NPS: scoring and evaluating the statistical significance of peptidic natural product–spectrum matches

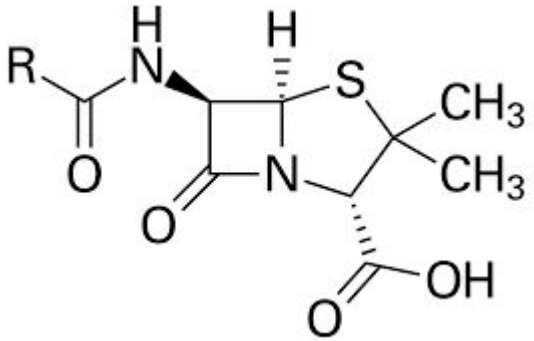
Alexey Gurevich

Center for Algorithmic Biotechnology
St. Petersburg State University, Russia

PNP basics

Peptidic Natural Products (PNPs):

- small bioactive peptidic molecules

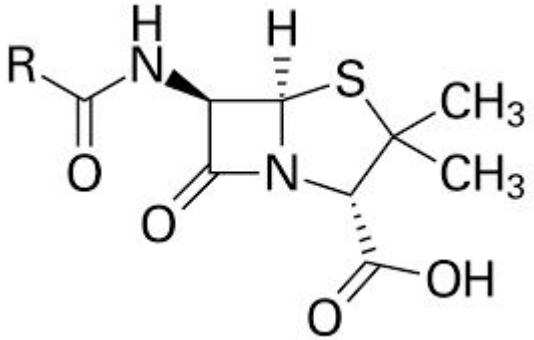


PNP basics

Peptidic Natural Products (PNPs):

- small bioactive peptidic molecules
- nonribosomal peptides (NRPs) (Marahiel et al., 1997) and

  ribosomally synthesized and post-translationally modified peptides (RiPPs) (Arnison et al., 2013)

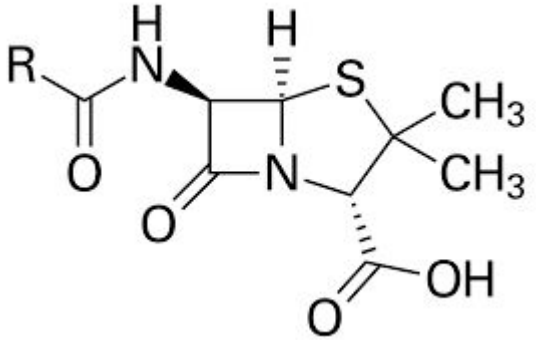


PNP basics

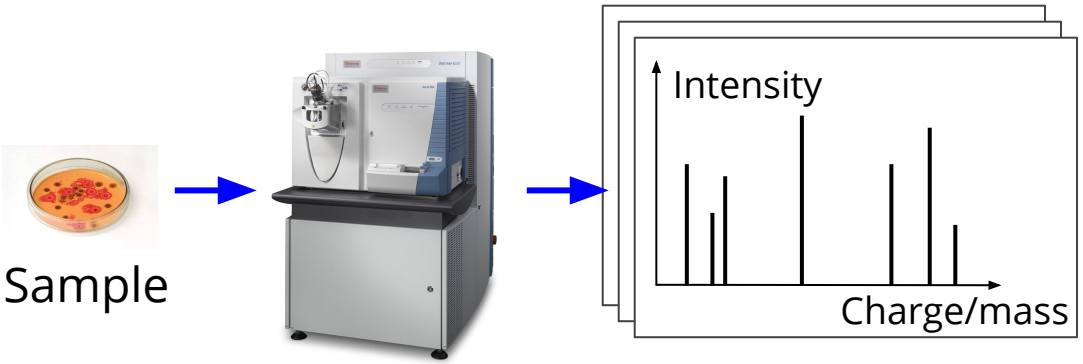
Peptidic Natural Products (PNPs):

- small bioactive peptidic molecules
- nonribosomal peptides (NRPs) (Marahiel et al., 1997) and

  ribosomally synthesized and post-translationally modified peptides (RiPPs) (Arnison et al., 2013)
- promising compounds in the drug research (including antibiotics, antitumor agents, ...)



