RNA Bioinformatics and High-Throughput Analysis Jena

# Prediction of B-Cell Epitopes

14 February 2019 / Maximilian Collatz

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA
INFECT CONTROL 2020
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
zwanzig20
PARTNERSCHAFT FÜR INNOVATION
UNTERNEHMEN REGION
Die BMBF-Innovationsinitiative Neue Länder
PHILIPPO·HESS·DO·DE·PIETAT·ET·LITER·OPT·MERITO·AUSP·SIGILLUM·SCHOLAE·MARPURGENSIS
15 27

# The problem

- Unknown vaccination status of patients
- Vaccination very expensive

⇒ Testing vaccination status is less expensive but still far from cheap

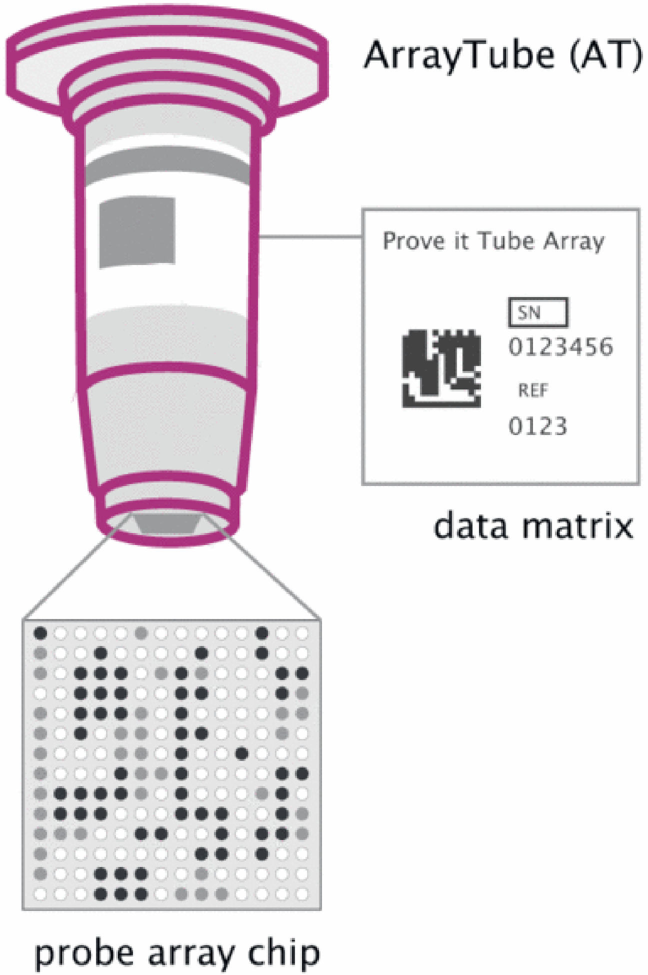
⇒ Time consuming

# The vision

- One chip for all
- All genes of the STIKO list
- Less expensive and much faster

