Some applications of machine learning to medicine

Pratik Worah

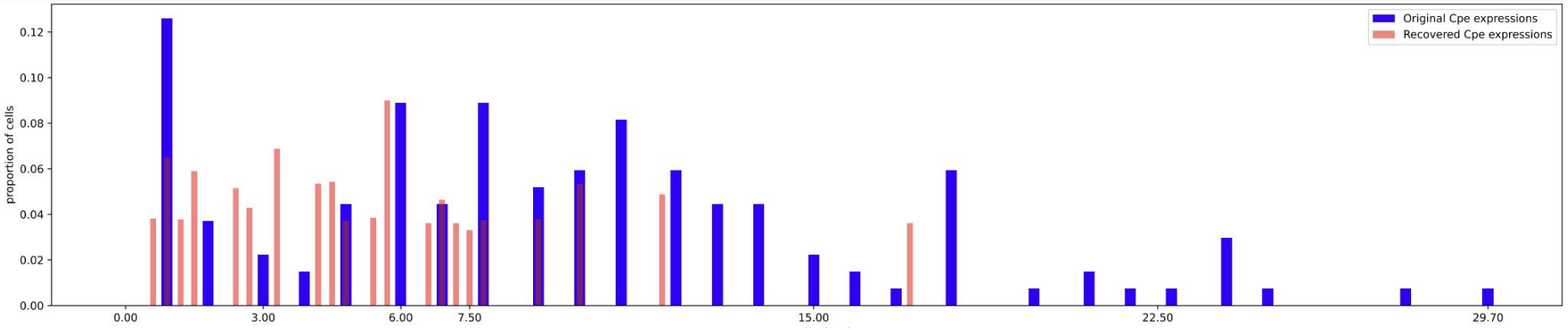
Broad themes:

- Based on abundance of data:
  - Lots of data available: neural networks.
    - Examples: predicting enzyme inhibitors using neural nets (NNs), cell annotation for flow-cytometry
  - Little training data available: algorithms like compressed sensing
    - Examples: designing genomic tests for “rare” diseases, recovering the interactome
- Based on application:
  - Clinical / directly patient related:
    - Examples: genomic tests, recovering single cell data …
  - Pharmaceutical:
    - Examples: drug discovery, flow cytometry …

Genomics

1. **Approximately recovering single cell distribution from aggregate measurements**
    a. Recovering scRNA-seq data from aggregate level hybridization assays.
    b. [W1] Recovering approximate single cell distribution from aggregate measurements, 2023
2. Recovering the interactome from scRNA-seq data
    a. Model gene translation as a Markov Chains (MC) and recover the transition matrix.
    b. [W2] Recovering a sparse linear dynamical system, 2023
3. Comparing recovered interactome with a reference
    a. Equivalent to hypothesis testing for MCs, but in high dimension with few cell samples.
    b. [SW] Designing optimal tests for slow converging Markov chains. Joint with C.Stein. IMLH (ICML), 2023

