Bagging of Neural Networks in Multitask Classification of Biological Activity for Nuclear Receptors

Maria Vladimirova

Moscow Institute of Physics and Technology

Supervisor: Vadim Strijov

Intelligent Data Processing: Theory and Applications
Barcelona, 2016

Problem

Purpose

Building an adequate model to predict the binding of two type molecules: ligands and receptors.

Basic classifier

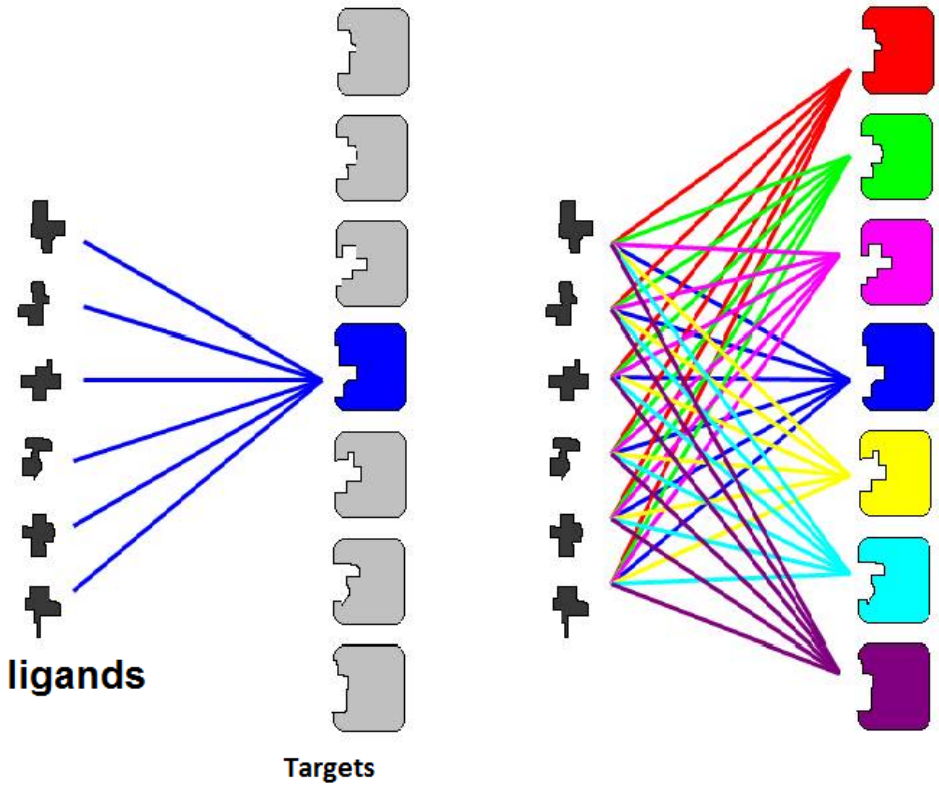
Two layer neural network model.

Proposal

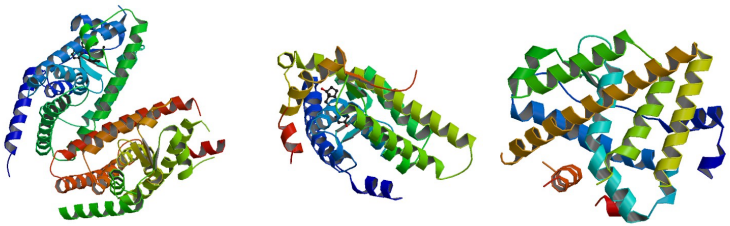
To improve the quality of classication we propose the multitask model of neural networks composition: the bagging procedure.

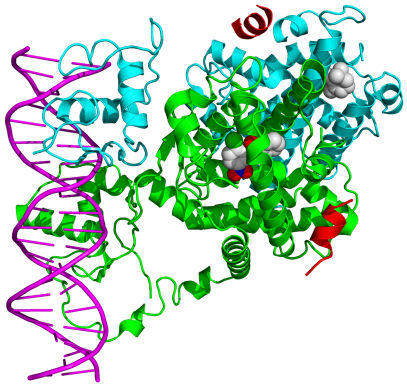
Biological activity for nuclear receptors

When a ligand binds to the nuclear receptor, this receptor binds with other proteins to regulate the expression of specific genes.



Ligand-receptor binding problems





Binding? No binding?

Prediction problems:

– 3D structure of the protein target is often unavailable

– Numerous experiments are conducted for certain receptors whereas almost nothing is studied for others

– The accuracy of ligand-based models quickly degrades when the number of known ligands decreases

Olexandr Isayev, Stephen J. Capuzzi et al. Qsar modeling of tox21 challenge stress response and nuclear receptor signaling toxicity assays, 2016.

Problems solutions

- **Computer modeling of 3D structure.**
  Robert D. Brown, Yvonne C. Martin. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding, 1997.
- **Quantitative structure–activity relationship (QSAR) modeling.**
  Liying Zhang, Hao Zhu, Tudor I. Oprea, Alexander Golbraikh, Alexander Tropsha. QSAR modeling of the blood-brain barrier permeability for diverse organic compounds, 2008.
- **The use of neural networks for researching problems.**
  Humberto Gonzalez-Diaz et al. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds 2007.

