18.337 Presentation

05/10/2023

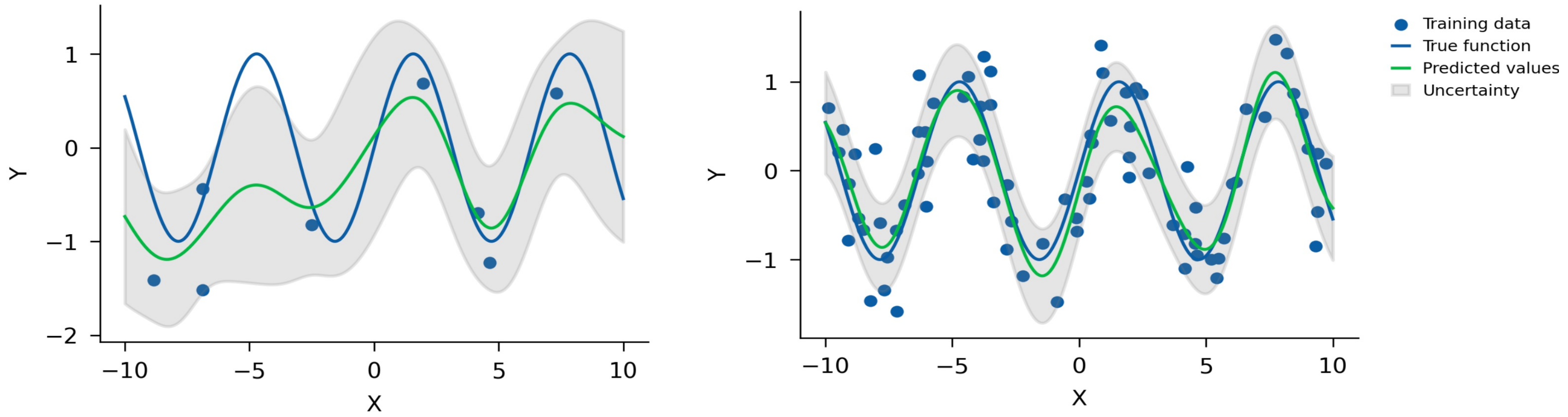
Uncertainty Quantification in High-throughput biological datasets with Gaussian Process

Thomas Cheng & Davy Deng

18.337 - Parallel Computing and Scientific Machine Learning

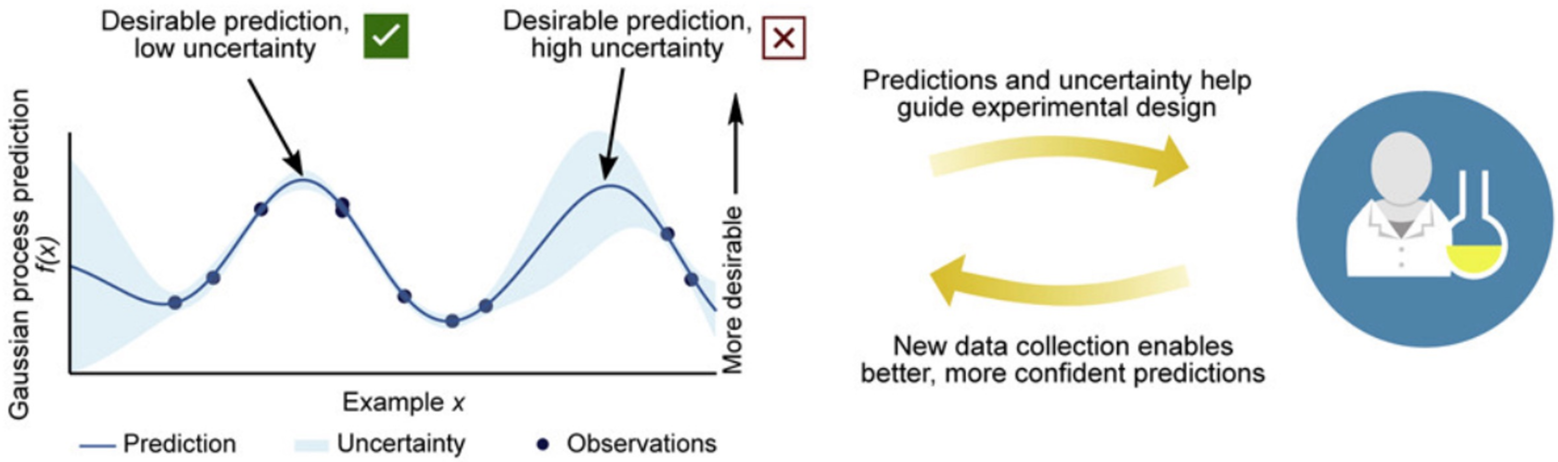
HARVARD MEDICAL SCHOOL
MASSACHUSETTS GENERAL HOSPITAL RESEARCH INSTITUTE
BROAD INSTITUTE

