Inference for Interpretable Machine Learning: Feature Importance and Beyond

Genevera I. Allen

Department of Statistics,
Center for Theoretical Neuroscience,
Zuckerman Institute,
Irving Institute,
Columbia University.

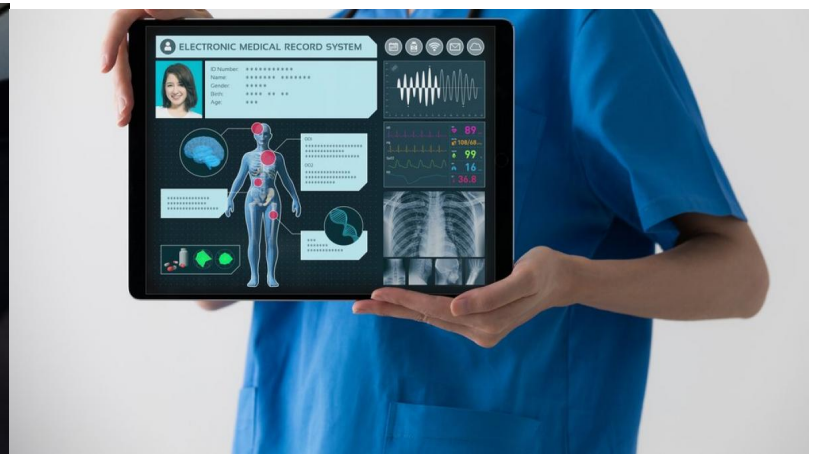
May 23, 2025

Interpretable Machine Learning

AI is being used to make critical decisions in our lives.

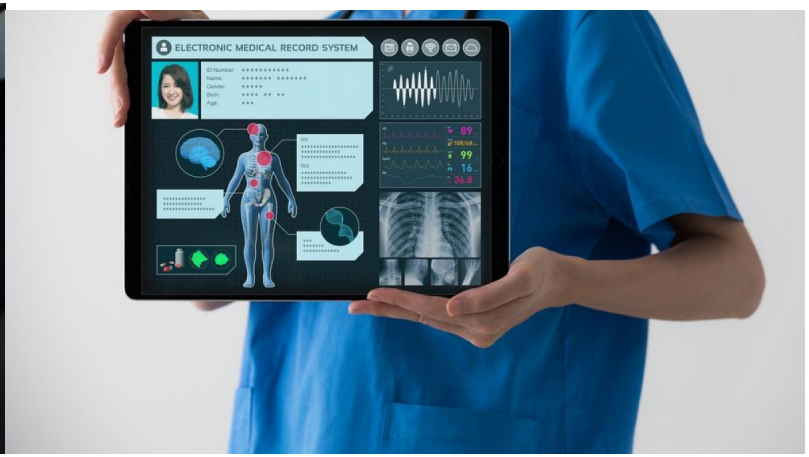
Interpretable Machine Learning

Can we trust AI (Machine Learning)?



Interpretable Machine Learning

Can we trust AI (Machine Learning)?



Make it interpretable or explainable!

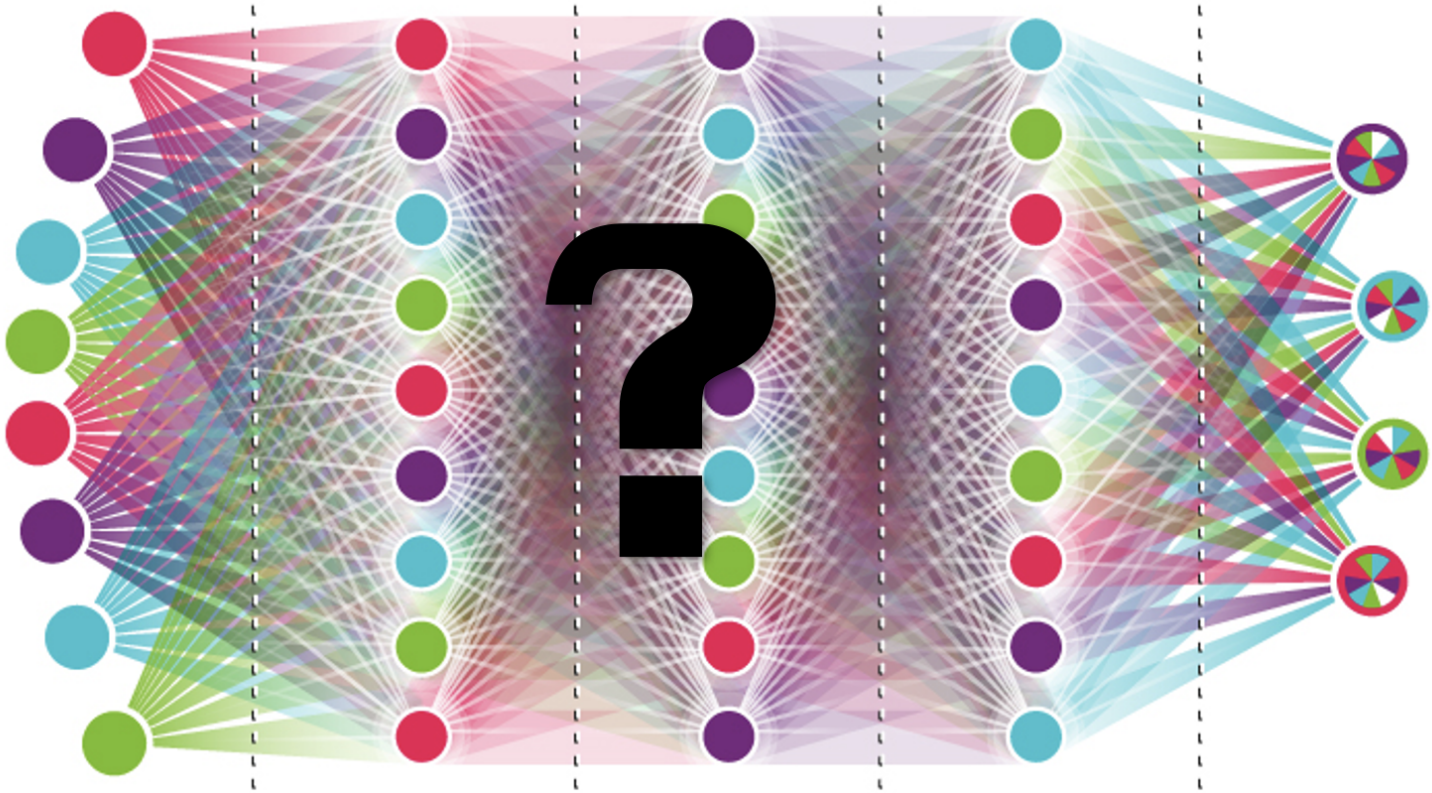
Interpretable Machine Learning

Can we trust AI (Machine Learning)?

Make it interpretable or explainable!

But, can we trust the interpretations?

Interpretable Machine Learning



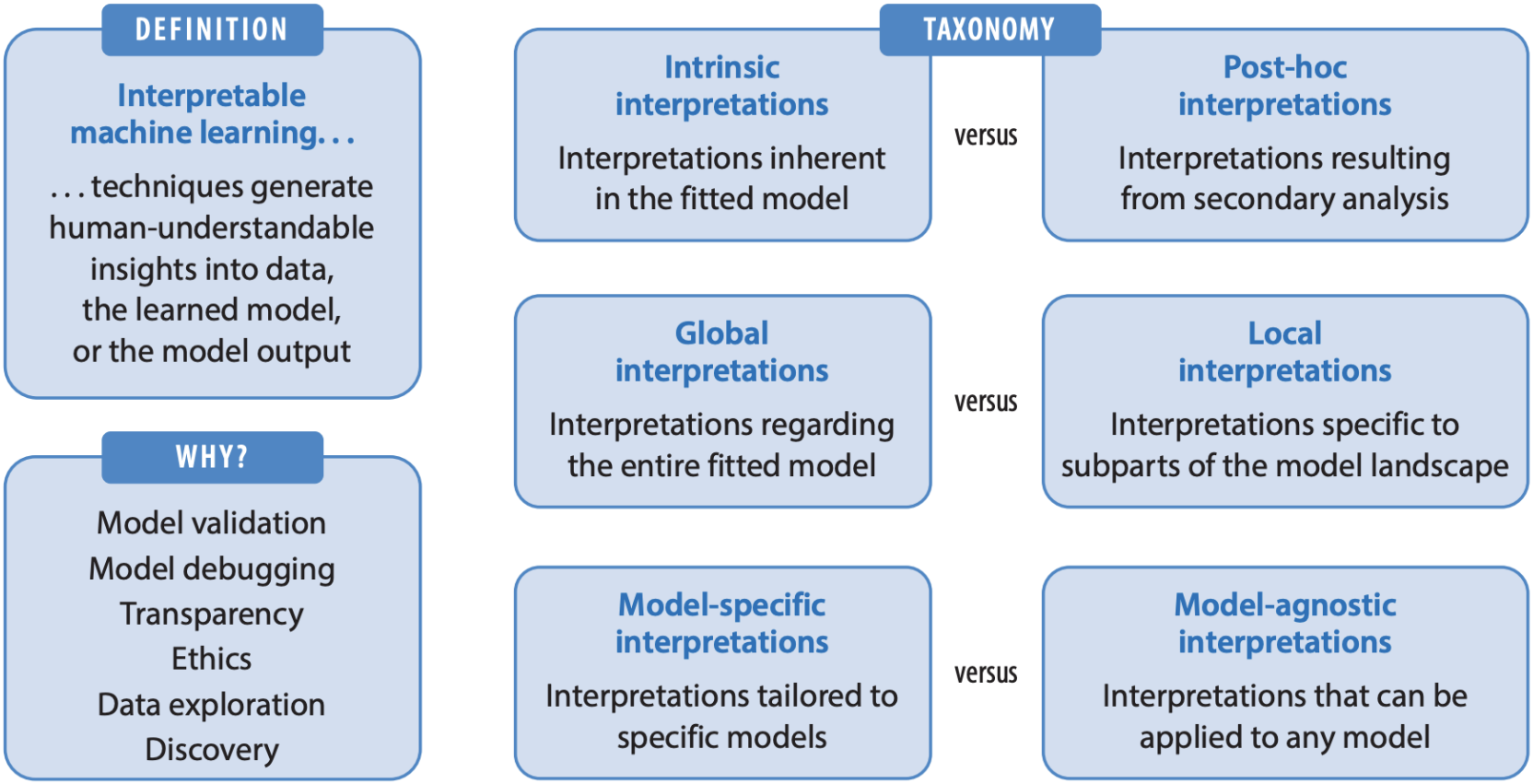
What goes on inside the black-box?

Interpretable Machine Learning

Interpretable Machine Learning

Machine learning (ML) techniques to find human understandable insights into the data, the ML model, or the ML model's output.

Interpretable Machine Learning



Allen et al., 2024

Our Focus: Interpreting Features

Feature Importance

How do the input features effect the model's output?

Focus: Feature Importance for **Tabular Data**.

Our Focus: Interpreting Features

Tons of measures of feature importance!

Model-Specific:

- Coefficients in GLMs, MDI for Trees, DeepLIFT, LRP, etc.

Model-Agnostic:

- Permutation (Konig et al., 2021).
  - Permute feature j and assess effect on predictions.
- Shapley values (Lundberg and Lee, 2017).
  - Assess the predictive value of adding feature j to all possible subsets of features.
- **Occlusion** (Covert et al., 2021).
  - Remove feature j and assess effect on predictions.

Our Focus: Interpreting Features

Tons of measures of feature importance!

Model-Specific:

- Coefficients in GLMs, MDI for Trees, DeepLIFT, LRP, etc.

Model-Agnostic:

- Permutation (Konig et al., 2021).
  - Permute feature j and assess effect on predictions.
- Shapley values (Lundberg and Lee, 2017).
  - Assess the predictive value of adding feature j to all possible subsets of features.
- **Occlusion** (Covert et al., 2021).
  - Remove feature j and assess effect on predictions.

Can we trust feature importance measures?

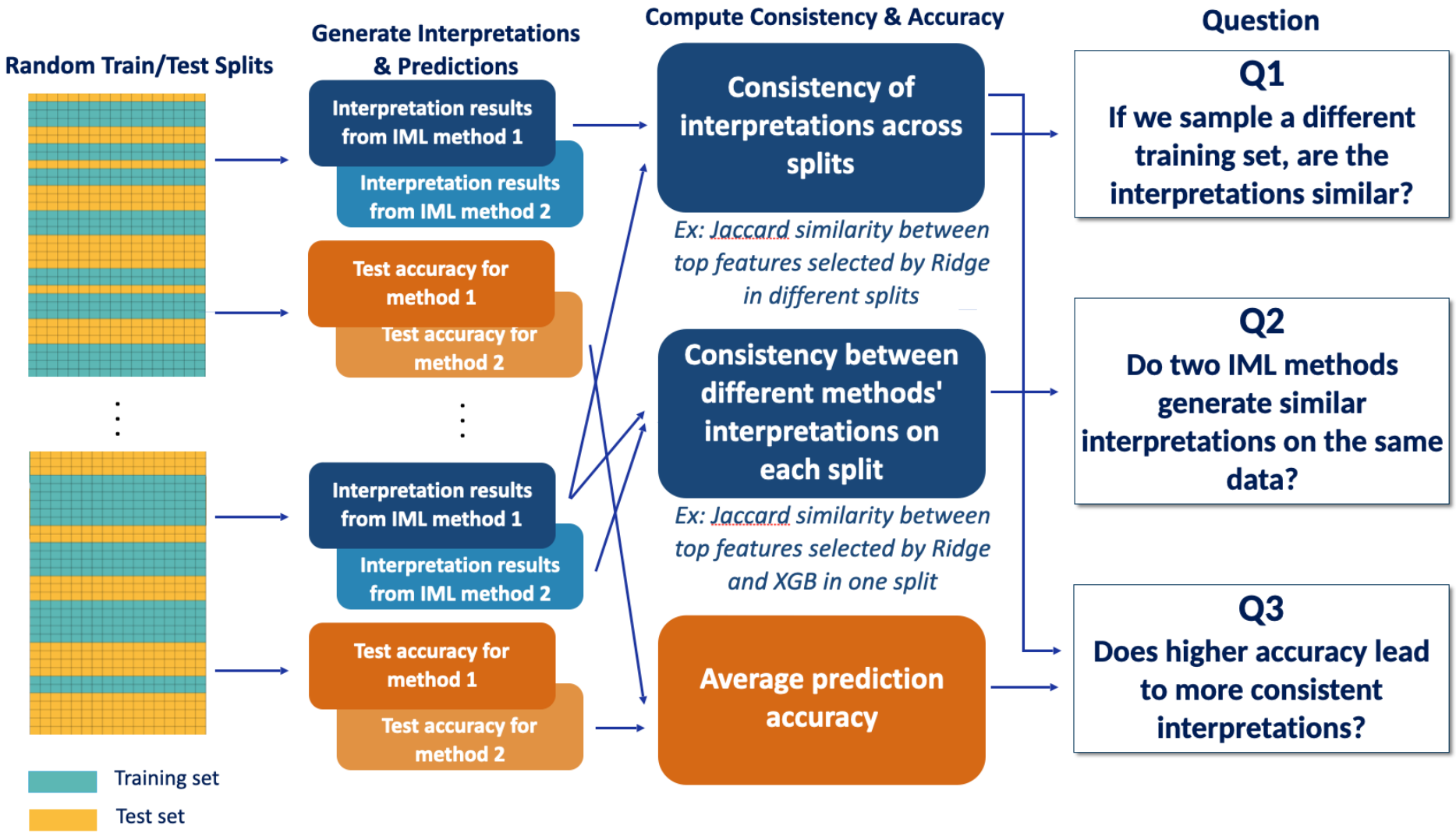
Results Preview: Empirical Reliability Study

Study Design: Approach: Test interpretation reliability through repeated random data perturbations.

- Data: 30+ Commonly used benchmark data sets chosen to vary n, p and other data properties.
- Data Perturbations: 10+ Scenarios like repeated train / test set splits, adding random noise, etc.
- IML Methods: 50+ popular IML methods applied.
- Consistency Metrics: 10+ consistency metrics employed to measure reliability.

