Seminar

Reporter: Lichuang Huang

Supervisor: Gang Cao

2022-03-28

part 1: literature sharing

Part 2: MCnebula

Where is a 'Lazy' method ?

What should a lazy analysis be?

part 3: pharmacology

END

part 1: literature sharing

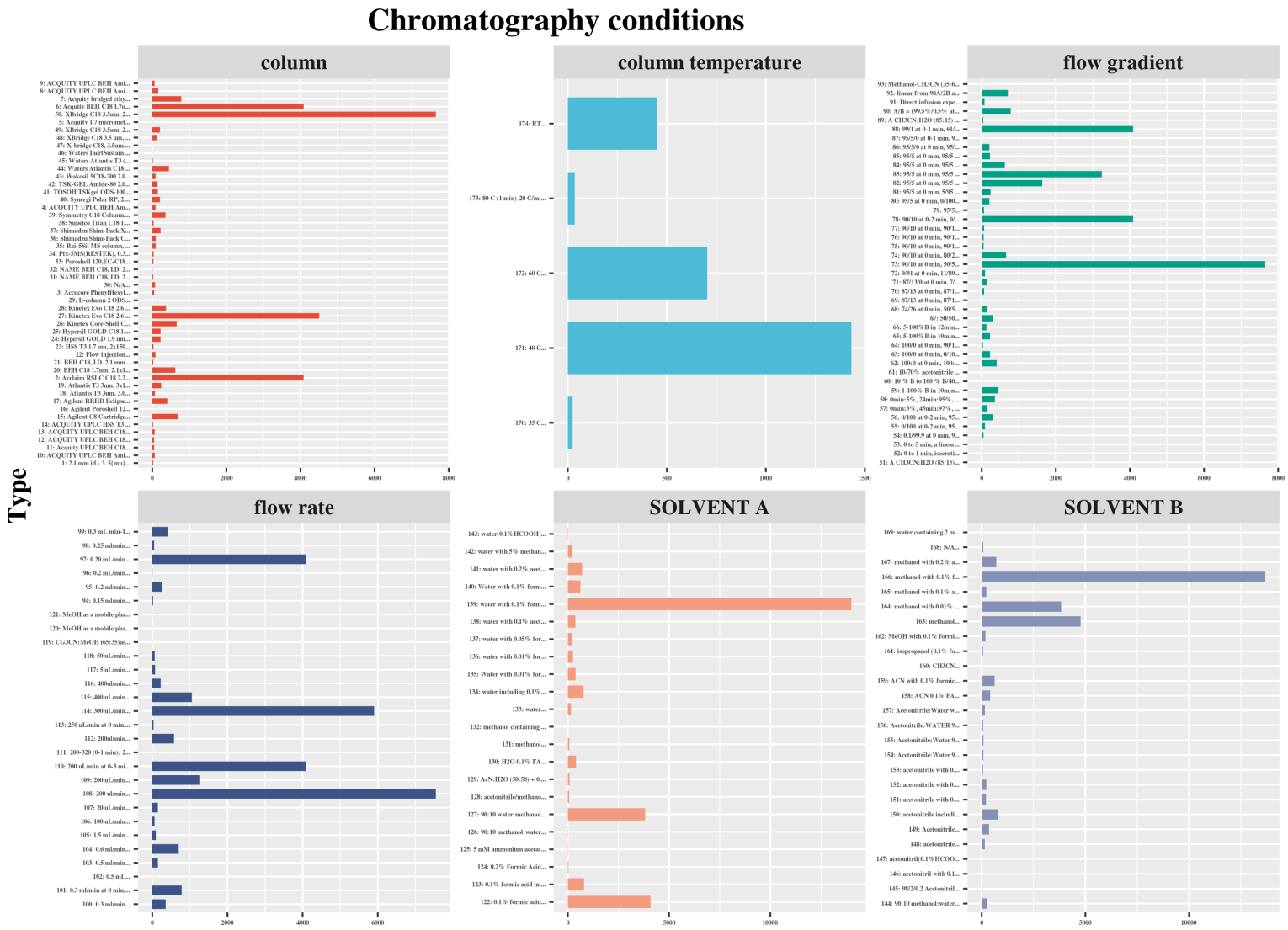
LC-MS: a multi-dimentional data analysis method

- Main dimention:
  - MS level 1
  - MS level 2

AS MS spectrum fail in ascertaining the molecular scaffold (geometrical or position isomers)

- Other dimention
  - retention time
  - CCS
  - ...

A variety range of chromatography condition



Retip: A machine learning method of LC-MS retention time prediction

Published in final edited form as:
Anal Chem. 2020 June 02; 92(11): 7515–7522. doi:10.1021/acs.analchem.9b05765.

Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics

Figure 2: retip

Retip: A machine learning method of LC-MS retention time prediction

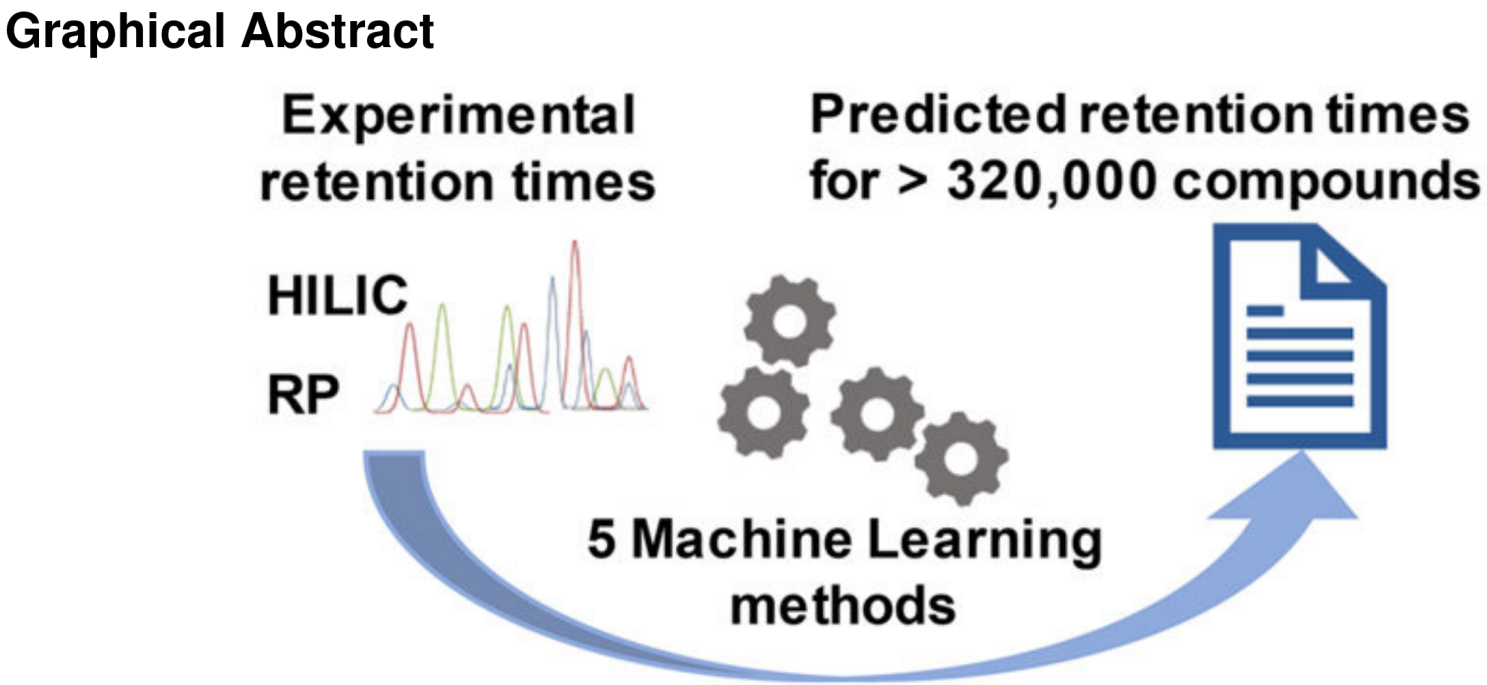


Figure 3: retip abstract

Retip: Mean absolute errors

Mean Absolute Errors (min) for RT Predictions

	HILIC			reversed phase LC		
	training	**test**	**validation**	**training**	**test**	**validation**
XGBoost	0.38	1.02	0.64	0.25	0.48	0.68
BRNN	0.37	1.05	0.60	0.43	0.51	0.76
RF	0.85	1.11	0.68	0.23	0.51	0.75
LightGBM	0.39	0.99	0.86	0.12	0.49	0.72
Keras	0.70	0.78	0.82	0.50	0.57	0.80

