# Enhancing pathogenicity prediction from structure

(with machine learning)

# Introduction

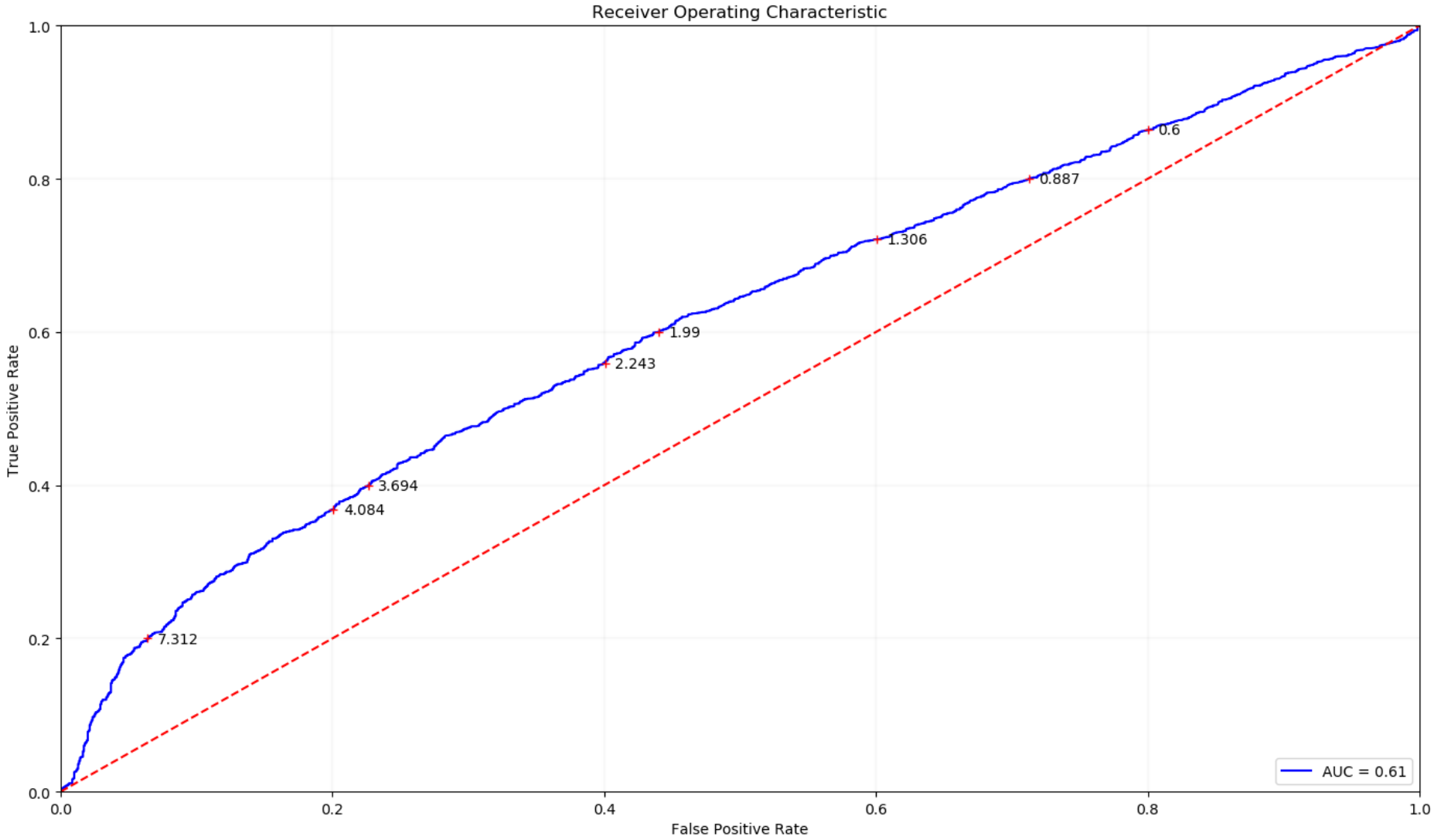
# Data

- 187 structures with 3454 variants of known significance
  - 2346 expected to be pathogenic (clinvar)
  - 1108 expected to be benign (gnomad)

- Collected from the pdbmap database with:
  - Crystal structure has to be available with resolution <2.5Å and length between 100 and 450
  - At least 3 different benign and 3 different pathogenic variants
  - Benign variants need to have maf>0.0001
  - Conflict between variants: benign if maf>0.05, pathogenic otherwise
  - Only 1 chain per structure is used and only 1 structure per uniprot

