End-to-End Phenotype Prediction using Data Independent Acquisition Mass Spectrometry Tensor

Fangfei Zhang[1]#, Shaoyang Yu[2]#, Lirong Wu[3]#, Zelin Zang[3], Yaoting Sun[1], Yi Xiao[1], Stan Z. Li[3]†, Zhongzhi Luan[2]†, Tiannan Guo[1]†

[1]Guomics Laboratory of Proteomics Big data, School of Life Sciences, Westlake University, Hangzhou, China
[2]Sino-German Joint Software Institute (JSI), Beihang University, Beijing, China
[3]Center for AI Research and Innovation (CAIRI), School of Engineering, Westlake University, Hangzhou, China

First authors
† Corresponding authors

西湖大學 WESTLAKE UNIVERSITY

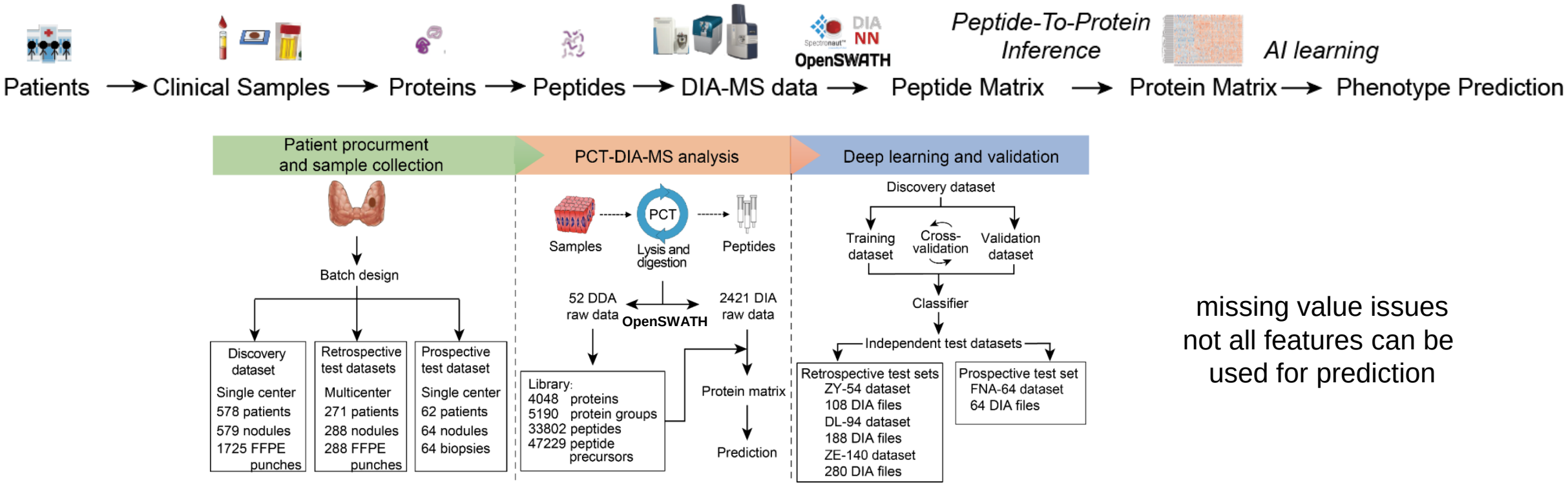
GUOMICS since 2017

www.guomics.com

ASMS REBOOT June 1 - 15 2020

- Introduction
  - Phenotype prediction by high-throughput DIA proteomics for clinical application
  - DIA-MS based proteomics and software tools: a glimpse in 2020
  - New analysis scheme– DIA Tensor (DIAT)
- Result
  - Construction of DIAT
  - Characterization of DIAT
  - Deep learning framework for DIAT
  - Phenotype prediction with DIAT
    - Classification of tumor/non-tumor on a hepatocellular carcinoma cohort
    - Classification of tumor/non-tumor on a thyroid carcinoma cohort

Deep Learning of high-throughput DIA proteomics from clinical cohorts



missing value issues
not all features can be
used for prediction

Protein Classifier for Thyroid Nodules Learned from Rapidly Acquired Proteotypes
Sun et al. MedRxived: April 14, 2020:10.1101/2020.04.09.20059741