Computational MS/MS basics

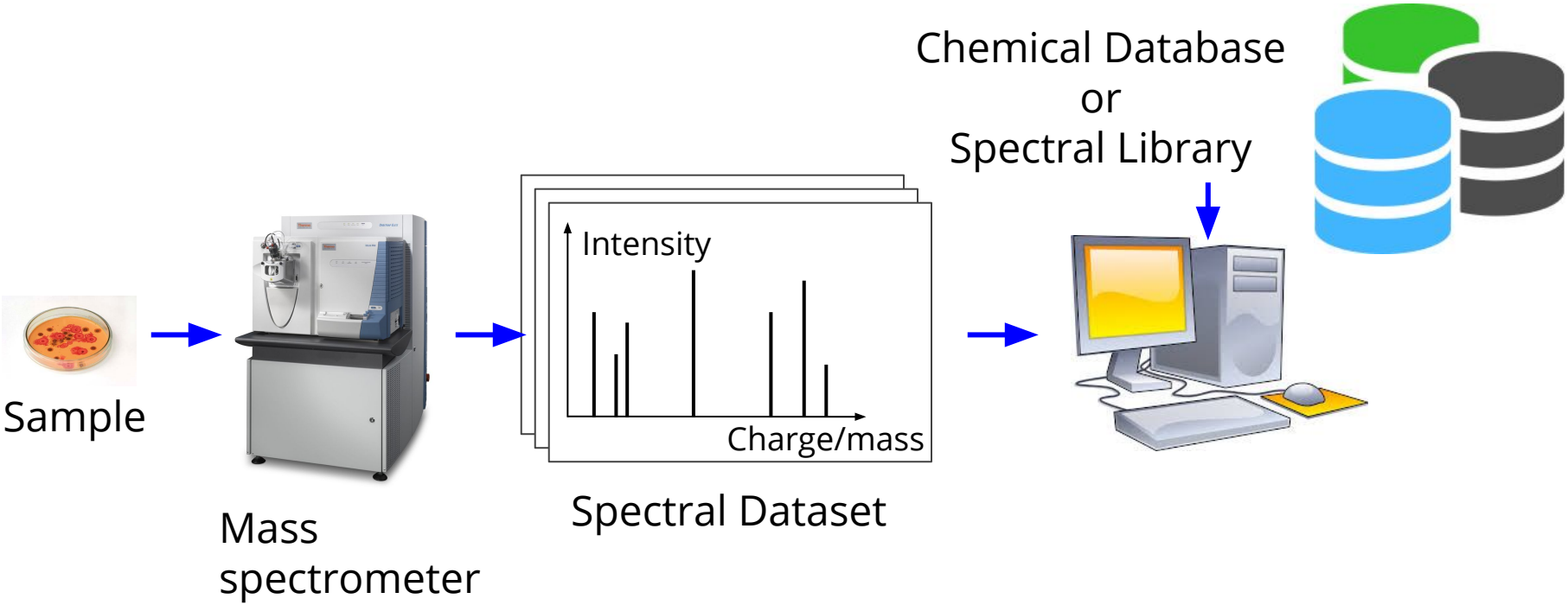


Sample

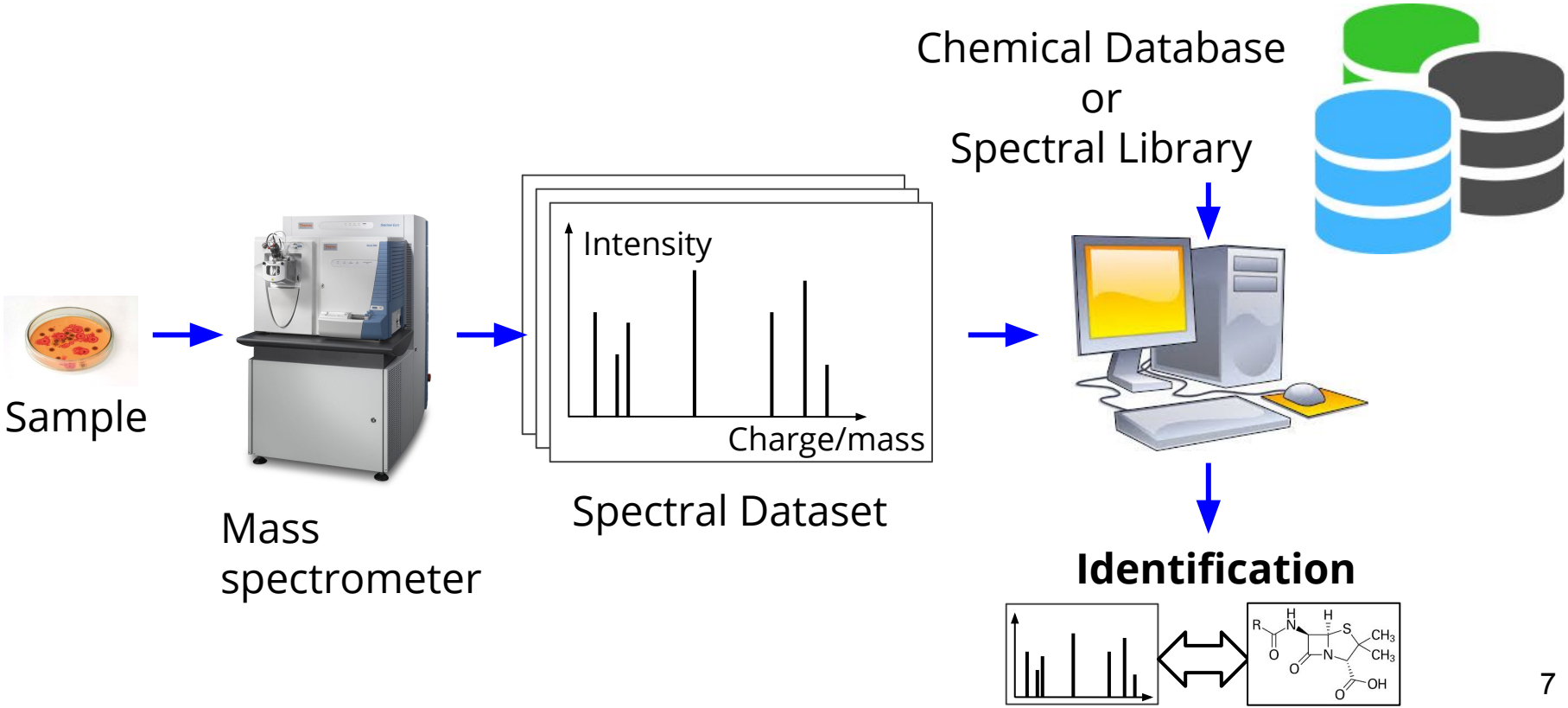
Mass spectrometer

Spectral Dataset

Computational MS/MS basics



Computational MS/MS basics



Computational MS/MS basics

[Almost] solved in traditional proteomics (linear peptides):

- **reference-based** identification
  - spectral libraries: SpectraST (Lam et al, 2007), Tremolo (Wang and Bandeira, 2014), etc
  - chemical databases: Mascot (Perkins et al, 1999), MS-GF+ (Kim and Pevzner, 2014), etc

Computational MS/MS basics

[Almost] solved in traditional proteomics (linear peptides):

- **reference-based** identification
  - spectral libraries: SpectraST (Lam et al, 2007), Tremolo (Wang and Bandeira, 2014), etc
  - chemical databases: Mascot (Perkins et al, 1999), MS-GF+ (Kim and Pevzner, 2014), etc
- ***de novo*** peptide sequencing: Pepnovo (Frank and Pevzner, 2005), NovoHMM (Fischer et al, 2005), etc

Challenges of PNP identification

Traditional proteomics tools are **not applicable** to PNPs due to:

Challenges of PNP identification

Traditional proteomics tools are **not applicable** to PNPs due to:

- nonlinear structure (cyclic, branch-cyclic, etc)

Challenges of PNP identification

Traditional proteomics tools are **not applicable** to PNPs due to:

- nonlinear structure (cyclic, branch-cyclic, etc)
- presence of many post-translational modifications and nonstandard amino acids

Challenges of PNP identification

Traditional proteomics tools are **not applicable** to PNPs due to:

- nonlinear structure (cyclic, branch-cyclic, etc)
- presence of many post-translational modifications and nonstandard amino acids
- small size of PNP spectral libraries

Challenges of PNP identification

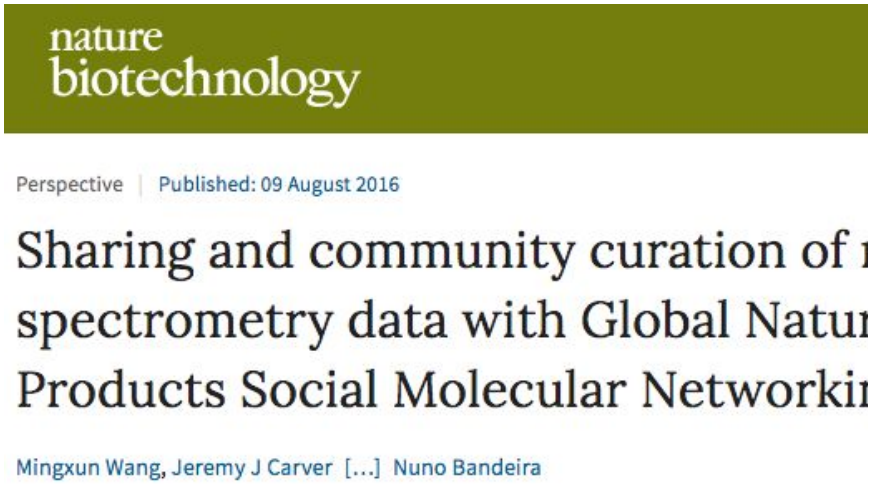
Traditional proteomics tools are **not applicable** to PNPs due to:

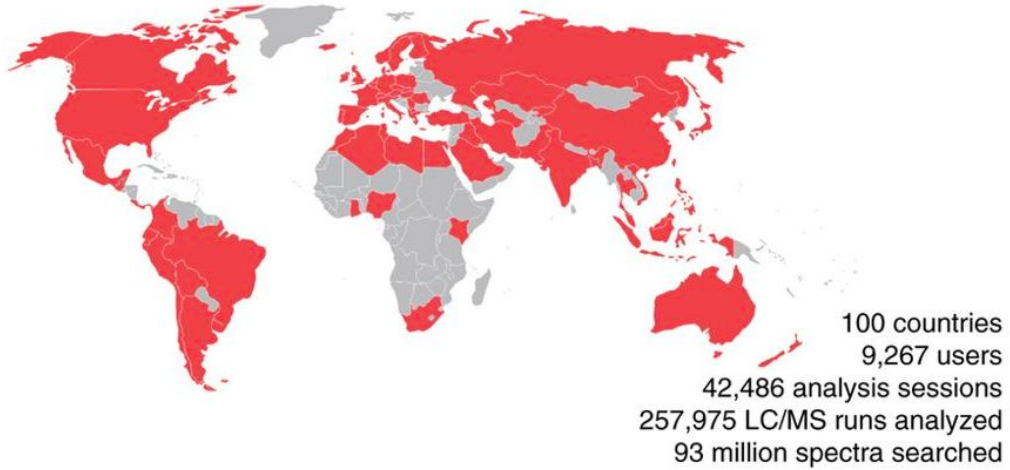
- nonlinear structure (cyclic, branch-cyclic, etc)
- presence of many post-translational modifications and nonstandard amino acids
- small size of PNP spectral libraries

Until recently there were no tools for MS-based high-throughput PNP identification

PNP identification tools

- **GNPS** (Wang et al, 2016)
  platform for sharing and analysing natural products MS/MS data (includes NP spectral library)



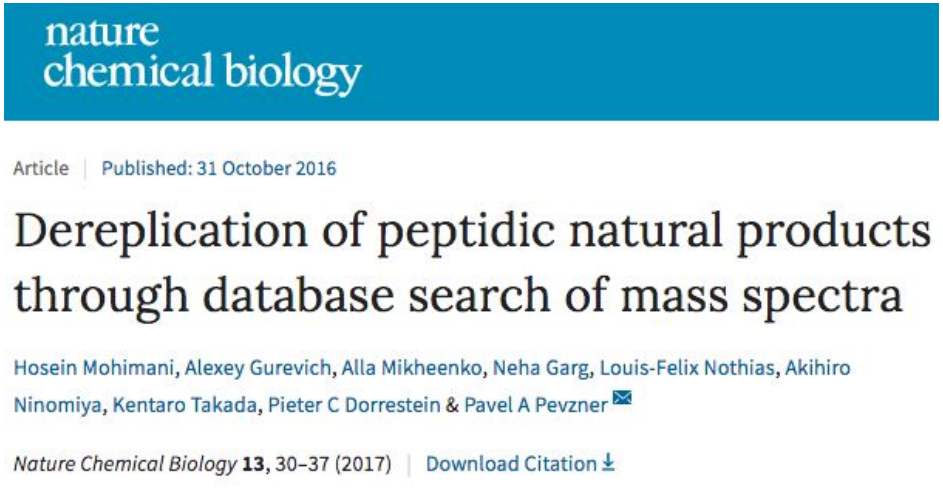


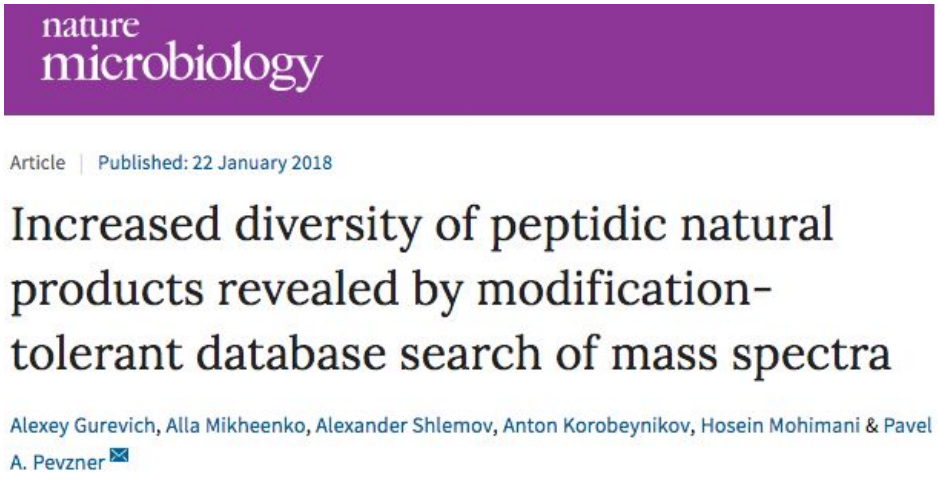
PNP identification tools

- **Dereplicator** (Mohimani et al, 2017)
  in silico Dereplication of PNP via database search

