Netherlands Bioinformatics Centre

Sequence-based prediction of protein secretion success in *Aspergillus niger*

Bastiaan van den Berg

23-09-2010

TU Delft Delft University of Technology

DSM

Introduction

- *Aspergillus niger* is widely used for industrial enzyme production
  - It has a high secretion capacity
  - Its fermentation is “generally recognized as safe” (GRAS)

Introduction

- *Aspergillus niger* is widely used for industrial enzyme production
  - It has a high secretion capacity
  - Its fermentation is “generally recognized as safe” (GRAS)

fermentor

pectinases

wine clarification

Introduction

- *Aspergillus niger* is widely used for industrial enzyme production
  - It has a high secretion capacity
  - Its fermentation is “generally recognized as safe” (GRAS)

fermentor

pectinases

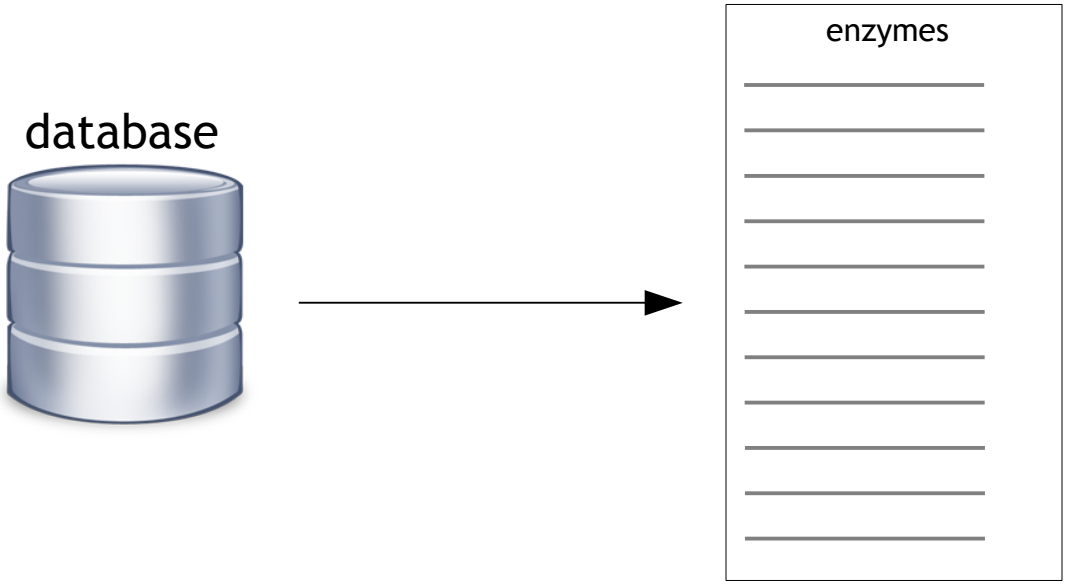
glucoamylase

wine clarification

high fructose corn syrup

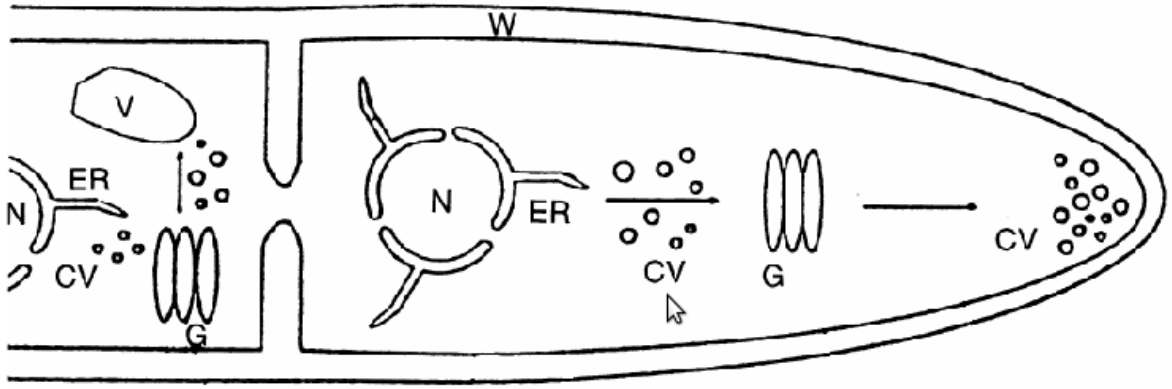
Introduction

- *Enzyme production:* gene over-expression using a strong constitutive promoter
- *Enzyme selection:* proteins from genome databases with predicted signal peptide



Introduction

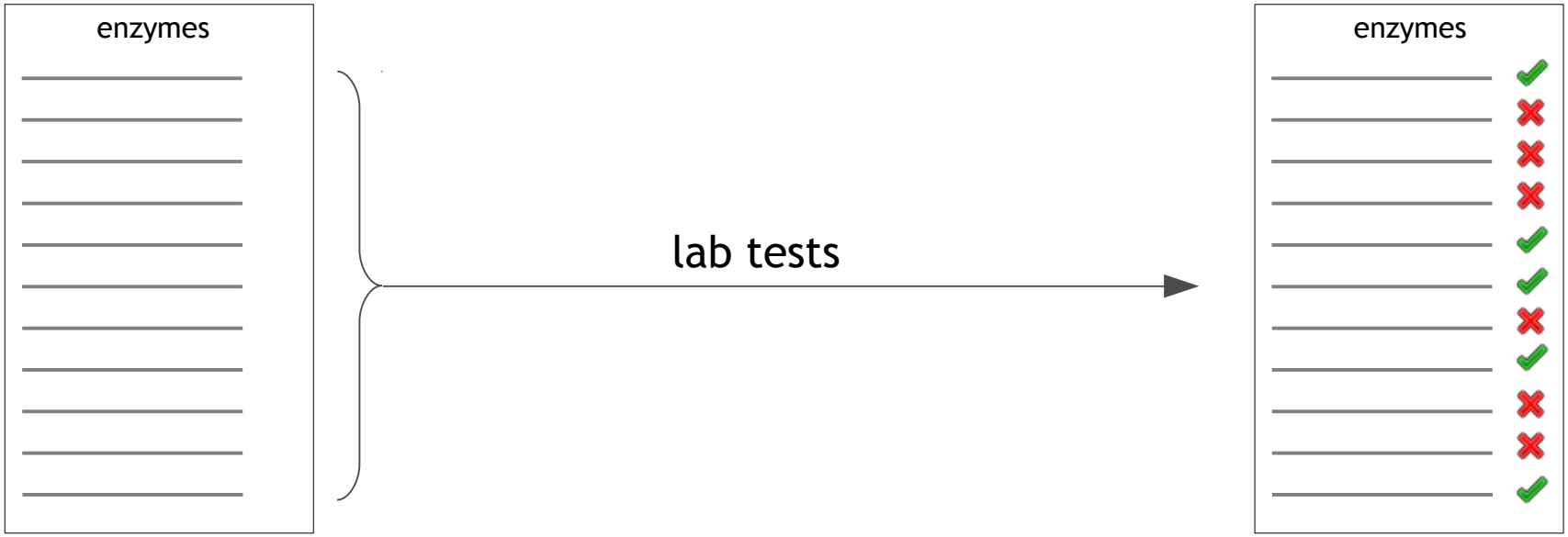
- A protein with an n-terminal signal peptide does not guarantee successful production and secretion when over-expressed
  - Protein can be translocated to different location
  - Over-expression can cause protein misfolding



