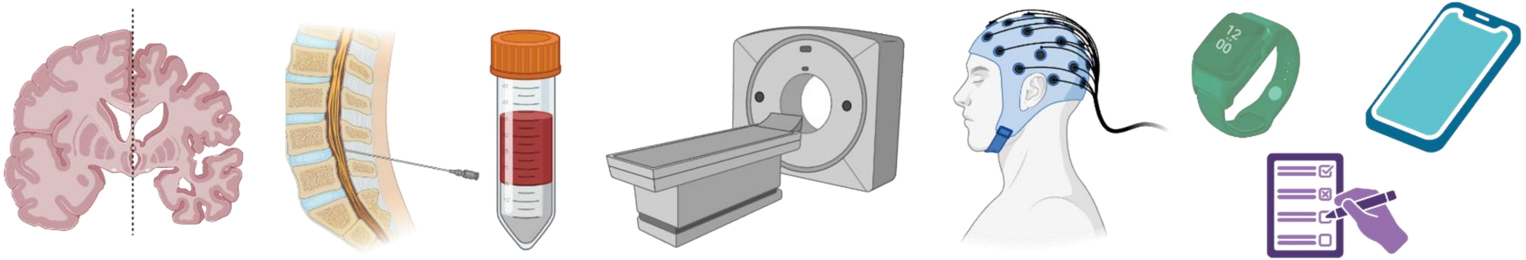
Statistical Interpretation of High-Dimensional Prediction Models using Conditional Permutation Importance

BBS Seminar

Denis Engemann,
Biomarker & Experimental Medicine Leader, Senior Scientist,
Roche Pharma Research & Early Development (pRED),
Neuroscience & Rare Diseases

Credits:
Ahmad Chamma, PhD, Inria Paris-Saclay
Bertrand Thirion, PhD, Inria Paris-Saclay

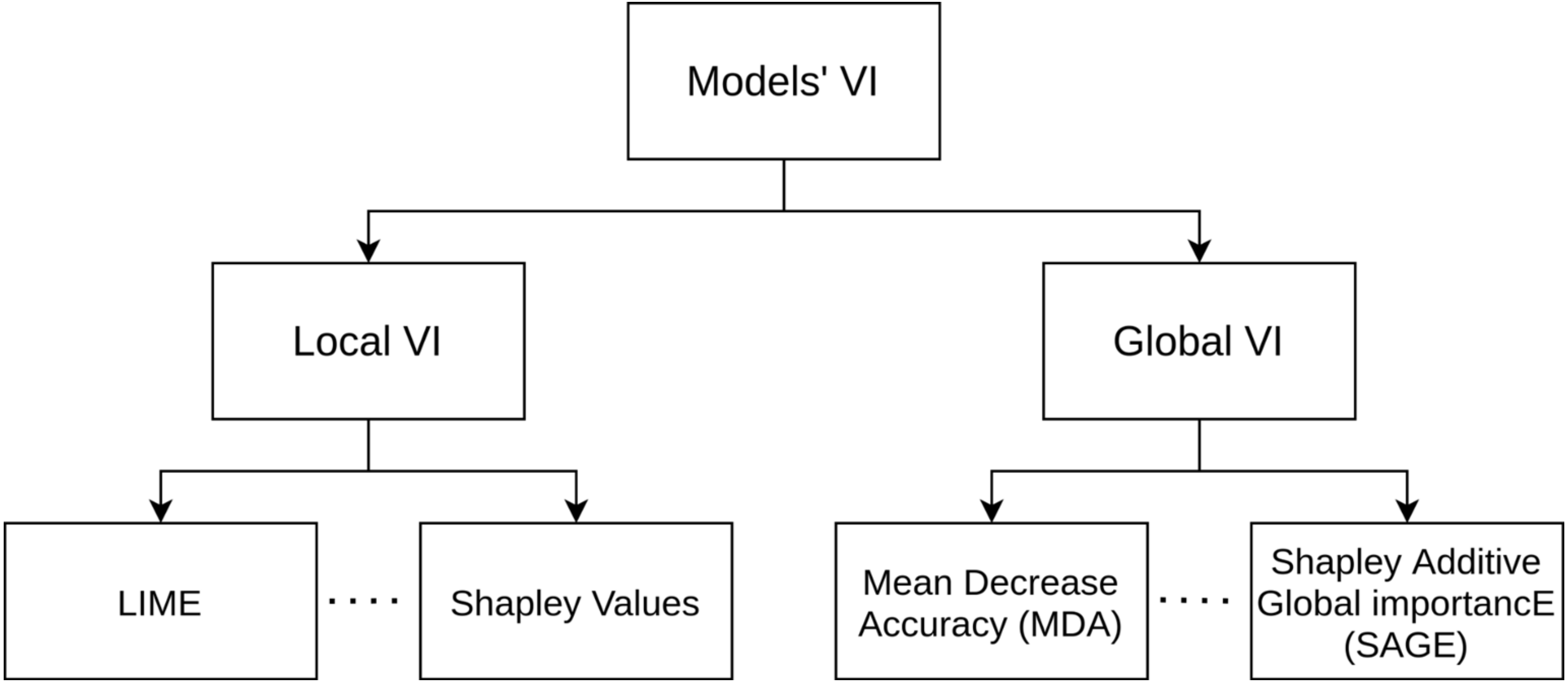
Context:
Multimodal High-Dimensional Data for Biomarker Discovery



- High-dimensional biological data hold promise for novel biomarkers
- ML is an excellent framework for flexible function estimation with heterogeneous data
- Use strength of stochastic optimization for building complex custom models (aka deep learning)
- Need for statistical decision rules constraining interpretation of ML results
- Statistical approach to variable importance literature needed

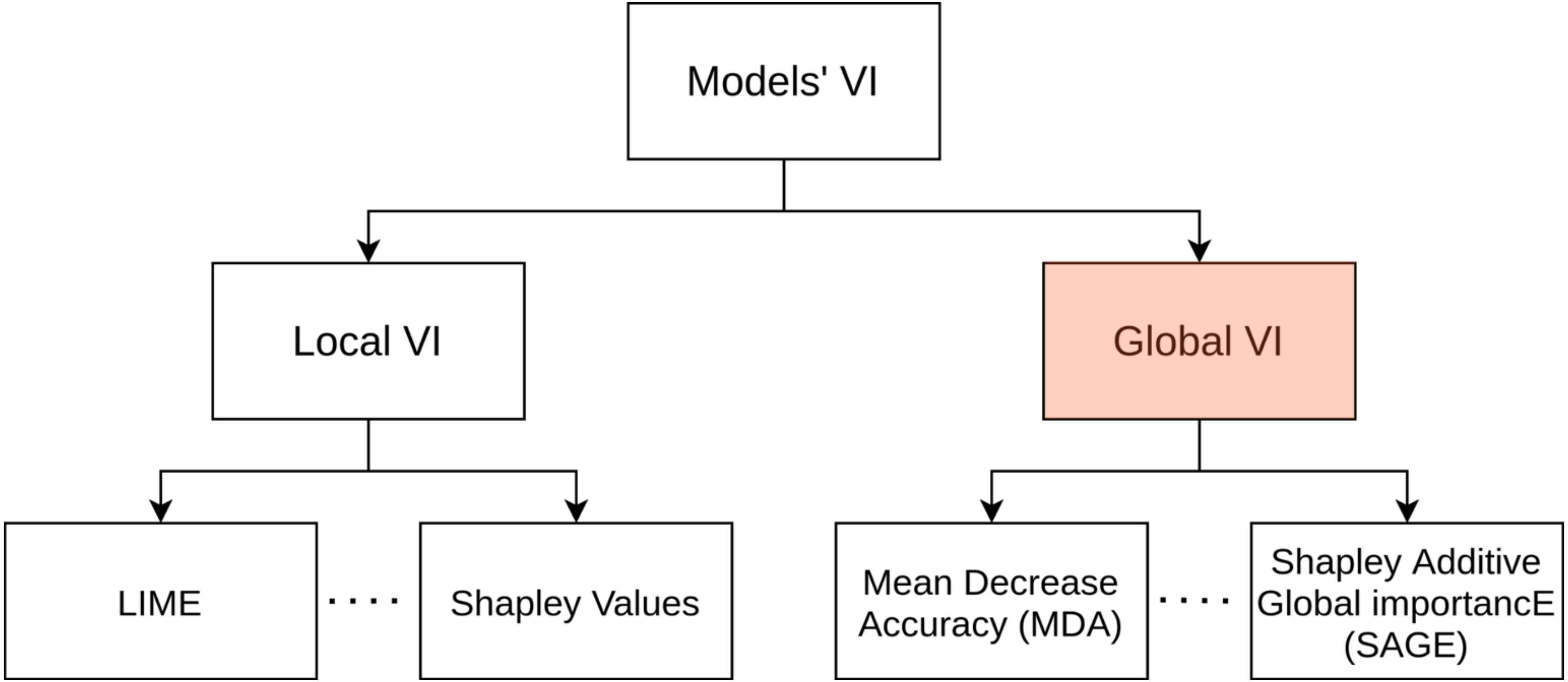
Variable importance: Estimating the influence inputs on model predictions

[e.g. Hooker et al 2018, arXiv:1806.10758; Zien et al 2009, Lecture Notes in Computer Science]