PNP identification tools

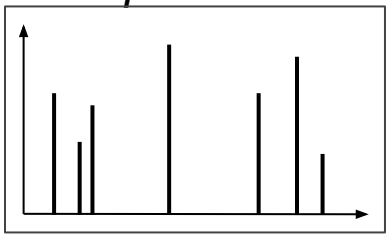
- **Dereplicator** (Mohimani et al, 2017)
  in silico Dereplication of PNP via database search
- **VarQuest** (Gurevich et al, 2018)
  variable dereplication



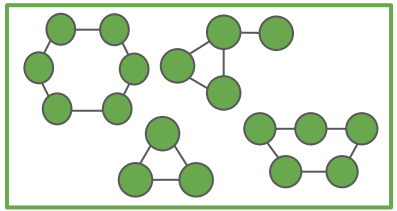


Dereplicator pipeline

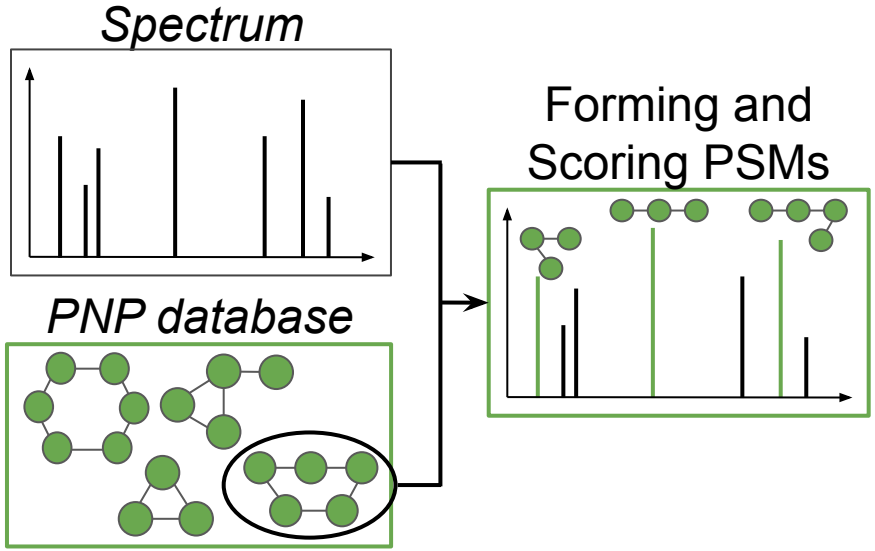
Spectrum



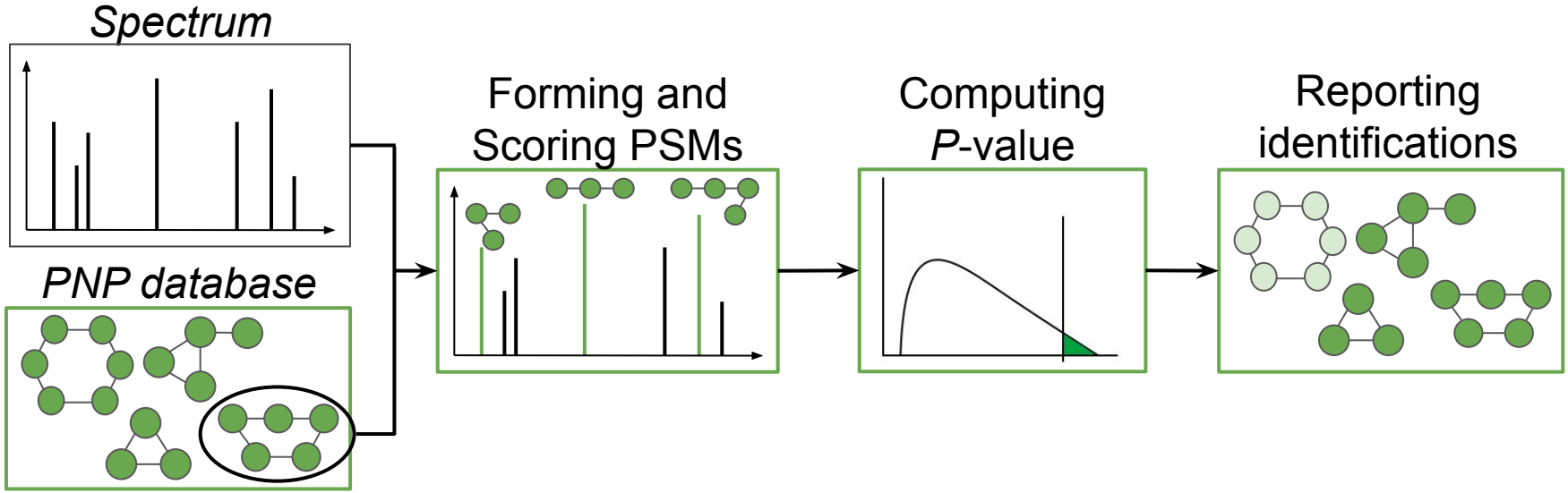
PNP database



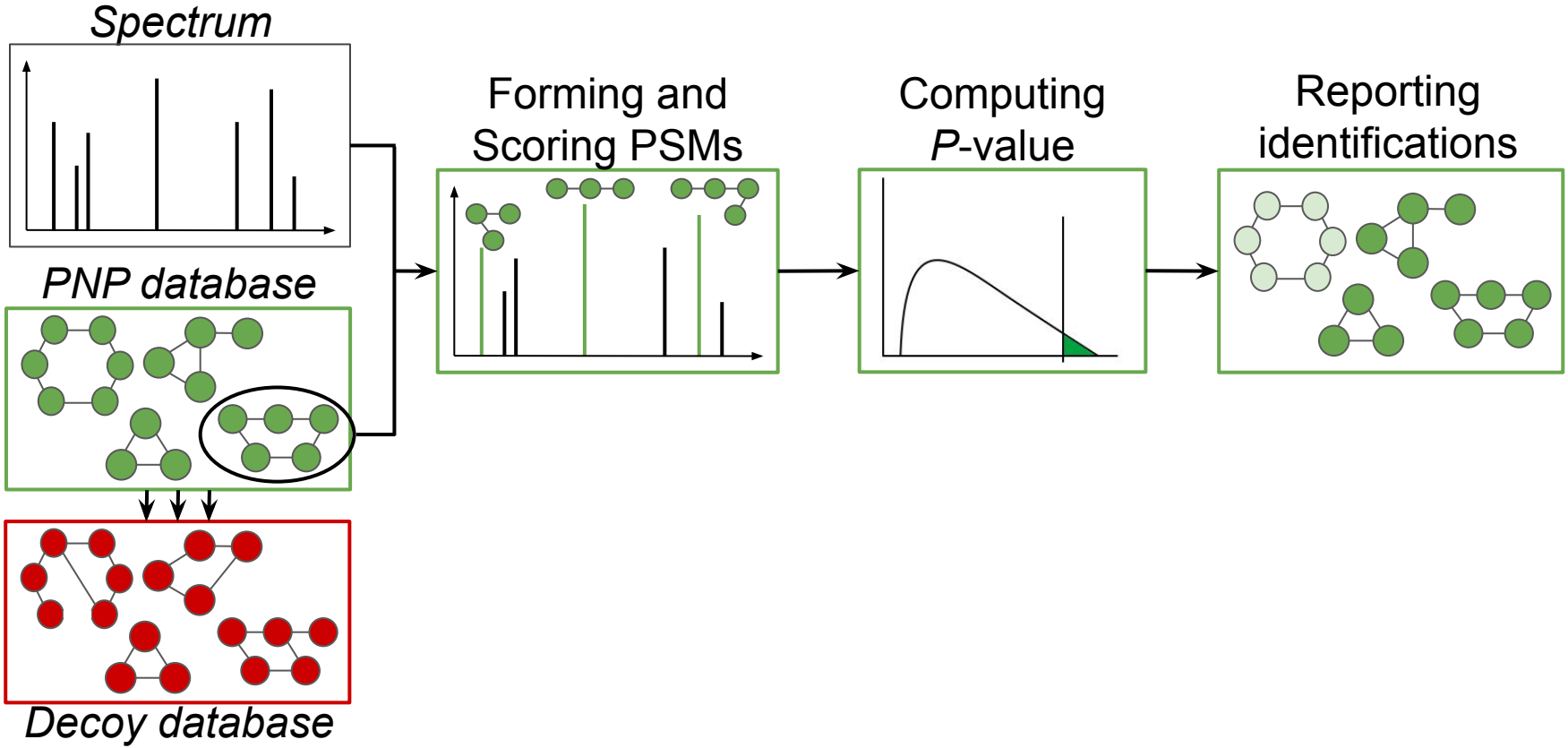