REVIEW

Proteomics
Proteomics and Systems Biology
www.proteomics-journal.com

Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020

*Fangfei Zhang, Weigang Ge, Guan Ruan, Xue Cai, and Tiannan Guo**

SWATH/DIA Acquisition Scheme

Shotgun CID	Purvine et al., 2003
DIA	Venable et al., 2004
MS^E	Plumb et al., 2006
PAcIFIC	Panchaud et al., 2009
XDIA	Carvalho et al., 2010
SWATH	Gillet et al., 2012
MSX	Egertson et al., 2013
SONAR	Moseley et al., 2018
WiSIM	Koopmans et al., 2018
BoxCar DIA	Meier et al., 2018
ScanningSWATH	Messner et al., 2019
diaPASEF	Meier et al., 2019
pulseDIA	Cai et al, 2020

SWATH/DIA Software Tools

Spectral Library Building Tools

Experimental Spectra Search				
Comet Eng et al., 2013	X!Tandem Craig et al., 2003	Mascot Brosch et al., 2009	Sequest Eng et al., 1994	OMMSA Brewis et al., 2010
MS GF+ Kim et al., 2014	ProteinPilot Shilov et al., 2007	pFind Chi et al., 2018	MSFragger Kong et al., 2017	MaxQuant Cox et al., 2008

Library Building		
SpectraST Lam et al., 2007	Skyline MacLean et al., 2010	Pulsar Spectronaut

Computationally Predicted Spectral Library		
Deep Mass Tiwary et al., 2019	Prosit Gessulat et al., 2019	pDeep Zhou et al., 2017