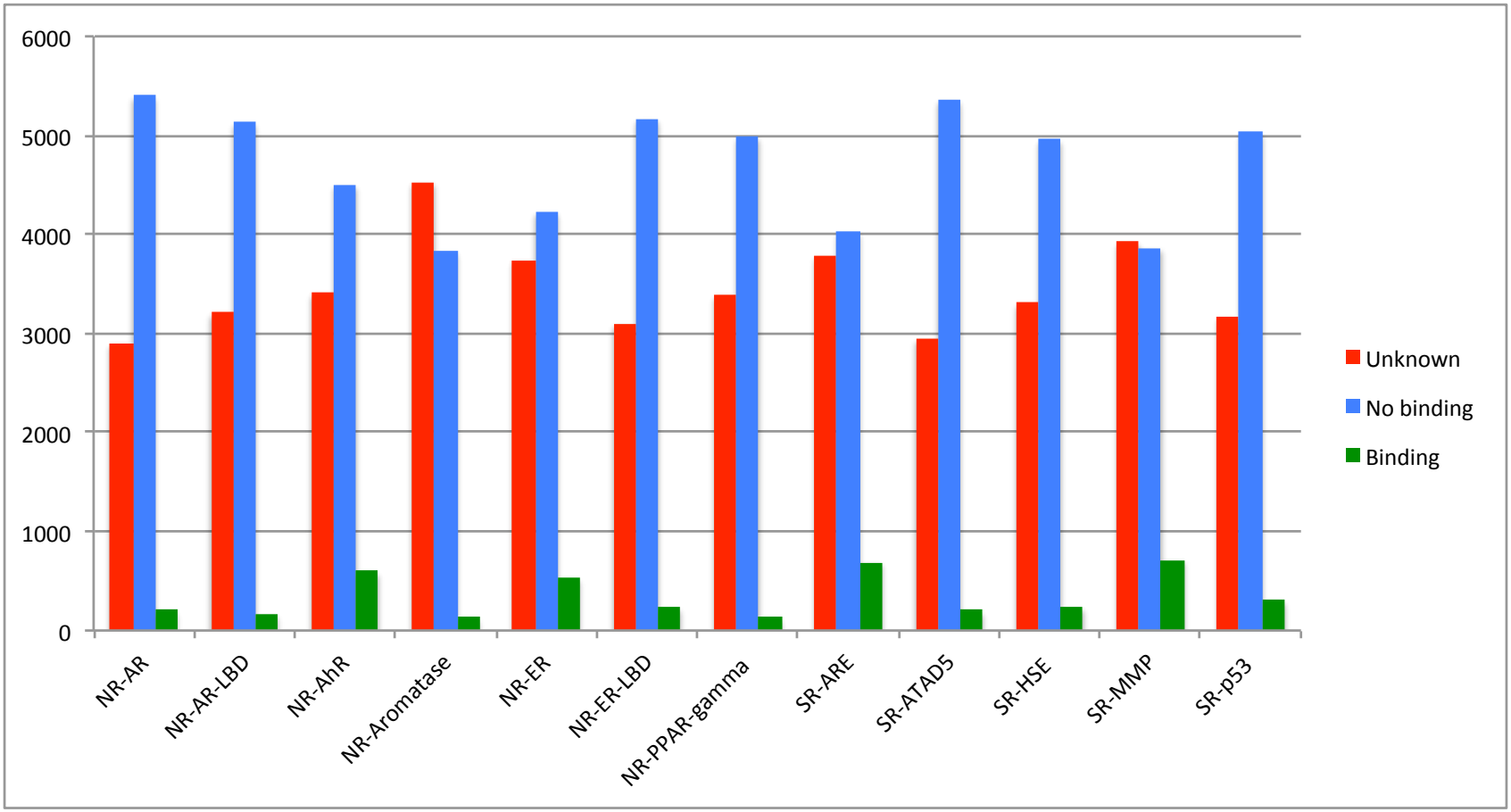
Dataset *Ligand-receptor binding*

Dataset from *Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill*.

- 8513 ligands;
- each ligand has 185 features generated with computer program;
- features are ligand chemical descriptors;
- consist of about 10000 chemical compounds, which are tested on 12 receptors: *NR-AR, NR-AR-LBD, NR-AhR, NR-Aromatase, NR-ER, NR-ER-LBD, NR-PPAR-gamma, SR-ARE, SR-ATAD5, SR-HSE, SR-MMP, SR-p53*;
- possible answers: 0 — no binding, 1 — binding, □ — missing value.

Number of ligand-receptor bindings

The table shows the number of objects with missing value, active compound and unactive compound for each receptor.
There are **16%** of fully marked up objects.



Problem statement

Given

Sample $\mathfrak{L} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, $\mathbf{x}_i \in \mathbb{R}^K$ — objects,
$\mathbf{y}_i \in \{0, 1, \square\}^M$ — M receptors output vector,
$N = 8513, K = 185, M = 12$.

Find

Algorithm $f : \mathbb{R}^K \to \{0, 1\}^M$ that compares each input vector to vector with answers on every receptor.

Cases

- Single-task: for every $m \in \{1, \ldots, M\}$ train separate model
- Multi-task: train joint model with vector output

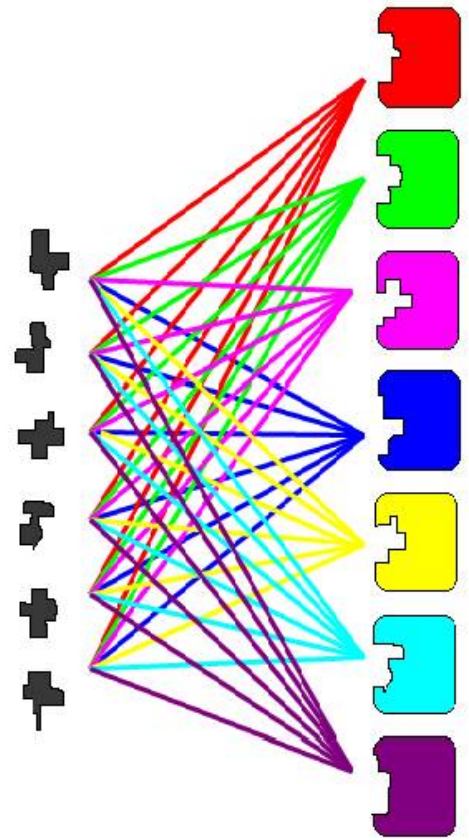
Multi-task learning: learns related problems together