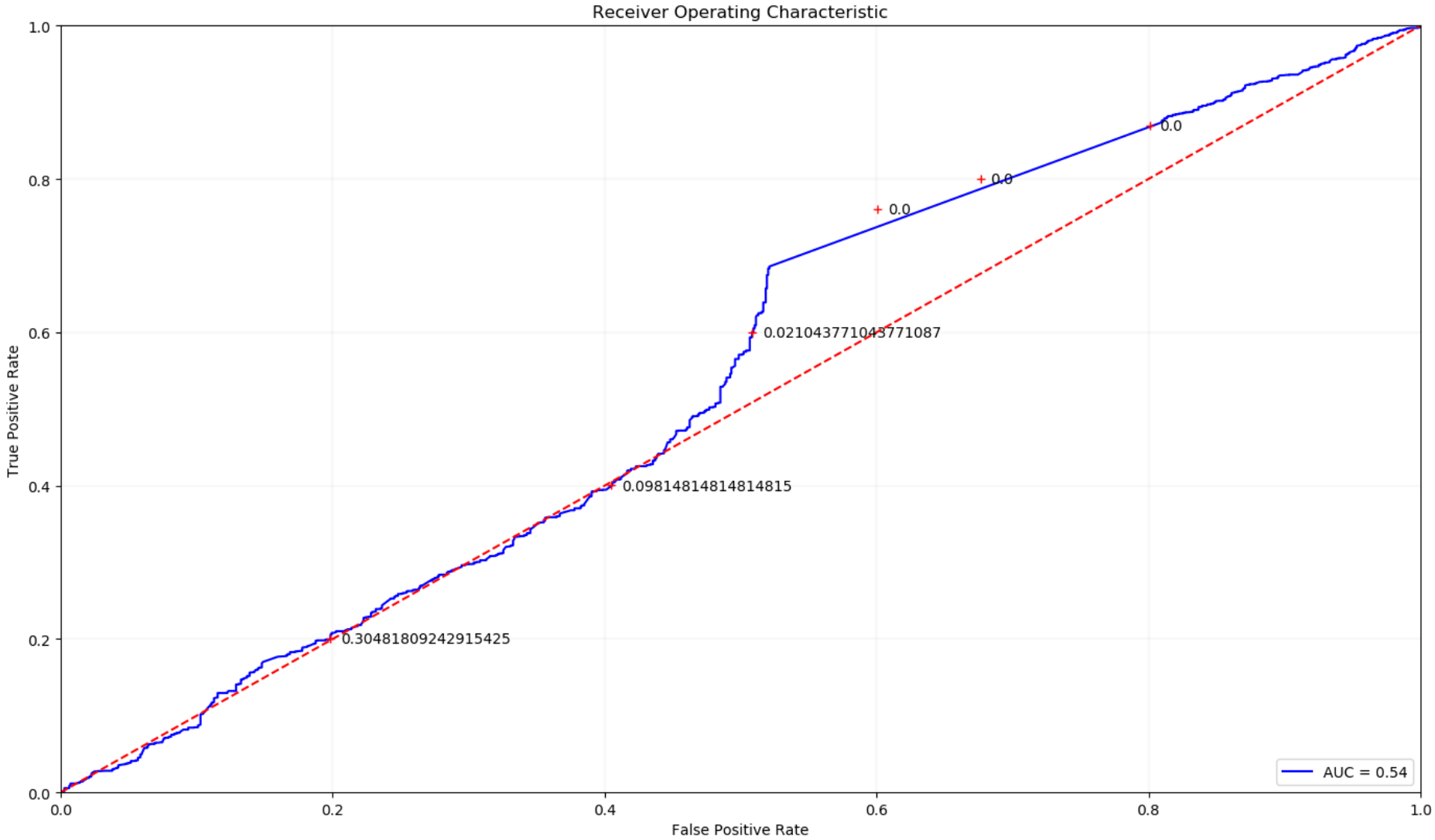
# Feature – ΔΔG

# Feature – ΔΔG


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
7.312
4.084
3.694
2.243
1.99
1.306
0.887
0.6
AUC = 0.61

# Feature – Pathprox

# Feature – Pathprox



# Feature – Uniprot Annotations

# Feature – Uniprot Annotations

| Molecule processing | Binding region | Amino acid modification | Covelant bonding | Sites |
|---|---|---|---|---|
| Transit peptide | Nucleotide binding | Modified residue | Disulfide bond | Binding site |
| Signal | Zinc finger | Lipidation | Cross-link | Metal binding |
| Initiator methionine | Calcium binding | Glycosylation | | Active site |
| Propeptide | DNA binding | | | Site |
| | Motif | | | |

- Helix
- Turn
- Beta strand
- Transmembrane regions

- 2838 annotations were suitable, 922 variants have none, 2230 have one, 298 have two and 4 have three