STIKO – Ständige Impfkommission

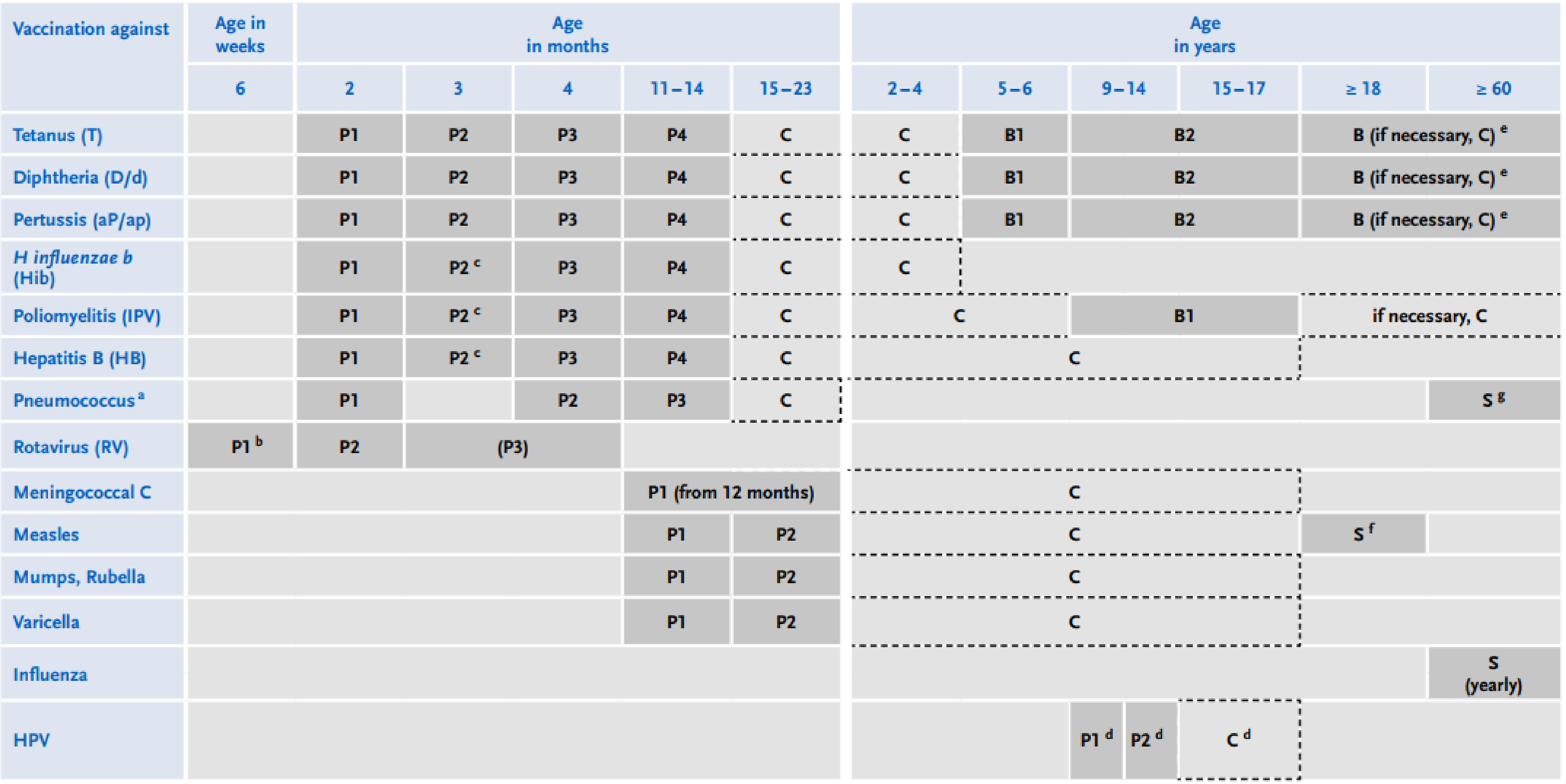
(Standing Committee on Vaccination)



# STIKO list December 2018

- Influenza
- Tetanus
- Measles
- Mumps
- HPV
- …

Table 1: Immunisation schedule (vaccinations) for infants, children, adolescents and adults



| Vaccination against | Age in weeks | Age in months | | | | | Age in years | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 6 | 2 | 3 | 4 | 11–14 | 15–23 | 2–4 | 5–6 | 9–14 | 15–17 | ≥ 18 | ≥ 60 |
| Tetanus (T) | | P1 | P2 | P3 | P4 | C | C | B1 | B2 | | B (if necessary, C)[e] | |
| Diphtheria (D/d) | | P1 | P2 | P3 | P4 | C | C | B1 | B2 | | B (if necessary, C)[e] | |
| Pertussis (aP/ap) | | P1 | P2 | P3 | P4 | C | C | B1 | B2 | | B (if necessary, C)[e] | |
| *H influenzae b* (Hib) | | P1 | P2[c] | P3 | P4 | C | C | | | | | |
| Poliomyelitis (IPV) | | P1 | P2[c] | P3 | P4 | C | C | | B1 | | if necessary, C | |
| Hepatitis B (HB) | | P1 | P2[c] | P3 | P4 | C | C | | | | | |
| Pneumococcus[a] | | P1 | | P2 | P3 | C | | | | | | S[g] |
| Rotavirus (RV) | P1[b] | P2 | (P3) | | | | | | | | | |
| Meningococcal C | | | | | P1 (from 12 months) | | C | | | | | |
| Measles | | | | | P1 | P2 | C | | | | S[f] | |
| Mumps, Rubella | | | | | P1 | P2 | C | | | | | |
| Varicella | | | | | P1 | P2 | C | | | | | |
| Influenza | | | | | | | | | | | | S (yearly) |
| HPV | | | | | | | | | P1[d] P2[d] | C[d] | | |

# The Fantastic Four

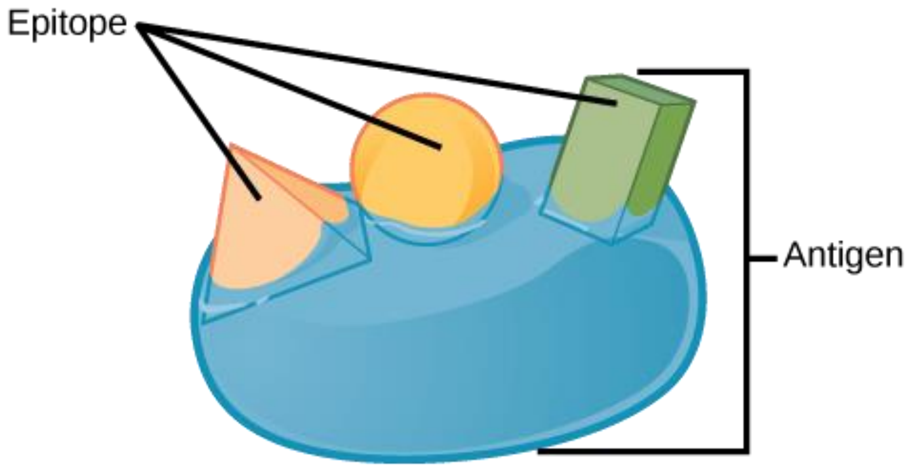
FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

