RUHR-UNIVERSITÄT BOCHUM

EXPLORING THE LIMITS OF THE TRYPTIC SEARCH SPACE USING PROTEIN-GRAPHS

PRODI

Dominik Lux - Protein-Graphs and Retrieval of Statistics

mpc MEDIZINISCHES PROTEOM CENTER

RUHR UNIVERSITÄT BOCHUM

RUB

Introduction

Me:

- M.Sc. in Computer Science
- (PhD-Student)

GitHub: https://github.com/Luxxii
E-Mail: dominik.lux@rub.de

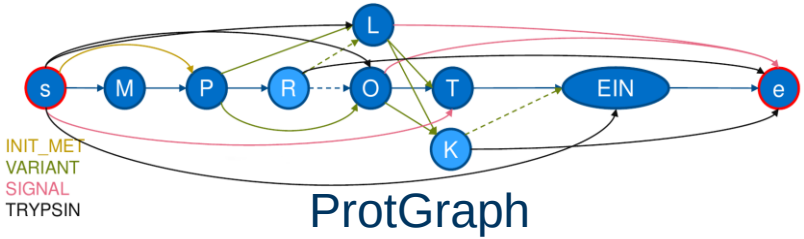
Work in: MPC (Medical Proteome Center) / Ruhr-University Bochum / Germany

- With:

Julian Uszkoreit

Dirk Winkelhardt

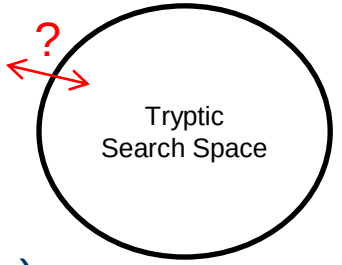
Mostly working on:



ProtGraph

RUHR UNIVERSITÄT BOCHUM RUB

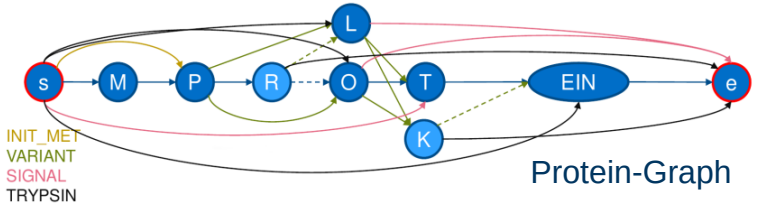
Overview and Motivation



Topic: Exploring the limits of the tryptic search space

- Retrieving the upper bound of peptides from a protein (efficiently)
- Deeper look into the upper bound of peptides present in a database
  - With and without features (explained later)

→ With the help of proteins-graphs



Protein-Graph

Why could this be important?

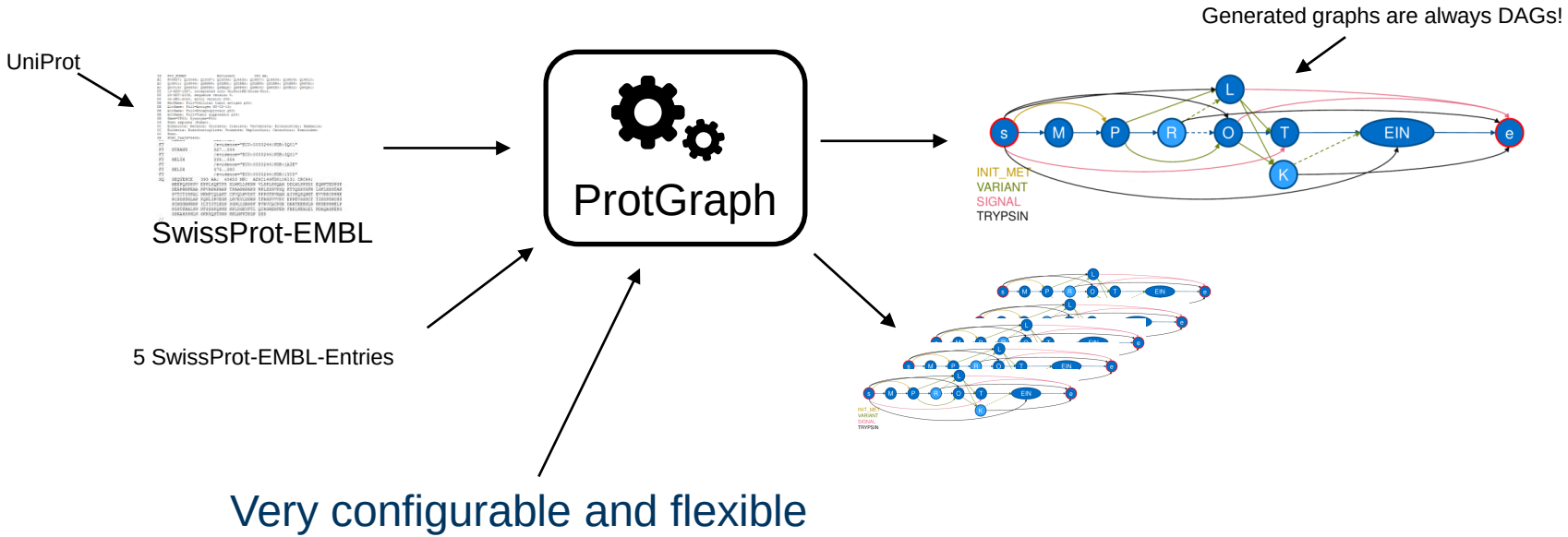
- Check possibility to generate a FASTA (proteogenomics?)
- Show complexity of the tryptic search space (is it searchable?)