# Feature – Uniprot Annotations

| Molecule processing | Binding region | Amino acid modification | Covelant bonding | Sites |
|---|---|---|---|---|
| Transit peptide | Nucleotide binding | Modified residue | Disulfide bond | Binding site |
| Signal | Zinc finger | Lipidation | Cross-link | Metal binding |
| Initiator methionine | Calcium binding | Glycosylation | | Active site |
| Propeptide | DNA binding | | | Site |
| | Motif | | | |

- Helix
- Turn
- Beta strand
- Transmembrane regions

- 2838 annotations were suitable, 922 variants have none, 2230 have one, 298 have two and 4 have three

# Feature – Uniprot Annotations

| Molecule processing | Binding region | Amino acid modification | Covelant bonding | Sites |
|---|---|---|---|---|
| Transit peptide | Nucleotide binding | Modified residue | Disulfide bond | Binding site |
| Signal | Zinc finger | Lipidation | Cross-link | Metal binding |
| Initiator methionine | Calcium binding | Glycosylation | | Active site |
| Propeptide | DNA binding | | | Site |
| | Motif | | | |

- Helix
- Turn
- Beta strand
- Transmembrane regions

# Feature – Uniprot Annotations

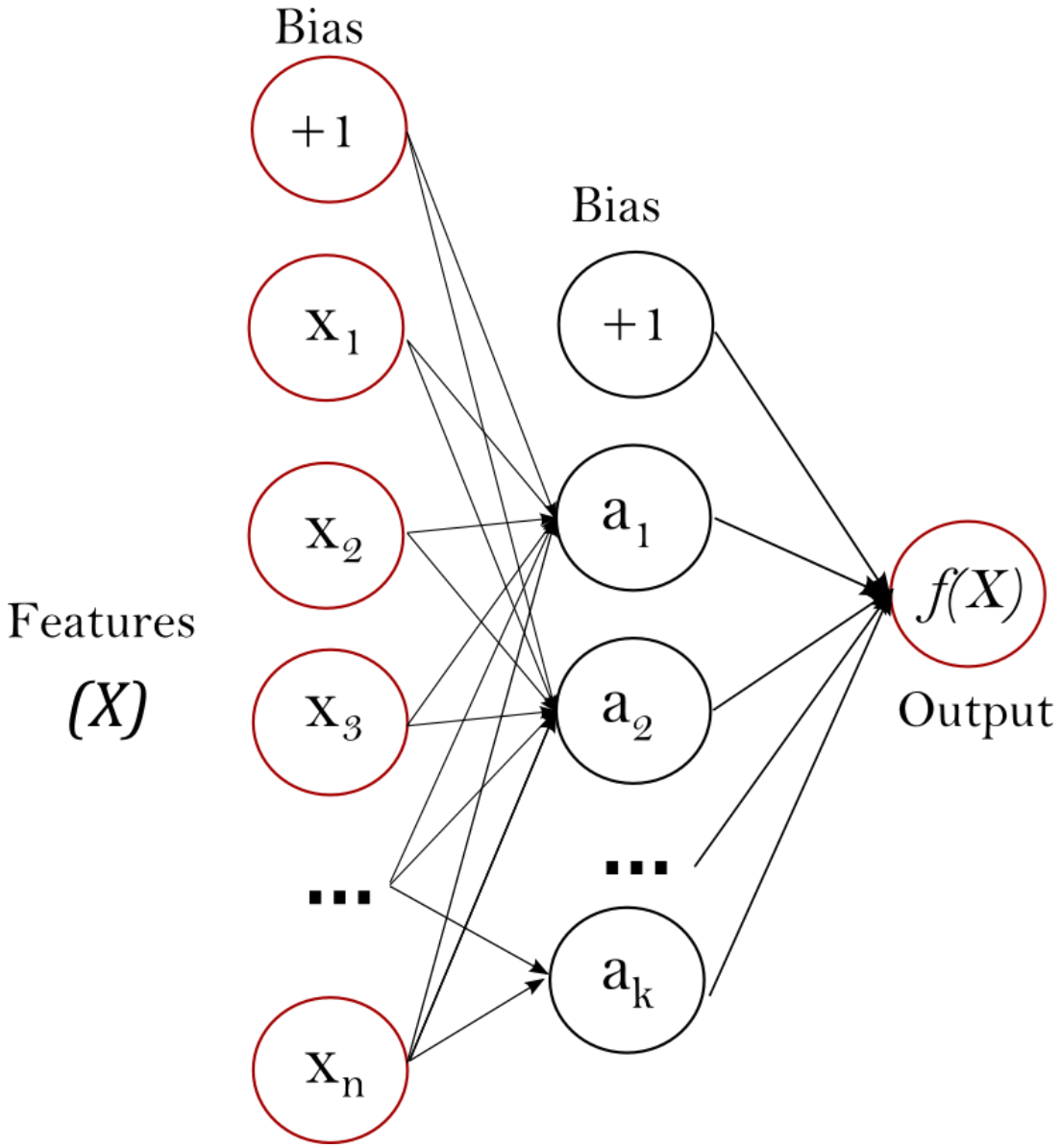
| Molecule processing | Binding region | Amino acid modification | Covelant bonding | Sites |
|---|---|---|---|---|
| Transit peptide | Nucleotide binding | Modified residue | Disulfide bond | Binding site |
| Signal | Zinc finger | Lipidation | Cross-link | Metal binding |
| Initiator methionine | Calcium binding | Glycosylation | | Active site |
| Propeptide | DNA binding | | | Site |
| | Motif | | | |