Abbott

virion\serion

UNIVERSITÄTS KLINIKUM Jena

# Background



**Immunodominance**

- Proteins that evoke the immune response
- Often exposed surface proteins

**Epitope**

- Region of the immunodominant protein that is recognized by the immune system
- Continuous or discontinuous
- Length: 8-21 amino acids

# Properties of antigens

- Protein disorder (IUPRED)
- Surface accessibility (RaptorX)
- Protein secondary structure (RaptorX)

- Localization (DeepLoc)

- Polymorphic regions (own pipeline)

# Epitope prediction by protein structure

**IUPRED – disorder tendency**

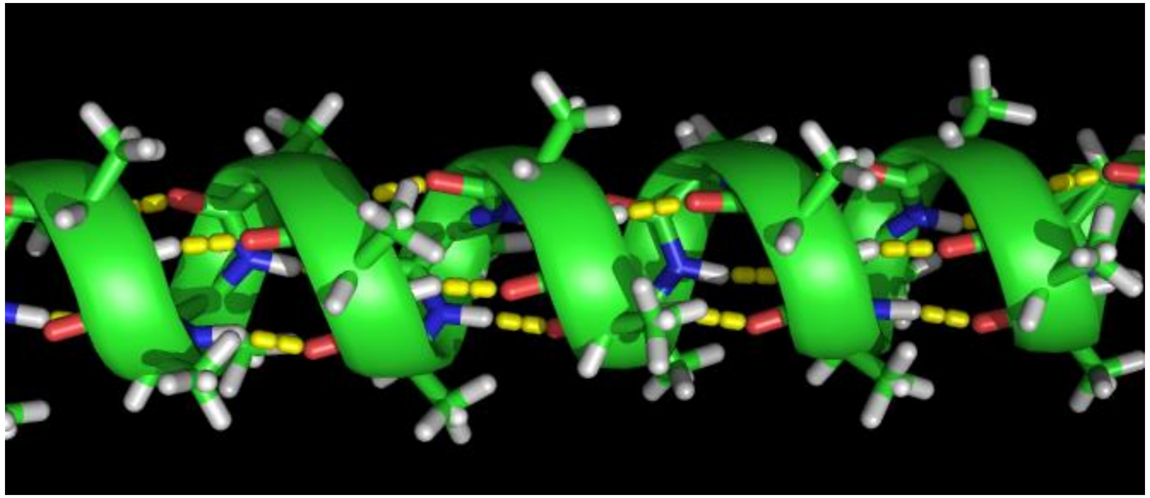
- Intrinsically Unstructured proteins without secondary structure
- Epitope regions are IU-regions