Problem formulation

- Probability distributions $p_1, \ldots, p_M$ on $\mathbb{R}^K \times \mathbb{R}$;
- Data: $\mathbf{d}_i = \left((\mathbf{x}_i^1, y_i^1), \ldots, (\mathbf{x}_i^m, y_i^m)\right) \sim p_i^m$, $m = 1, \ldots, M$;
- Learn predicting models $f(\boldsymbol{\theta}_1), \ldots, f(\boldsymbol{\theta}_M)$, such, that

$$\underset{[\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m] \in \boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{M} \sum_{m=1}^{M} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i^m, f(\mathbf{x}_i^m, \boldsymbol{\theta}_i)),$$

- $\boldsymbol{\theta}$ encourages "common structure" of the tasks.



ligands targets

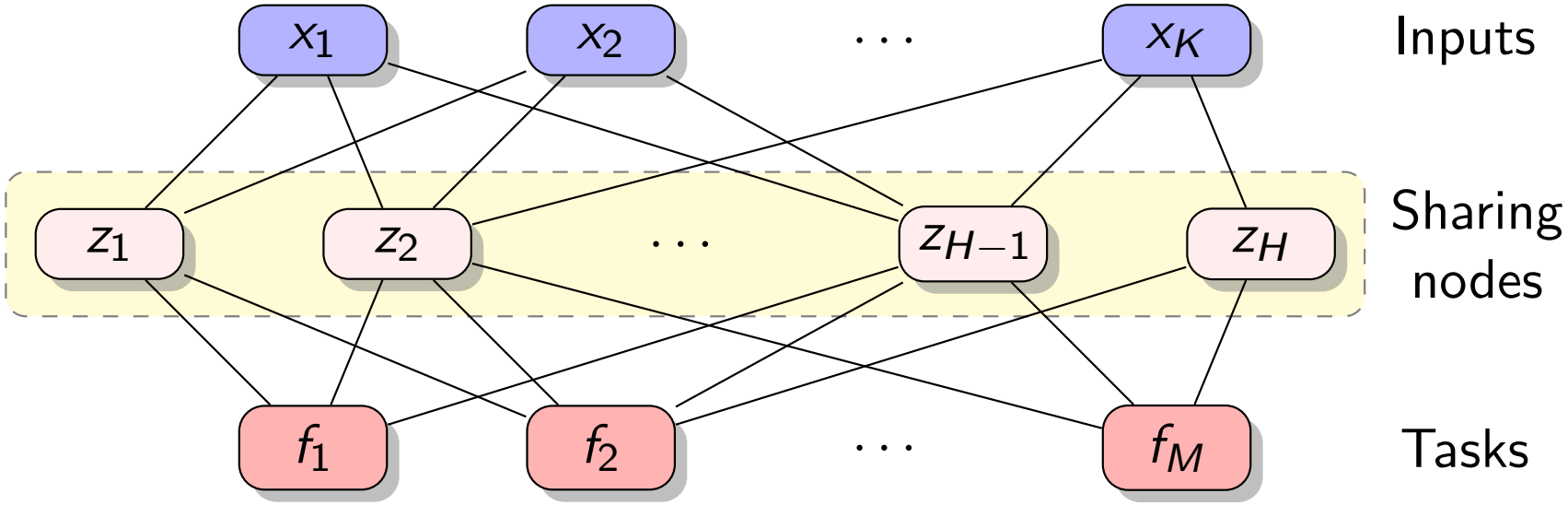
Basic classifier

Two layer neural network

$$\mathbf{z}(\mathbf{x}, \theta) = \mathbf{W}_2^T \mathbf{tanh}(\mathbf{W}_1^T \mathbf{x}) : \mathbb{R}^K \to \mathbb{R}^H,$$

$$\mathbf{f}(\mathbf{x}, \theta) = \frac{1}{1 + \mathbf{exp}\big(-\mathbf{z}(\mathbf{x}, \theta)\big)} : \mathbb{R}^K \to [0,1]^M,$$

$\theta = \mathbf{vec}(\mathbf{W}_1^T | \mathbf{W}_2^T)$ — vector of neural network parametrs,
H — number of hidden neurons.