Gaussian Process (GP) for uncertainty quantification



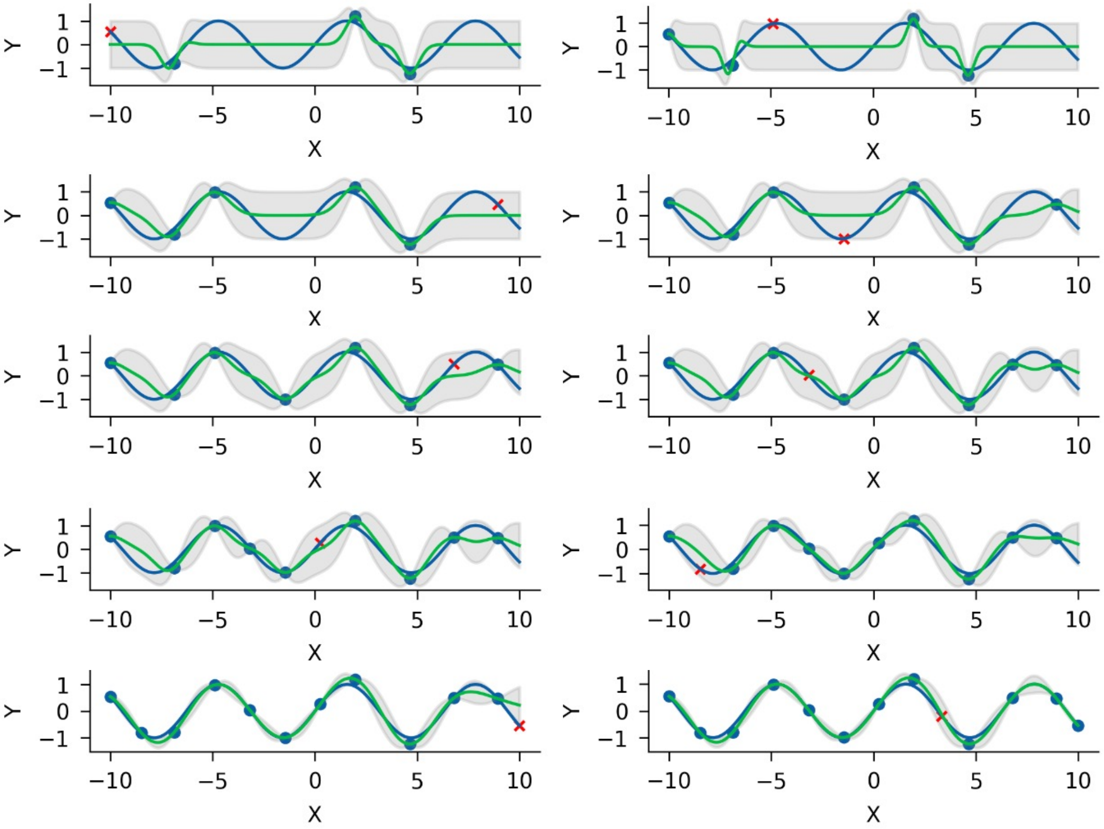
Hie et al, *Cell Systems* (2020)