- Helix
- Turn
- Beta strand
- Transmembrane regions

- **Added instead**: Resolution and length of structure

# Classifier

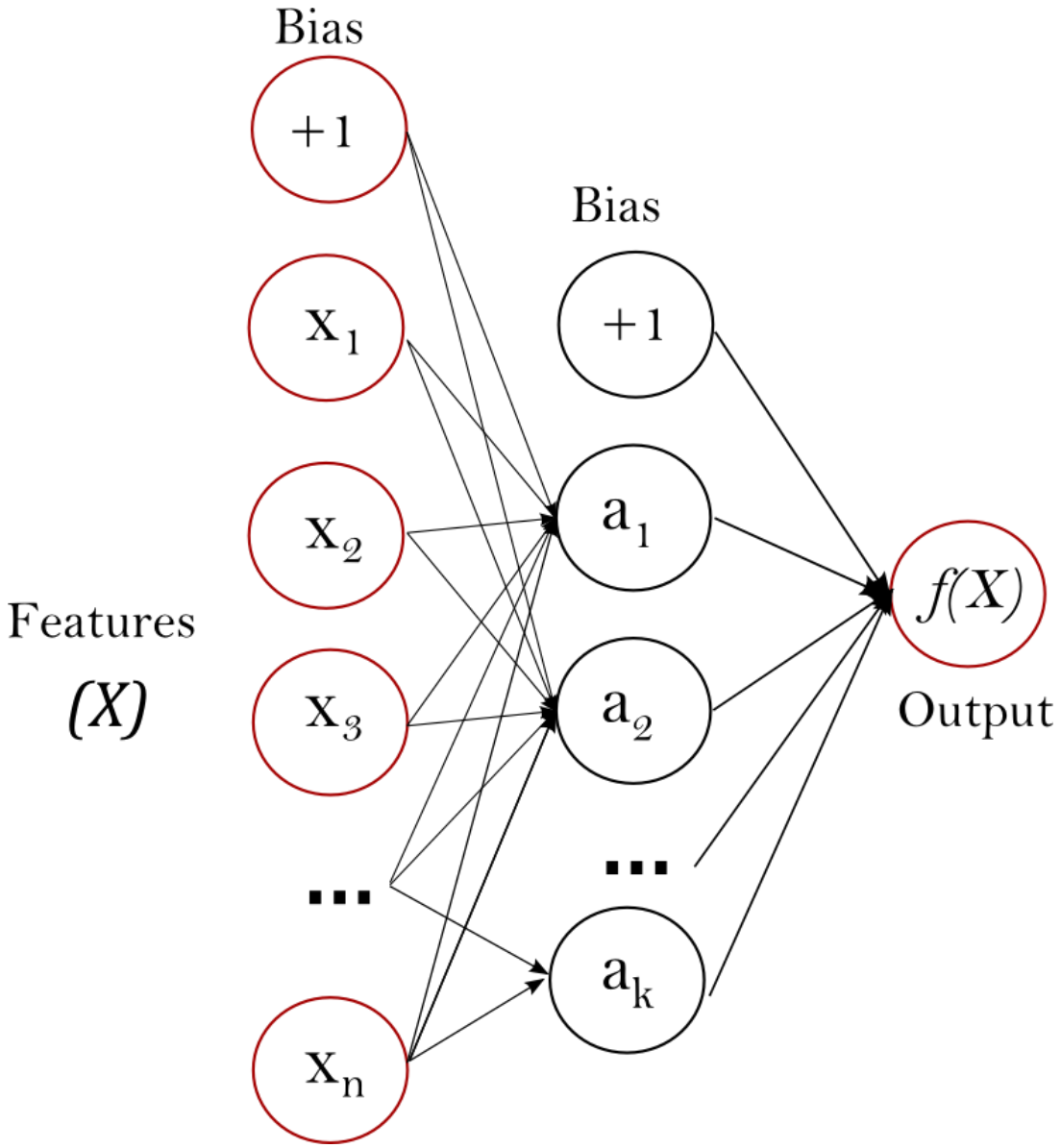
- Basic Function: $f(\cdot): R^m \rightarrow R^o$
- Data: $X = x_1, x_2, \ldots, x_m, y$
- Input Layer: $\{x_i | x_1, x_2, \ldots, x_m\}$
- Value at node $i$ in layer $j$:
  $w_{j,i,1}a_{j-1,i,1} + w_{j,i,2}a_{j-1,i,2} + \ldots + w_{j,i,m}a_{j-1,i,m} + Bias_{j,i}$
- Activation function: $g(\cdot): R \rightarrow R$