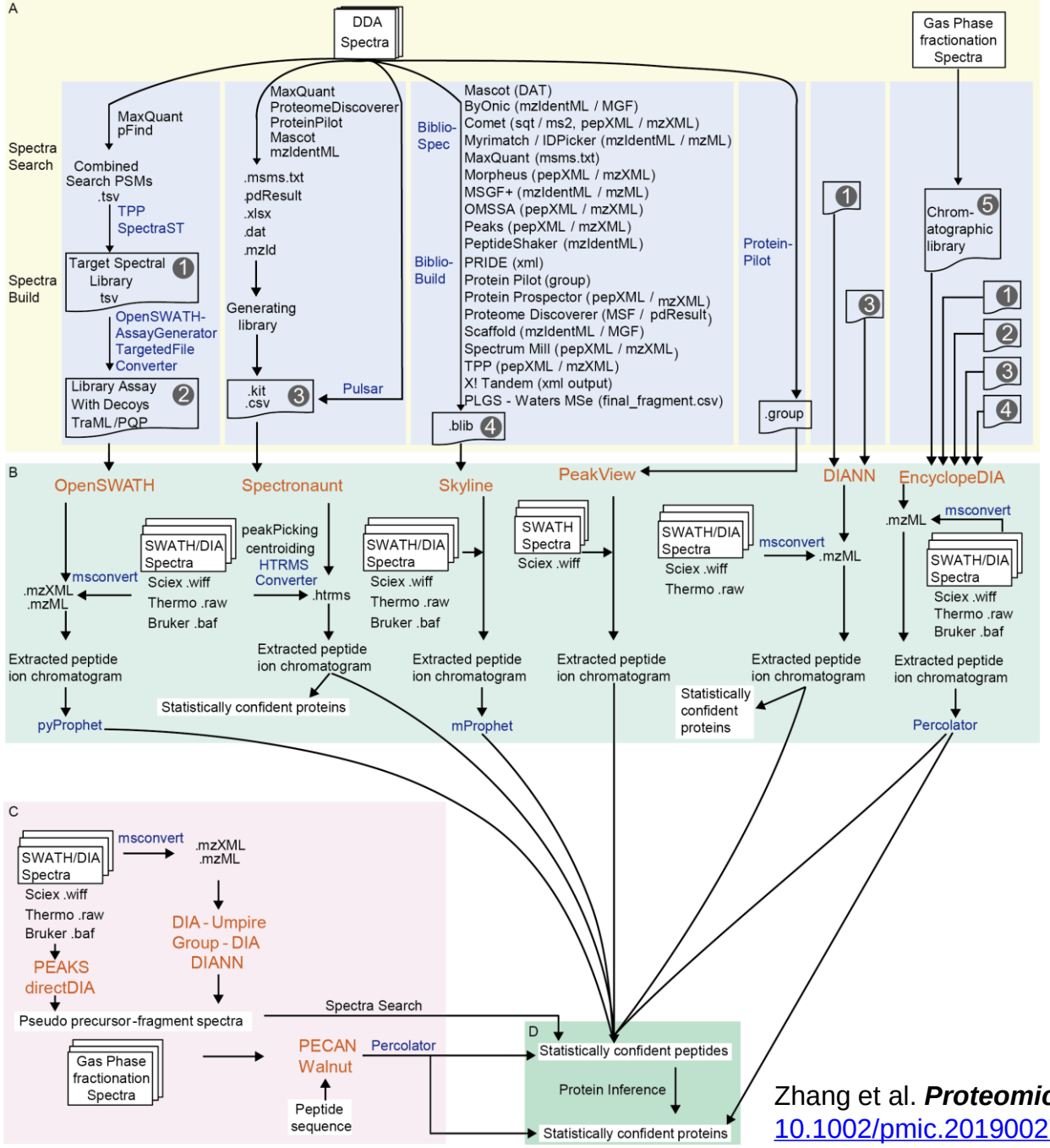
DIA Analysis Tools

Library-based Analysis				
OpenSWATH Roest et al., 2014	Spectronaut Bruderer et al., 2015	EncyclopeDIA Searle et al., 2018	Skyline MacLean et al., 2010	PeakView Gillet et al., 2012
Spector Peckner et al., 2018	Proteome Discoverer Thermo	PIQED Meyer et al., 2017	DIA-NN Demichev et al., 2018	

Library-free Analysis				
directDIA Spectronaut	Peaks Tran et al., 2019	DIA-Umpire Tsou et al., 2015	Group-DIA Li et al., 2015	PECAN Ting et al., 2017

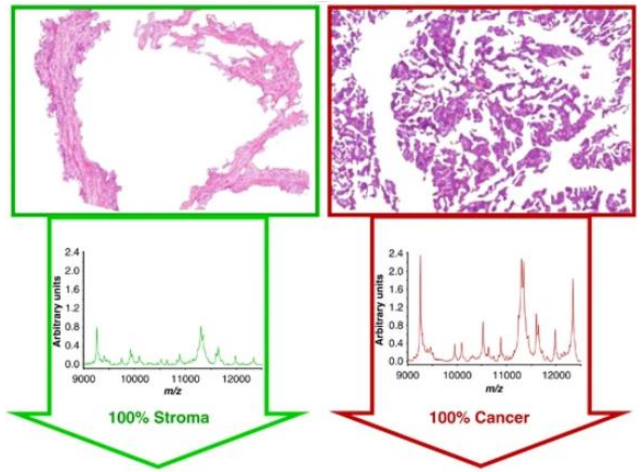
Statistical Validation		
mProphet Reiter et al., 2011	pyProphet Teleman et al., 2015	Percolator Kall et al., 2007

◆ SCIEX Triple TOF5600/6600 ▲ Thermo Orbitrap (QE-HF, QE-HFX, Lumos) ● Bruker timsTOF Pro
◆ SCIEX Triple TOF 6600+ ▲ Thermo LTQ ★ Waters Xevo G2-XS QTof



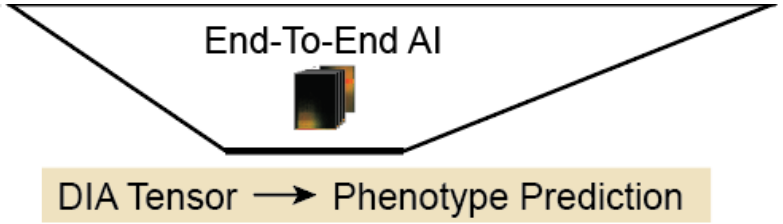
Zhang et al. ***Proteomics*** (2020)
10.1002/pmic.201900276

A novel analysis procedure and format DIA Tensor (DIAT)