ProtGraph is available on:
Github: https://github.com/mpc-bioinformatics/ProtGraph
PyPI: pip install protgraph
BioConda: conda install -c bioconda protgraph

ProtGraph in a Nutshell

→ No explanation of how graphs are generated

In a Nutshell:



Very configurable and flexible

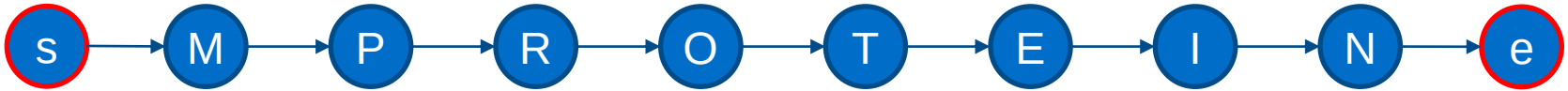
RUHR UNIVERSITÄT BOCHUM RUB

Possible Protein-Graphs

Snippet from SwissProt-EMBL:

```
SQ   SEQUENCE   8 AA;  988 MW;  XXXXXXXXXXXXXXXX CRC64;
     MPROTEIN
```

ProtGraph Parameters:
“None”



Possible Protein-Graphs

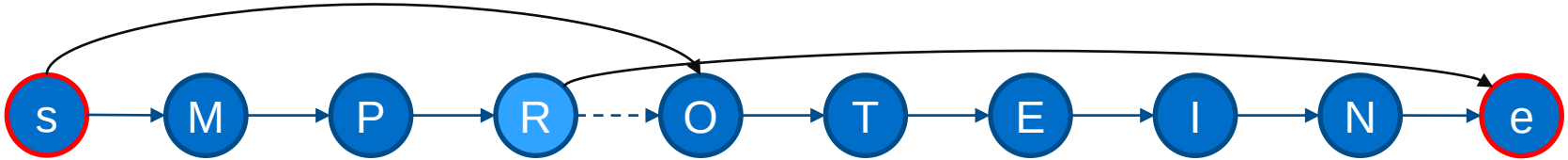
Snippet from SwissProt-EMBL:

```
SQ   SEQUENCE   8 AA;  988 MW;  XXXXXXXXXXXXXXXX CRC64;
     MPROTEIN
```

ProtGraph Parameters:
“None”, Trypsin-Digestion

Peptides:

s → MPR → e
s → OTEIN → e
s → MPROTEIN → e (1 miscleavage)



(graph contains 3 peptides)

TRYPSIN

Possible Protein-Graphs

Snippet from SwissProt-EMBL:

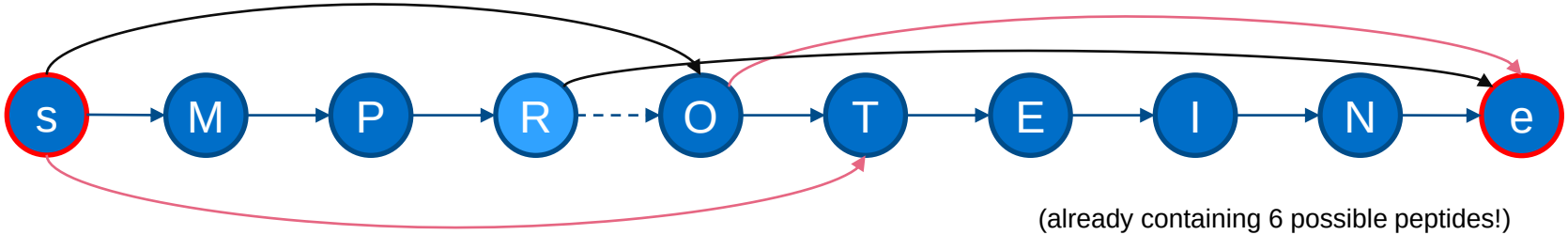
```
SQ   SEQUENCE   8 AA;  988 MW;  XXXXXXXXXXXXXXXX CRC64;
     MPROTEIN
```

ProtGraph Parameters:
Signal-Peptide, Trypsin-Digestion

Information derived from SP-EMBL:

```
FT   SIGNAL          1..4
FT                   /evidence="EXAMPLE Signal Peptide"
```

FeatureTable, hence features



(already containing 6 possible peptides!)