Source: https://scikit-learn.org/stable/modules/neural_networks_supervised.html

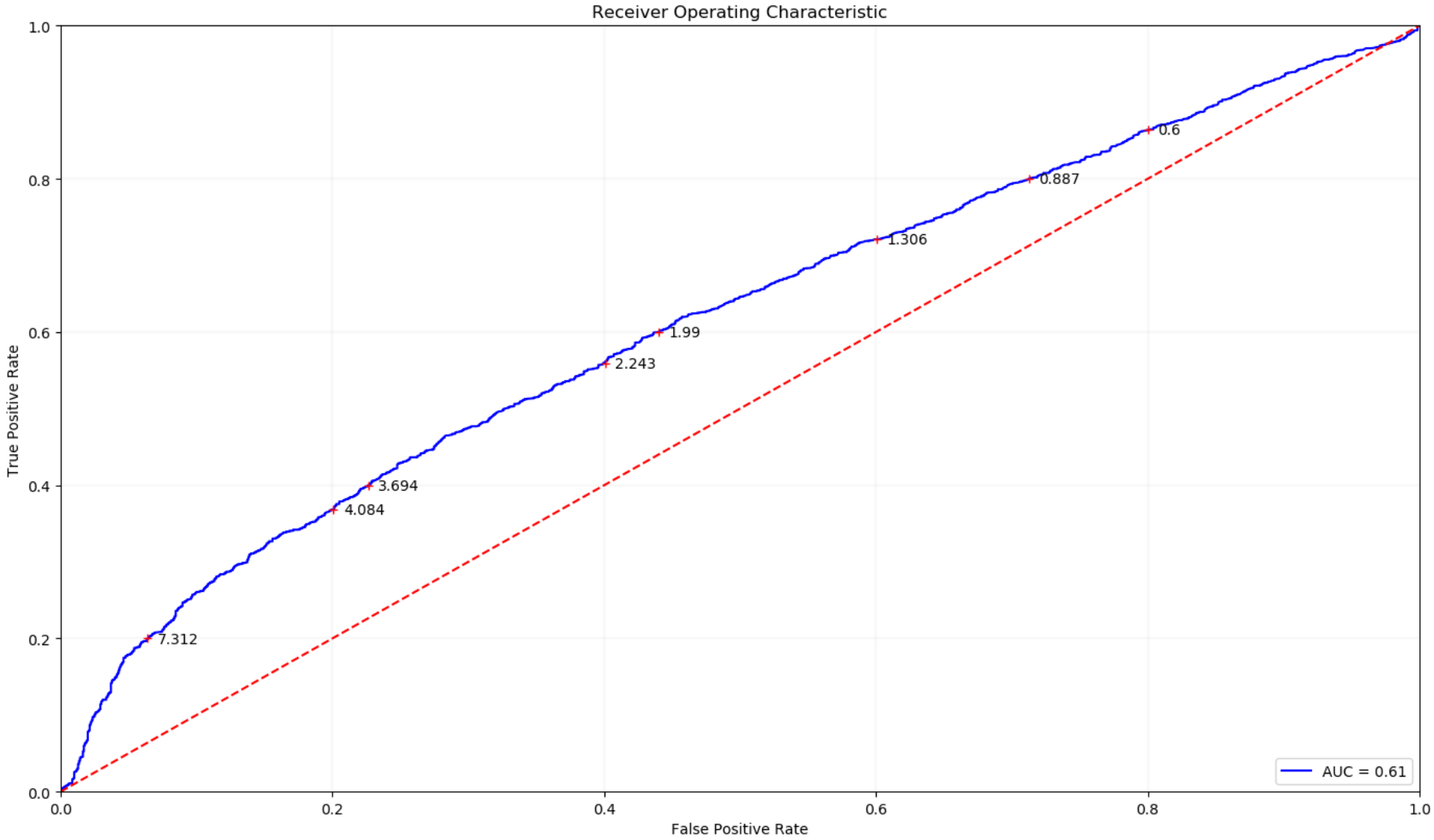
# Classifier

- **Input needs to be scaled!**

