university of groningen

Winning the Rat Race

Sub-challenge 2:
Inter Species Prediction of Protein Phosphorylation

Gyan Bhanot — Rutgers University

Michael Biehl — University of Groningen

Adel Dayarian — Kavli Institute of Theoretical Physis

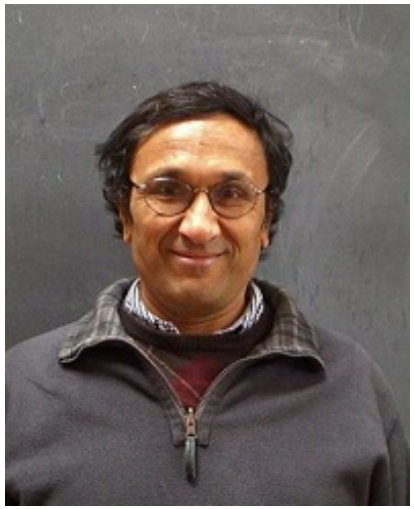
Sahand Hormoz — UC Santa Barbara

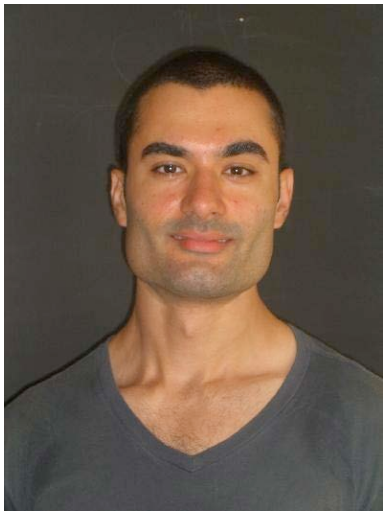
university of groningen

The team

Gyan Bhanot
Rutgers University and
Inst. for Adv. Study, Princeton

Sahand Hormoz **Adel Dayarian**
Kavli Institute for Theoretical Physics,
University of California, Santa Barbara





overview

set-up: phosphorylation data only

naïve prediction:

assume “human ≈ rat” (in terms of phosphorylation)

machine learning approach:

Learning Vector Quantization (LVQ) classifier

combination of predictions:

based on validation procedures to estimate performances

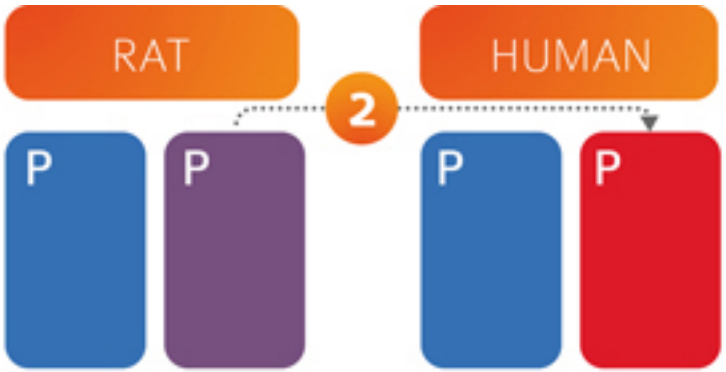
sub-challenge set-up

www.sbvimprover.com

Legend:

- P — Phosphorylation
- GEx — Gene expression
- A B — Stimulus subset
- Not provided
- Provided data
- Predicted data
- Provided after 1 July

sub-challenge set-up

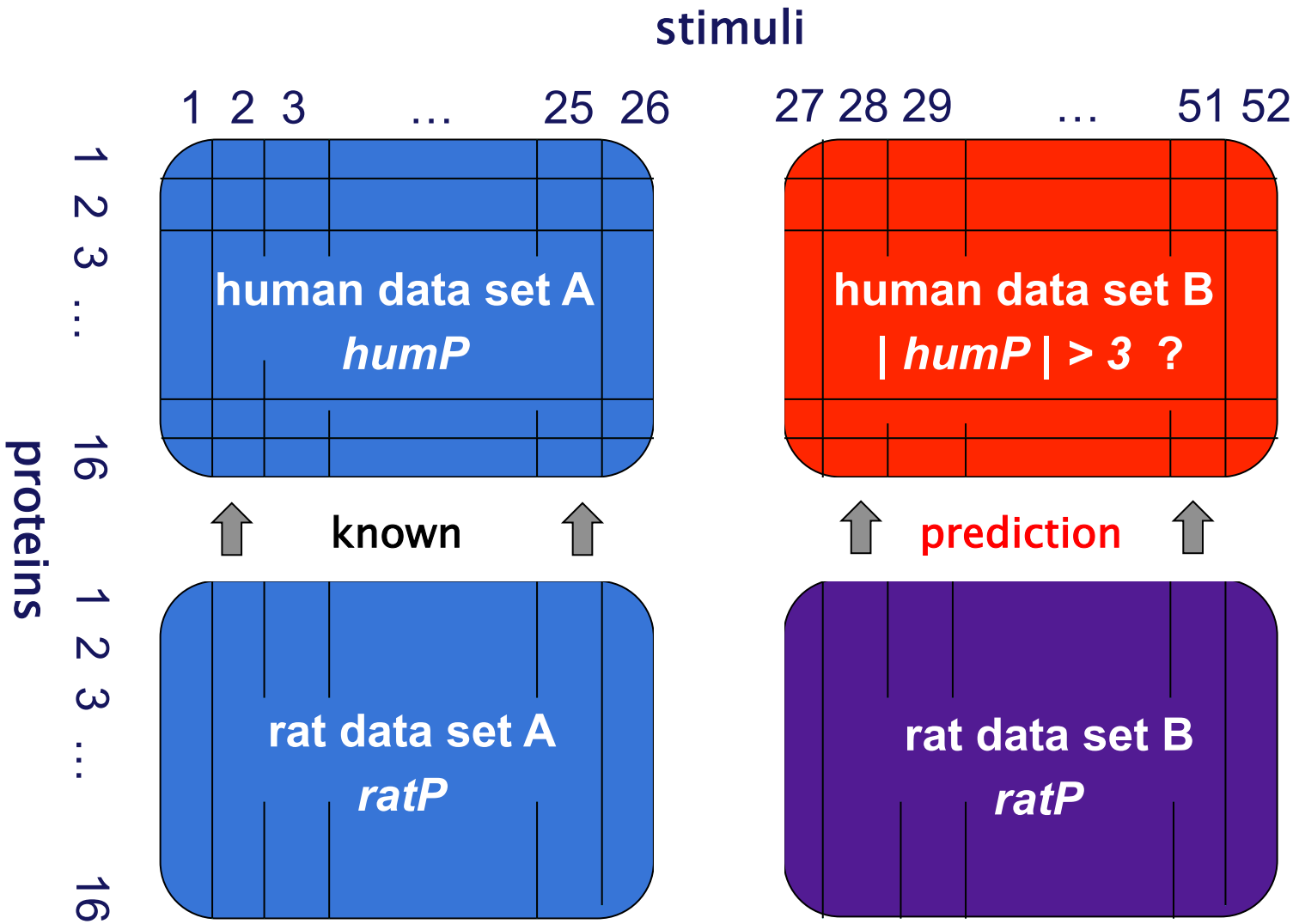


www.sbvimprover.com

preprocessing: median of 3 replicates
maximum absolute level of (@5 min, @25 min)