Probabilies of ligand-receptor binding

Assumption

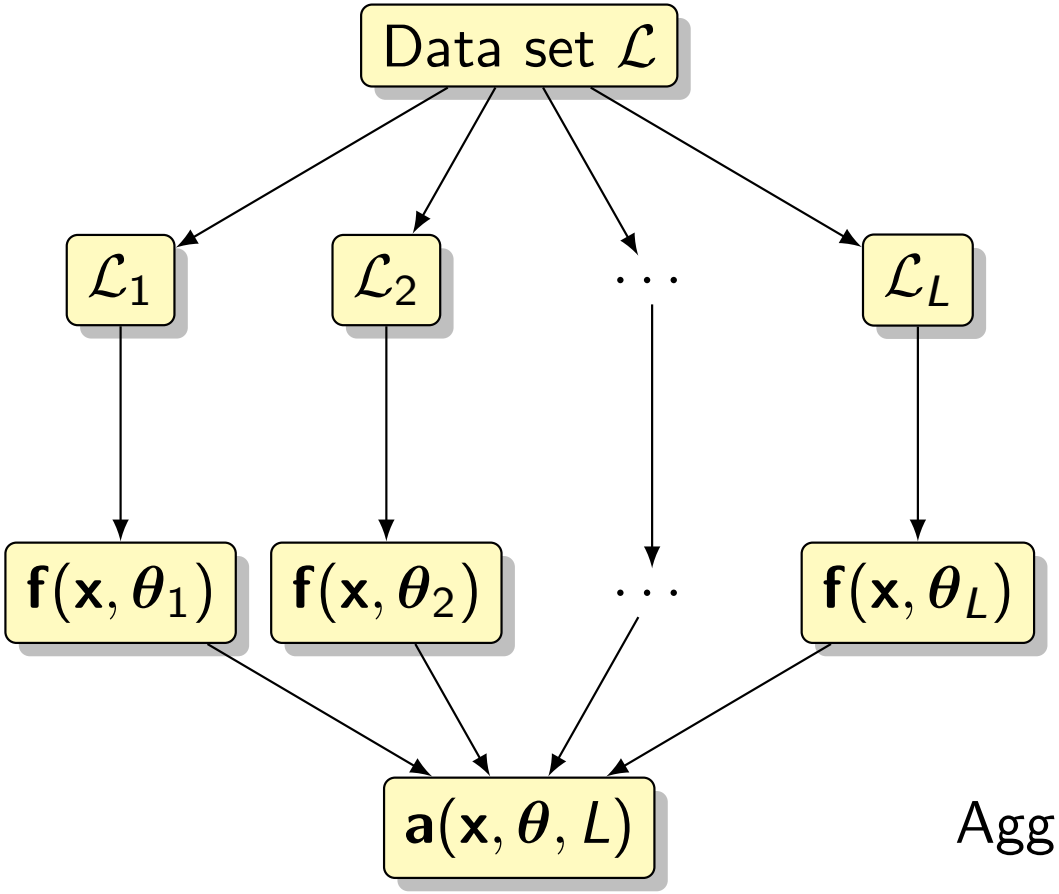
y is a random vector realization, each component of which has a Bernoulli distribution.

Statement

Elements $\mathbf{f}(\mathbf{x}, \mathbf{w})$ are probabilities of that the ligand $\mathbf{x}$ binds to receptors:

$$\mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) = \begin{bmatrix} P(y_1 = 1|\mathbf{x}, \boldsymbol{\theta}) \\ P(y_2 = 1|\mathbf{x}, \boldsymbol{\theta}) \\ \vdots \\ P(y_M = 1|\mathbf{x}, \boldsymbol{\theta}) \end{bmatrix}.$$

Bagging improves the quality of classification



Bootstrap L times: 63% of the unique elements in the subset

Train L estimators

Aggregate answers of estimators

Classification model: simple voting

$$\mathbf{a}(\mathbf{x}, \theta, L) = \sum_{\ell=1}^{L} \pi_\ell \mathbf{f}(\mathbf{x}, \theta^\ell), \ \pi_\ell = \frac{1}{L}.$$

Optimization of error functions

$a^m(\mathbf{x}_i, \theta, L)$ — the m-th element of classifier's answer on object $\mathbf{x}_i$,
y_i^m — the m-th element of trully answer on object $\mathbf{x}_i$.

Mean squared error

$$\mathcal{L}_1(\theta, \mathbf{x}_i, \mathbf{y}_i) = \frac{1}{2} \sum_{m=1}^{M} \left(a^m(\mathbf{x}_i, \theta, L) - y_i^m\right)^2.$$

Cross-entropy error

$$\mathcal{L}_2(\theta, \mathbf{x}_i, \mathbf{y}_i) = -\sum_{m=1}^{M} y_i^m \log P(y_i^m = 1|\mathbf{x}_i, \theta) + \\ + (1 - y_i^m) \log P(1 - y_i^m = 1|\mathbf{x}_i, \theta).$$

