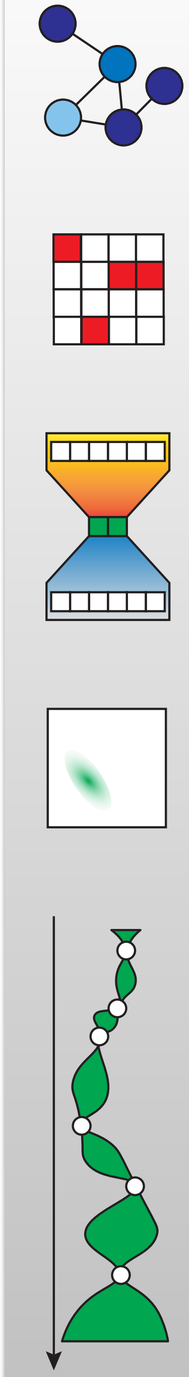
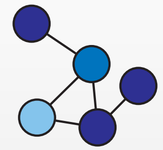
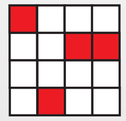
Deep Learning Methods for Learning Phenotype from Genotype

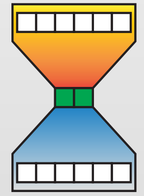
Eric J. Ma
Dept. Biological Engineering, MIT

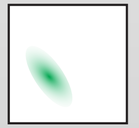


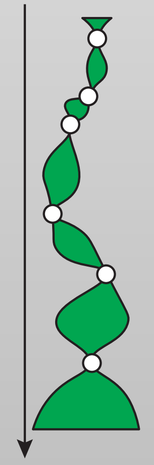
Automatic discovery of structural features predictive of phenotype





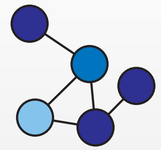


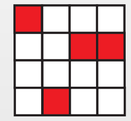


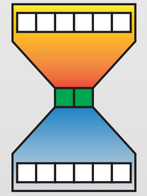


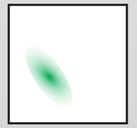
Sequence regression ignores positional interactions

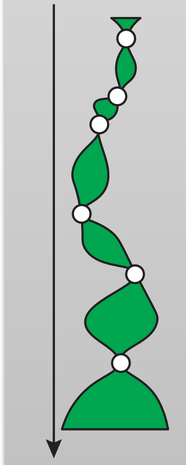
Sequence	DR
PQVTLWQKPIVTIKIGG	2.4
PQVTLWQRPIVTIKIGG	3.8
PQVTLWQRPLVTIKIGG	9.4
PQVTLWQRPIVTIKIGG	3.5