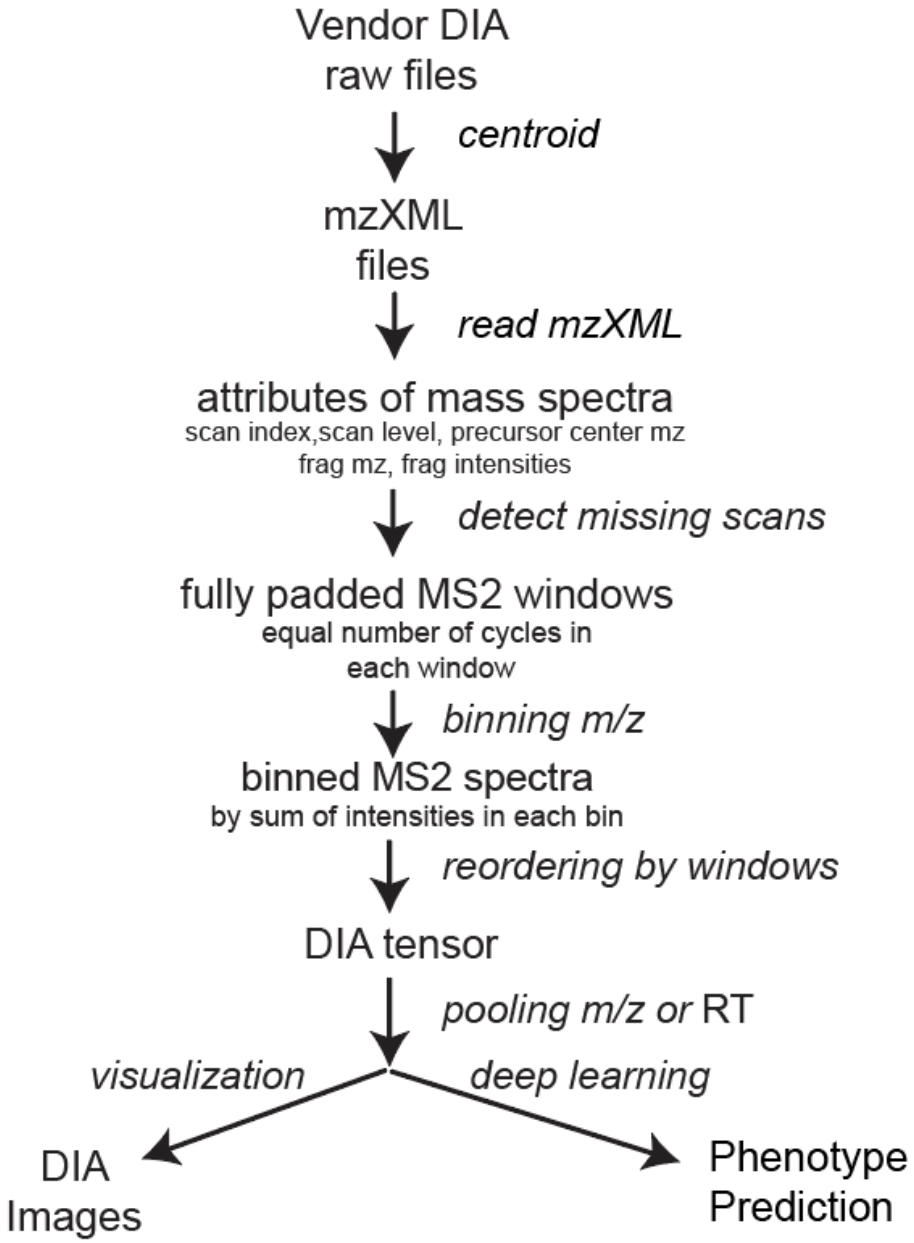
MALDI Mass Imaging

Aichler et al. ***Laboratory Investigation*** (2015)