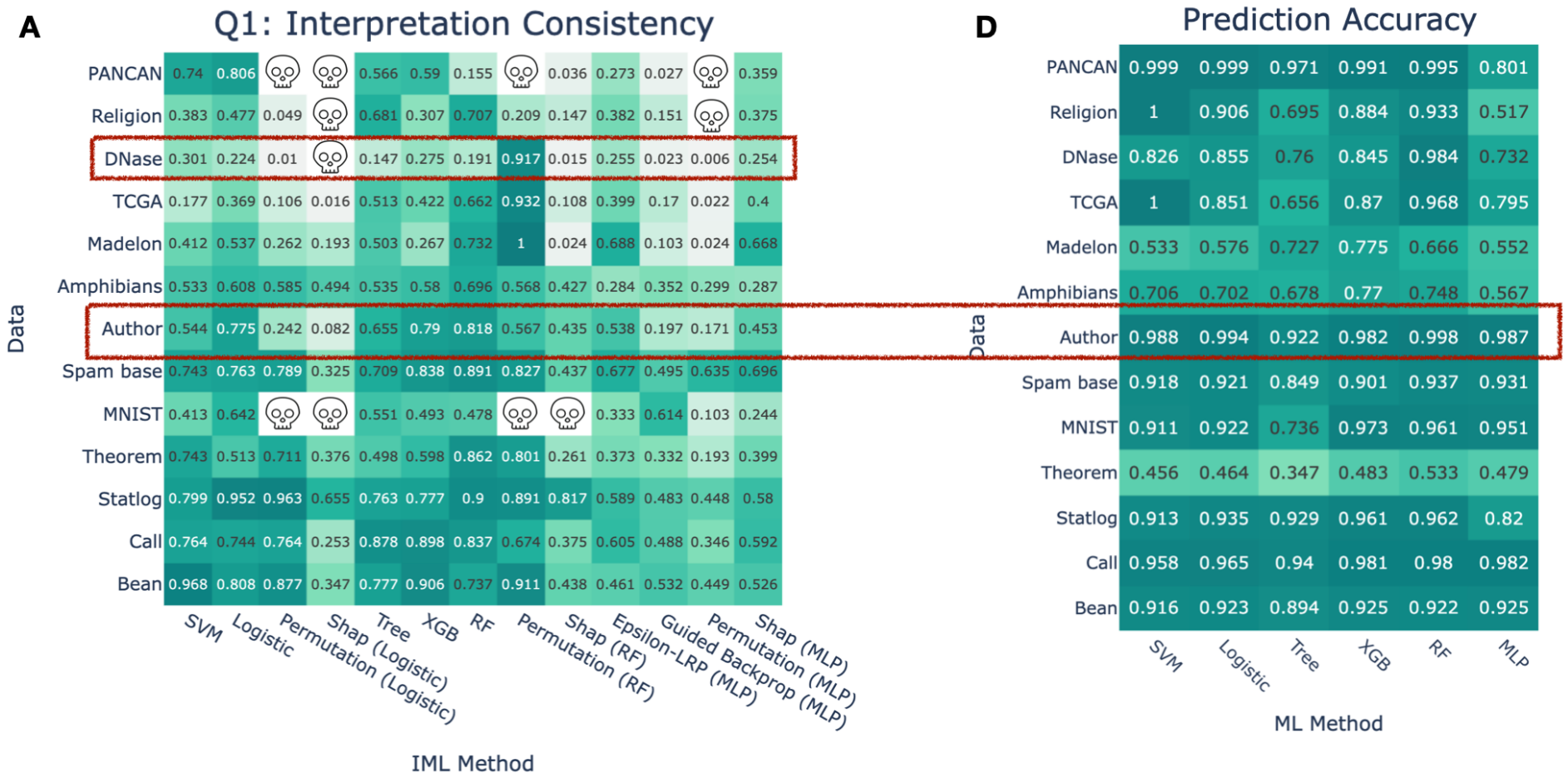
Results Preview: Empirical Reliability Study

Study Design:



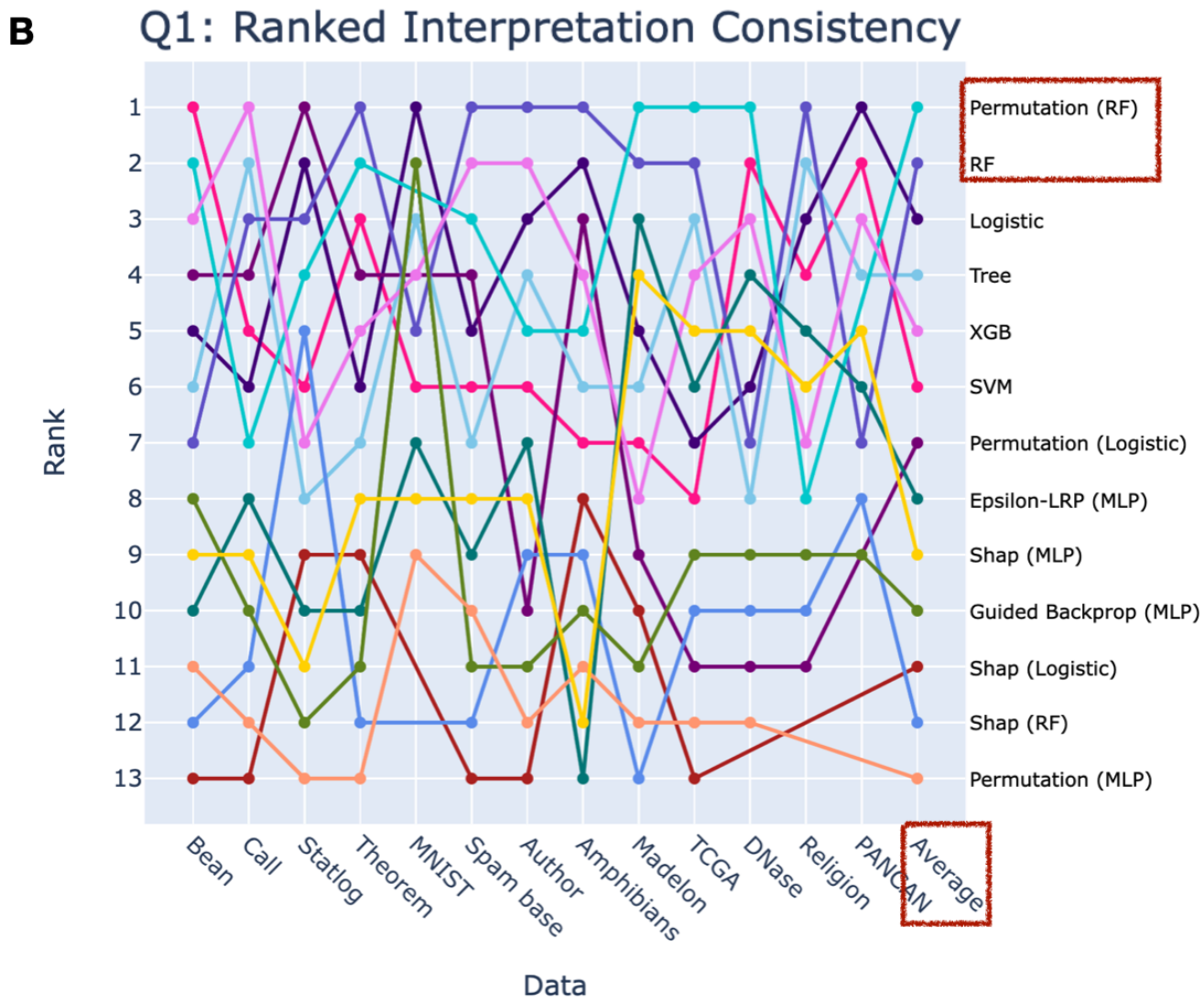
Results Preview: Empirical Reliability Study

Feature Importance for Classification: Internal Consistency



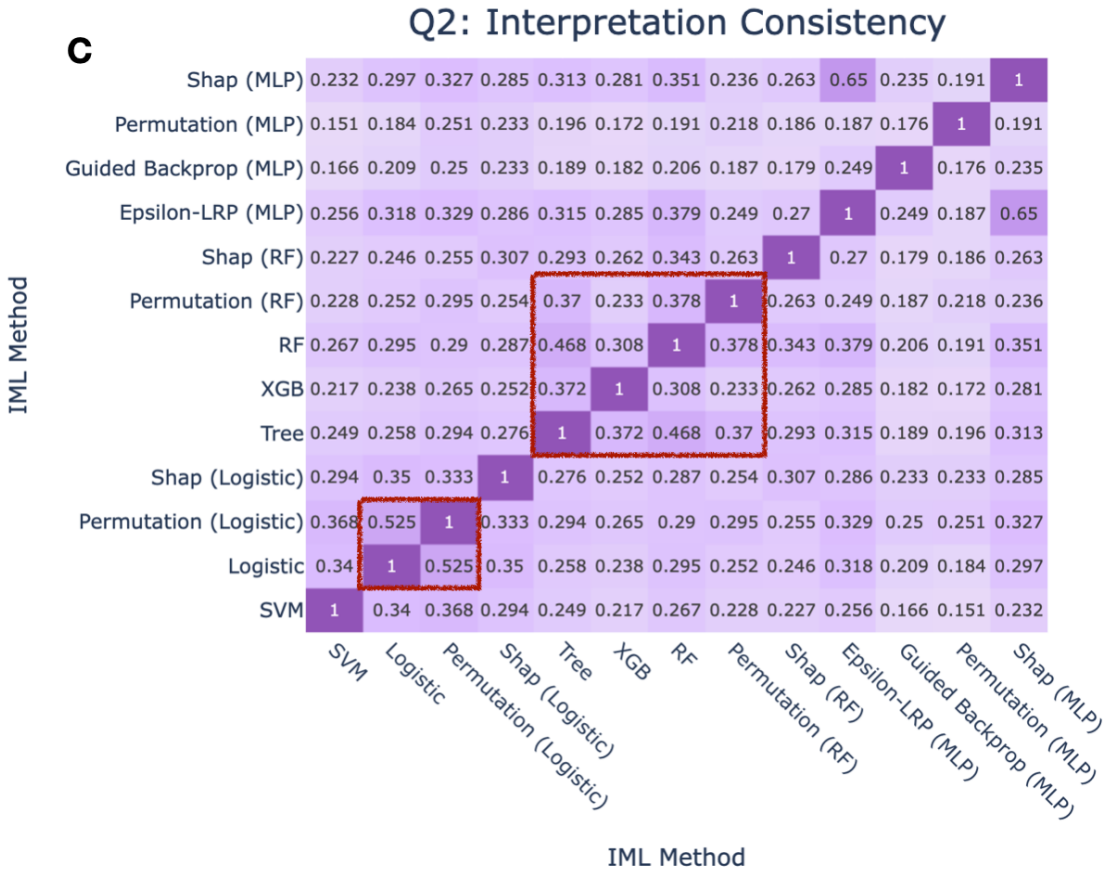
Results Preview: Empirical Reliability Study

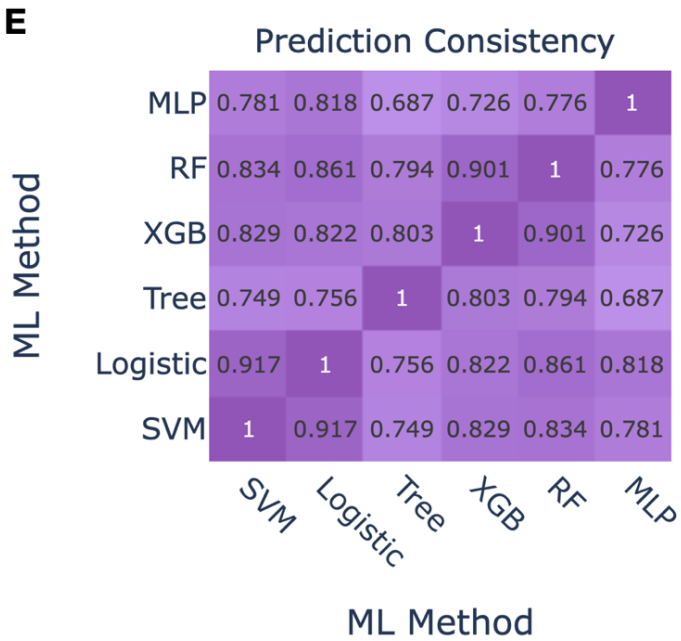
Feature Importance for Classification: Internal Consistency



Results Preview: Empirical Reliability Study

Feature Importance for Classification: Between-Method Consistency





Results Preview: Empirical Reliability Study

Interpret with caution!

What can we do?

Uncertainty Quantification (Statistical Inference) for Feature Importance?

Our Goal

Our Goal

We seek to provide **confidence intervals for feature interpretations** for any learned ML model (**model-agnostic**) and without making any modeling or data assumptions (**distribution-free**).

1. Feature Importance.
2. Higher-Order Feature Interaction Importance.

Focus: Occlusion-based importance for Tabular Data.

1 Motivation

2 Inference for Feature Importance

3 Inference for Higher-Order Feature Interactions

Background: Feature Importance Inference

Distribution-Free & Model-Agnostic Inference for Feature Importance?

Related: Conformal Inference

- Conformal (predictive) inference yields valid distribution-free & model-agnostic confidence intervals for predictions (Shafer and Vovk, 2008).

Related: Post Selection Inference

- Valid inference after model selection for model-specific ML approaches (Taylor and Tibshiarni, 2015).

Background: Feature Importance Inference

ML Feature Importance:

- How does feature j affect $\hat{f}(x)$?
- How does my specific ML model use feature j to make predictions?
- Property of the specific ML model-data interaction.

Population Feature Importance:

- How does feature j affect $f^*(x)$?
- Assumes a population (data-generating) model (f^*).
- Property of data and assumed population model, not a specific ML model.

Background: Feature Importance Inference

ML Feature Importance:

- Inference via Occlusion: LOCO! (Lei et al., 2018; Rinaldo et al., 2019).

Population Feature Importance:

- All of classical statistics!
- Inference in ML context via occlusion: (Shah and Peters, 2020; Zhang and Janson, 2020; Williamson et al., 2022; Williamson and Feng, 2020; Lundberg et al., 2018; and many more).
- Requires strong assumptions on the ML model employed ($\hat{f}$) and/or knowledge of the data distribution or population model (f^*) (Shah and Peters, 2020).

Background: LOCO

Leave-One-Covariate-Out (LOCO) Inference:

- (Lei et al., 2018; Rinaldo et al., 2019)

Idea: Inference on feature occlusion scores based on data-splitting and conformal inference.

Background: LOCO

Leave-One-Covariate-Out (LOCO) Inference:

- (Lei et al., 2018; Rinaldo et al., 2019)

Inference target:

$$\Delta_j^*(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left[\text{Err}\left(Y, \hat{f}^{-j}(X_{-j}; \boldsymbol{X}_{-j}, \boldsymbol{Y})\right) - \text{Err}\left(Y, \hat{f}(X; \boldsymbol{X}, \boldsymbol{Y})\right) | \boldsymbol{X}, \boldsymbol{Y}\right]$$

$\hat{f}(X; \boldsymbol{X}, \boldsymbol{Y})$ denotes the model trained on set $(\boldsymbol{X}, \boldsymbol{Y})$ and evaluated at new point X and Err() is any prediction error metric.

- Difference in expected prediction error with feature j vs. without feature j.
- $\Delta_j^*(\boldsymbol{X}, \boldsymbol{Y})$ is positive when feature j improves predictions.
- Interpretation: How important is feature j when deploying the trained model $\hat{f}$?

Background: LOCO

Leave-One-Covariate-Out (LOCO) Inference:

- (Lei et al., 2018; Rinaldo et al., 2019)

Inference target:

$$\Delta_j^{*,split}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left[\text{Err}\left(Y_2, \hat{f}^{-j}(X_{2,-j}; \boldsymbol{X}_{1,-j}, \boldsymbol{Y}_1)\right) - \text{Err}\left(Y_2, \hat{f}(X_2; \boldsymbol{X}_1, \boldsymbol{Y}_1)\right) | \boldsymbol{X}_1, \boldsymbol{Y}_1\right]$$

Approach: Split data into D_1 and D_2; fit predictive models to D_1, & and use D_2 to evaluate prediction error to construct confidence intervals.

