WYSS INSTITUTE

Embedding Models

For Data Digest, Discovery and Design

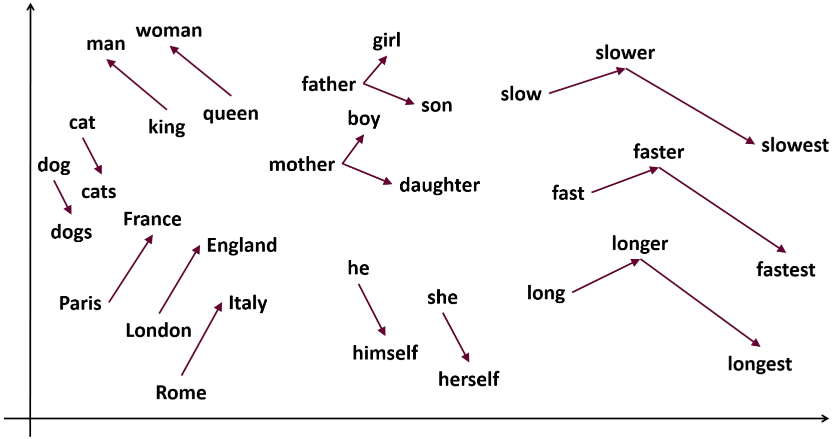
Sahil Loomba

Acknowledgements
Mohammed Eslami, Hugh Haddox, Luki Goldschmidt, Hamed Eramian

DARPA SD2E

Embedding Models

- Encode arbitrary entities into a d-dimensional vector space
    - Closed under simple vector algebra
- Incorporate vast amounts of prior “unsupervised” knowledge in databases
- Constrain complex models; make up for “supervised” data in few-sample and high-dimensional settings
- Enable downstream predictive machine learning models by providing “resolved” feature spaces for discovery and design of biological parts



word2vec (Mikolov et al. 2013)

Embedding Models

Embedding biological parts important to SD2 program, such as:

Circuits (as ontologies): `TetR “negatively regulate” TetA`

Proteins (as sequences): `VPLLGLY...`

Genes (as sequences): `AATCGGTA...`

Embedding Models

Embedding biological parts important to SD2 program, such as:

Circuits (as ontologies): `TetR “negatively regulate” TetA`

ecoli2vec

yeast2vec

Proteins (as sequences): `VPLLGLY...`

prot2vec

Genes (as sequences): `AATCGGTA...`

ribo2vec

(Aside on Visualizing High-Dimensional Spaces)

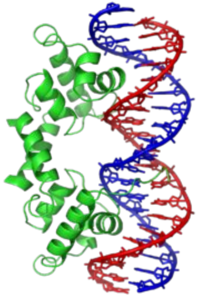
You’ll be seeing many plots of 10-100 dimensional spaces!

We’ll employ two useful methods that reduce dimensionality to 2-3:

1. Principal Component Analysis (PCA): captures dimensions of maximum variance in the original feature space
2. t-distributed Stochastic Neighbor Embedding (t-SNE): captures local neighborhood information in the original feature space

ecoli2vec

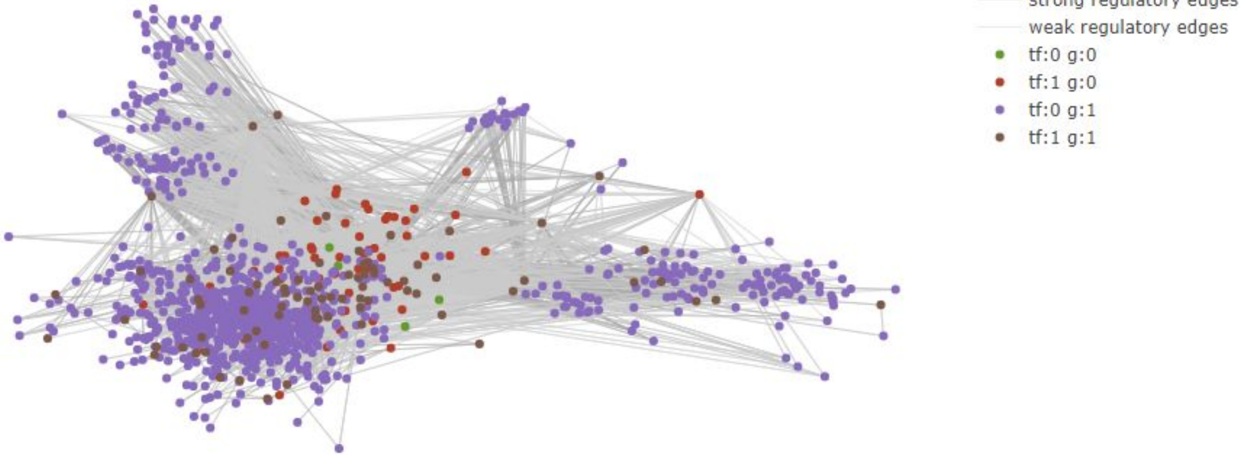
- Intuition: learn key biomolecular interactions in E. coli as a host
- Input: strong regulatory relationships across 147 TFs and 1033 genes in E. coli from RegulonDB
    - Relationships are signed, directed, and symmetrized
        - V1: <tf> “regulates” <gene>; <tf> “positively regulates” <gene>
        - V2: <seq> “sequence of” <entity>
        - V3: <entity> “binds to” <entity>



ecoli2vec

- Intuition: learn key biomolecular interactions in E. coli as a host
- Input: relationships across 147 TFs & 1033 genes in E. coli from RegulonDB
- Output: 100-dimensional embeddings of TFs, Genes, Relationships