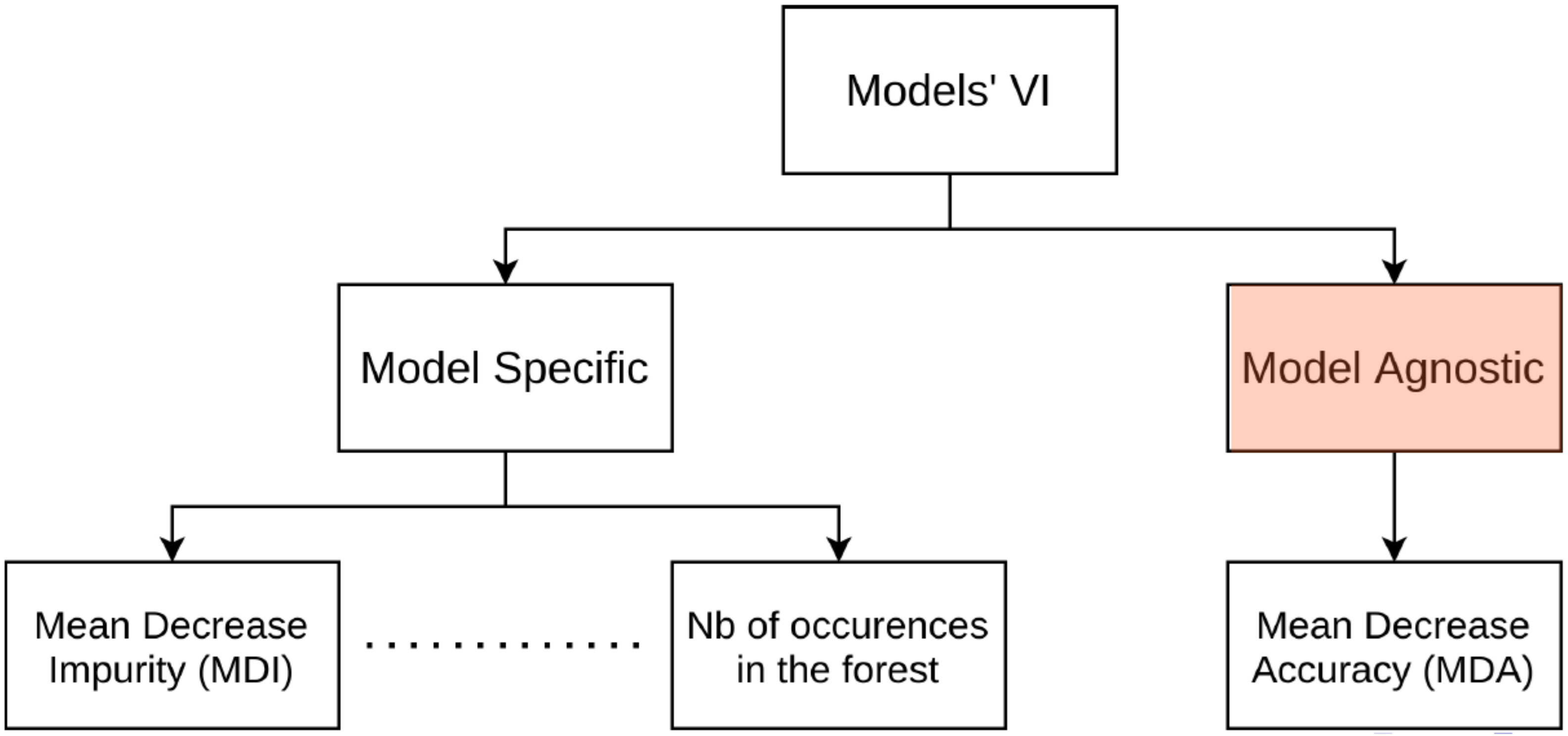
Variable importance: Estimating the influence inputs on model predictions

[e.g. Hooker et al 2018, arXiv:1806.10758; Zien et al 2009, Lecture Notes in Computer Science]



Variable importance: Estimating the influence inputs on model predictions

[e.g. Hooker et al 2018, arXiv:1806.10758; Zien et al 2009, Lecture Notes in Computer Science]

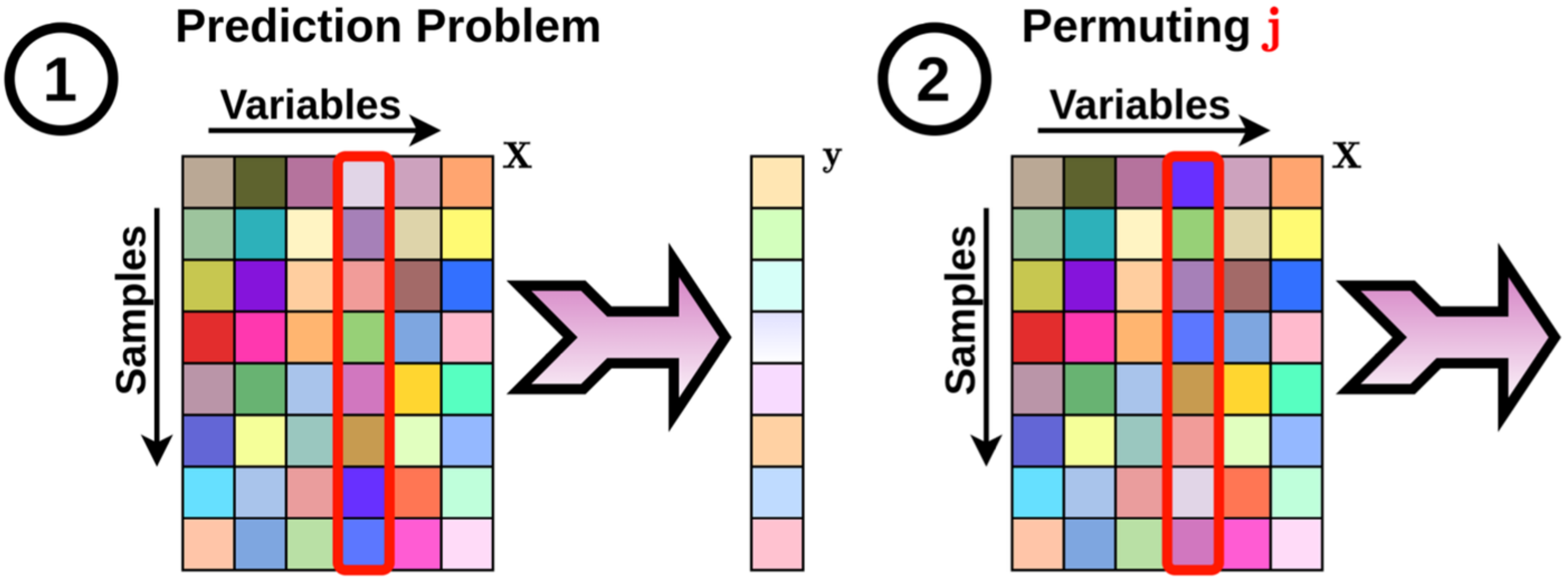


Variable importance:
Estimating the influence inputs on model predictions

Prefer variable importance with statistical guarantees

- **Ideal goal:** find all relevant variables and don't pick up irrelevant variables -> control false discovery rate [Candes et al 2017, J Royal Stat Soc]
- **Impact:** Critical for discover work and study design to pick up the good biomarker candidates
- Simplifying liability management and cut down development time by using **statistical guarantees**
- E.g. guarantees obviate excessive sensitivity analyses

Permutation importance
[Breiman, Machine Learning, 2001]

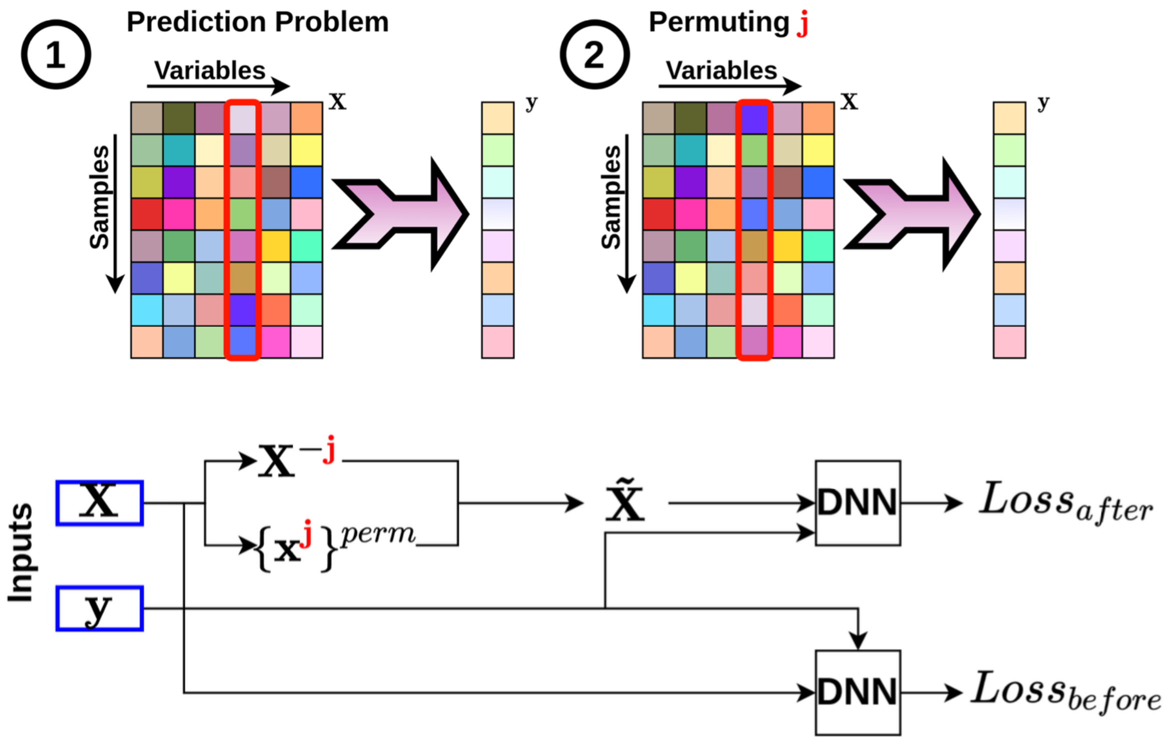


Is variable **j** important?