Gouka et al, 1997

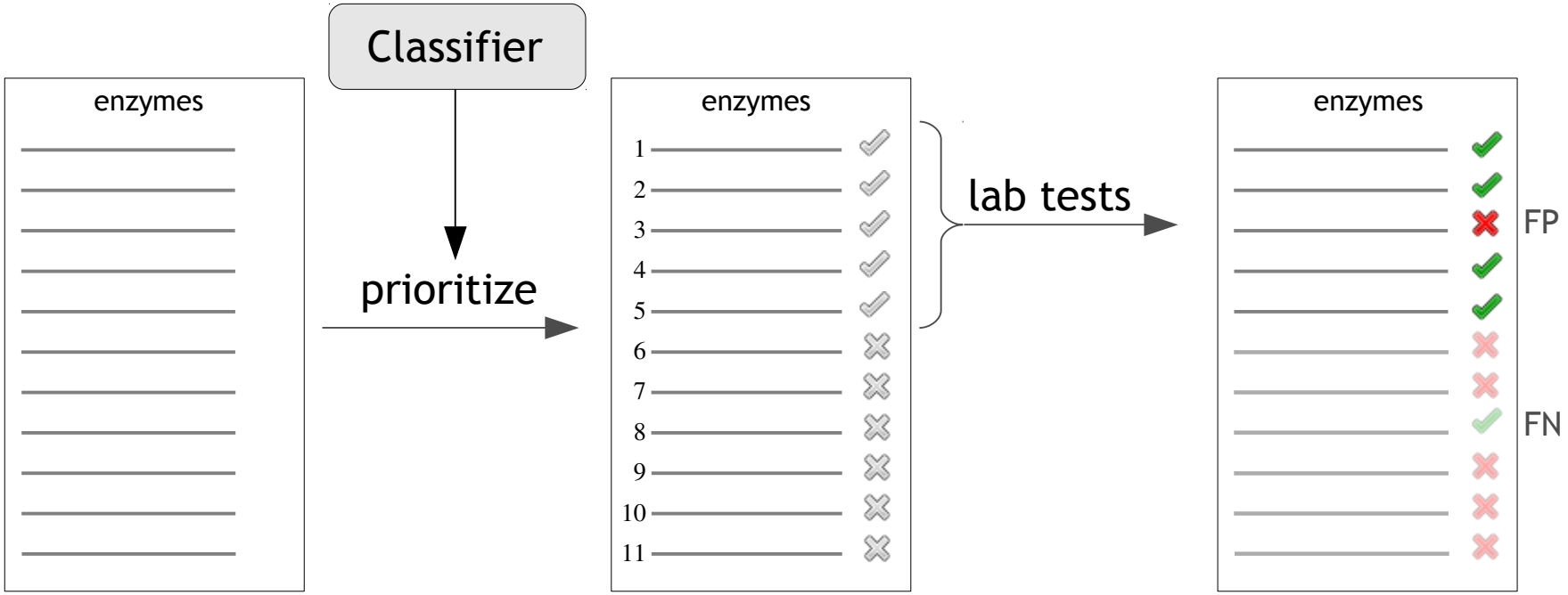
Introduction

- **Costly** lab work needed to test for successful production of over-expressed enzymes
  - secretion
  - high production rate



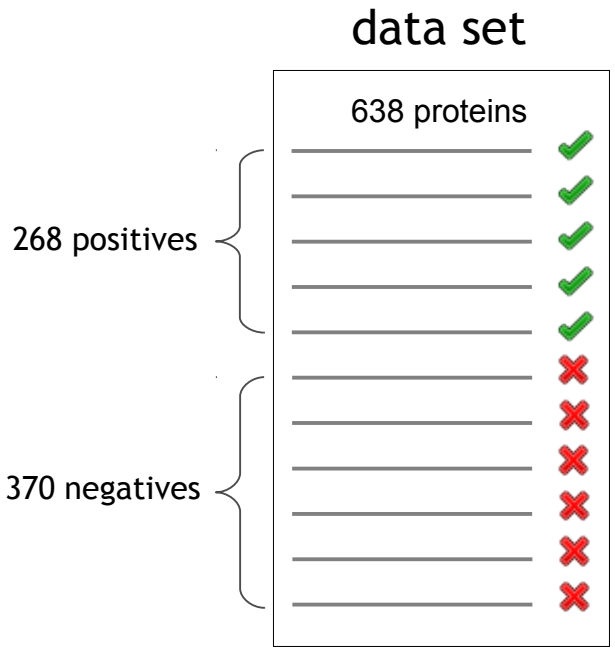
Introduction

- A classifier can greatly reduce the amount of lab work, at the cost of missing some successful enzymes



Data set

- Data set with success scores of 638 over-expressed *A. niger* proteins is used to train and test a classifier