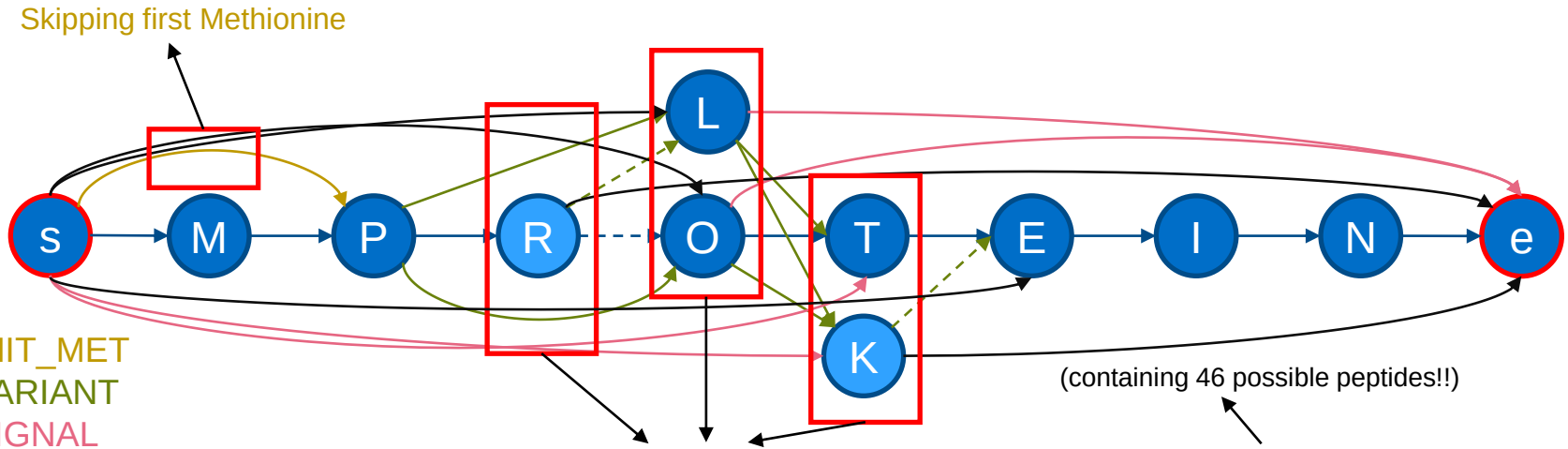
SIGNAL
TRYPSIN

Possible Protein-Graphs

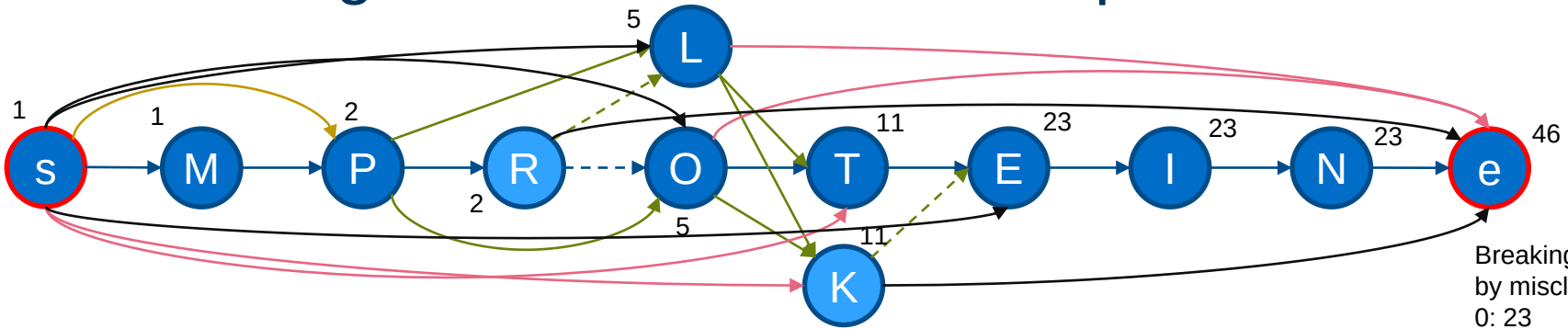
Snippet from SwissProt-EMBL:

```
SQ   SEQUENCE   8 AA;  988 MW;  XXXXXXXXXXXXXXX CRC64;
     MPROTEIN
```

ProtGraph Parameters:
Signal-Peptide, Variants, Initiator Methionine, Trypsin-Digestion



Calculating Number of Possible Peptides/Paths



Breaking it down by miscleavages:
0: 23
1: 19
2: 4

→ Use the topological order of the graph with dynamic programming!

Caution:

- A path from s to e can already represent a peptide from another path s to e
  - It is an upper bound for peptides contained in a graph
- #Possible Pep. contains peptides with up to “infinite many miscleavages”

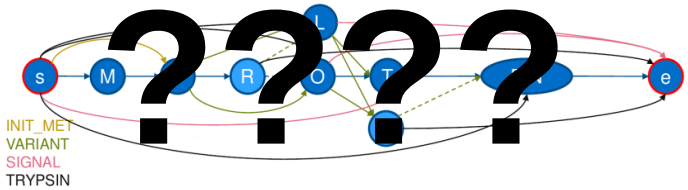
Protein-Graph Generation

ProtGraph parameters summarized:

- Digestion (fixed: Trypsin)
- Isoforms
- Variants
- Initiator Methionine
- Signal Peptides

- Mutagens
- Conflicts
- Aminoacid-Ambiguity-Resolving

Resolving:
B → D, N
J → I, L
Z → Q, E



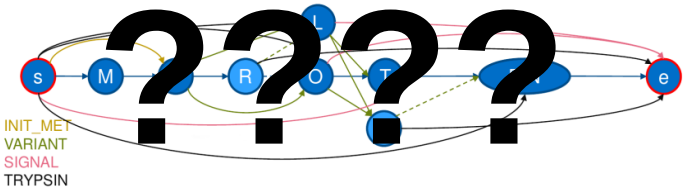
Protein-Graph Generation

ProtGraph parameters summarized:

- Digestion (fixed: Trypsin)
- ~~Isoforms~~
- Variants
- ~~Initiator Methionine~~
- Signal Peptides

- Mutagens
- ~~Conflicts~~
- Aminoacid-Ambiguity-Resolving

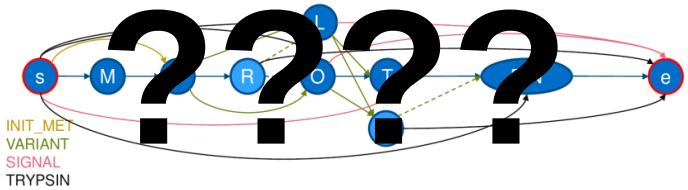
Resolving:
B → D, N
~~J → I, L~~
Z → Q, E



Protein-Graph Generation

ProtGraph parameters summarized:

- Digestion (fixed: Trypsin)
- Isoforms
- Variants
- Initiator Methionine
- Signal Peptides