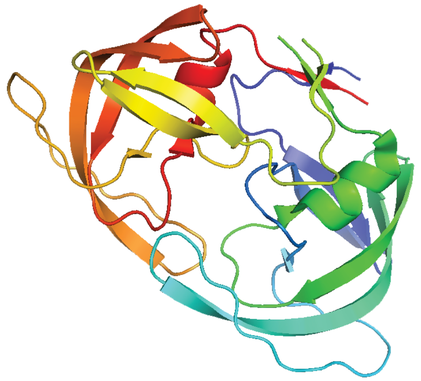


Proteins have a natural graph (network) representation

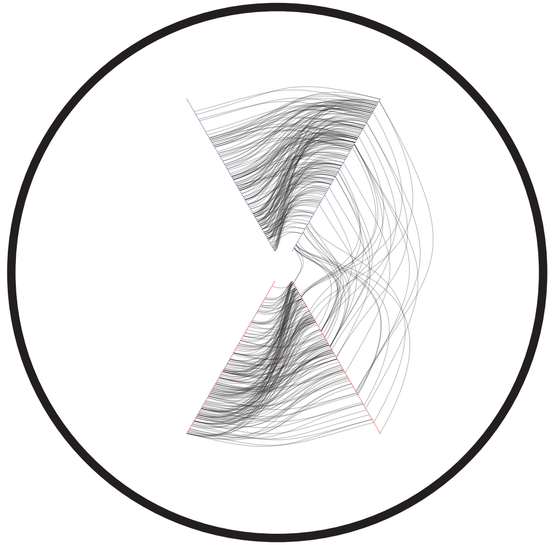
Sequence

PQVTLWQKPIVTIKIGG
PQVTLWQRPIVTIKIGG
PQVTLWQRPLVTIKIGG
PQVTLWQRPIVTIKIGG

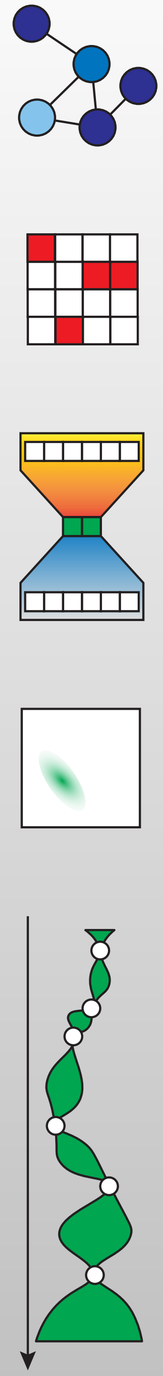
Homology Model



Network Representation

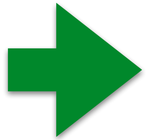


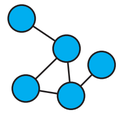
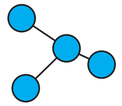
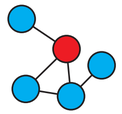
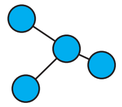
Nodes: Amino acids
Edges: Biochemical interactions

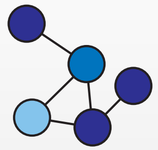


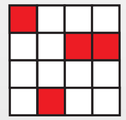
Can graph regression on proteins help us interpret structure better?

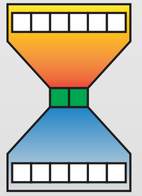
Sequence	DR
PQVTLWQ**K**PIVTIKIGG	2.4
PQVTLWQRPIVTIKIGG	3.8
PQVTLWQRP**L**VTIKIGG	9.4
PQVTLWQRPIVTIKIGG	3.5

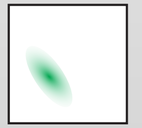


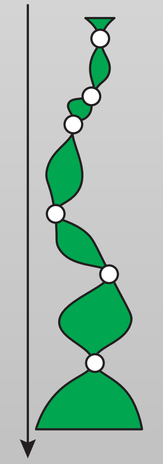
Structure Graph	DR
	2.4
	3.8
	9.4
	3.5







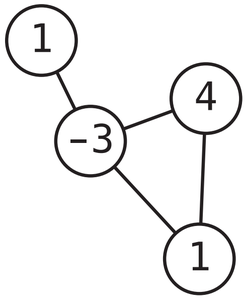


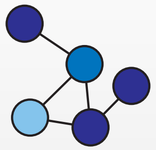


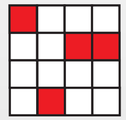
Graph convolutions let us compute unique fingerprints for distance graphs

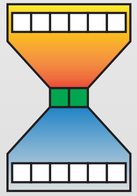
Input Convolution Non-Linearity Fingerprint