Dereplicator pipeline



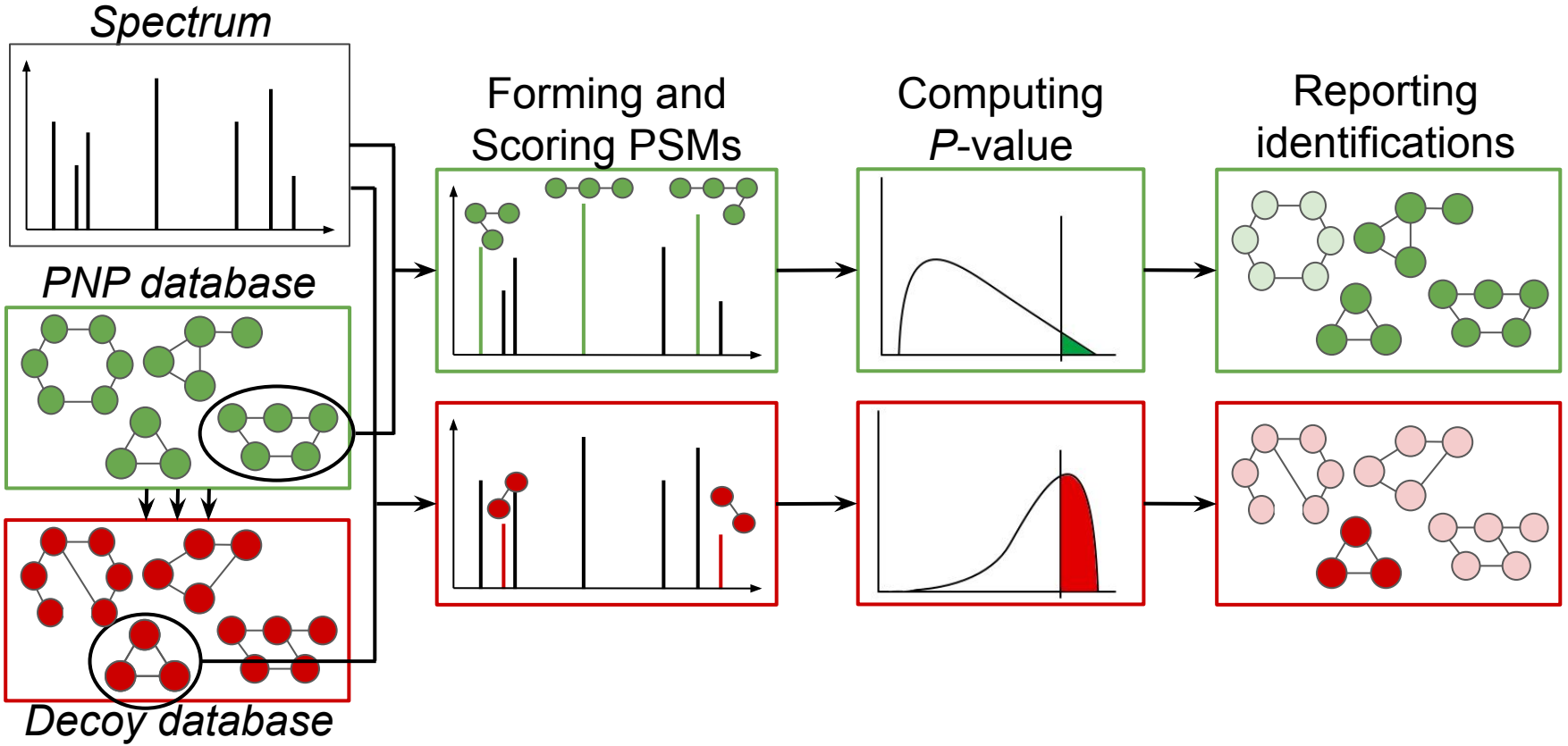
Dereplicator pipeline



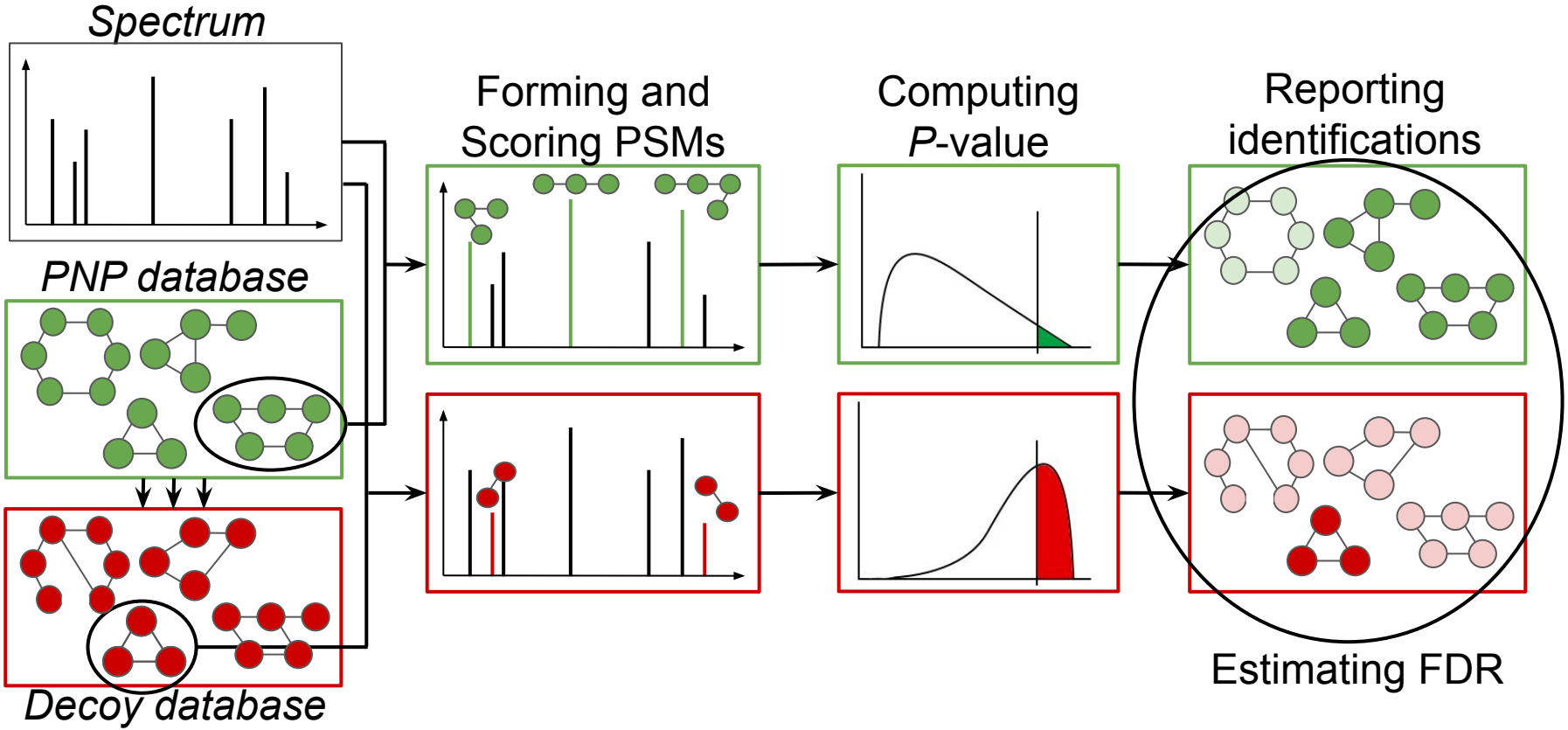
Dereplicator pipeline



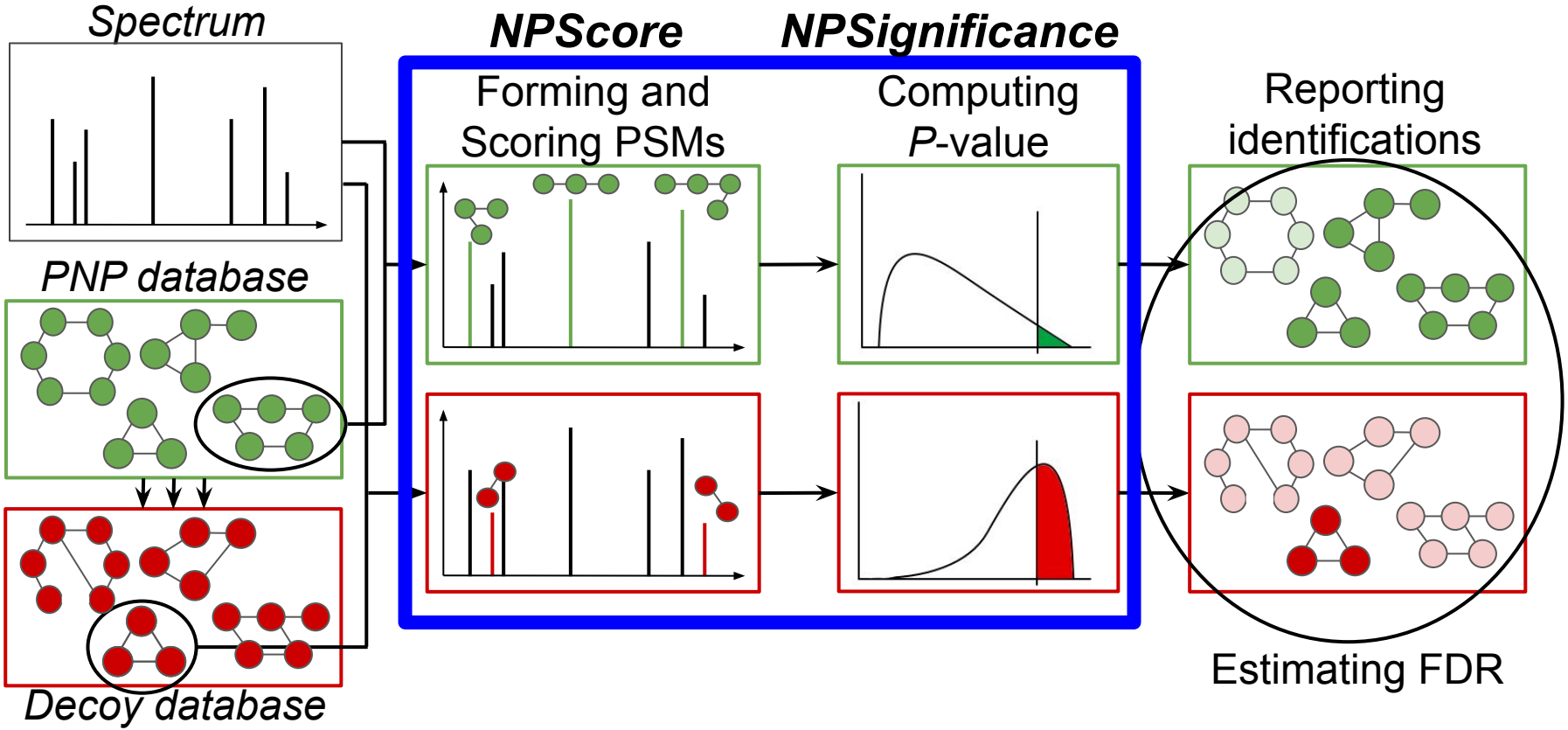
Dereplicator pipeline



Dereplicator pipeline

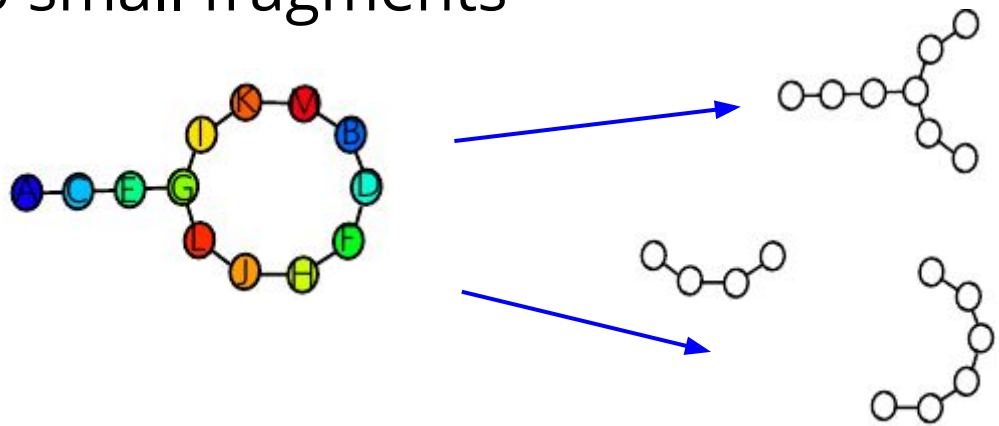


Dereplicator pipeline



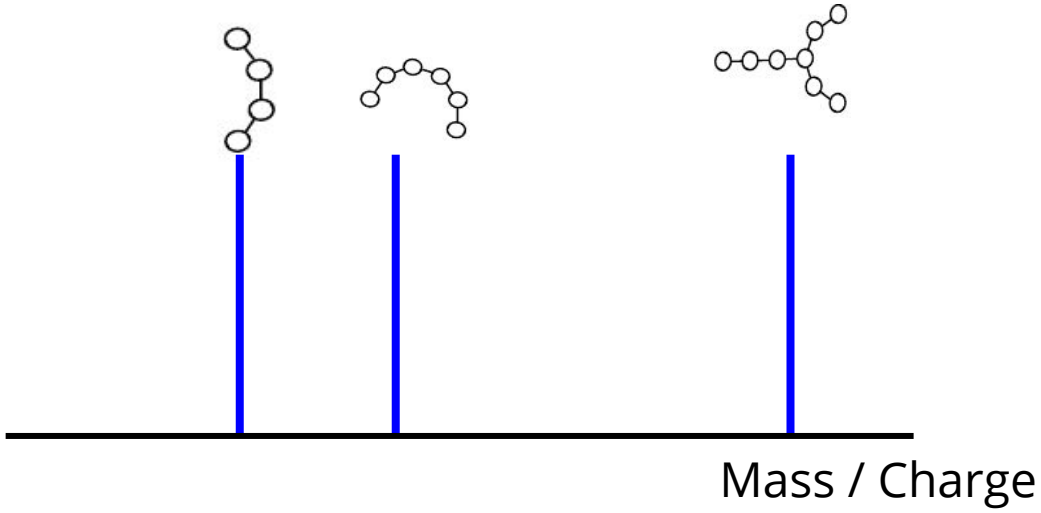
Theoretical spectrum