Evidential active learning

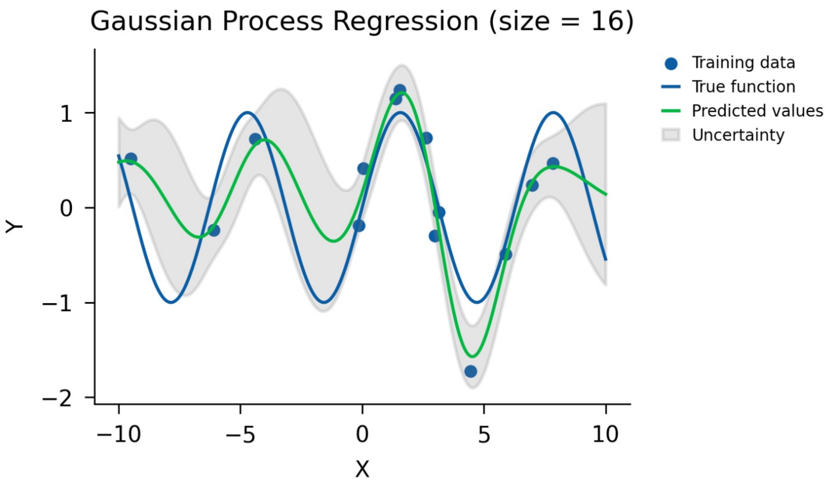


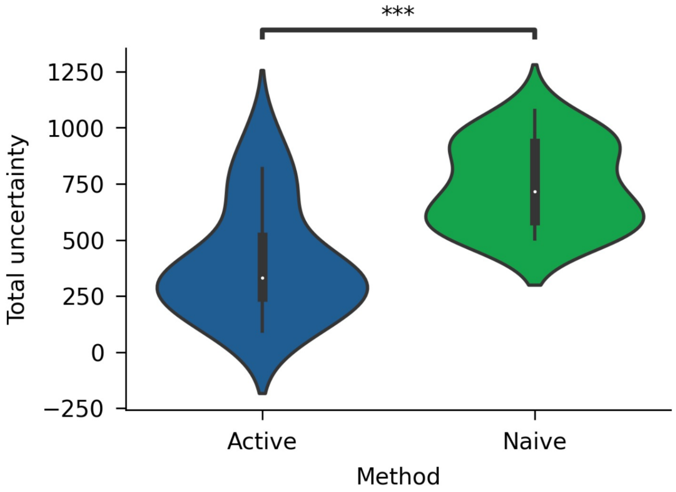
Hie et al, *Cell Systems* (2020)

Active learning using GP



Active learning using GP reduces overall uncertainty





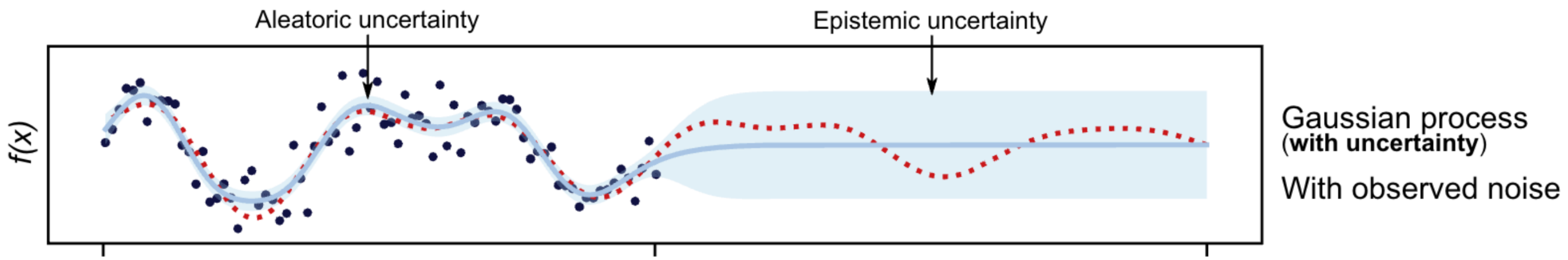
Random

Experiments

Gaussian Processes in Julia

Feature	AGP.jl	Stheno.jl	GP.jl
Sparse GP	✓	✖	✓
Custom prior Mean	✓	✓	✓
Hyperparam. Opt.	✓	?	✓
MultiOutput	✓	✓	✖
Online	✓	✖	✖