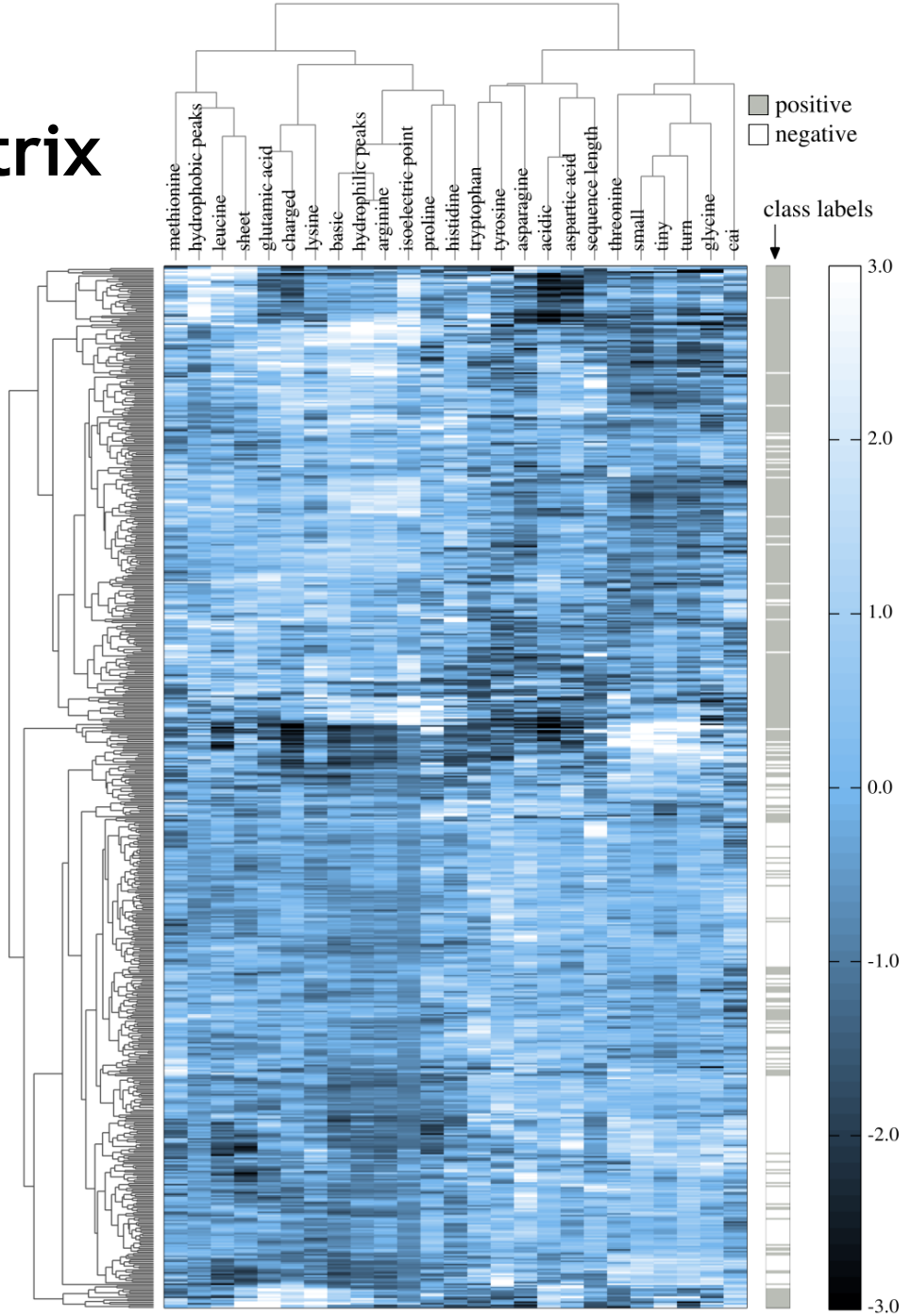
Features

- 39 sequence based features calculated
- 25 features with p-value < 0.001 (two-sample *t*-test) selected for classification

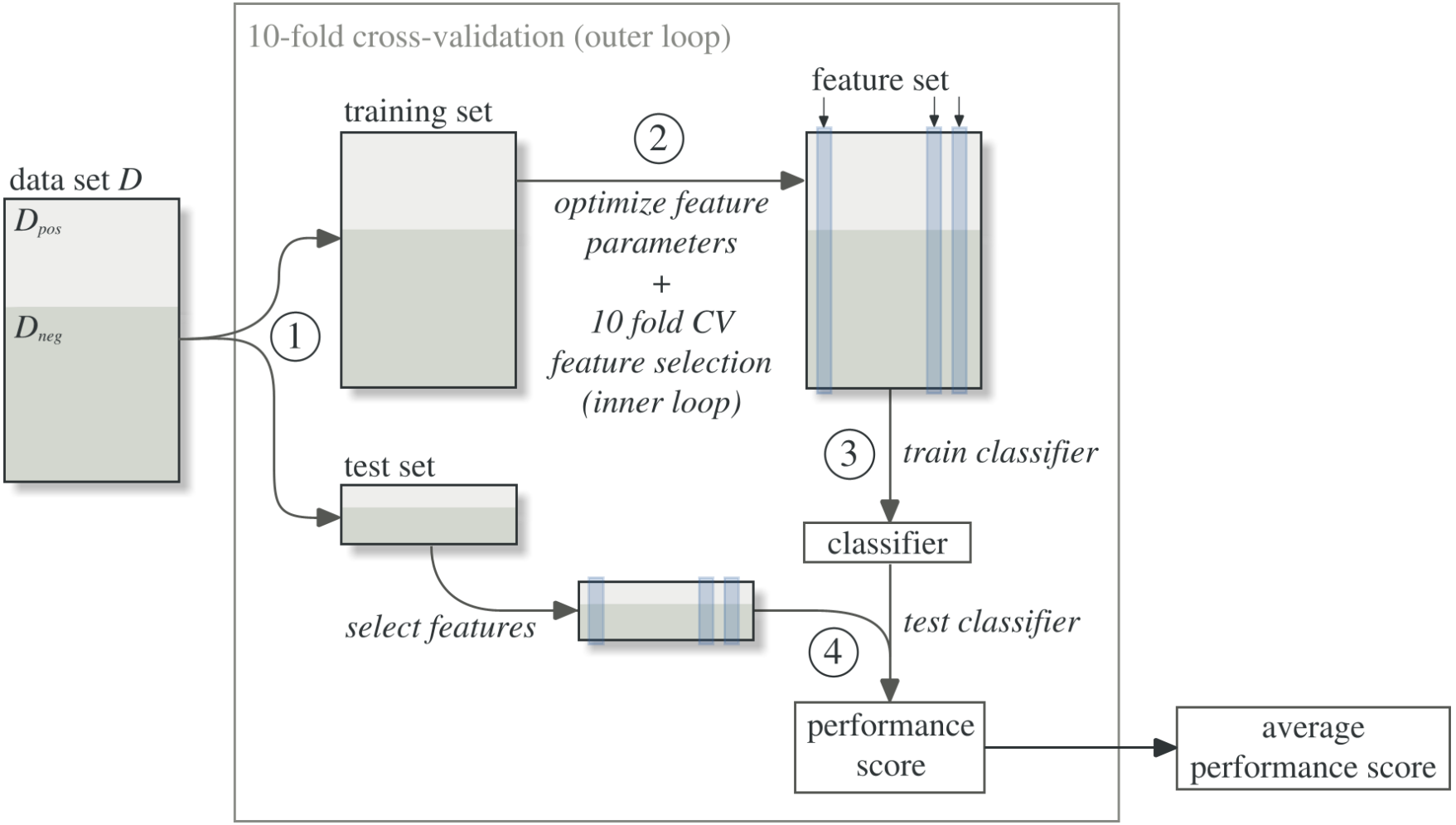
Nucleotide composition	4
GC-content	1
CAI index	1
Amino acid composition	20
Amino acid set composition	8
Hydrophobic/phylic peaks	2
GRAVY-score	1
Isoelectric point (pI)	1
Sequence length	1

Clustered feature matrix

- Hierarchical clustering of the proteins and features already shows class separability