target: recommended threshold of 3.0
for phosphorylation

data


stimuli
1 2 3 … 25 26
27 28 29 … 51 52
proteins
1 2 3 … 16
human data set A
humP
human data set B
| humP | > 3 ?
known
prediction
1 2 3 … 16
rat data set A
ratP
rat data set B
ratP

naïve prediction

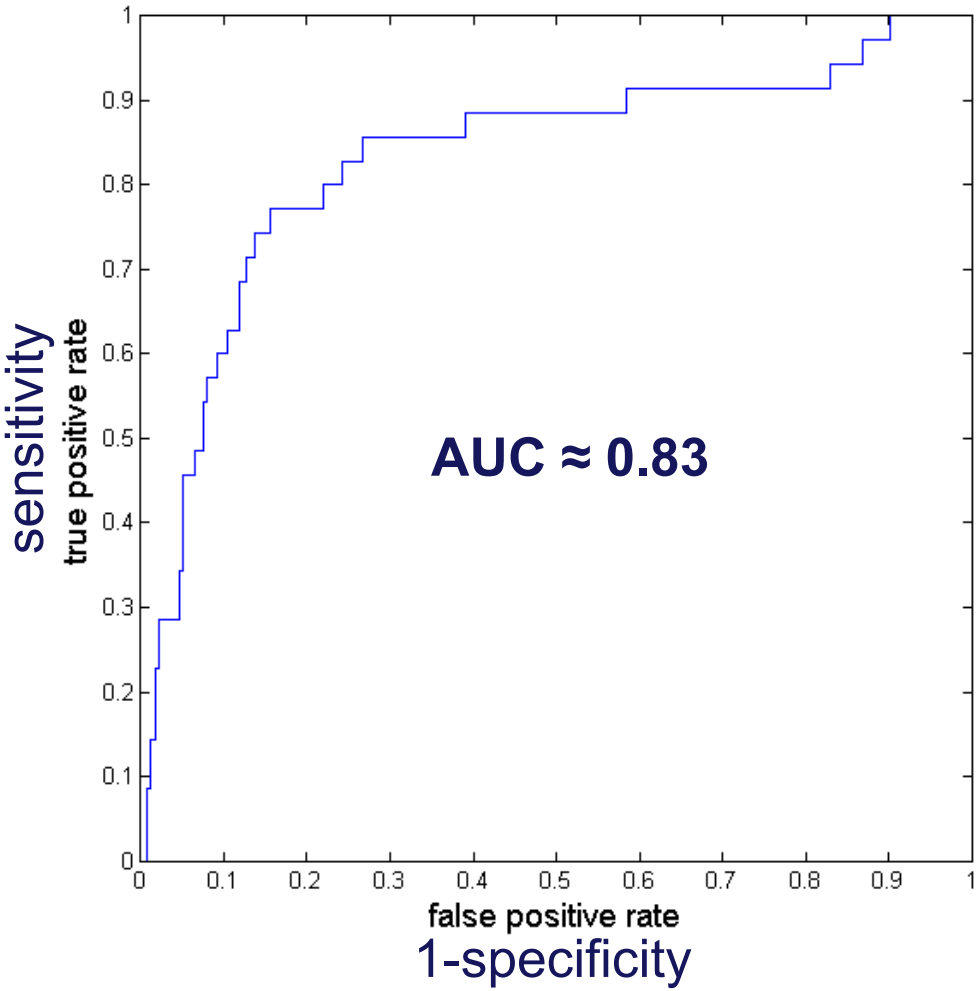
assume similar activation in both species: "**human ≈ rat**"

prediction *score*, corresponding to **threshold 3** for activation

$$c_{naive} = \frac{1}{2}\left[1 + \tanh\left(\frac{|ratP| - 3}{5}\right)\right] \quad \in [0, 1]$$

- precise (monotonic!) form is irrelevant for ROC, PR etc.
- threshold 0.5 for crisp classification
- here: scaling factor yields values well-spread in [0,1]