Figure 4: retip prediction error

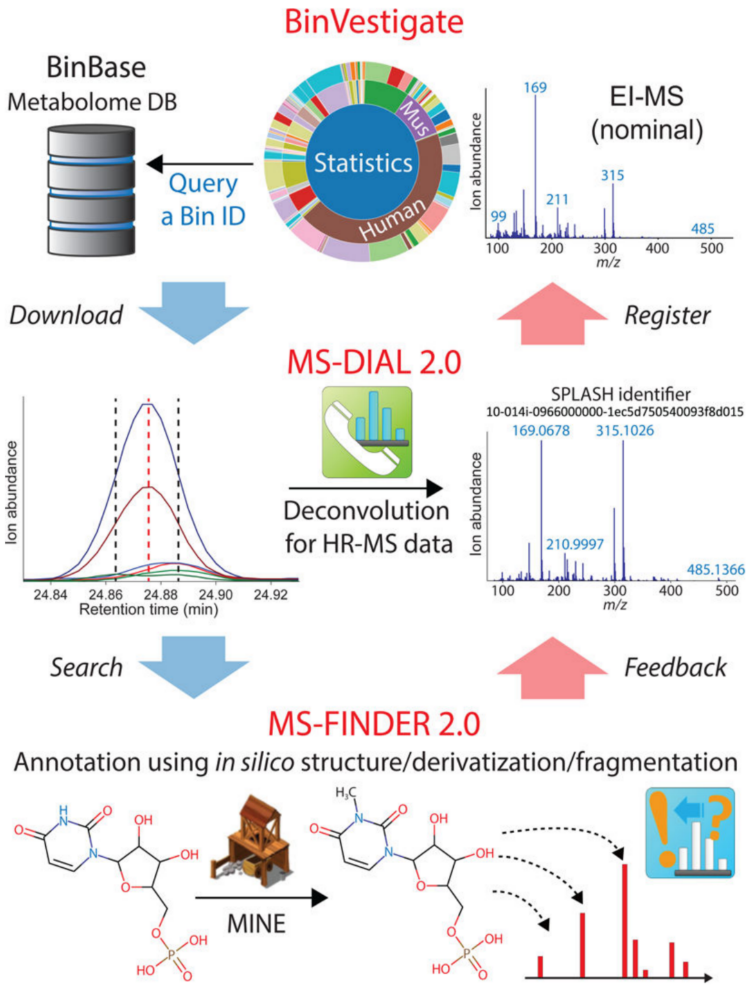
Retip: shrink the scope of structure library

Reduction in Candidate Chemical Search Space for Blood Plasma Experimental MS/MS Spectra[1]

	without RT filtering		with ±1 min RT filtering	
no. of isomers	no. of formulas	no. of isomers	no. of formulas	no. of isomers
1	580	580	300	300
2	281	562	152	304
3	241	723	93	279
4	155	620	48	192
5	102	510	21	105
6	87	522	32	192
7	87	609	14	98
8	47	376	9	72
9	37	333	9	81
10	385	3850	125	1250
total	2002	8685	803	2873