- ~~Mutagens~~
- ~~Conflicts~~
- Aminoacid-Ambiguity-Resolving

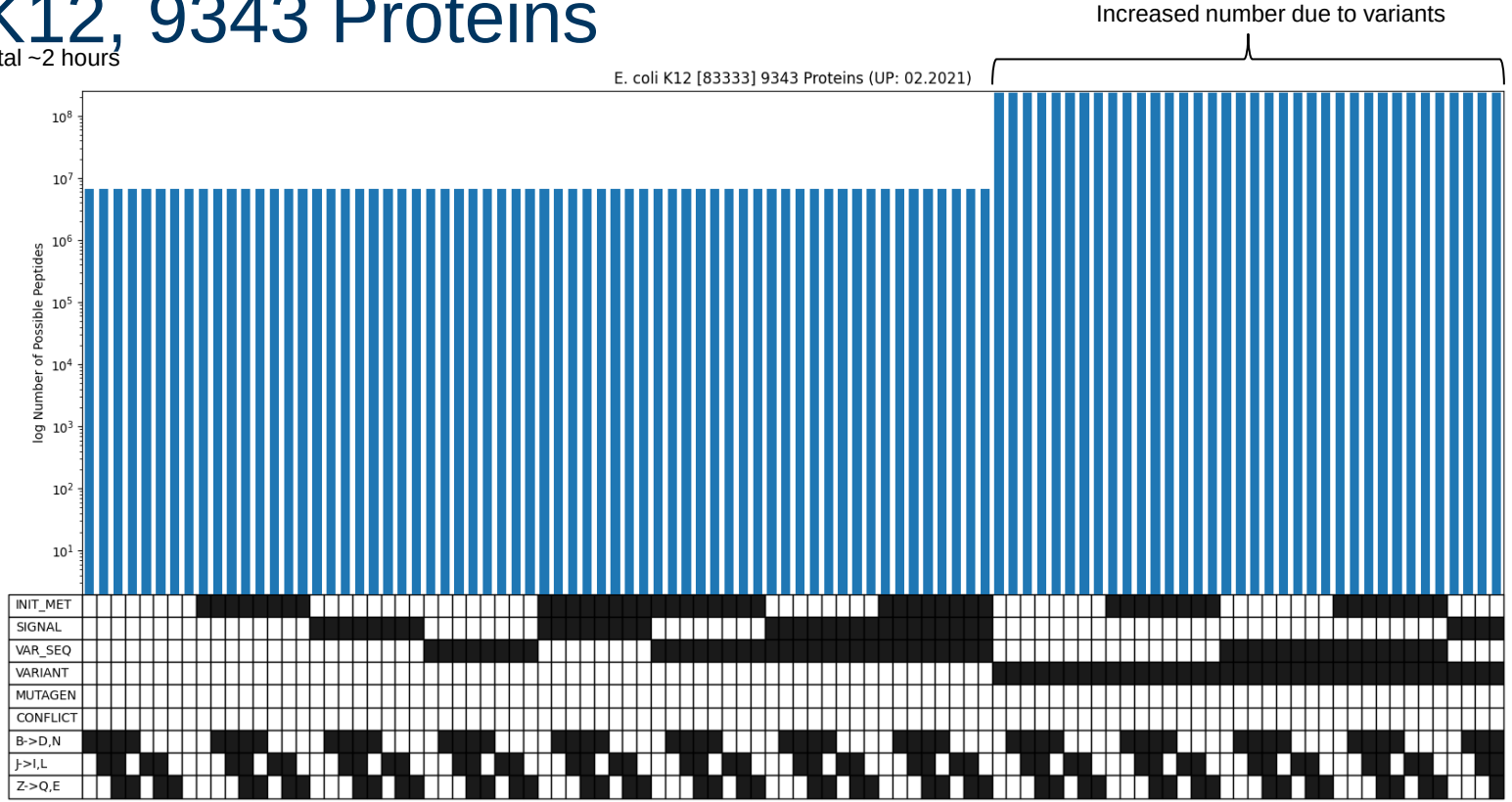
Resolving:
B → D, N
J → I, L
Z → Q, E

- There are in total: $2^9 = 512$ possibilities how ProtGraph can generate graphs
- Retrieve the #Possible Peptides (and #Possible Peptides by miscleavages) on specific datasets

E. Coli K12, 9343 Proteins

Calculation needed in total ~2 hours

- Bars are sorted in ascending order
- Size of search space is very similar while including specific features
- Resolving ambiguity only increase the search space slightly