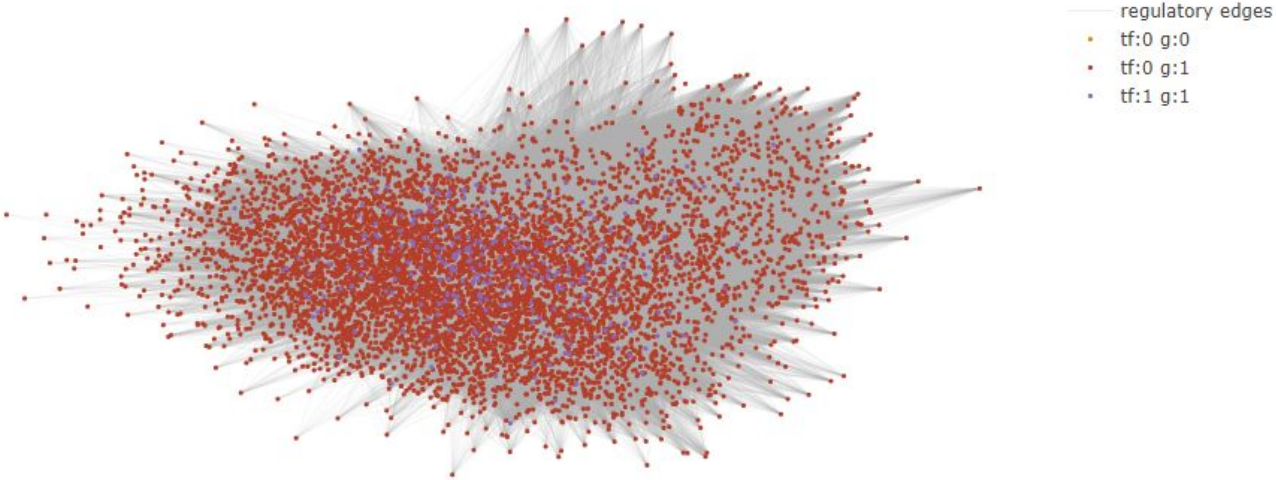
2D pca plot of ecoli2vec fwd embeddings



yeast2vec

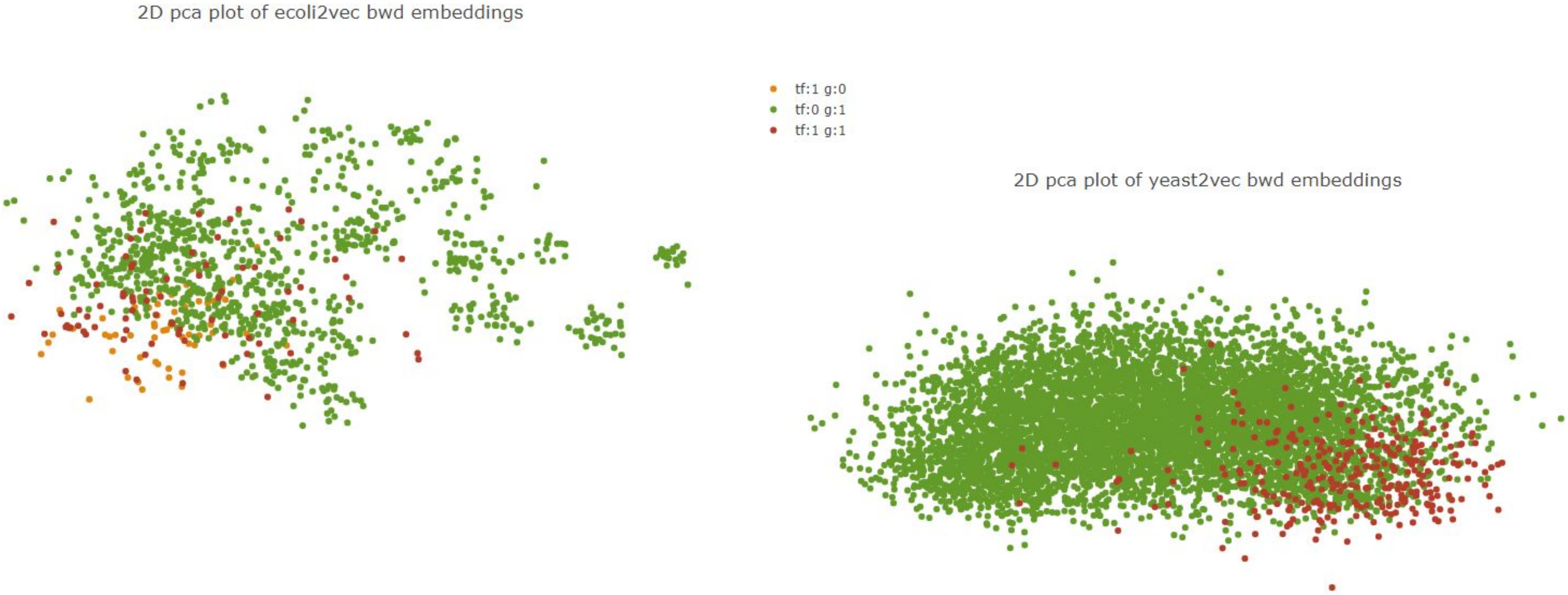
- Intuition: learn key biomolecular interactions in Yeast as a host
- Input: relationships across 307 TFs & 6725 genes in Yeast from Yeastract
- Output: 100-dimensional embeddings of TFs, Genes, Relationships

2D pca plot of yeast2vec fwd embeddings



ecoli2vec & yeast2vec

PCA on embedding shows “TFness” being captured

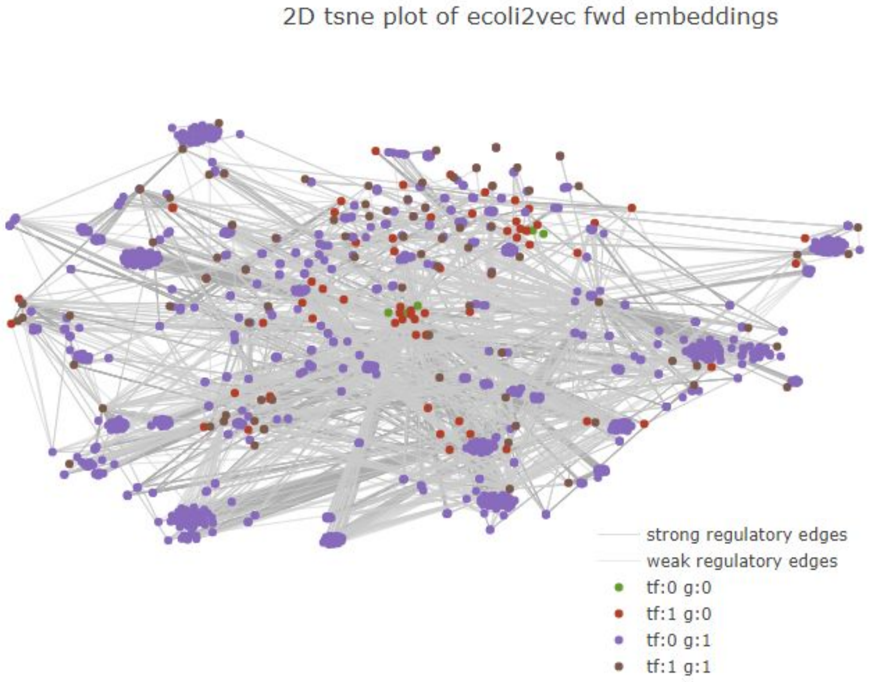


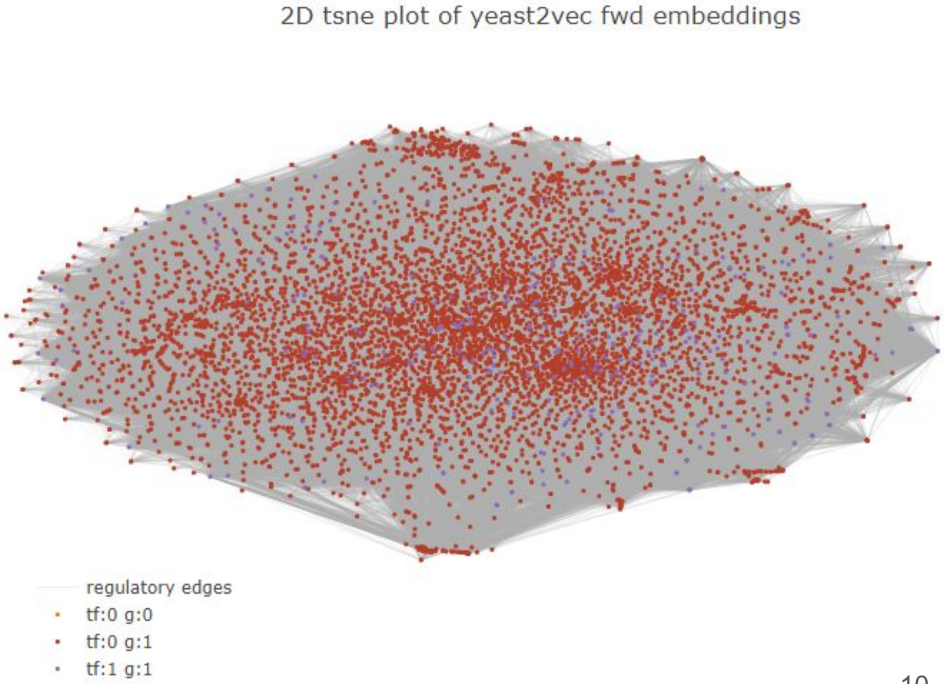
2D pca plot of ecoli2vec bwd embeddings

2D pca plot of yeast2vec bwd embeddings

ecoli2vec & yeast2vec

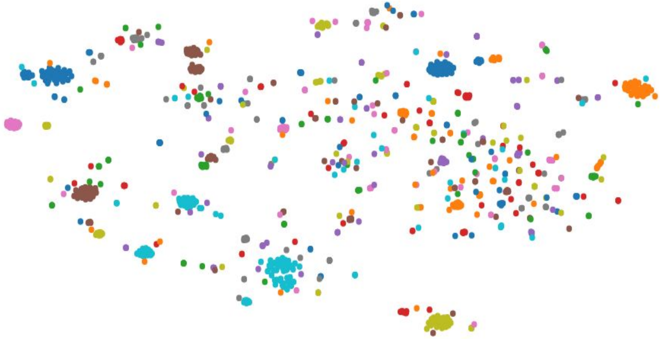
t-SNE on embedding shows “modularity” being captured


