Interpretable deep learning models to predict protein phenotype from genotype

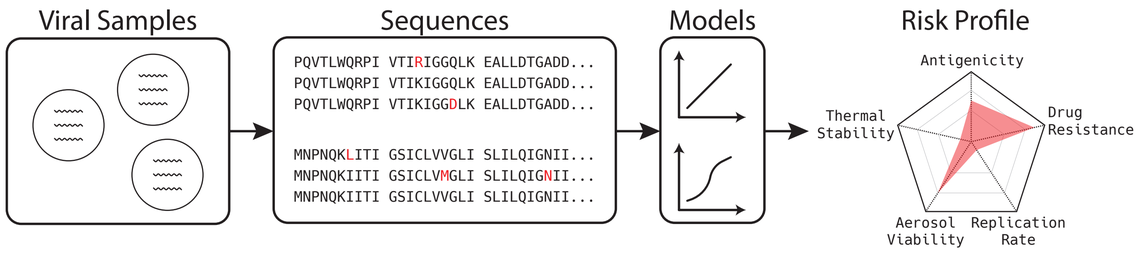
Eric J. Ma[1], David K. Duvenaud[2], Jonathan A. Runstadler[1,3]
[1]Department of Biological Engineering & [3]Division of Comparative Medicine, MIT
[2]School of Engineering and Applied Science, Harvard University

Introduction

Problems

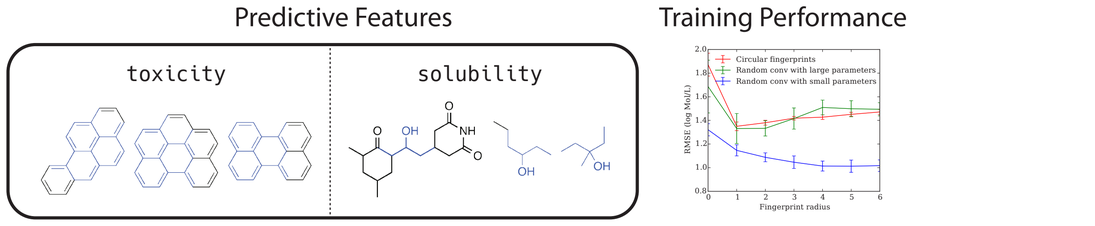
1. Pathogen risk determination is currently based on non-standardized measurements and simple heuristics.
2. Mapping from genotype to phenotype is complex.
3. Lack of standardized measurements hampers systematic study & reproducibility.
4. Current machine learning models cannot regress on inputs of variable length.

Vision



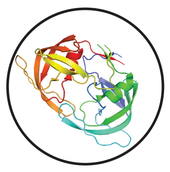
- Real-time dashboard for influenza surveillance
- Risk profile informs tailored interventions

Prior Work



- Duvenaud et. al., 2016 (arXiv): prediction of chemical properties on chemical graphs
- Genotype: chemical structure; phenotype: chemical property.
- Applications in drug screening, toxicity prediction etc.

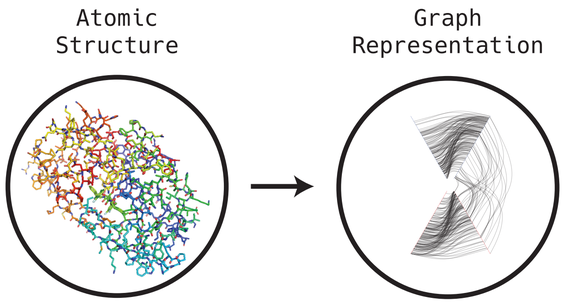
Goals



- Data set: HIV-1 Protease, Stanford HIV Drug Resistance Database
- Train convolutional network on protein graph.
- Develop software package for generalized graph regression