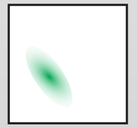
Scalars

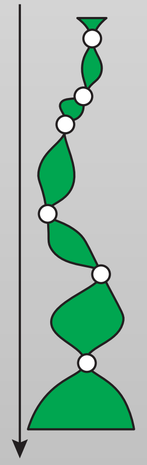




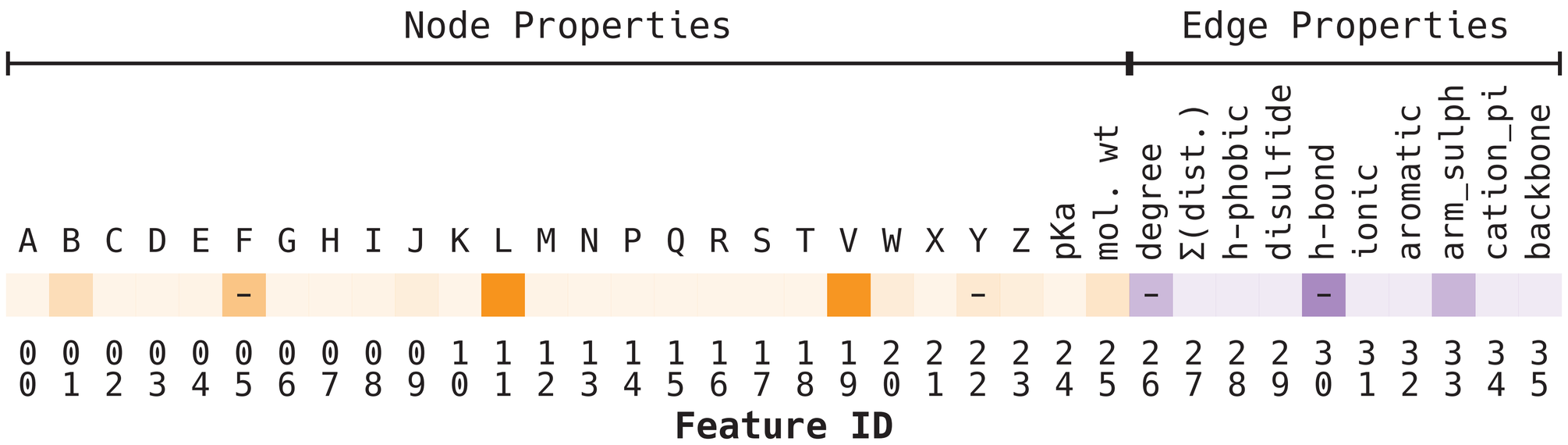


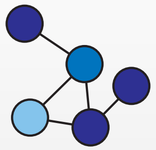


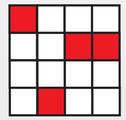


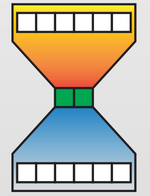


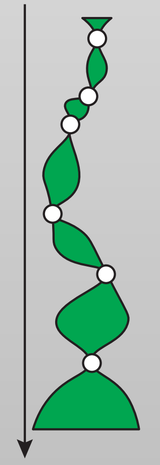
Graph convolutions let us compute unique fingerprints for distance graphs