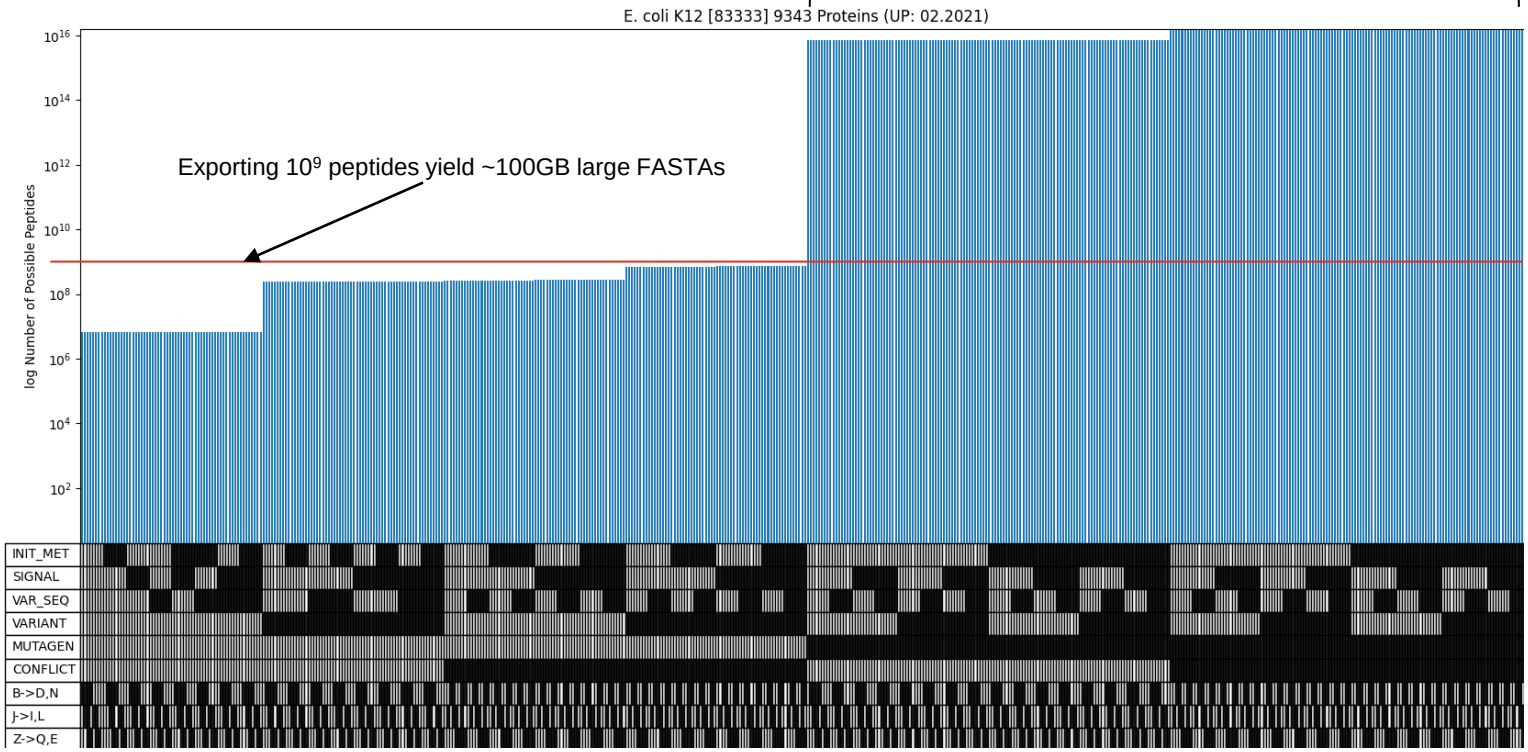


RUHR UNIVERSITÄT BOCHUM RUB

E. Coli K12, 9343 Proteins

Calculation needed in total ~2 hours

Large increase while using Mutagens



- Bars are sorted in ascending order
- Size of search space is very similar while including specific features
- Resolving ambiguity only increase the search space slightly
- Exporting FASTAs using all features is not feasible for E. Colit K12

Only change the search space slightly

Applied Feature ■
Not Applied Feature □

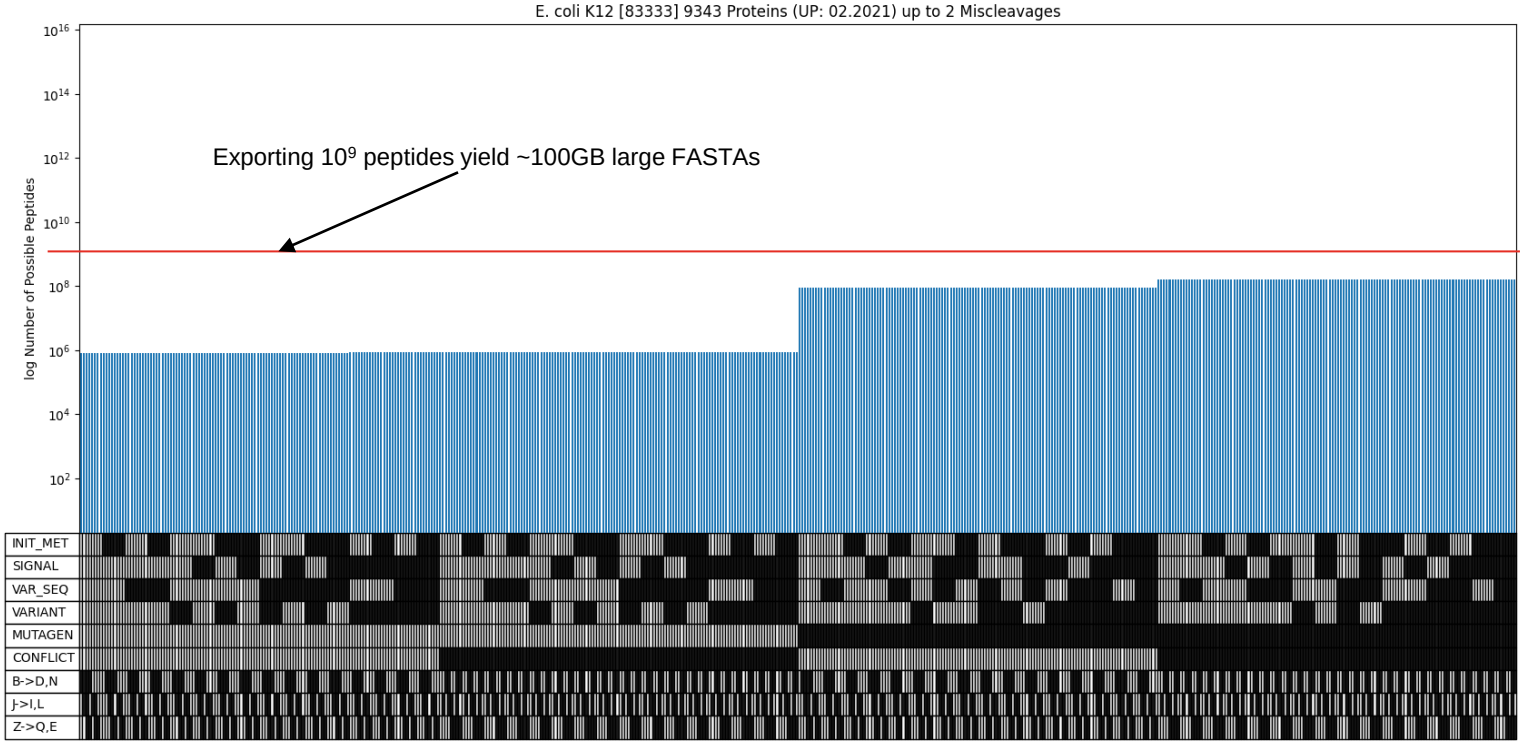
E. Coli K12, 9343 Proteins

Calculation needed in total ~2 hours

Large increase while using Mutagens

- Export is feasible if only using peptides with up to 2 miscleavages

→ E.coli K12 can be searched with all its features with FASTA files!