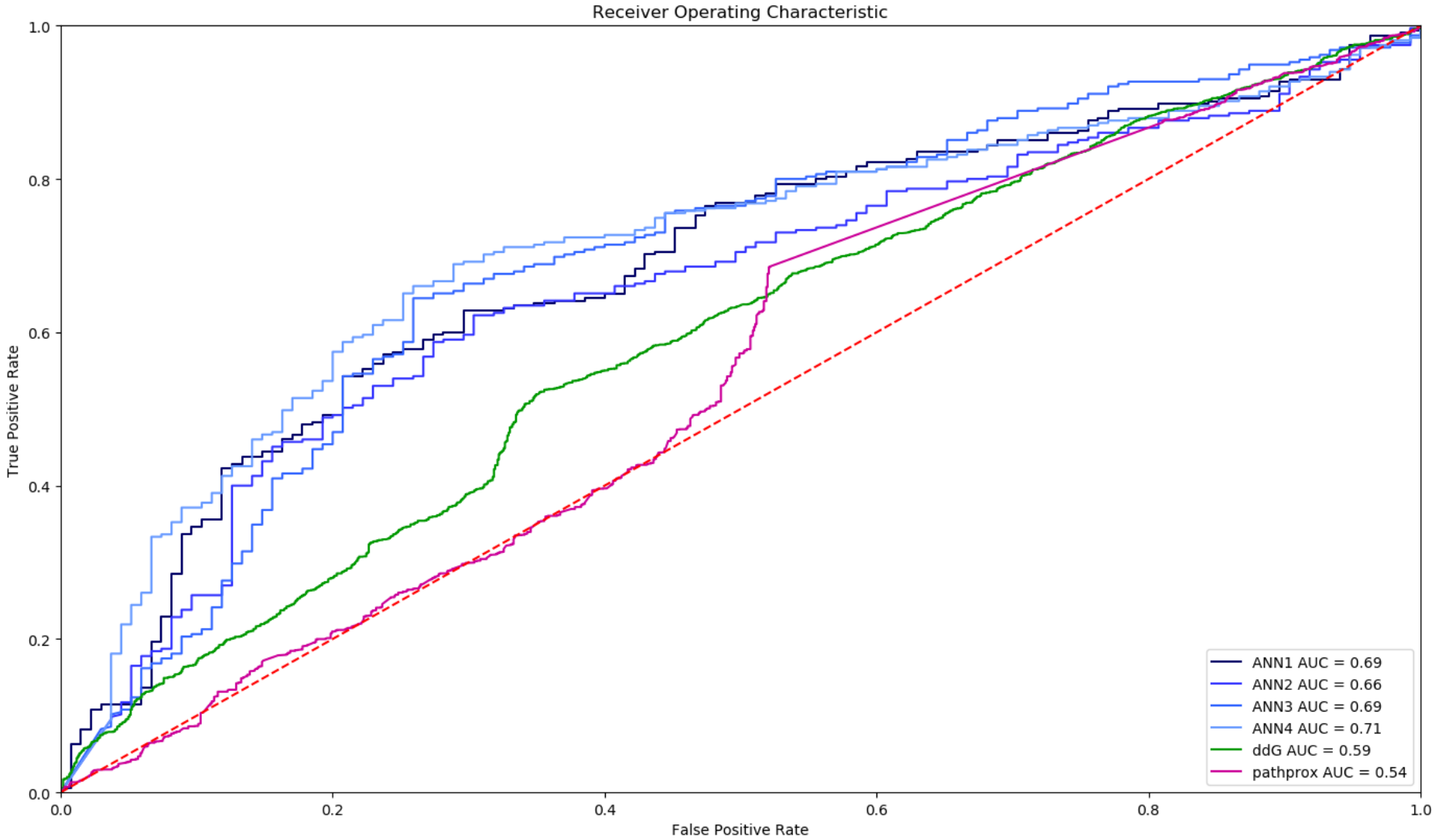



Source: https://scikit-learn.org/stable/modules/neural_networks_supervised.html