Figure 5: retip application

Retip: integrate in MSfinder



Retip: R package

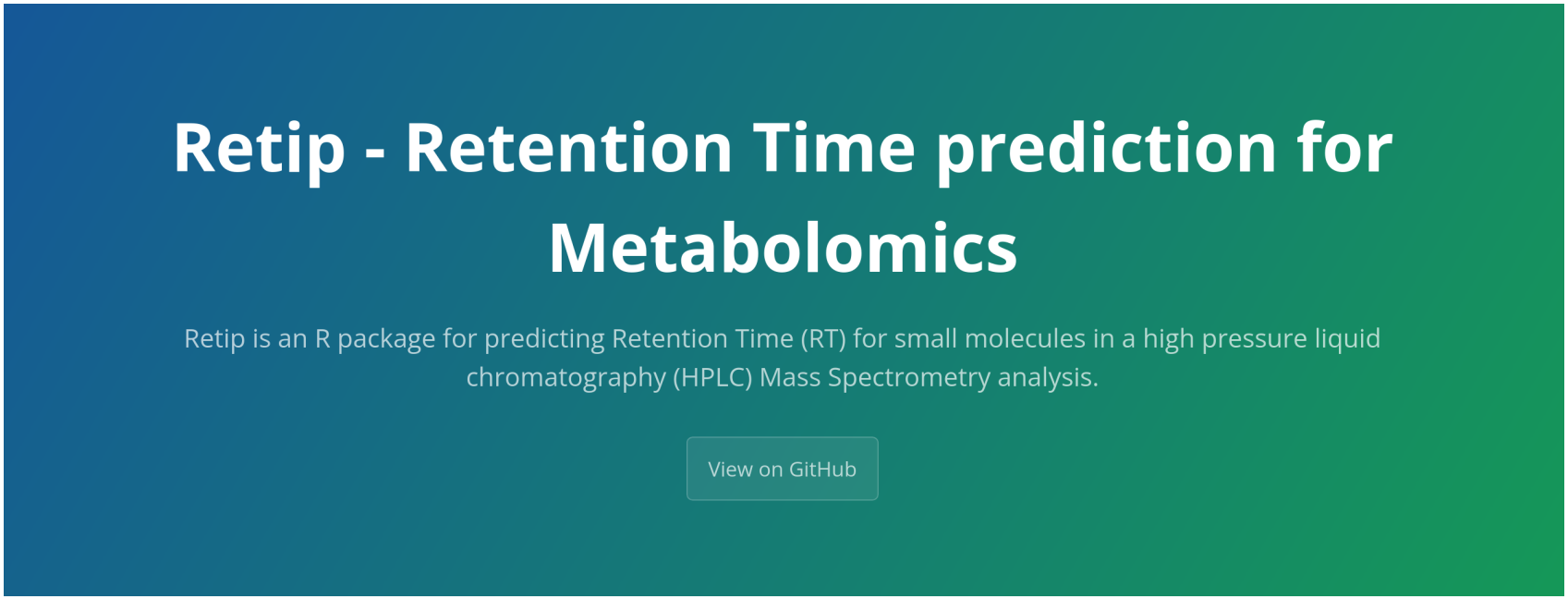


Figure 7: retip in R

Retip: workflow

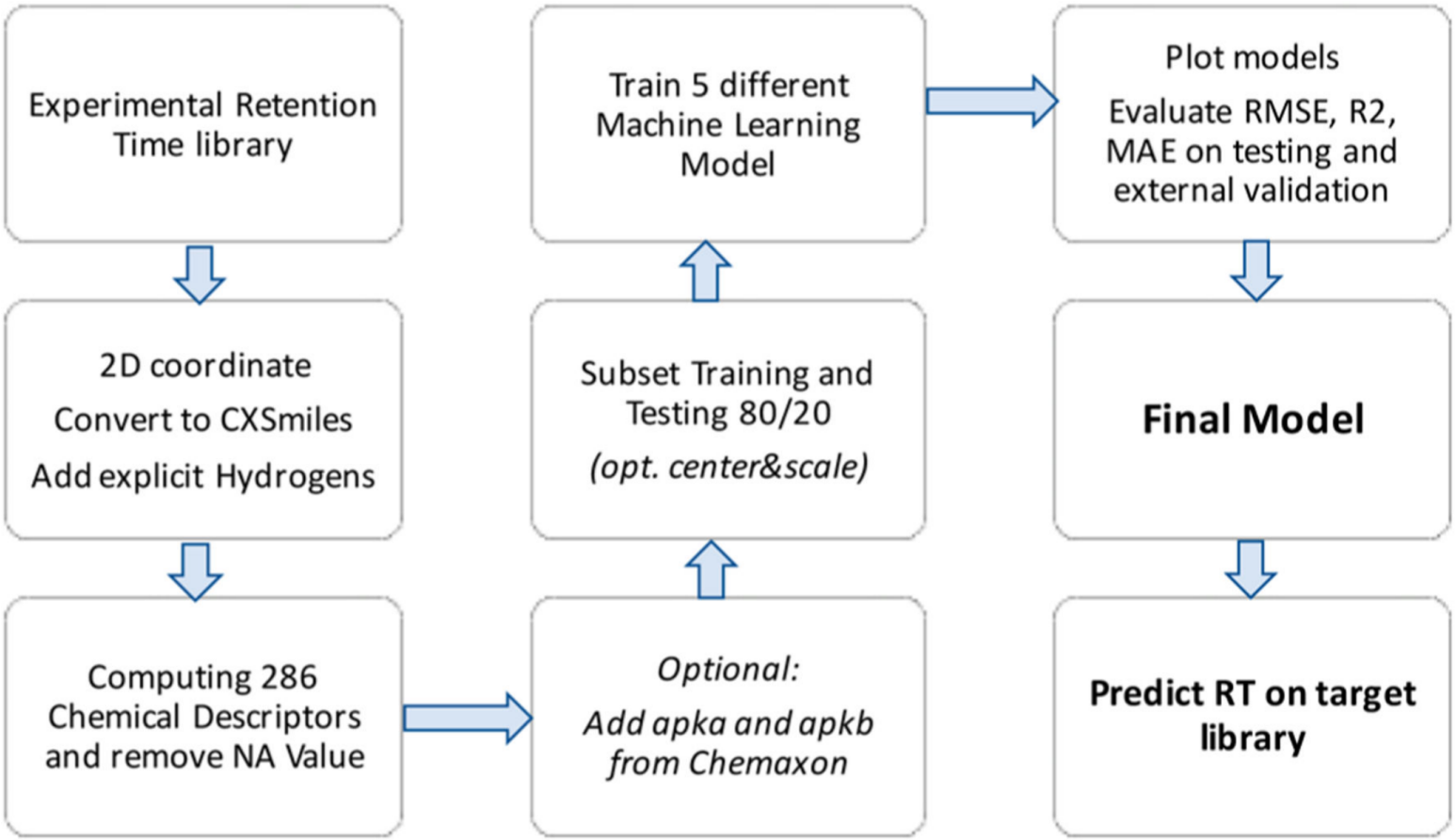


Figure 8: retip workflow

Summary

1. An awsome method for utilizing RT and filter structure candidates
   - computional
   - conversion available for any chromatography condition
   - public available R package
2. A method possibly be integrated in MCnebula

Part 2: MCnebula

Research background

Untarget LC-MS analysis

Researcher have to check structure and spectrum one by one

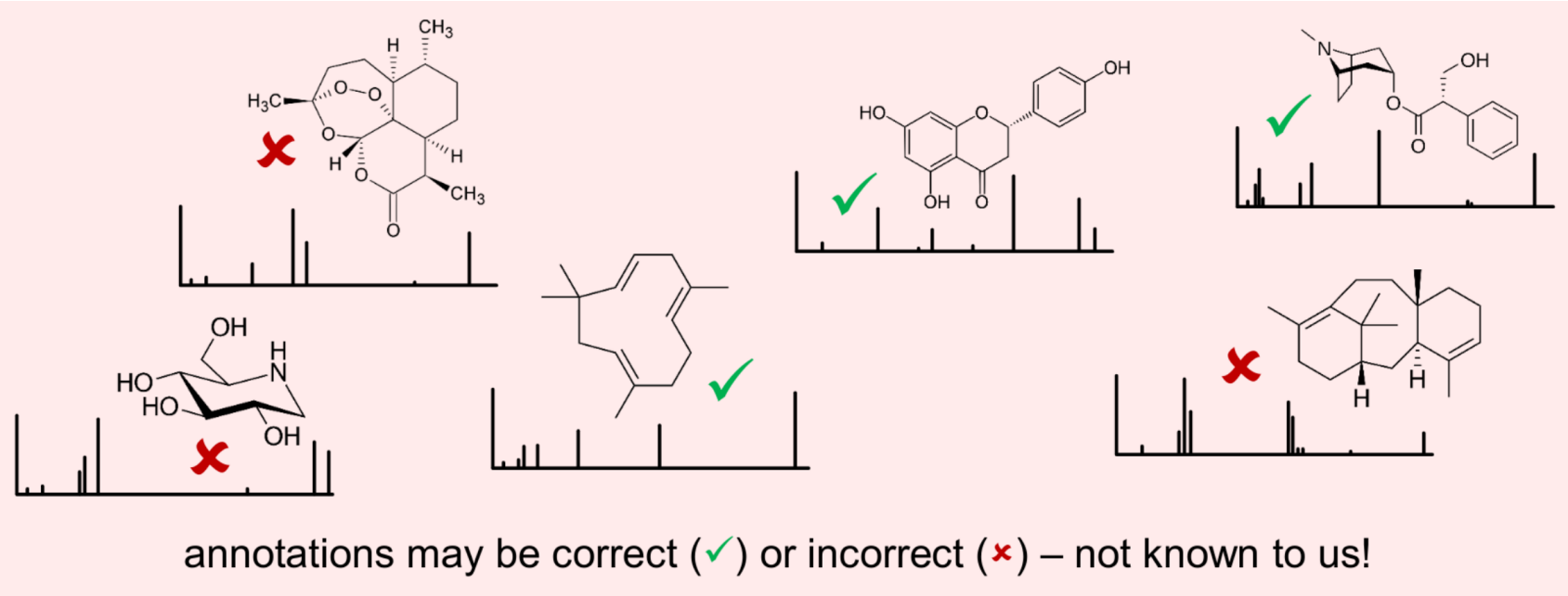
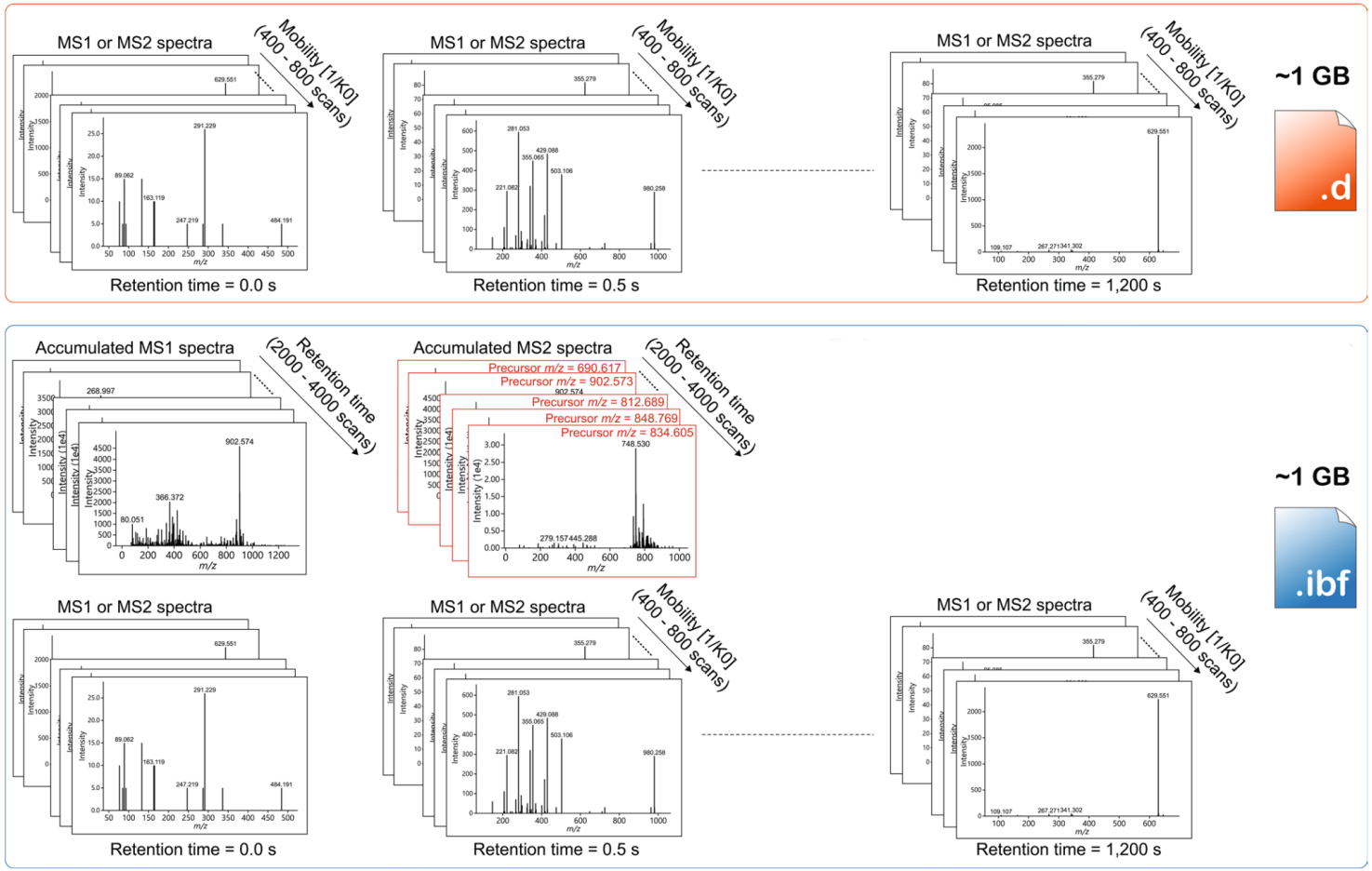