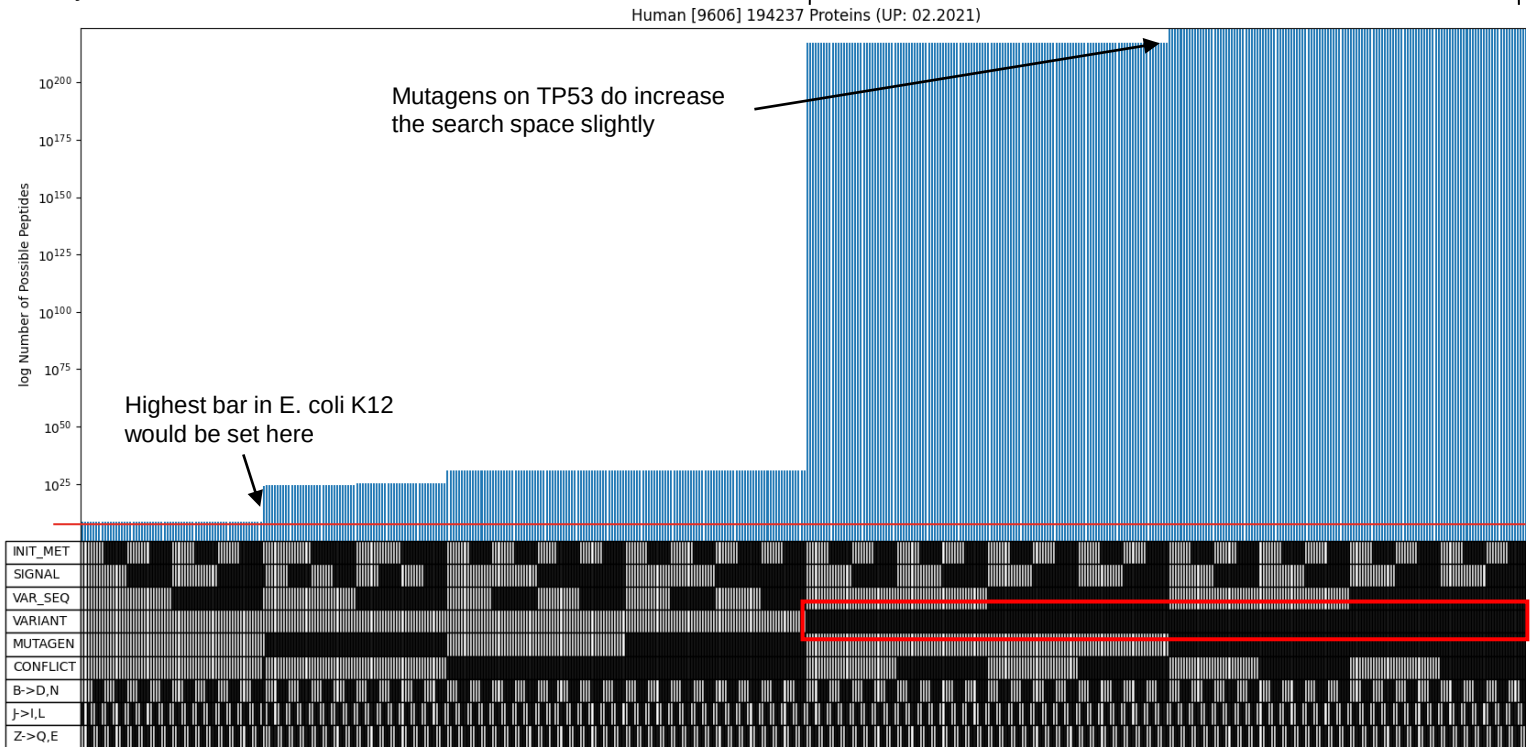
Only change the search space slightly

Applied Feature ■
Not Applied Feature □

Human, 194 237 Proteins

Calculation needed in total ~4 Days

- Main factor for the large search space:
  - Variants
- Search space is dominated by a single Protein:
  - P04637
  - TP53
- TP53 has highest number of variants (~1000)