Permute **j** on testing data and track performance change of model

Permutation importance is alive and well

[Breiman, Machine Learning, 2001]

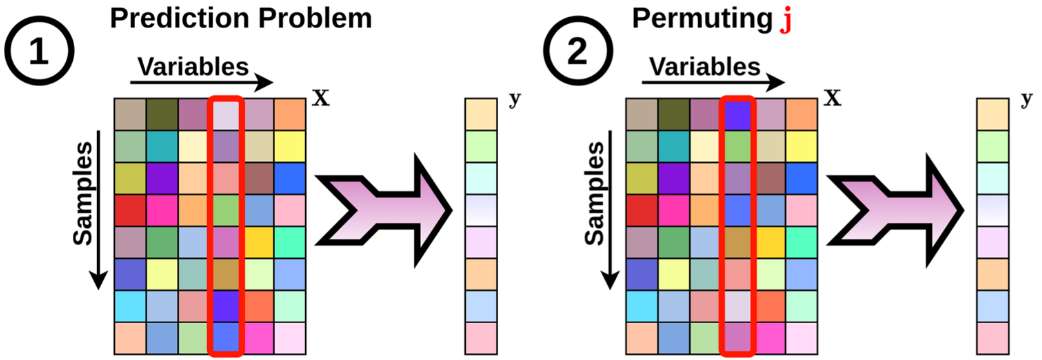


Modern flavors of permutation importance in life science context

- recent integration in artificial neural network architecture (permfit) and successful application in large genetics datasets [Mi et al 2021, Nat Comms]
- directly focus on tracking loss function of model after permutations
- Statistically valid p-values

Permutation importance is alive and well

[Breiman, Machine Learning, 2001]



Modern flavors of permutation importance in life-science context

- recent integration in artificial neural network architecture (permfit) and successful application in large genetics datasets [Mi et al 2021, Nat Comms]

nature COMMUNICATIONS

ARTICLE

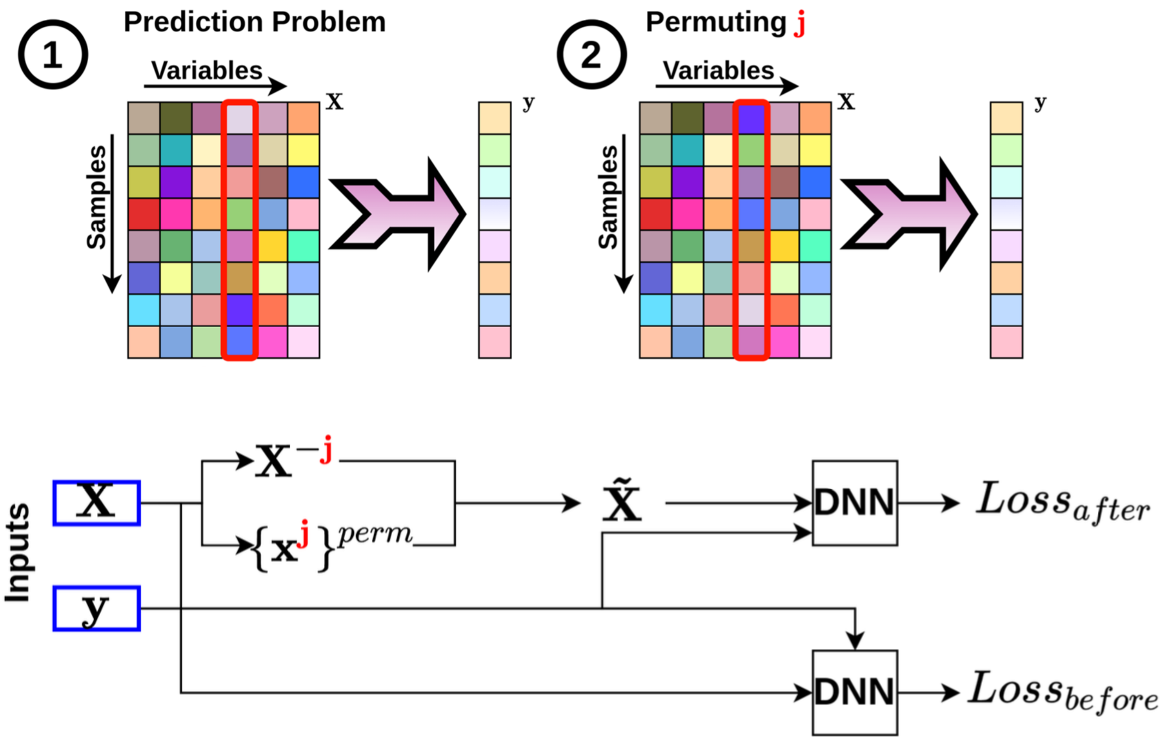
https://doi.org/10.1038/s41467-021-22756-2 OPEN

Check for updates

Permutation-based identification of important biomarkers for complex diseases via machine learning models

Xinlei Mi[1], Baiming Zou[2], Fei Zou[2] & Jianhua Hu[3]

Permutation importance is alive and well
[Breiman, Machine Learning, 2001]



Modern flavors of permutation importance in life-science context

- recent integration in artificial neural network architecture (permfit) and successful application in large genetics datasets [Mi et al 2021, Nat Comms]
- track focus loss function after permutations
- statistically valid p-values
- **breaks if variables are correlated!** [Chamma et al 2023]

proceedings.neurips.cc

Statistically Valid Variable Importance Assessment through Conditional Permutations

Part of Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Main Conference Track

Bibtex | Paper | Supplemental

Authors

Ahmad CHAMMA, Denis A. Engemann, Bertrand Thirion

Report an Issue | Name Change Policy

Paper #1

[Chamma, Engemann, Thirion, 2023, NeurIPS]

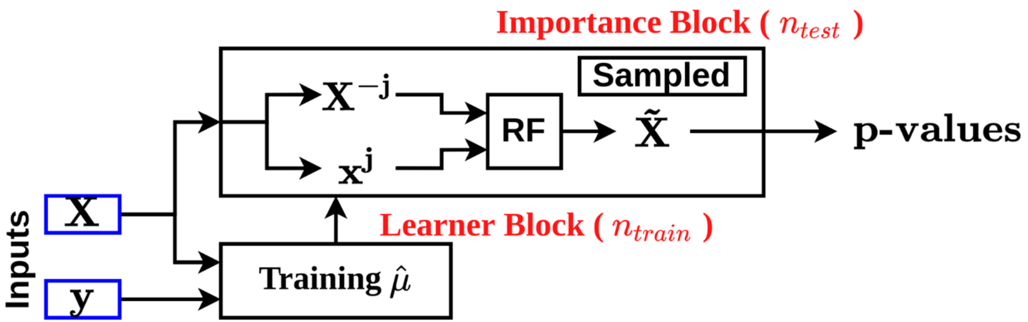
Conditional permutation importance (CPI)
[Chamma, Engemann, Thirion, 2023, NeurIPS]

- Let $\epsilon^{j} = \mathbf{x}^{j}$ - $\hat{\mathbf{x}}^{j}$ with $\hat{\mathbf{x}}^{j} = \mathbb{E}(\mathbf{x}^{j}|\mathbf{X}^{-j})$

Samling $\mathbf{x}^{j}$ from the conditional distribution

$$\tilde{\mathbf{x}}^{j} = \hat{\mathbf{x}}^{j} + \{\epsilon^{j}\}^{perm}$$

Why? The dependency between the variable of interest and the remaining variables is preserved.

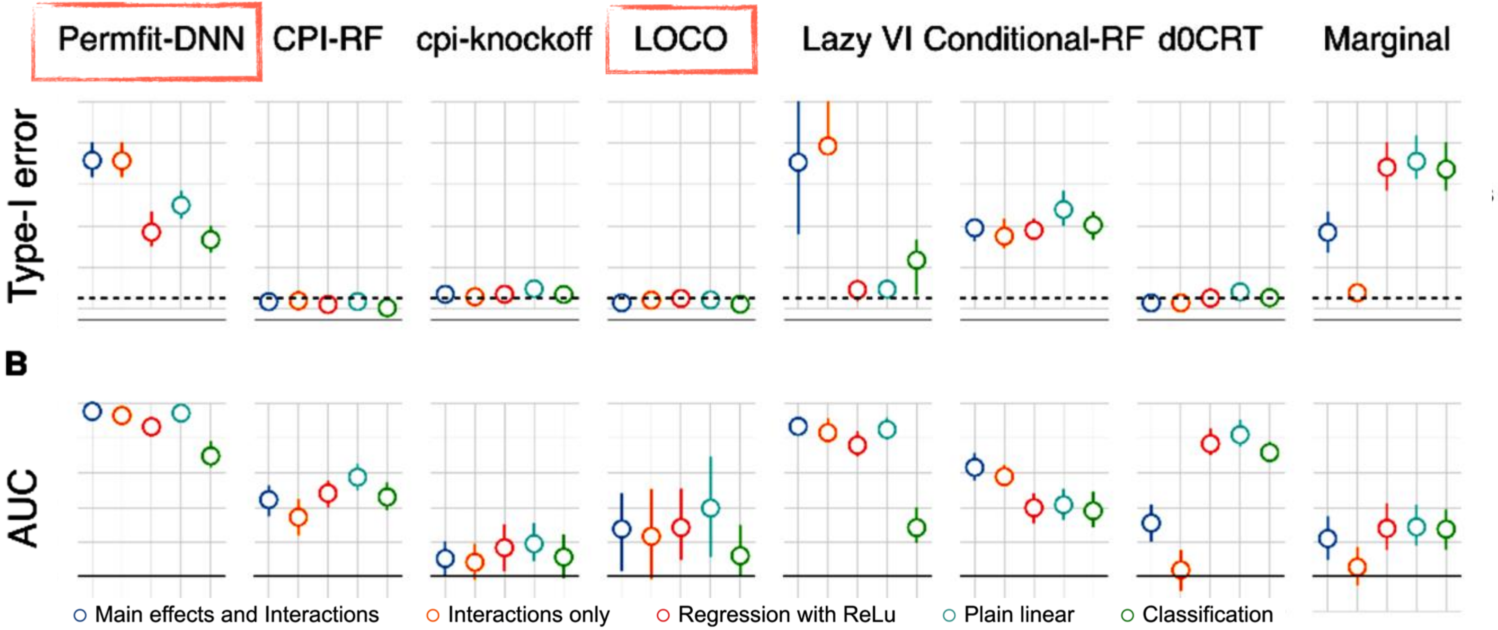


In a nutshell

- Statistically valid p-values **even if variables are correlated!**
- Fast because we can use approximate estimator during sampling phase (e.g. random forest) and avoid refitting (cf. vs LOCO approach)
- Converges to permfit if variables are uncorrelated
- Developed VS DNN architecture but