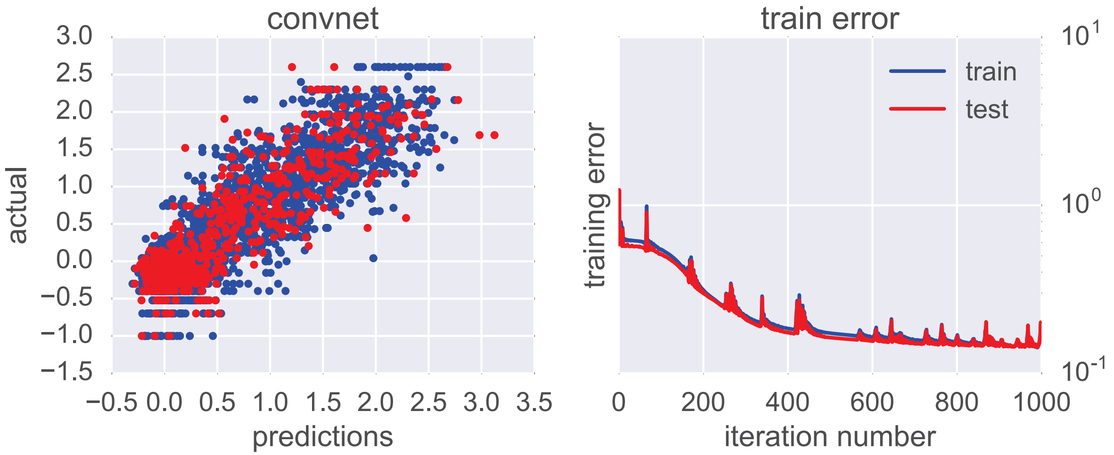


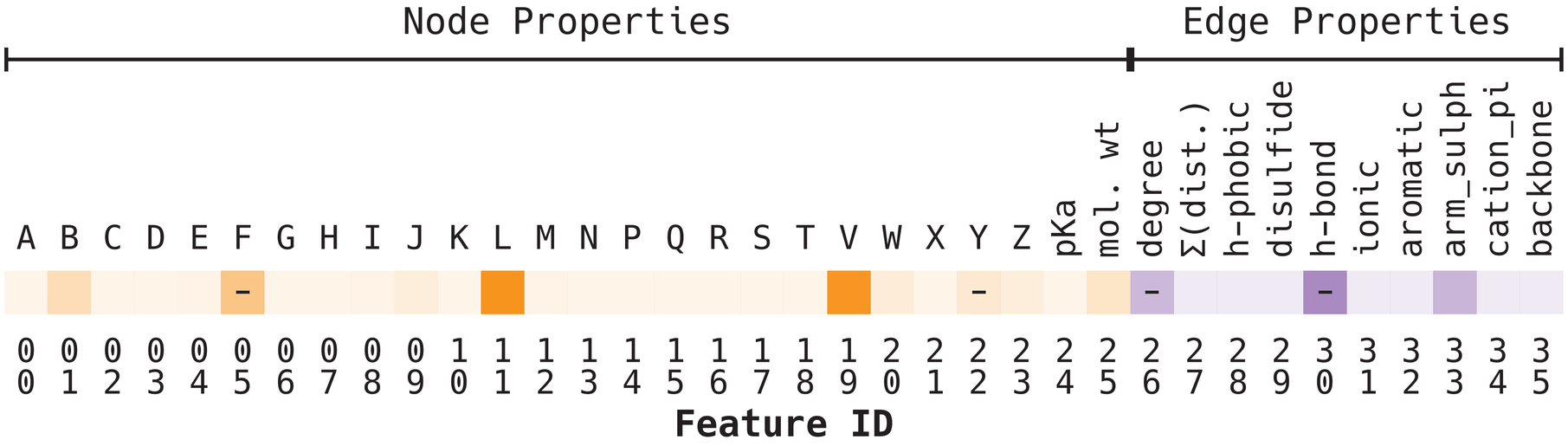




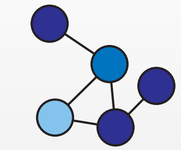
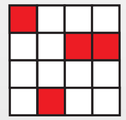
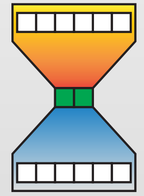
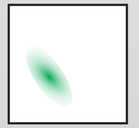
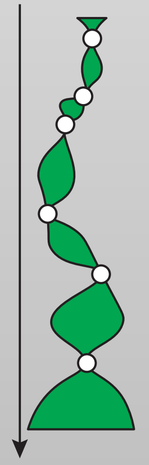


Graph convolutions automatically learn HIV protease a.a. neighbourhoods responsible for drug resistance