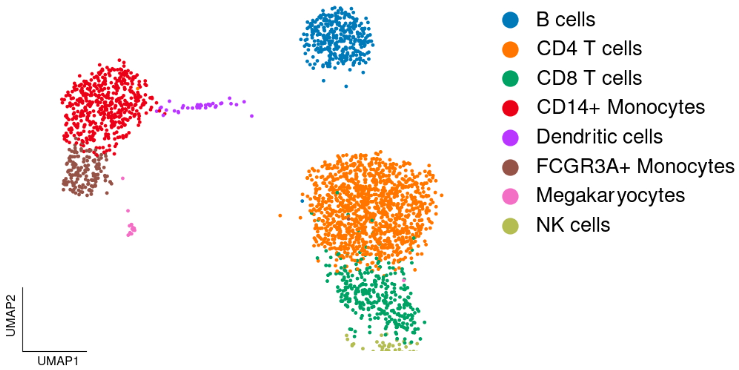
Types of uncertainty



SMILLIE LAB

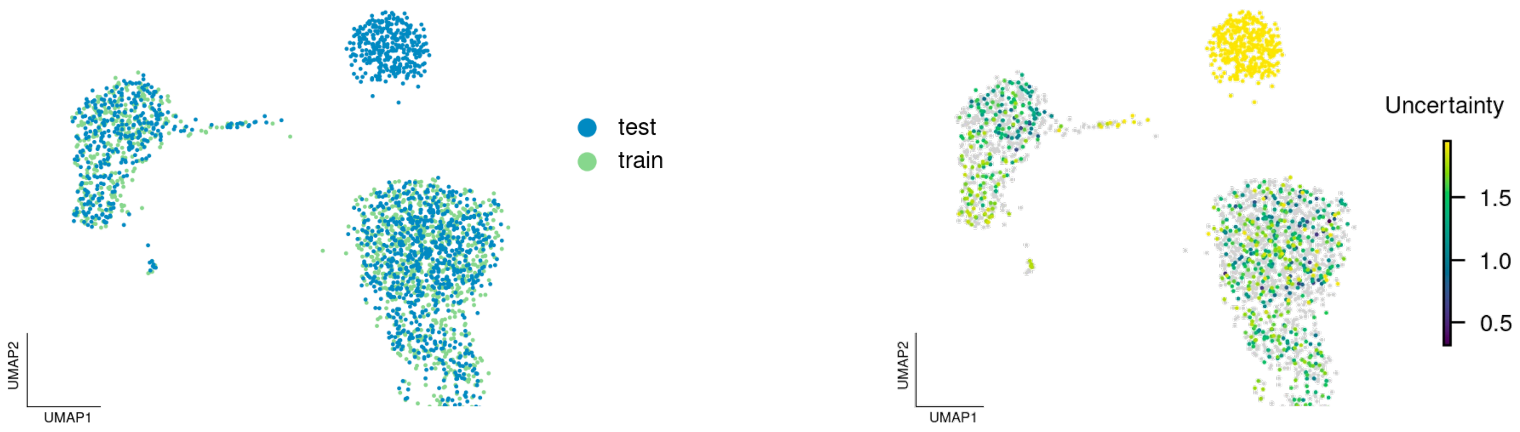
Application 1: Single cell RNA-seq

- Captures cell-type transcriptomic heterogeneity in health and disease
- Quantifying uncertainty may
  - Identify batch effects
  - Identify rare condition-specific cell populations



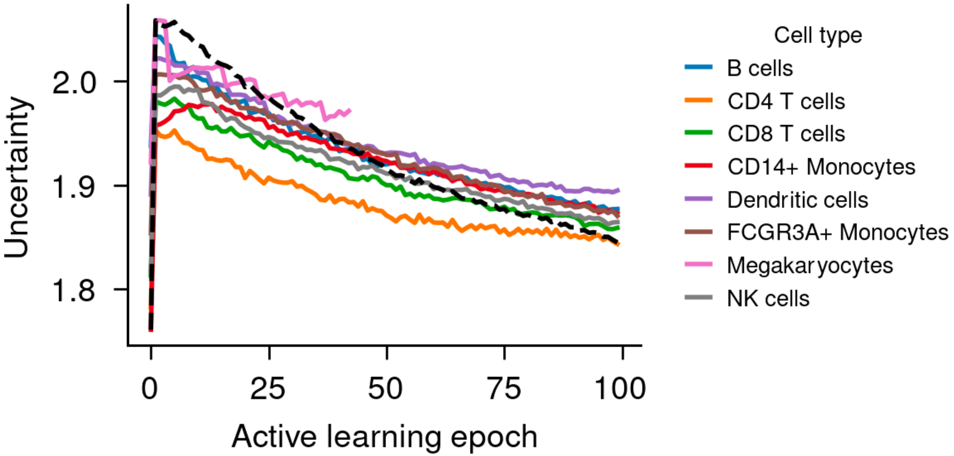
Gaussian Processes reveals epistemic uncertainty