Figure 9: MS/MS analysis

Research backgroud

However, there is a data disaster



Where is a 'Lazy' method ?

Extensive method development

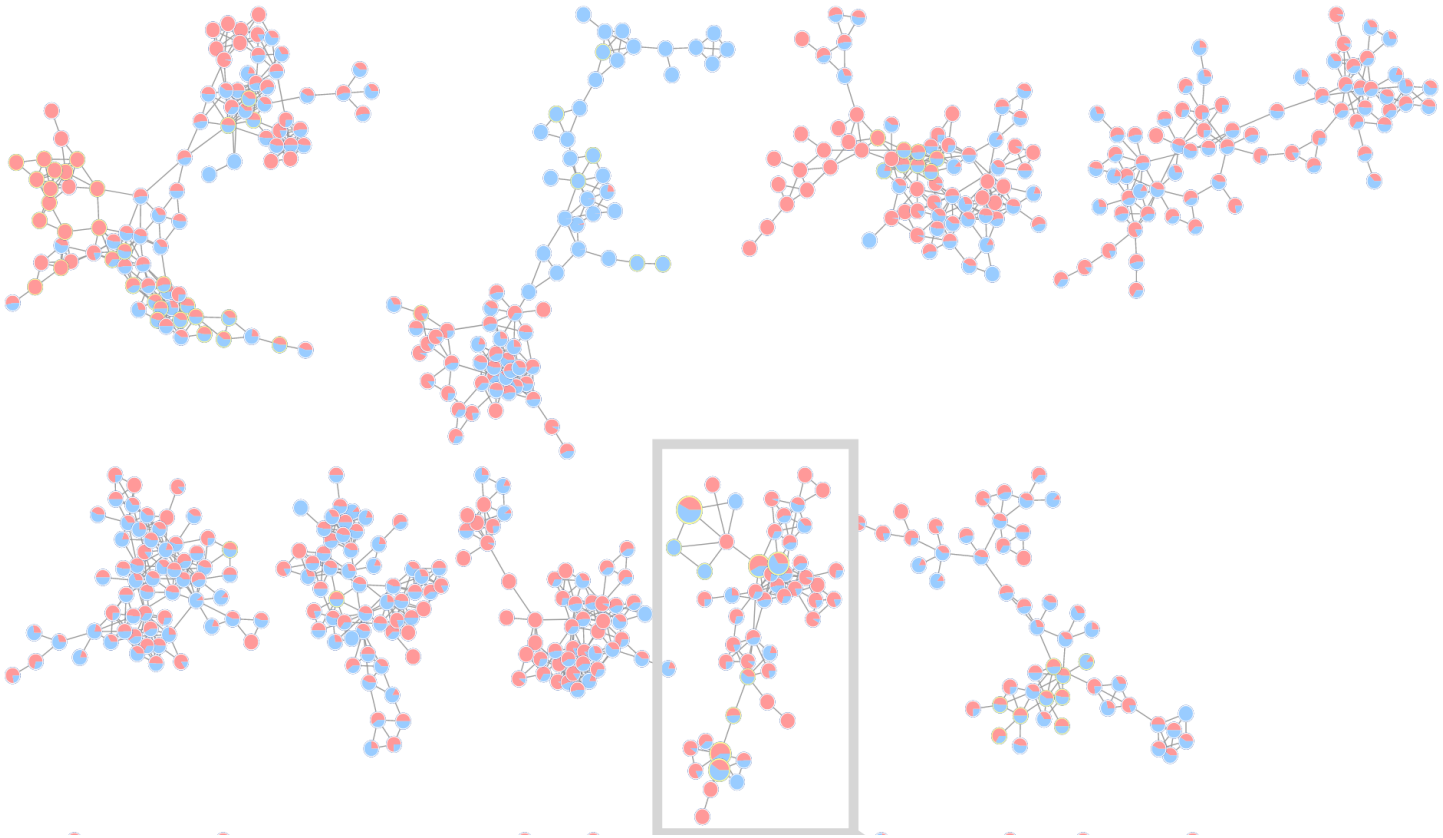
- MS dataset process tool (Integration tool)
  - MZmine2 (recently, MZmine3 update)
  - OpenMS (Python tool)
  - XCMS (R tool)
  - commercial tools. . . . . .

Publish in Nature (method, biotechnology, communication. . . . .):

- GNPS Monopoly
  - molecular networking (MN)
  - Feature based molecular networking (FBMN)
  - Qmistree, IIMN, ...
- CompMass
  - MAdial, MSfinder, MRMPROBES/MRMDIFF
- SIRIUS
  - SIRIUS, ZODIAC, CSI:fingerID, CANOPUS, COSMIC. . . . . .
- . . . . . .