**RaptorX – protein structure**

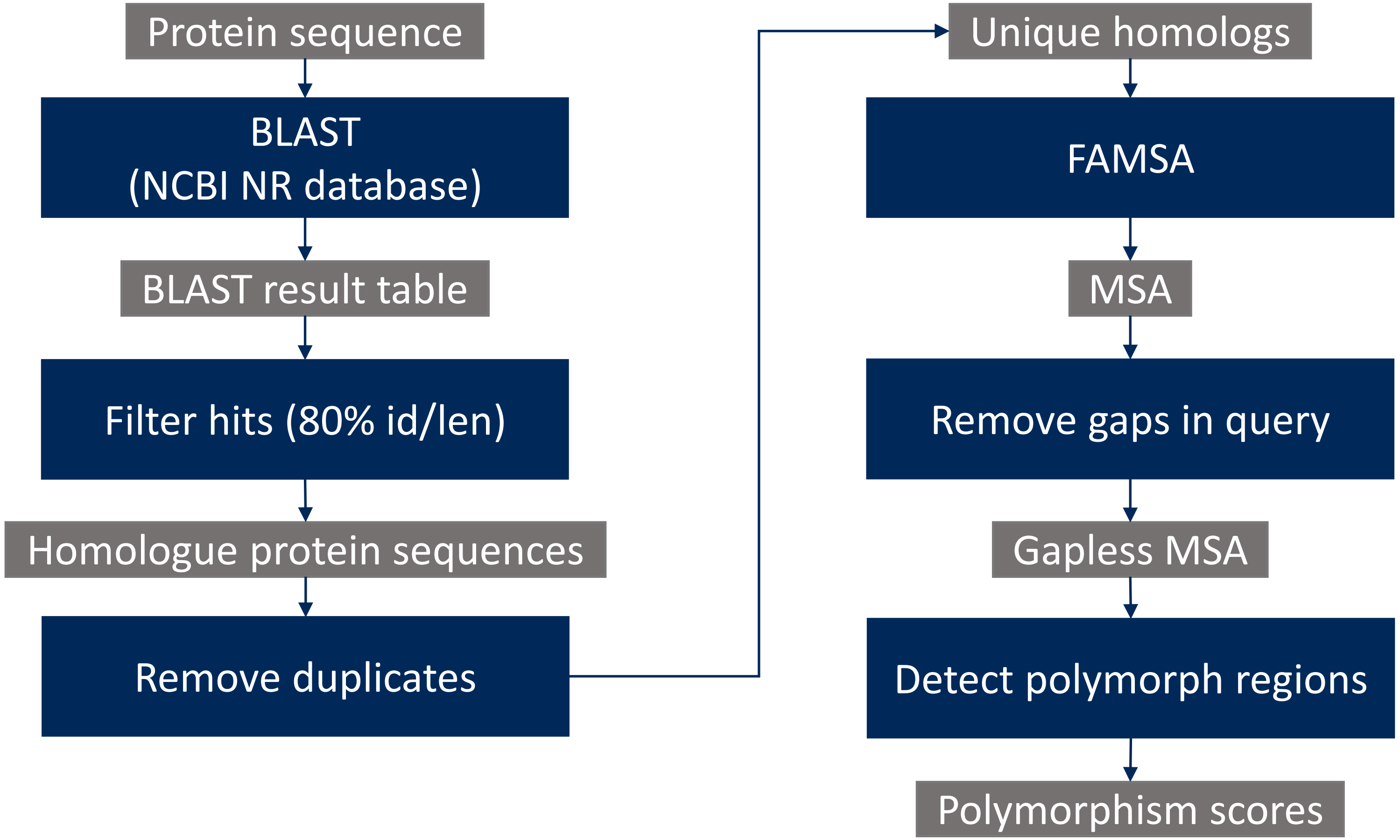
- β-sheets
- α-helixes
- Random coil
- Surface accessibility