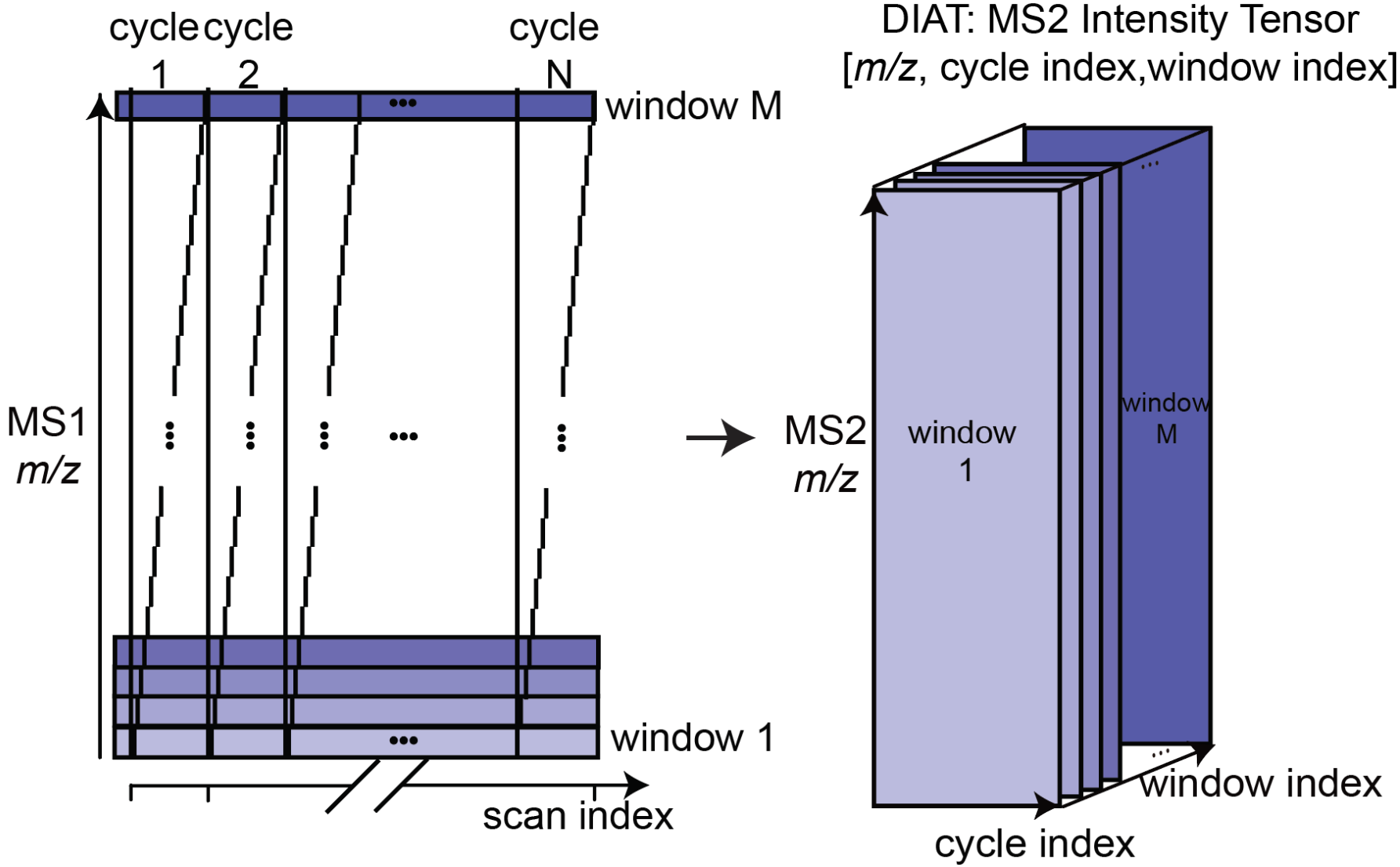


- E2E Phenotype prediction
- Transfer a pretrained DL network from computer vision