2D tsne plot of ecoli2vec fwd embeddings

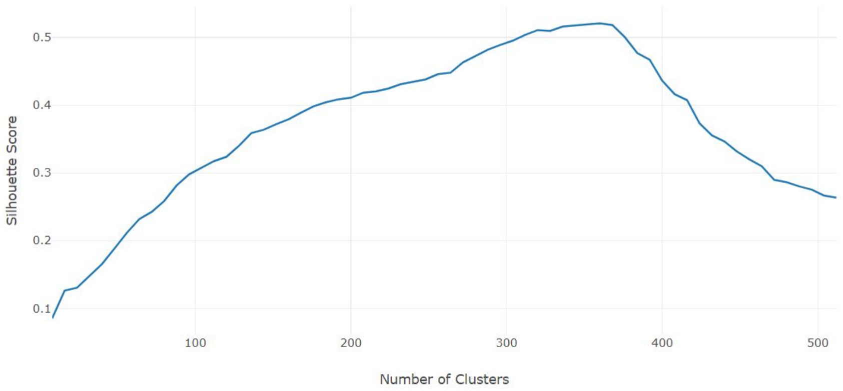

2D tsne plot of yeast2vec fwd embeddings

ecoli2vec | Data Digest QC for SD2 Program

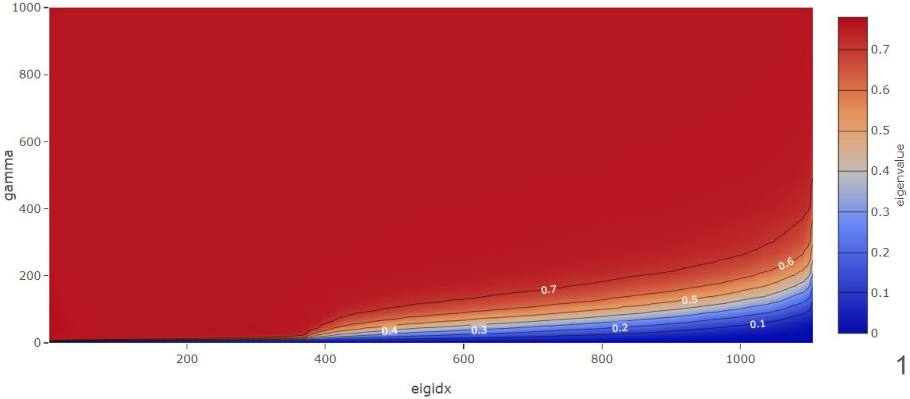
- Embeddings "condense" knowledge, if used appropriately can compensate for data
- Cluster genes in ecoli2vec space: discovered 360 "genetic modules" that might be co-dependent in the expression space



Silhouette Scores with varying Number of Clusters for "Complete" Agglomerative Clustering

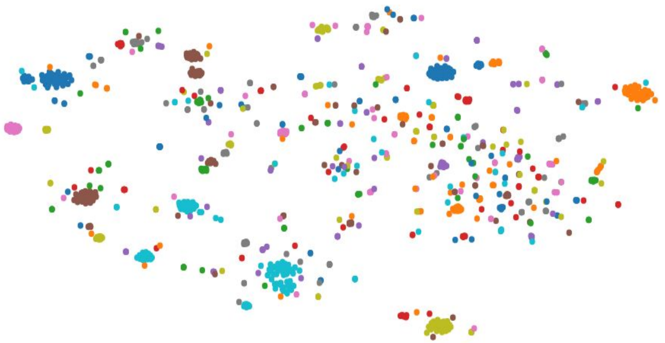


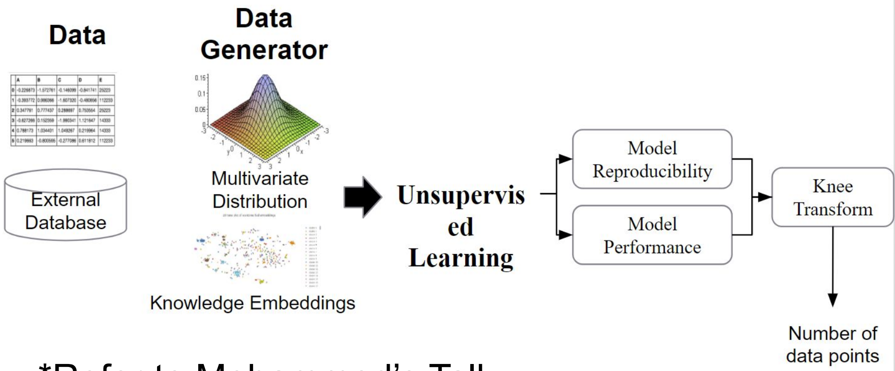
Eigenspectral Curves of Laplacian Matrix with varying Perplexity



ecoli2vec | Data Digest QC for SD2 Program

- Embeddings “condense” knowledge, if used appropriately can compensate for data
- Cluster genes in ecoli2vec space: discovered 360 “genetic modules” that might be co-dependent in the expression space
- Feed dependencies to Mohammed’s Power Analysis pipeline to estimate required number of replicates
- Mohammed’s analysis reveals: use of embeddings seem to impose conditions that indeed require fewer data replicates on the Q0 Rule30 CP