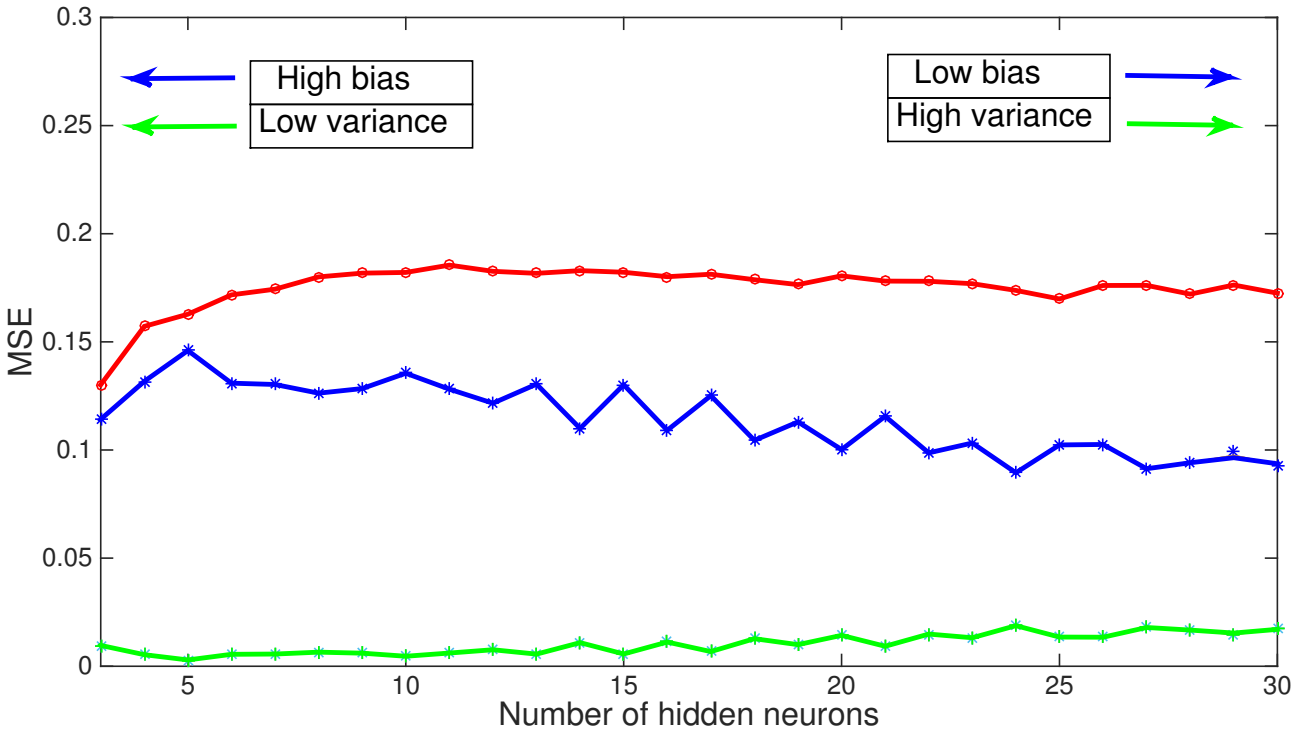
Optimization problem

$$\hat{\theta}, \hat{L} = \underset{\theta, L}{\operatorname{argmin}} \sum_{i=1}^{|\mathfrak{L}|} \mathcal{L}(\theta, \mathbf{x}_i, \mathbf{y}_i), \ \mathcal{L} \in \{\mathcal{L}_1, \mathcal{L}_2\}.$$

Decomposion of the expected error helps to improve the data fitting proccess

Bias-variance tradeoff

$$\mathcal{L}(a) = \text{Noise} + \text{Bias} + \text{Variance}.$$



Bias, variance and average error for neural networks depending on number of hidden neurons.

Mean squared error (MSE)

The MSE of an estimator y with respect to an unknown $f(x)$:

$$R(f) = \mathrm{E}_{x,y}\left[\left(y - f(x)\right)^2\right].$$

Condition for a minimum:

$$\operatorname*{argmin}_f R(f) = P(y = 1|x) = \mathrm{E}[y|x].$$

The main measure of quality for any algorithm a:

$$\mathcal{L}(a) = \mathrm{E}_{\mathfrak{L}}\left[\mathrm{E}_{x,y}\left[\left(y - a(x, \mathfrak{L})\right)^2\right]\right].$$

- Bagging unchanges bias.
- Bagging reduces variance L times.

Cross-entropy function

$y = \{0, 1\}$ — binary classes, p — trully probability $P(y = 1|x)$,
f — hypothesys probability $P(y = 1|x)$.
Cross-entropy corresponds Kullback–Leibler divergence:

$$D_{\mathrm{KL}}(p, f) = p \ln \frac{p}{f} + (1 - p) \ln \frac{1 - p}{1 - f}.$$

Problem statement:

$$\underset{f \in [0,1]}{\operatorname{argmin}} \mathrm{E}_{x,y} \Big[D_{\mathrm{KL}}(y, f) \Big] = f^*(x), \;\; f^*(x) = \frac{1}{Z} \exp\Big(\mathrm{E}_{x,y} \big[\ln f(x) \big] \Big).$$

The main measure of quality for any algorithm a:

$$\mathcal{L}(a) = \mathrm{E}_{\mathfrak{L}} \Big[\mathrm{E}_{x,y} \Big[D_{\mathrm{KL}}(y, a(x, \mathfrak{L})) \Big] \Big].$$

- Bagging unchanges bias.
- Bagging reduces variance.

Modified BackPropagation

Gradient method of optimization: $w := w - \eta \nabla \mathcal{L}(\mathbf{w}, \mathbf{x}, \mathbf{y})$.

BackProp:

forward pass:
predict with neural network;
compute error at the output units;
if there is a missing value in the data, assume error is equal to zero;

backward pass:
compute error at the hidden units in dependence on error function;

weight update: gradient method.

Computational experiment

Purposes

- Apply the proposed algorithm on real data.
- Compare results from the multitask model with the single model.

Dataset

Sample $\mathfrak{L} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, $\mathbf{x}_i \in \mathbb{R}^K$ — objects,
$\mathbf{y}_i \in \{0, 1, \square\}^M$ — M receptors output vector,
$N = 8513, K = 185, M = 12$.