Recovering scRNA-seq from aggregate measurements

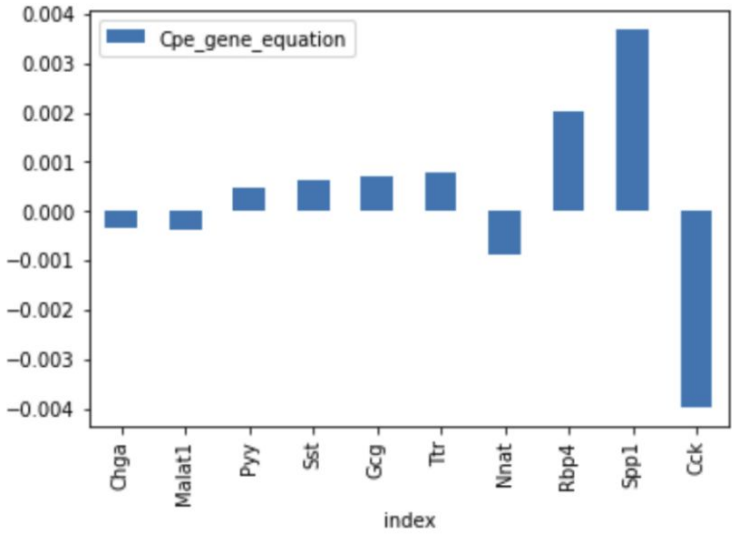


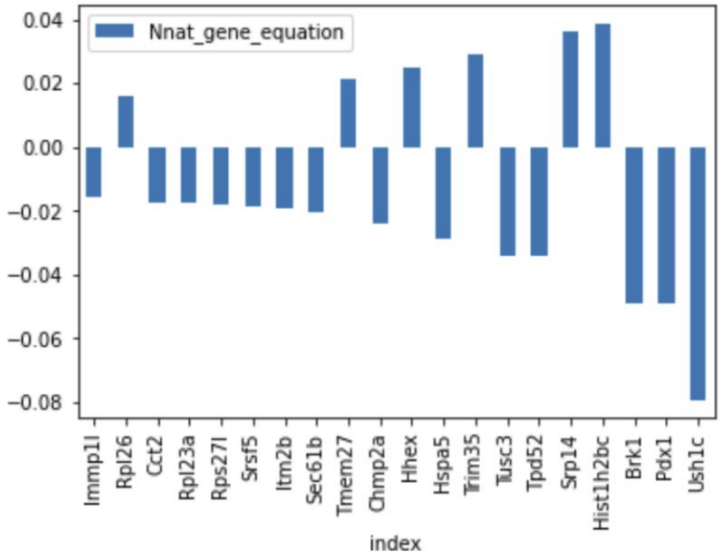
- Carboxypeptidase E (Cpe): gene for enzyme involved in synthesis of insulin, glucagon etc
- Sample: 4000 cells from murine pancreas [Bastidas-Ponce et al., Development 2019]
- Blue bars are recovered from data using RNA-seq type assays, while Red bars are actual scRNA-seq data

Genomics

1. Approximately recovering single cell distribution from aggregate measurements
    a. Recovering scRNA-seq data from aggregate level hybridization assays.
    b. [W1] Recovering approximate single cell distribution from aggregate measurements, 2023
2. **Recovering the interactome from scRNA-seq data**
    a. Model gene translation as a Markov Chains (MC) and recover the transition matrix.
    b. [W2] Recovering a sparse linear dynamical system, 2023
3. Comparing recovered interactome with a reference
    a. Equivalent to hypothesis testing for MCs, but in high dimension with few cell samples.
    b. [SW] Designing optimal tests for slow converging Markov chains. Joint with C.Stein. IMLH (ICML), 2023

Genes that directly influence other genes



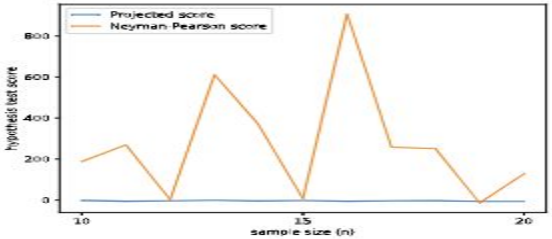


- Y-axis: (inverted) influence of gene on secretion of (1) Cpe, (2) Nnat
- Data from [Bastidas-Ponce et al. Devlopment 2019] and [Bergen et al., Nature 2020]

Genomics

1. Approximately recovering single cell distribution from aggregate measurements
    a. Recovering scRNA-seq data from aggregate level hybridization assays.
    b. [W1] Recovering approximate single cell distribution from aggregate measurements, 2023
2. Recovering the interactome from scRNA-seq data
    a. Model gene translation as a Markov Chains (MC) and recover the transition matrix.
    b. [W2] Recovering a sparse linear dynamical system, 2023
3. **Comparing recovered interactome with a reference**
    a. Equivalent to hypothesis testing for MCs, but in high dimension with few cell samples.
    b. [SW] Designing optimal tests for slow converging Markov chains. Joint with C.Stein. IMLH (ICML), 2023

Comparing two interactomes



(a) Neyman-Pearson scores from small sample (projected) and unprojected tests

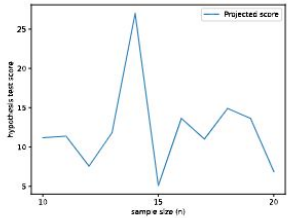


Figure 2. High positive values of the score suggest that the empirical distribution is from beta cells (the ground truth). Here we can distinguish between the two cell types (beta cells vs ductal cells) using just 10-20 cell samples of Cpe expressions.