Background: LOCO

Leave-One-Covariate-Out (LOCO) Inference:

- (Lei et al., 2018; Rinaldo et al., 2019)

Advantages:

- Model-agnostic (only studied for regression).
- Agnostic to the form of contribution from feature j.
- Assumption-light & Distribution-free.

Background: LOCO

Leave-One-Covariate-Out (LOCO) Inference:

- (Lei et al., 2018; Rinaldo et al., 2019)

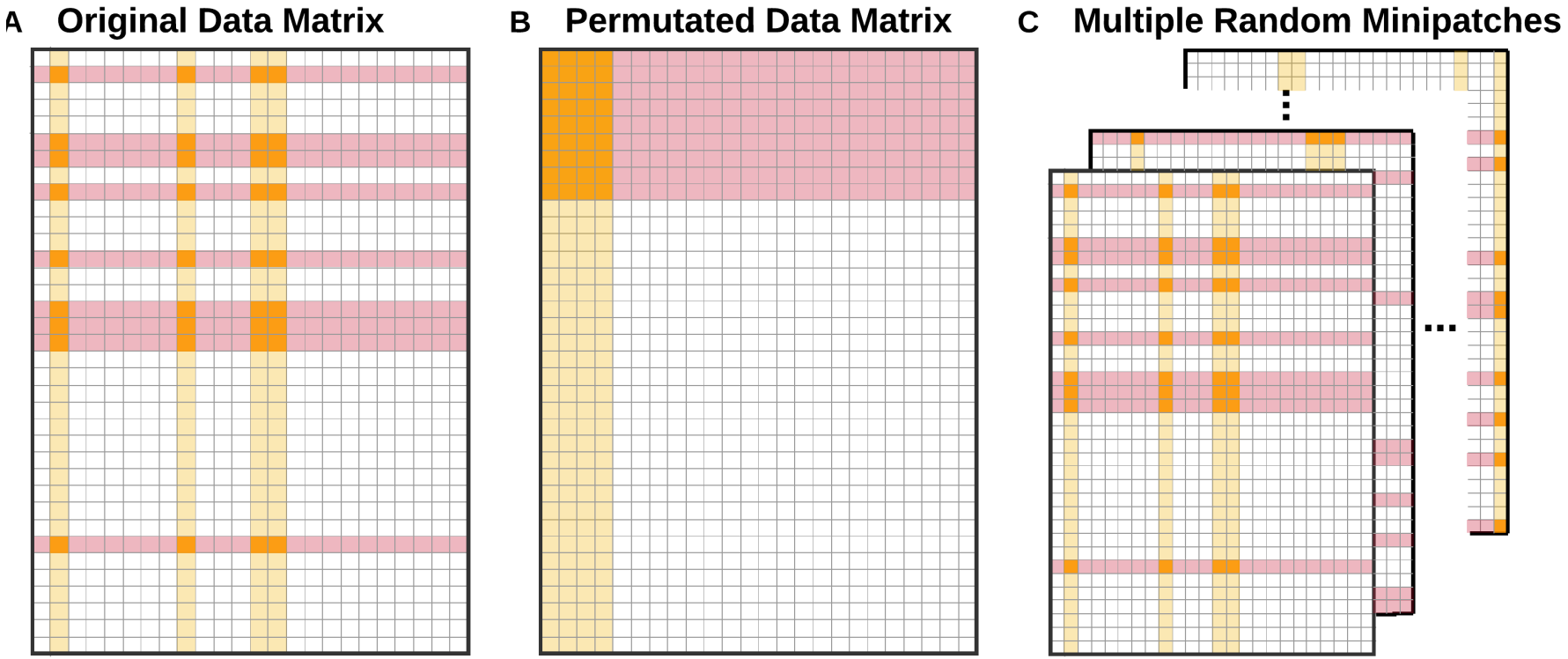
Advantages:

- Model-agnostic (only studied for regression).
- Agnostic to the form of contribution from feature j.
- Assumption-light & Distribution-free.

Disadvantages:

- Data splitting loses statistical power.
- Computationally prohibitive.
- Inference depends on random data split & not all available data.

Minipatch Ensemble Learning



Idea: Minipatches are tiny subsamples of both observations and features.
Double bagging! (Yao & Allen, 2019; Toghani & Allen, 2021)

Minipatch Ensemble Learning

Name:

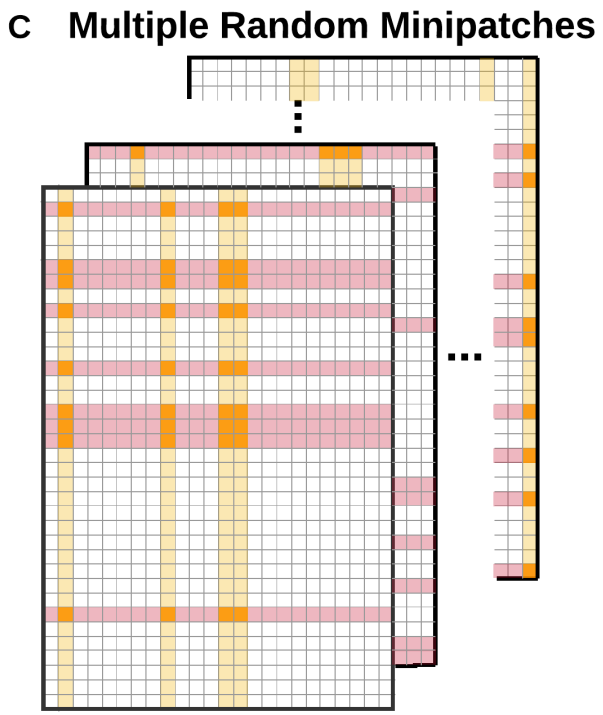
- Minibatches (tiny subsamples of observations).
- Patches (image processing).

Inspiration:

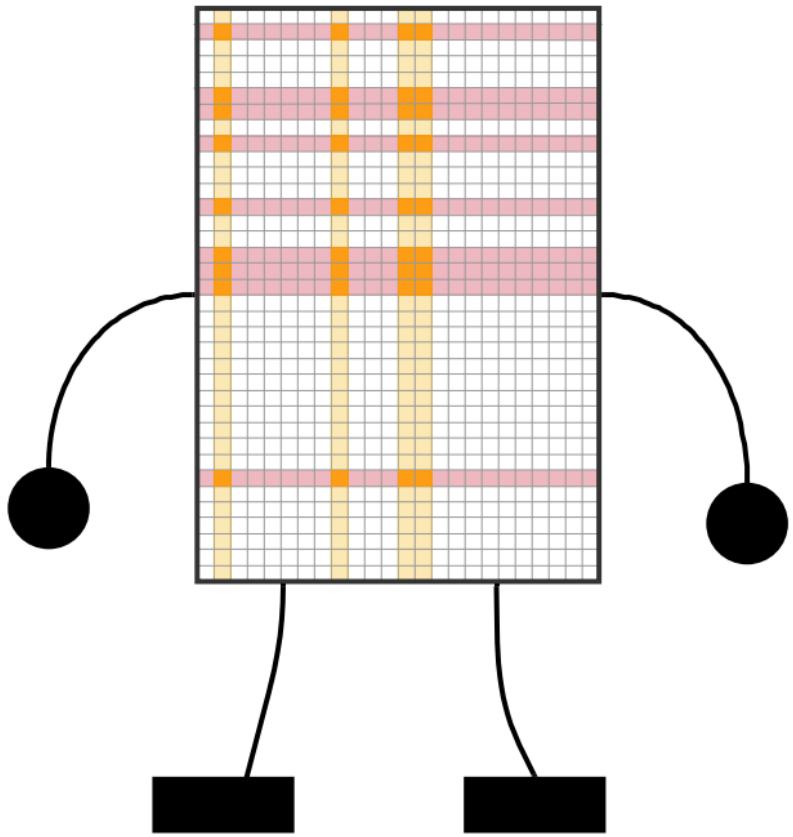
- Random Forests (Louppe and Geurts, 2012).
- Stochastic Optimization & Dropout.
- Stability Selection.

Advantages:

- Fast, distributed & memory efficient.
- Increased ensemble diversity.



Minipatch Ensemble Learning



Can we use minipatches for LOCO inference?

Minipatch LOCO

Step 1:

- Fit minipatch ensemble.

Minipatch LOCO

Step 1:

- Fit minipatch ensemble.

Inference Approach:

- Calculate the **LOO** (leave-one-observation-out) Predictor: $\hat{f}_{-i}(X_i)$.
  - Ensemble minipatches without observation i.
- Calculate the **LOCO-LOO** Predictor: $\hat{f}_{-i}^{-j}(X_i)$.
  - Ensemble minipatches without observation i and without feature j.
- Feature Occlusion Scores:

$$\hat{\Delta}_j(X_i, Y_i) = \text{Error}(Y_i, \hat{f}_{-i}^{-j}(X_i)) - \text{Error}(Y_i, \hat{f}_{-i}(X_i)).$$

- Construct asymptotically normal confidence intervals for Δ_j.

Minipatch LOCO

Step 1:

- Fit minipatch ensemble.

Inference Approach:

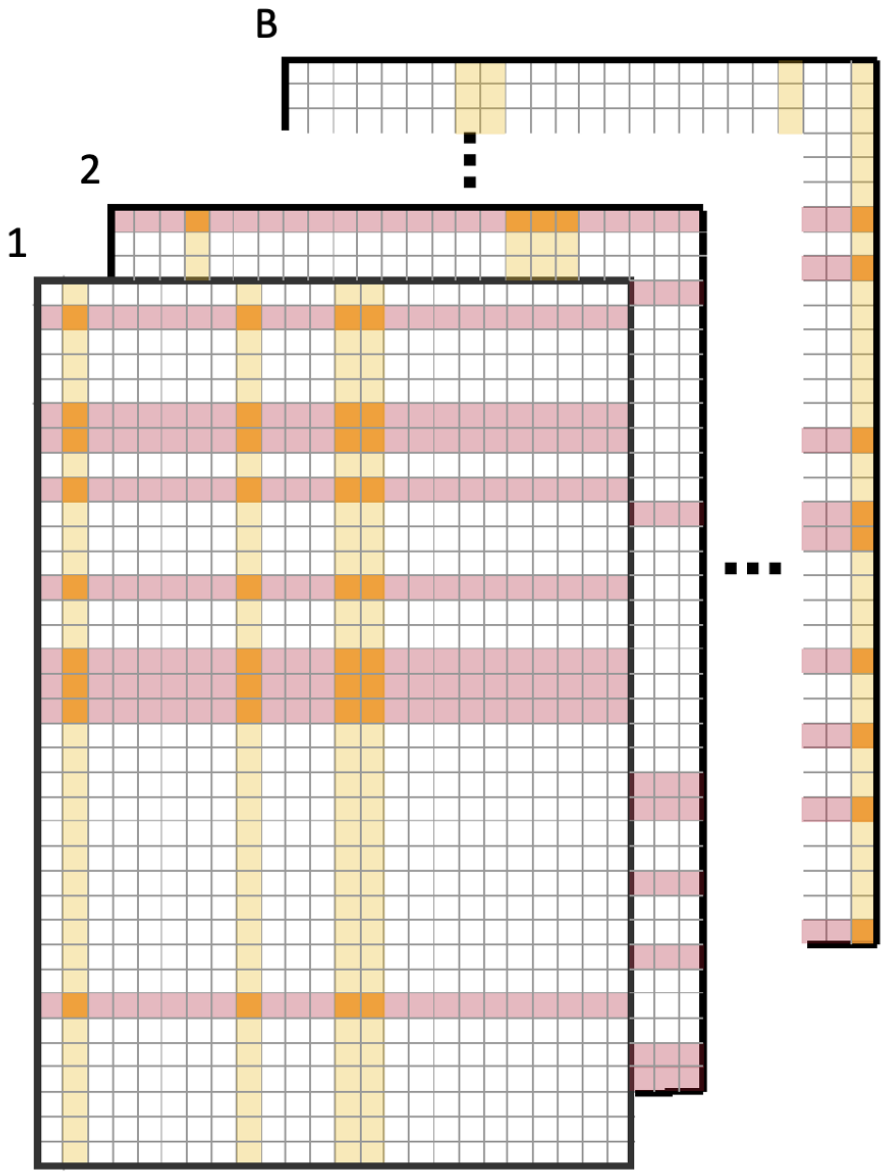
- Calculate the **LOO** (leave-one-observation-out) Predictor: $\hat{f}_{-i}(X_i)$.
  - Ensemble minipatches without observation i.
- Calculate the **LOCO-LOO** Predictor: $\hat{f}_{-i}^{-j}(X_i)$.
  - Ensemble minipatches without observation i and without feature j.
- Feature Occlusion Scores:

$$\hat{\Delta}_j(X_i, Y_i) = \text{Error}(Y_i, \hat{f}_{-i}^{-j}(X_i)) - \text{Error}(Y_i, \hat{f}_{-i}(X_i)).$$

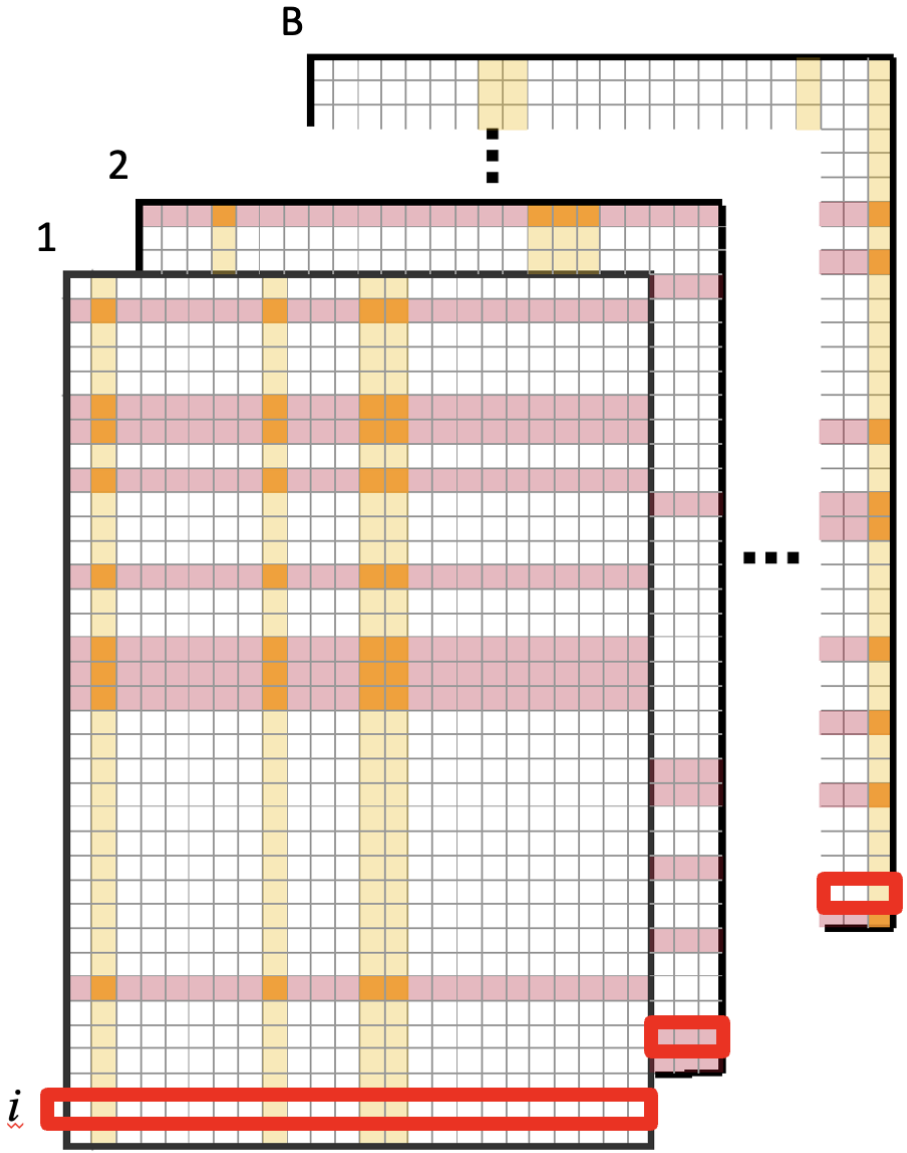
- Construct asymptotically normal confidence intervals for Δ_j.