The most popular or sophisticated method for MS analysis

A visualization strategy: GNPS molecular networking



The most popular or sophisticated method for MS analysis

A machine learning method for compound prediction

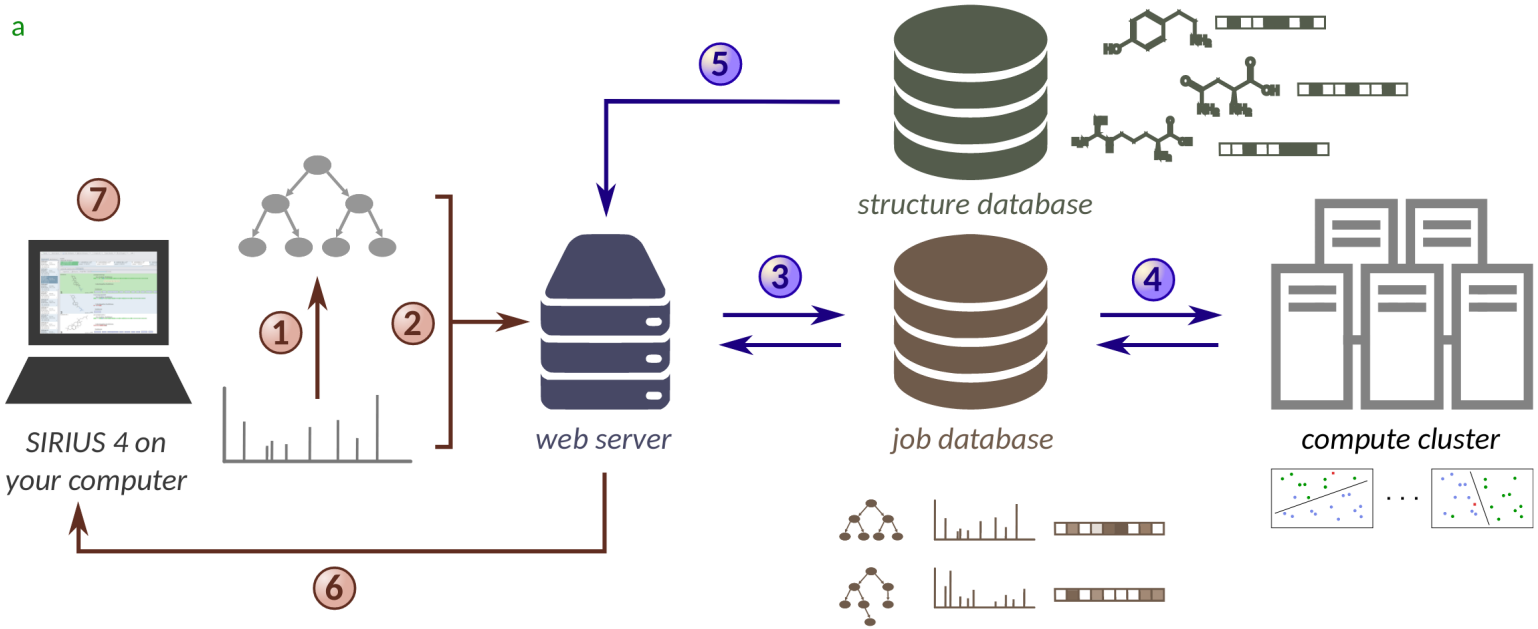


Figure 12: SIRIUS workflow

The most popular or sophisticated method for MS analysis

A comprehensive platform

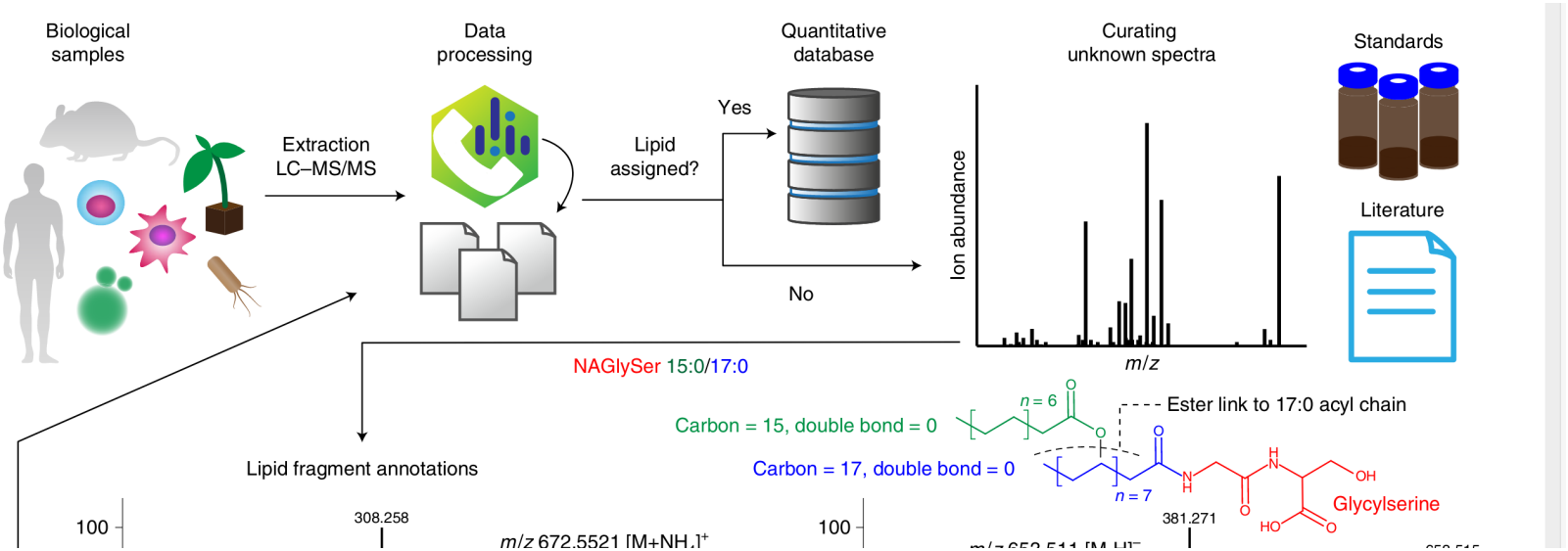


Figure 13: compMass

What should a lazy analysis be?

A method:

1. Computational workflow
   - almost computer do all
2. Comprehensive structure database
   - involve all public available database
3. High accuracy
4. Predictive potency
   - compound without MS/MS spectrum library
   - unknown compound should possibly be cover
5. The most intuitive visualization
   - what is the dataset tell us?
6. Fast collating
   - easily collate metadata for identified compound
7. ...