Large-scale benchmarking of variable importance methods: Accurate detection and ranking (AUC) VS false positives (type-1 error)

Standard permutations:

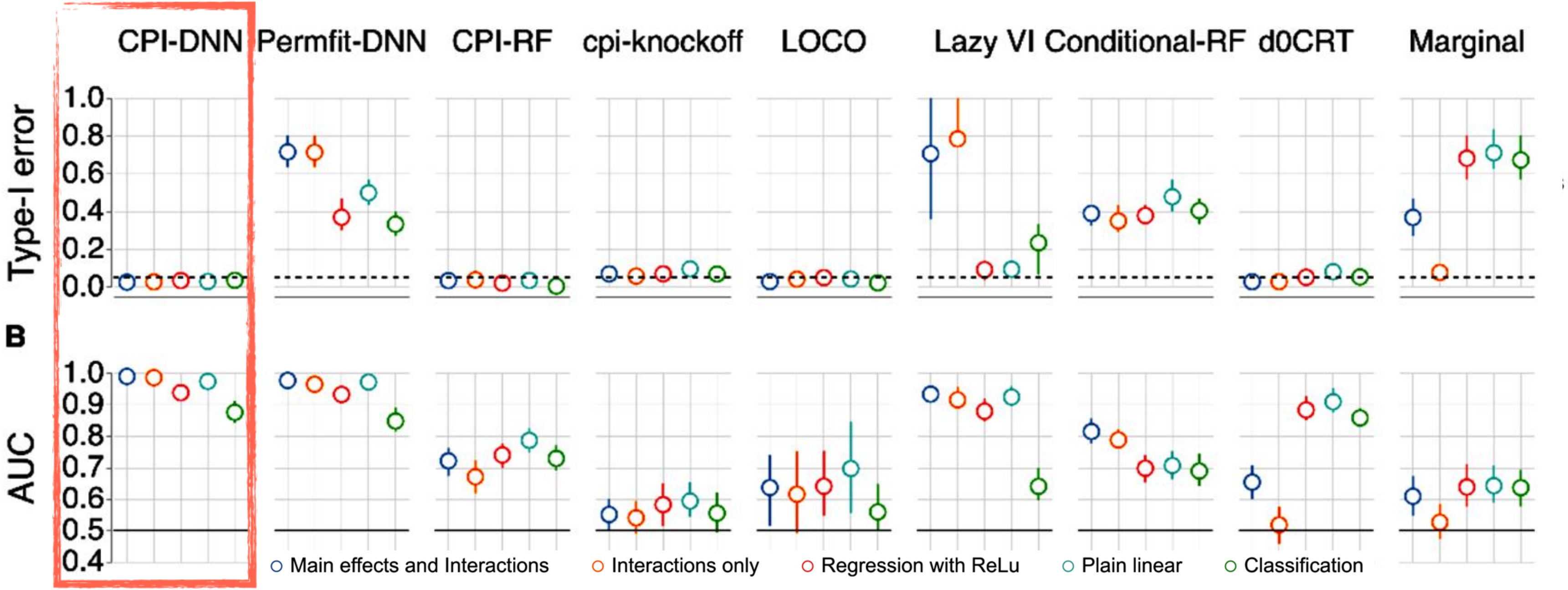


[Chamma, Engemann & Thirion, 2023, NeurIPS]

Observation: approaches tend to be either good at Type-1 error or AUC

CPI-DNN is good at ranking while avoiding false positives! Other methods good at either detecting OR controlling type-1 error

Standard permutations: Decomposition of variable for conditional permutation:

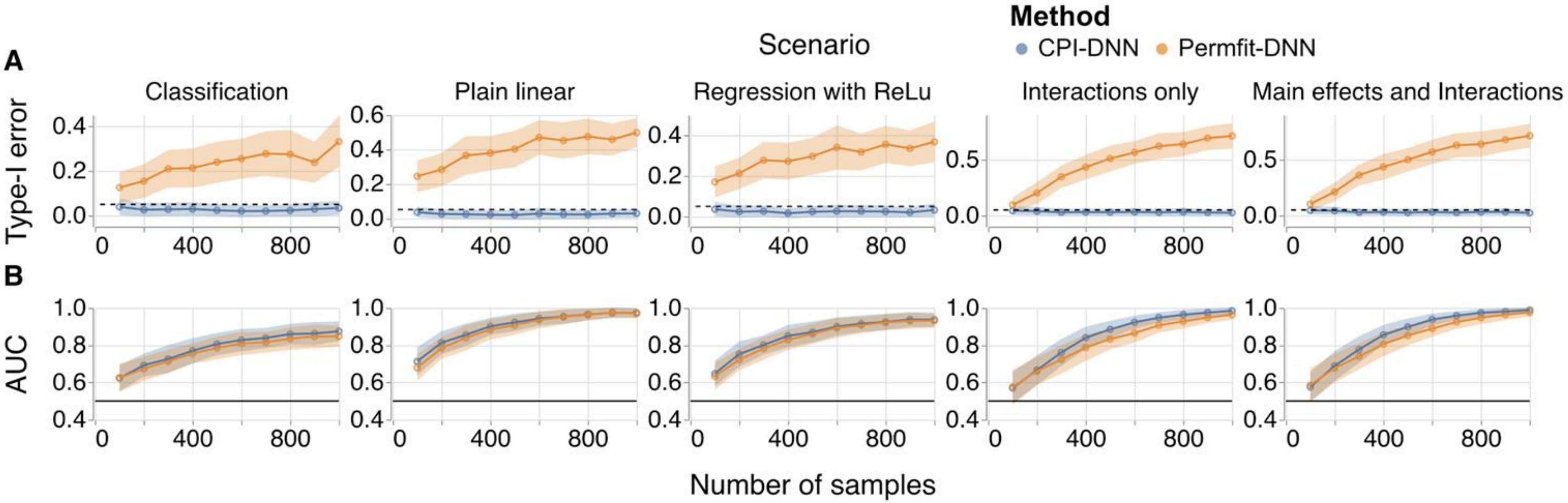


[Chamma, Engemann & Thirion, 2023, NeurIPS]

Proposed method highly sensitive & controlling type-1 error

Need for variable importance measures with support for correlated variables and in the large-scale biomedical setting CPI-DNN

Standard permutations: Decomposition of variable for conditional permutation:



[Chamma, Engemann & Thirion, 2023, NeurIPS]