Free computationally! (After Step 1)

Minipatch LOCO

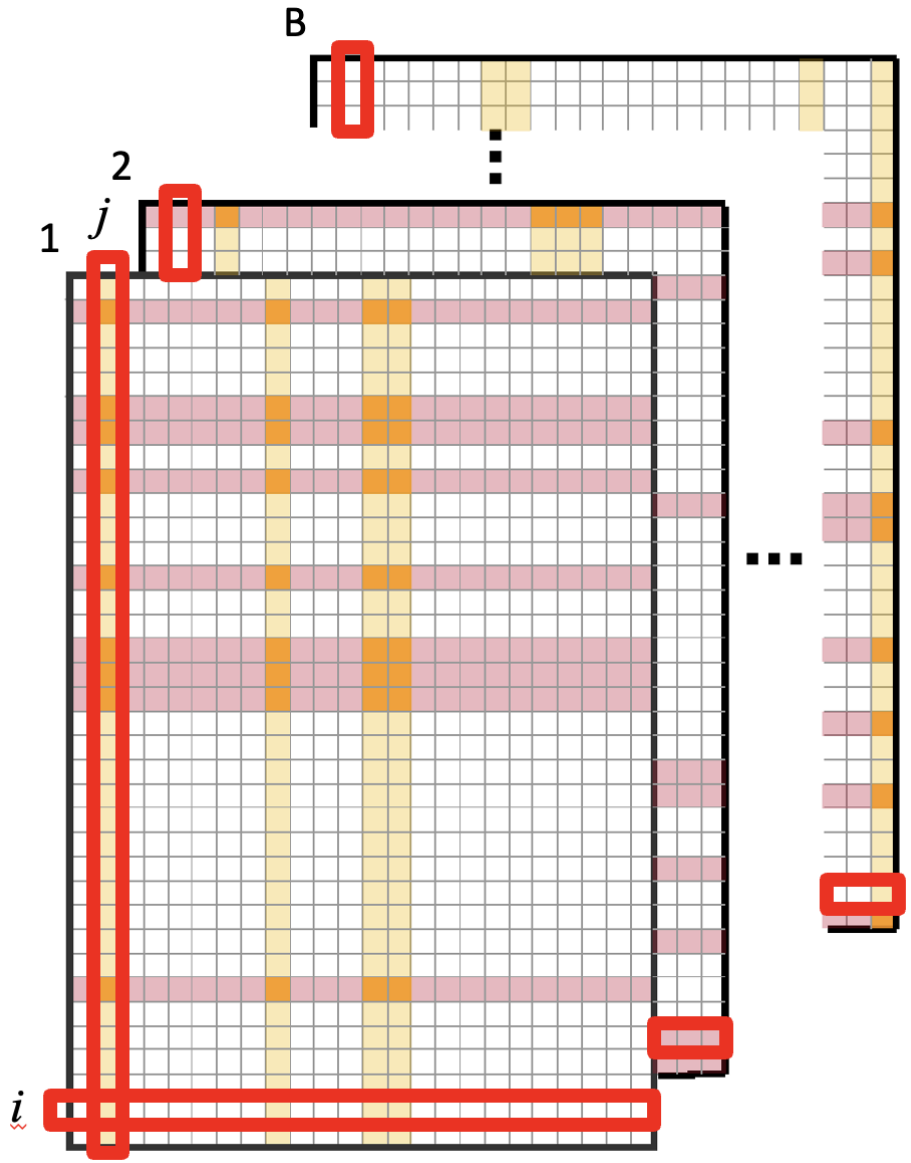


Minipatch LOCO



$$\hat{f}_{-i} = \hat{f}^{(1)} + \ldots + \hat{f}^{(B)}$$

Minipatch LOCO



$$\hat{f}_{-i} = \hat{f}^{(1)} + \ldots + \hat{f}^{(B)}$$

$$\hat{f}_{-i}^{-j} = \ldots + \hat{f}^{(B)}$$

Minipatch LOCO

Minipatch Feature Occlusion Scores:

$$\hat{\Delta}_j(X_i, Y_i) = \text{Error}(Y_i, \hat{f}_{-i}^{-j}(X_i)) - \text{Error}(Y_i, \hat{f}_{-i}(X_i)).$$

$1 - \alpha$ **Confidence Interval for** Δ_j^*:

$$\hat{\mathbb{C}}_j = \left[\bar{\Delta}_j - \frac{z_{\alpha/2}\hat{\sigma}_j}{\sqrt{N}}, \bar{\Delta}_j + \frac{z_{\alpha/2}\hat{\sigma}_j}{\sqrt{N}} \right]$$

where $\bar{\Delta}_j = \frac{1}{N}\sum_{i=1}^n \Delta_j(X_i, Y_i)$ is the sample mean and $\hat{\sigma}_j = \sqrt{\frac{\sum_{i=1}^n (\Delta_j(X_i, Y_i) - \bar{\Delta}_j)^2}{N-1}}$ is the sample standard deviation.

Minipatch LOCO Validity

Do Minipatch LOCO confidence intervals have valid coverage?

Challenges:

1. $\hat{\Delta}_j(X_i, Y_i)$'s are dependent.
   - Each pair of $\hat{\Delta}$'s depends on same $N-2$ samples.
   - Existing dependent Central Limit Theorems do not apply with this extreme dependency.
2. Same data is used for training & inference!
   - New type of selective inference problem.

Minipatch LOCO Validity

A1. $\mathrm{Error}()$ satisfies a Lipschitz condition with constant L.
- Trivially / easily satisfied.

A2. Bounded minipatch predictions: $||\hat{f}_{I,F}(X) - \hat{f}_{I',F'}(X)|| \leq B$.
- Automatically satisfied for classification; satisfied under light assumptions for regression.

A3. Minipatch size: $n = o(\frac{\sigma_j}{LB}\sqrt{N})$.

A4. Minipatch number: $K \gg (\frac{L^2B^2N}{\sigma_j^2} + \frac{LB\sqrt{N}}{\sigma_j} + 1)\log(N)$.

Minipatch LOCO Validity

Theorem

Under assumptions A1-A4,

$$\sqrt{N}\hat{\sigma}_j^{-1}(\bar{\Delta}_j - \Delta_j^*) \xrightarrow{d} \mathcal{N}(0,1).$$

Corollary

$$\lim_{N\to\infty} \mathbb{P}(\Delta_j^* \in \hat{\mathbb{C}}_j) = 1 - \alpha.$$

Minipatch LOCO confidence intervals have valid asymptotic coverage!

Minipatch LOCO Validity

Proof Sketch:

- Recent paper (Bayle et al., 2020) shows CLT for cross-validation error, if training algorithm is stable.
- Minipatch LOO estimates similar to leave-one-out cross-validation error.
- Minipatch ensembles are stable with any base model!
- Characterize conditions for MP stability via bounded differences, MP number, MP size.

Minipatch LOCO Validity

Algorithmic Stability = Algorithm Robustness?

Algorithmic Stability [Elisseeff et al., 2005; Kim & Barber, 2023]

$$P\left(|\hat{f}_n(X_{n+1}) - \hat{f}_{n-1}(X_{n+1})| > \epsilon\right) \leq \delta$$

- Predictions from algorithm are robust to removal of an individual data point.
- Algorithmic Stability important for:
  - Generalizability, conformal inference, selective inference, and more.
- Most statistical / ML algorithms are NOT stable.
  - Ex: (Generalized) linear models, decision trees, neural networks, etc.

- Minipatches & Ensembling make ANY base model stable!

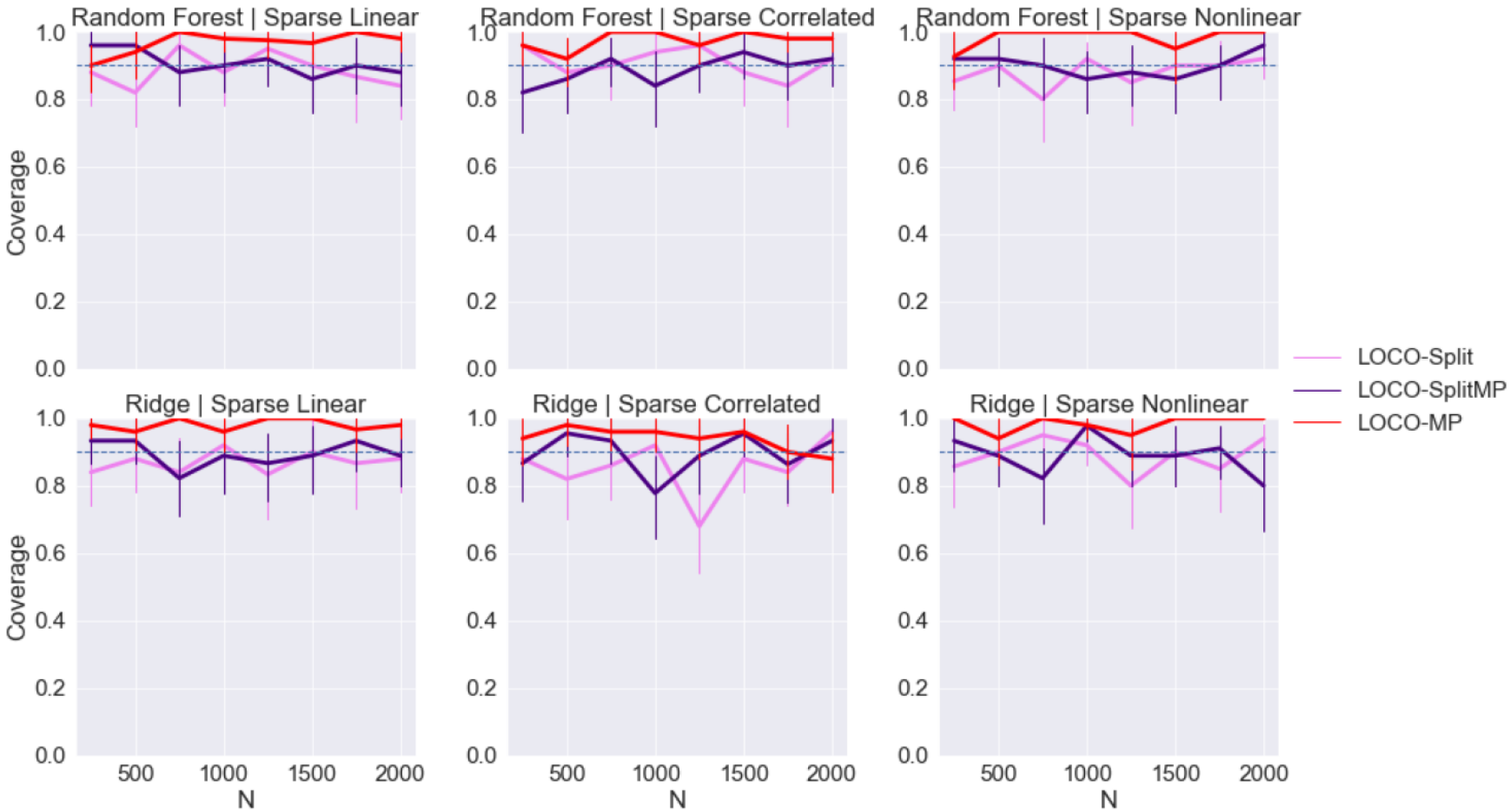
Minipatch LOCO Validity

Other Theoretical Extensions & Contributions:

- Variance barrier solving vanishing variance problem.
- Alternative assumptions that permit minipatches of any size.
- Valid coverage after hyperparameter tuning (minipatch size).
- Analysis of connection to population feature importance and conditional independence testing under certain (linear) models.
- **Relation to post selection inference**.
- Analysis and discussion of minipatch LOCO inference with correlated features.

Simulation Studies

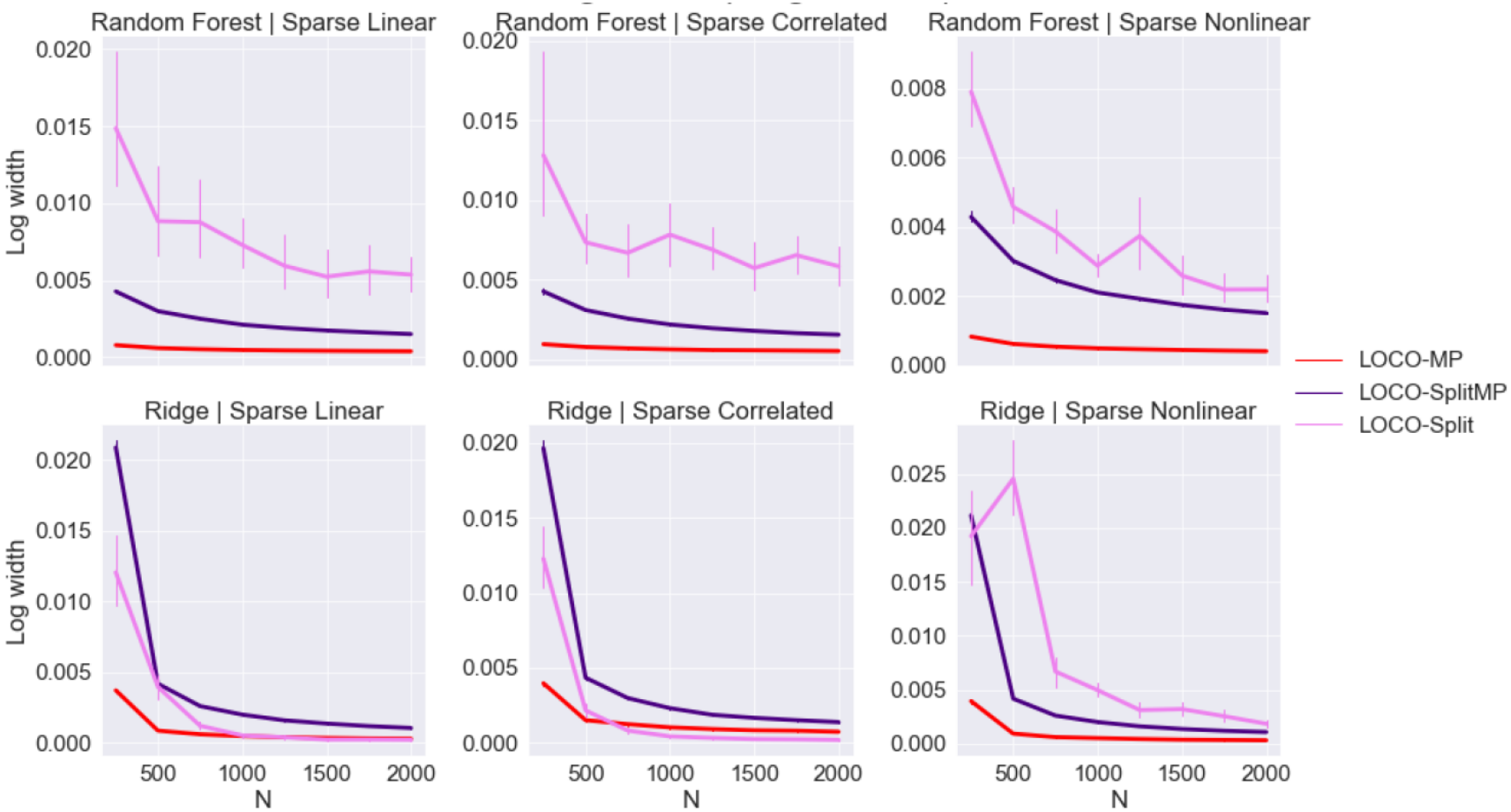
Theory Validation: Coverage.



Coverage for regression simulations for a null feature; $M = 200$, varying N; LOCO-MP with $m = \sqrt{M}$, $n = \sqrt{N}$, $K = 10,000$.