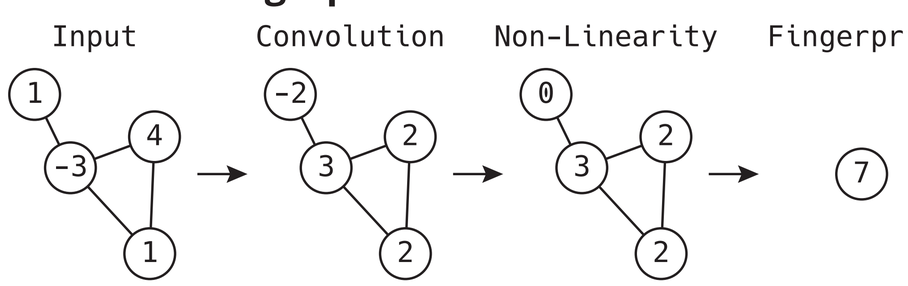
Deep Learning Algorithm

Protein Graphs



- Nodes: Amino acids with their identities
- Edges: Biochemical interactions between amino acids
- Biochemical knowledge of each node is encoded as fixed vector.

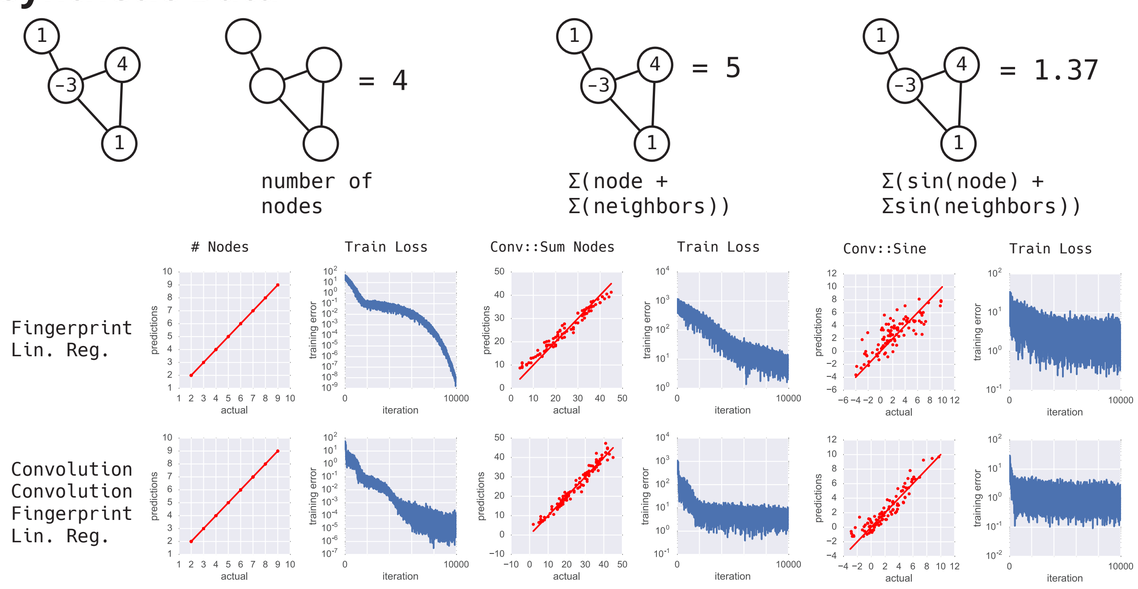
Convolution & Fingerprint



- Convolutions capture local structure of graph
- Non-linearities allow modelling of arbitrary functions
- Fingerprints represent a fixed-length representation of underlying graph.
- Graphs with identical nodes and edges will have identical fingerprints.