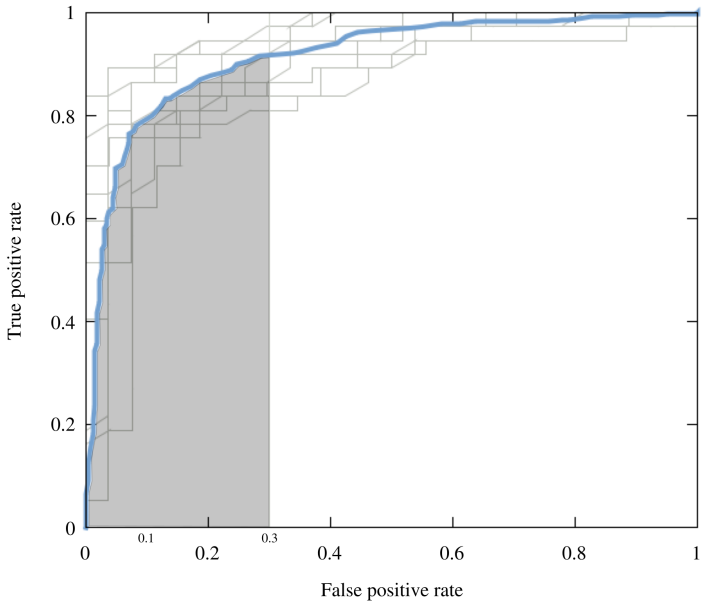


Training and validation protocol

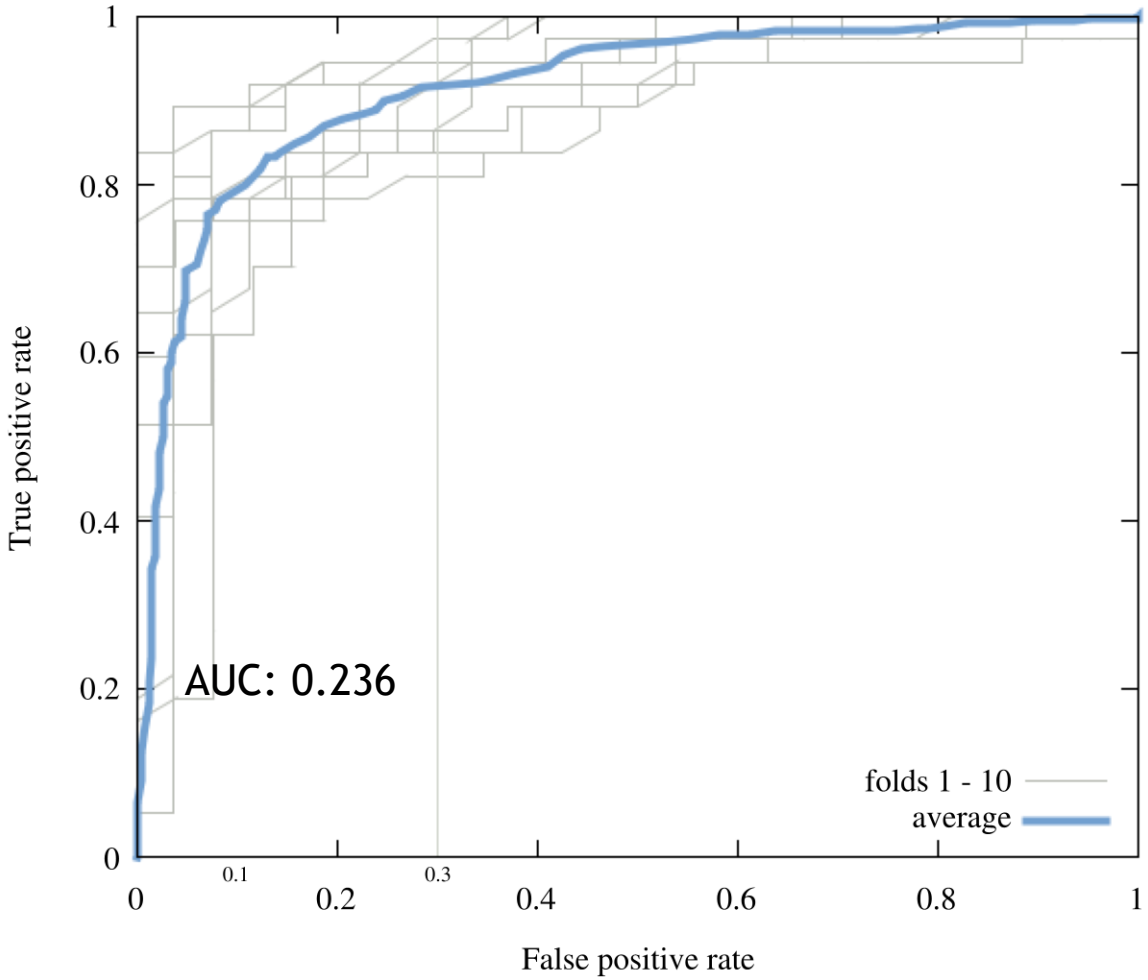


Performance evaluation

- The area under the curve (AUC) of the range 0% to 30% false positive rate of the ROC curve is used as performance measure, because we are mostly interested in low false positive rates.