Simulation Studies

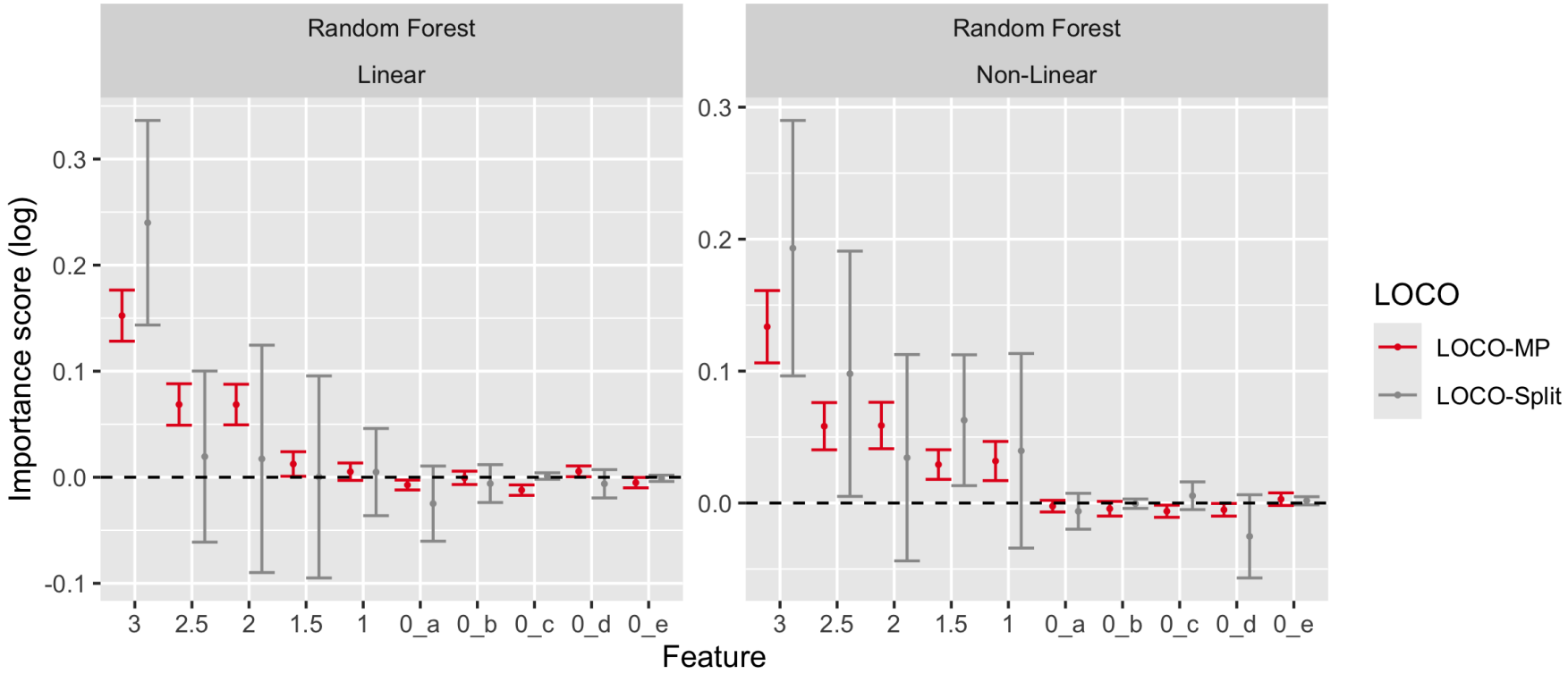
Interval Width:



Log interval width for regression simulations for a null feature; $M = 200$, varying N; LOCO-MP with $m = \sqrt{M}$, $n = \sqrt{N}$, $K = 10,000$.

Simulation Studies

Confidence Intervals:



Regression simulation $M = 200$ with 5 signal features and $N = 500$; LOCO-MP with $m = .5$, $n = .5$, $K = 1,000$.

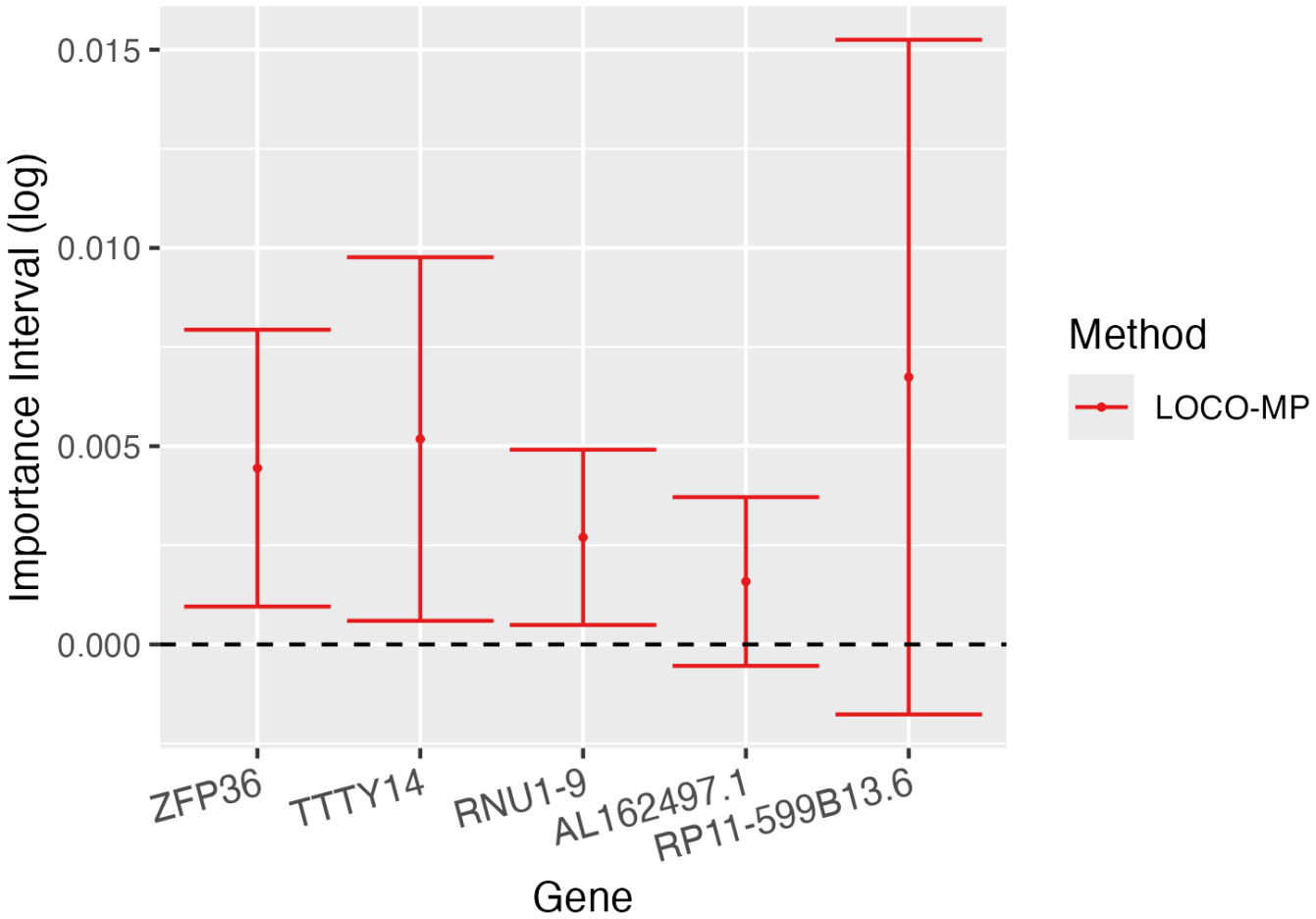
Results Preview: Genomics of Cognitive Decline

ROSMAP Study: Longitudinal clinical-pathological cohort study of aging and Alzheimer's Disease.

- Data: $n = 507$ subjects with clinical, imaging and post-mortem genomics data.
- Task: Predict a measure of global cognition proximate to death based on post-mortem gene expression (bulk RNA-seq).
- Model: Random Forest.
- Gene filtering: Most variable genes ($p = 86$).

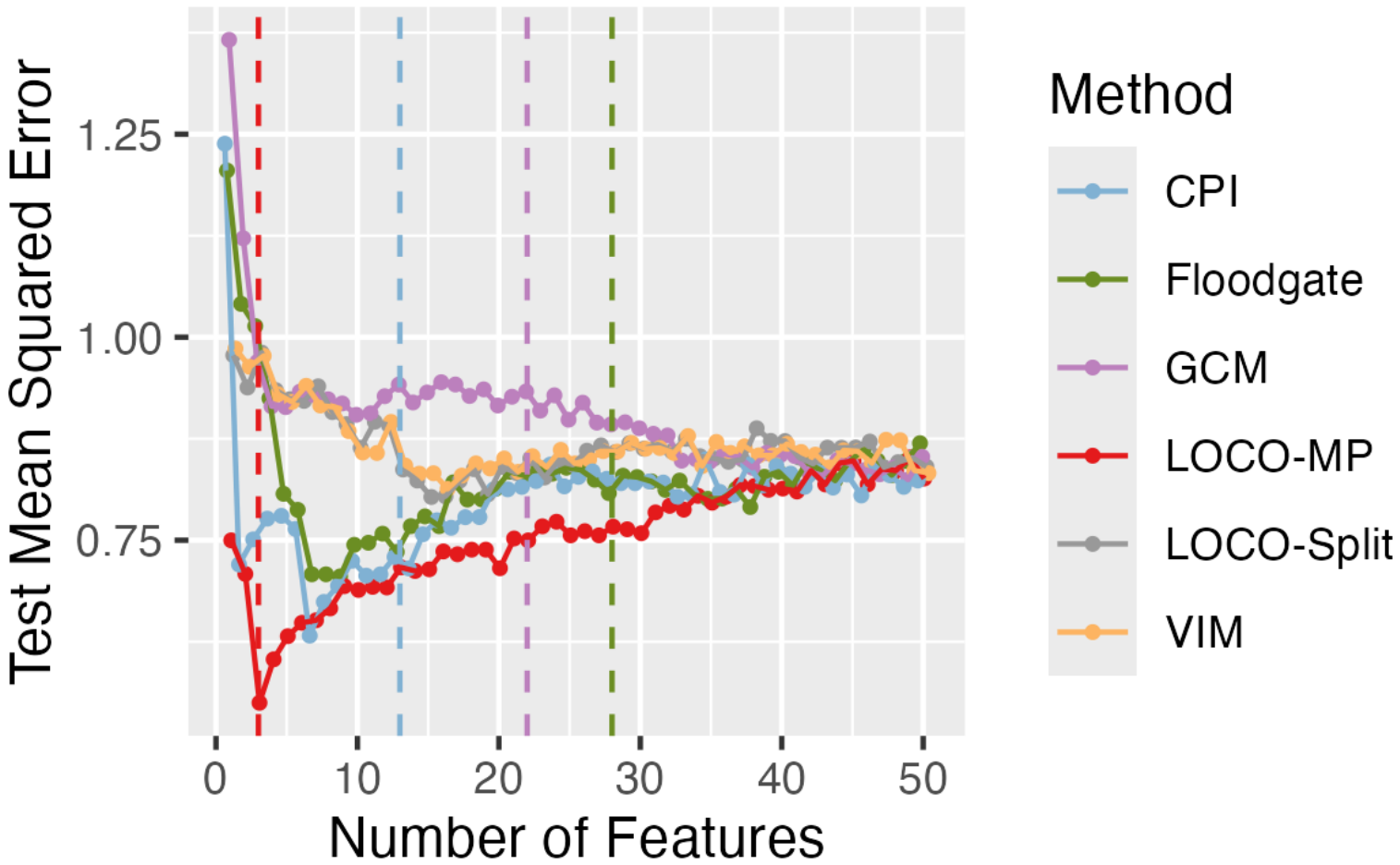
Results Preview: Genomics of Cognitive Decline

Our Results:



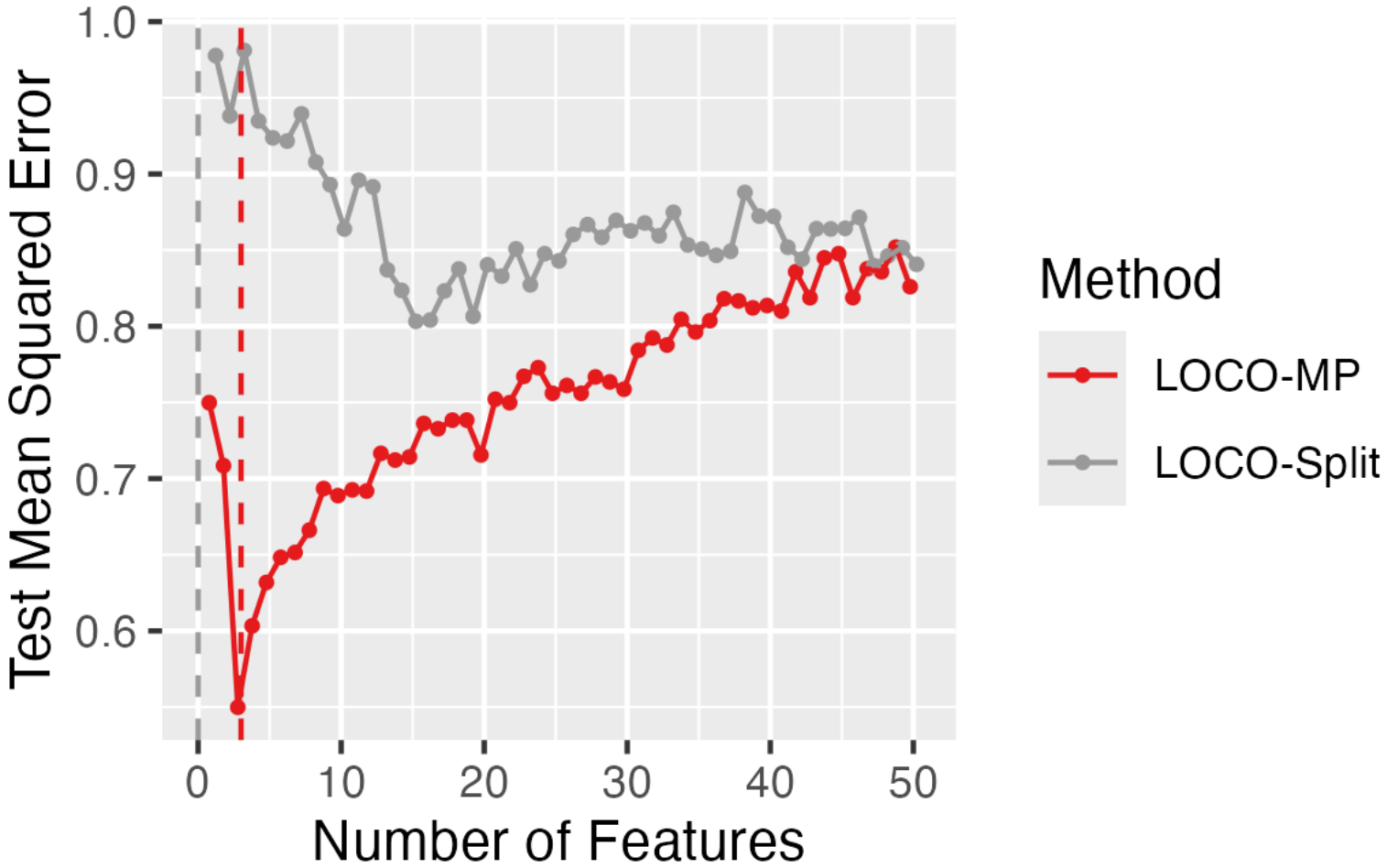
Results Preview: Genomics of Cognitive Decline

Comparative Results (Ablation):



Results Preview: Genomics of Cognitive Decline

Comparative Results (Ablation):



Results Preview: Genomics of Cognitive Decline

Genes Discovered:

- ZFP36: Up-regulated in AD patients (Guan et al., 2022)
- TTTY14: Y-Chromosome up-regulated in male microglia in AD (Wang et al., 2024).
- RNU1-9: long non-coding RNA (lncRNA) with no known links to AD; several have recently linked lncRNAs to neurodegenerative diseases (Huang et al.,2024; Black et al., 2024; Mosquera-Heredia et al., 2024).
  - New target to study!

Feature Interaction Importance

Goal

Can we find important (higher-order) feature interactions?
And, can we quantify their uncertainty?

Why Feature Interactions?

- "Omics" (Genomics, proteomics, chemometrics, and etc.)
- Materials science.
- Drug discovery.
- Clinical diagnostics.
- Advertising.

Feature Interaction Importance

Goal

Can we find important (higher-order) feature interactions?
And, can we quantify their uncertainty?