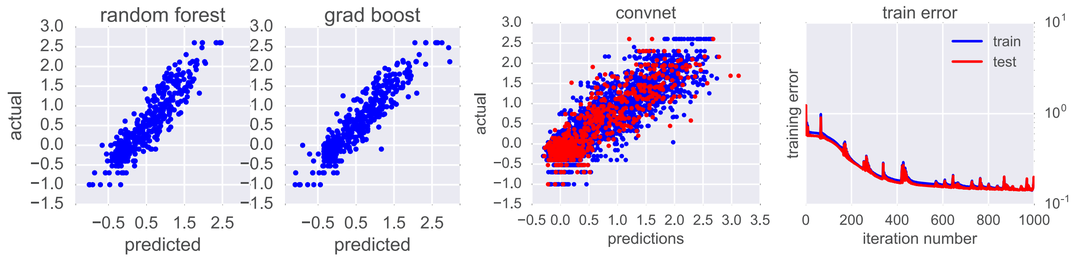
Results

Synthetic Data



- Learn mathematical transforms on top of integer graphs.
- Deeper networks converge in fewer iterations with smaller error.
- Graph-based convolutional neural nets work on simplemathematical functions.

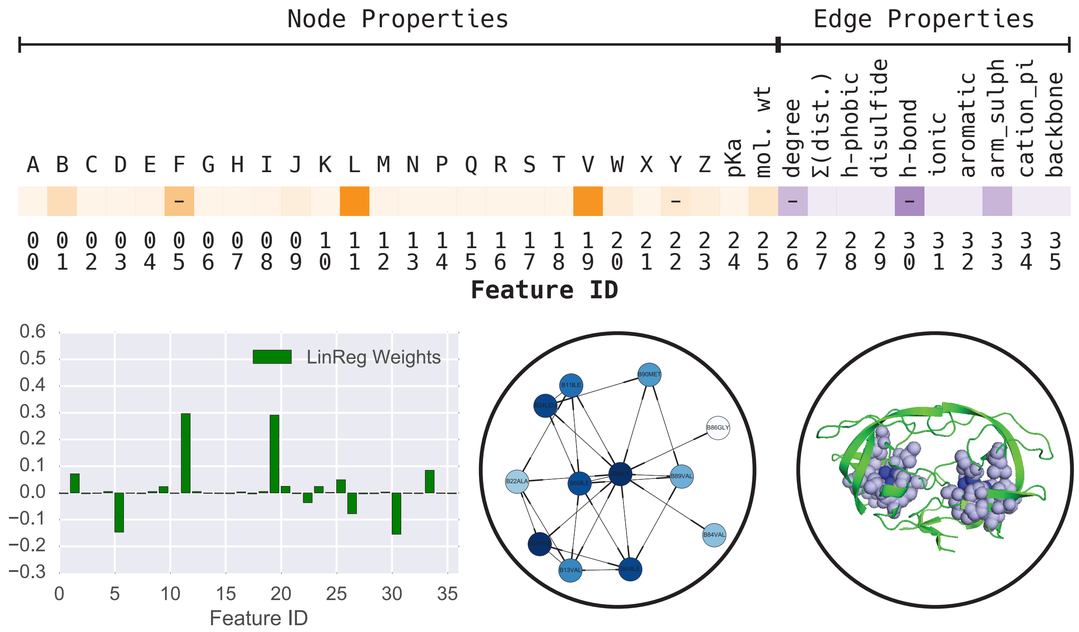
HIV Protease Drug Resistance



Model	Error
Random Forest	0.09
Gradient Boost	0.07
Convolution	0.14

- Simple convolutional model comparable with random forest and gradient boost baseline models.
- Very little overfitting.

Visualizing Learned Features



- Able to recapitulate known mutations involved in FPV drug resistance.
- Interpretable: hydrophobic network of amino acids implicated in FPV resistance.
- (left) Dark nodes: highly activating; light nodes: weakly activating
- (right) Green ribbon: backbone; Dark blue spheres: top activating nodes; Light blue spheres: neighbors.

Future Work

- Learning Capacity: neural network architecture improvements; prevent overfitting.
- Interpretability: better visualizations of convolutional feature maps.
- Applications: pathogen genomic surveillance, chemical surveillance.

MIT BE
BIOLOGICAL ENGINEERING

CEIRS
NIAID Centers of Excellence for Influenza Research and Surveillance

MIT CEHS

The Division of Comparative Medicine
MIT - DCM Building 16, Room 825
77 Massachusetts Avenue
Cambridge, MA 02139