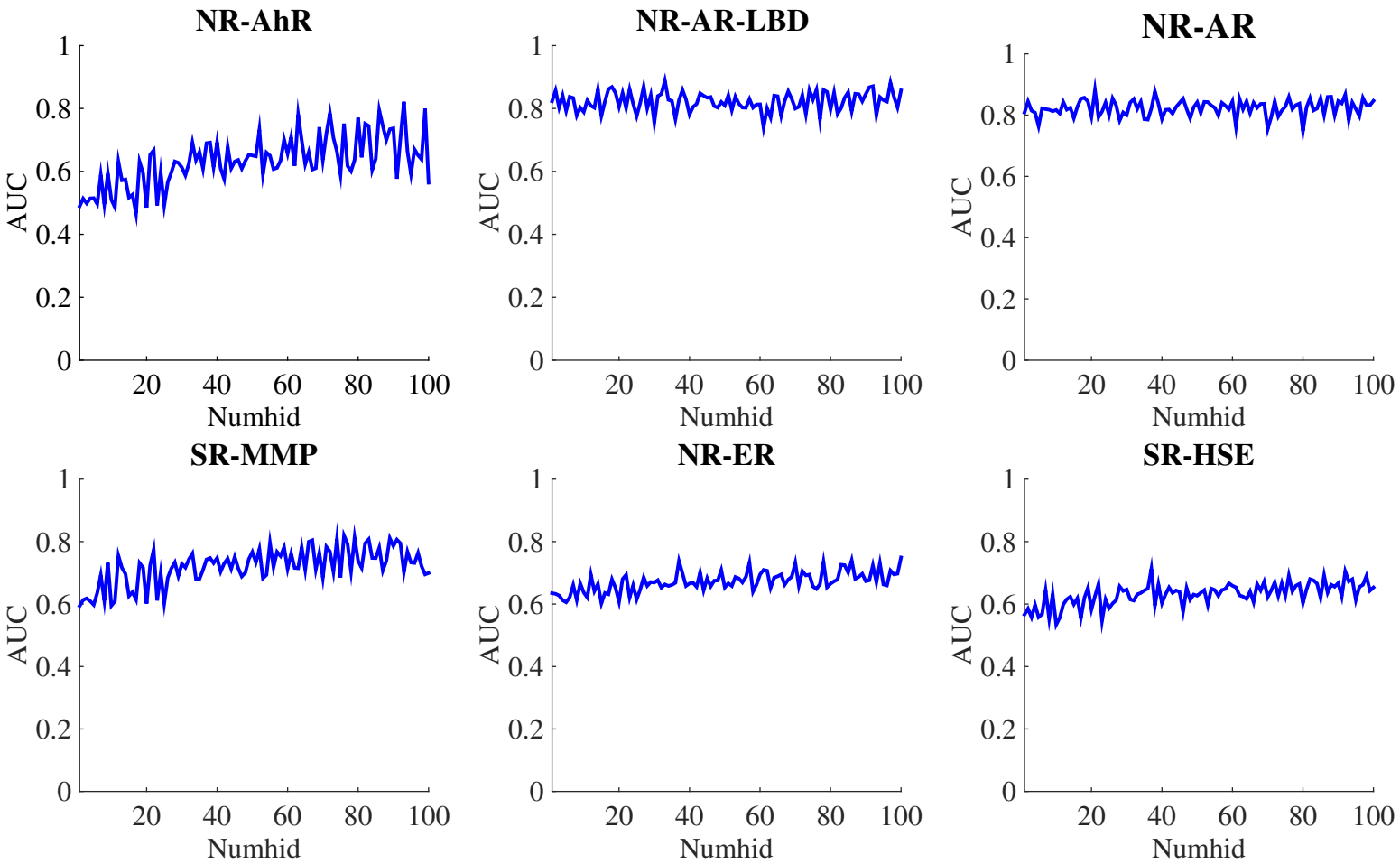
Problem

Minimaze loss function $Q(\theta, L|\mathfrak{L}) = \sum_{i=1}^{|\mathfrak{L}|} \mathcal{L}(\theta, \mathbf{x}_i, \mathbf{y}_i)$, where
$\mathcal{L}$ — mean squared error or cross-entropy error.

Quality measure

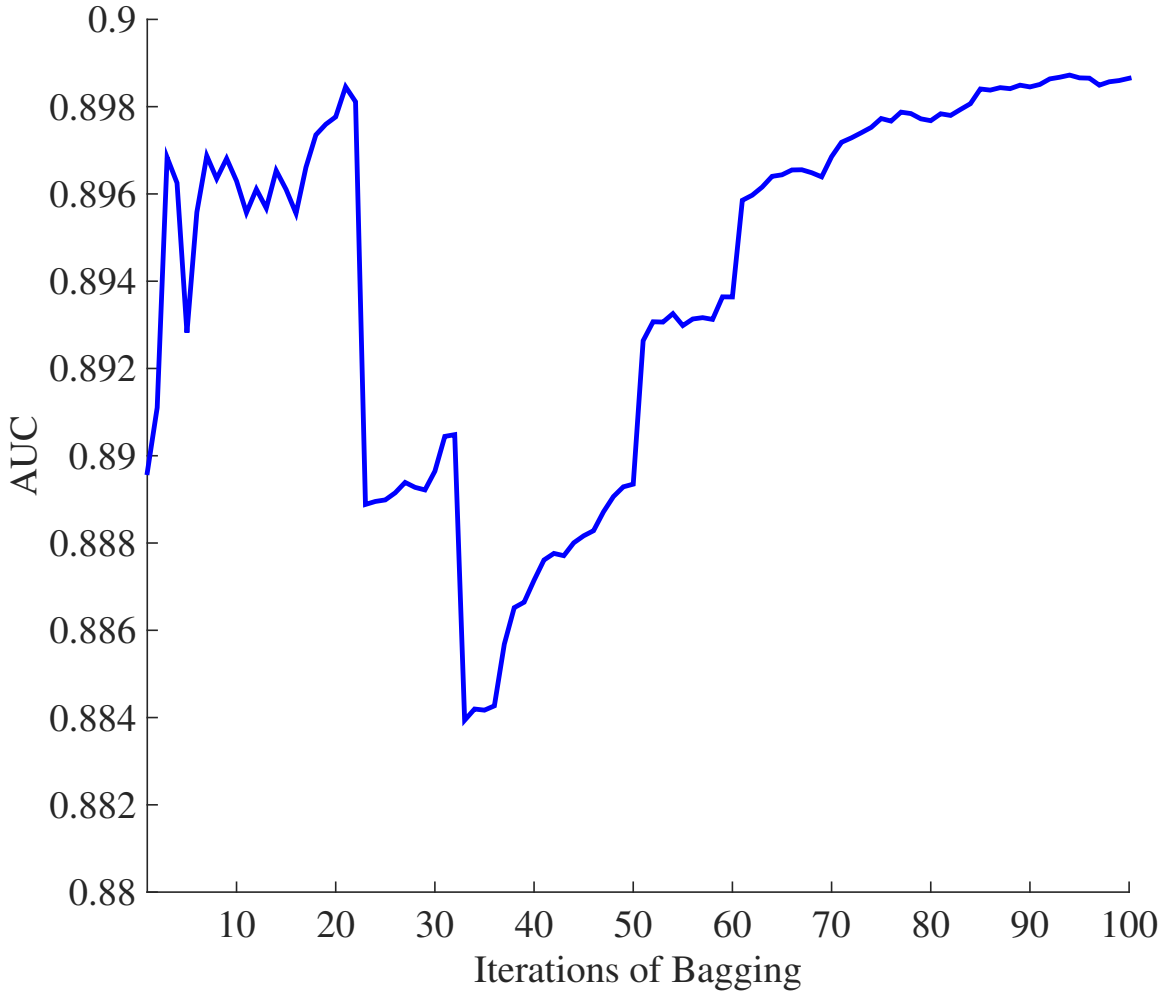
AUC, 5-fold cross-validation.