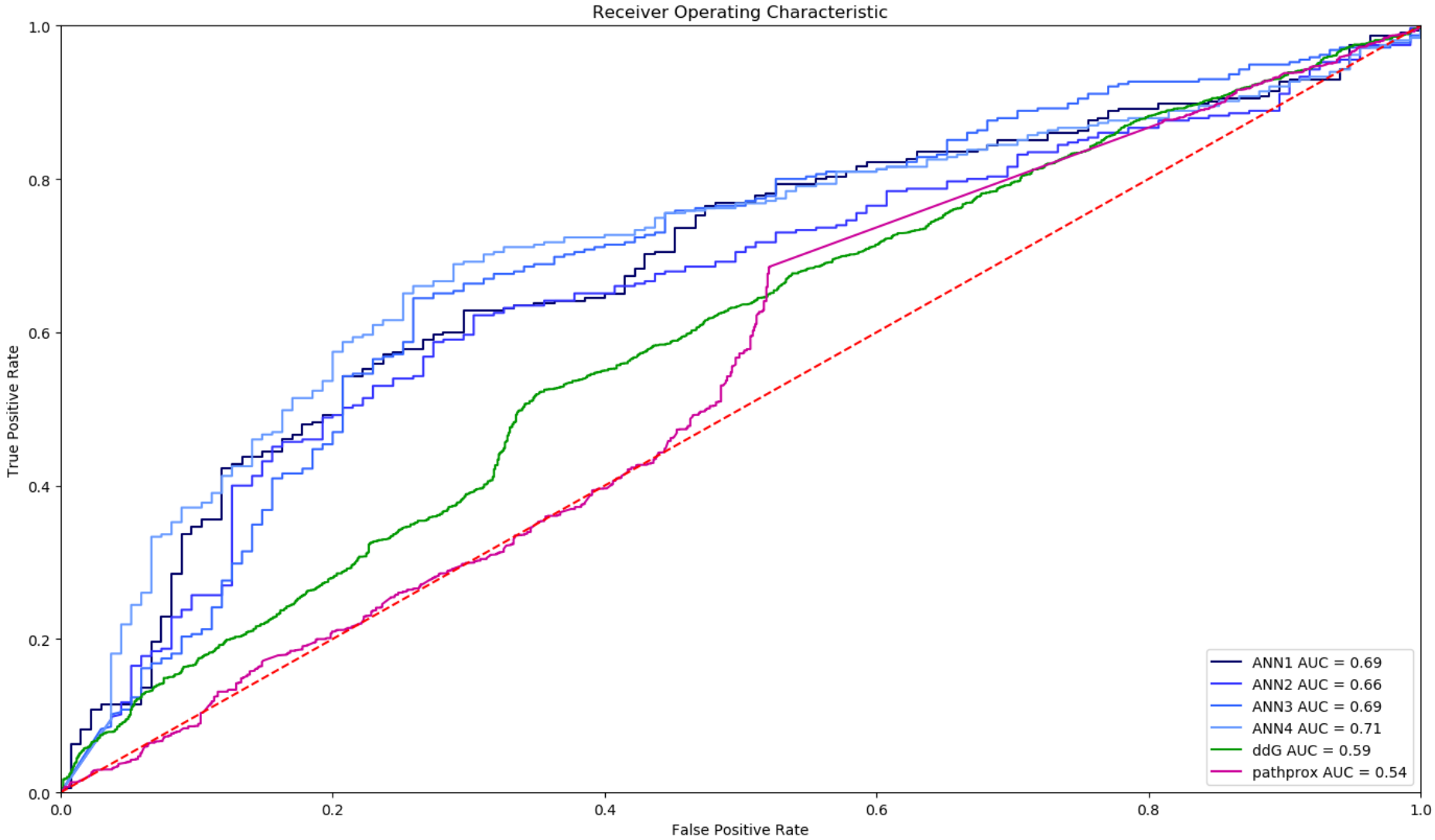
# Classifier


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.6
0.887
1.306
1.99
2.243
3.694
4.084
7.312
AUC = 0.61