Data

External Database

Data Generator

Multivariate Distribution

Knowledge Embeddings

Unsupervised Learning

Model Reproducibility

Model Performance

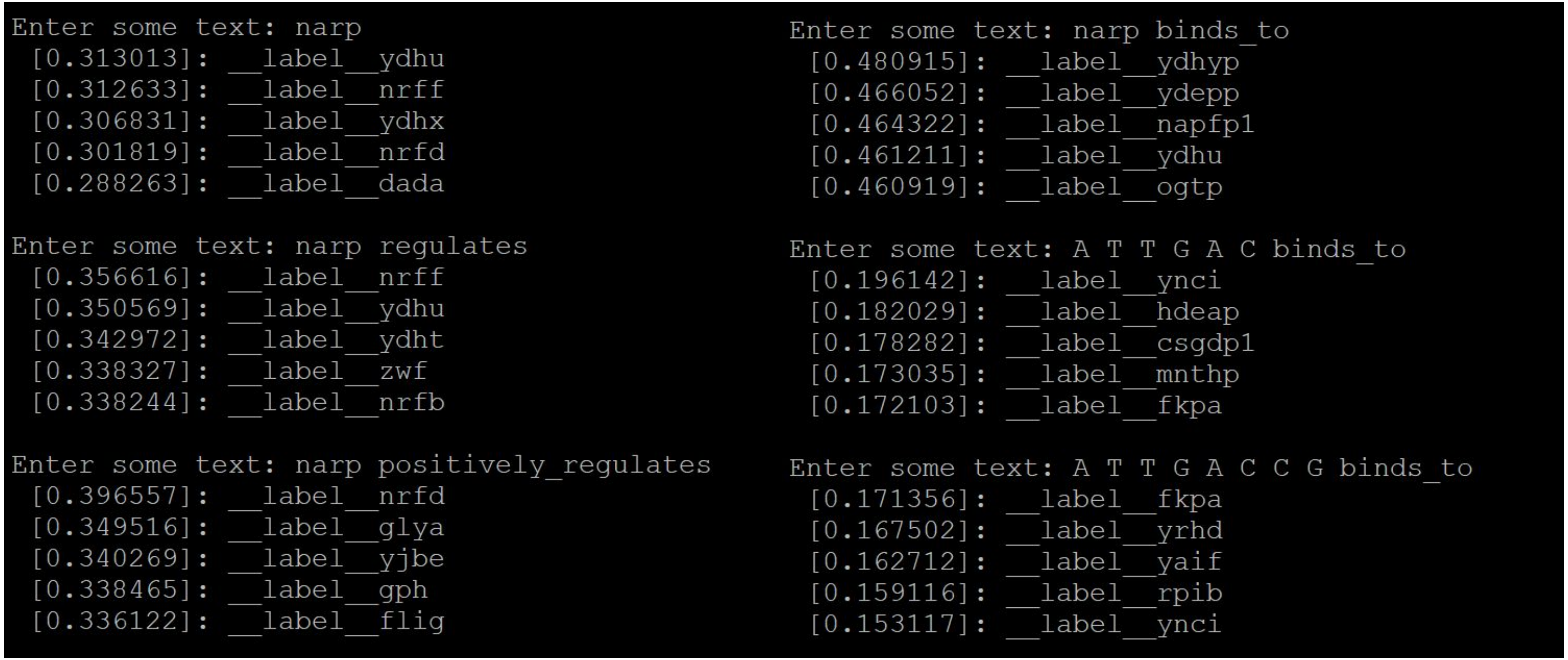
Knee Transform

Number of data points

*Refer to Mohammed’s Talk

host2vec | Novel Host Chassis Challenge Problem

Issue queries to “discover” interactions; such as “narp regulates ?”



```
Enter some text: narp
 [0.313013]:  __label__ydhu
 [0.312633]:  __label__nrff
 [0.306831]:  __label__ydhx
 [0.301819]:  __label__nrfd
 [0.288263]:  __label__dada

Enter some text: narp regulates
 [0.356616]:  __label__nrff
 [0.350569]:  __label__ydhu
 [0.342972]:  __label__ydht
 [0.338327]:  __label__zwf
 [0.338244]:  __label__nrfb

Enter some text: narp positively_regulates
 [0.396557]:  __label__nrfd
 [0.349516]:  __label__glya
 [0.340269]:  __label__yjbe
 [0.338465]:  __label__gph
 [0.336122]:  __label__flig
```

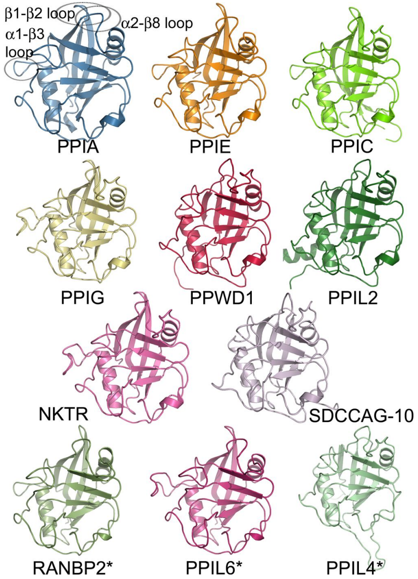
```
Enter some text: narp binds_to
 [0.480915]:  __label__ydhyp
 [0.466052]:  __label__ydepp
 [0.464322]:  __label__napfp1
 [0.461211]:  __label__ydhu
 [0.460919]:  __label__ogtp

Enter some text: A T T G A C binds_to
 [0.196142]:  __label__ynci
 [0.182029]:  __label__hdeap
 [0.178282]:  __label__csgdp1
 [0.173035]:  __label__mnthp
 [0.172103]:  __label__fkpa

Enter some text: A T T G A C C G binds_to
 [0.171356]:  __label__fkpa
 [0.167502]:  __label__yrhd
 [0.162712]:  __label__yaif
 [0.159116]:  __label__rpib
 [0.153117]:  __label__ynci
```

prot2vec

- Intuition: learn the space of “natural” proteins
- Input: 93,588 amino acid sequences across the Human Proteome on UniProt
- Output: 100-dimensional embeddings of arbitrary amino acid sequences

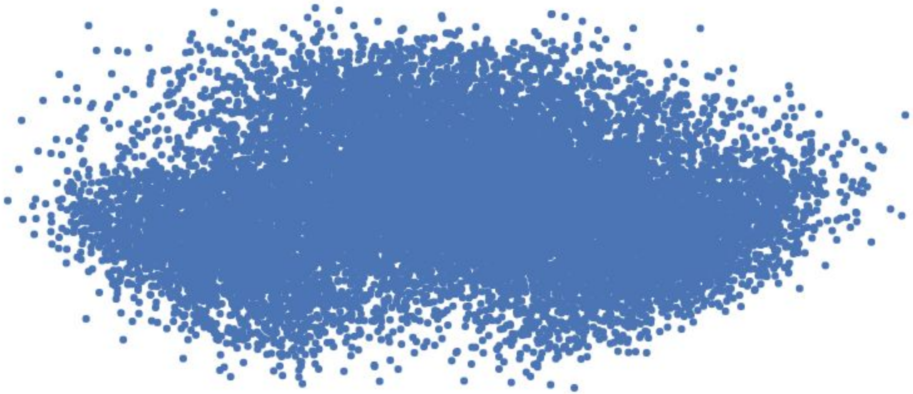
prot2vec

- Intuition: learn the space of “natural” proteins
- Input: 93,588 amino acid sequences across the Human Proteome on UniProt
- Output: 100-dimensional embeddings of arbitrary amino acid sequences
- Test on new data: Protein Family Prediction
  - 324,017 sequences from SwissProt across 7027 families
  - High accuracy of 0.732 on the simplest 1-nearest-neighbor classifier

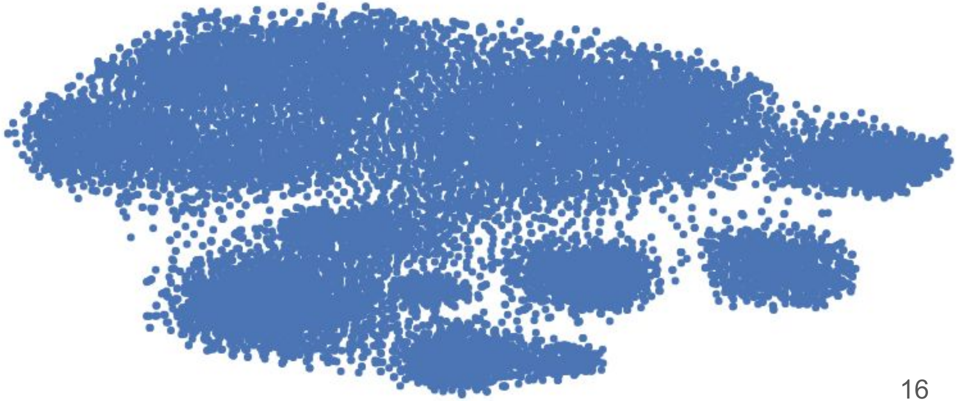


prot2vec | Protein Stability Challenge Problem

- Intuition: query learnt pro2vec space for embeddings of ProtStab sequences
- Test Input to pro2vec: 16,174 design sequences from IPD Database for SD2
- Output: 100-dimensional embeddings of design sequences in ProtStab CP