MCnebula Emerged

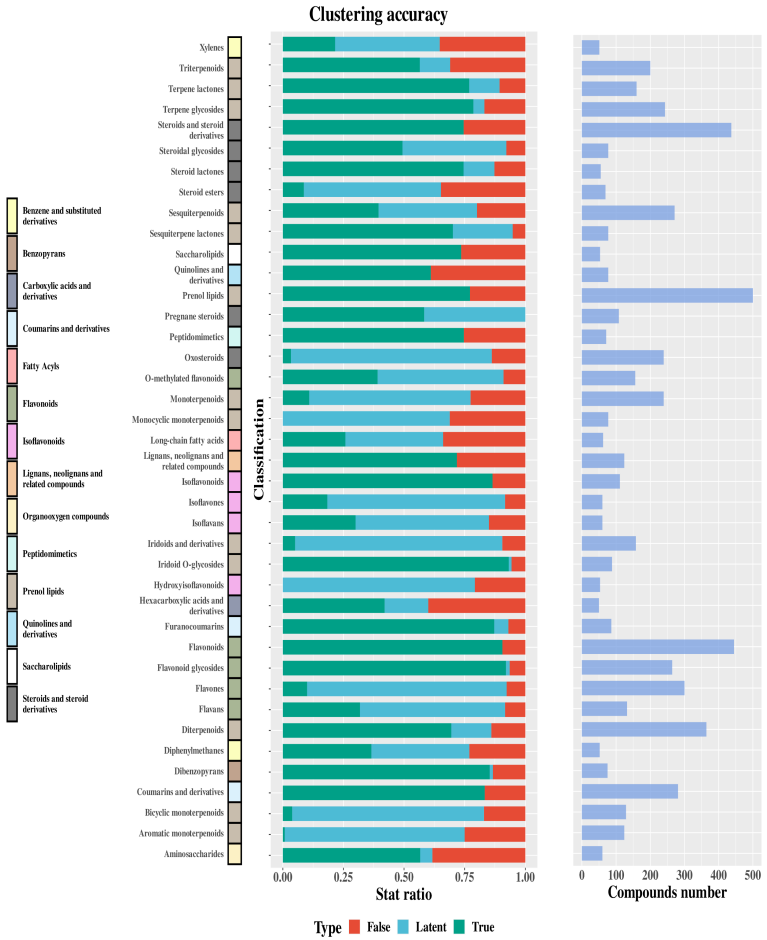


Figure 14: MCnubula

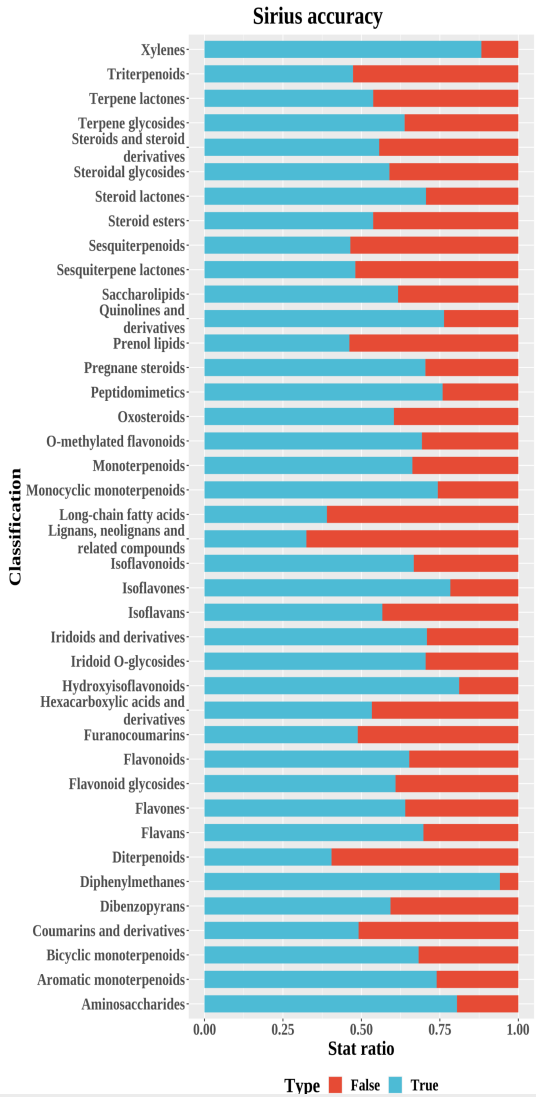
Inovation

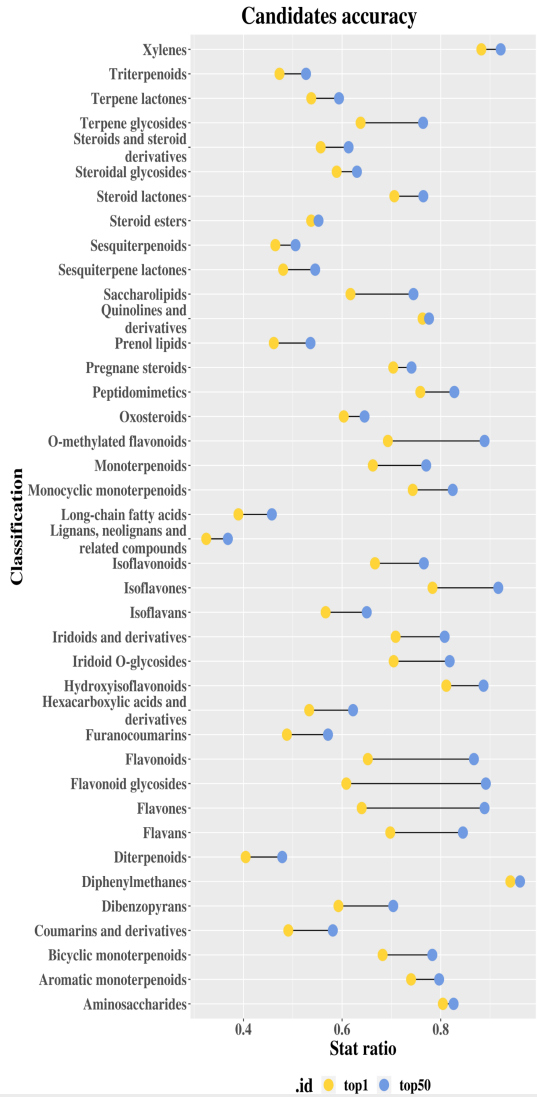
1. Improve the compound prediction accuracy
   - via clustering in classification
   - via considering retention time
   - ...
2. High accuracy classes clustering for MS/MS annotation
   - over 80% accuracy clustering, even unknown compound (no structure information)
3. Intuitive compound classes distribution in network visualization
   - each class involves a sub-nebula to explore the compound annotation
4. MCnebula algorithm integrated in R
   - cover SIRIUS LC-MS workflow analysis into R pipline
5. A wide range of applicability
   - not be confined to metabolome identification
   - not be confined to spectrum library, but structure library

Recent progress

1. Stat clustering accuracy
2. Stat identification accuracy
3. Simulate isotopes pattern and re-stat accuracy
4. Design re-rank method
5. Prepare public data for further validation
   - format MoNA, MassBank, etc. spectrum data as mgf
   - retention time (simulate)
   - retention time (true positive)

Sirius accuracy





The premise of improving accuracy

1. Classes prediction accuracy greater than structure prediction
2. Some correct structures are in candidates
3. Chromatography can be implemented into identification

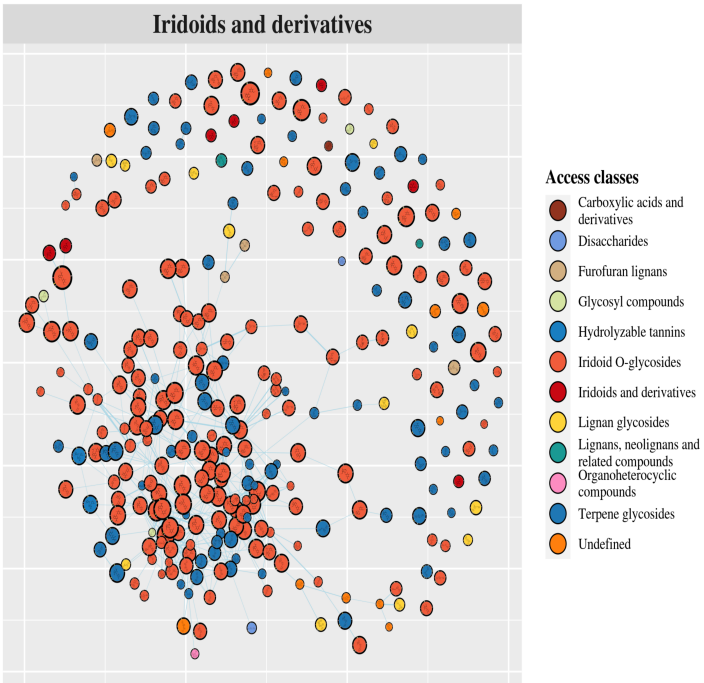
Implementation method

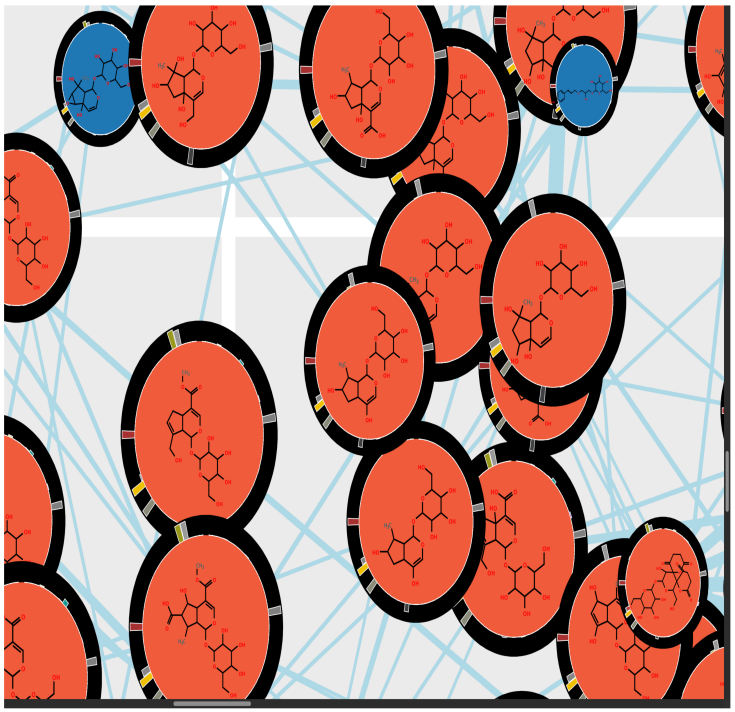
- Collate candidates from identification
- Consider structure similarity in a class
- Integrate retention time (integrate Retip RT prediction)

An assumption: half truth

Based on SIRIUS identification accuracy, assuming that the half part of top score idicate true results

Those identified structures are considered as reference compounds





Re-rank method

Harf as reference compound

- Perform Retip learning, then predict all RT of candidates and filter the impossible.
- Based on structure similarity in a class, perform clustering in this class and get score.
- Integrate all scores and rerank.