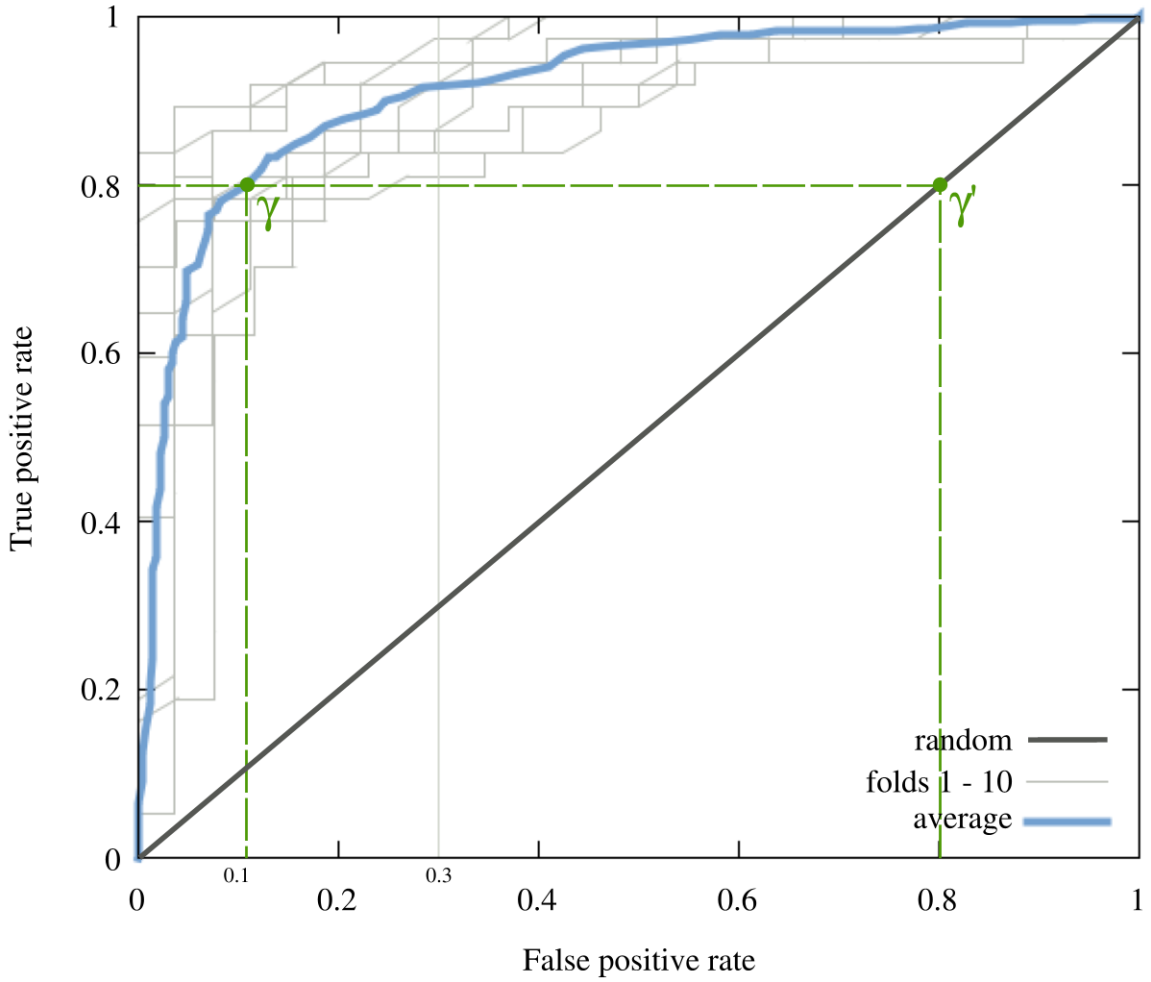
Results: linear discriminant classifier



Results: example

Example situation:
A list of 100 enzymes:
42 positives and
58 negatives

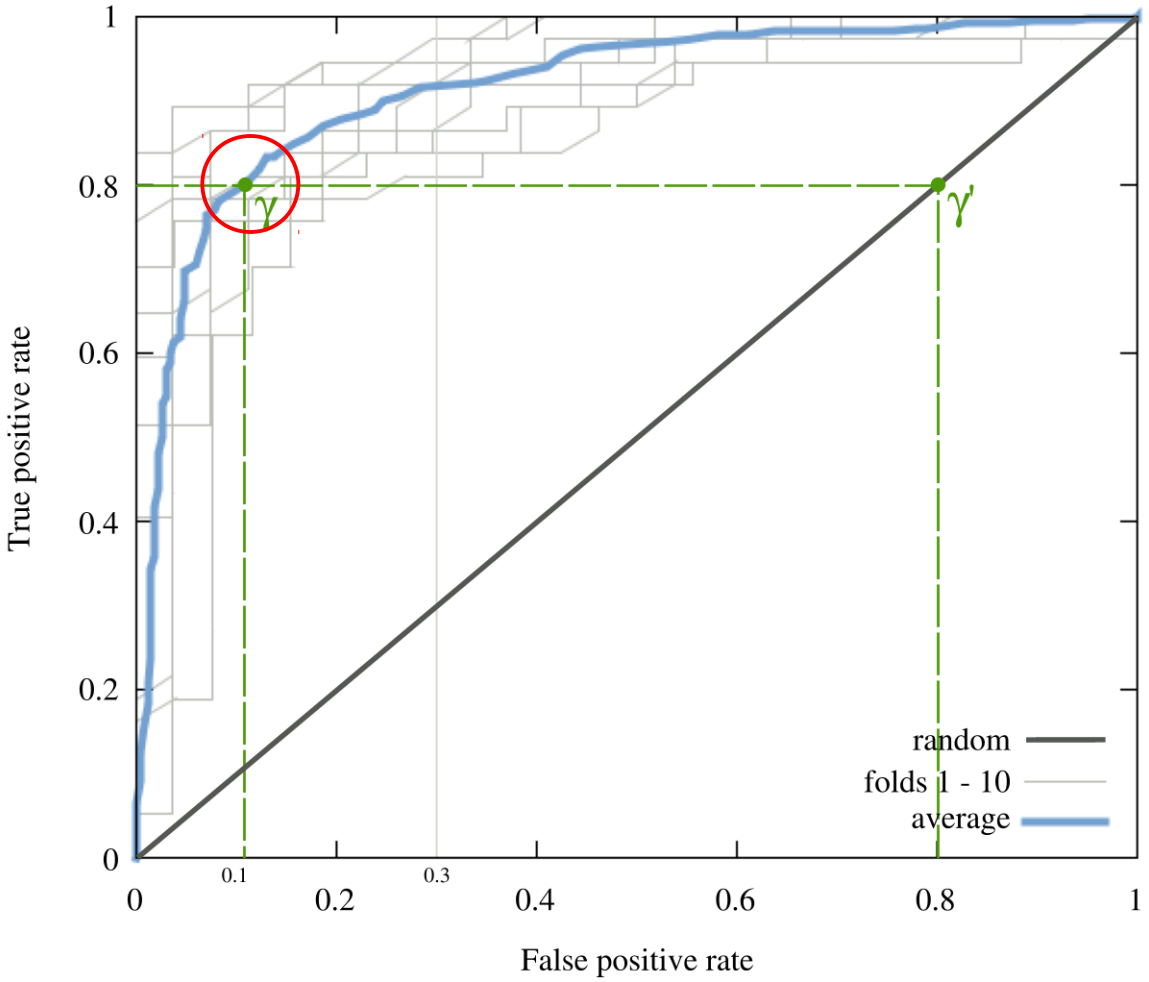
		γ
Classifier	TP-rate	*0.80*
	TP	*34*
		γ
	FP-rate	*0.10*
	FP	*6*
	lab tests	*40*
		γ'
Random selection	FP-rate	*0.80*
	FP	*46*
	lab tests	*80*
$\text{lab tests}_{\text{random}} / \text{lab tests}_{\text{classifier}}$		*2.00*