Only change the search space slightly

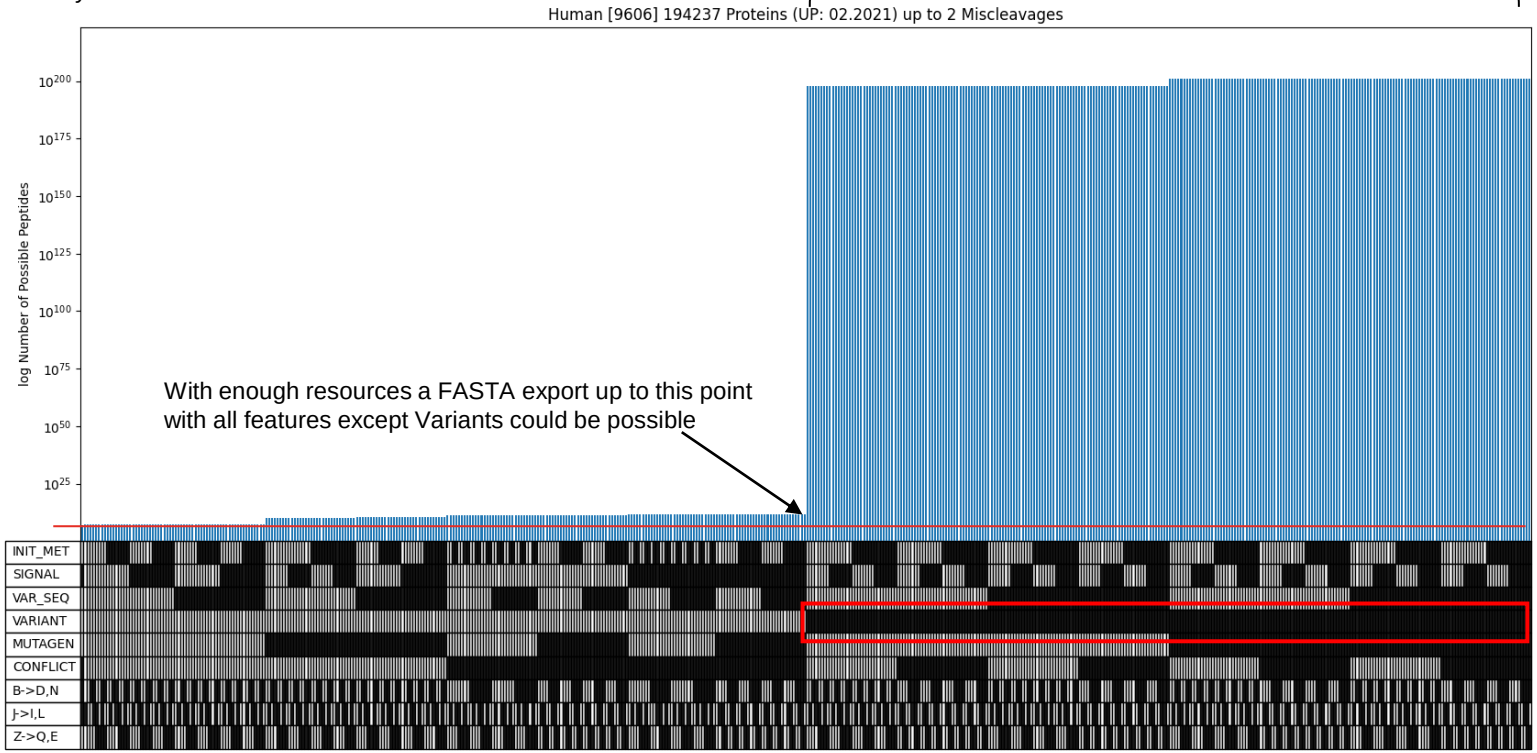
Applied Feature ■
Not Applied Feature □

Human, 194 237 Proteins

Calculation needed in total ~4 Days

Search space “explodes” while using Variants

- Main factor for the large search space:
  - Variants
- Search space is dominated by a single Protein:
  - P04637
  - TP53
- TP53 has highest number of variants (~1000)
- Not feasible to export FASTAs!

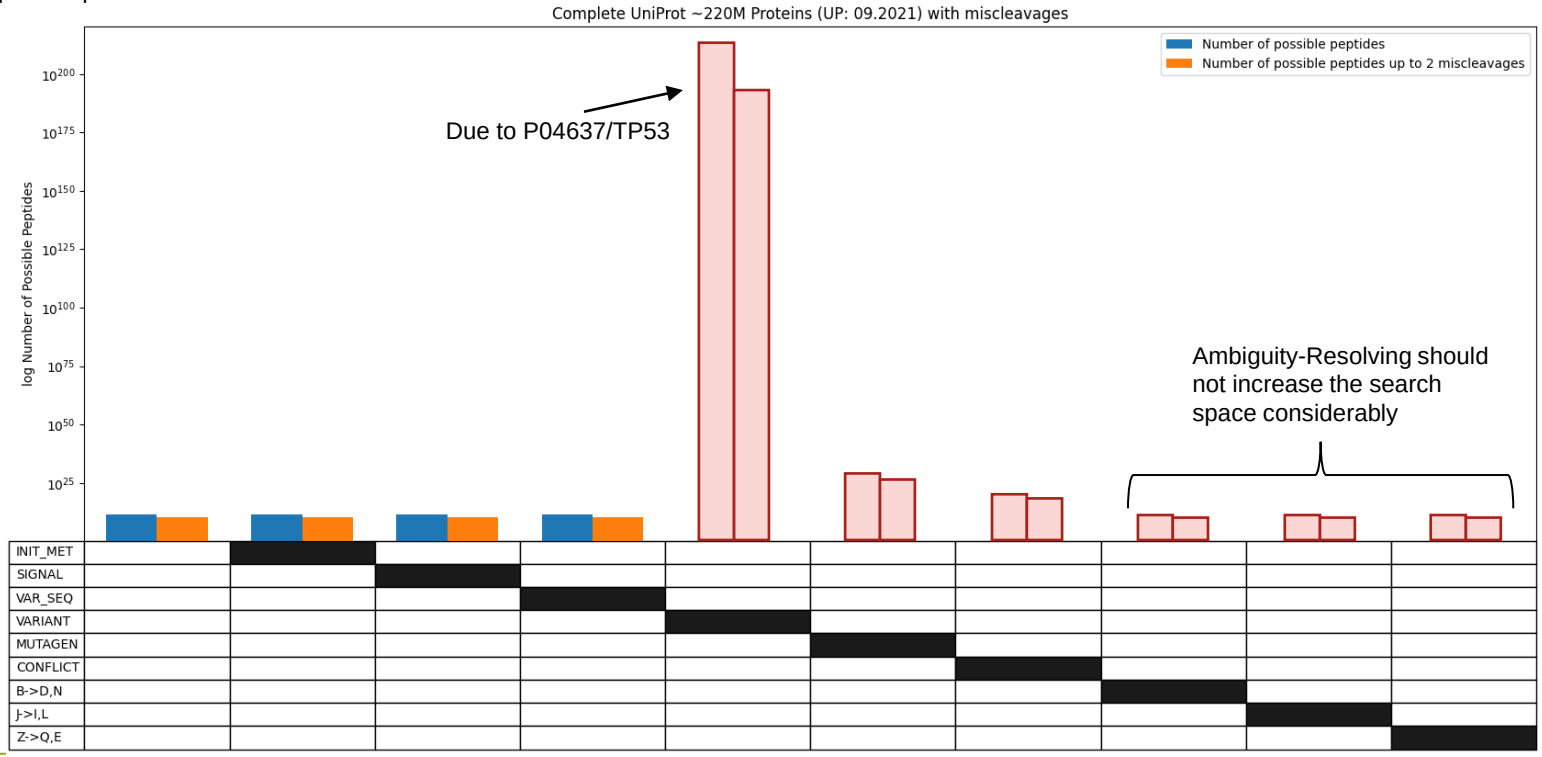


RUHR UNIVERSITÄT BOCHUM RUB

Complete UniProt, ~220M Proteins

Calculation needed ~26h per bar-pair

- Incomplete results
- Blue/Orange:
  - Calculated
- Red:
  - Assumption
- P04637/TP53 still dominates the search space while using variants