naïve prediction



ROC

with respect to the full panel

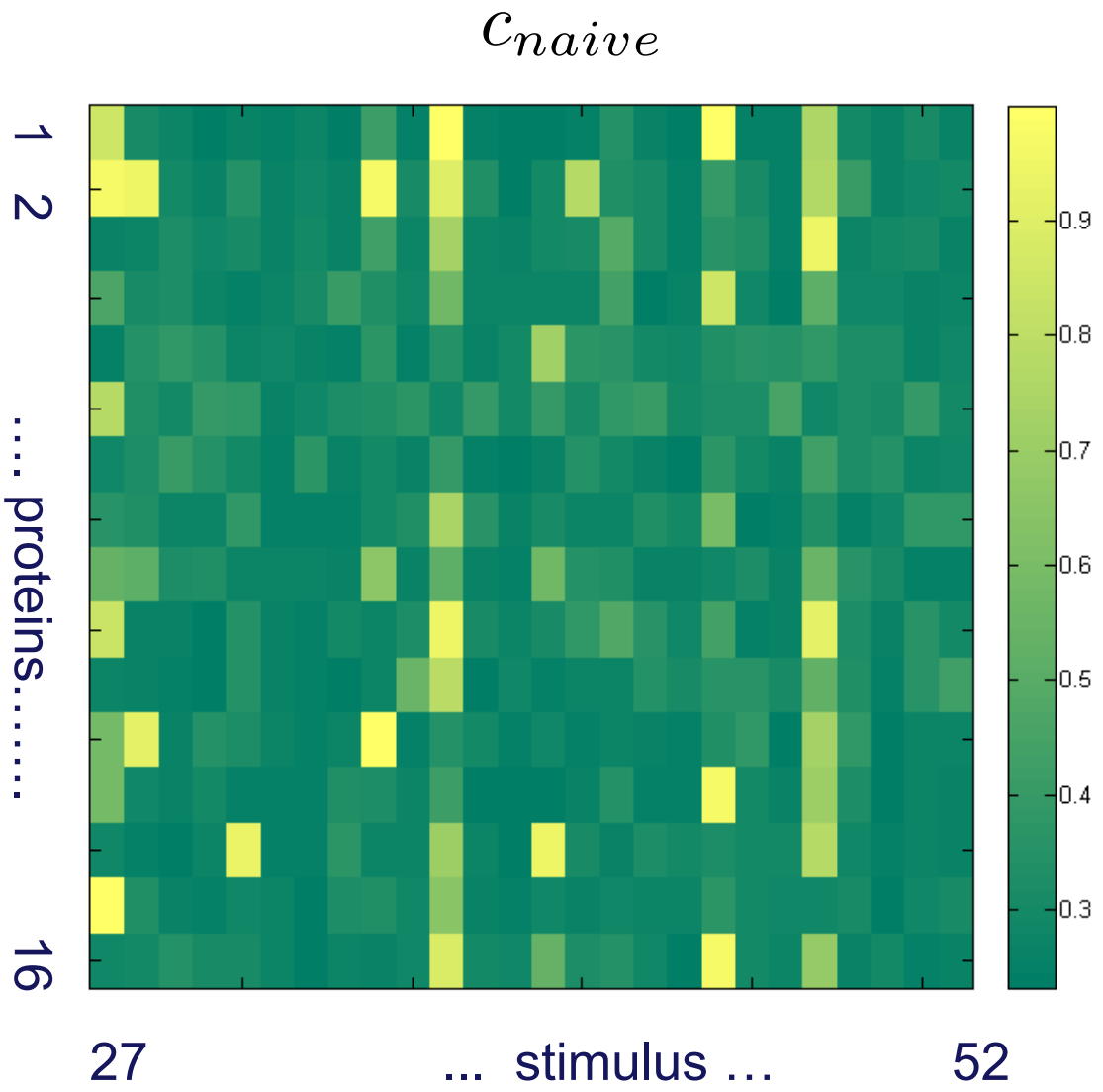
(416 predictions) of

| *humP* | > 3

AUC: probability for $c_{naive}(+) > c_{naive}(-)$

for a random pair of 1 positive and 1 negative sample

naïve prediction

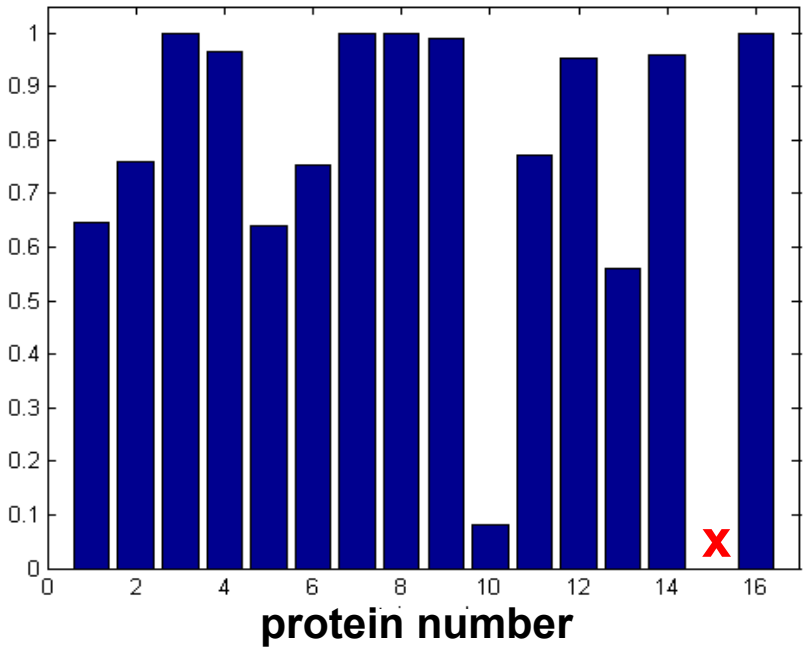


color-coded certainty

for $|humP|>3$

in **data set B**

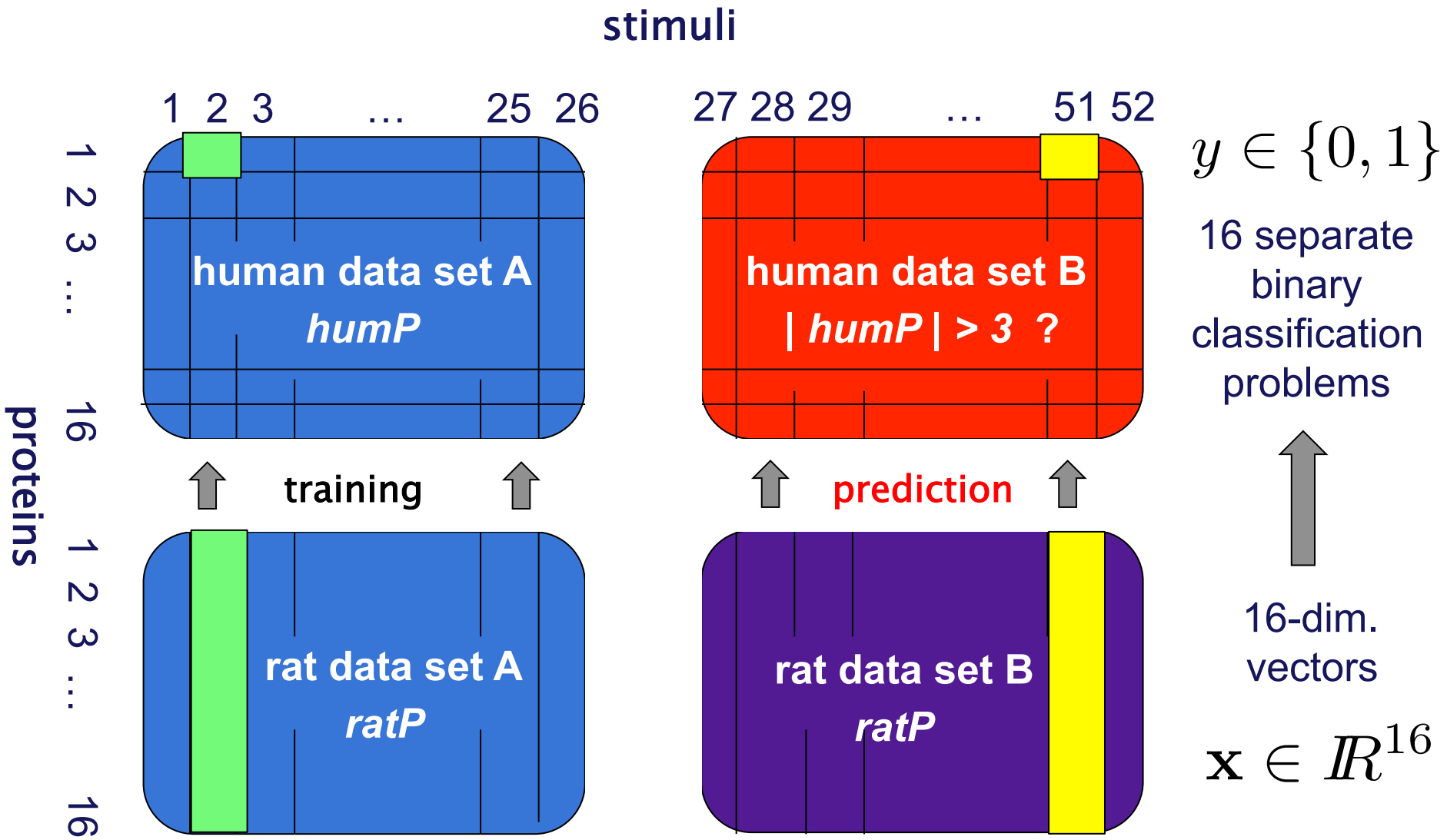
naïve prediction



protein specific AUC (ROC)
for prediction of **data set A**

x (only negative samples)