Implement a new score for re-rank

Structure candidates re-rank score:

$$S_{re-rank} = (S_{simi} \times W_{simi}) + (S_{cluster} \times W_{cluster}) + (S_{RT} \times W_{RT})$$

part 3: pharmacology

E. ulmoides stir-fried with sailing water contribute to target treatment of kidney

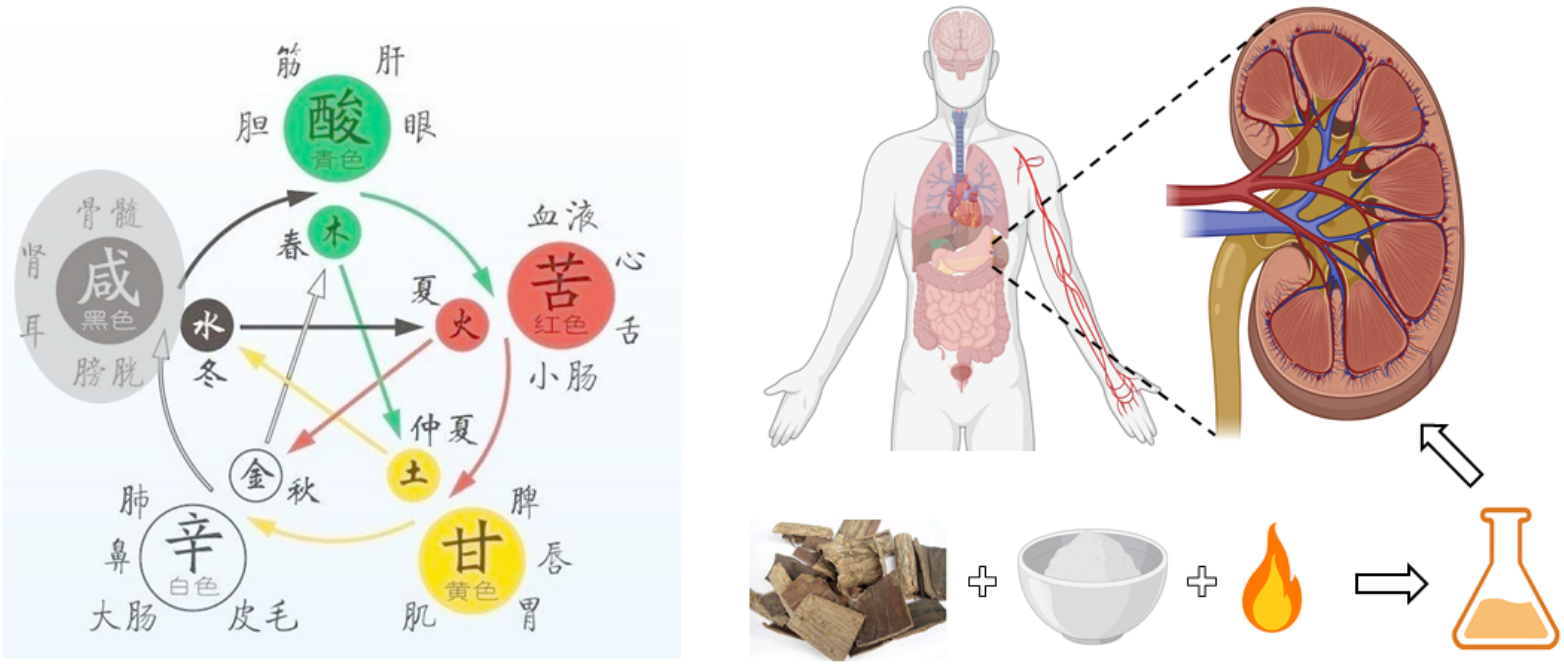
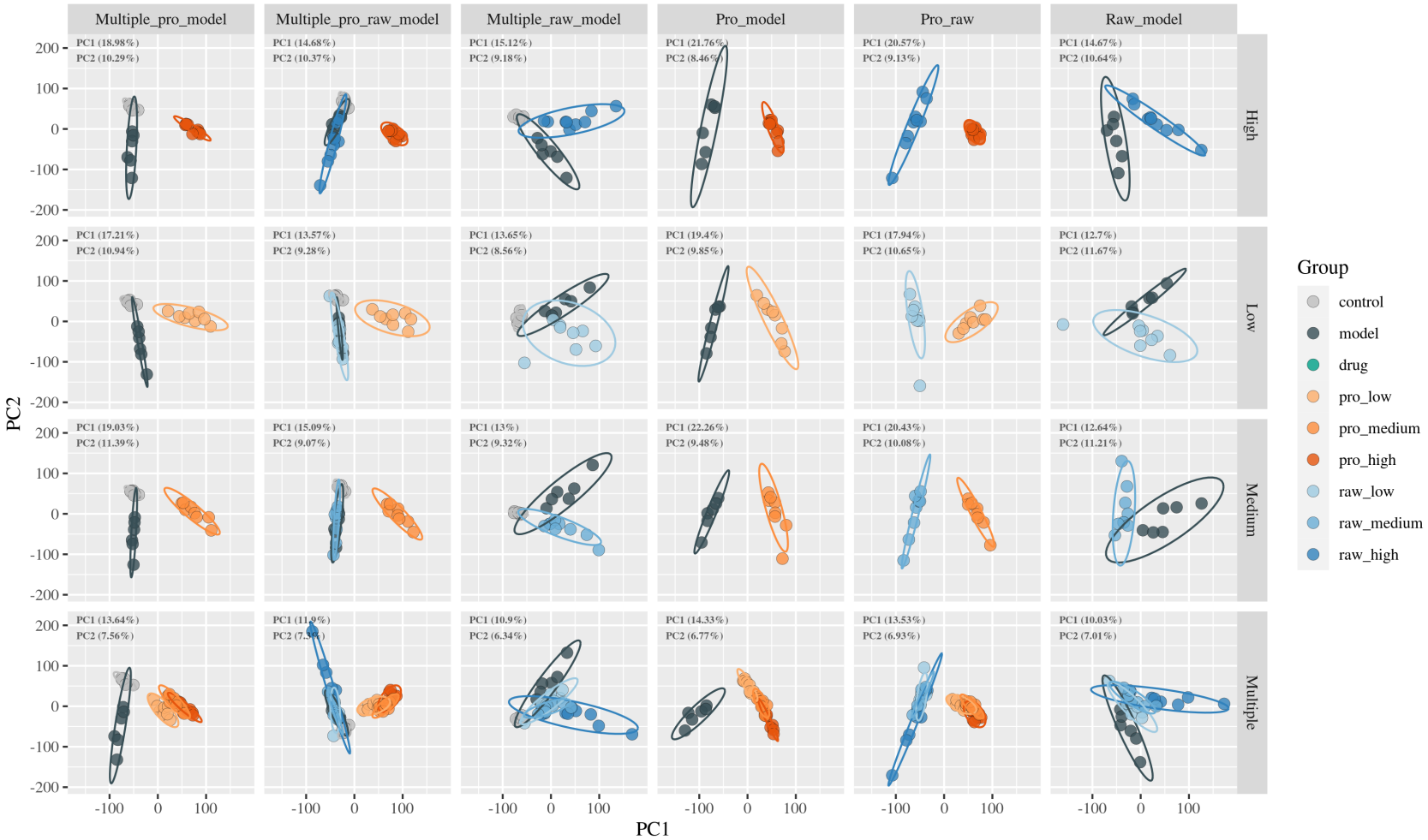


Figure 16: stir-fried with sailing water

Serum metabolome PCA (ion mode: neg)