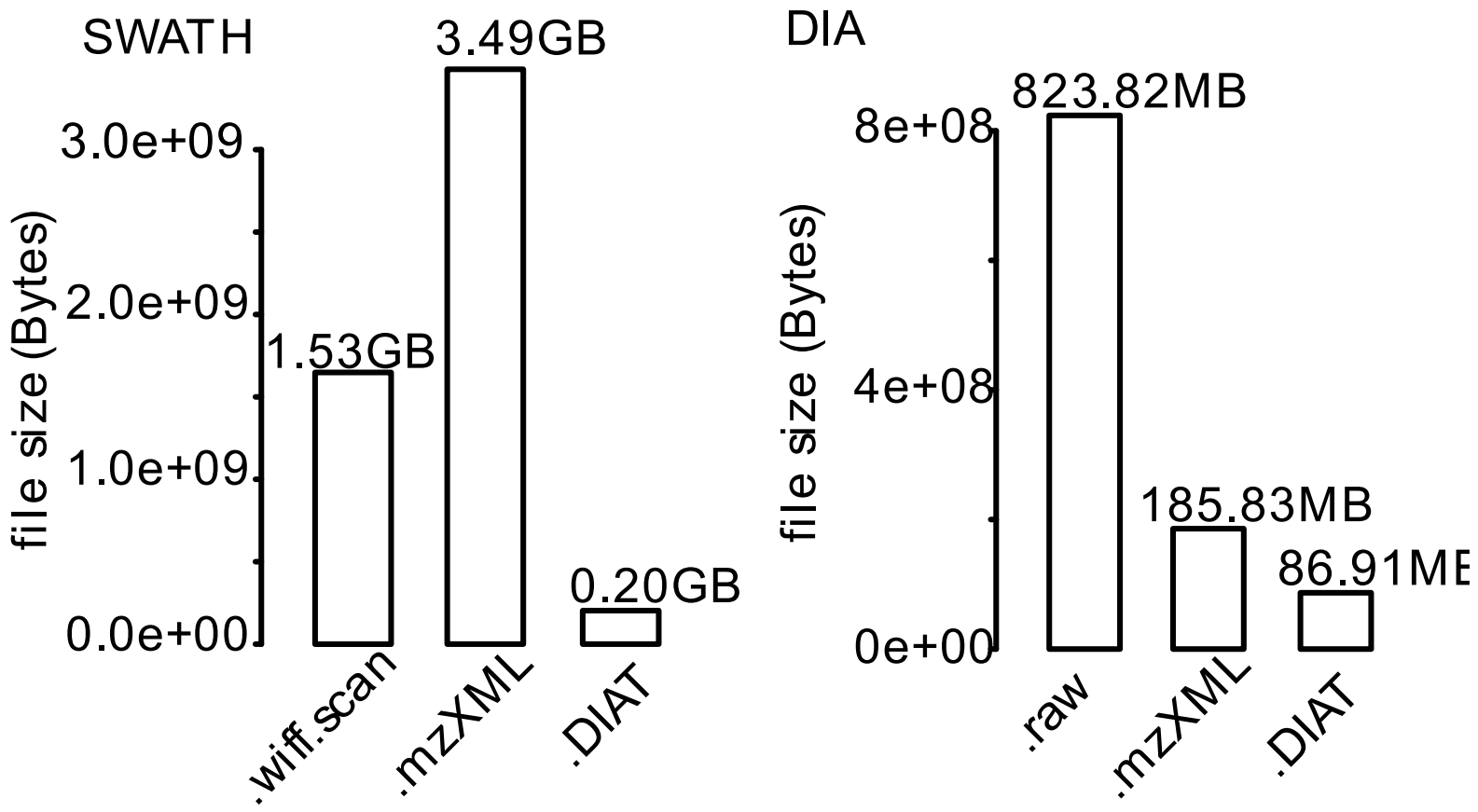
Construction of DIAT

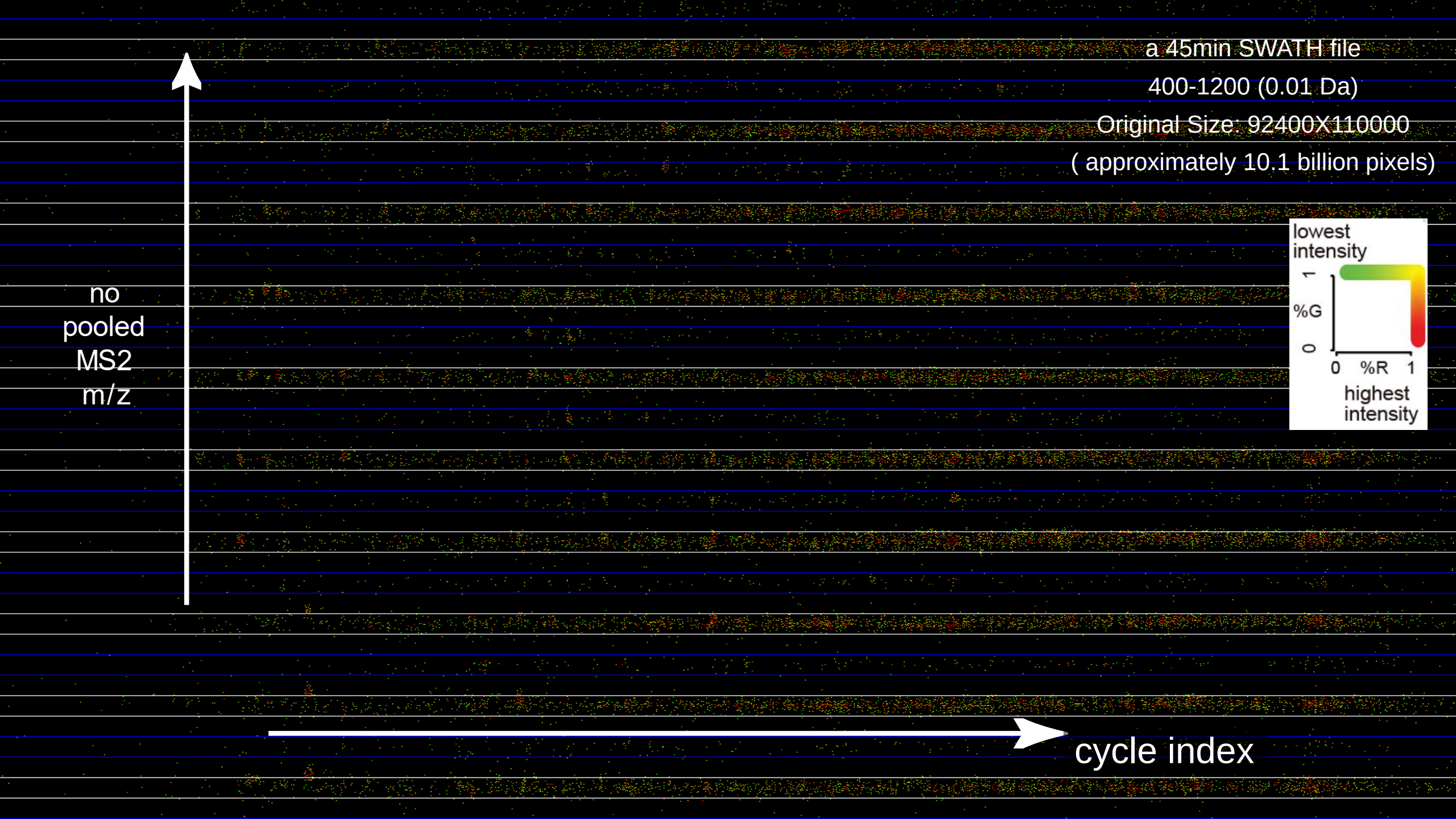


Construction of DIAT

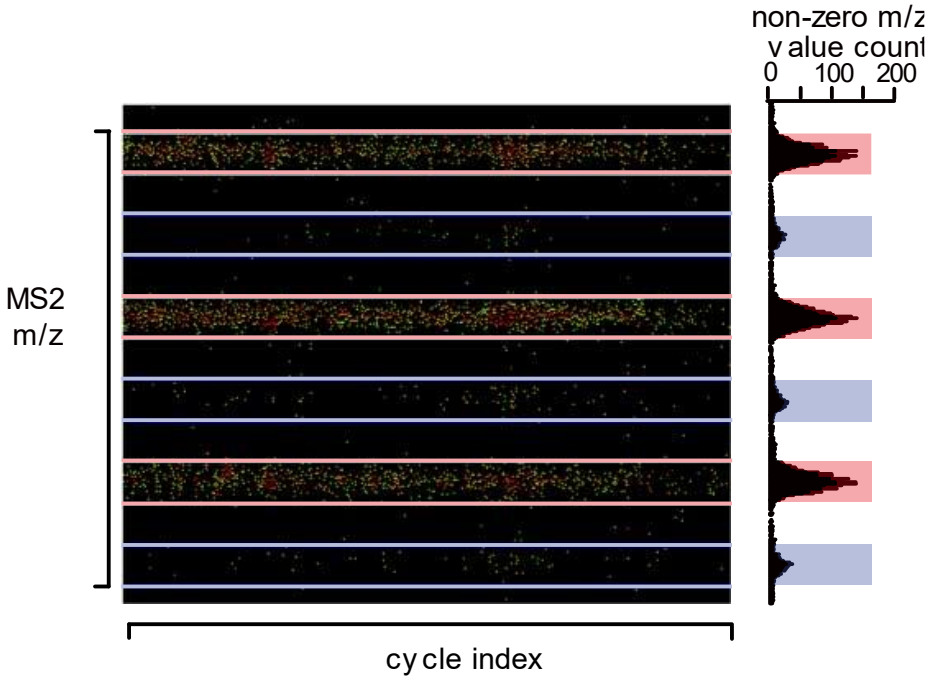


Construction of DIAT




a 45min SWATH file
400-1200 (0.01 Da)
Original Size: 92400X110000