Existing Approaches:

- H-Statistic *(Friedman and Popescu, 2008)*.
  - Computationally Burdensome.
- Shapley Interaction Scores *(Ribeiro et al., 2016; Sundararajan et al., 2020; Tsai et al., 2023)*.
  - Computationally Burdensome.
- Interaction Forests *(Basu et al., 2018)*.
  - Only for Random Forests.

No statistical inference!

iLOCO: Interaction LOCO

Leave-Two-Out:

$$\Delta^*_{j,k}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left[\text{Err}\left(Y, \hat{f}^{-(j,k)}(X_{-(j,k)})\right) - \text{Err}\left(Y, \hat{f}(X)\right) \mid \boldsymbol{X}, \boldsymbol{Y}\right]$$

- $\Delta^*_{j,k}$ positive when leaving features j and k out increases the prediction error.

iLOCO: Interaction LOCO

Leave-Two-Out:

$$\Delta^*_{j,k}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left[\text{Err}\left(Y, \hat{f}^{-(j,k)}(X_{-(j,k)})\right) - \text{Err}\left(Y, \hat{f}(X)\right) \mid \boldsymbol{X}, \boldsymbol{Y}\right]$$

- $\Delta^*_{j,k}$ positive when leaving features j and k out increases the prediction error.

Definition: iLOCO

$$\text{iLOCO}^*_{j,k} = \Delta^*_j + \Delta^*_k - \Delta^*_{j,k}.$$

iLOCO: Interaction LOCO

Definition: iLOCO

$$\text{iLOCO}^*_{j,k} = \Delta^*_j + \Delta^*_k - \Delta^*_{j,k}.$$

$$\text{iLOCO}^*_{j,k} = \mathbb{E}\Bigg[\left(\text{Err}\left(Y, \hat{f}^{-j}\right) + \text{Err}\left(Y, \hat{f}^{-k}\right) - \text{Err}\left(Y, \hat{f}^{-(j,k)}\right) \right) - \text{Err}\left(Y, \hat{f}\right) \mid \boldsymbol{X}, \boldsymbol{Y} \Bigg]$$

- **Theorem:** If we assume a functional ANOVA decomposition, $\text{iLOCO}^*_{j,k}$ is equal to the contribution of all terms containing the j, k pairwise interaction.

iLOCO: Interaction LOCO

Definition: iLOCO

$$\text{iLOCO}^*_{j,k} = \Delta^*_j + \Delta^*_k - \Delta^*_{j,k}.$$

- Positive values of $\text{iLOCO}^*_{j,k}$ indicate an important interaction.
- Negative values of $\text{iLOCO}^*_{j,k}$ indicate two important features that are correlated, but don't interact.
  - Existing feature importance metrics don't work well for correlated features (*Verdinelli and Wasserman, 2024*).

iLOCO: Interaction LOCO

Definition: Higher-Order iLOCO

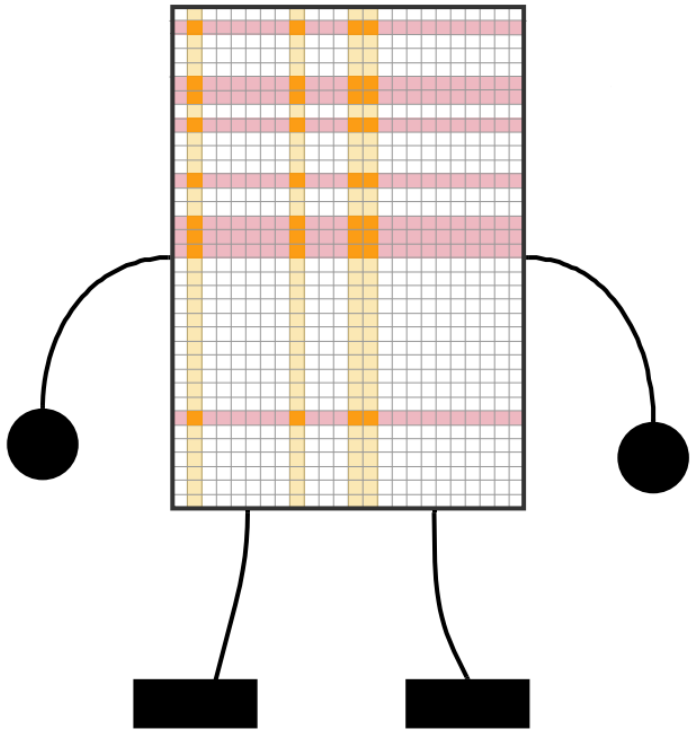
$$\text{iLOCO}_S^* = \sum_{T \subseteq S} (-1)^{|S|-|T|} \Delta_T^*.$$

where S is a subset of features.

- First higher-order feature interaction metric.

iLOCO: Interaction LOCO

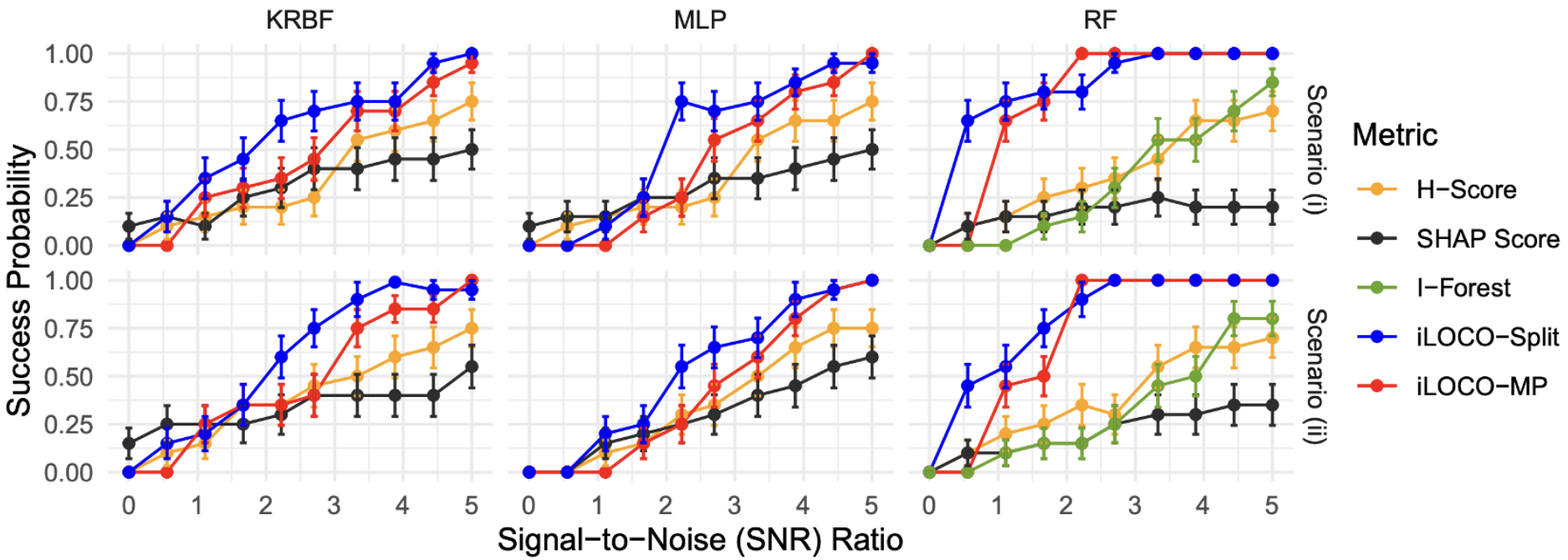
iLOCO Estimation & Inference:



- Fast computation (distributed).
- Model-agnostic, distribution-free, & assumption-light statistical inference.

Simulation Results Preview

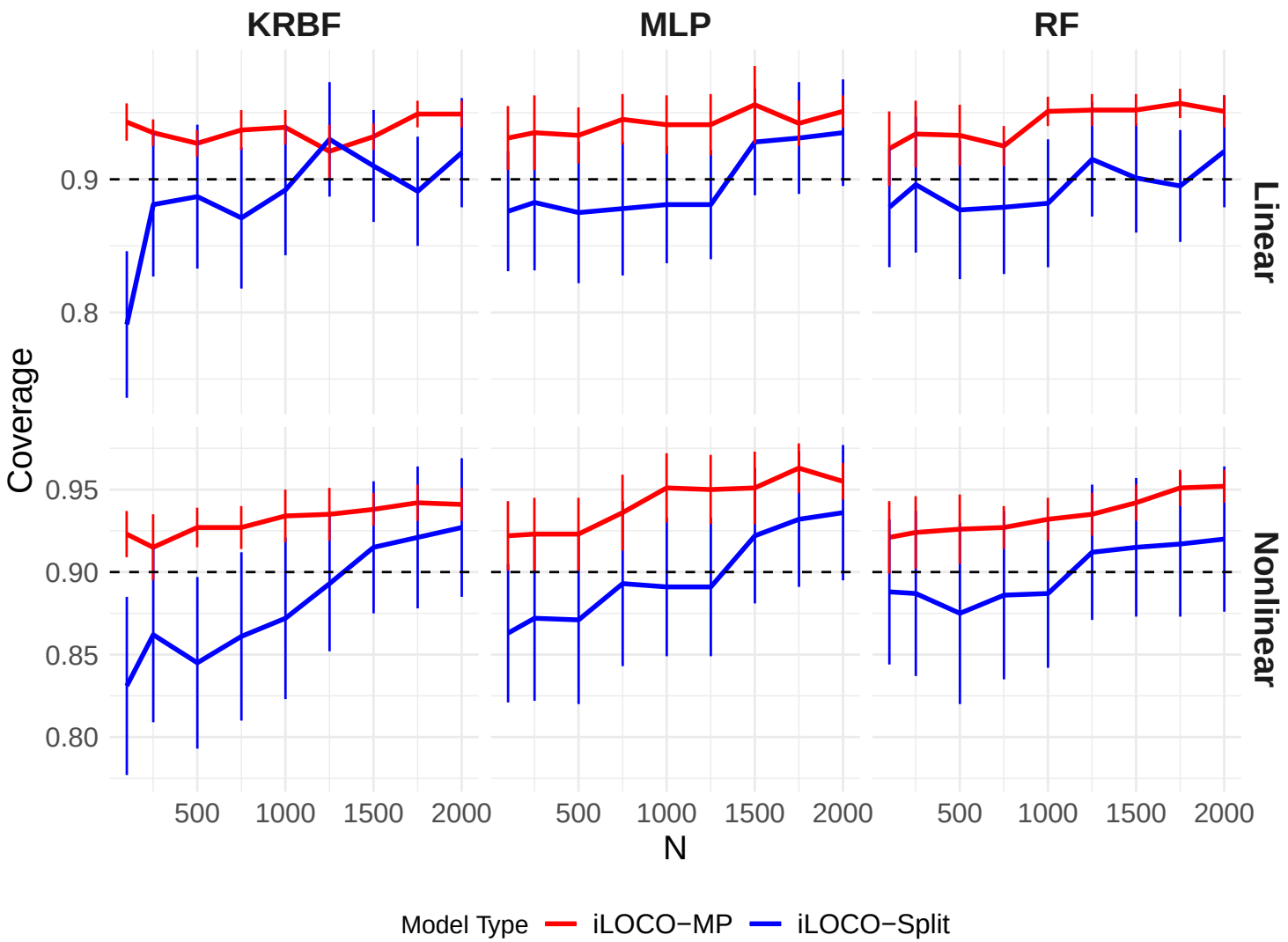
Interaction Detection:



Computational Timing: With only $M = 20$ features, iLOCO-MP took ≈20 seconds vs. >24 hours for competitors to run.

Simulation Results Preview

Confidence Interval Coverage:



Results: Cocktail Ingredients

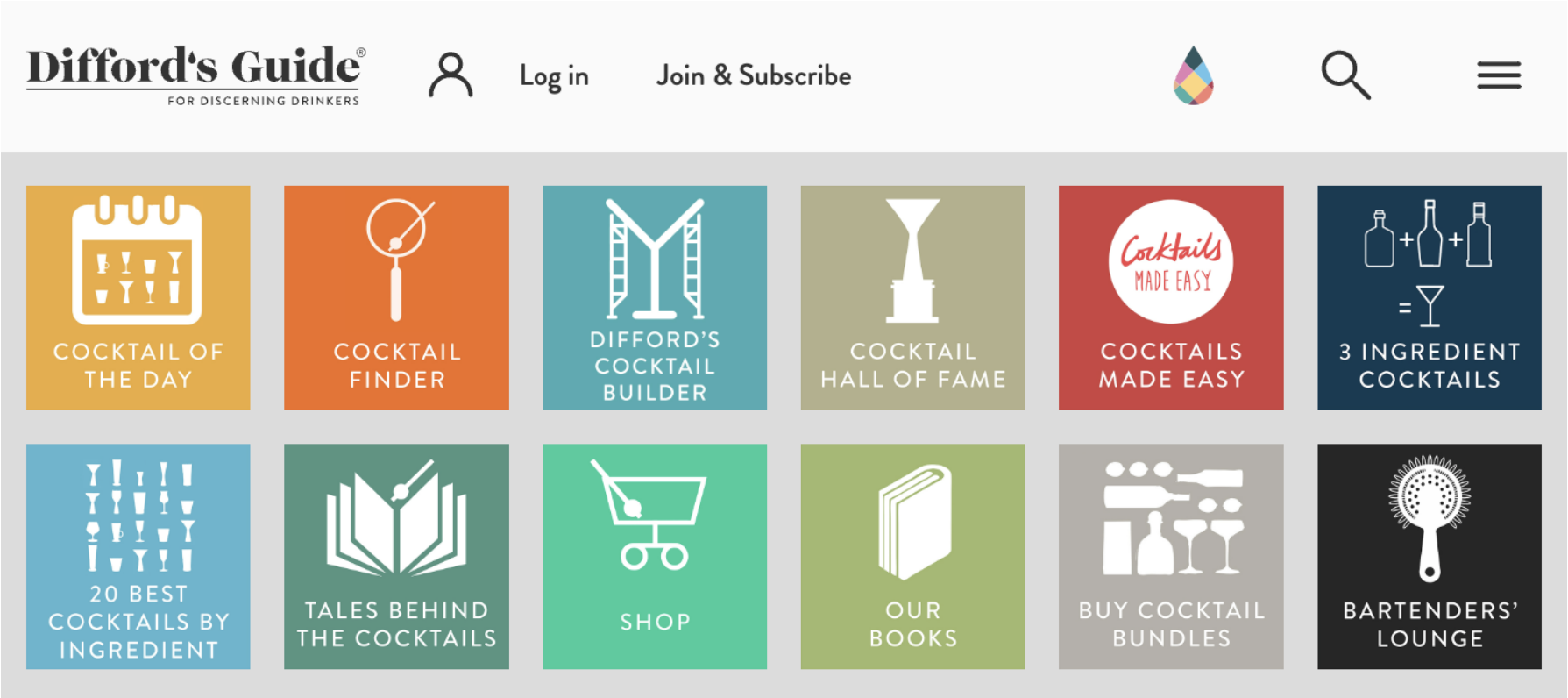
Question

Which cocktail ingredients or combinations of cocktail ingredients yield the best drinks?