Proposed method robust across #samples & generative scenarios

Deep-dive into CPI-DNN - complexity

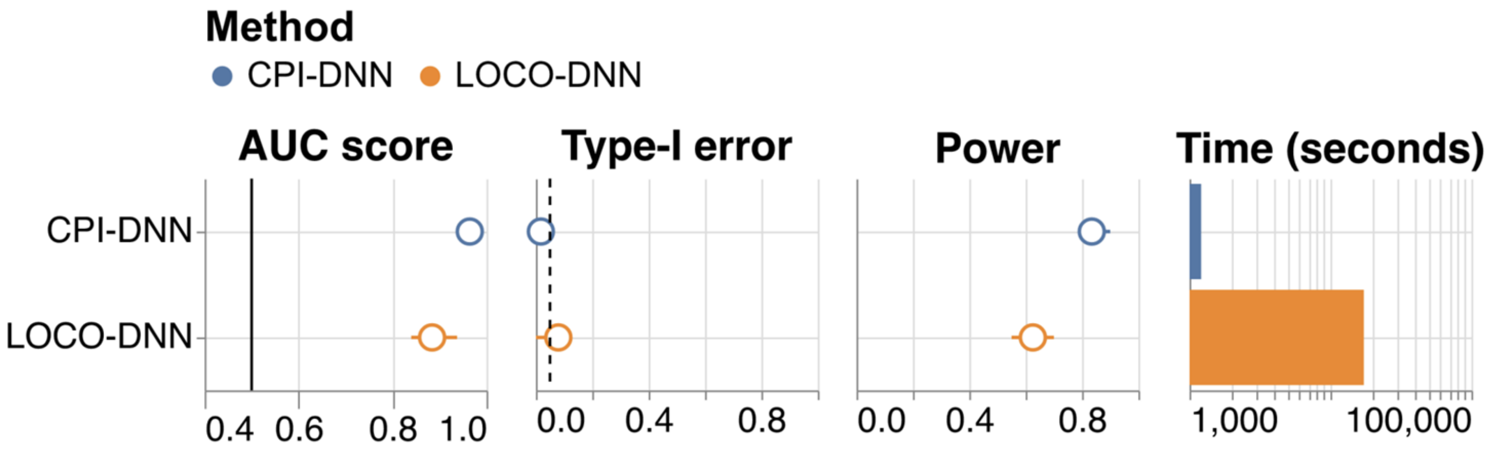
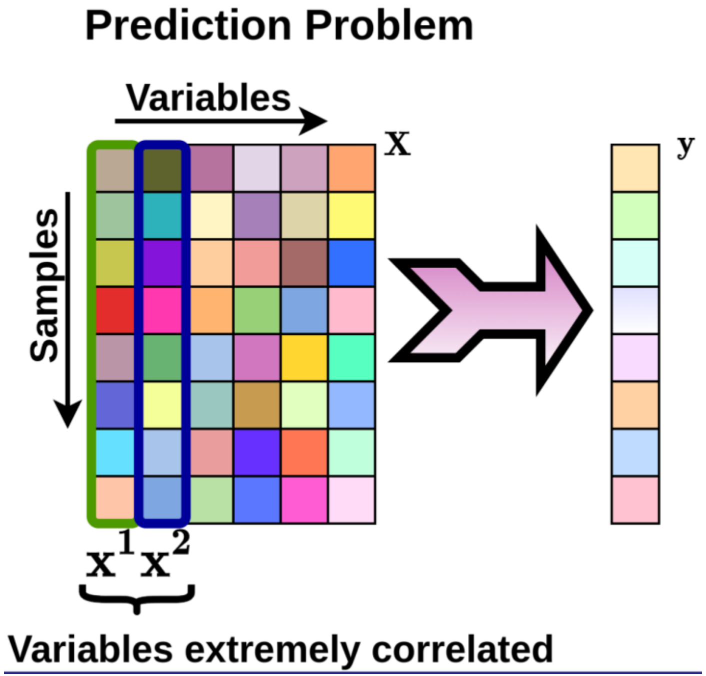


Figure S2: ***CPI-DNN* vs *LOCO-DNN***: Performance at detecting important variables on simulated data with $n = 1000$, $p = 50$ and $\rho = 0.8$ in terms of **(AUC score)**, **Type-I error**, **Power** and **Time**. Dashed line: targeted type-I error rate. Solid line: chance level.

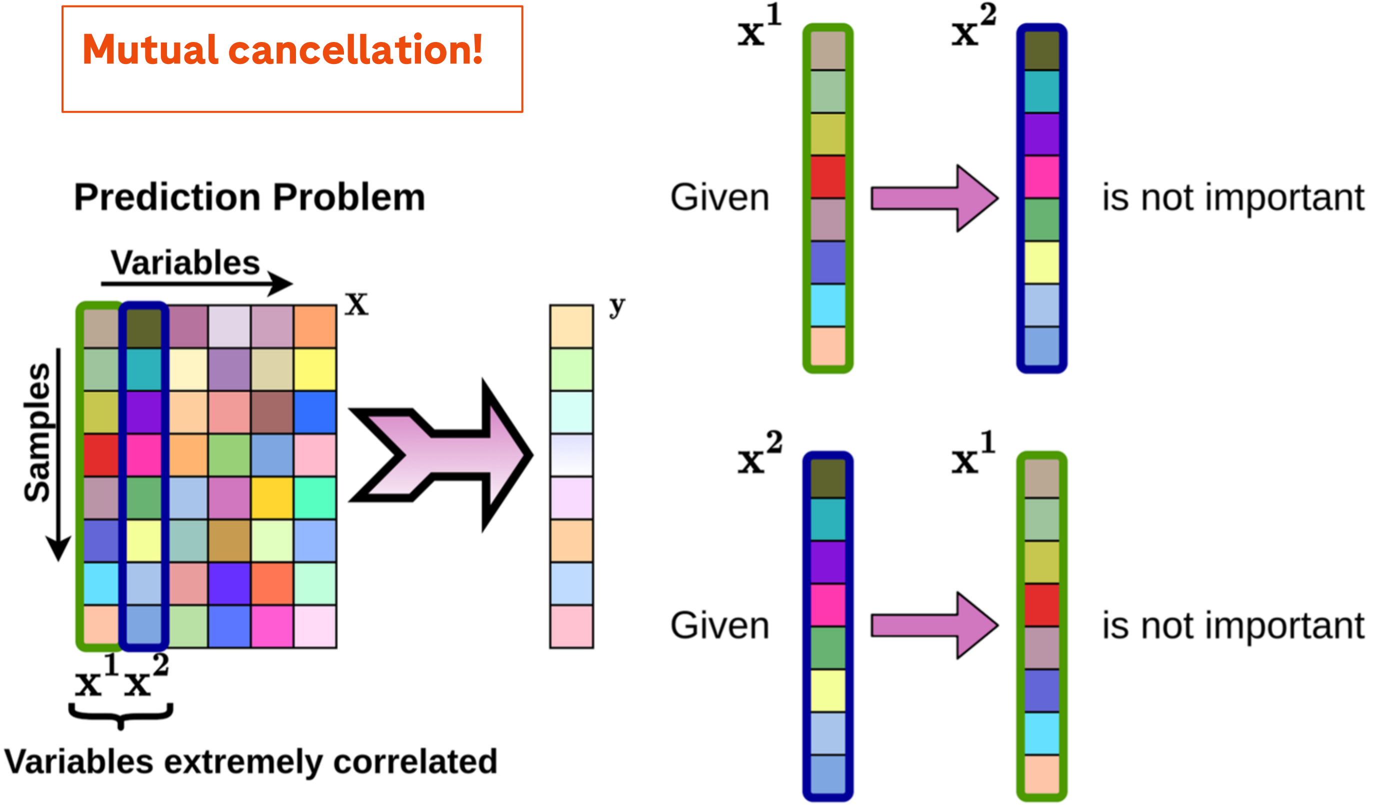
[Chamma, Engemann & Thirion, 2023, NeurIPS]

Limits of conditional inference



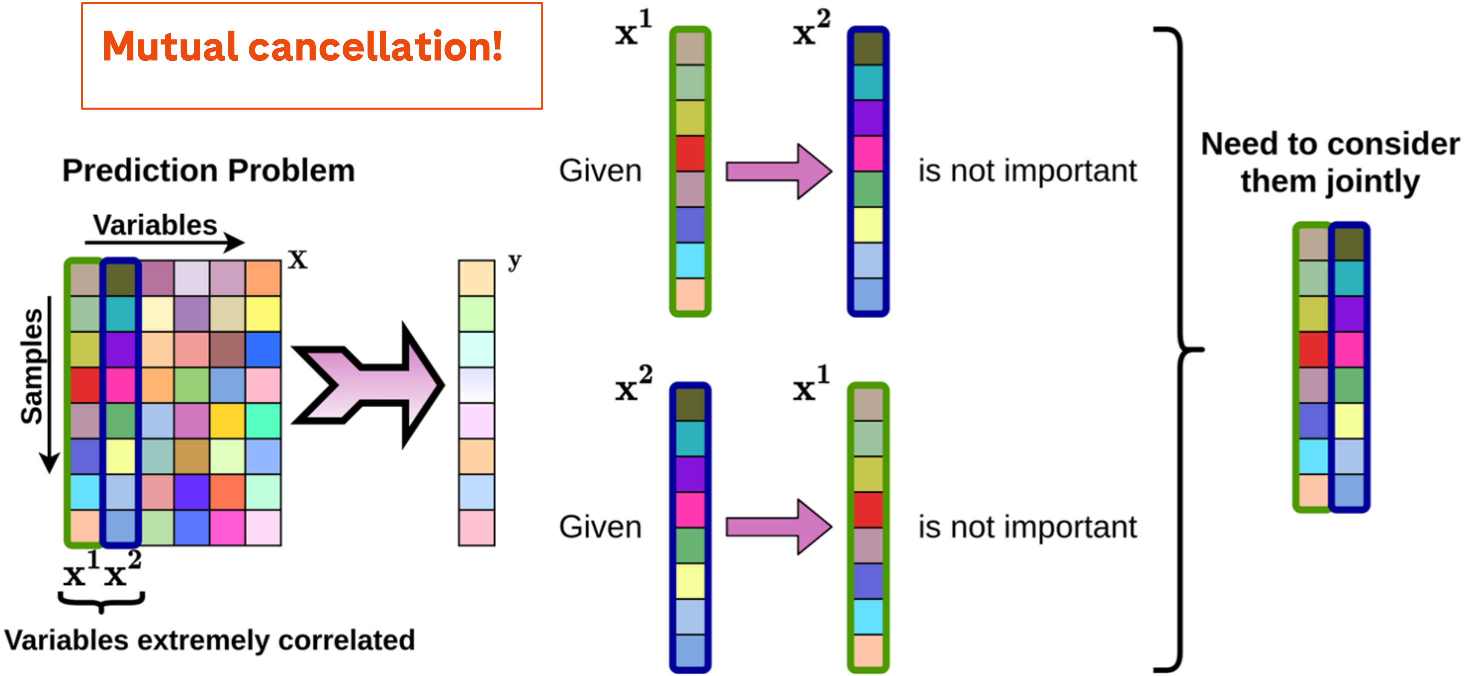
Limits of conditional inference

Mutual cancellation!



Limits of conditional inference – grouping to the rescue?