Forecasting fast-evolving pathogen sequence trajectory

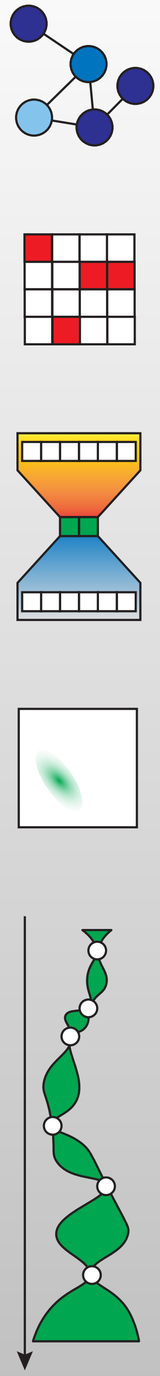
Forecasting sequence evolution is difficult because sequence space is discrete

Combinatorics problem

10 a.a.: 20^{10} theoretical space

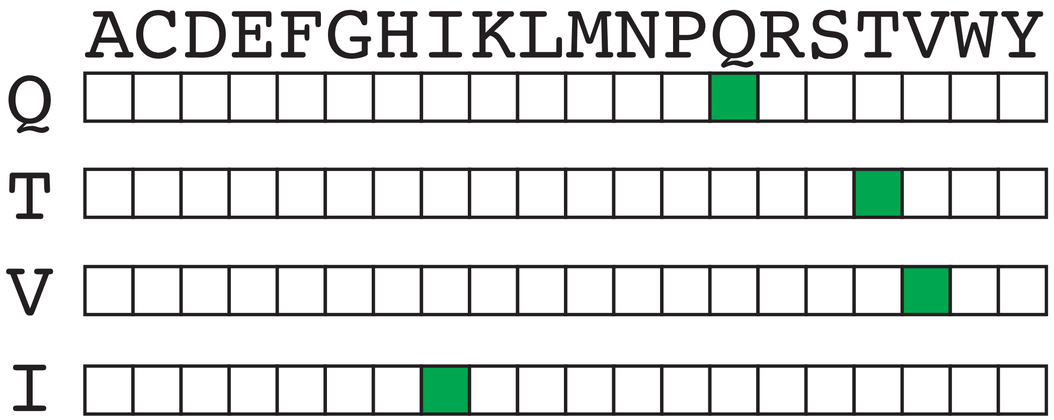
Deep mutational scans: 10^5 empirical space

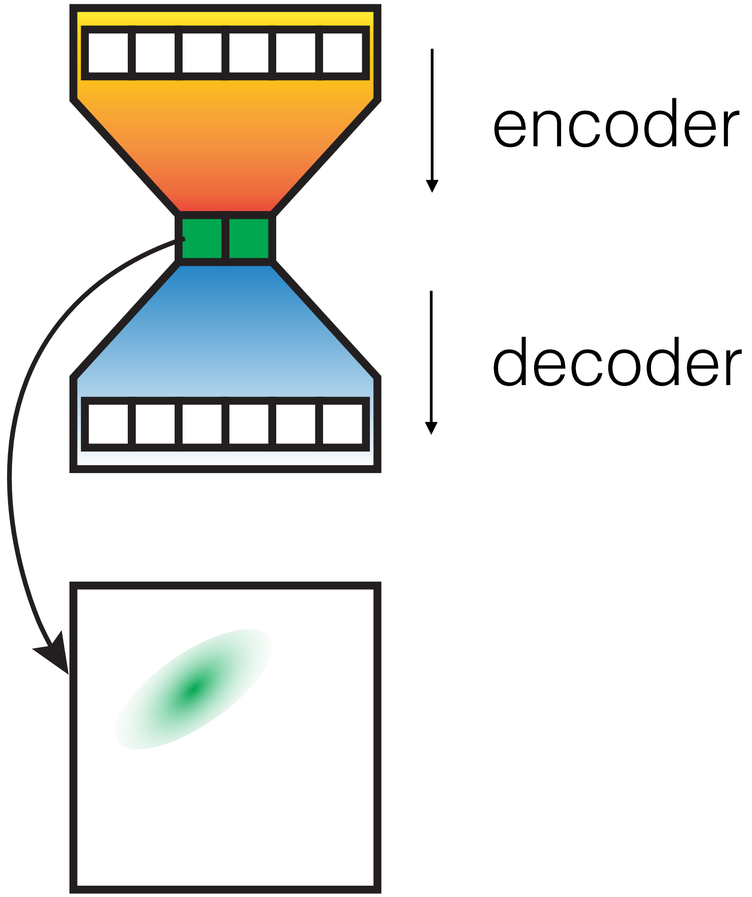
Mutational **direction** cannot be predicted easily

