RUHR-UNIVERSITÄT BOCHUM

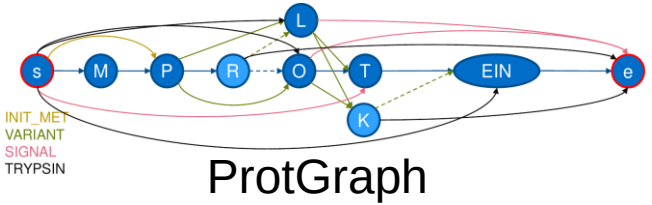
PROTGRAPH

A GRAPH APPROACH FOR REPRESENTING PROTEINS AND PEPTIDES

Dominik Lux

RUHR UNIVERSITÄT BOCHUM

ProtGraph



ProtGraph

→ Created at the MPC

→ Written/Usable with Python >=3.6

Available on:

GitHub: https://github.com/mpc-bioinformatics/ProtGraph

PyPI: pip install protgraph

BioConda: conda install protgraph --channel bioconda

→ Introducing ProtGraph: internal generation of protein-graphs / interesting results

Motivation

- Proteomic-Identification-Workflows usually use following FASTAs for search engines:
  - Canonical sequences (sometimes with isoform sequences)

→ FASTA does not provide information about Variants/Mutagens/Conflicts/etc …

→ Why is it not used? Data is publicly available in UniProt!

→ **Our goal**: Provide an up-to-date FASTA, including such information

→ ProtGraph arose, utilizing a graph structure, allowing to generate such FASTAs

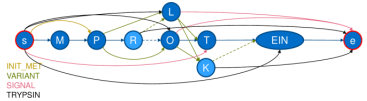
→ Side-Effect: Generating FASTA-files in some cases is not feasible!

→ Side-Effect: Allows for a deeper look into the (tryptic) search space

ProtGraph (Swiss-Prot/EMBL)

SwissProt-EMBL

ProtGraph



- SwissProt-EMBL provides needed information
  - Variants/Signal Peptides/etc…
- EMBL is a standardized data format for DNA (and proteins)

  (European Molecular Biology Laboratory)
- UniProt: Each Entry is saved in a special EMBL-Format

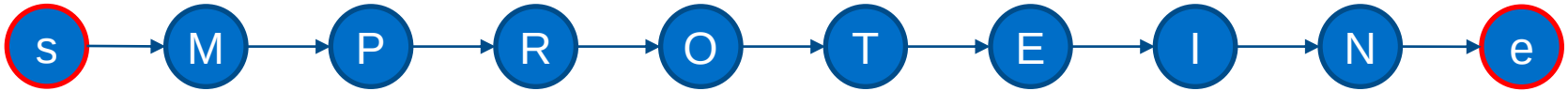
  (UniProt follows EMBL as closely as possible: www.uniport.org/docs/userman.htm)

Swiss-Prot/EMBL to Protein-Graph

Checklist:
Canonical

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

→ Add the canonical sequence
  → Generate the initial graph: Each amino acid = node

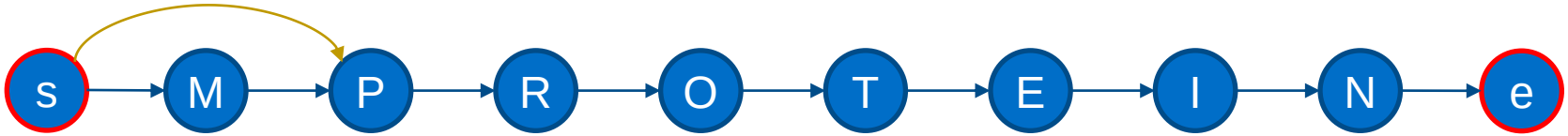


Swiss-Prot/EMBL to Protein-Graph

```
FT   INIT_MET        1
FT                   /note="Removed"
FT                   /evidence="EXAMPLE Initiator Methionine"
```

Checklist:
Canonical
Init. Met.