The re-collate serum metabolome dataset



MetaboAnalysis pathway enrichment

Compound identified from MetaboAnalysis

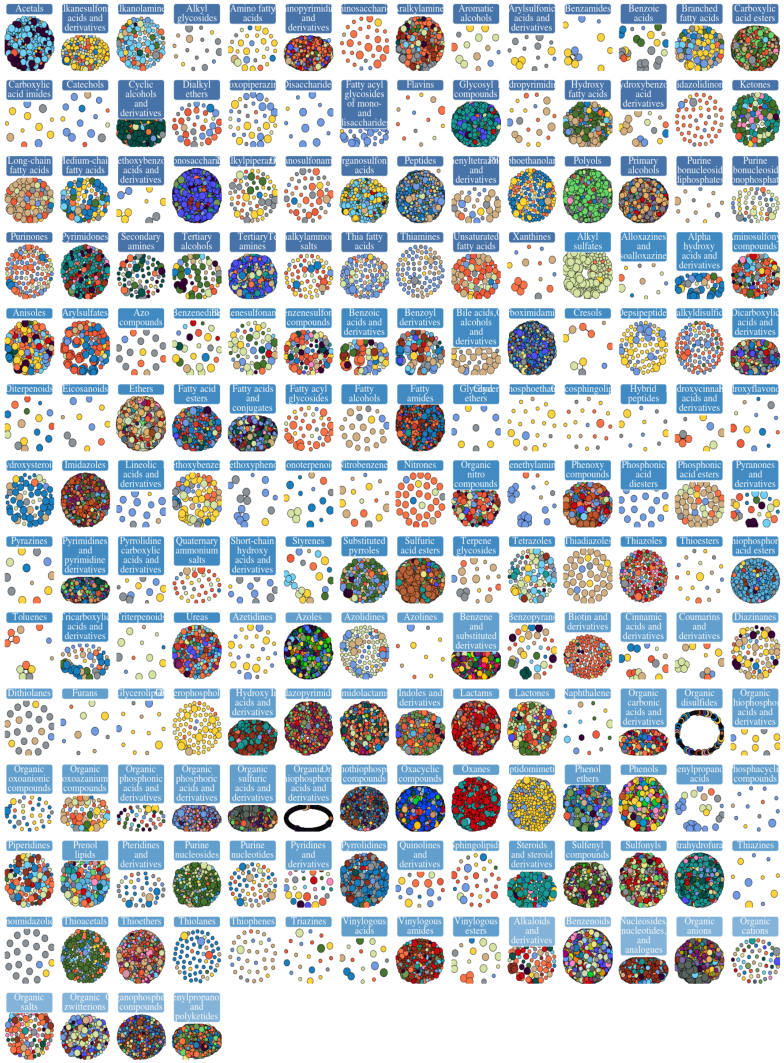
Compounds summary[1]

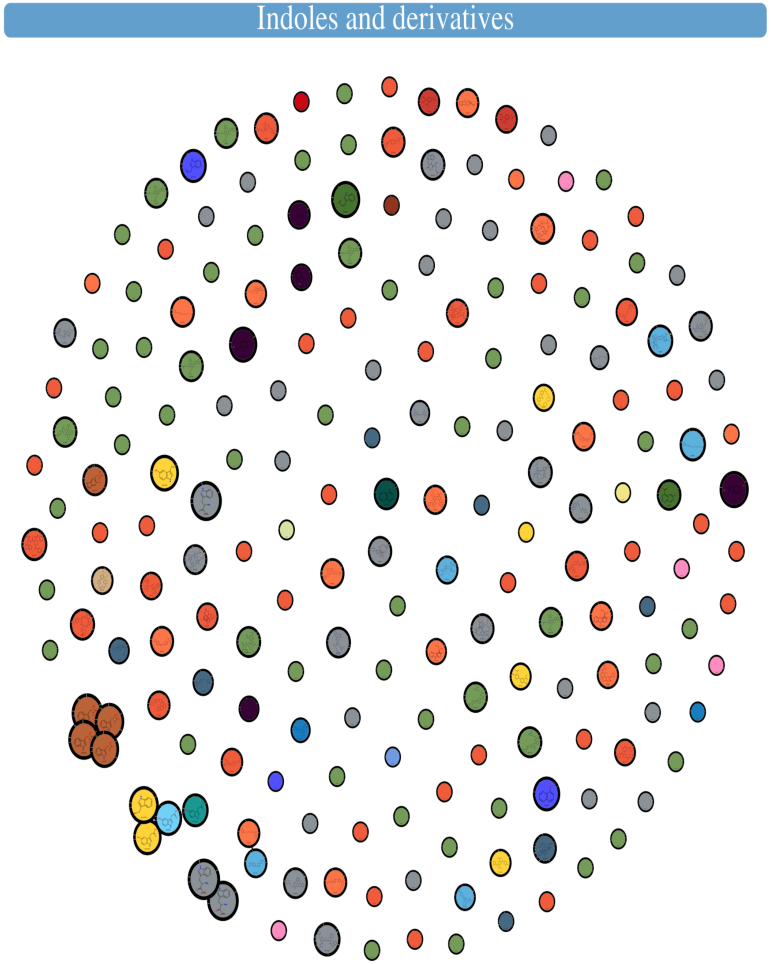
LC-MS

			Pro_model			Raw_model			Pro_raw		
	Name	Vip	p_value	q_value	log2.fc	p_value	q_value	log2.fc	p_value	q_value	log2.fc
Tyrosine metabolism ---- Gamma: 0.0014346 ---- Hits.sig: 19											
17773	1,2-dehydrosalsolinol	2.09	3.156972e-04	2.801997e-03	-2.40	1.401373e-01	0.269757469	-1.03	2.623607e-01	1.612208e-01	-1.38
3661	2-Hydroxy-3-phenylpropenoate;	1.86	5.861307e-04	4.120221e-03	-2.44	1.863859e-02	0.155105834	-1.22	3.931628e-02	4.714590e-02	-1.22
10749	2-Hydroxyphenylacetate;	1.81	8.923267e-04	5.375483e-03	-3.38	6.949672e-03	0.116504767	-1.83	1.306798e-02	2.272734e-02	-1.54
16722	2-Hydroxyphenylacetate;	1.95	3.789563e-01	2.235120e-01	0.61	3.198618e-04	0.033910360	-2.72	7.709450e-03	1.583724e-02	3.33
3130	2-Phenylacetamide;	2.17	4.787005e-05	8.091609e-04	-9.61	5.794610e-01	0.454909205	0.37	4.349195e-03	1.076423e-02	-9.98
3011	3-Methoxy-4-hydroxyphenylglycolaldehyde	1.85	6.057294e-04	4.205147e-03	-1.62	1.899659e-02	0.155731836	-0.92	1.833706e-02	2.838744e-02	-0.70
3258	3-Methoxy-4-hydroxyphenylglycolaldehyde	2.04	5.136352e-04	3.796729e-03	-1.33	2.858462e-03	0.085351344	-1.43	7.884504e-01	3.359960e-01	0.11
3661	3-Methoxy-4-hydroxyphenylglycolaldehyde	1.86	5.861307e-04	4.120221e-03	-2.44	1.863859e-02	0.155105834	-1.22	3.931628e-02	4.714590e-02	-1.22
3085	3,4-Dihydroxy-L-phenylalanine;	2.17	8.833535e-03	2.197970e-02	0.92	4.369825e-03	0.099304131	1.62	4.032123e-02	4.790680e-02	-0.70
3132	3,4-Dihydroxy-L-phenylalanine;	2.04	2.479301e-01	1.704938e-01	0.34	2.030467e-02	0.158109139	1.08	2.884917e-02	3.856373e-02	-0.74
17251	3,4-Dihydroxy-L-phenylalanine;	2.10	1.238889e-03	6.636721e-03	1.03	1.047508e-01	0.245452263	1.15	7.558880e-01	3.266524e-01	-0.13

Figure 19: compound

MCnebula analysis





Schedule

- MCnebula
  1. Job done of re-rank method in R codes
  2. Simulate re-rank method in identification
  3. Compare clustering effectiveness with Qmistree and MolNetEnhancer
- Renal LC-MS data collate and analysis

END

Thank you