(approximately 10.1 billion pixels)
lowest intensity
1
%G
0
0
%R
1
highest intensity
no pooled MS2 m/z
cycle index

Characterization of DIAT



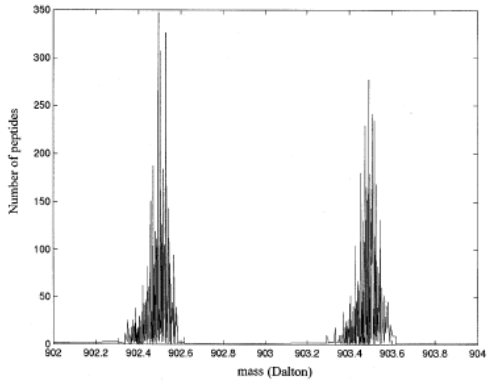
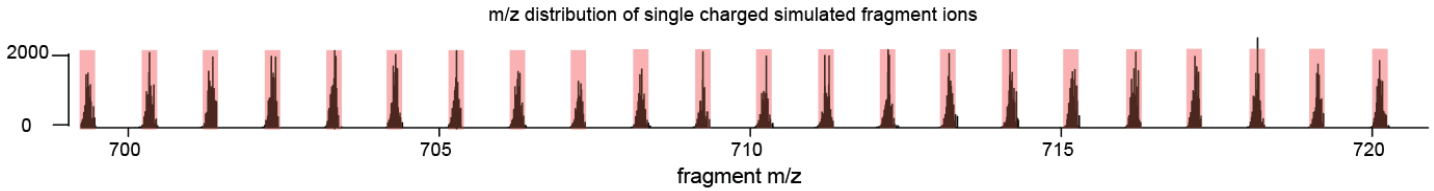
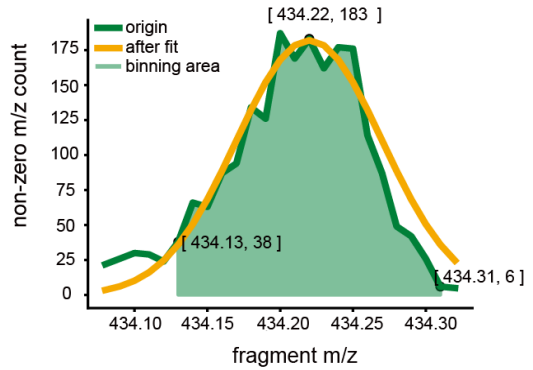