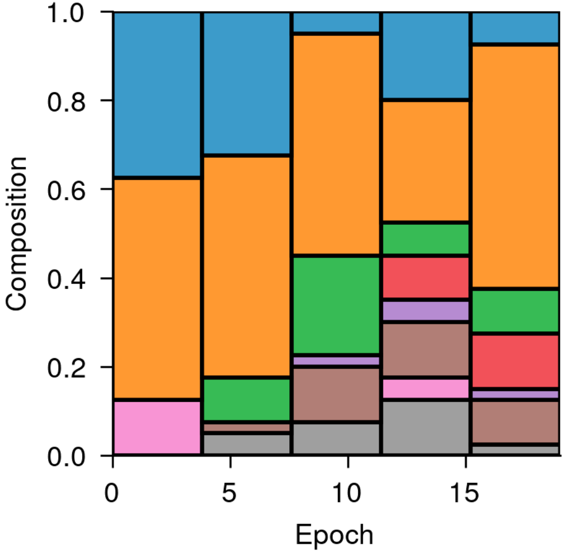
- Due to data hold out in training set (B cells)



Active learning

- Nominate new cells for acquisition
  - Dominated by held out cells

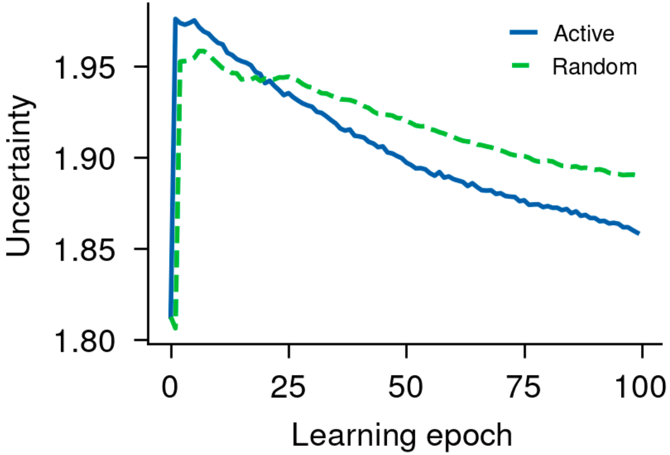




Active learning

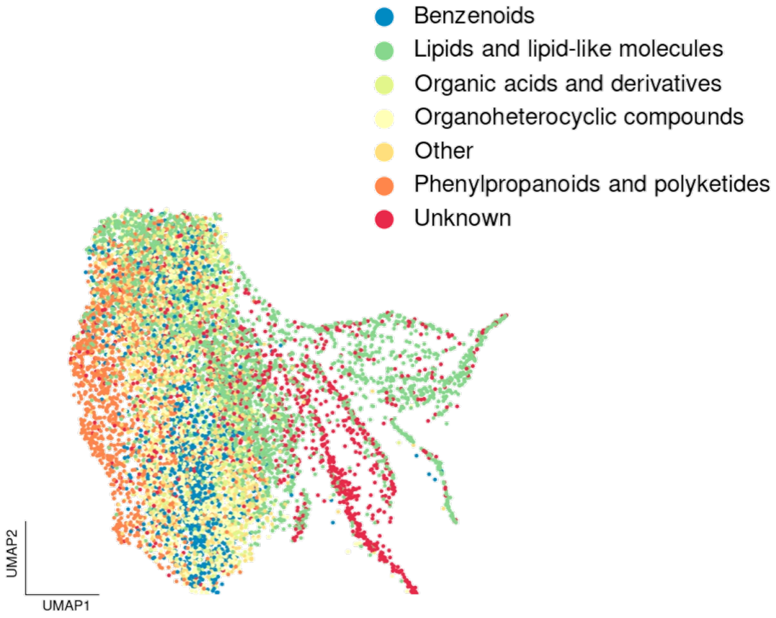
- Active learning out-performs random

- In practice can guide new sample acquisition for expensive/hard-to obtain samples, e.g.
  - Perturb-seq
  - Tumor core-needle biopsy