machine learning approach



university of groningen

Learning Vector Quantization

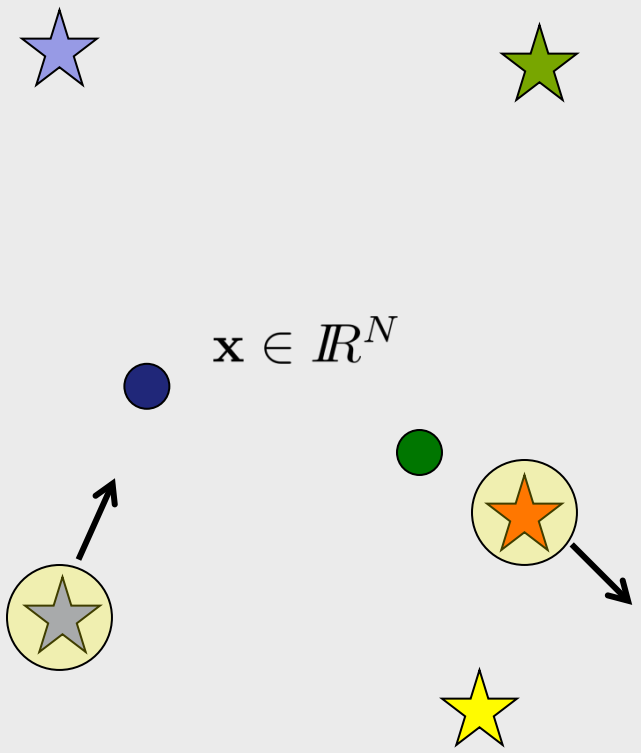
N-dimensional data, feature vectors

· identification of **prototype vectors** from labeled example data

· distance based **classification** (e.g. Euclidean)

heuristic scheme: **LVQ1** [Kohonen, 1990]

- **initialize** prototype vectors for different classes
- present a **single example**
- identify the **winner** (closest prototype)
- move the winner
  - closer **towards the data** (same class)
  - **away from the data** (different class)
- present training data many times



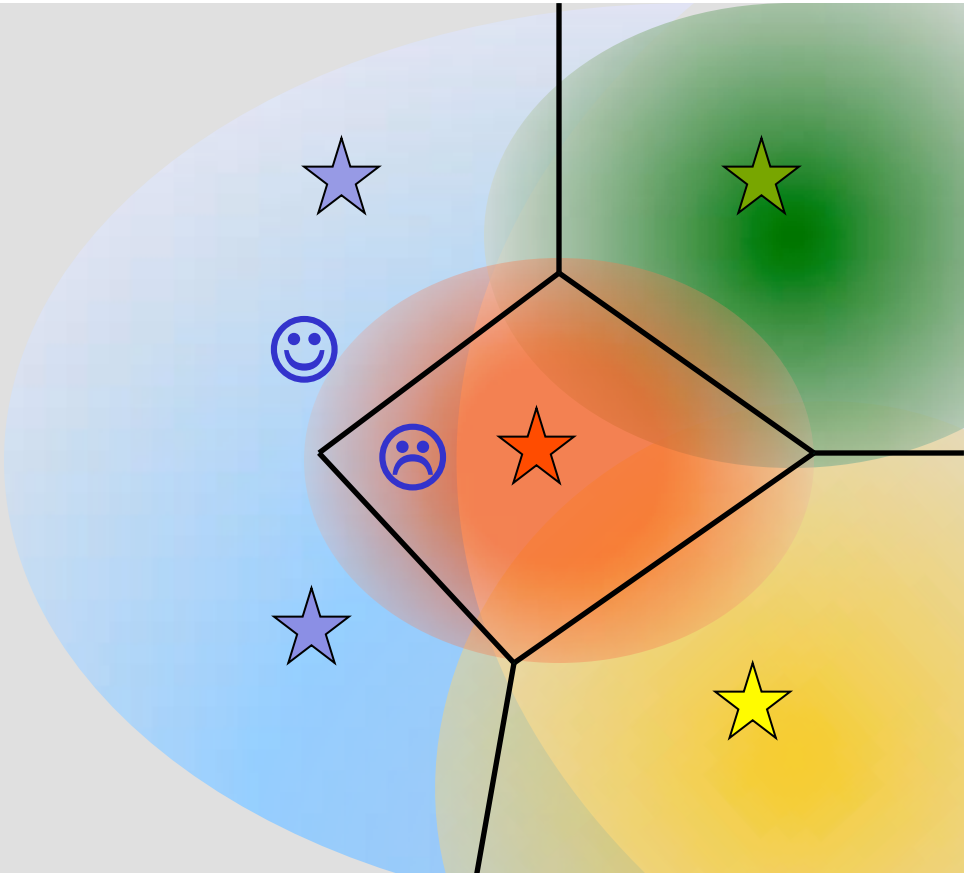
university of groningen

Learning Vector Quantization

N-dimensional data, feature vectors

- identification of **prototype vectors** from labeled example data
- distance based **classification** (e.g. Euclidean)

- **distance-based classification**
  [here: Euclidean distances]
- **tesselation of feature space**
  [piece-wise linear]
- aim: **discrimination of classes**
  (≠ vector quantization
  or density estimation)
- **generalization ability**
  correct classification of *new* data



LVQ prediction

here: 16-dim. data $\mathbf{x} \in I\!R^{16} \quad y(\mathbf{x}) \in \{0, 1\}$

one prototype per class $\mathbf{w}^0, \mathbf{w}^1 \in I\!R^{16}$

Nearest prototype classification:

$$d^0 = \sum_i (w_i^0 - x_i)^2$$

$$d^1 = \sum_i (w_i^1 - x_i)^2$$

$$y(\mathbf{x}) = \begin{cases} 1 & \text{if } d^1(\mathbf{x}) < d^0(\mathbf{x}) \\ 0 & \text{else.} \end{cases}$$