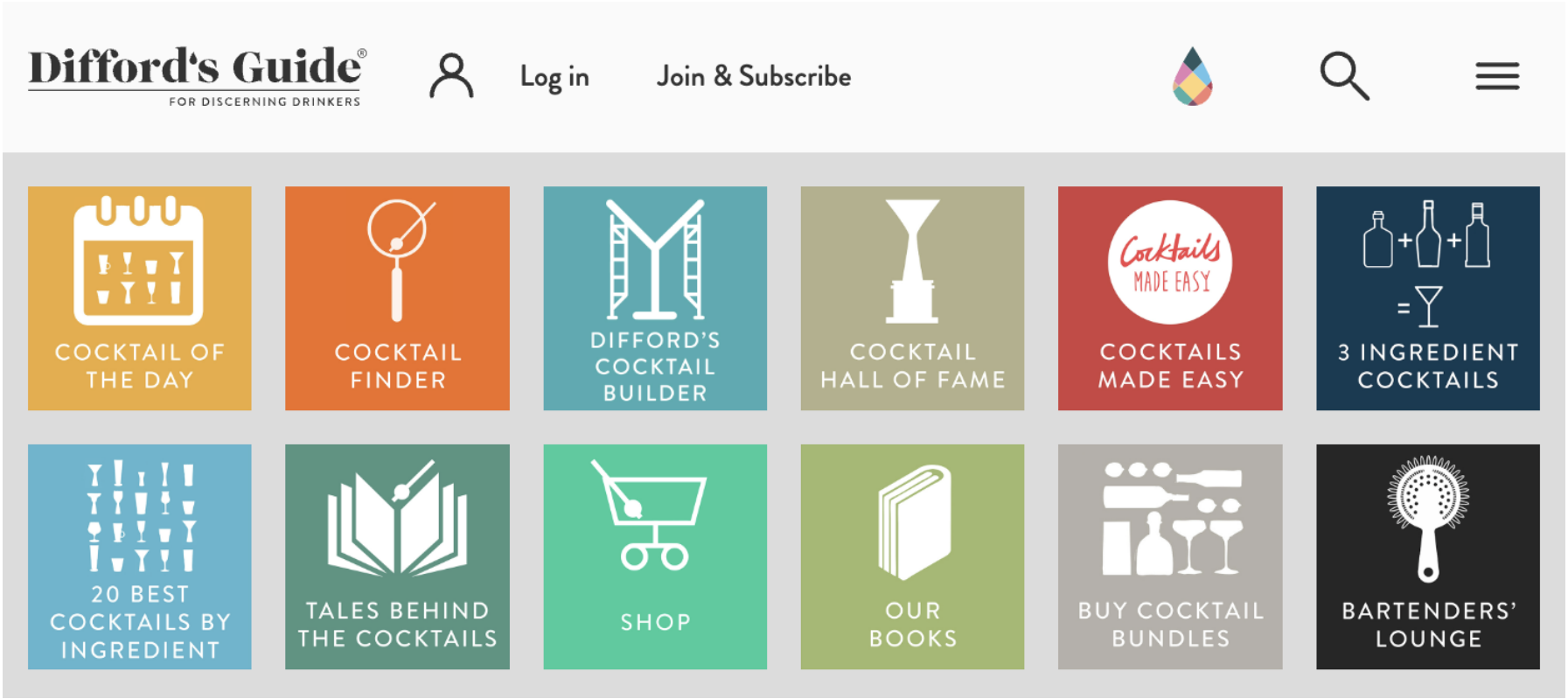
Results: Cocktail Ingredients

Question

Which cocktail ingredients or combinations of cocktail ingredients yield the best drinks?



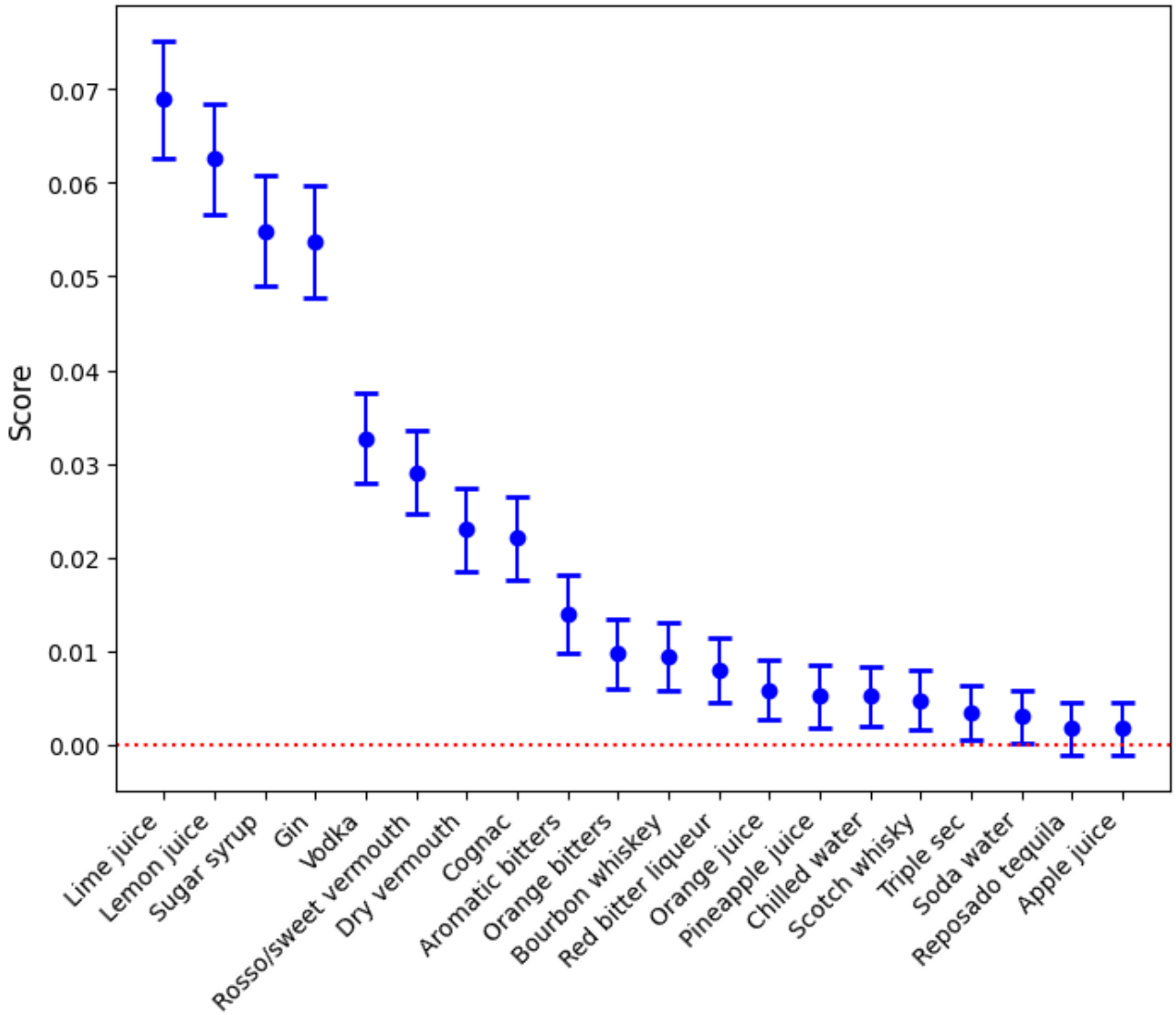
Results: Cocktail Ingredients



- Web-scraped **Difford's Guide** website.
- $N = 15975$ cocktails and studied the top $M = 100$ cocktail ingredients.
- Goal: Use the ingredients to predict the *Difforfd's Guide* Rating.
- Model: Minipatch ensemble of deep neural networks.

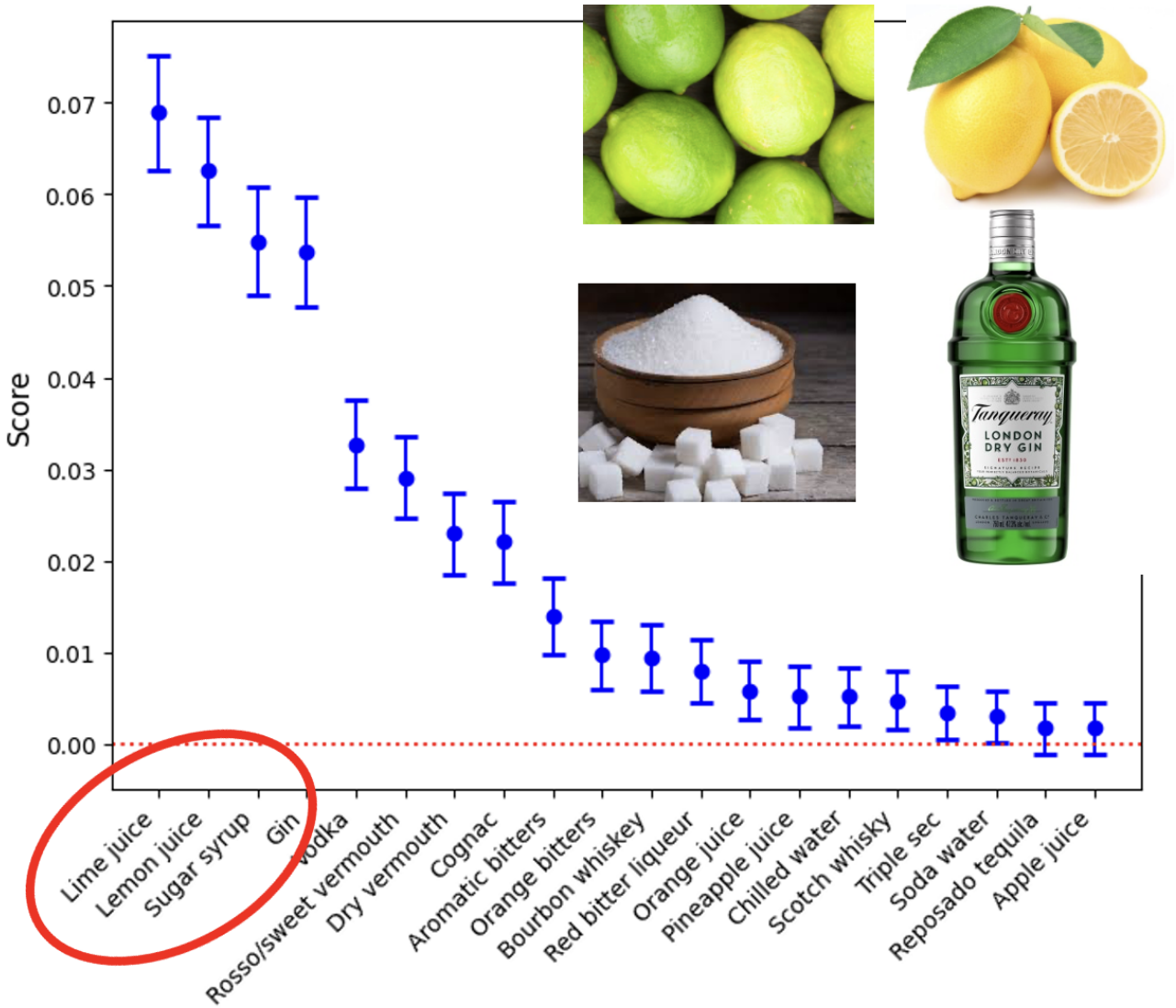
Results: Cocktail Ingredients

Inference for Feature Importance (LOCO-MP) Results:



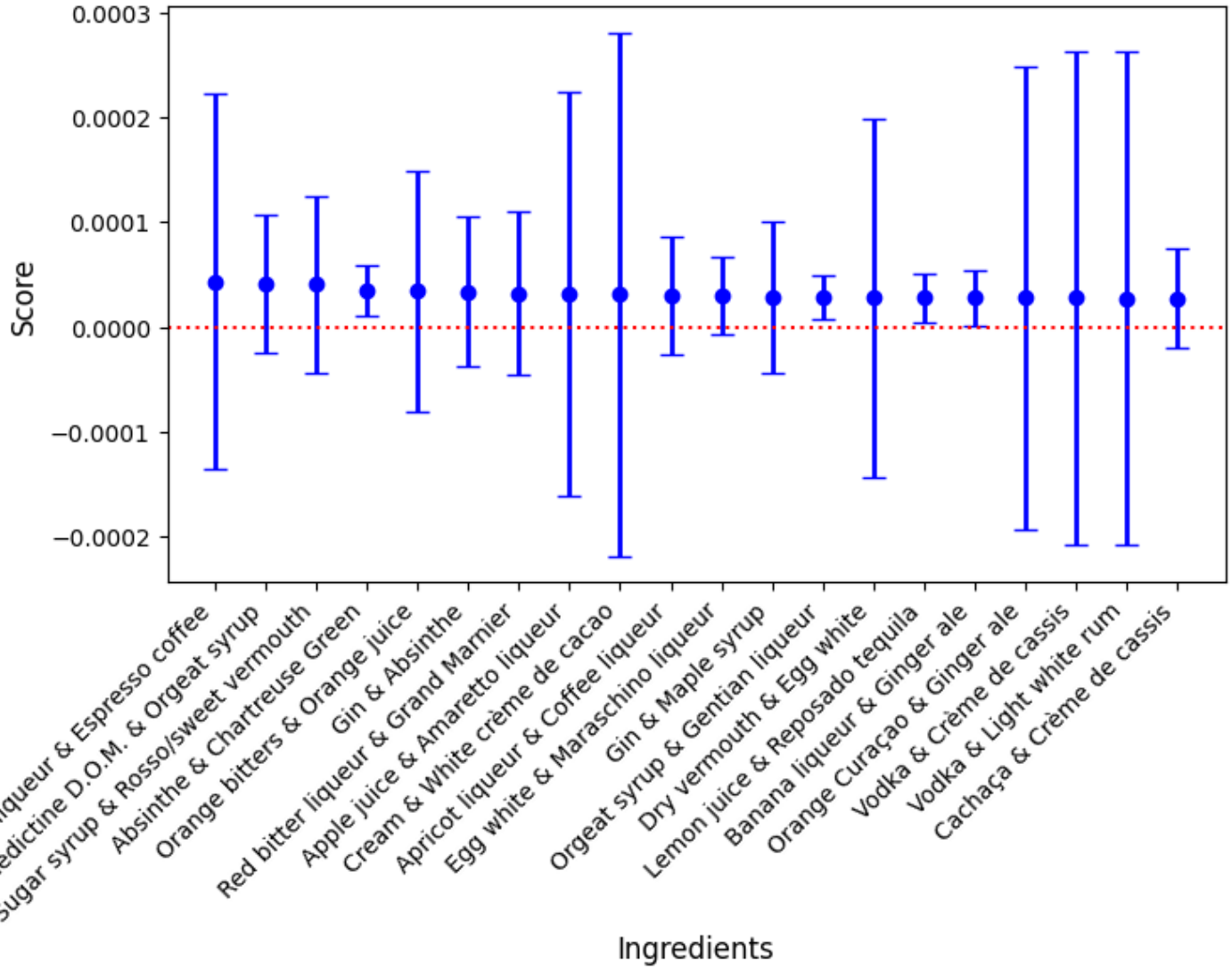
Results: Cocktail Ingredients

Inference for Feature Importance (LOCO-MP) Results:



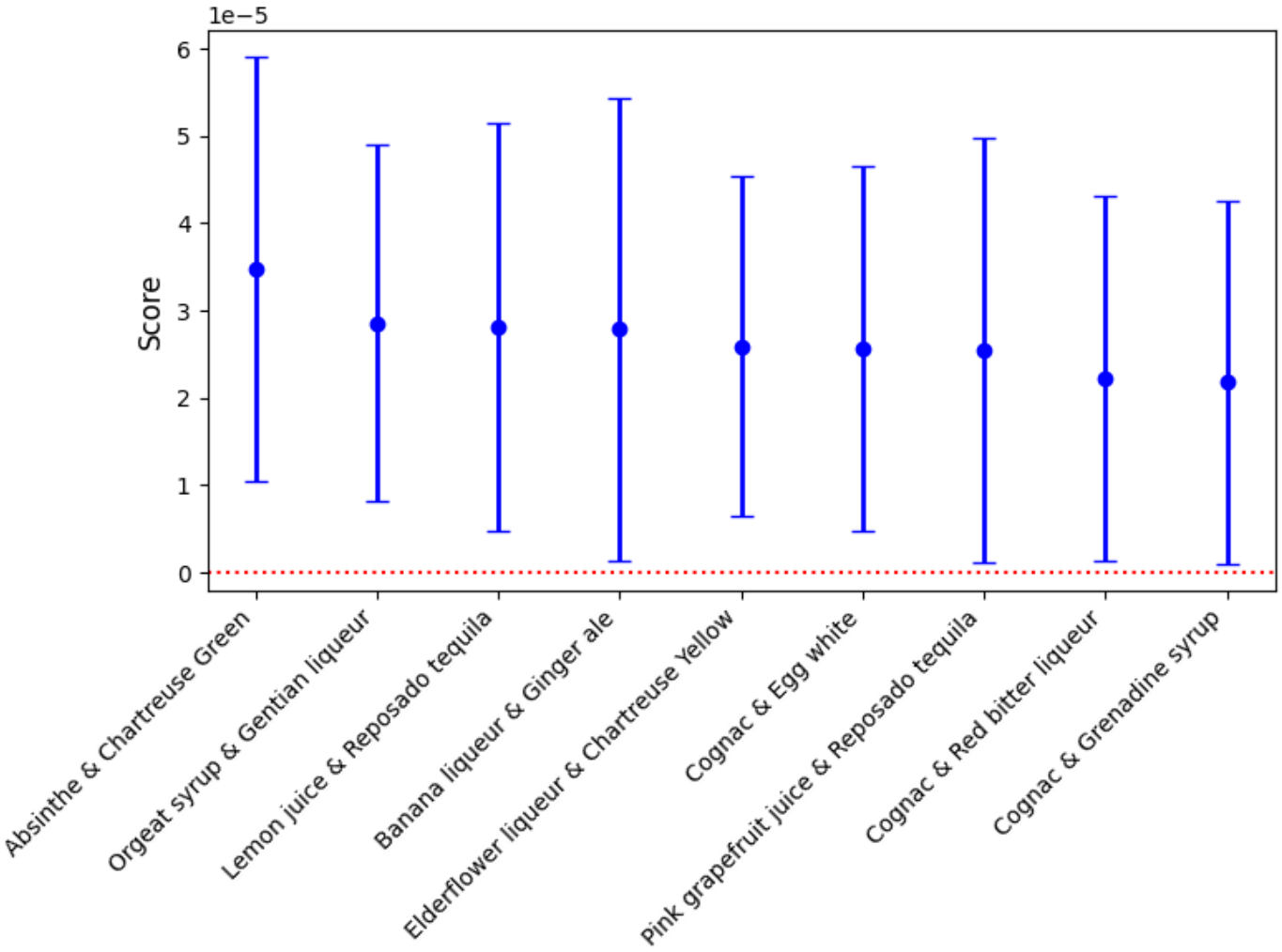
Results: Cocktail Ingredients

Inference for Feature Interactions (iLOCO-MP) Results:



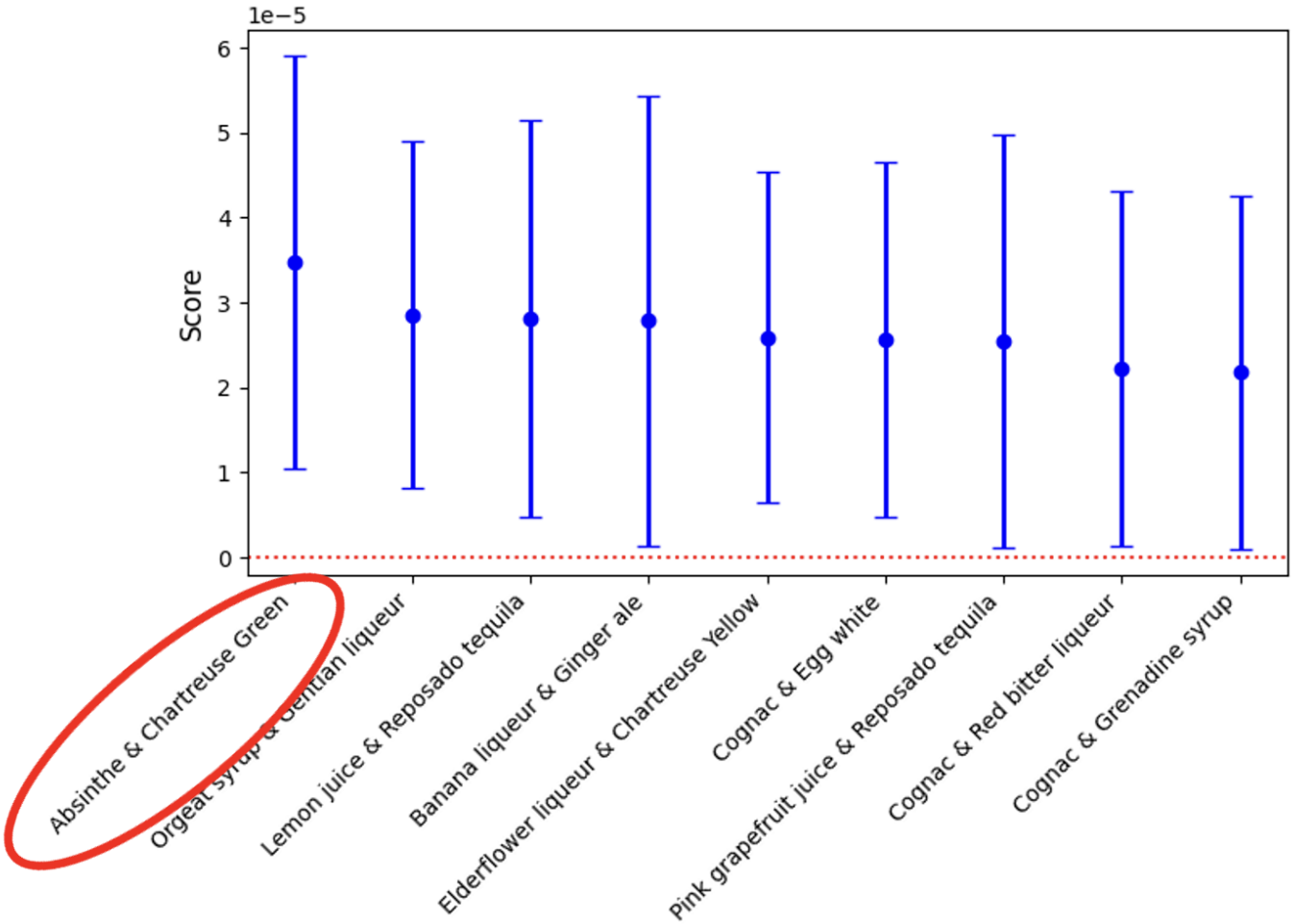
Results: Cocktail Ingredients

Inference for Feature Interactions (iLOCO-MP) Results:



Results: Cocktail Ingredients

Inference for Feature Interactions (iLOCO-MP) Results:



Results: Cocktail Ingredients

Inference for Feature Interactions (iLOCO-MP) Results:

Hearn Cocktail

★★★★½ / ★★★★☆

- Irish whiskey
- Rosso/sweet vermouth
- **Chartreuse Green**
- Orinoco bitters
- **Absinthe**
- Orange bitters

Parmenides

★★★★½ / ★★★★½

- Cognac (brandy)
- **Chartreuse Green**
- Chardonnay wine
- Lemon juice
- Sugar syrup (2:1)
- **Absinthe**

Neptune's Wrath

★★★★½ / ★★★★½

- Gin
- **Absinthe**
- Lemon juice
- Sugar syrup (2:1)
- Egg white (pasteurised)
- **Chartreuse Green**

Grapevyne

★★★★½ / ★★★☆☆

- Green grapes (seedless)
- Vodka
- Ambrato vermouth
- Grappa
- **Chartreuse Green**
- **Absinthe**

Results: Cocktail Ingredients

Inference for Feature Interactions (iLOCO-MP) Results:

Churchillian

★★★★½ / ★★★★☆

- **Absinthe**
- Cognac (brandy)
- Scotch whisky
- Cynar/carciofo amaro
- **Chartreuse Green**
- Maraschino liqueur

St. Patrick's Day

★★★★½ / ★★★½☆

- Irish whiskey
- **Chartreuse Green**
- Green crème de menthe
- Lime juice
- Sugar syrup (2:1)
- **Absinthe**
- Soda (club soda) water

Red Thorn

★★★★½ / ★★★★☆

- Raspberries
- Irish whiskey
- Dry vermouth
- Raspberry syrup
- Lemon juice
- **Chartreuse Green**
- **Absinthe**

The Burly French Man

★★★★☆ / ★★★☆☆

- **Absinthe**
- Cognac (brandy)
- Rosso/sweet vermouth
- **Chartreuse Green**
- Damson liqueur
- Orange bitters
- Aromatic bitters

Results: Cocktail Ingredients

Inference for Feature Interactions (iLOCO-MP) Results:

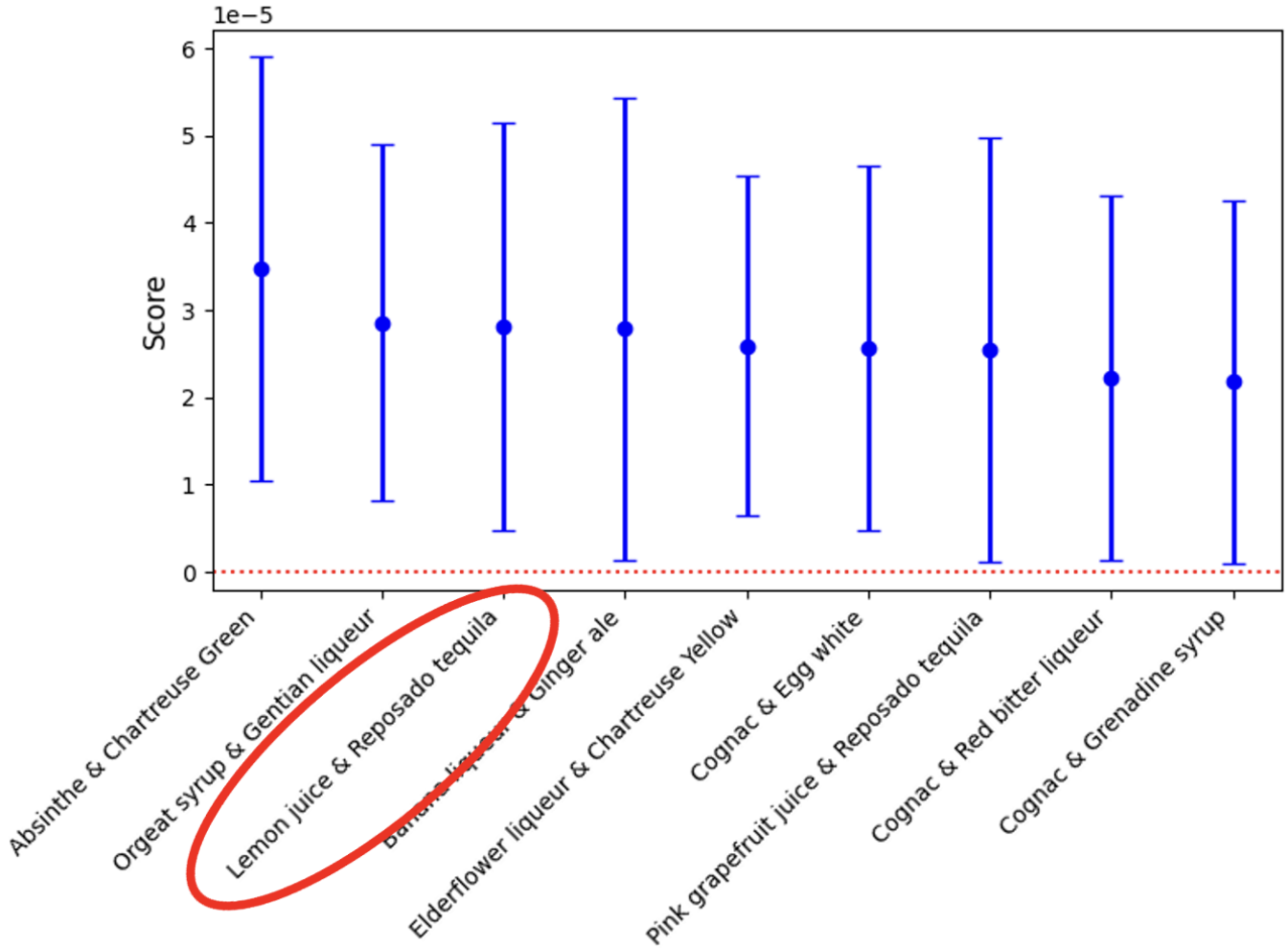
Neptune's Wrath

★★★★½ / ★★★★½

- Gin
- ✓ Absinthe
- Lemon juice
- Sugar syrup (2:1)
- Egg white (pasteurised)
- ✓ Chartreuse Green

Results: Cocktail Ingredients

Inference for Feature Interactions (iLOCO-MP) Results:



Results: Cocktail Ingredients

Inference for Feature Interactions (iLOCO-MP) Results:

21st Century (Difford's recipe)

★★★★★ / ★★★★☆

- **Reposado tequila**
- White crème de cacao
- Aromatized wine
- Blanc quinquina/kina
- **Lemon juice**

Corpse Reviver No.4

★★★★★ / ★★★★½

- Absinthe
- **Reposado tequila**
- Aromatized wine
- Triple sec
- **Lemon juice**
- Sugar syrup (2:1)

Los Pacaminos Margarita

★★★★★ / ★★★★☆

- **Reposado tequila**
- Blanco tequila
- Mezcal
- Triple sec
- Corn/maize liqueur
- Mexican corn whisky
- Lime juice
- **Lemon juice**
- Agave syrup
- Margarita Bitters

Swedish Margarita

★★★★½ / ★★★½☆

- **Reposado tequila**
- Swedish Punsch
- Lime juice
- **Lemon juice**

Results: Cocktail Ingredients

Inference for Feature Interactions (iLOCO-MP) Results:

Acapulco No.2

★★★★½ / ★★★★☆

- Reposado tequila
- Strega liqueur
- Lemon juice
- Egg white (pasteurised)

Tequila Basil Lemonade

★★★★½ / ★★★★☆

- Reposado tequila
- Basil leaves
- Sugar syrup (2:1)
- Lemon juice
- Soda (club soda) water

Dried Meadow Flower

★★★★½ / ★★★★☆

- Reposado tequila
- Elderflower liqueur
- Gentian liqueur
- Lemon juice
- Soda (club soda) water

Sunny Disposition Highball

★★★★½ / ★★★½☆

- Reposado tequila
- Apricot liqueur
- Gentian liqueur
- Lemon juice
- Bitter lemon tonic

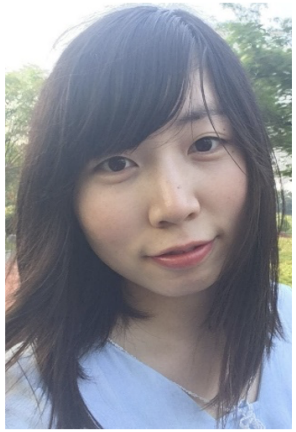
Summary

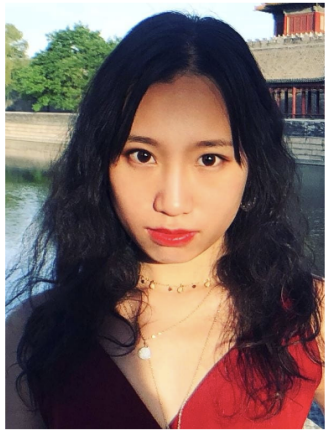
Summary

- Uncertainty quantification for ML interpretations is critical.
  - Interpreting features.
- Minipatch LOCO inference for feature importance.
  - Model-agnostic, distribution-free, & assumption-light.
  - Avoids data-splitting.
  - Inference on $\hat{f}()$.
  - Theoretical contributions might be of independent interest.
- Minipatch ensembles give free algorithmic stability!
- iLOCO for feature interaction importance.
  - Fast & powerful metric.
  - Inference via minipatches.
- Tons of future statistical work on inference for interpretable ML!

Acknowledgments

- **Lili Zheng**, UIUC.
- **Luqin Gan**, PhD Rice.
- **Camille Little**, PhD Candidate, Rice.
- Michael Weylandt, Baruch College.
- Tarek Zikry, Columbia.





NSF

NIH National Institutes of Health

NeuroNex

Key References

G. I. Allen, L. Gan, and L. Zheng, "Interpretable Machine Learning for Discovery: Statistical Challenges and Opportunities", *Annual Review of Statistics and Its Applications*, **11**:97–121, 2024.

L. Gan, T. Zikry, and G. I. Allen, "Are machine learning interpretations reliable? A stability study", (Submitted) arXiv:2505.15728, 2025+.

L. Gan*, L. Zheng* and G. I. Allen, "Model-Agnostic Confidence Intervals for Feature Importance: A Fast and Powerful Approach Using Minipatch Ensembles", (Submitted), arXiv:2206.02088, 2025+. * *Denotes equal contribution.*

C. Little, L. Zheng and G. I. Allen, "iLOCO: Distribution-Free Inference for Feature Interactions", (Submitted), arXiv:2502.06661, 2025+.

Thank You!