Mutual cancellation!



ojs.aaai.org

Variable Importance in High-Dimensional Settings Requires Grouping

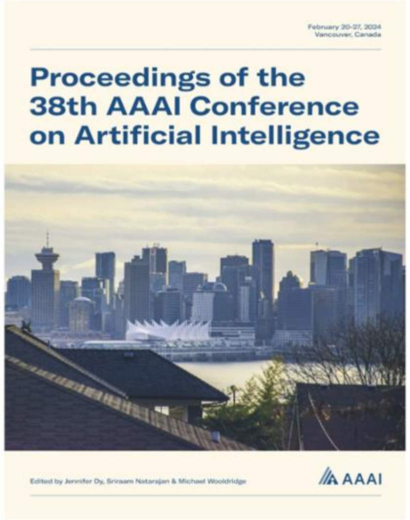
Ahmad Chamma
Inria-Saclay, Palaiseau, France Université Paris-Saclay CEA Saclay

Bertrand Thirion
Inria-Saclay, Palaiseau, France Université Paris-Saclay CEA Saclay

Denis Engemann
Roche Pharma Research and Early Development, Neuroscience and Rare Diseases, Roche Innovation Center Basel, F. Hoffmann–La Roche Ltd., Basel, Switzerland

DOI: https://doi.org/10.1609/aaai.v38i10.28997

Keywords: ML: Transparent, Interpretable, Explainable ML, ML: Classification and Regression, ML: Deep Learning Algorithms, ML: Dimensionality Reduction/Feature Selection, ML: Ensemble Methods



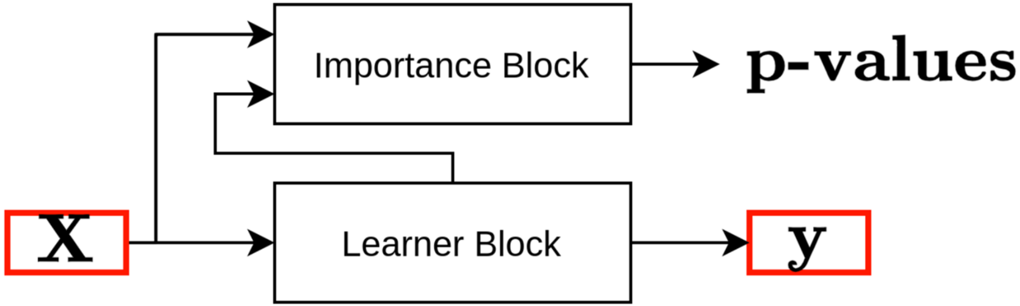
Roche

Paper #2

[Chamma, Thirion, Engemann, 2024, AAAI]

Block-based Conditional Permutation Importance (CPI)

[Chamma, Thirion, Engemann, 2024, AAAI]

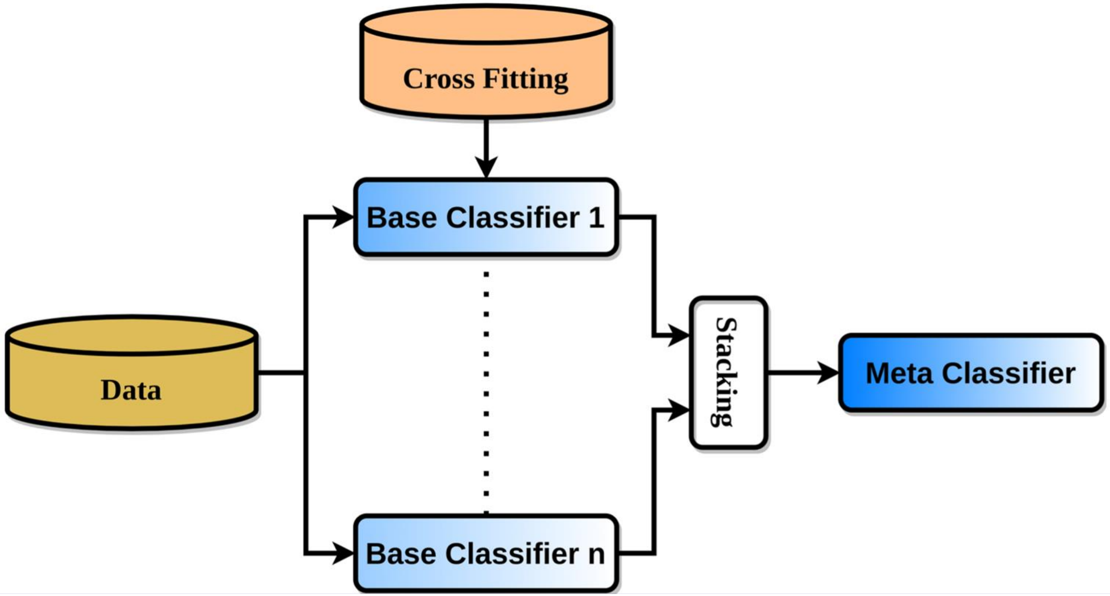


In a nutshell

- Statistically valid p-values **per block**
- Speed gains through **internal stacking**
- Converges to CPI if group size = 1 and to permfit if variables are uncorrelated
- Developed Vs DNN architecture but flexible design

Make use of stacking

[Chamma, Thirion, Engemann, 2024, AAAI]

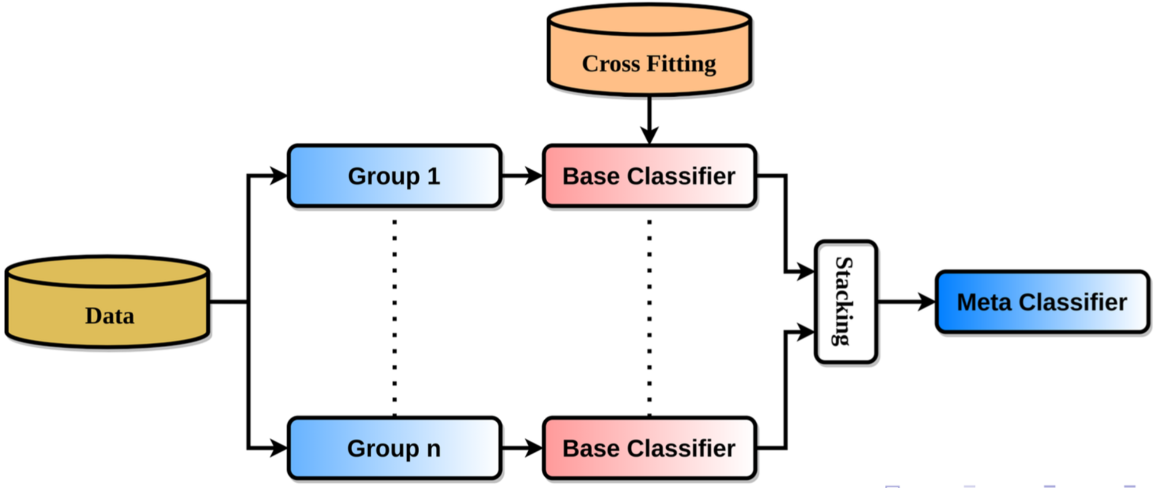


Stacking

- **Original idea:** Enhancing predictions by stacking multiple prediction models [Wolpert, Neural Networks, 1992]

Make use of stacking

[Chamma, Thirion, Engemann, 2024, AAAI]

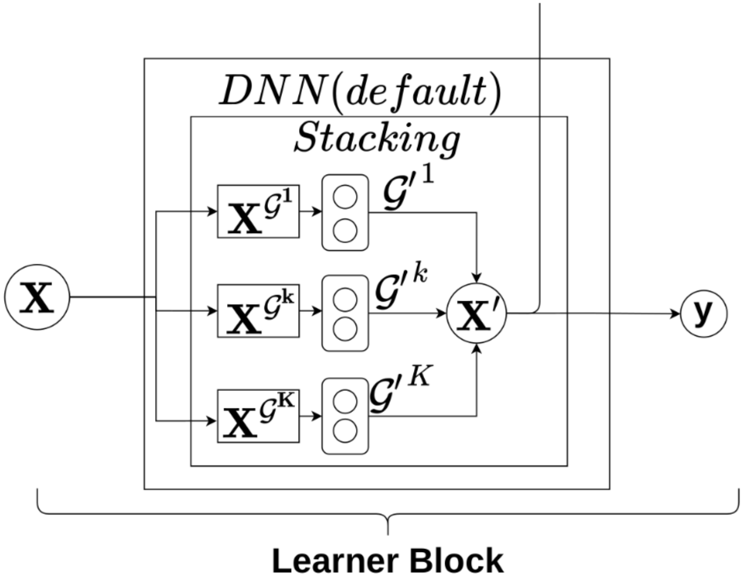


Stacking

- **Original idea:** Enhancing predictions by stacking multiple prediction models [Wolpert, Neural Networks, 1992]
- **Adaptation**: Combine multiple input domains and groups of variables [Rahim et al 2015, Liem et al 2017, Engemann et al 2020, ...]