LVQ prediction

prediction *score / certainty* for activation

$$c_{LVQ} = \frac{1}{2}\left[1 + \tanh\left(\frac{d^0 - d^1}{200}\right)\right]$$

- precise (monotonic!) form is irrelevant for ROC, PR etc.
- crisp classification for threshold 0.5
- here: scaling factor yields range of values similar to naïve prediction

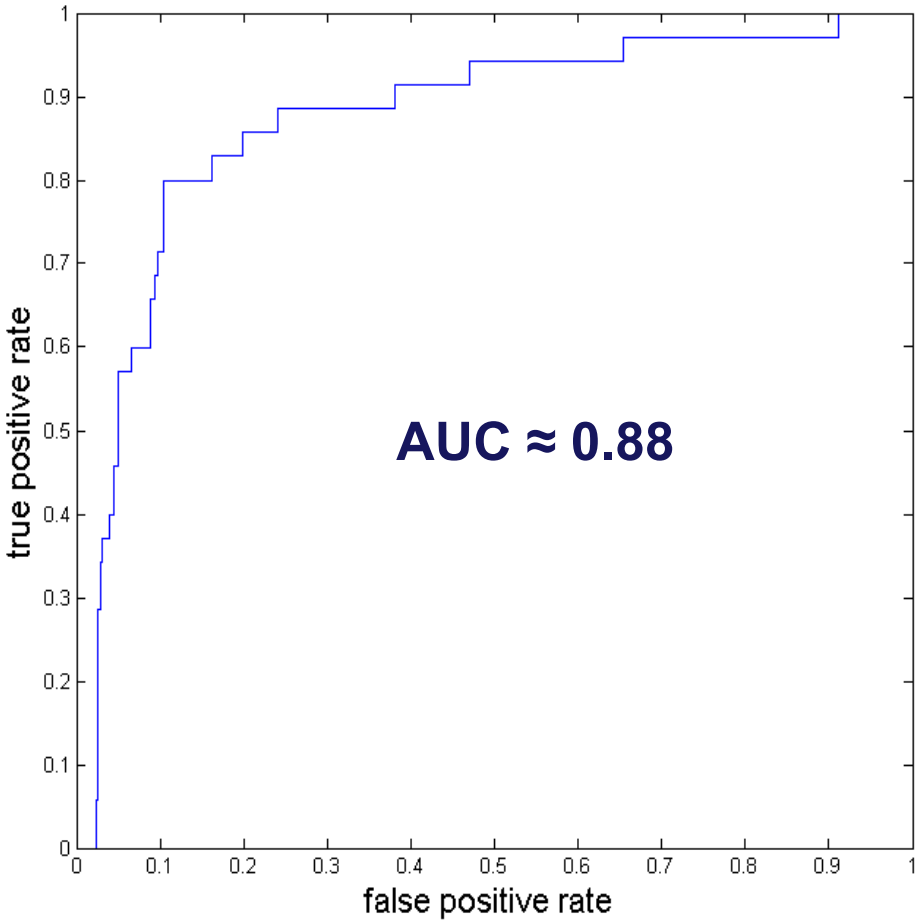
validation: 26 Leave-One-Out training processes:

split data set A (complete) in 25 training / 1 test sample

(if training set is all negative: accept naïve prediction)

