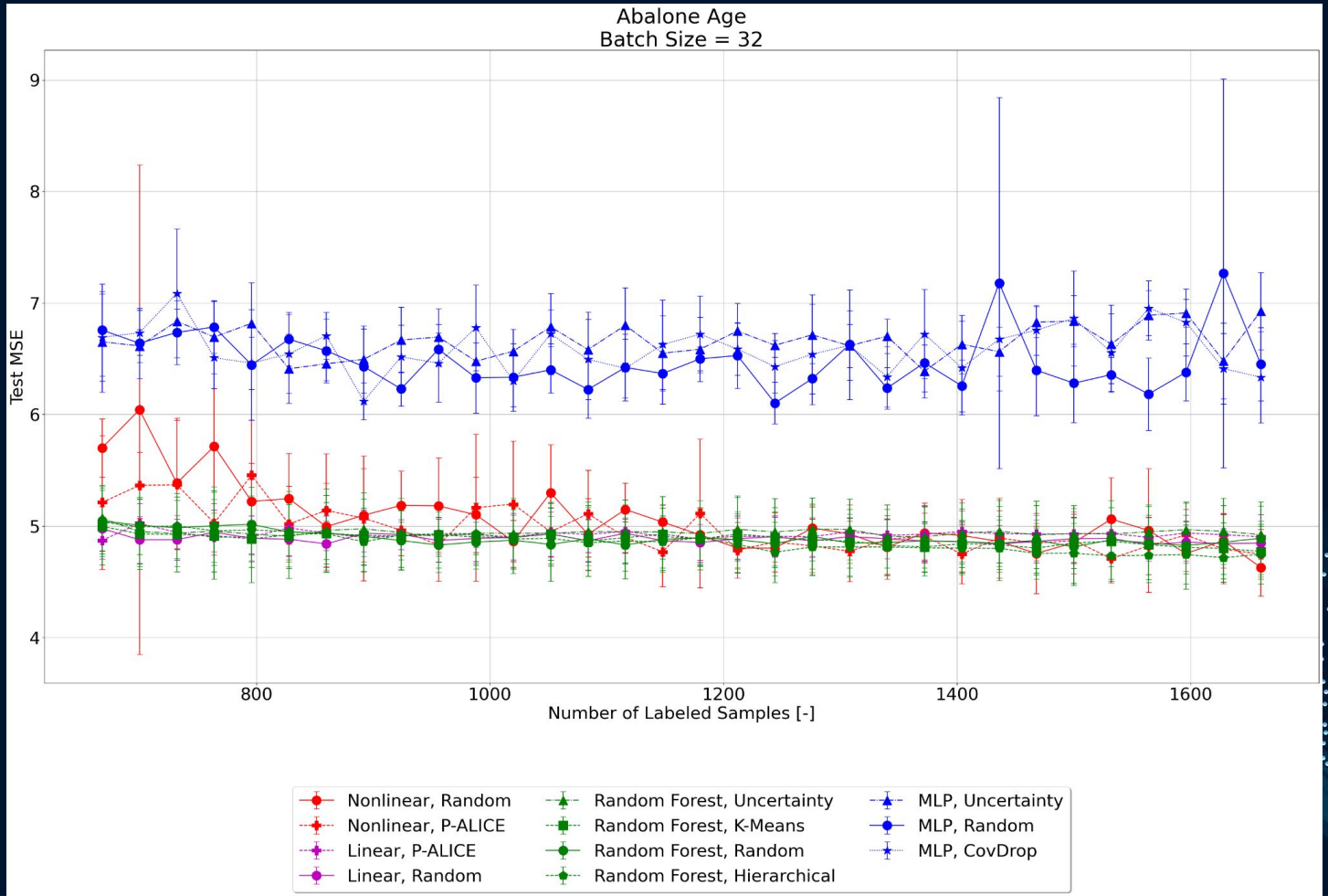
Presentation Structure

- Motivation
- Datasets
- Methods
- Hypotheses
- **Results**
- Conclusions

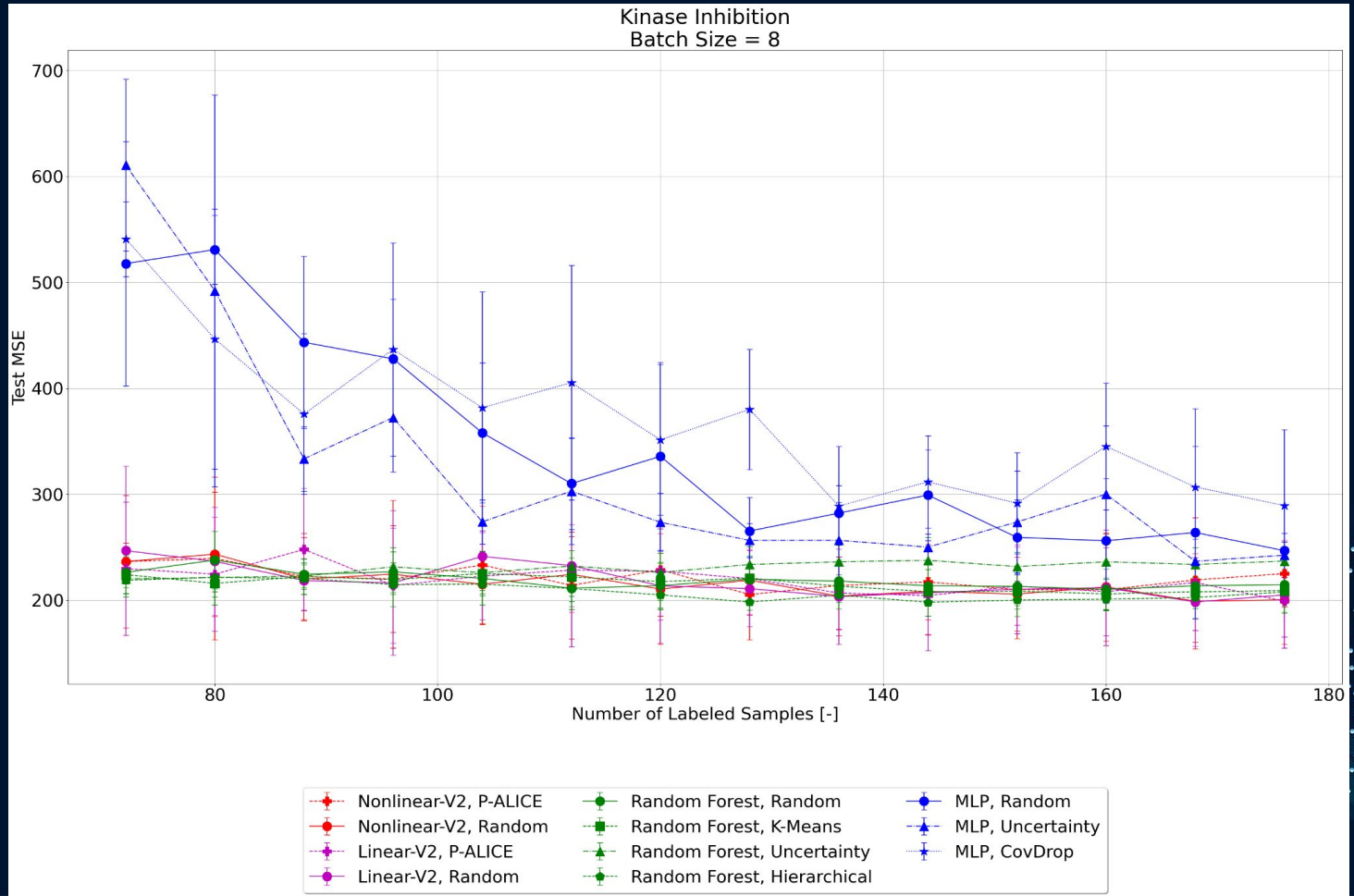
Results Logd74



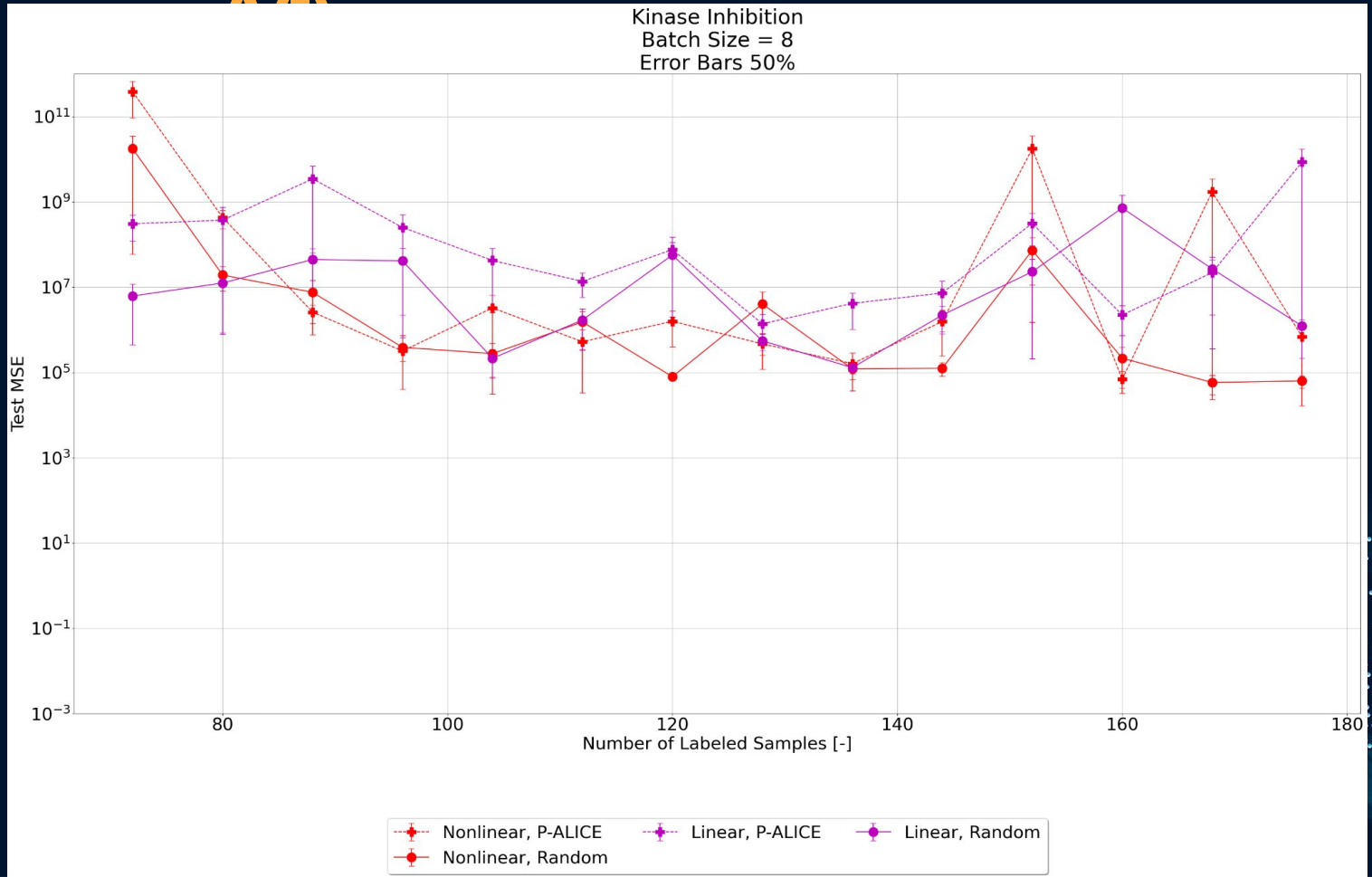
Results Abalone



Results Kinase Inhibition (V2)



Results Kinase Inhibition



Presentation Structure

- Motivation
- Datasets
- Methods
- Hypotheses
- Results
- **Conclusions**

Conclusions

- Base learners can be a confounding variable when evaluating active learning methods across different base learners
- Real datasets can be rewarding to work with, but summarizing their differences can be challenging
- Let us know if you have ideas about how we can improve our analysis!

References

[1] https://link.springer.com/article/10.1007/s10994-009-5100-3
[2] https://statweb.rutgers.edu/buyske/591/lect11.pdf
[3] https://www.researchgate.net/publication/254196943_Using_Random_Forest_to_Learn_Imbalanced_Data#:~:text=Lastly%2C%20ensemble%20methods%2C%20such%20as,class%20%5B7%2C%2018%5D.
[4] https://elifesciences.org/reviewed-preprints/89679