Make use of stacking
[Chamma, Thirion, Engemann, 2024, AAAI]

G: Original group, G′: Linear projected group

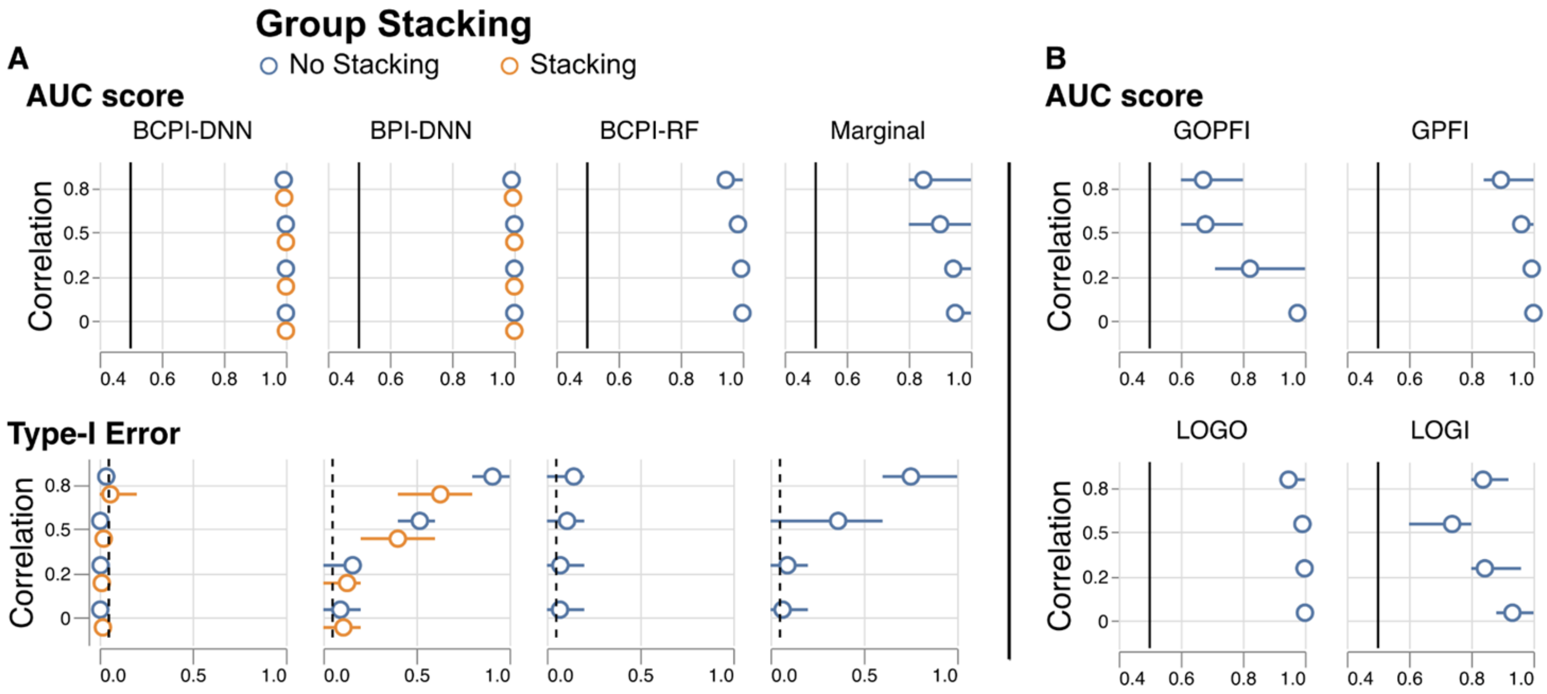


Stacking

- **Original idea:** Enhancing predictions by stacking multiple prediction models [Wolpert, Neural Networks, 1992]
- **Adaptation**: Combine multiple input domains and groups of variables [Rahim et al 2015, Liem et al 2017, Engemann et al 2020, ...]
- **New:** Integrate stacking into DNN architecture as linear sublayer

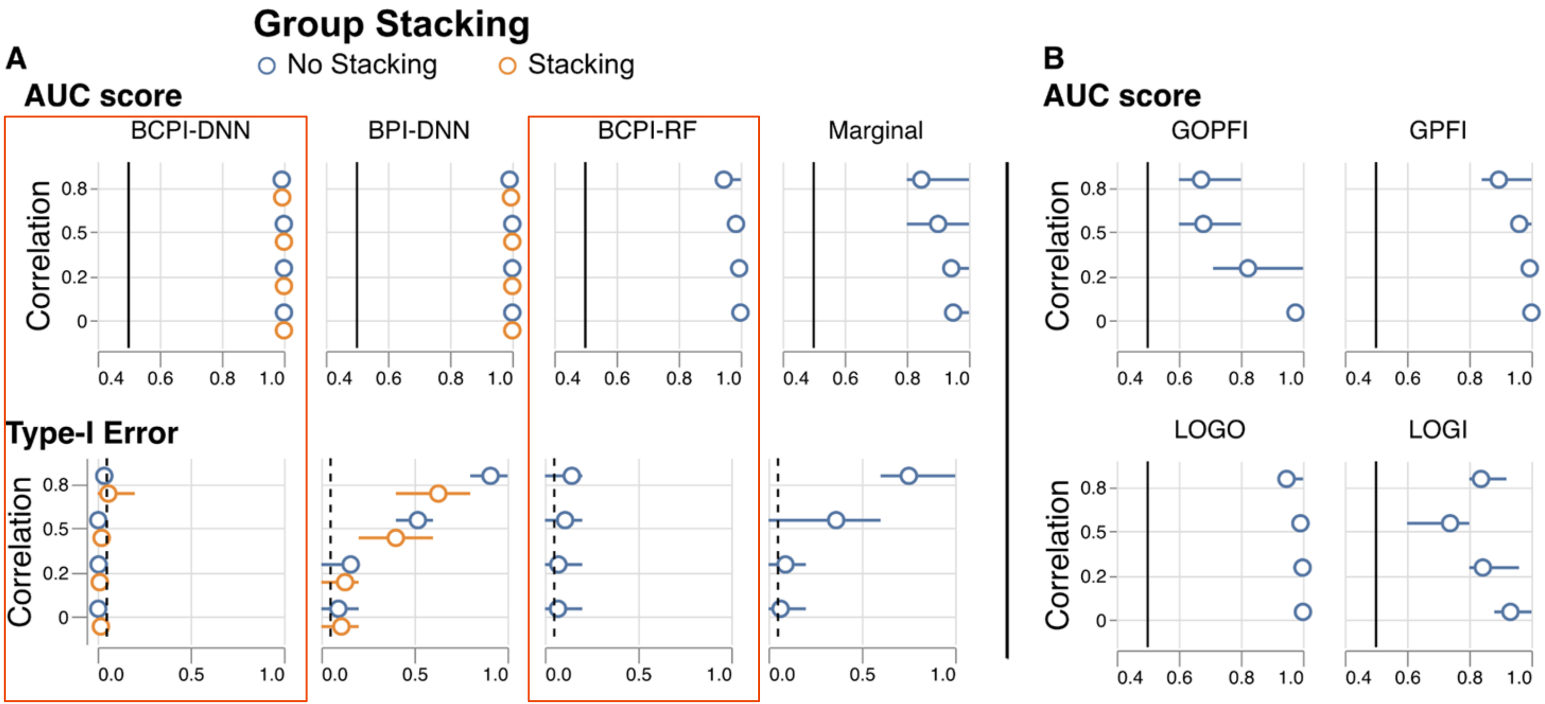
BCPI: correct block ranking & controlling type-1 error

[Chamma, Thirion, Engemann, 2024, AAAI]



BCPI: correct block ranking & controlling type-1 error

[Chamma, Thirion, Engemann, 2024, AAAI]



BCPI: speed gains through internal stacking

[Chamma, Thirion, Engemann, 2024, AAAI]

Stacking improves computation times

BCPI: speed gains through internal stacking

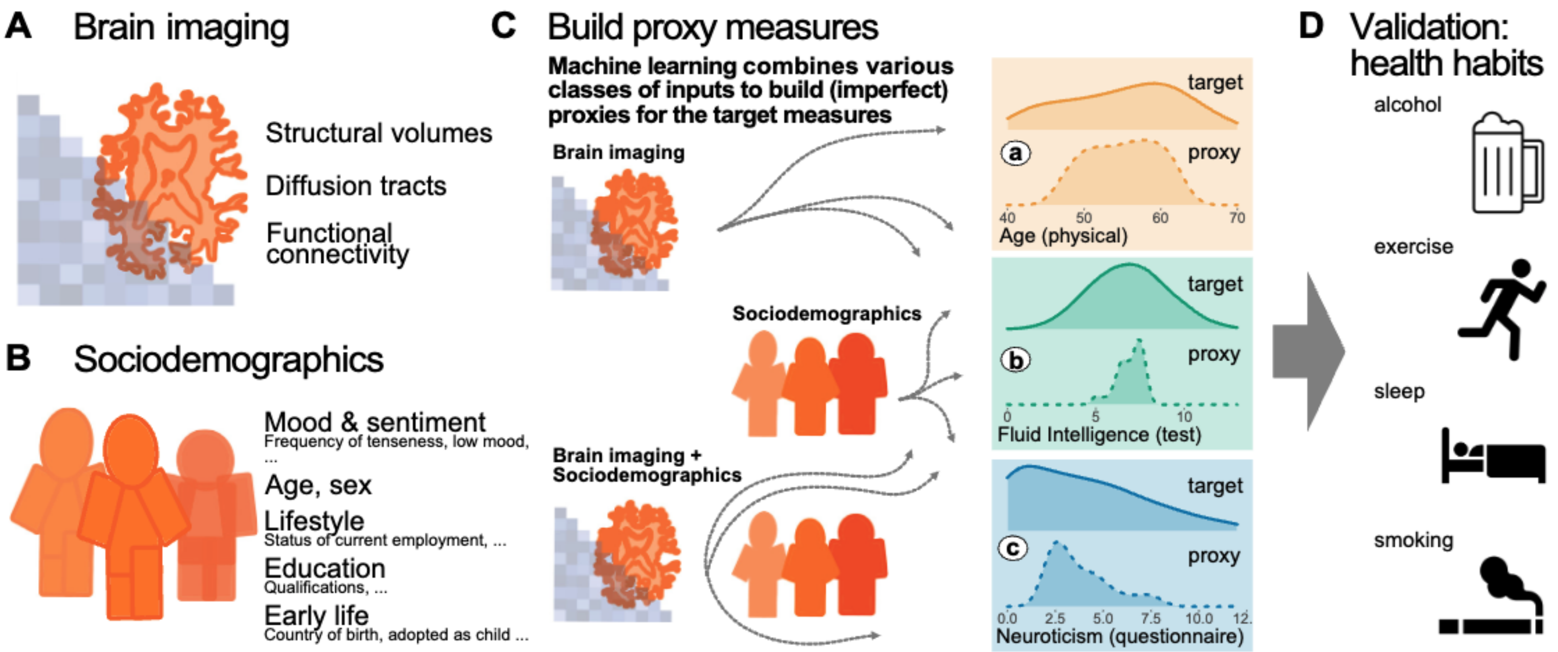
[Chamma, Thirion, Engemann, 2024, AAAI]