Results: example

Example situation:
A list of 100 enzymes:
42 positives and
58 negatives

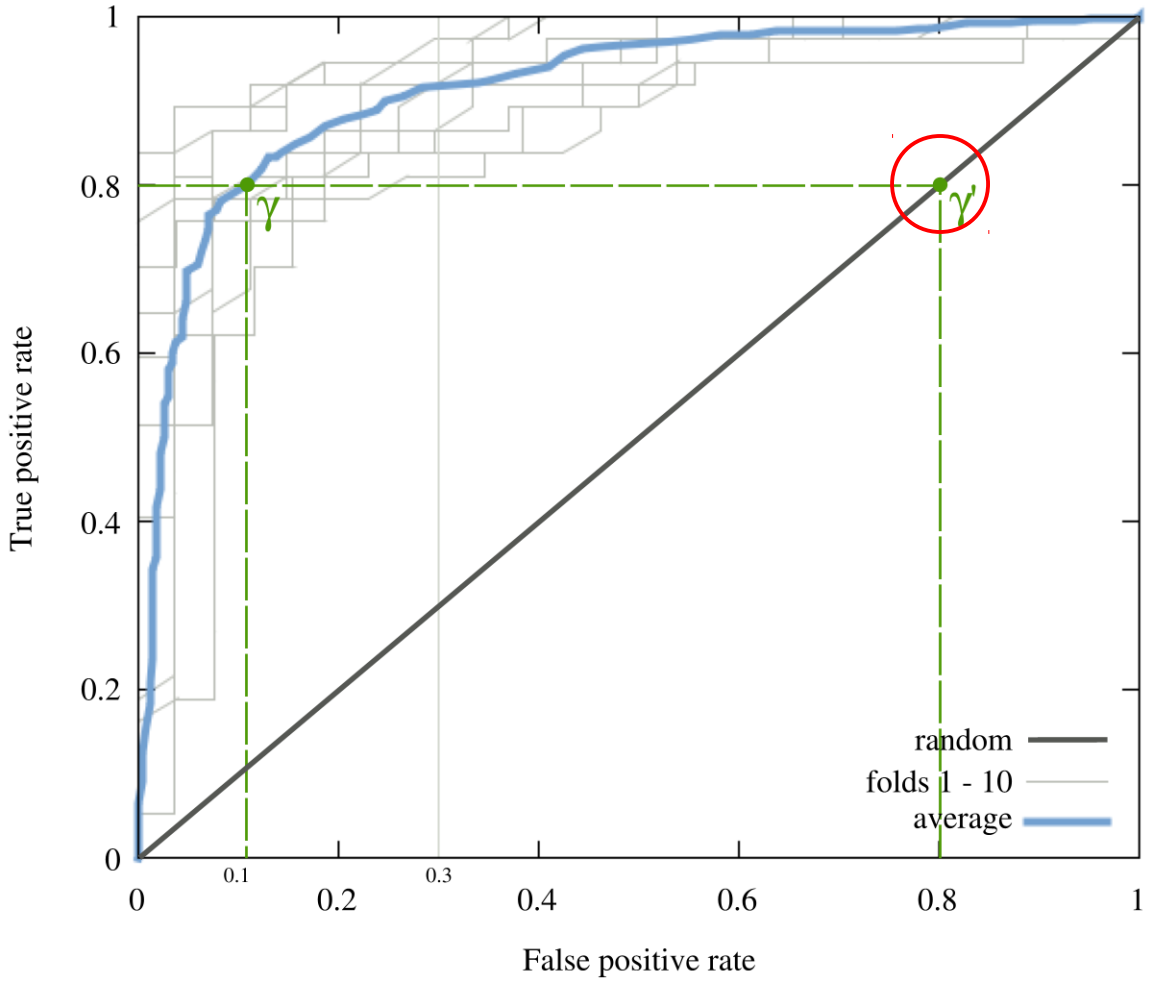
		γ
Classifier	TP-rate	*0.80*
	TP	*34*
		γ
	FP-rate	*0.10*
	FP	*6*
	lab tests	*40*
		γ'
Random selection	FP-rate	*0.80*
	FP	*46*
	lab tests	*80*
$\text{lab tests}_{\text{random}}$ / $\text{lab tests}_{\text{classifier}}$		*2.00*



Results: example

Example situation:
A list of 100 enzymes:
42 positives and
58 negatives

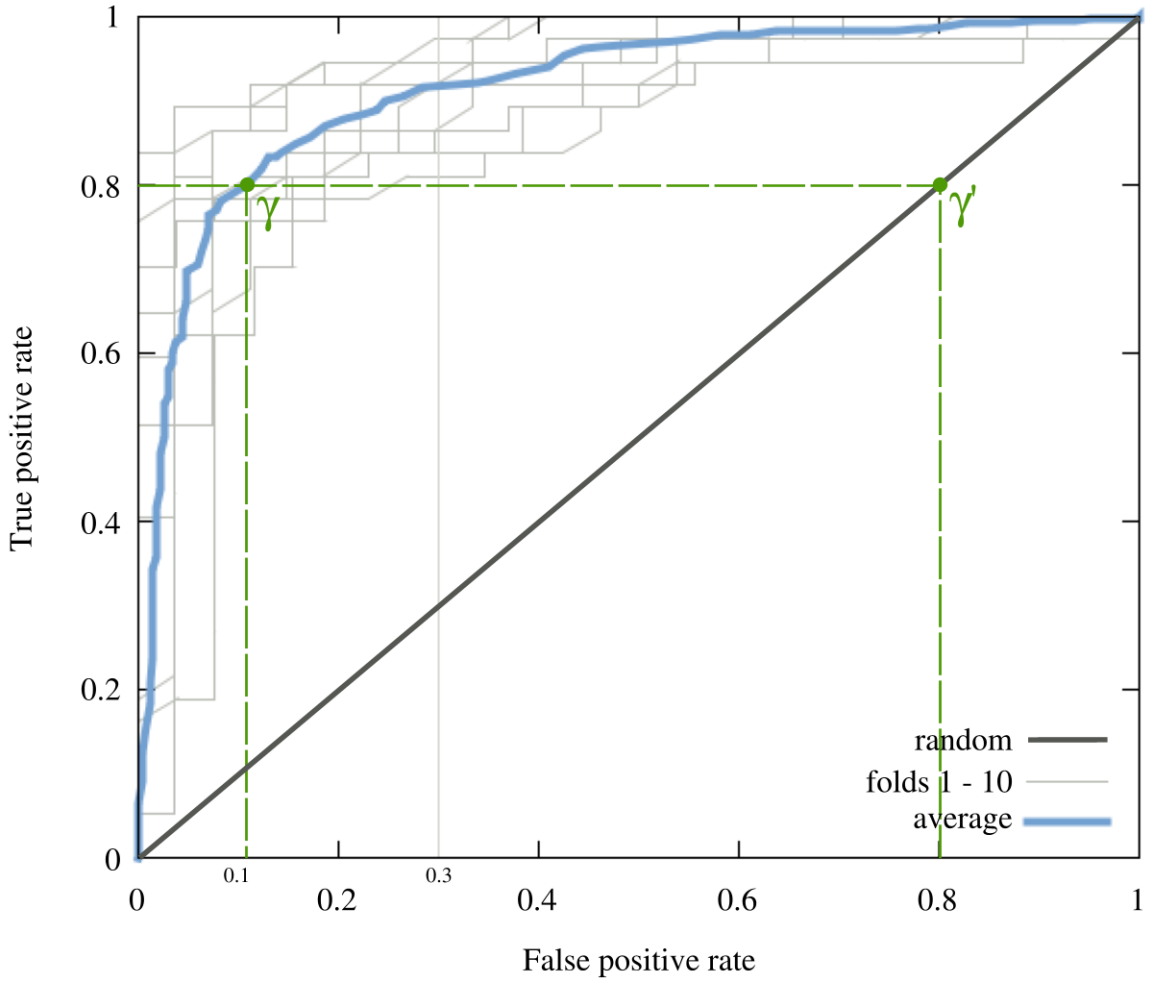
		γ
	TP-rate	*0.80*
	TP	*34*
		γ
Classifier	FP-rate	*0.10*
	FP	*6*
	lab tests	*40*
		γ'
Random selection	FP-rate	*0.80*
	FP	*46*
	lab tests	*80*
$\text{lab tests}_{\text{random}}$ / $\text{lab tests}_{\text{classifier}}$		*2.00*