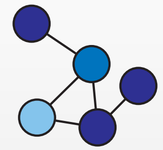


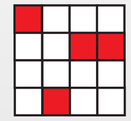
Variational auto-encoders provide a path towards continuous space

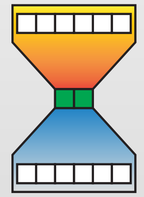
```
SQQTVIPNIGSRPRVRN
SQQAVIPNIGSRPRIRD
SQQAVIPNIGSRPRIRD
SQQAVIPNIGSRPRIRD
```

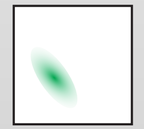


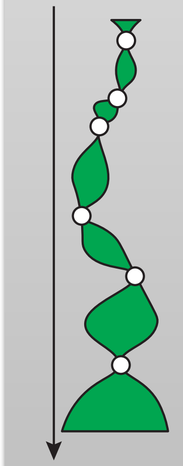




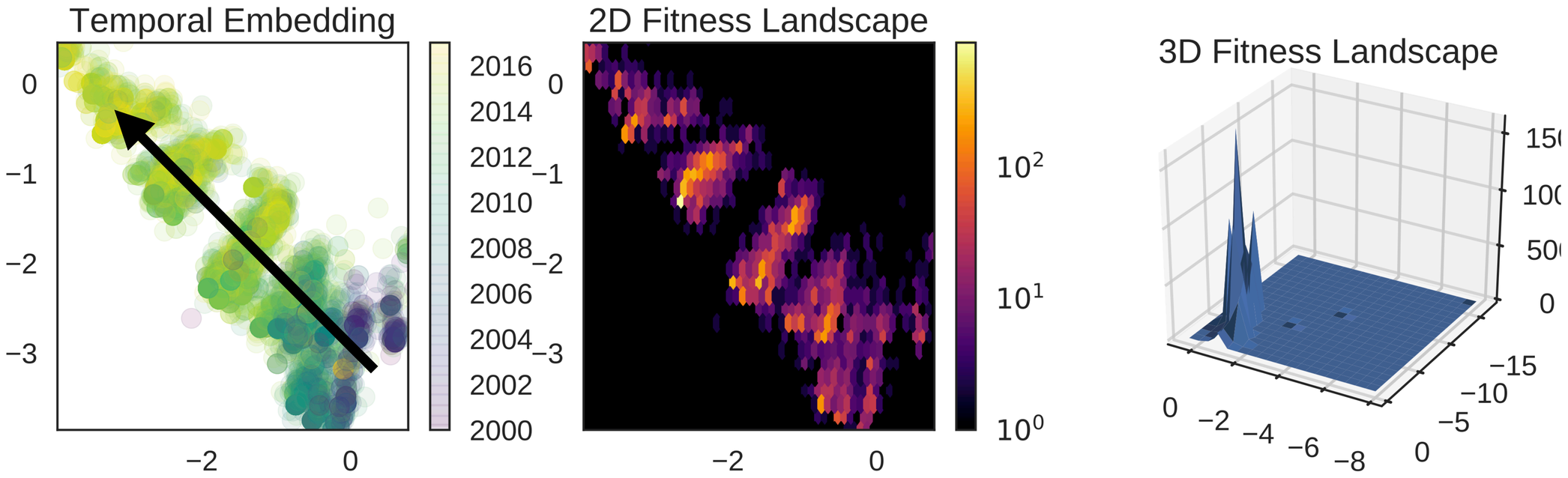




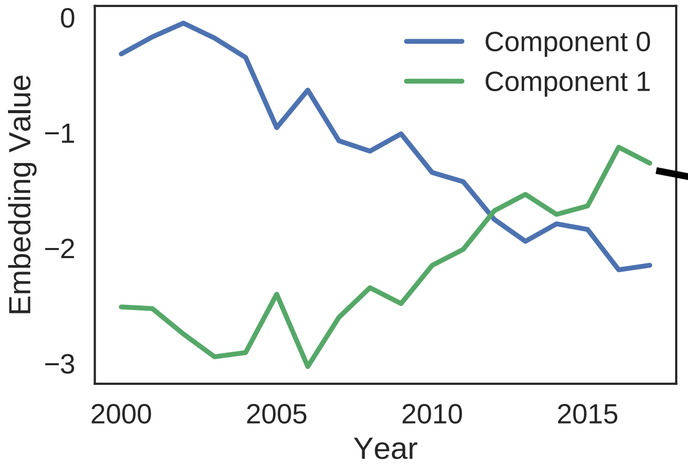


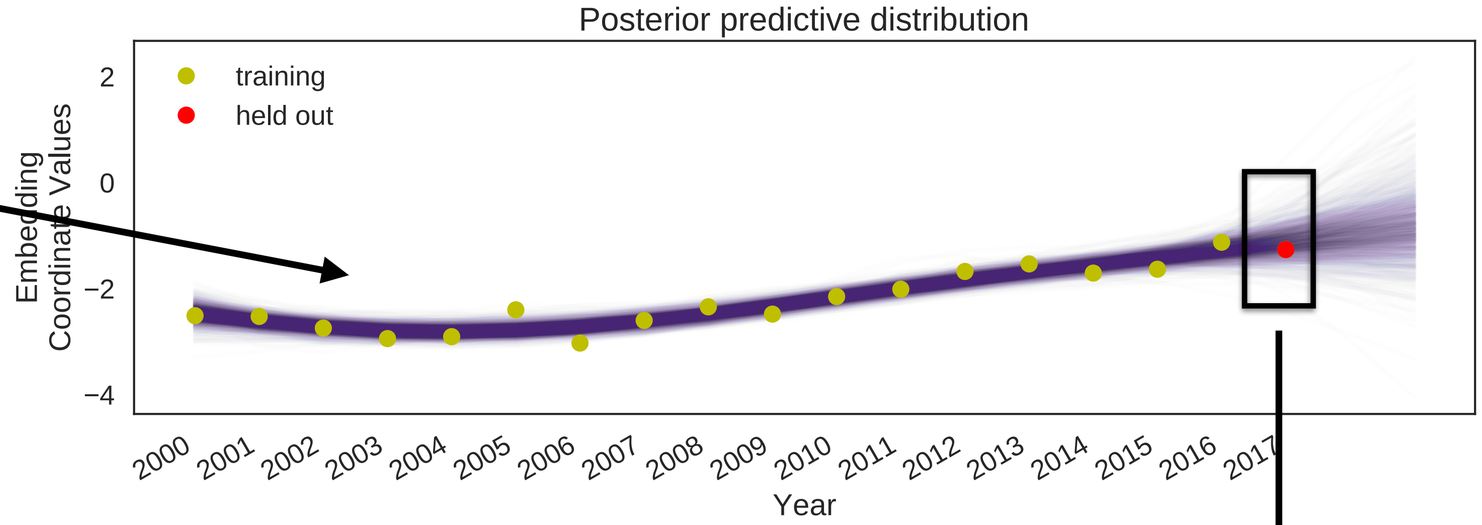


Influenza H3N2 protein evolution has direction in continuous space



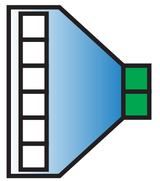
Gaussian process regression enables probabilistic forecasting & sampling of new sequence space