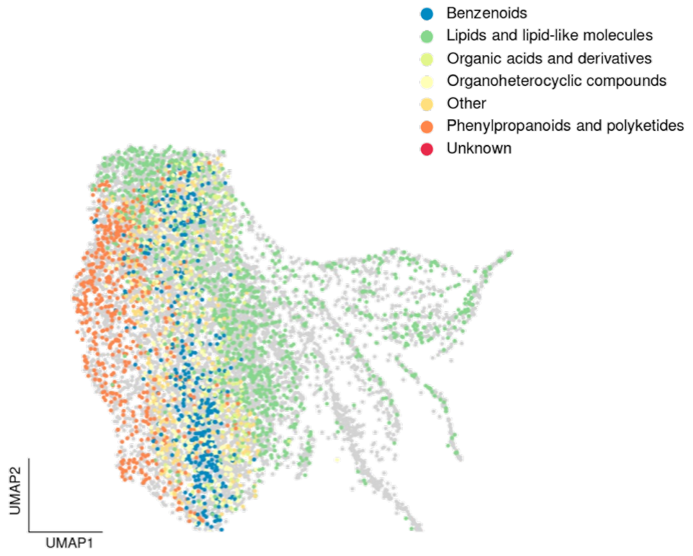
Application 2: Metabolomics

- Notoriously difficult in machine learning
- 90% compounds unannotated
- Motivating problem:
  - Quantify uncertainty to nominate low confidence compounds
  - Generate reference spectra experimentally subjected to experimental budget



Gaussian Processes reveals aleatoric uncertainty

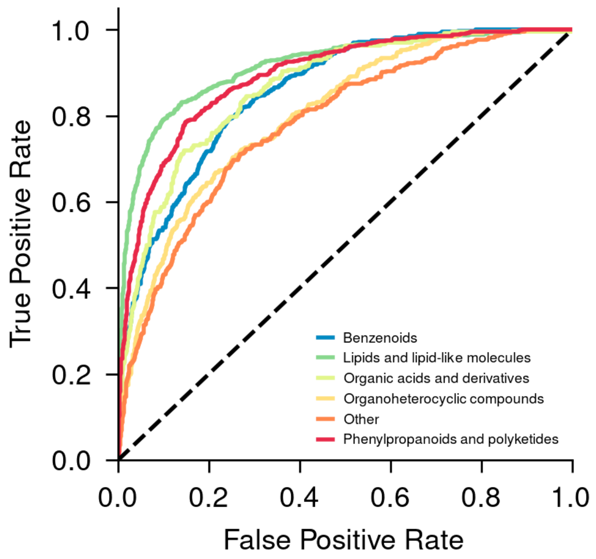
- Due to high intrinsic noise in data and/or poor data labelling



Application of Gaussian Processes

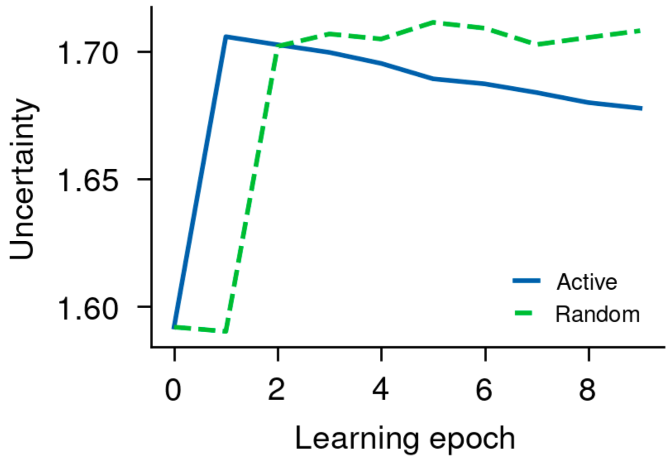
- Good classification performance despite low data quality

- Comparable with SVM
  - Standard practice in metabolomics



Application of Gaussian Processes

- Good classification performance despite low data quality
- Comparable with SVM
  - Standard practice in metabolomics
- Computationally expensive and numerically unstable
- Active learning out-perform random



Future work

- Formulate as Bayesian optimization
- In-domain kernel design
- Explore different native Julia implementations
- Better delineate uncertainty in active learning
  - epistemic uncertainty vs aleatoric uncertainty

Extra Slides