Stacking improves computation times

while preserving type-1 error control and high block-ranking performance

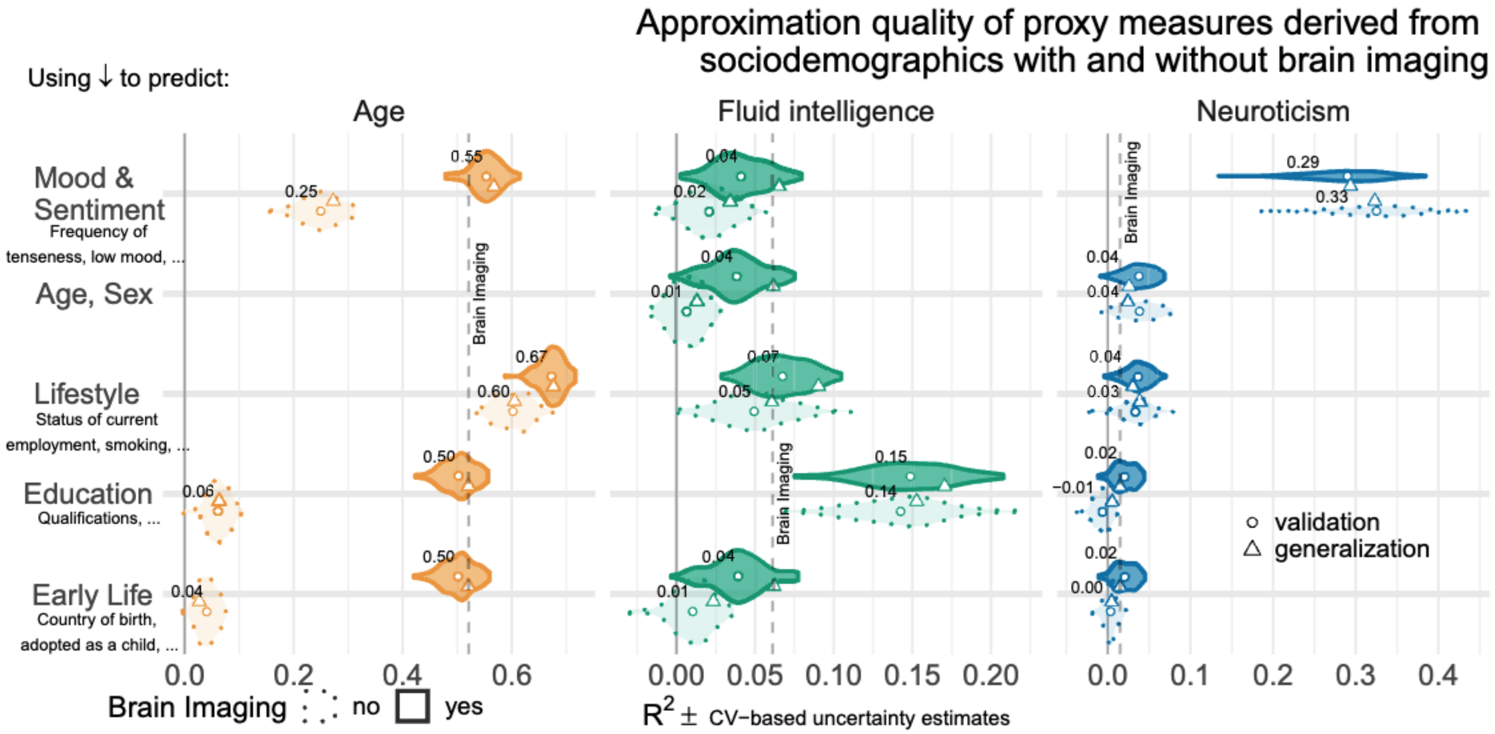
Empirical example: Proxy measures of mental health?

[Dadi, … & Engemann, 2021, GigaScience]



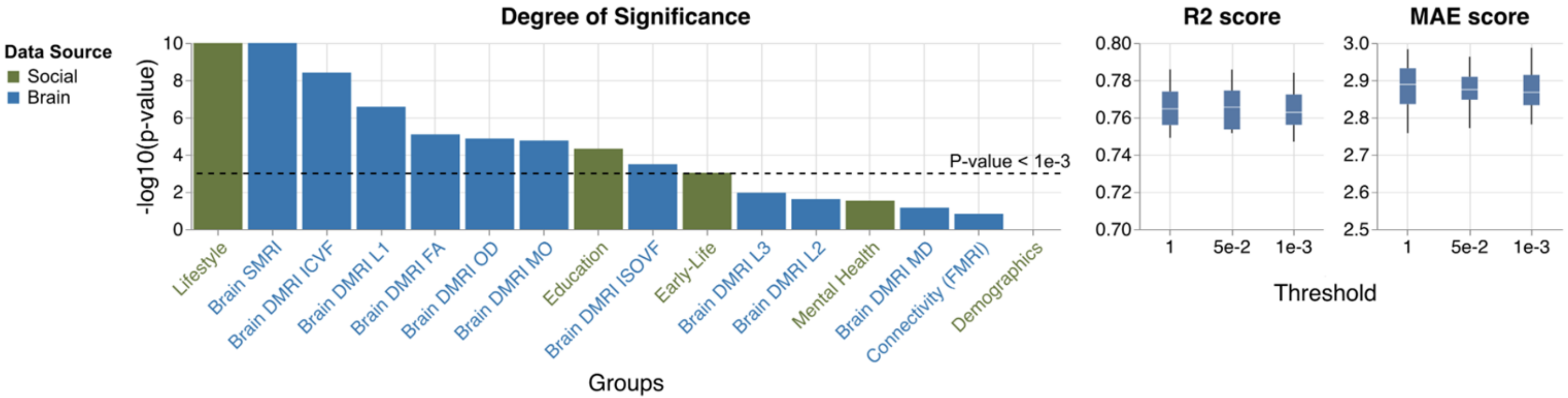
Empirical example: Proxy measures of mental health?

[Dadi, … & Engemann, 2021, GigaScience]



Dadi et al. 2021 revisited: BCPI for fine-grained inference

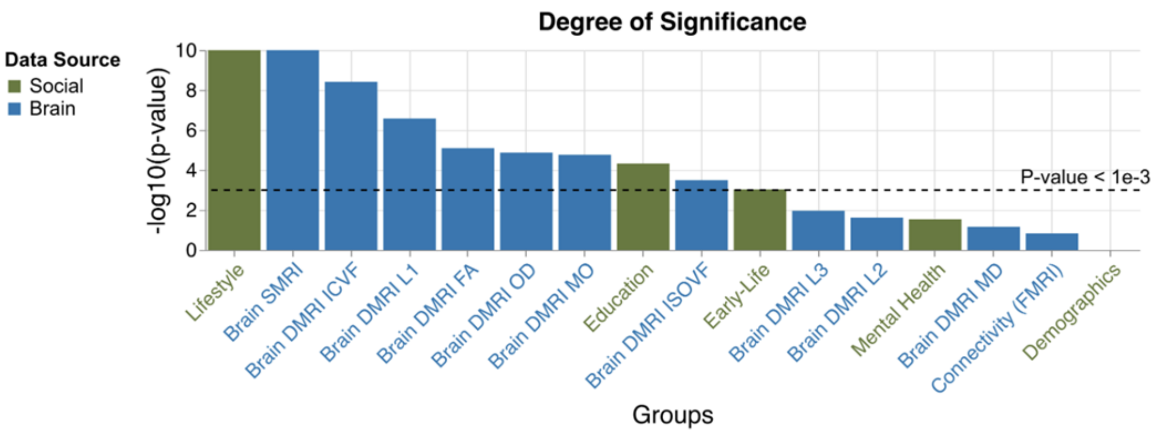
[Chamma, Thirion, Engemann, 2024, AAAI]

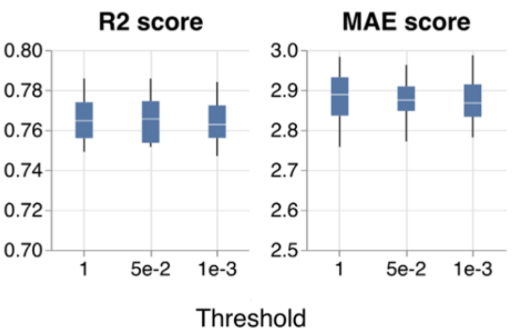


BCPI for age prediction: lifestyle factors, anatomical & diffusion MRI & education provide non-redundant information

Dadi et al. 2021 revisited: BCPI for fine-grained inference

[Chamma, Thirion, Engemann, 2024, AAAI]





BCPI for age prediction: lifestyle factors, anatomical & diffusion MRI & education provide non-redundant information

BCPI for variable selection: reduced model (cross-fitted) preserves prediction performance

Take home messages

Conditional permutation importance methods

- **CPI** plus expressive base learner (e.g. DNN) provides strong detection-performance with type-1 error control in the presence of **correlated variables**
- Faster than e.g. LOCO methods
- **BCPI** extends and generalizes this behavior to **high-dimensional structured** data with **extreme correlations** via group-level inference
- **Flexible toolbox**: plug your own models