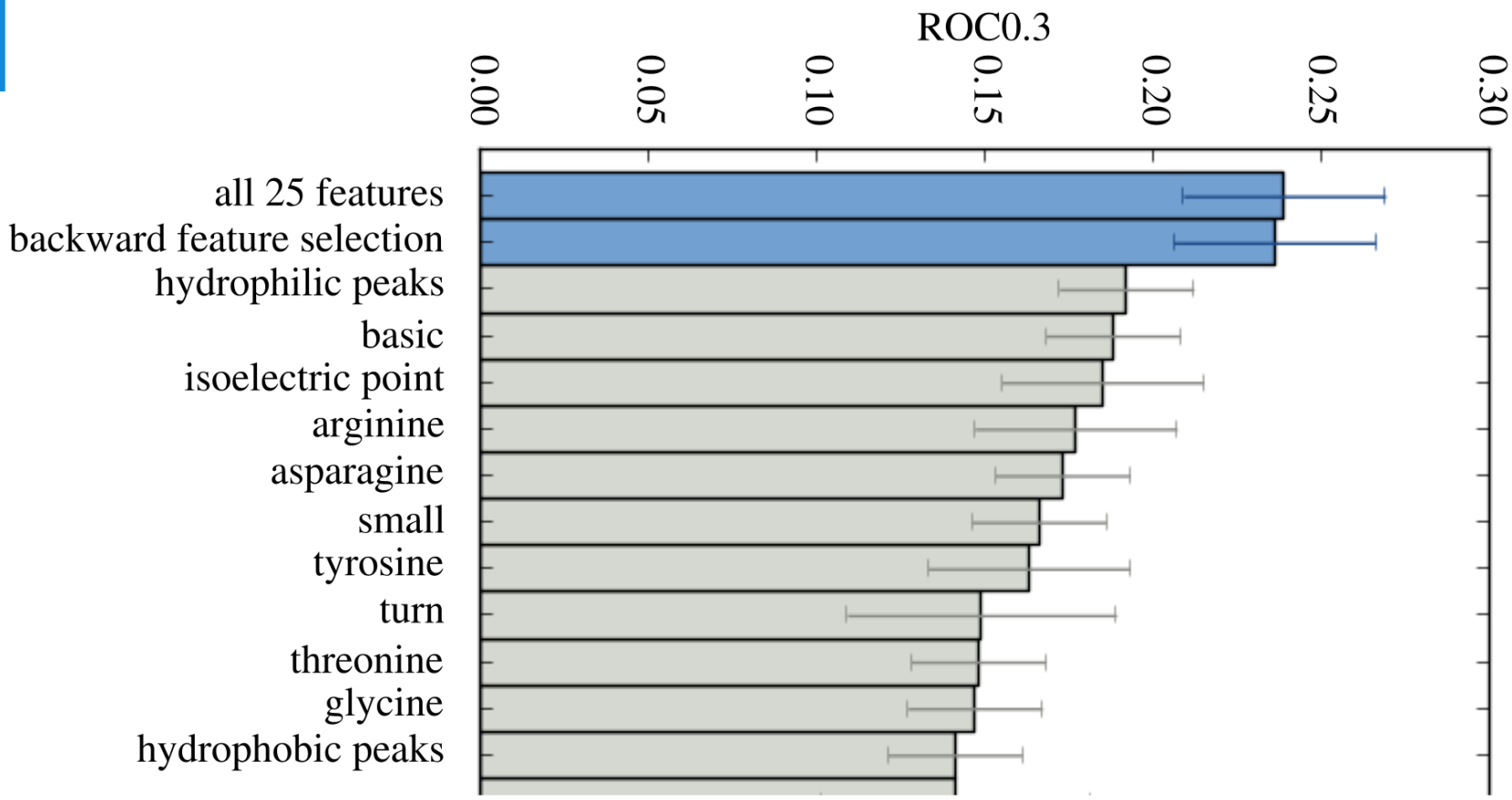
Results: example

Example situation:
A list of 100 enzymes:
42 positives and
58 negatives

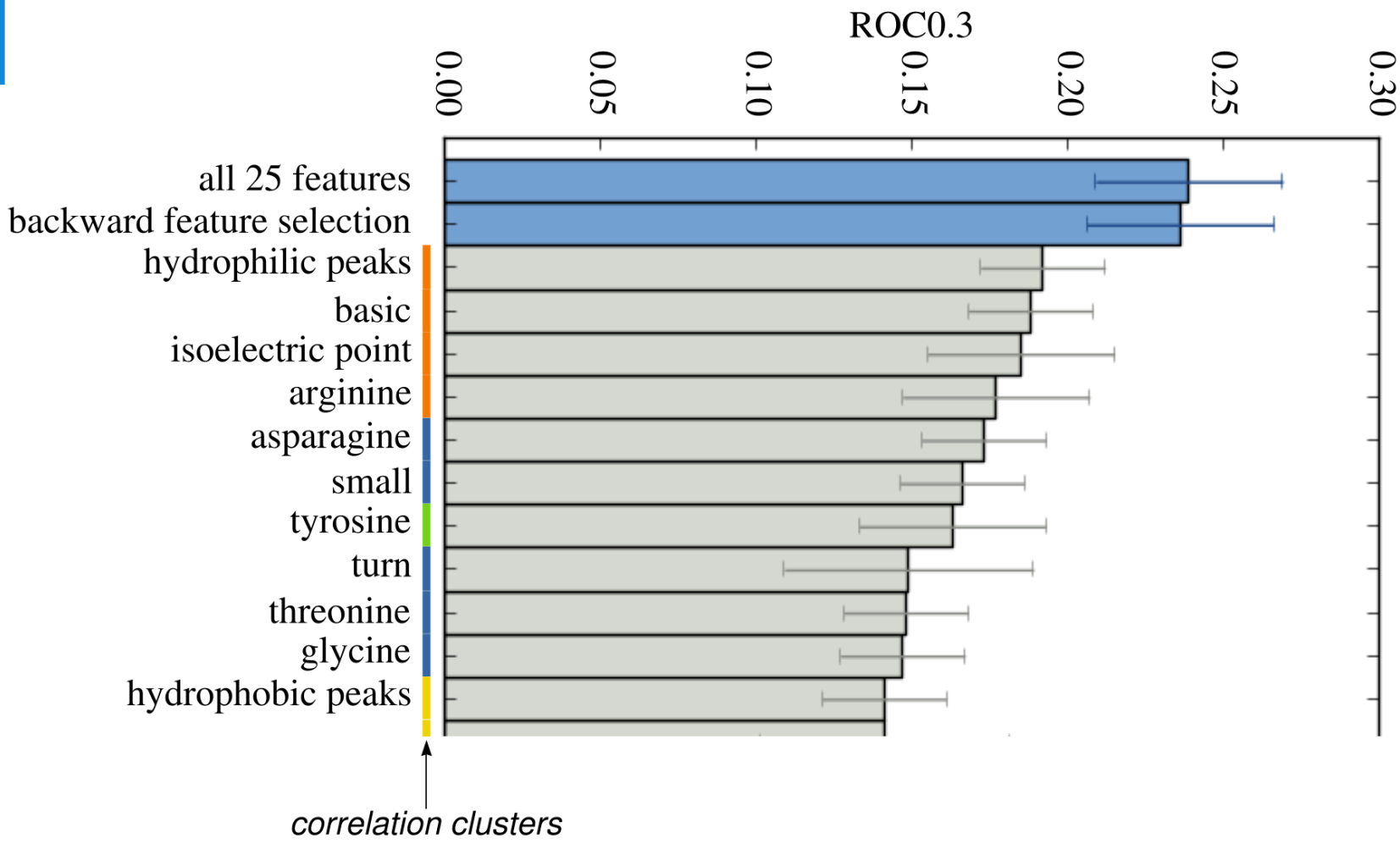
		γ
Classifier	TP-rate	*0.80*
	TP	*34*
		γ
	FP-rate	*0.10*
	FP	*6*
	lab tests	*40*
		γ'
Random selection	FP-rate	*0.80*
	FP	*46*
	lab tests	*80*
lab tests$_{random}$ / lab tests$_{classifier}$		*2.00*