**DeepLoc**

- Cellular localization prediction



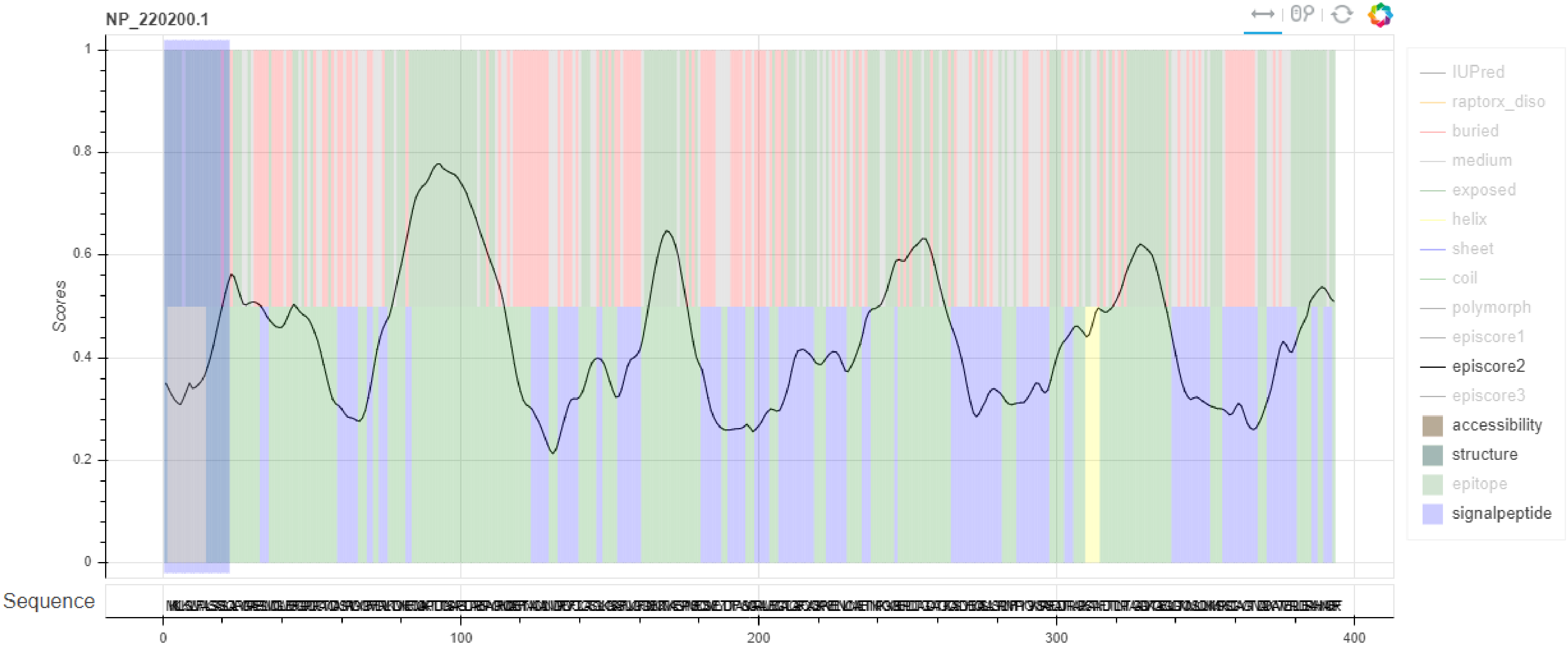
# Detection of polymorphism scores

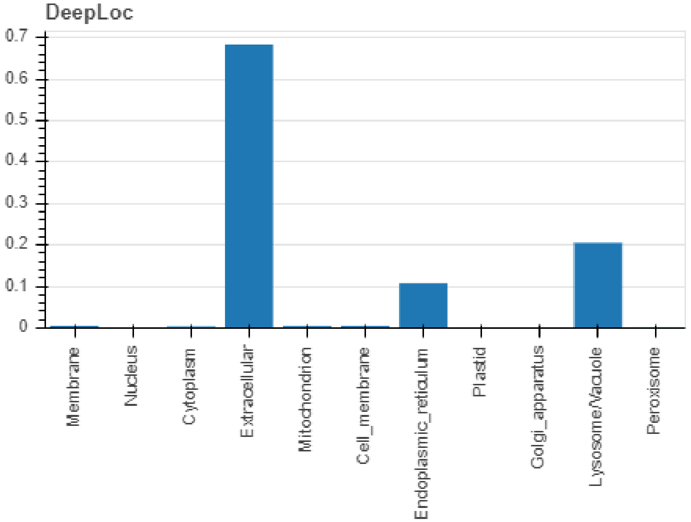


# Visualization requirements

- Large proteins
- Protein disorder
- Accessibility
  - Buried
  - Medium
  - Exposed
- Structure
  - β-sheets
  - α-helixes
  - Coil
- Polymorphism score