→ Adding Initiator Methionine



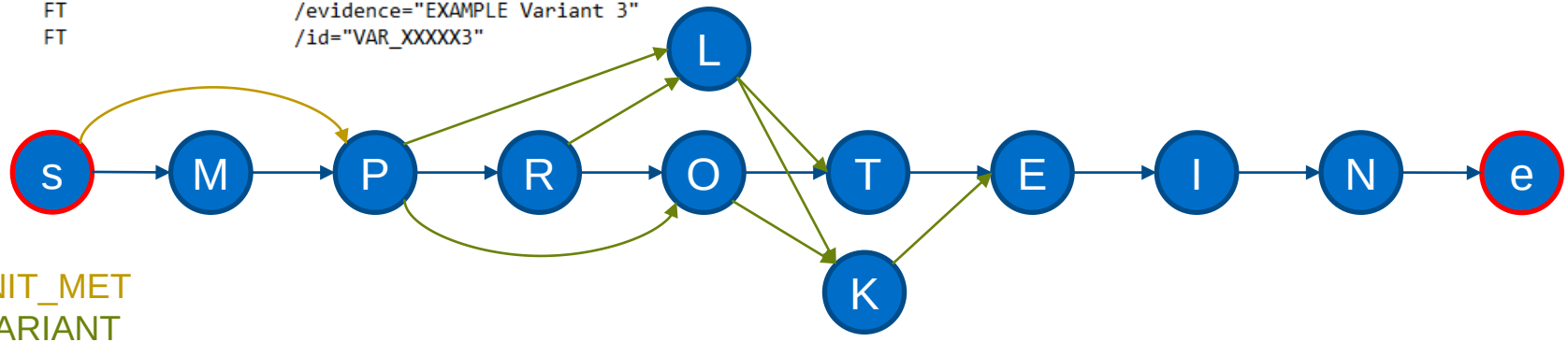
INIT_MET

Swiss-Prot/EMBL to Protein-Graph

```
FT   VARIANT         3
FT                   /note="Missing (in Example)"
FT                   /evidence="EXAMPLE Variant 1"
FT                   /id="VAR_XXXXX1"
FT   VARIANT         4
FT                   /note="O -> L (in Example)"
FT                   /evidence="EXAMPLE Variant 2"
FT                   /id="VAR_XXXXX2"
FT   VARIANT         5
FT                   /note="T -> K (in Example)"
FT                   /evidence="EXAMPLE Variant 3"
FT                   /id="VAR_XXXXX3"
```