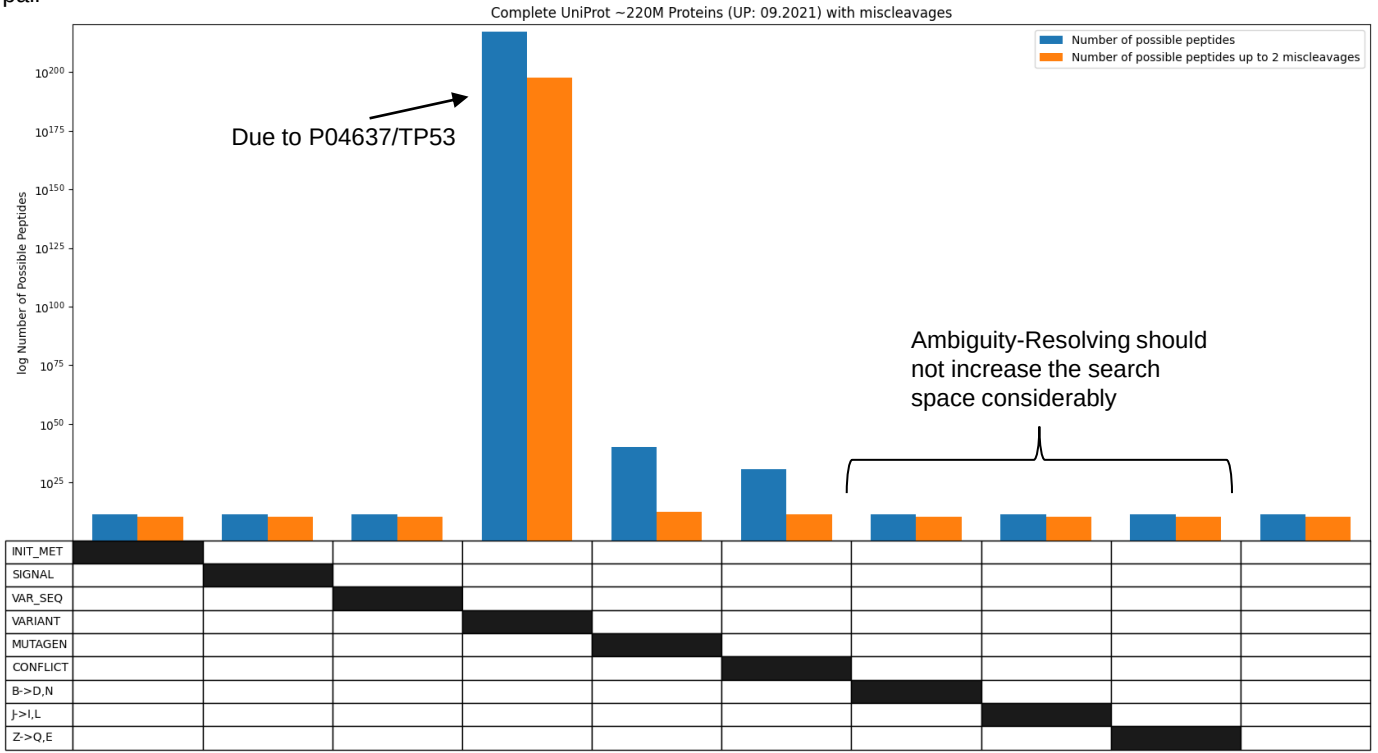


RUHR UNIVERSITÄT BOCHUM RUB

Complete UniProt, ~220M Proteins (updated slide)

Calculation needed ~26h per bar-pair

- Incomplete results
- Blue/Orange:
  - Calculated
- Red:
  - Assumption
- P04637/TP53 still dominates the search space while using variants



Applied Feature
Not Applied Feature

RUHR UNIVERSITÄT BOCHUM
RUB

Conclusion

- → Graphs can contain huge amounts of peptides while being small in size
  - → Up to $2^{\#Nodes - 2}$ many peptides can be represented in a graph

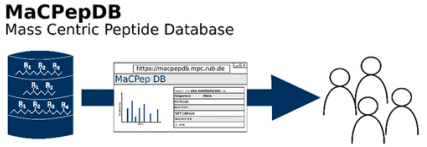
- → The number of Variants (and probably Mutagens) increase the search space drastically to unmanageable amounts of peptides
  - → FASTA-Exports may not be feasible!

- → FASTA-Exports may be possible for small datasets (E.G. E. coli K12)
  - → … and can be searched with (preliminary results)

Conclusion for future projects

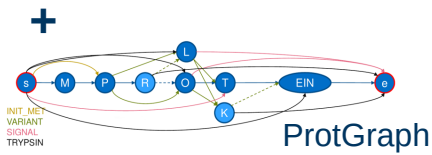
→ Increase the size of MacPepDB to cover more peptides

- Resolving Ambiguity / Signal Peptides / Isoforms / Initiator Methionine can be included safely without exploding in size



→ These statistics could be used protein-wise

- … for designing (limits on) algorithms on protein-graphs
  - E.G. Useful for extracting peptides directly from graphs
- … to create a hybrid approach between MacPepDB and ProtGraph
  - … tackling long running peptide-queries for the complete UniProt-Dataset



Acknowledgement

Medical Proteome Center:

- PD Dr. Martin Eisenacher
- **Dr. Julian Uszkoreit**
- Dr. Michael Turewicz
- Dr. Markus Stepath
- **Dirk Winkelhardt**
- Daniel Kleefisch
- Karin Schork
- Sai Spoorti Ramesh

Thank You for your attention!

F953N-19

FoRUM

SPONSORED BY THE

Federal Ministry of Education and Research

FKZ 031 A 534A

de.NBI

GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

PRODI