2D pca plot of prot2vec embeddings of Protein Stability Dataset

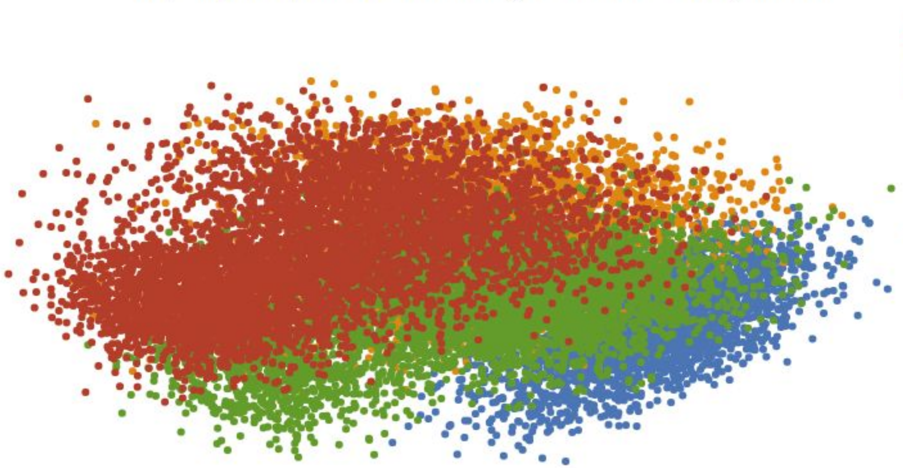


2D tsne plot of prot2vec embeddings of Protein Stability Dataset

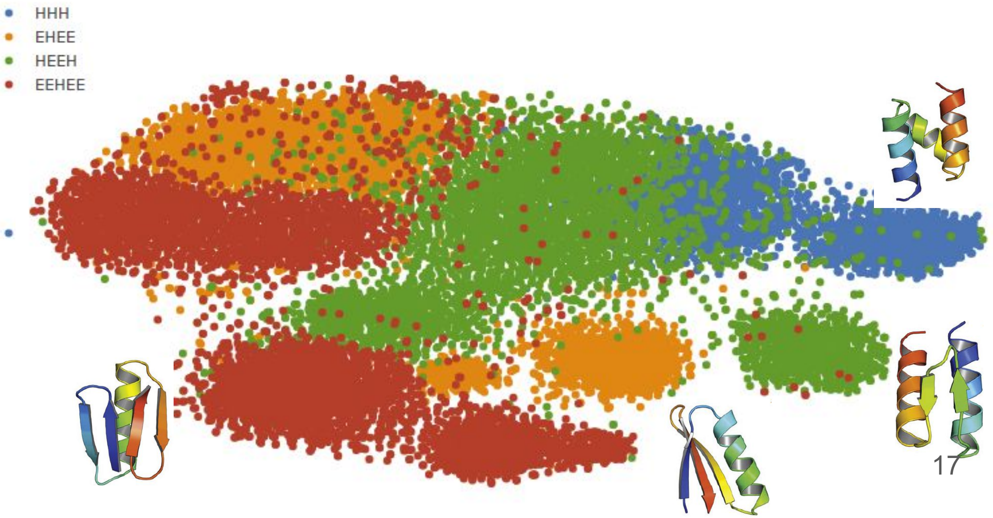
prot2vec | Protein Stability Challenge Problem

- Intuition: query learnt pro2vec space for embeddings of ProtStab sequences
- Test Input to pro2vec: 16,174 design sequences from IPD Database for SD2
- Output: 100-dimensional embeddings of design sequences in ProtStab CP
- Interpretation: designs cluster by protein topology*

2D pca plot of prot2vec embeddings of Protein Stability Dataset



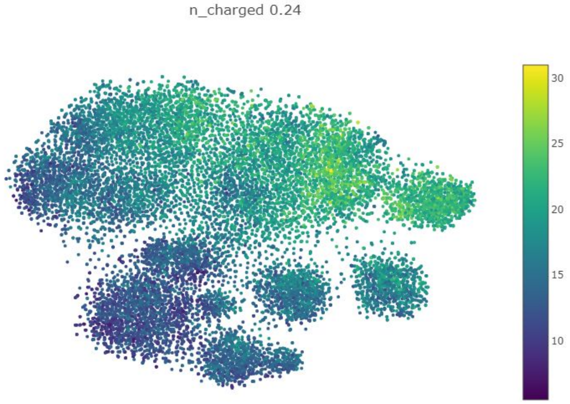
2D tsne plot of prot2vec embeddings of Protein Stability Dataset



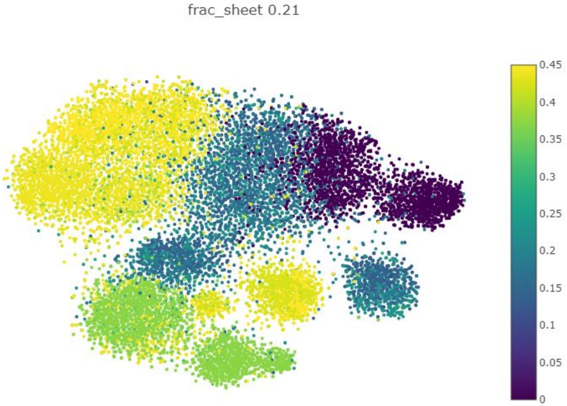
*remember, prot2vec never saw any ProtStab data!

prot2vec | Protein Stability Challenge Problem

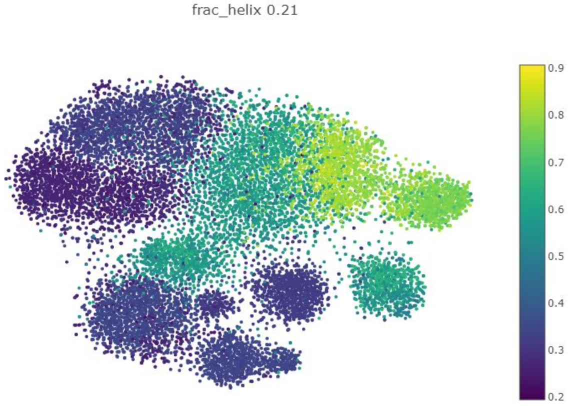
Several protein design metrics correlate with prot2vec embeddings, some of which also happen to be upranked metrics in Rocklin et al. (2017)



Total Charge



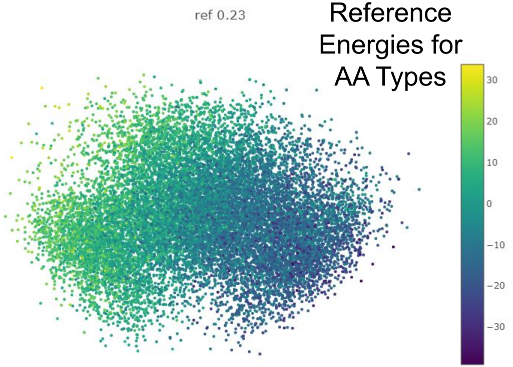
Fraction of Sheet



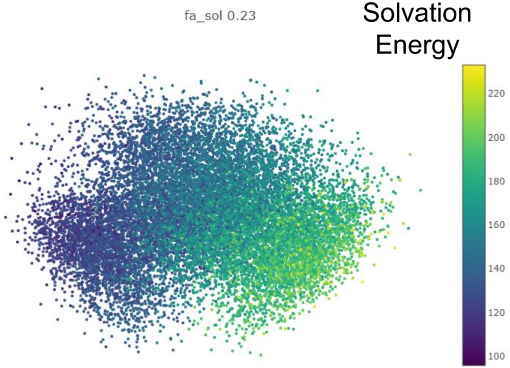
Fraction of Helix

(t-SNE)

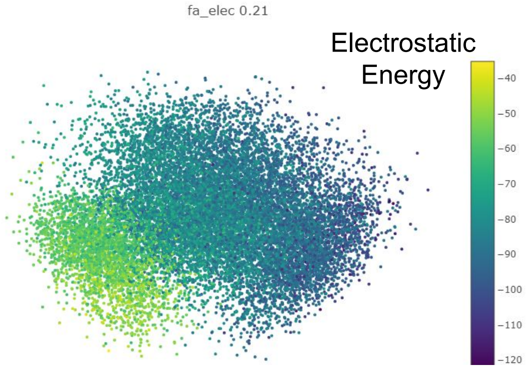
prot2vec | Protein Stability Challenge Problem