→ Adding Feature Variants
- 2 Types: Insertion/Deletion

Checklist:
- Canonical
- Init. Met.
- Variants



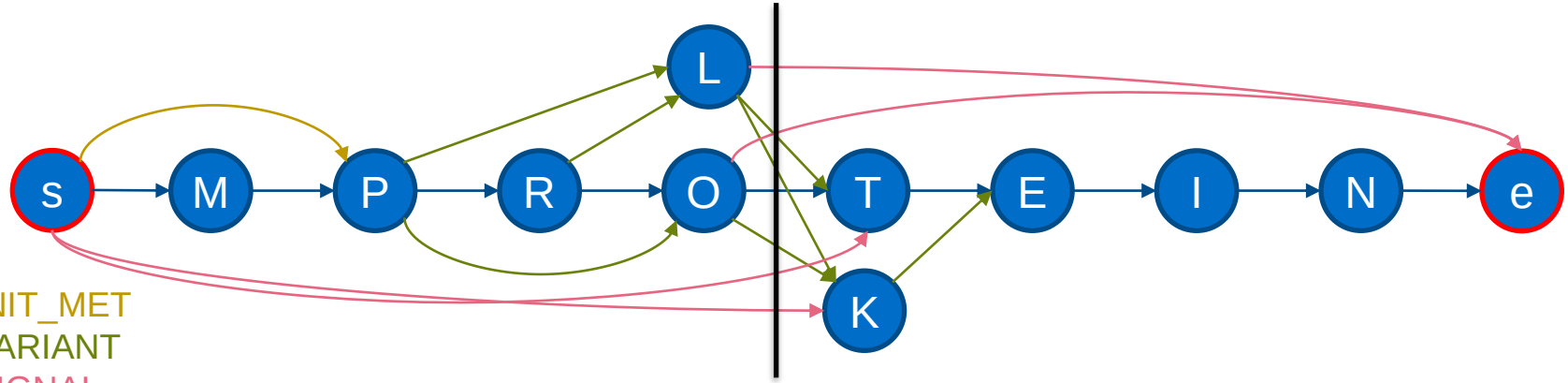
INIT_MET
VARIANT

RUHR UNIVERSITÄT BOCHUM RUB

Swiss-Prot/EMBL to Protein-Graph

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

→ Adding Feature Signal Peptide



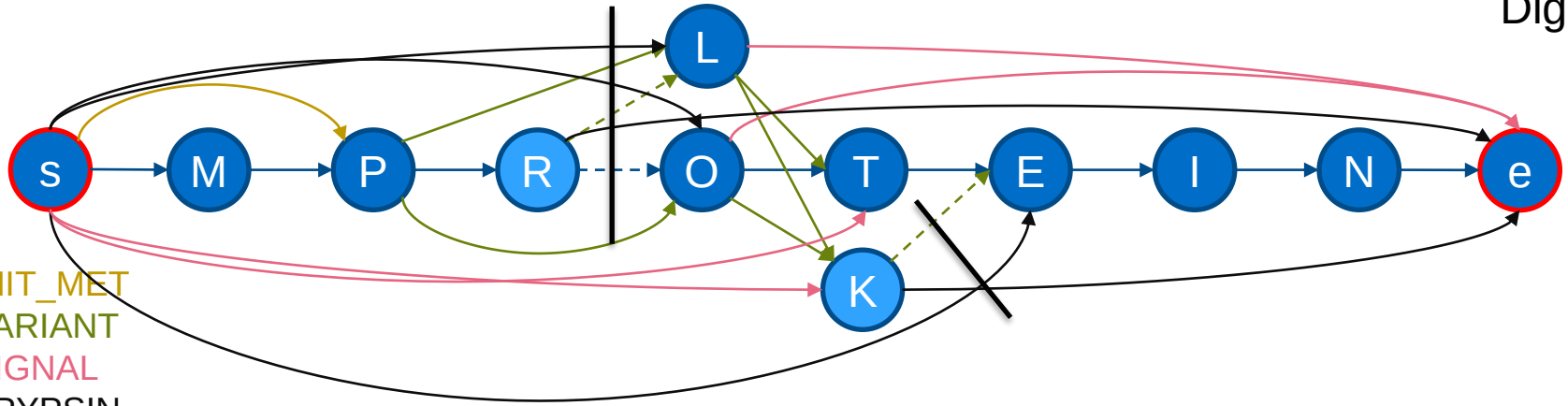
INIT_MET
VARIANT
SIGNAL

Checklist:
- Canonical
- Init. Met.
- Variants
- Signal Pep.

Swiss-Prot/EMBL to Protein-Graph

- → Digestion with arbitrary rules are possible
  - → E.G.: Trypsin
    (cut after R and K expect if the next amino acid is P)

Checklist:
- Canonical
- Init. Met.
- Variants
- Signal Pep.
- Digestion

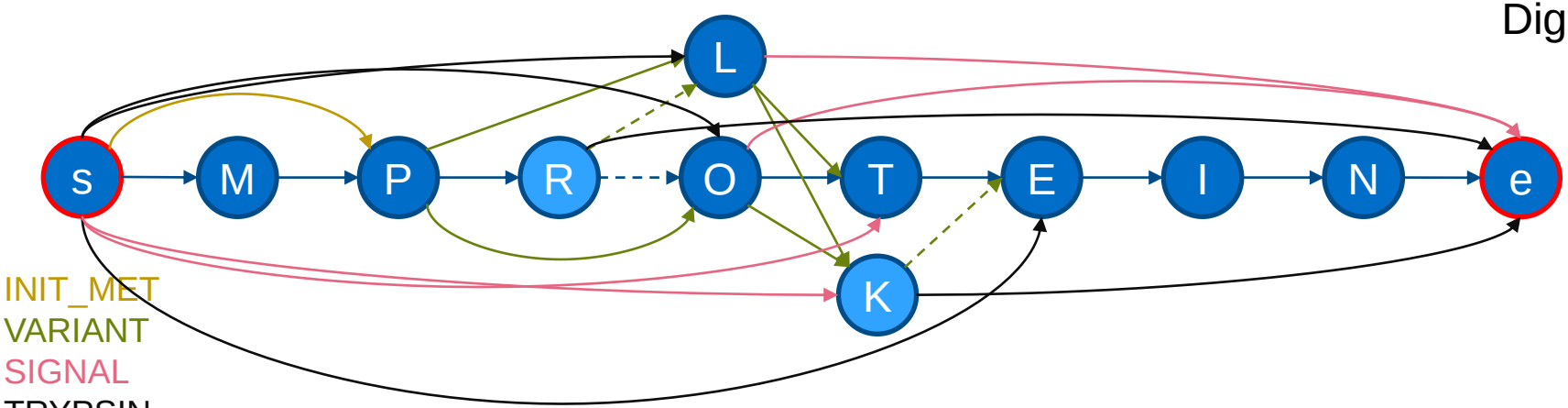


RUHR UNIVERSITÄT BOCHUM RUB

Swiss-Prot/EMBL to Protein-Graph

ShowCase:

Checklist:
- Canonical
- Init. Met.
- Variants
- Signal Pep.
- Digestion

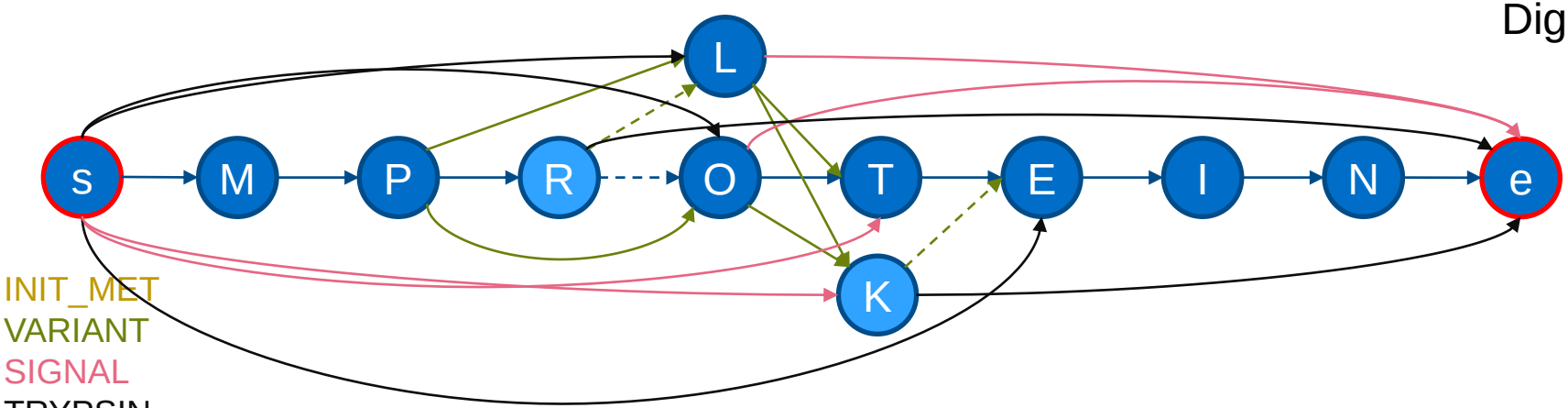


RUHR UNIVERSITÄT BOCHUM RUB

Swiss-Prot/EMBL to Protein-Graph

ShowCase:

Checklist:
- Canonical
- ~~Init. Met.~~
- Variants
- Signal Pep.
- Digestion

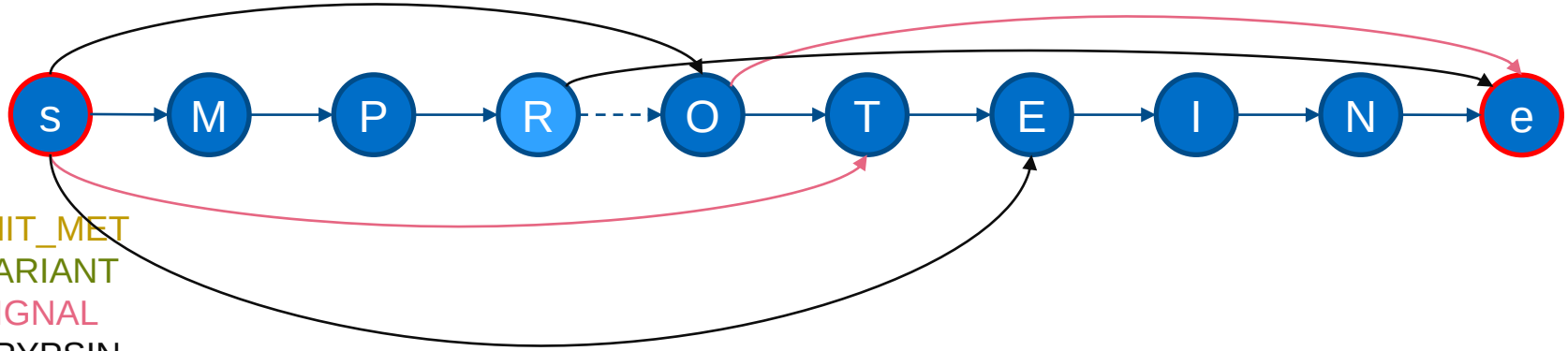


RUHR UNIVERSITÄT BOCHUM RUB

Swiss-Prot/EMBL to Protein-Graph

ShowCase:

Checklist:
- Canonical
- ~~Init. Met.~~
- ~~Variants~~
- Signal Pep.
- Digestion

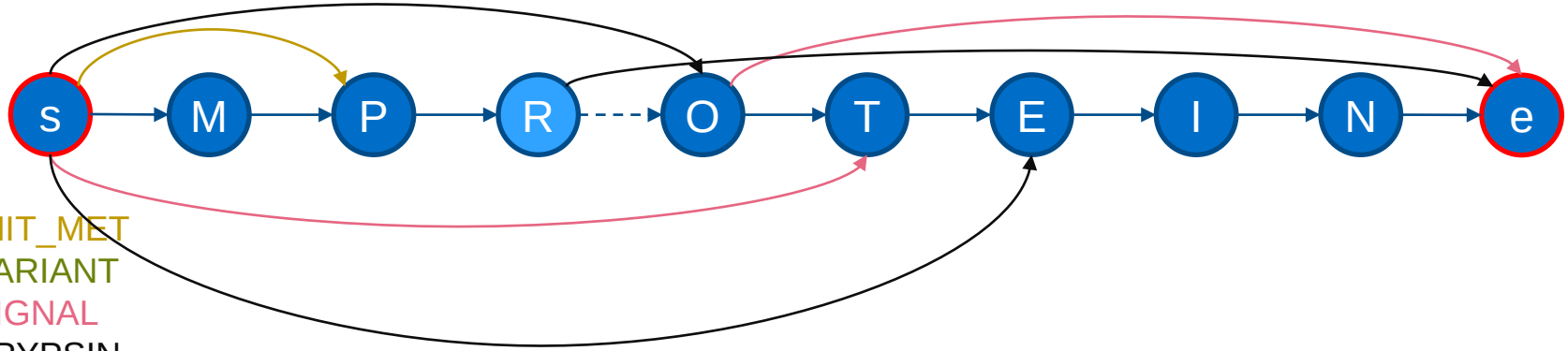


RUHR UNIVERSITÄT BOCHUM RUB

Swiss-Prot/EMBL to Protein-Graph

ShowCase:

Checklist:
- Canonical
- Init. Met.
- ~~Variants~~
- Signal Pep.
- Digestion

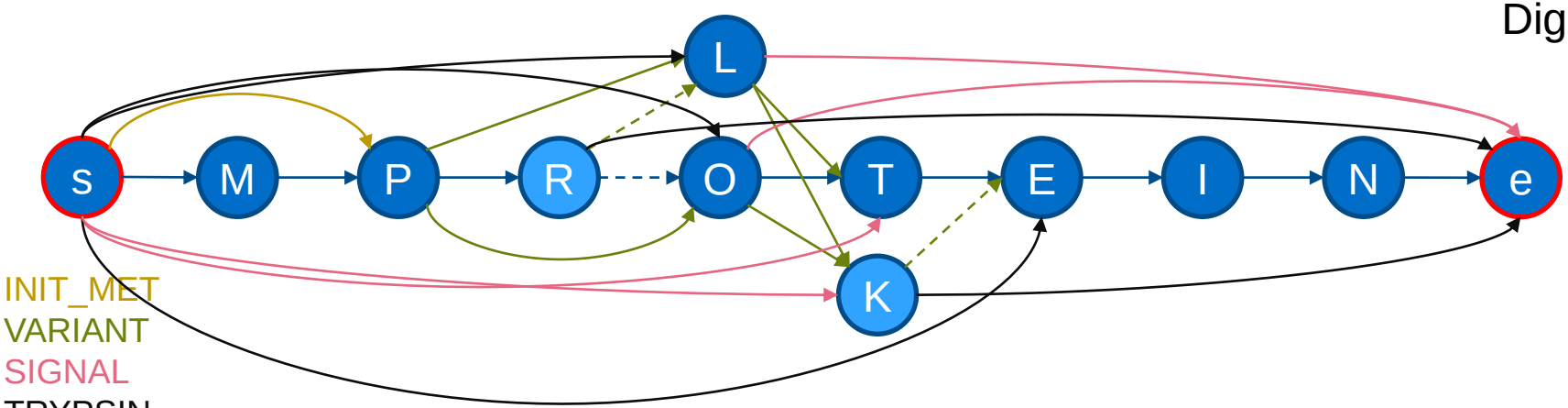


RUHR UNIVERSITÄT BOCHUM RUB

Swiss-Prot/EMBL to Protein-Graph

ShowCase:

Checklist:
- Canonical
- Init. Met.
- Variants
- Signal Pep.
- Digestion

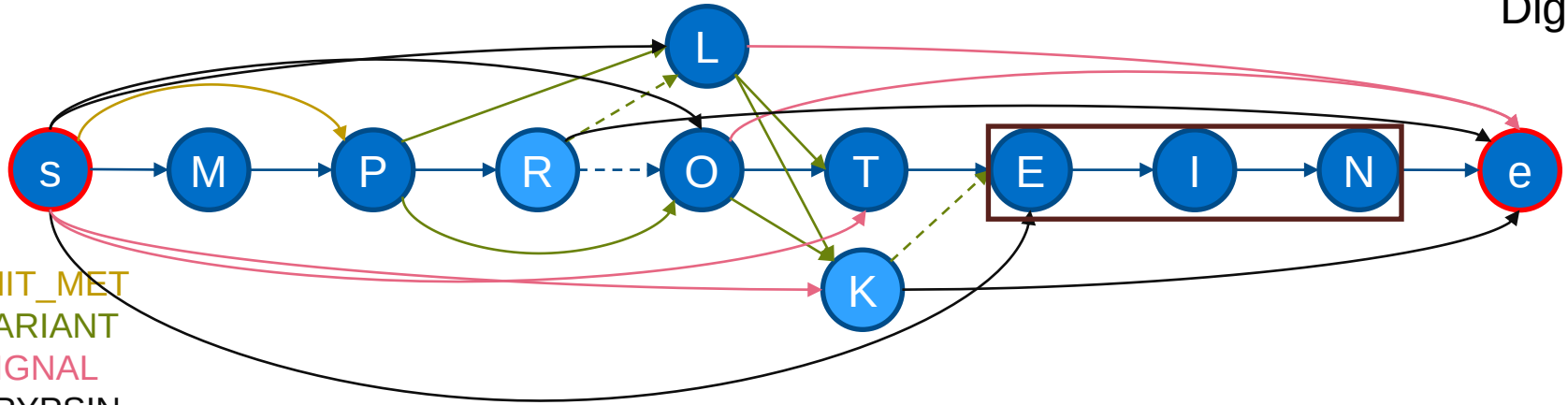


RUHR UNIVERSITÄT BOCHUM RUB

Swiss-Prot/EMBL to Protein-Graph

→ Optimize graph by concatenating chains

Checklist:
- Canonical
- Init. Met.
- Variants
- Signal Pep.
- Digestion



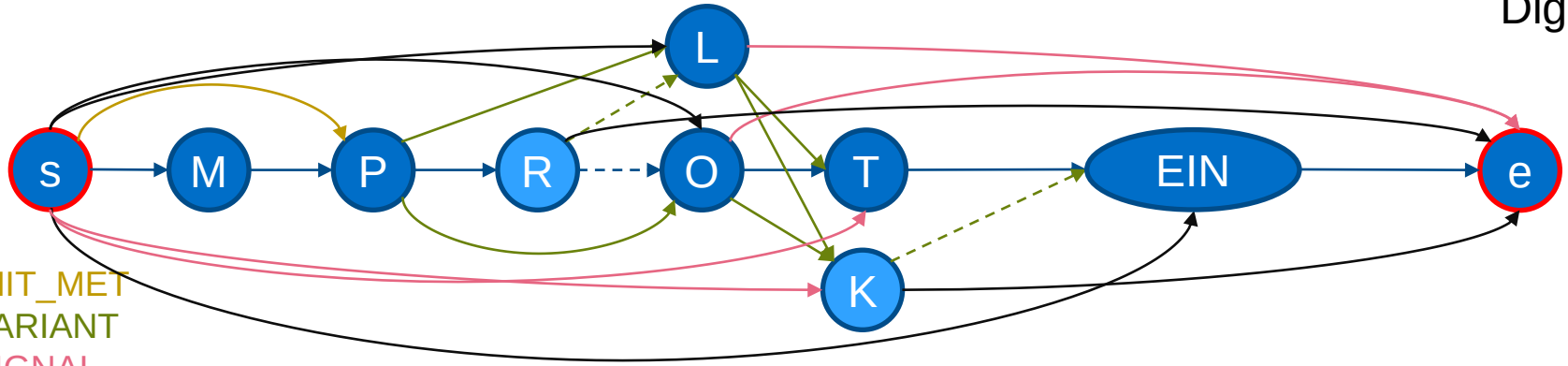
INIT_MET
VARIANT
SIGNAL
TRYPSIN

RUHR UNIVERSITÄT BOCHUM RUB

Swiss-Prot/EMBL to Protein-Graph

→ Optimize graph by concatenating chains

Checklist:
- Canonical
- Init. Met.
- Variants
- Signal Pep.
- Digestion

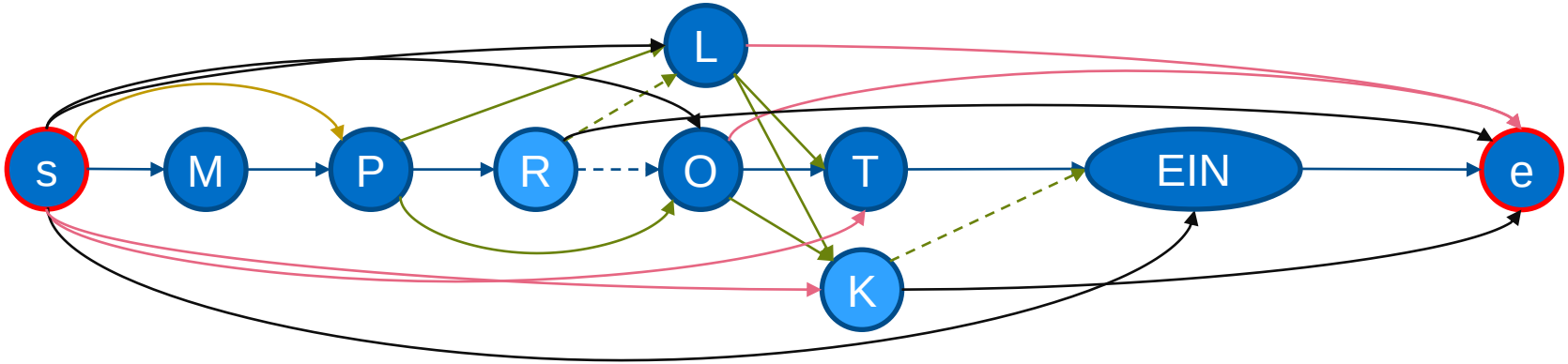


INIT_MET
VARIANT
SIGNAL
TRYPSIN

→ Reduces the graph size drastically for most proteins

RUHR UNIVERSITÄT BOCHUM RUB

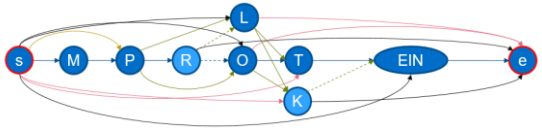
Information Retrieval with Top. Order



- ProtGraph generates in all cases Directed and Acyclic Graphs (DAGs)
- Topological Order:
  - s → M → P → R → O → L → T → K → EIN → e
    (One of multiple possible solutions!)