Dependence AUC on number of hidden layer neurons H



Choose a number H from which the accuracy in all receptors is approximately constant.

Dependence AUC on number of bagging iterations

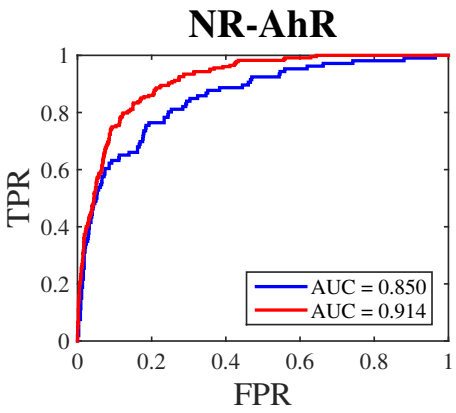


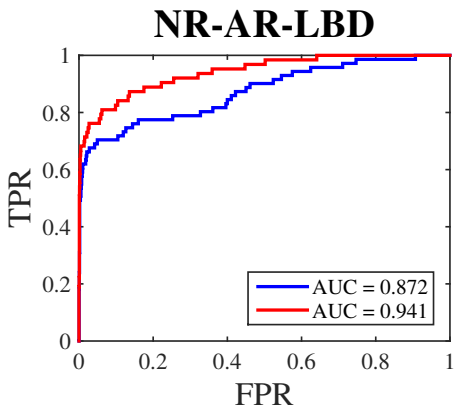
Choose a number L from which the accuracy in all receptors is stabilized.

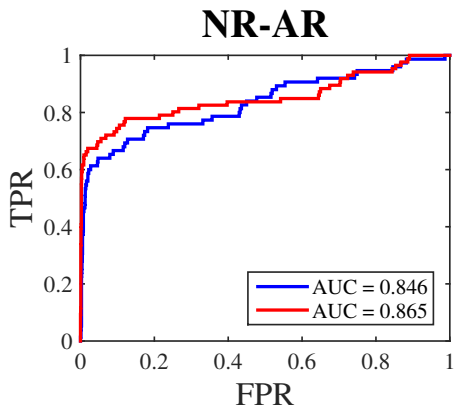
Results

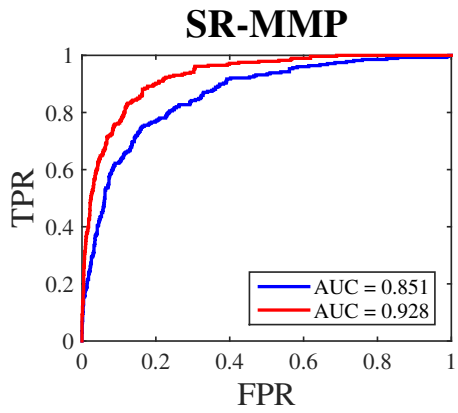
Receptor	Neural Networks	Bagging (cross-entropy)	Bagging (squared)
NR-AhR	0.8589 ± 0.0216	**0.9089 ± 0.0210**	**0.9088 ± 0.0174**
NR-AR-LBD	0.8725 ± 0.0455	**0.9138 ± 0.0064**	**0.9207 ± 0.0458**
NR-AR	0.8456 ± 0.0294	**0.8658 ± 0.0129**	**0.8704 ± 0.0166**
SR-MMP	0.8512 ± 0.0483	**0.9132 ± 0.0110**	**0.9161 ± 0.0109**
NR-ER	0.7585 ± 0.0726	**0.8109 ± 0.0329**	**0.8151 ± 0.0253**
SR-HSE	0.7189 ± 0.0583	**0.8274 ± 0.0193**	**0.8380 ± 0.0347**
SR-p53	0.7345 ± 0.0838	**0.8532 ± 0.0257**	**0.8585 ± 0.0204**
NR-PPAR-gamma	0.7610 ± 0.0725	**0.8435 ± 0.0437**	**0.8539 ± 0.0171**
SR-ARE	0.7698 ± 0.0307	**0.8265 ± 0.0208**	**0.8268 ± 0.0076**
NR-Aromatase	0.7808 ± 0.0482	**0.8697 ± 0.0308**	**0.8676 ± 0.0218**
SR-ATAD5	0.8338 ± 0.0714	**0.8682 ± 0.0187**	**0.8629 ± 0.0332**
NR-ER-LBD	0.8299 ± 0.0241	**0.8917 ± 0.0267**	**0.8884 ± 0.0168**