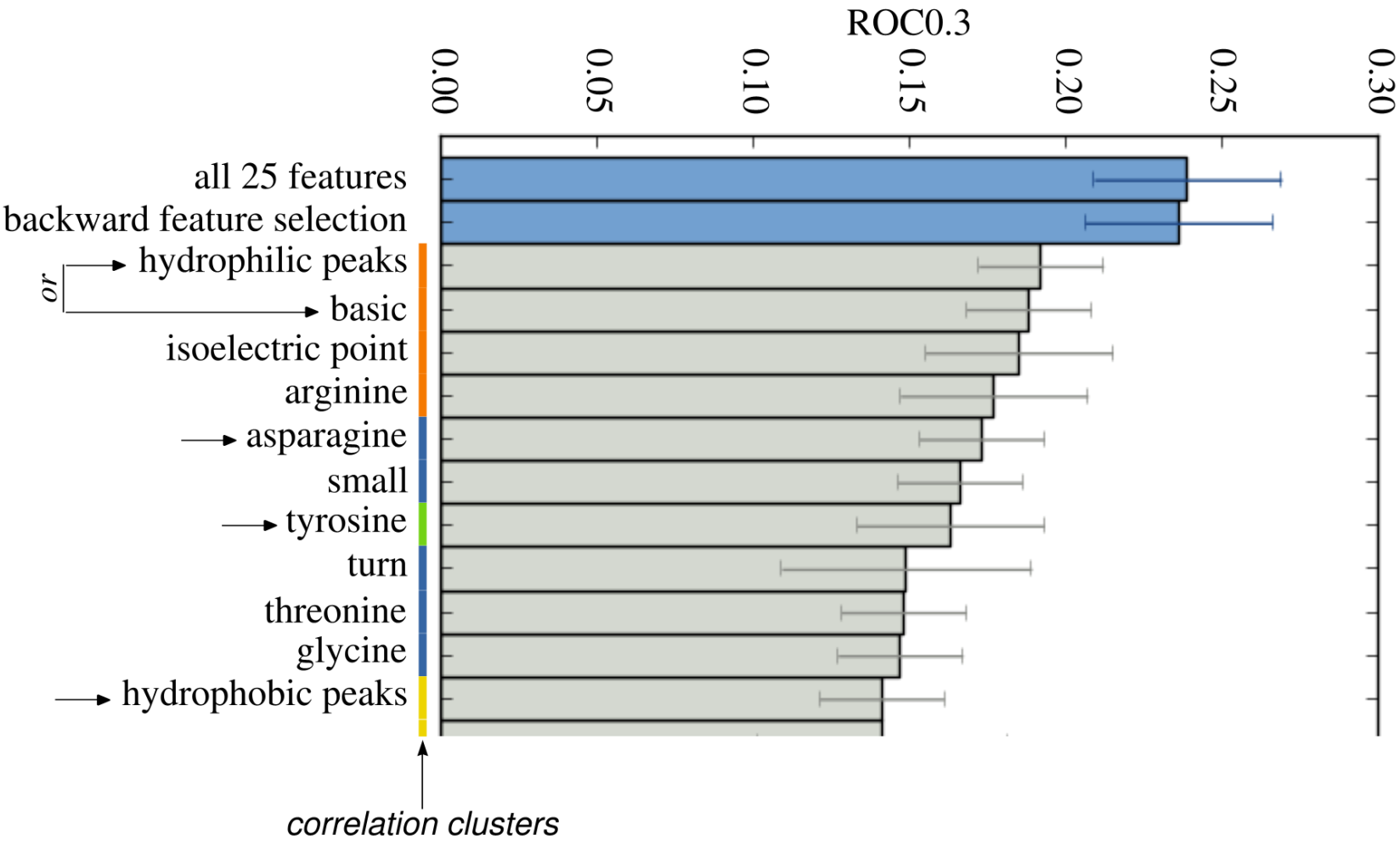
Results


ROC0.3
0.00
0.05
0.10
0.15
0.20
0.25
0.30
all 25 features
backward feature selection
hydrophilic peaks
basic
isoelectric point
arginine
asparagine
small
tyrosine
turn
threonine
glycine
hydrophobic peaks

Results


ROC0.3
0.00
0.05
0.10
0.15
0.20
0.25
0.30
all 25 features
backward feature selection
hydrophilic peaks
basic
isoelectric point
arginine
asparagine
small
tyrosine
turn
threonine
glycine
hydrophobic peaks
correlation clusters

Results


ROC0.3
0.00
0.05
0.10
0.15
0.20
0.25
0.30
all 25 features
backward feature selection
hydrophilic peaks
or
basic
isoelectric point
arginine
asparagine
small
tyrosine
turn
threonine
glycine
hydrophobic peaks
correlation clusters

Further research

- Separation n-terminal signal peptide and rest of the sequence
- Taking into account predicted secondary structure and solvent accessibility
- Using string kernels to include information on the location of amino acid
- Same study on a data set with heterologous proteins

Acknowledgements

nbic netherlands bioinformatics centre

DSM

Hans Roubos, Liang Wu, Herman Pel

TU Delft Delft University of Technology

Jurgen Nijkamp, Dick de Ridder, Marcel Reinders