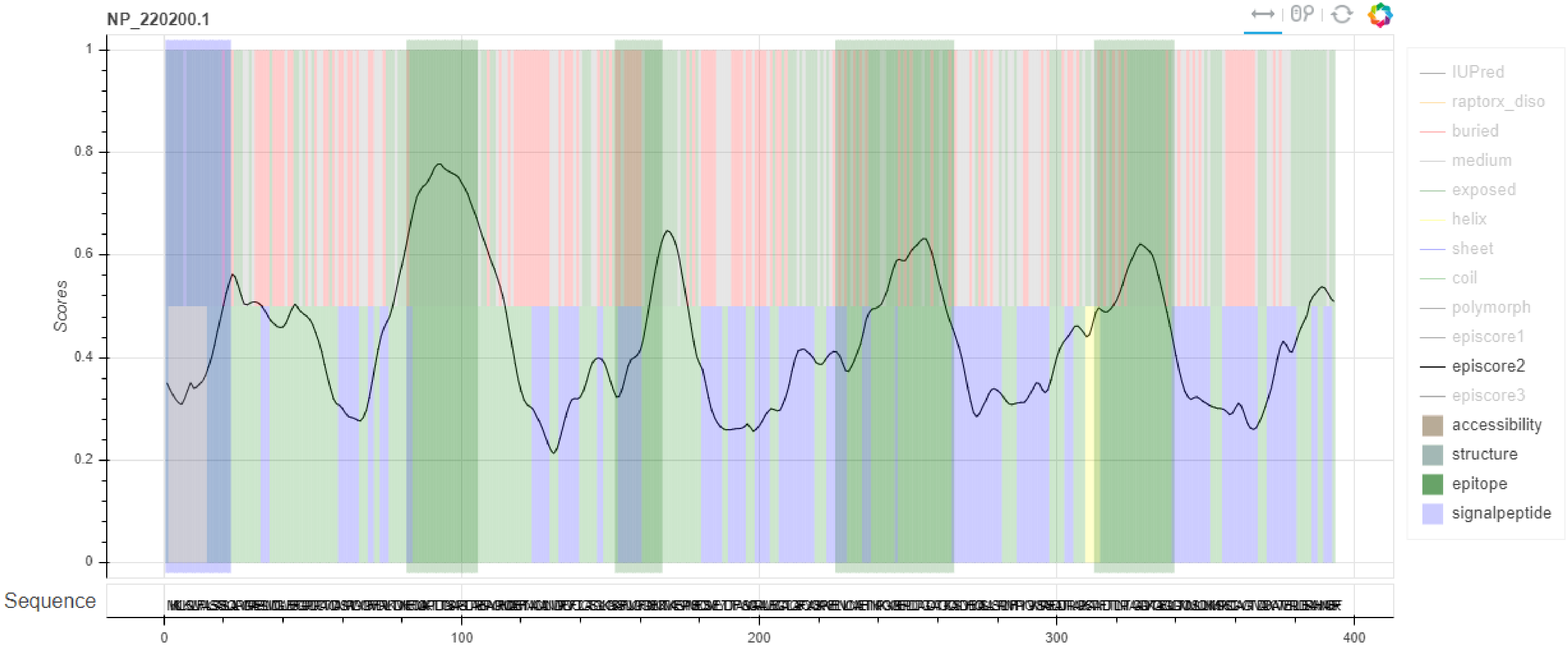
→ Interactive HTML plot

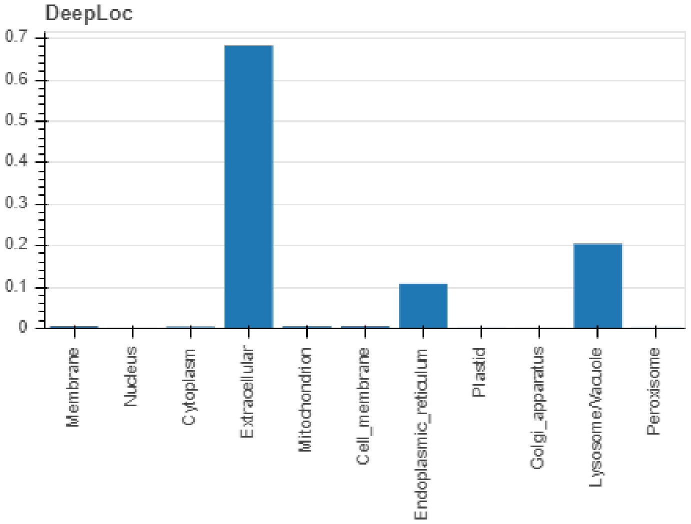




SignalP results: SP='YES' Cleavage site between pos. 22 and 23: LQA-LP D=0.898 D-cutoff=0.570 Networks=SignalP-noTM