After collision in mass spectrometer a molecule is split into small fragments

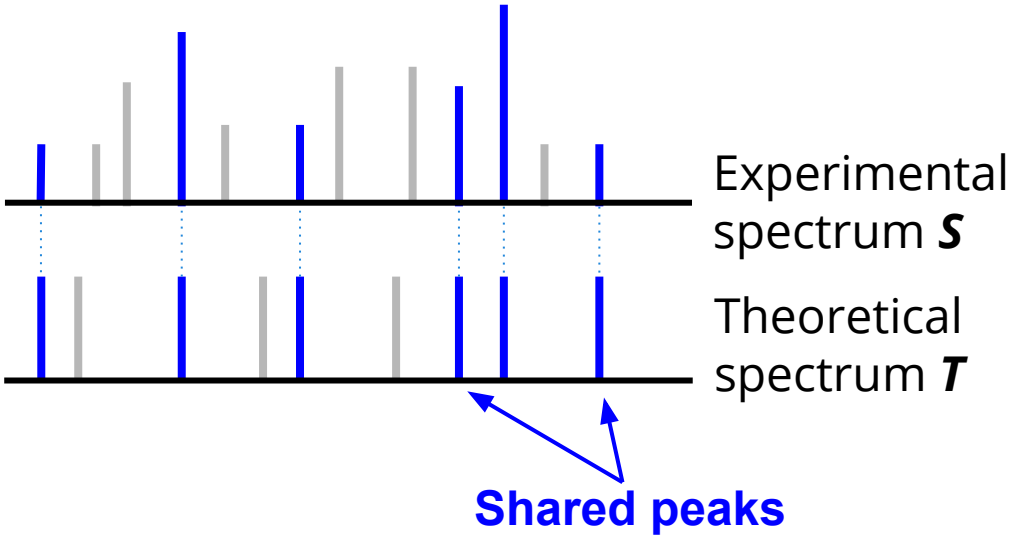


Theoretical spectrum

Given a molecule structure, we model this process obtaining molecule's ***Theoretical Spectrum***

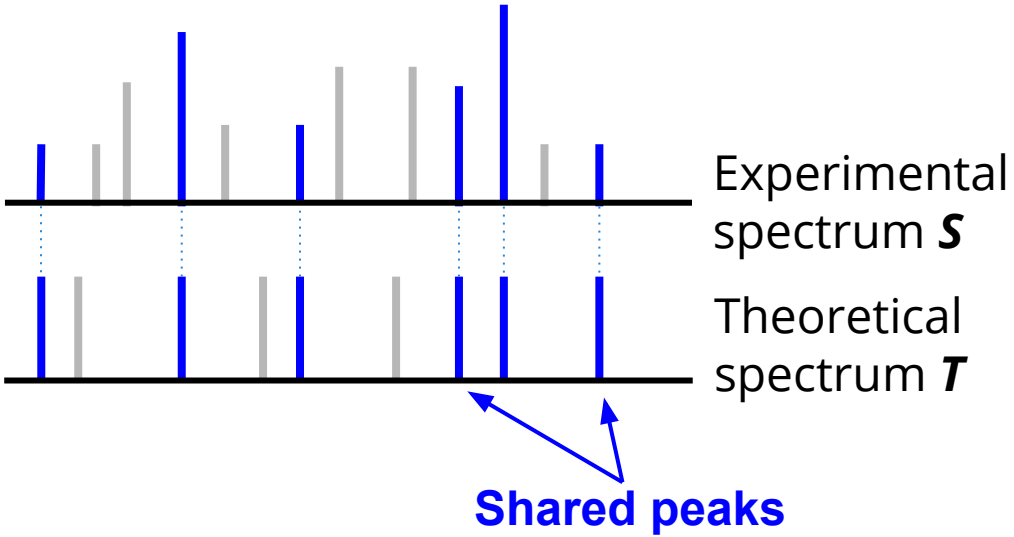


Scoring PSMs: baseline



Baseline scoring: $\mathrm{SPCScore}(S, T) = \#\{\text{Shared peaks}\}$

Scoring PSMs: baseline



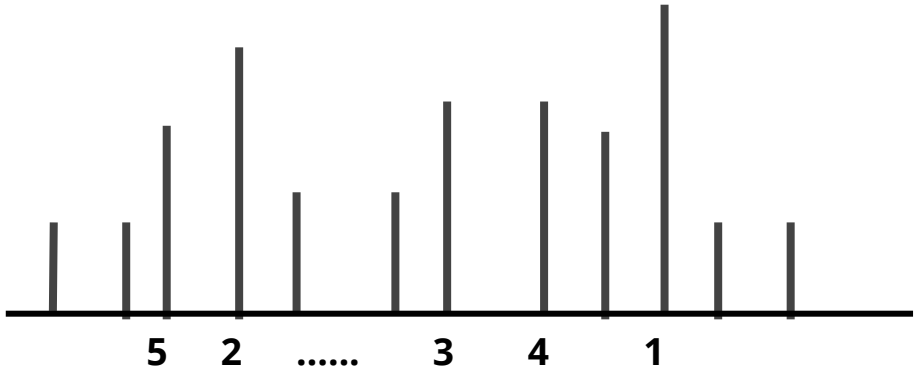
Limitations:

- Same score for high- and low-intensity peaks
- Very primitive fragmentation model (e.g. neutral losses are not considered)

Baseline scoring: $\mathrm{SPCScore}(S, T) = \#\{\text{Shared peaks}\}$

NPScore: basic ideas

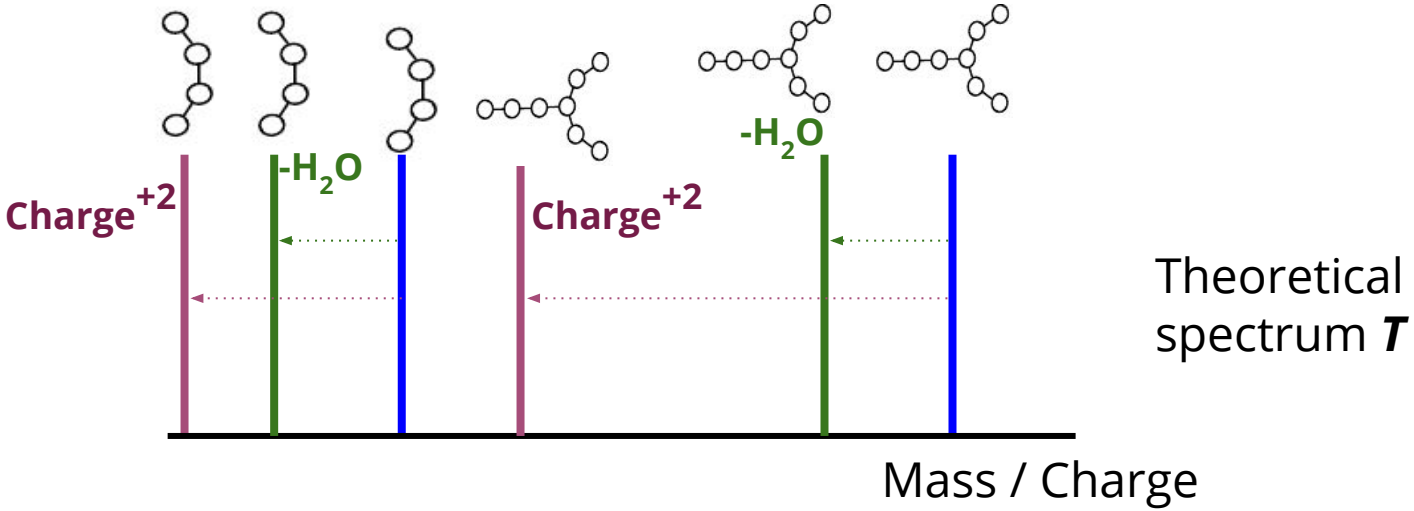
- Considers peak intensities



Experimental spectrum ***S***

NPScore: basic ideas

- Considers peak intensities
- Considers additional ion types



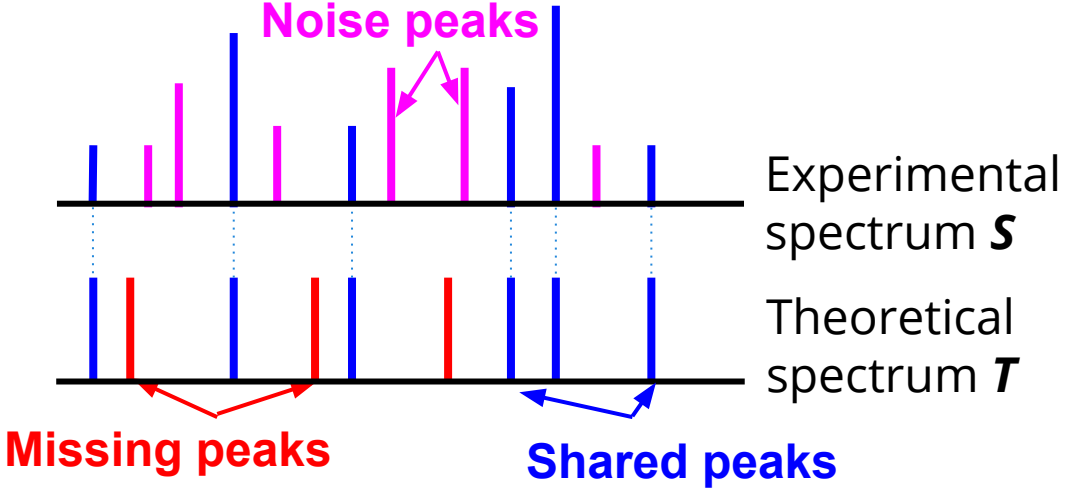
NPScore: basic ideas

- Considers peak intensities
- Considers additional ion types
- Has individual scoring weights for each (rank, ion type) pair

rank \ ion type	primary	isotopic shift	H_2O loss	…
1	$w_{1,1}$	$w_{1,2}$	…	
2	$w_{2,1}$	…		
…	…			

NPScore: basic ideas

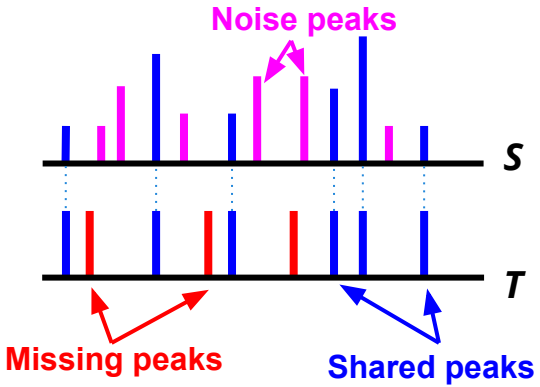
- Considers peak intensities
- Considers additional ion types
- Has individual scoring weights for each (rank, ion type) pair
- Considers both shared and missing peaks



NPScore: algorithm

Scoring is based on the following generative model:
(Dančík et al, 1999; Frank & Pevzner, 2005;...):

All peaks are generated independently, and



NPScore: algorithm

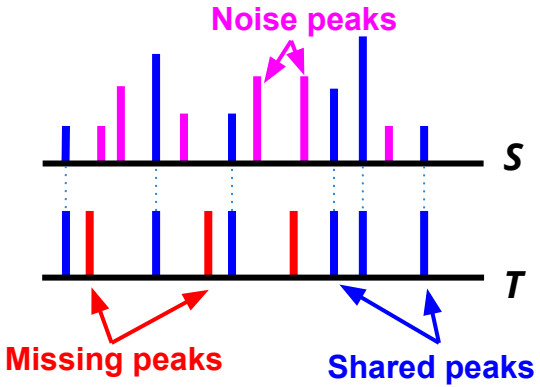
Scoring is based on the following generative model:
(Dančík et al, 1999; Frank & Pevzner, 2005;...):

All peaks are generated independently, and

T-Peak generates **S-Peak** with **Prob(*rank* | *ion*)**

T-Peak is missing with **Prob(*0* | *ion*)**

Noise generates **S-Peak** with **Prob(*rank* | NULL)**



NPScore: algorithm

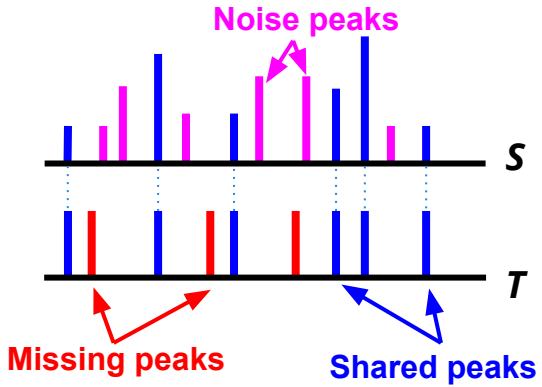
Scoring is based on the following generative model:
(Dančík et al, 1999; Frank & Pevzner, 2005;...):

All peaks are generated independently, and

T-Peak generates **S-Peak** with **Prob(*rank* | *ion*)**

T-Peak is missing with **Prob(*0* | *ion*)**

Noise generates **S-Peak** with **Prob(*rank* | NULL)**



$$\mathrm{NPScore}(S,T) = \log \frac{\mathrm{Prob}(S|T)}{\mathrm{Prob}(S|\varnothing)}$$

NPScore: algorithm

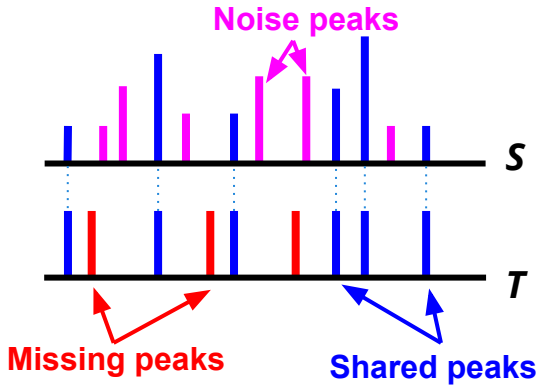
Scoring is based on the following generative model:
(Dančík et al, 1999; Frank & Pevzner, 2005;...):