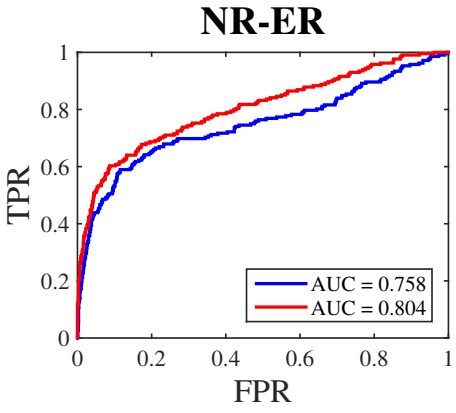
Bagging of neural networks ROC-curve

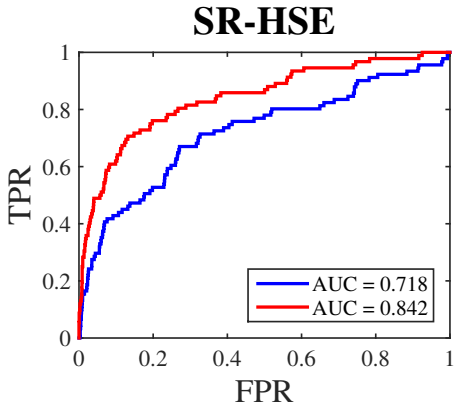


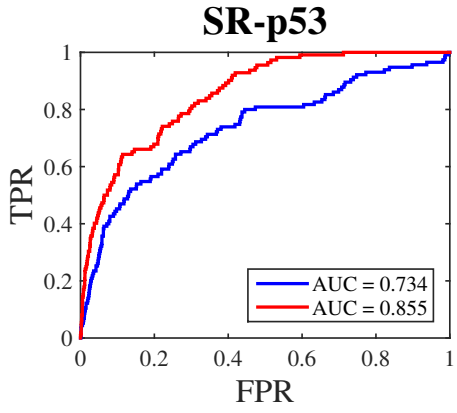


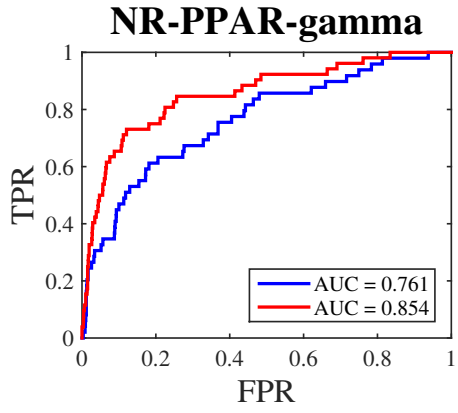


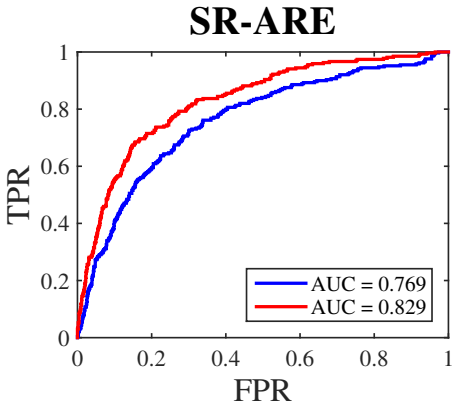


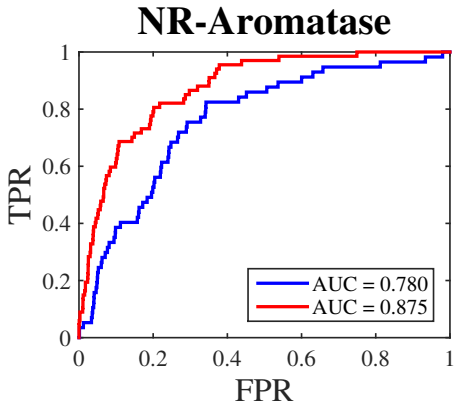


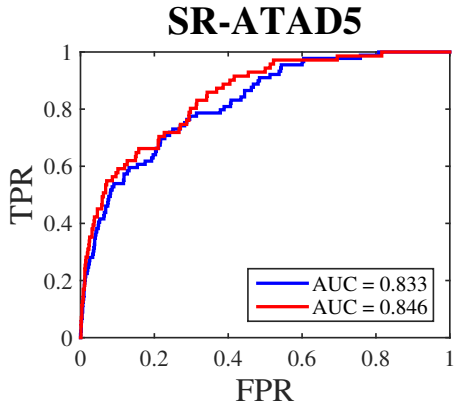


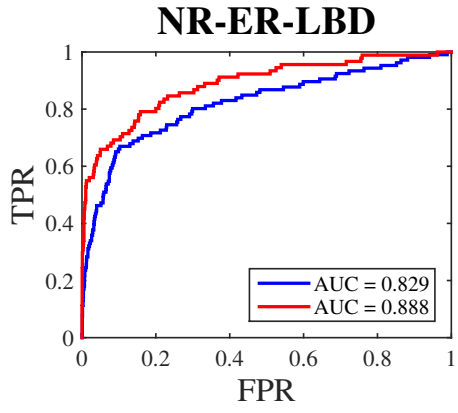












Conclusion

Obtained dependence ligand-receptor binding on molecule chemical features.

Improved prediction quality by

- considering the multitask problems of linear and logistic regressions with squared and cross-entropy loss functions;
- analyzing error decomposition into bias and variance terms;
- using bagging of neural networks.