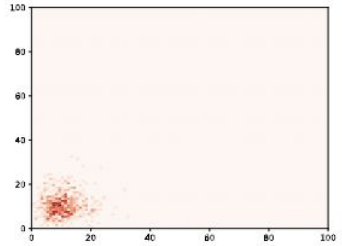


Figure 3. Heat map for the transition matrix of Cpe expression in beta cells. Darker areas indicate higher values.

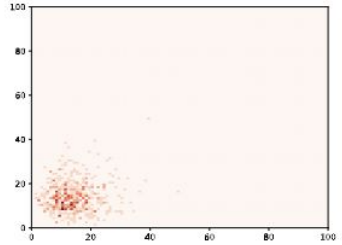


Figure 4. Heat map for the transition matrix of Cpe expression in alpha cells. Darker areas indicate higher values.

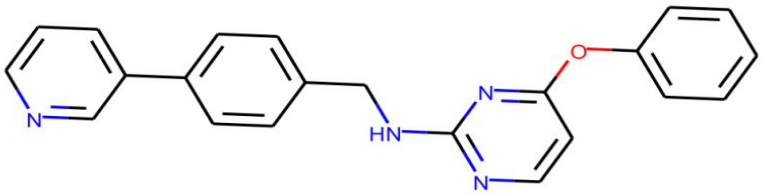
Genomics

1. Approximately recovering single cell distribution from aggregate measurements
    a. Recovering scRNA-seq data from aggregate level hybridization assays.
    b. [W1] Recovering approximate single cell distribution from aggregate measurements, 2023
2. Recovering the interactome from scRNA-seq data
    a. Model gene translation as a Markov Chains (MC) and recover the transition matrix.
    b. [W2] Recovering a sparse linear dynamical system, 2023
3. Comparing recovered interactome with a reference
    a. Equivalent to hypothesis testing for MCs, but in high dimension with few cell samples.
    b. [SW] Designing optimal tests for slow converging Markov chains. Joint with C.Stein. IMLH (ICML), 2023

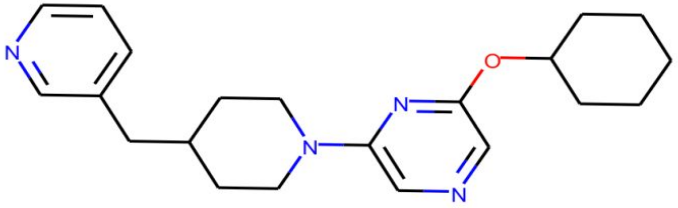
Pharmaceutical

1. **Train NNs to learn and predict selective and potent enzyme inhibitors**
    a. Example: predict small molecule inhibitors to MNK2 that are non-inhibitors for MNK1.
    b. [TW] Enhancing small molecule selectivity using Wasserstein distance based reweighing. Joint with W.Torng, 2022
2. Train NNs to annotate cells in a stream
    a. Example: in flow cytometry, the goal is to classify cells based on fluorescence
    b. [FJMW] Learning rate under distribution shift (with an application to flow cytometry). Joint with M.Fahrbach, A.Javanmard, V.Mirrokni. ICML, 2023