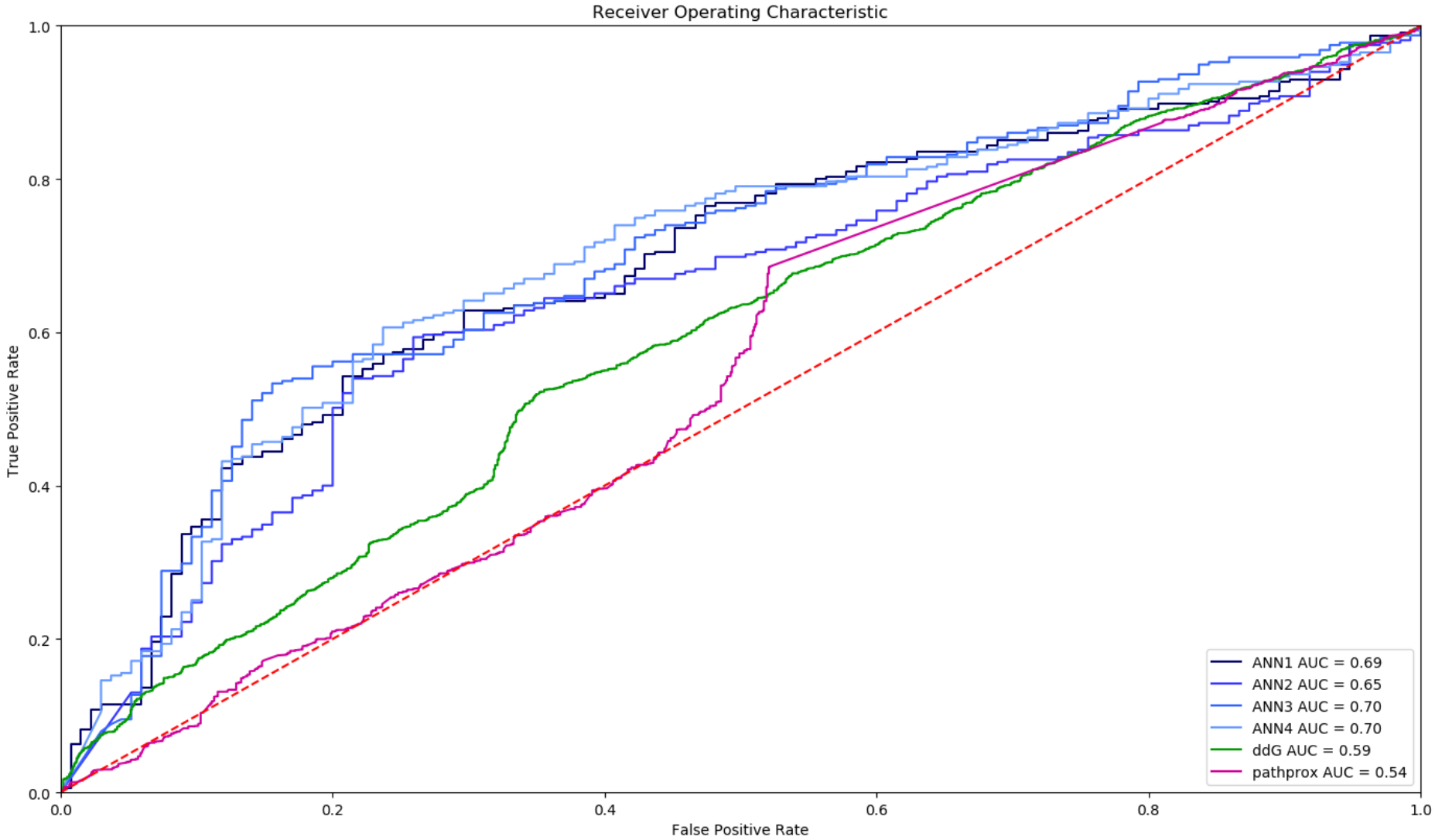
# Classifier Training



# Classifier Training



# Classifier Training



# Classifier 100_100_100_100_21000

# Classifier 100_100_100_100_21000

| Resolution | Length | ddG | pathprox | Beta strand | Binding region | Covalent bonding | Helix | Turn |
|---|---|---|---|---|---|---|---|---|
| 0.464 | 0.347 | 0.535 | 0.579 | 0.502 | 0.506 | 0.485 | 0.579 | 0.503 |
| 0.585 | 0.805 | 0.44 | 0.333 | 0.503 | 0.497 | 0.535 | 0.333 | 0.491 |

# Classifier 100_100_100_100_21000

| Resolution | Length | ddG | pathprox | Beta strand | Binding region | Covalent bonding | Helix | Turn |
|---|---|---|---|---|---|---|---|---|
| 0.464 | 0.347 | 0.535 | 0.579 | 0.502 | 0.506 | 0.485 | 0.579 | 0.503 |
| 0.585 | 0.805 | 0.44 | 0.333 | 0.503 | 0.497 | 0.535 | 0.333 | 0.491 |

| Resolution | Length | ddG | pathprox | Beta strand | Binding region | Covalent bonding | Helix | Turn |
|---|---|---|---|---|---|---|---|---|
| 0.313 | 0.24 | 0.359 | 0.39 | 0.339 | 0.339 | 0.328 | 0.39 | 0.34 |
| 0.363 | 0.436 | 0.317 | 0.286 | 0.337 | 0.338 | 0.348 | 0.286 | 0.336 |

# Conclusion and discussion