All peaks are generated independently, and

T-Peak generates **S-Peak** with **Prob(*rank* | *ion*)**

T-Peak is missing with **Prob(*0* | *ion*)**

Noise generates **S-Peak** with **Prob(*rank* | NULL)**



$$\text{NPScore}(S,T) = \log \frac{\text{Prob}(S|T)}{\text{Prob}(S|\varnothing)}$$

$$= \sum_{(rank,ion)\ \in\ \text{Shared}(S,T)} \log \frac{\text{Prob}(rank|ion)}{\text{Prob}(rank|NULL)} + \sum_{ion\ \in\ \text{Missing}(S,T)} \log \frac{\text{Prob}(0|ion)}{\text{Prob}(0|NULL)}$$

NPScore parameters

We obtain scoring parameters with statistical learning:

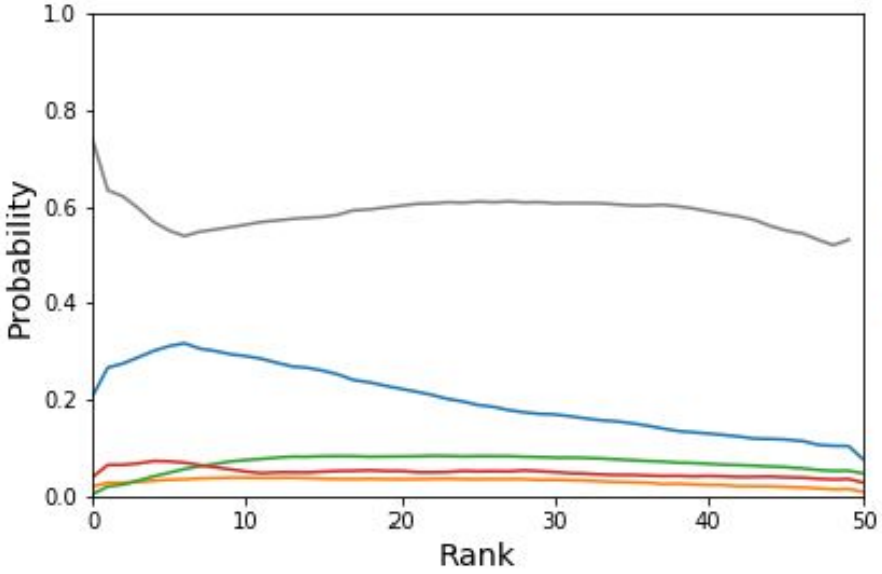
- Form a training dataset of high quality PSMs (e.g., by selecting highly confident Dereplicator matches)

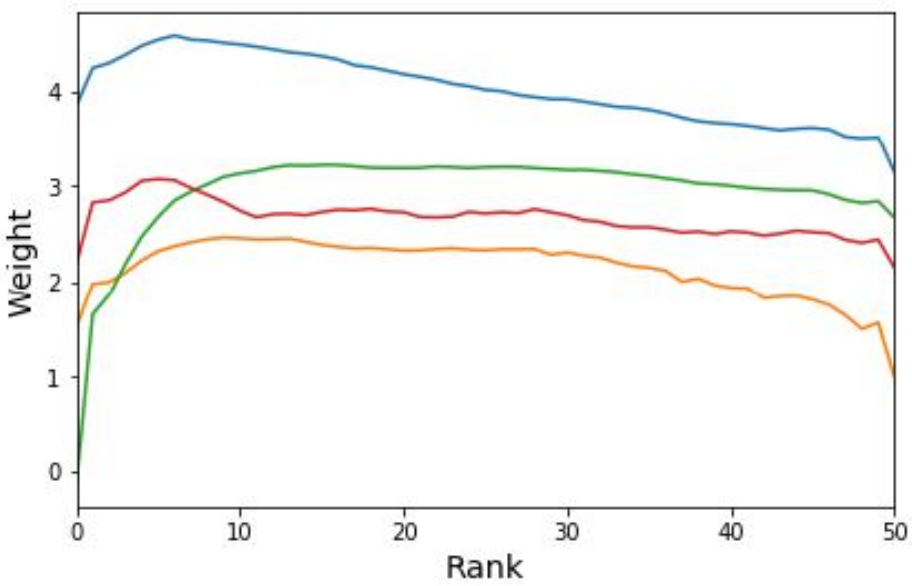
NPScore parameters

We obtain scoring parameters with statistical learning:

- Form a training dataset of high quality PSMs (e.g., by selecting highly confident Dereplicator matches)
- Statistically learn probabilities **Prob(*rank* | *ion*)** as frequencies of corresponding events (i.e., the ratio of peaks ranked ***rank*** and explained by a ion type ***ion*** to the total number of peaks ranked ***rank*** in the train dataset)

NPScore parameters





Learned probabilities

Resulting weights

Results

Benchmarking on linear dataset

Dataset: A draft map of the human proteome (MSV000079514; Kim et al., 2014),

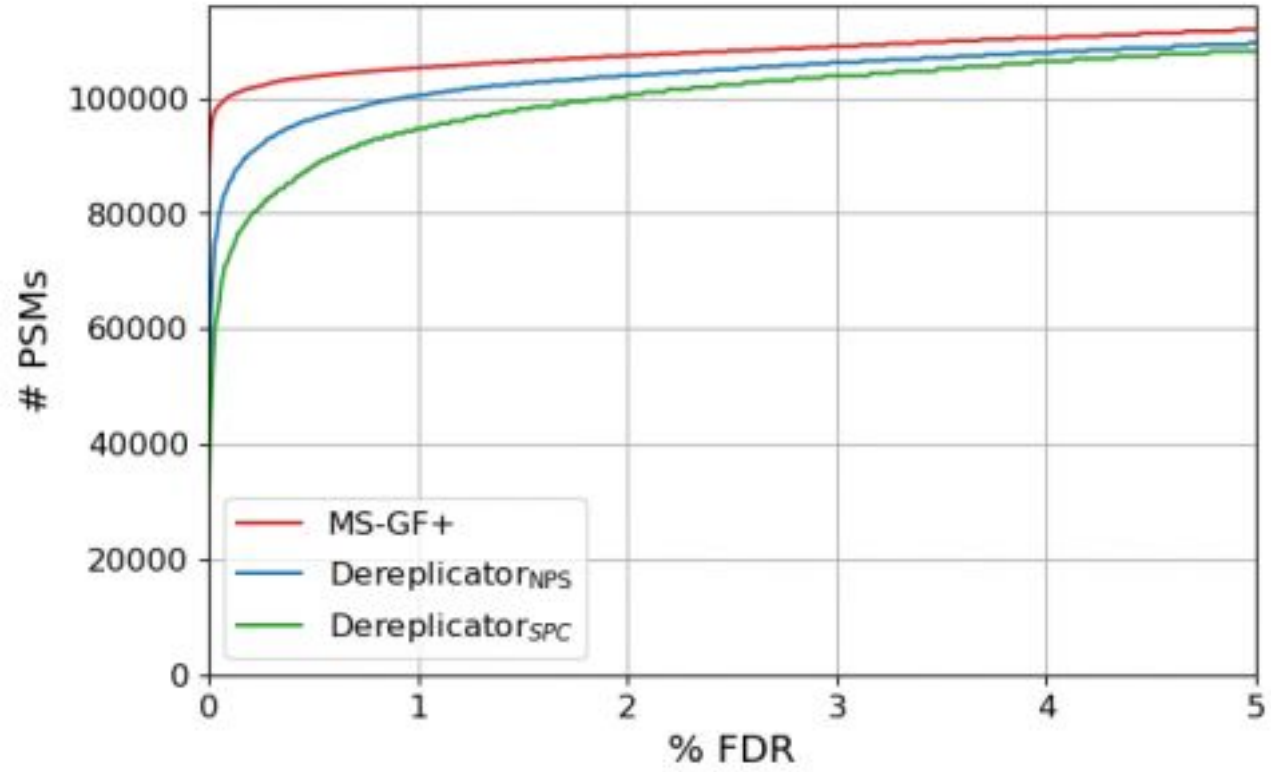
426k spectra (Kidney tissues only)

Peptide database: RefSeq Proteins, trypsin/P, no missed cleavages (**47k peptides**)

Instrument: LTQ Orbitrap Velos (HCD)

Train data: MS-GF+ identifications at FDR level 0% (17,794 PSMs from Heart tissues from MSV000079514)