prediction: *ensemble average* of certainties over the 26 LVQ systems

LVQ prediction



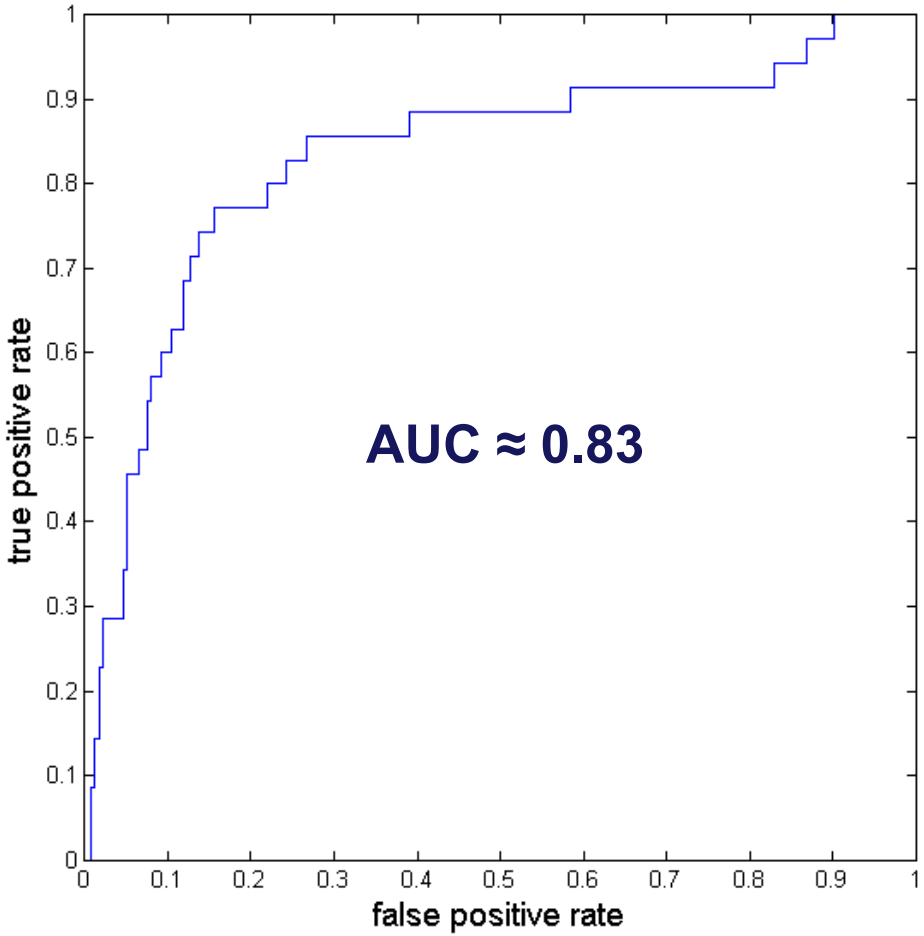
ROC

with respect to the full panel (416 predictions) of

$$|\, humP \,| > 3$$

obtained in the Leave-One-Out validation scheme

naïve prediction

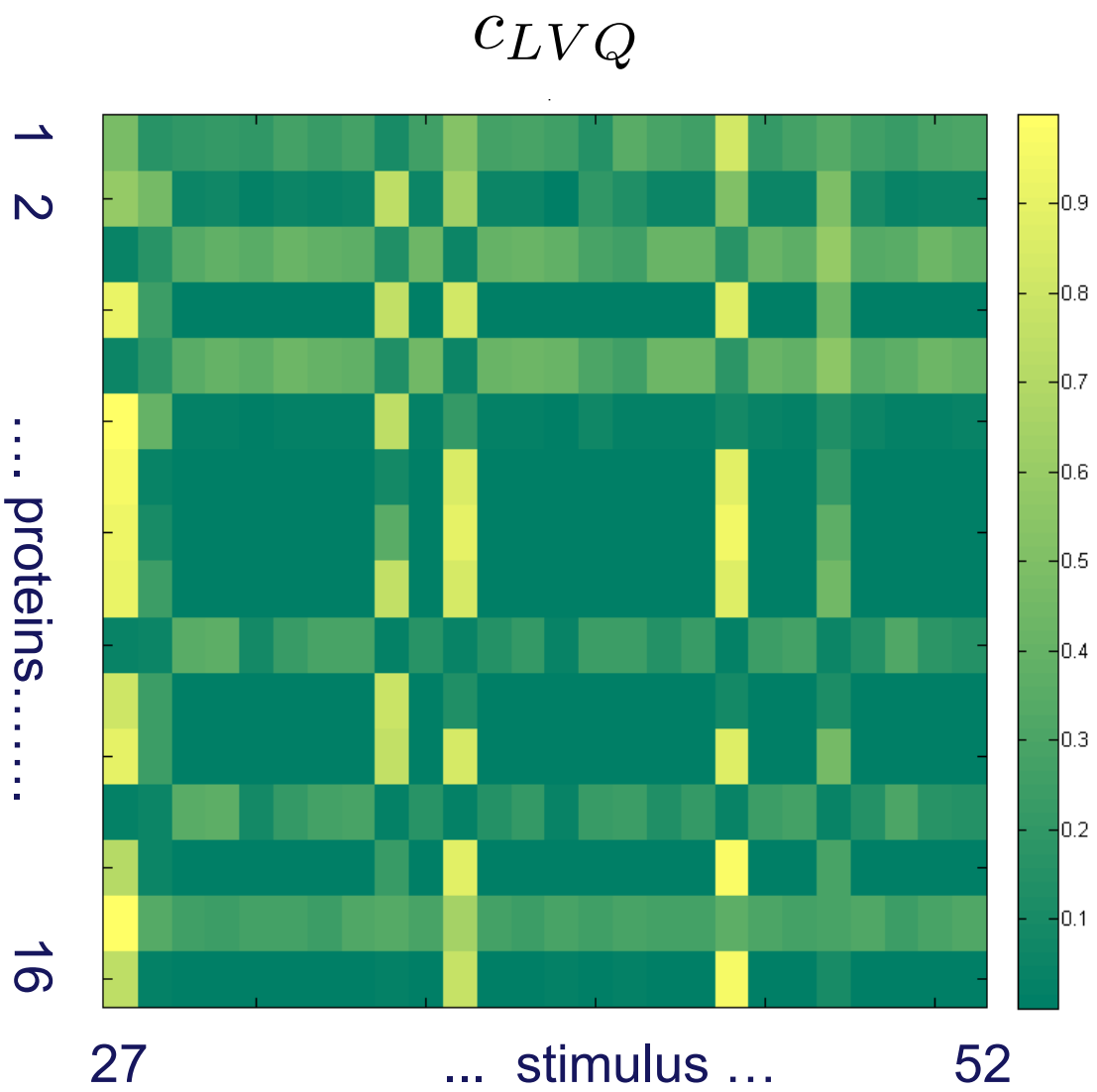


ROC

with respect to the full panel (416 predictions) of

$| humP | > 3$

LVQ prediction

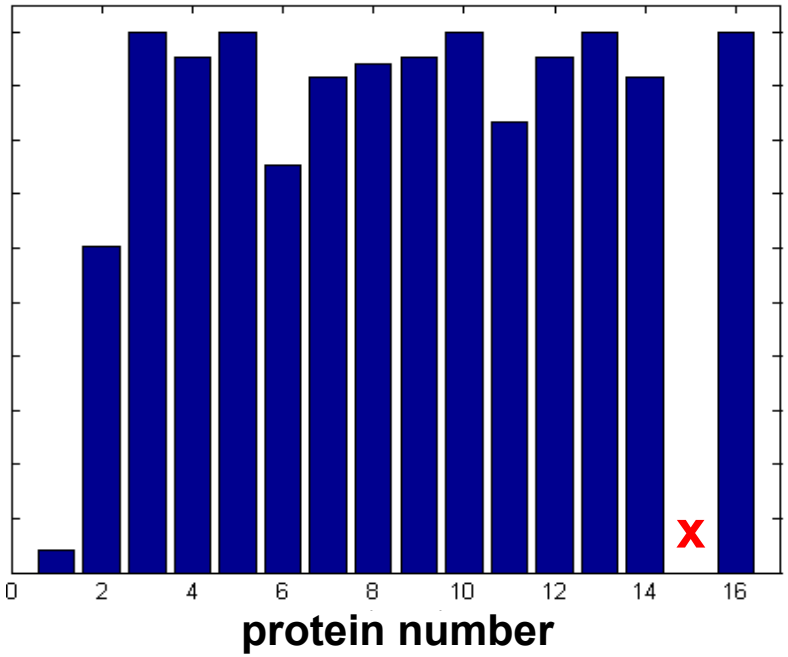


color-coded certainty

for $|humP|>3$

in **data set B**

LVQ prediction



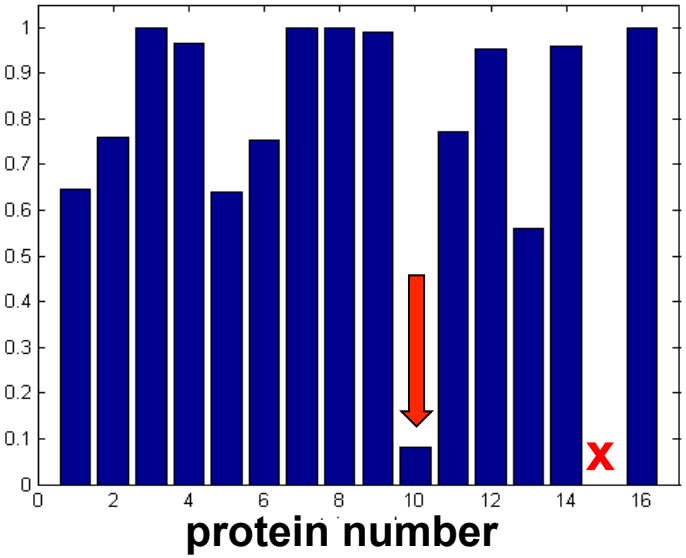
protein specific AUC (ROC) for prediction of **data set A**

x (only negative samples)