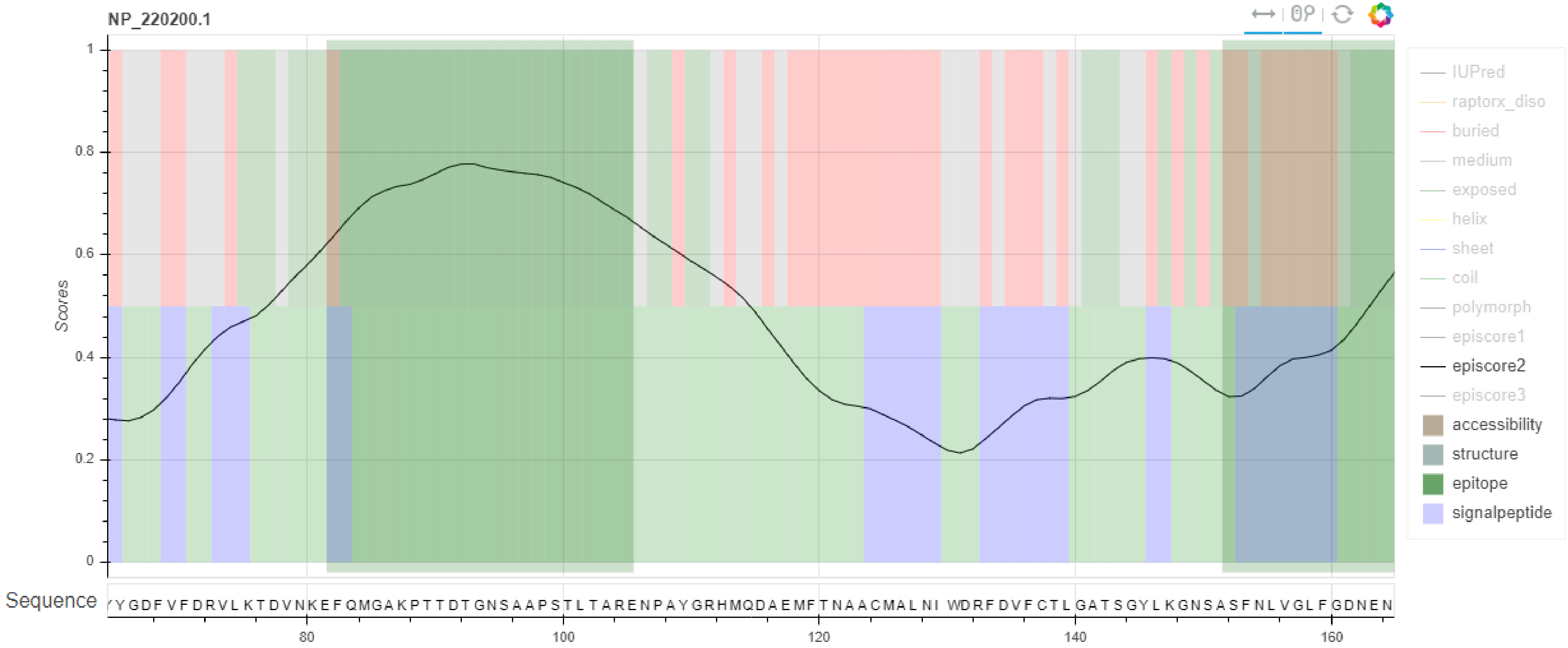
Homologs used for polymorph score: 412

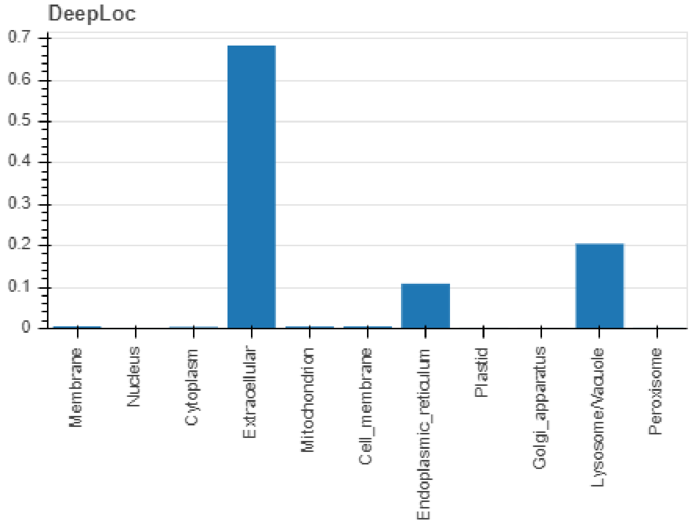




SignalP results: SP='YES' Cleavage site between pos. 22 and 23: LQA-LP D=0.898 D-cutoff=0.570 Networks=SignalP-noTM

Homologs used for polymorph score: 412





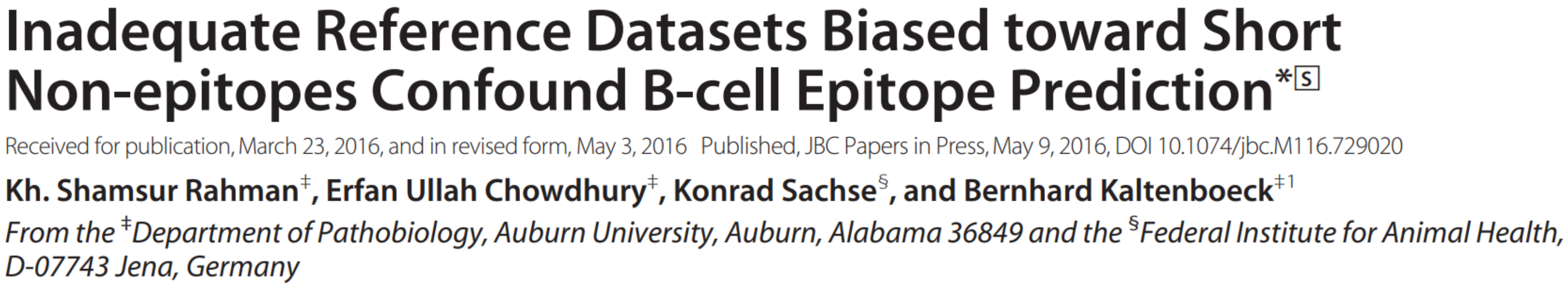
SignalP results: SP='YES' Cleavage site between pos. 22 and 23: LQA-LP D=0.898 D-cutoff=0.570 Networks=SignalP-noTM

Homologs used for polymorph score: 412

# Training the Deep Neural Network

- Input
    - Epitope sequence
    - Predicted structure information
- Output
    - Epitope, or non-epitope