→ Topological Orders allow to retrieve interesting statistics per protein!

Information Retrieval with Top. Order



→ Calculate the number of possible peptides in a protein* 46

→ Get the number of peptides by miscleavages*

#Mssclvgs	0	1	2
#Peptides	23	19	4

→ Get the number peptides by length**

Length	1	2	3	4	5	6	7	8
#Peptides	3	5	8	8	6	4	8	4

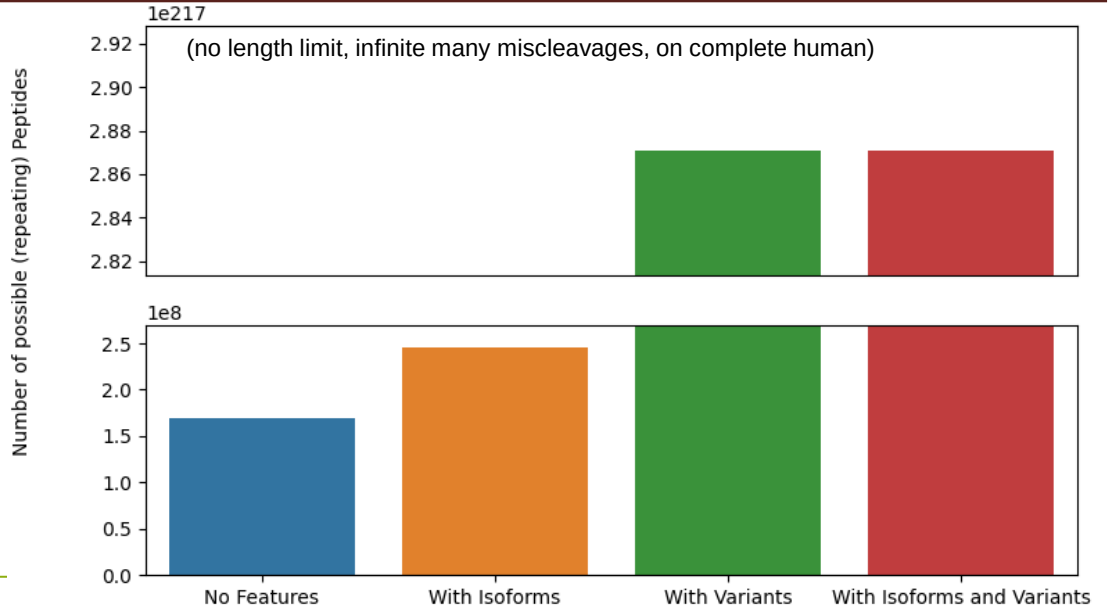
→ Application on real proteins!

* Peptides can occur multiple times in one protein

** Under the premise of “infinte” many miscleavages, counting amino acids

Information Retrieval with Top. Order

P09669 (COX6C Human):	120*
P73840 (SYUA Human):	1 385 365*
P04637 (P53 Human):	2.875E+217*



P53:

- Dominates the search space
- Has most variants
- Most researched protein?