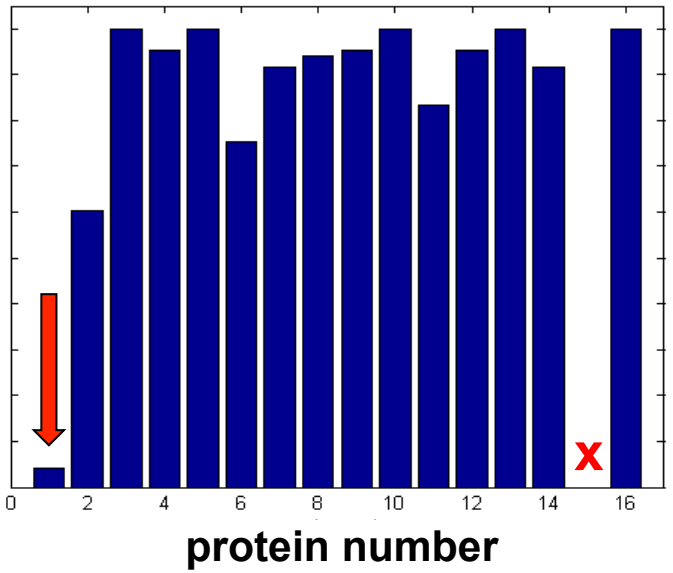
weighted combination

protein specific AUC (ROC) in data set A

c_{naive}



c_{LVQ}



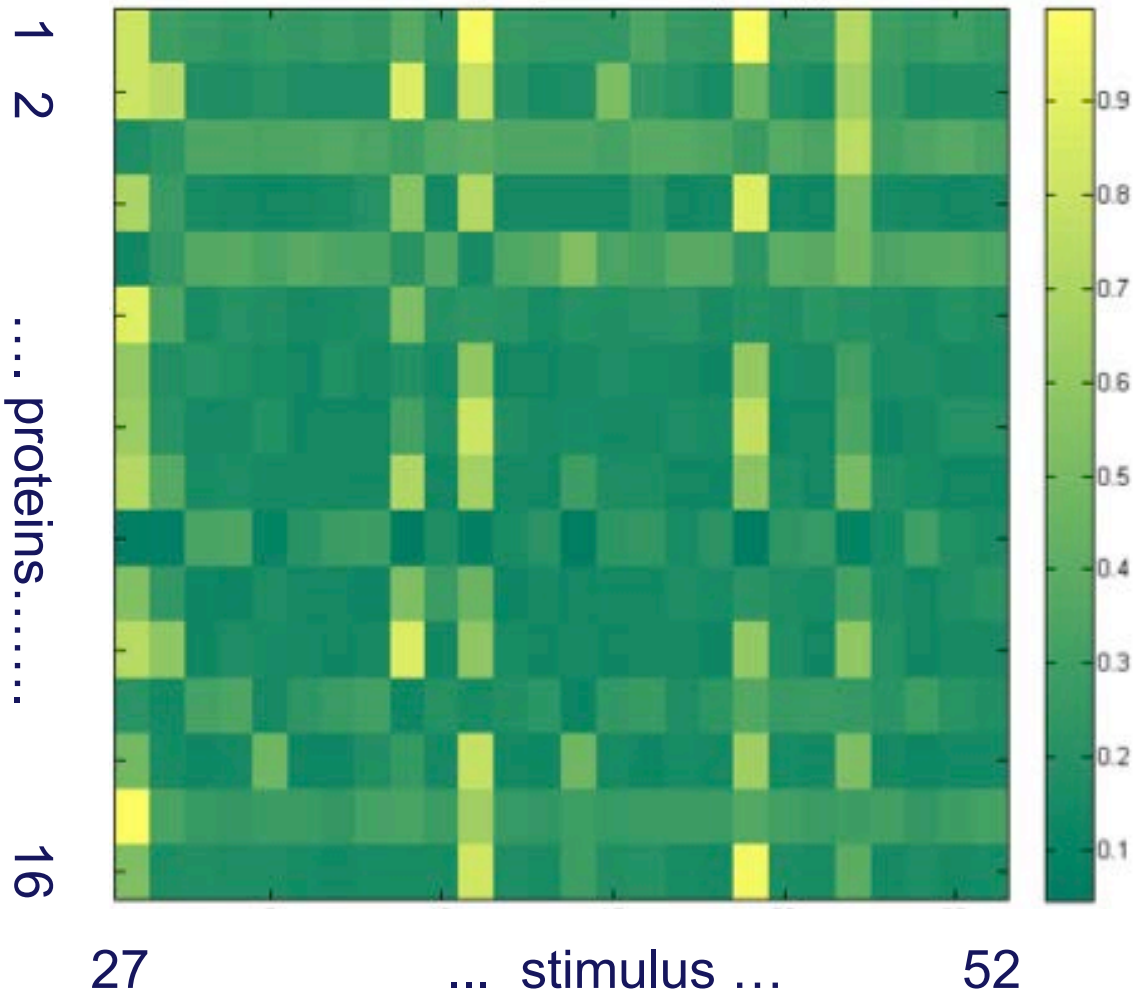
weighting scheme:

$$c_{final} = \frac{\mathrm{AUC}_{naive} \cdot c_{naive} + \mathrm{AUC}_{LVQ} \cdot c_{LVQ}}{\mathrm{AUC}_{naive} + \mathrm{AUC}_{LVQ}}$$

x (naïve prediction for protein 15)