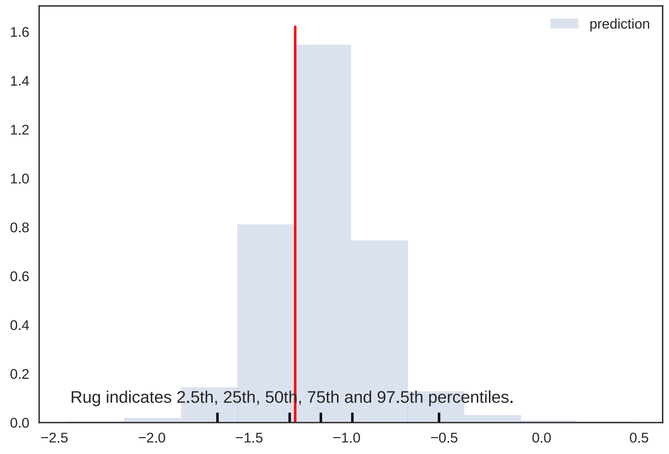




Pos. 176, 327 most likely to change.

Antigenic effects moderate.

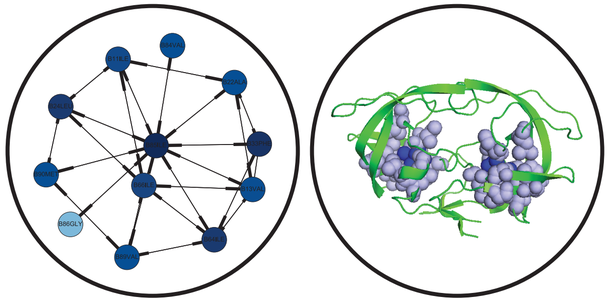


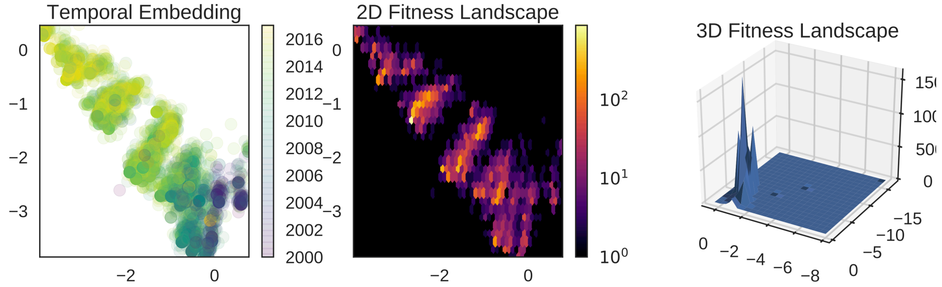


Take-Home Ideas

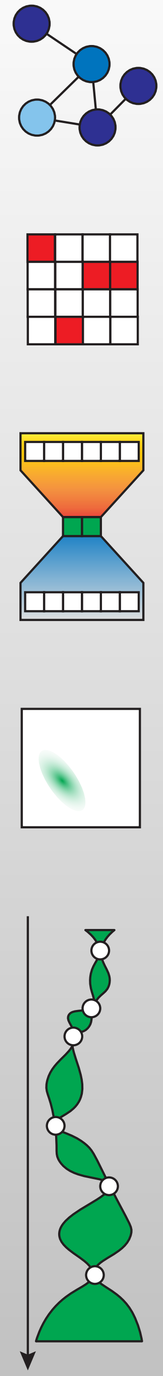
Graph convolutions: automatically learning structural determinants of phenotype

Variational autoencoders: forecasting fast-evolving pathogen sequence trajectory





Applications: CAD of biologics, pre-emptive vaccine development.



Acknowledgments

Jonathan A. Runstadler
Islam T. M. Hussein
Nichola J. Hill

Collaborators
David K. Duvenaud