→ Deeper look into P53

* Peptides can occur multiple times in one protein

Information Retrieval with Top. Order

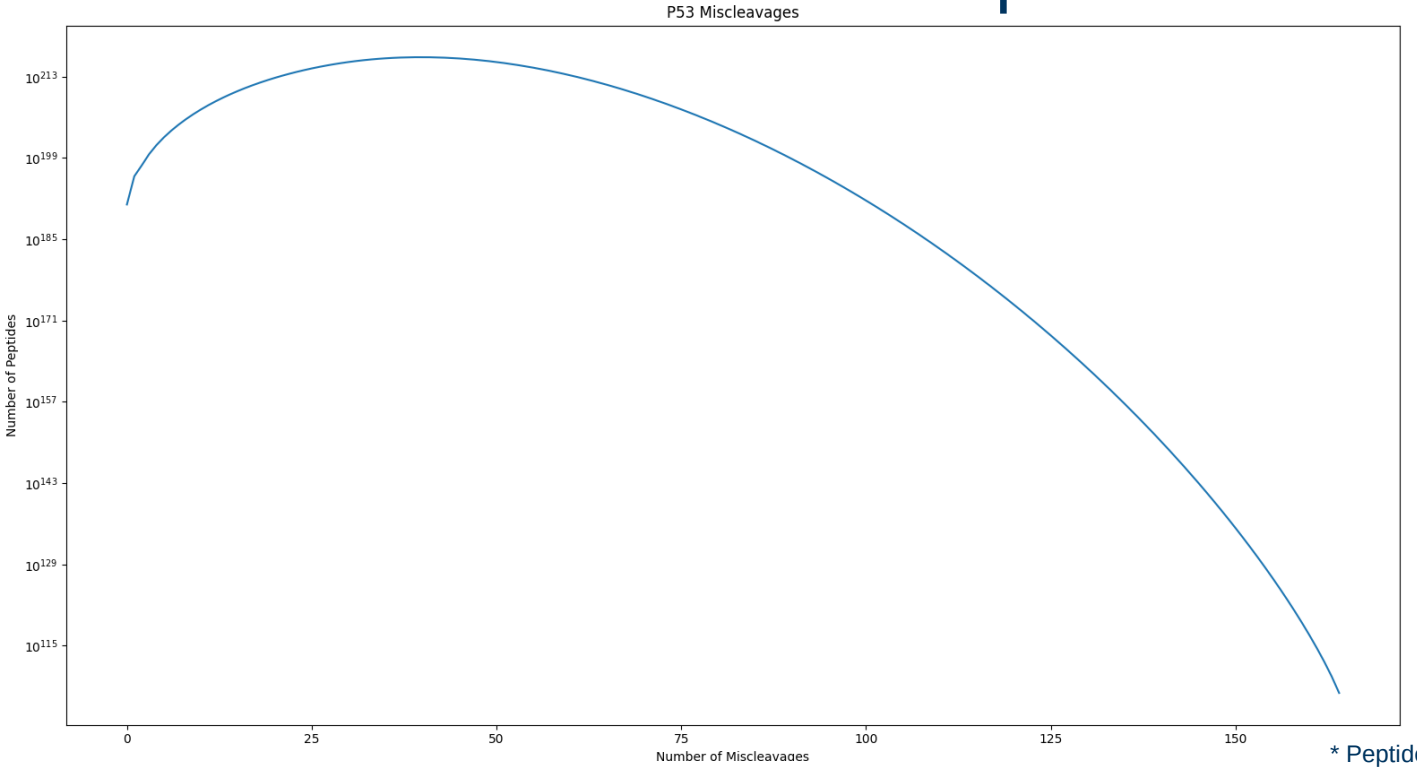


Figure:
Number of possible peptides* by miscleavages for P04637 (P53 Human)

* Peptides can occur multiple times in one protein

RUHR UNIVERSITÄT BOCHUM
RUB

Information Retrieval with Top. Order

#Mssclvgs	#Possible Peptides*
0	7 488 643 278
1	7 278 922 167
2	7 069 457 392
3	6 860 625 343
4	6 652 821 545
5	6 446 737 427
6	6 243 206 955
7	6 042 551 084
8	5 844 964 760
9	5 650 834 085

→ Calculate the number of all possible tryptically digested peptides by miscleavages

- On complete UniProt
- Using only the canonical sequence

(actual numbers can vary slightly depending on the version of UniProt if recalculated)

* Peptides can occur multiple times in one protein

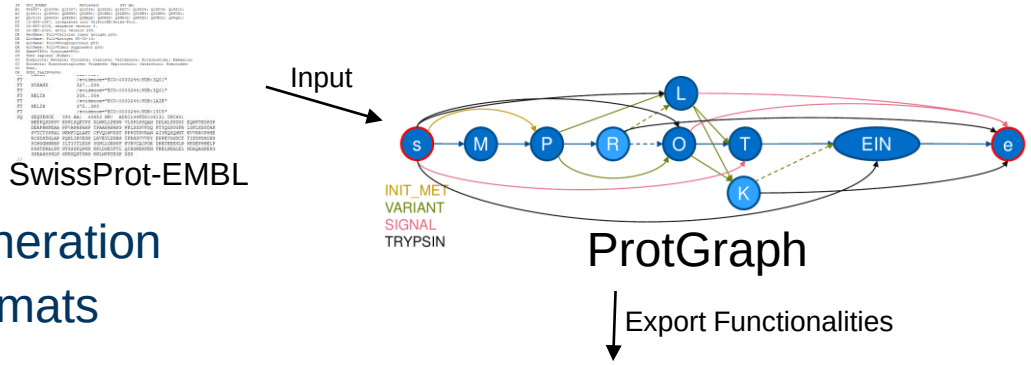
Outlook

ProtGraph:

- Efficient and quick graph generation
- Allows exports in various formats
  - Proteins/Peptides
  - Graphs
- Flexible
  - E.G. Selectable features to be applied on graphs

Future Work:

- Utilizing graphs directly
  - Querying graphs by weights of peptides



RUHR UNIVERSITÄT BOCHUM RUB

Acknowledgement

ProtGraph:
- Dominik Lux

ProtGraph Soul Support:
- PD Dr. Martin Eisenacher
- Dr. Julian Uszkoreit
- Dirk Winkelhardt

Thank you for your attention!

PRODI

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

F953N-19
FoRUM

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
FKZ 031 A 534A

elixir GERMANY

BioInfra.Prot
0010101000110

mpc MEDIZINISCHES PROTEOM CENTER
RUHR UNIVERSITÄT BOCHUM
RUB