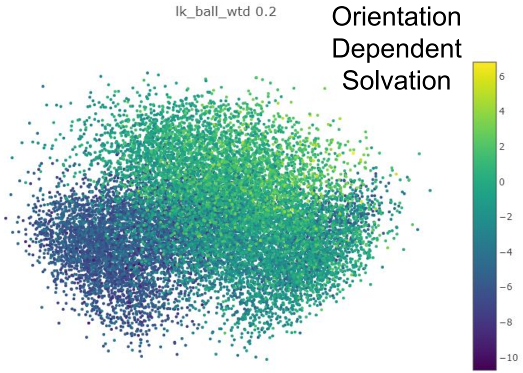
Reference Energies for AA Types



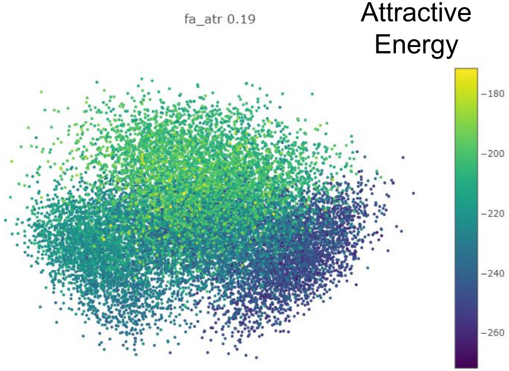
Solvation Energy



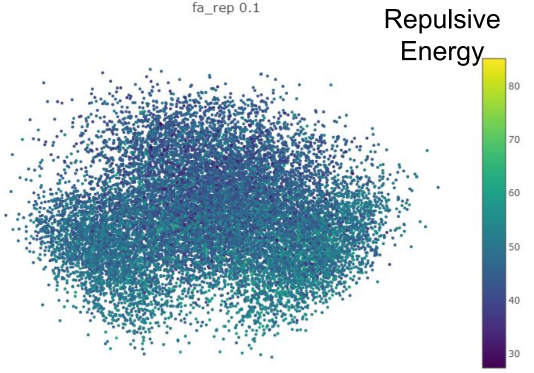
Electrostatic Energy



Orientation Dependent Solvation



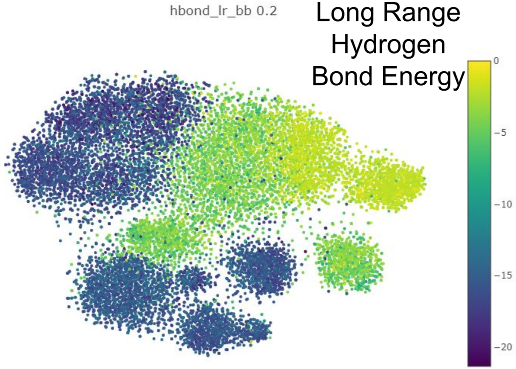
Attractive Energy



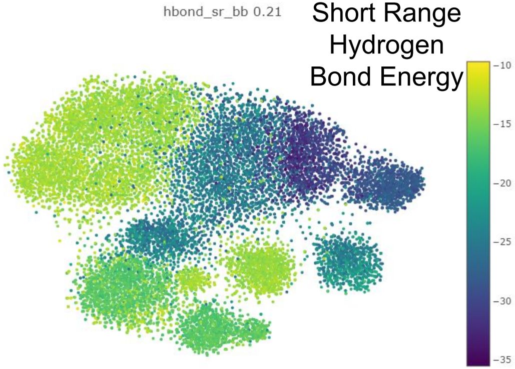
Repulsive Energy

(PCA)

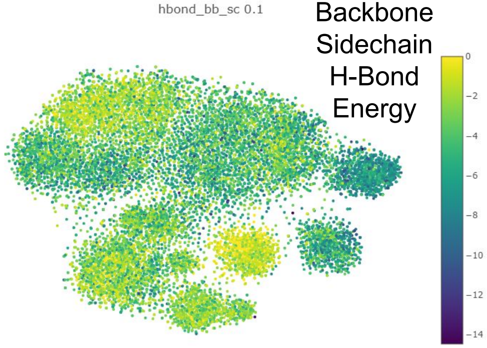
prot2vec | Protein Stability Challenge Problem



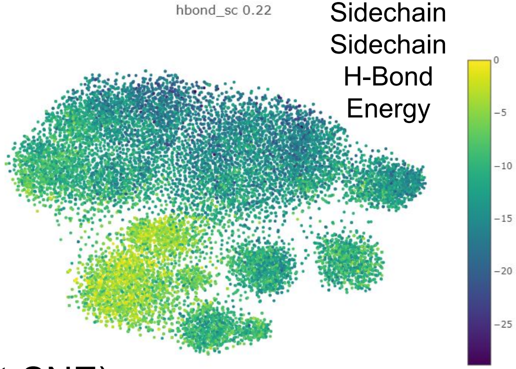
Long Range Hydrogen Bond Energy



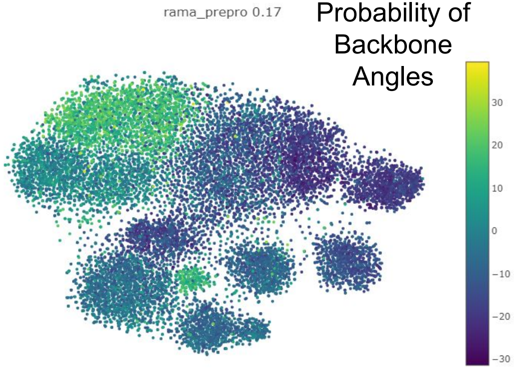
Short Range Hydrogen Bond Energy



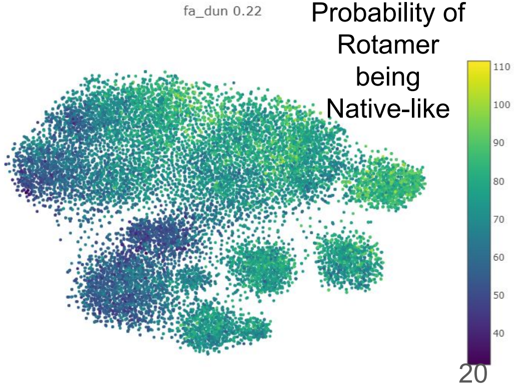
Backbone Sidechain H-Bond Energy



Sidechain Sidechain H-Bond Energy



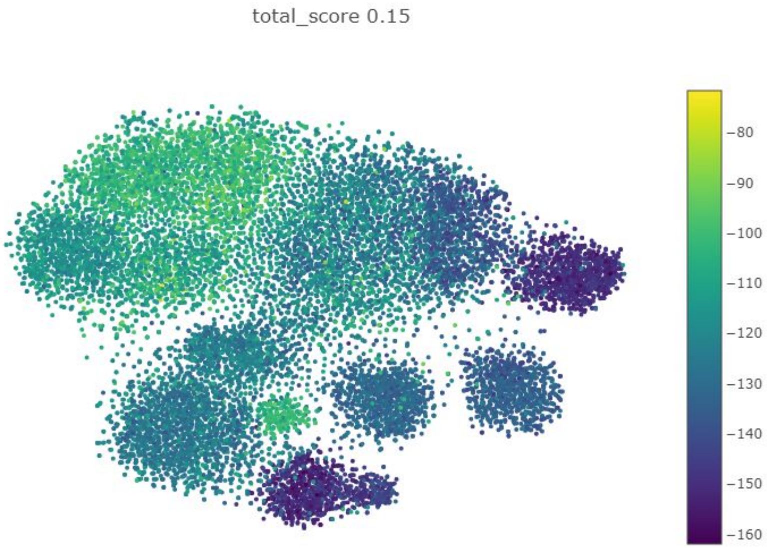
Probability of Backbone Angles



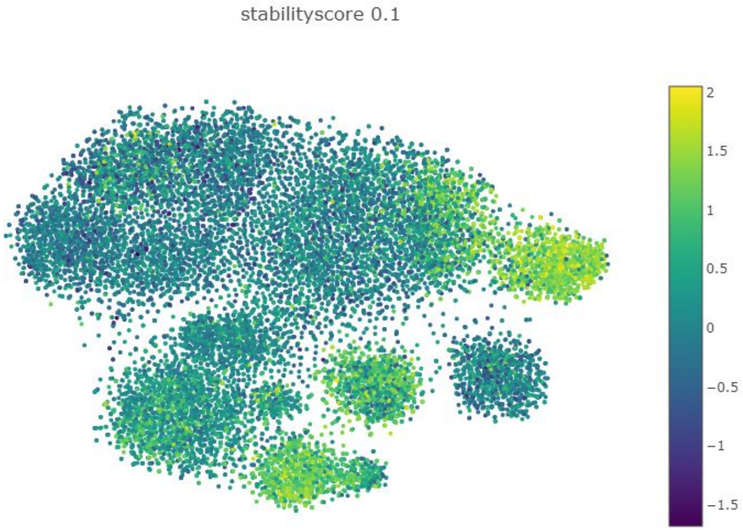
Probability of Rotamer being Native-like

(t-SNE)

prot2vec | Protein Stability Challenge Problem



Total Score from
Rosetta Energy Function
(Alford et al. 2017)