combined prediction

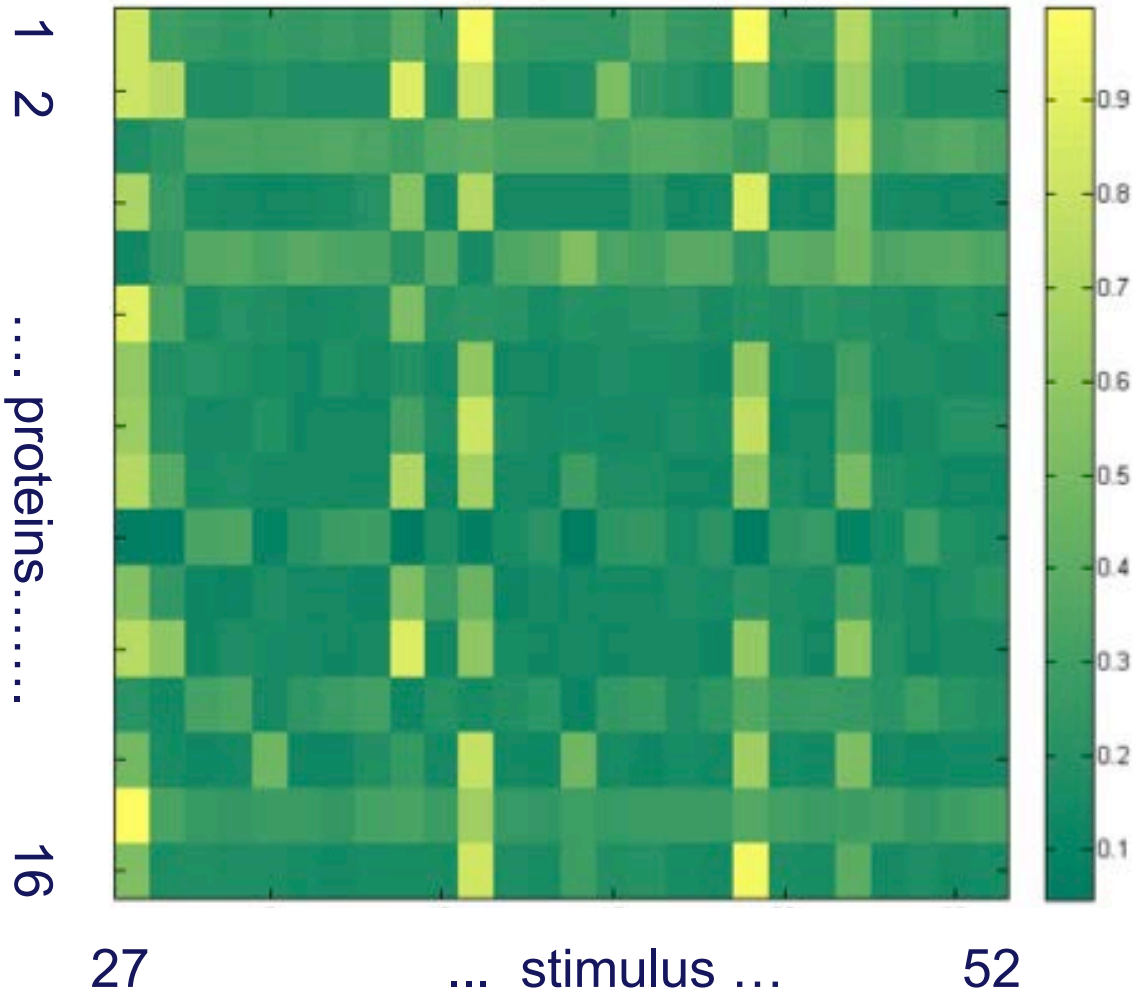
c_{final}



color-coded certainty for $|humP|>3$ in **data set B**

combined prediction

c_{final}



color-coded certainty for $|humP|>3$ in **data set B**

AUC (PR) = 0.54
Pearson = 0.75
BAC = 0.77

Summary

Analysis based on Phosphorylation data only

1) naïve prediction: **"human = rat"**
2) machine learning approach: **LVQ classifier**

Leave-One-Out procedure

combination of classifiers based on

protein-specific validation performance

Result:

(slight) improvement over naïve prediction is possible

Outlook

Extend machine learning analysis

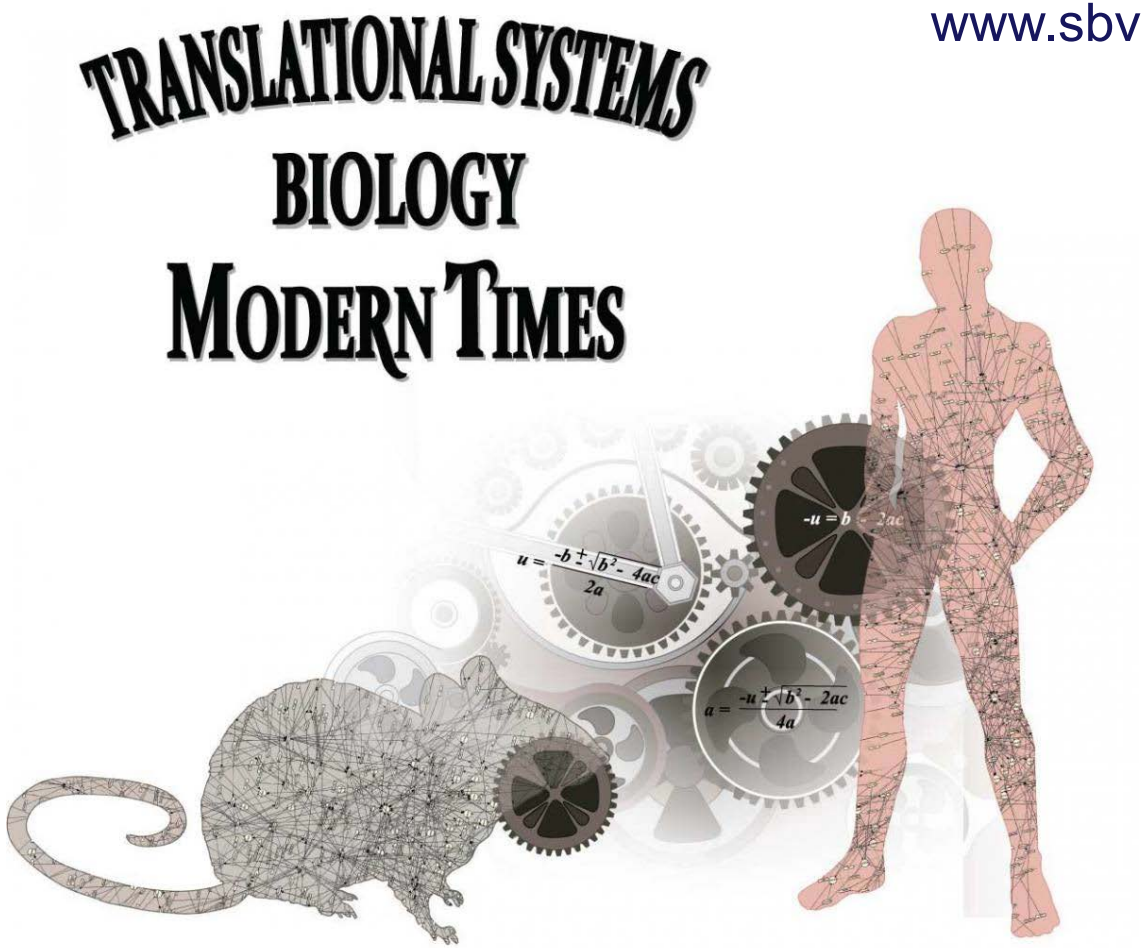
feature weighting schemes, e.g. Matrix Relevance LVQ (*)
should give further insight and help to understand,
for instance, protein-specific difficulties
(*) www.cs.rug.nl/biehl

Include gene expression data

try to understand the underlying mechanisms
and differences / common properties of rat/human
samples

Conclusion

www.sbvimprover.com



Rats and humans may be closer than we think!

Thanks

Thanks for
a great challenge!

Thanks for
real teamwork!

Thank you for your
attention!