# The dataset



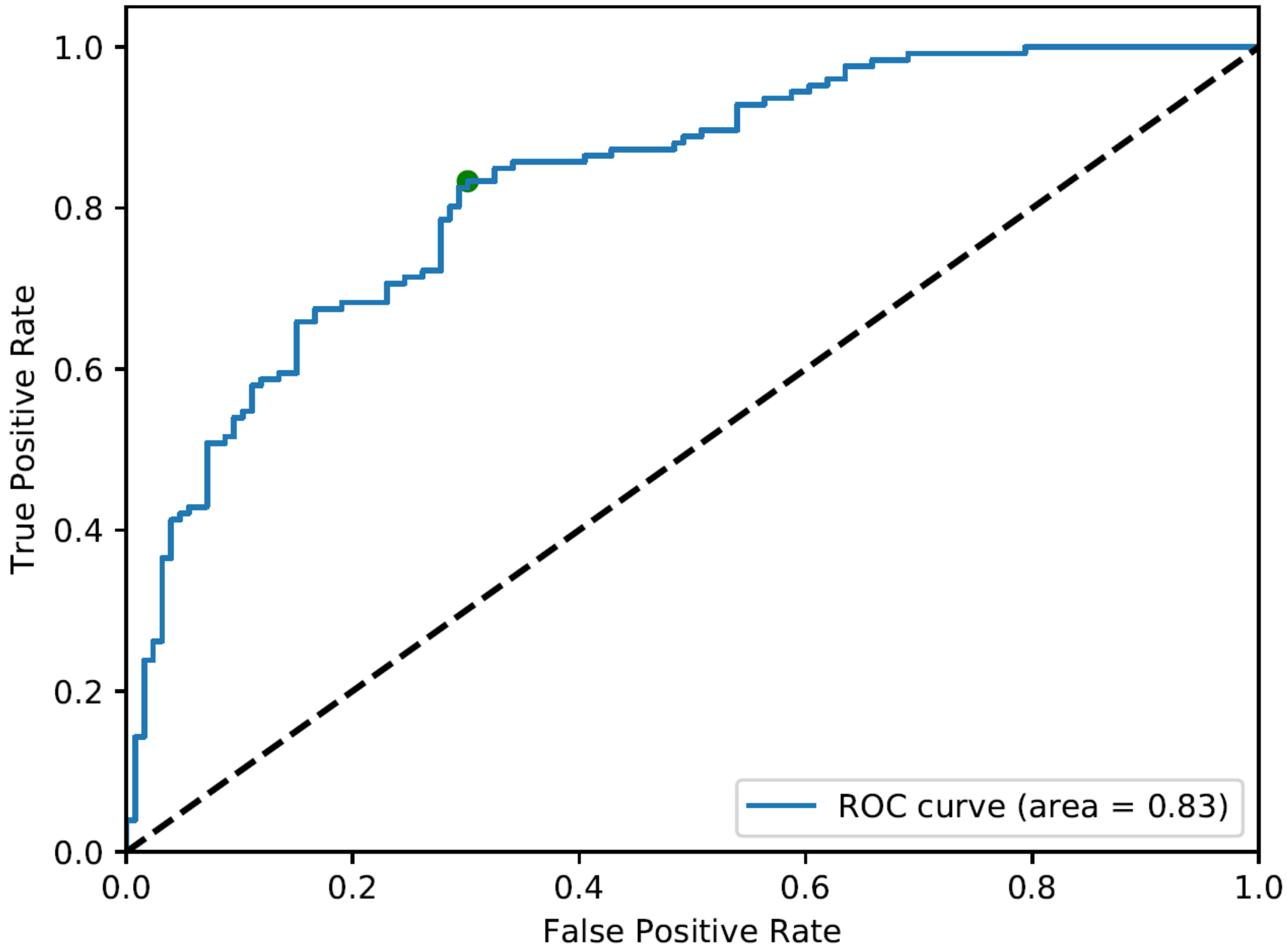
**Inadequate Reference Datasets Biased toward Short Non-epitopes Confound B-cell Epitope Prediction**[*][S]

Received for publication, March 23, 2016, and in revised form, May 3, 2016 Published, JBC Papers in Press, May 9, 2016, DOI 10.1074/jbc.M116.729020

**Kh. Shamsur Rahman**[‡], **Erfan Ullah Chowdhury**[‡], **Konrad Sachse**[§], **and Bernhard Kaltenboeck**[‡1]
*From the [‡]Department of Pathobiology, Auburn University, Auburn, Alabama 36849 and the [§]Federal Institute for Animal Health, D-07743 Jena, Germany*

- 630 confirmed epitopes
- 771 confirmed non-epitopes and random sequences

# Accuracy


1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.83)

# Outlook

- Fine-tune input and DNN parameters
- Evaluate prediction by spotting of predicted epitope sequences and verification with antibody sera
- Evaluation against other published epitopes

# Thanks

# The Deep Neural Network

- Deep Neural Network
  - Protein sequence
    - Two layers of LSTM with 50 nodes
    - One layer of 50 Dense nodes
  - Other parameters
    - One layer of 50 Dense nodes
  - Combined
    - One layer of 50 Dense nodes
    - One layer of 2 Dense nodes (epitope or non-epitope)