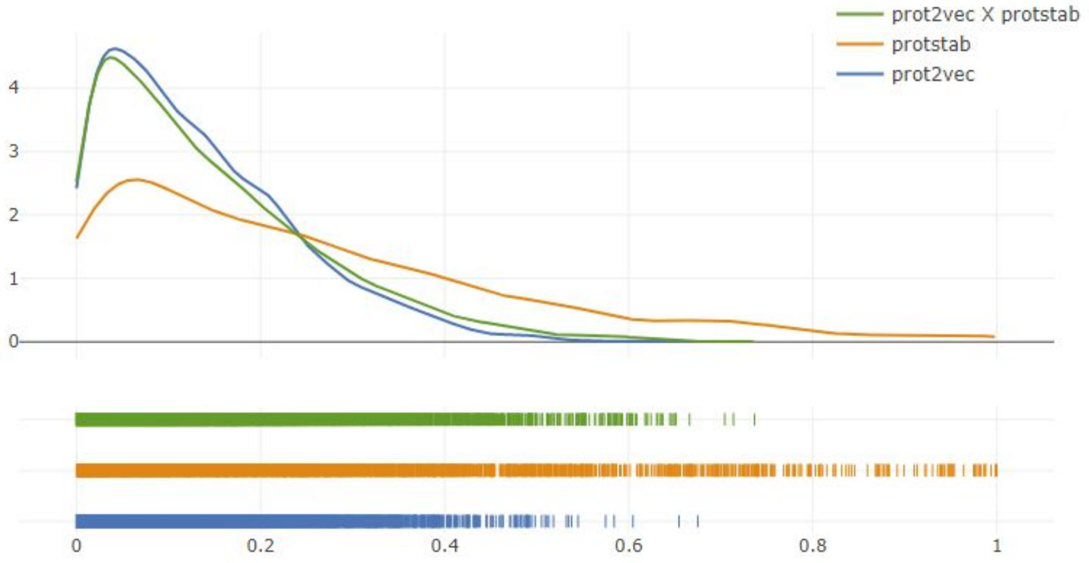
Protein Stability Score
(Rocklin et al. 2017)

(t-SNE)

prot2vec | Protein Stability Challenge Problem

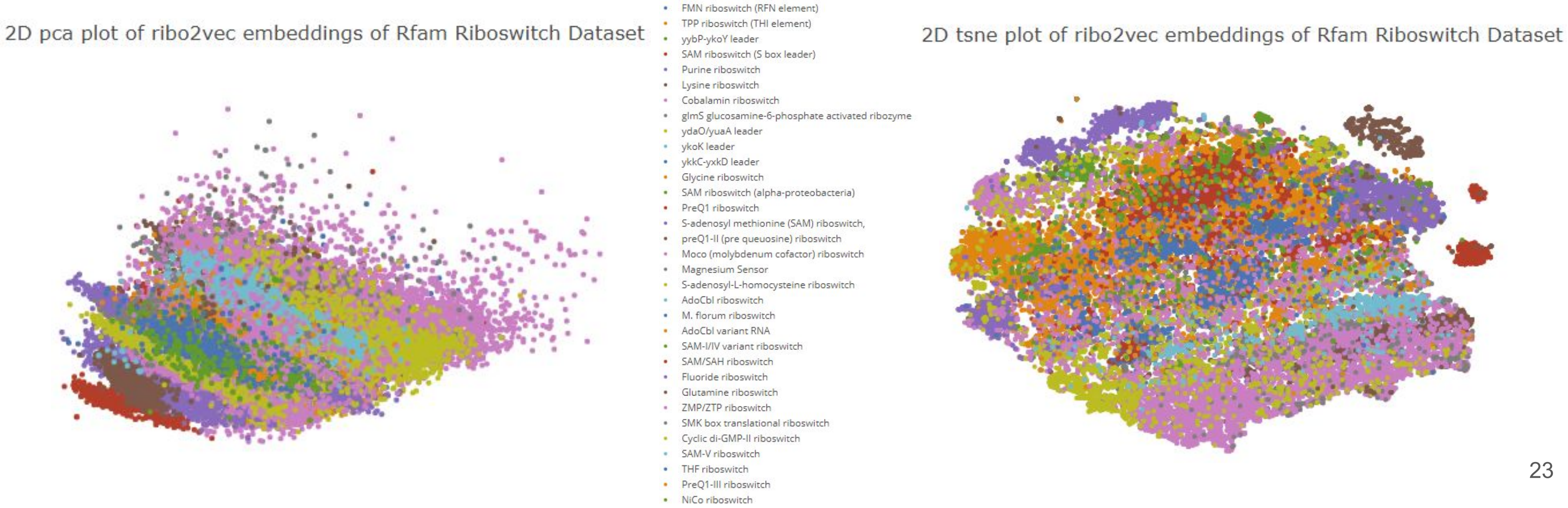
- Clearly, prot2vec captures key protein properties
  - Evidence of prot2vec: sequence is foundation of high-level protein properties
- ~115 features in the ProtStab challenge
- prot2vec as an extra “feature space” that reduces redundancy while capturing key protein properties

Distribution of Pairwise Dimension Correlations for protstab Dataset and Corresponding prot2vec Embedding



ribo2vec | Riboswitch Design Challenge Problem

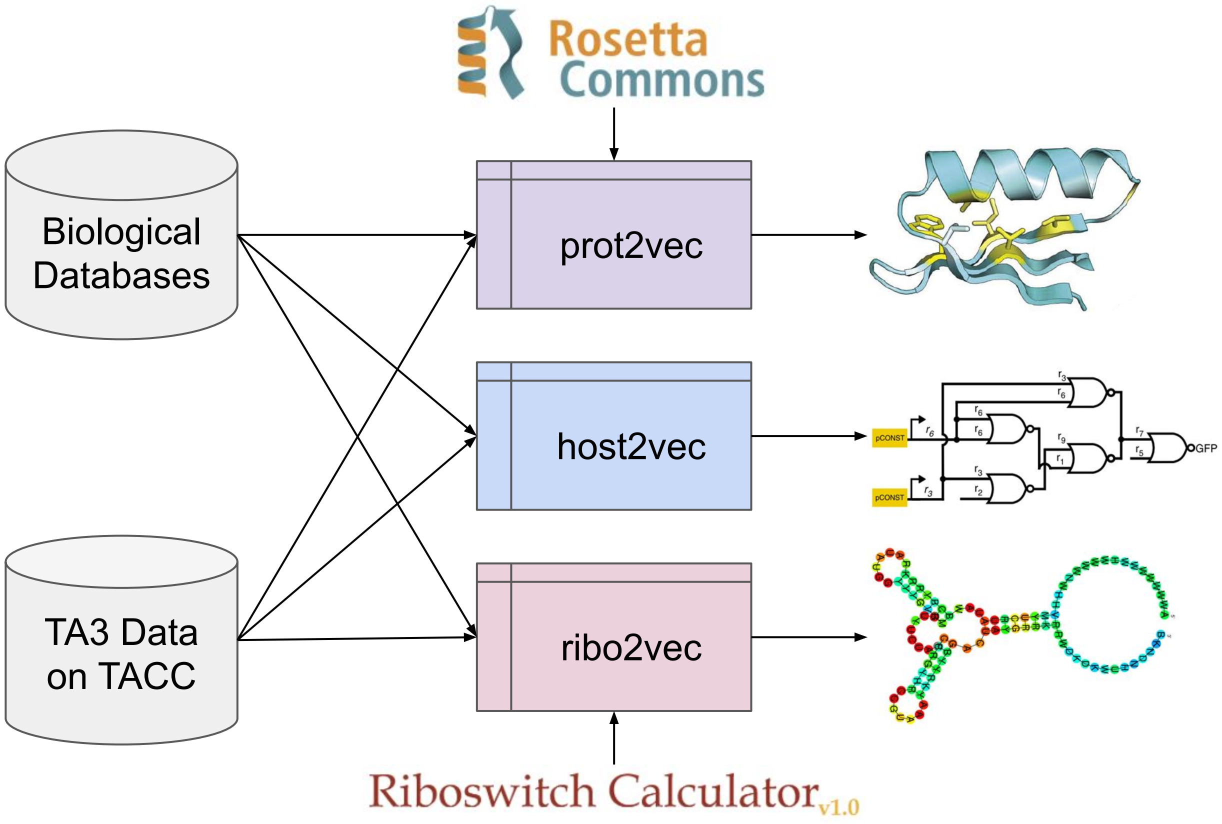
- Intuition: learn the space of mRNAs (aptamers) that bind to ligands
- Input: 49,159 nucleotide sequences across 33 riboswitch families on Rfam
- Output: 10-dimensional embeddings of arbitrary nucleotide sequences