Benchmarking on linear dataset



Benchmarking on PNP dataset

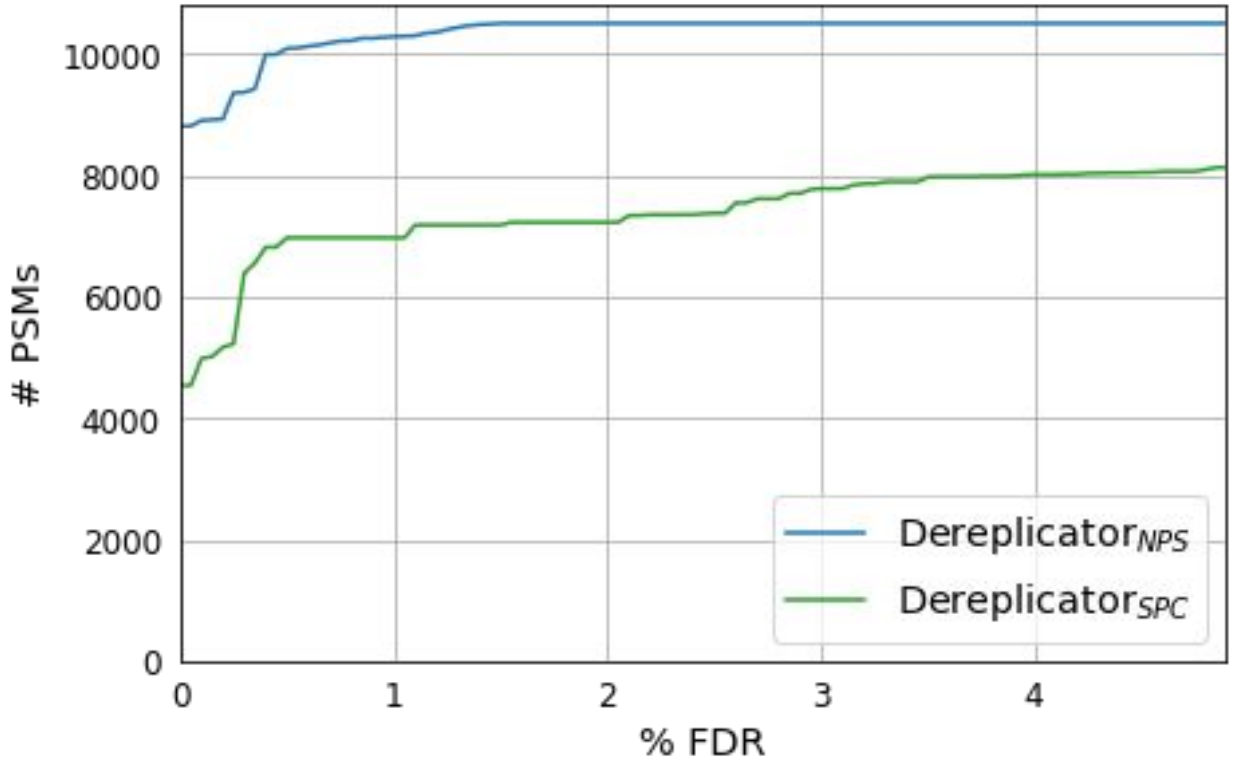
Dataset: GNPS spectral datasets (Wang et al., 2016)
16M spectra (13 high-resolution datasets)

Peptide database: *PNPdatabase* from AntiMarin, DNP, MiBIG, StreptomeDB (Gurevich et al, 2018)
(**5,021 PNPs**)

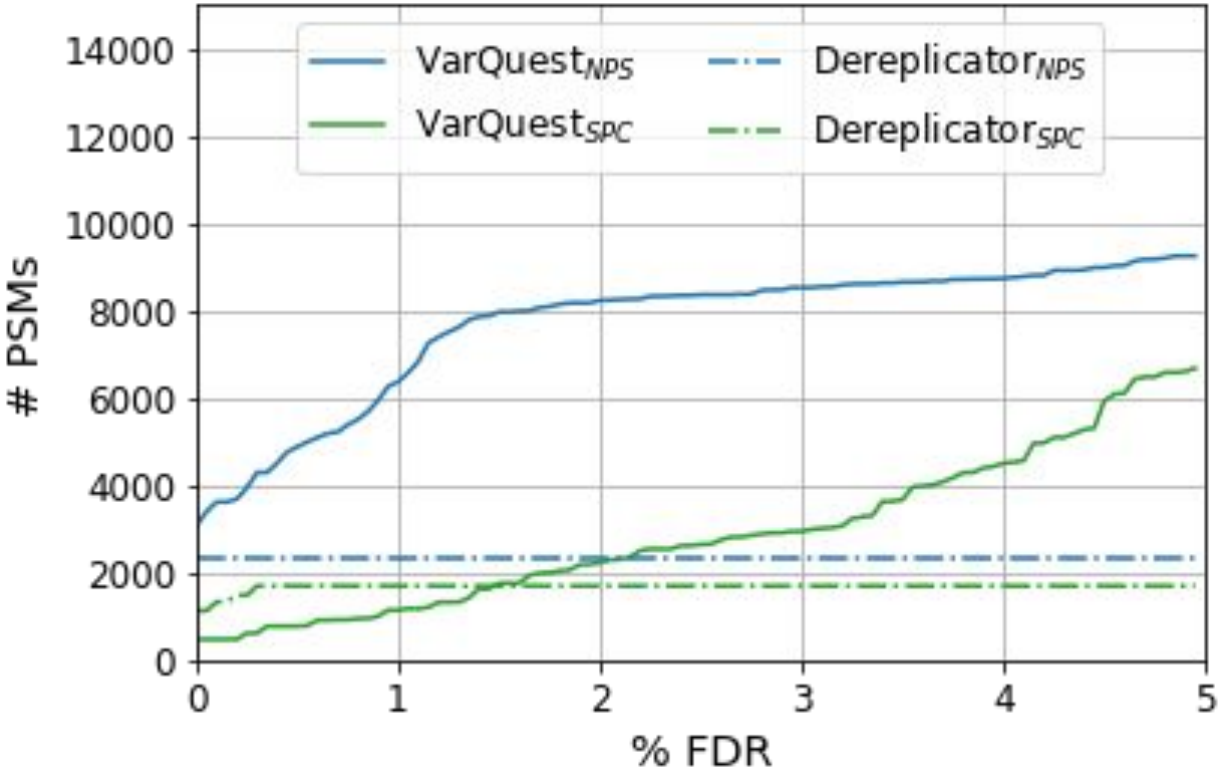
Instruments: Bruker micrOTOF-QII, Bruker maXis, LTQ-Orbitrap Velos, Agilent Q-TOF LC/MS

Train data: Dereplicator identifications at FDR level 0%
(14,757 PSMs from 120 GNPS datasets)

Benchmarking on PNP dataset



Benchmarking on PNP dataset: VarQuest



* Only 3 datasets:
1 Pseudomonas
2 Streptomyces
Total: 1.1M spectra

Acknowledgements


Azat Tagirdzhanov


Alexander Shlemov


Hosein Mohimani

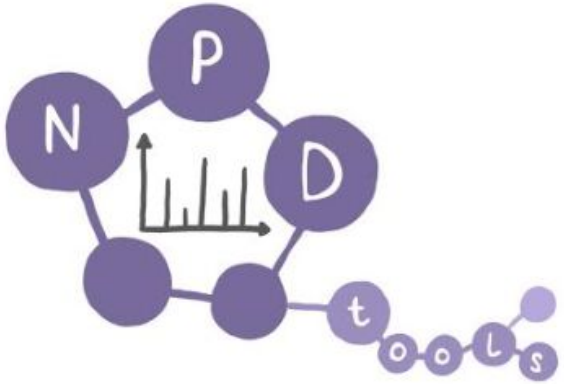

Anton Korobeynikov


Pavel Pevzner

Funding: Russian Science Foundation (grant #19-16-00049)

Thank you!

Questions?



Web: http://cab.spbu.ru/software/nps/

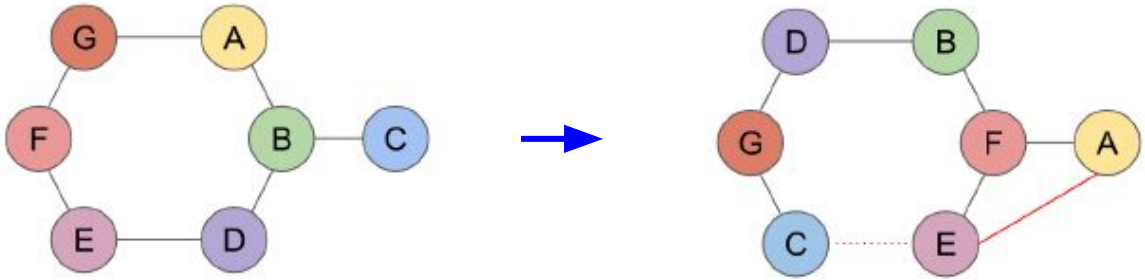
Email: npdtools.support@cab.spbu.ru

Paper: Tagirdzhanov *et al.*, *Bioinformatics*, <in press>, 2019

Supplementary slides

Decoy models

- **Linear peptides**
  Peptides from reversed protein sequencies
- **PNPs**
  AA shuffling + edge replacement



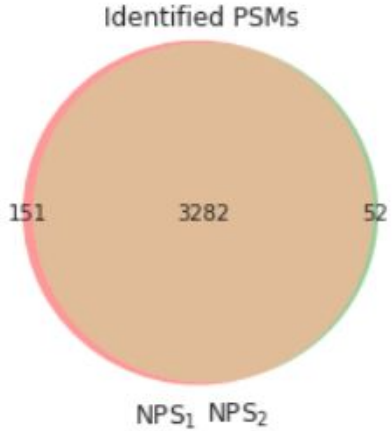
(Gurevich et al, 2018)

In case of overfitting concern

We randomly formed subsets $Train_{1,2}$ and Test:

- Do not contain spectra from the same spectral dataset
- Do not contain same compounds

Resulting identifications:



The table

Methods	Number of PSMs			Number of Peptides		
	P^{-10}	FDR 0%	FDR 1%	P^{-10}	FDR 0%	FDR 1%
Proteomics dataset						
MS-GF+	N/A	87,223	105,155	N/A	14,534	17,081
SPC	102,951	32,455	94,663	17,723	7,107	15,968
NPS	115,044	46,252	100,444	22,081	9,414	16,570
PNP datasets						
SPC	8,544	4,538	6,972	351	231	304
NPS	10,504	8,811	10,287	395	290	378