Inhibitors of MNK2 but not MNK1



C(NC=1N=CC=C(OC=2C=CC=CC2)N1)C=3C=CC(=CC3)C=4C=CC=NC4



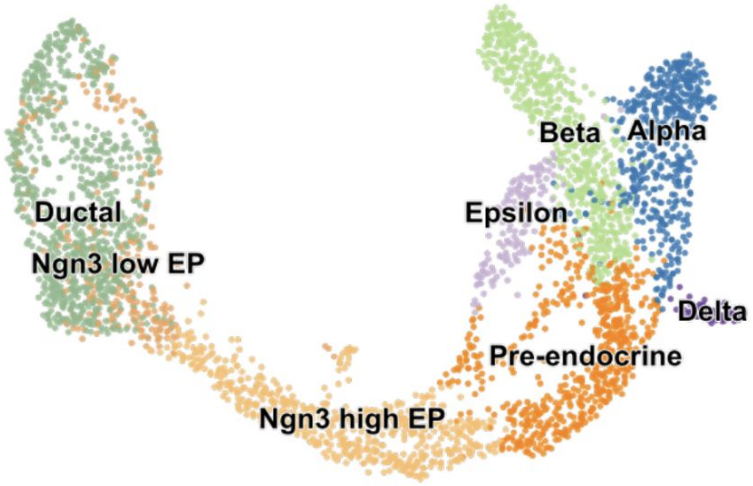
C(C1CCN(CC1)C=2C=NC=C(OC3CCCCC3)N2)C=4C=CC=NC4

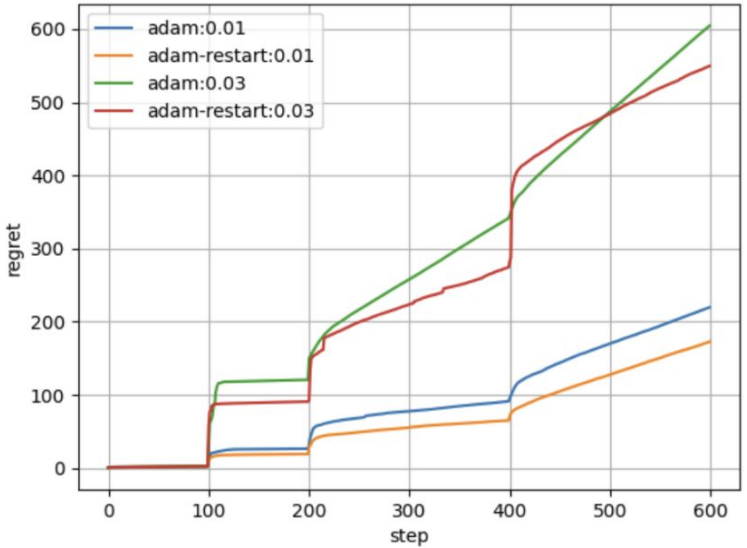
- 5 out of 100 compounds are expected to be inhibitors for MNK2 but not MNK1 at 10uM
- Can be tried for other enzymes, more than one pair of enzymes etc…

Pharmaceutical

1. Train NNs to learn and predict selective and potent enzyme inhibitors
    a. Example: predict small molecule inhibitors to MNK2 that are non-inhibitors for MNK1.
    b. [TW] Enhancing small molecule selectivity using Wasserstein distance based reweighing. Joint with W.Torng, 2022
2. **Train NNs to annotate cells in a stream**
    a. Example: in flow cytometry, the goal is to classify cells based on fluorescence
    b. [FJMW] Learning rate under distribution shift (with an application to flow cytometry). Joint with M.Fahrbach, A.Javanmard, V.Mirrokni. ICML, 2023

High throughput flow cytometry





Accuracy of cell annotation using NNs and **unlabelled** flow-cytometry data

Pharmaceutical

1. Train NNs to learn and predict selective and potent enzyme inhibitors
    a. Example: predict small molecule inhibitors to MNK2 that are non-inhibitors for MNK1.
    b. [TW] Enhancing small molecule selectivity using Wasserstein distance based reweighing. Joint with W.Torng, 2022
2. Train NNs to annotate cells in a stream
    a. Example: in flow cytometry, the goal is to classify cells based on fluorescence
    b. [FJMW] Learning rate under distribution shift (with an application to flow cytometry). Joint with M.Fahrbach, A.Javanmard, V.Mirrokni. ICML, 2023

References

- Papers/preprints by PW available at:
  - http://people.cs.uchicago.edu/~pworah/compbio.html
- [Bergen et al.]: Generalizing rna velocity to transient cell states through dynamical modeling. Nature Biotechnology, 2020
- [Bastidas-Ponce et al]: Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. Development, 2019.