2D pca plot of ribo2vec embeddings of Rfam Riboswitch Dataset

2D tsne plot of ribo2vec embeddings of Rfam Riboswitch Dataset

Embedding Models for SD2



Rosetta Commons

Biological Databases

TA3 Data on TACC

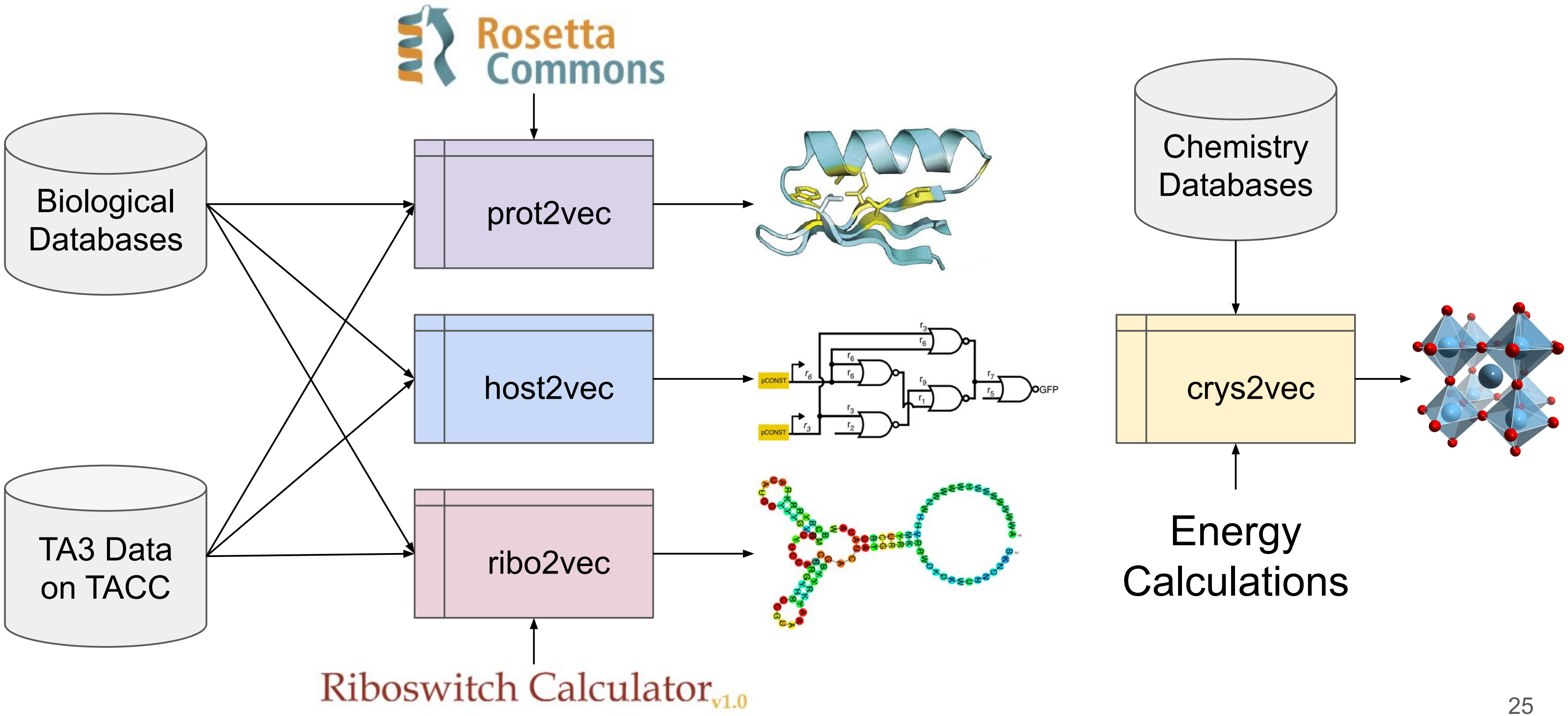
prot2vec

host2vec

ribo2vec

Riboswitch Calculator v1.0

Embedding Models for SD2



Rosetta Commons

Biological Databases

TA3 Data on TACC

prot2vec

host2vec

ribo2vec

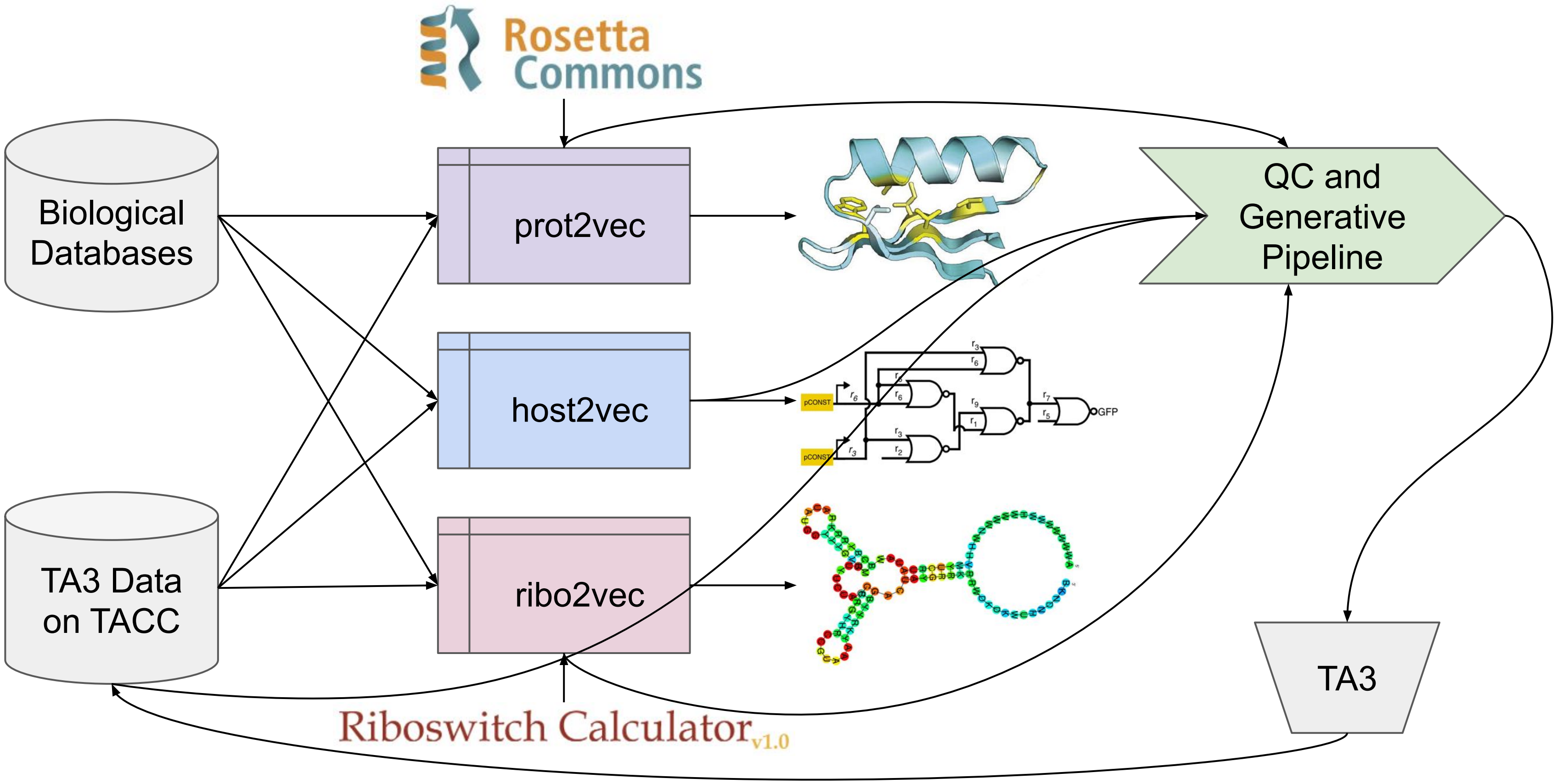
Riboswitch Calculator v1.0

Chemistry Databases

crys2vec

Energy Calculations

Embedding Models for SD2



Rosetta Commons

Biological Databases

TA3 Data on TACC

prot2vec

host2vec

ribo2vec

QC and Generative Pipeline

TA3

Riboswitch Calculator v1.0