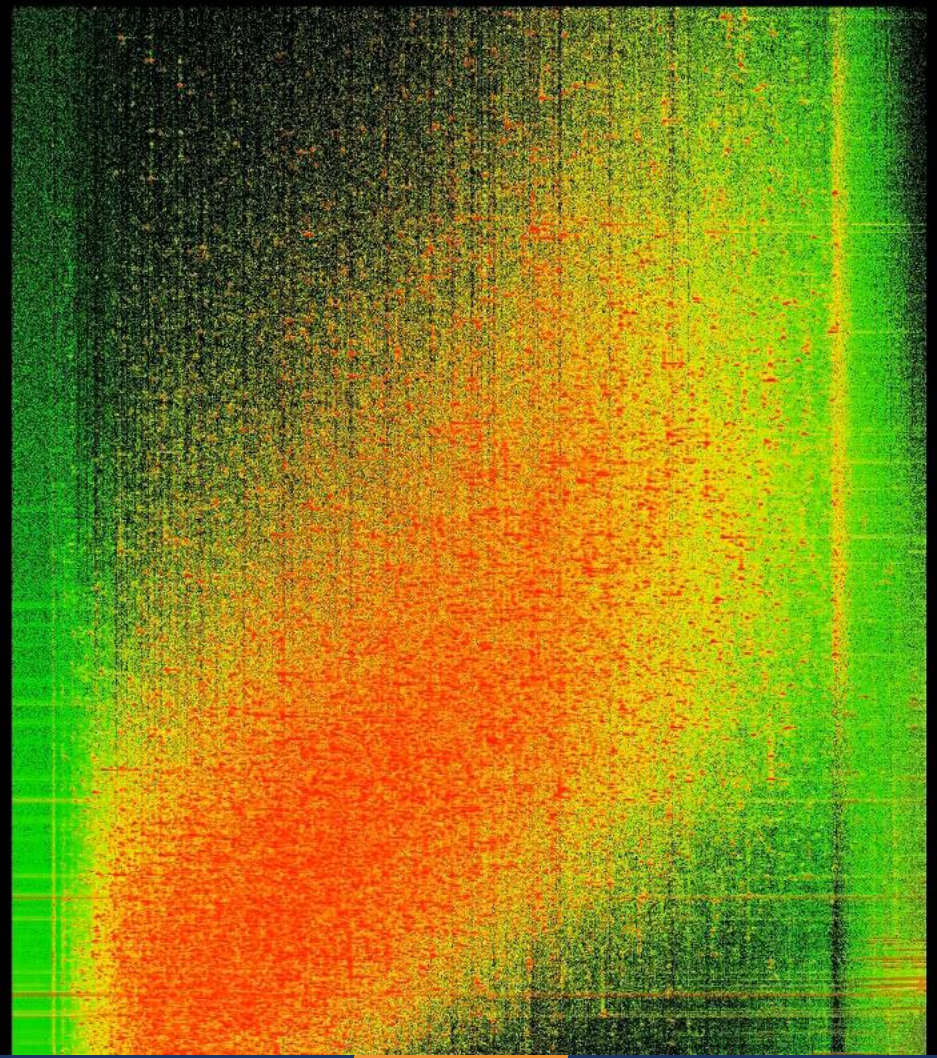
Figure 1. Number of peptides in SWISS-PROT of masses between 902 and 904 Da. The ordinate has a mass resolution of 10^{-7} Da.

Gay et al. ***Electrophoresis*** *20*, 3527-3534 (1999)

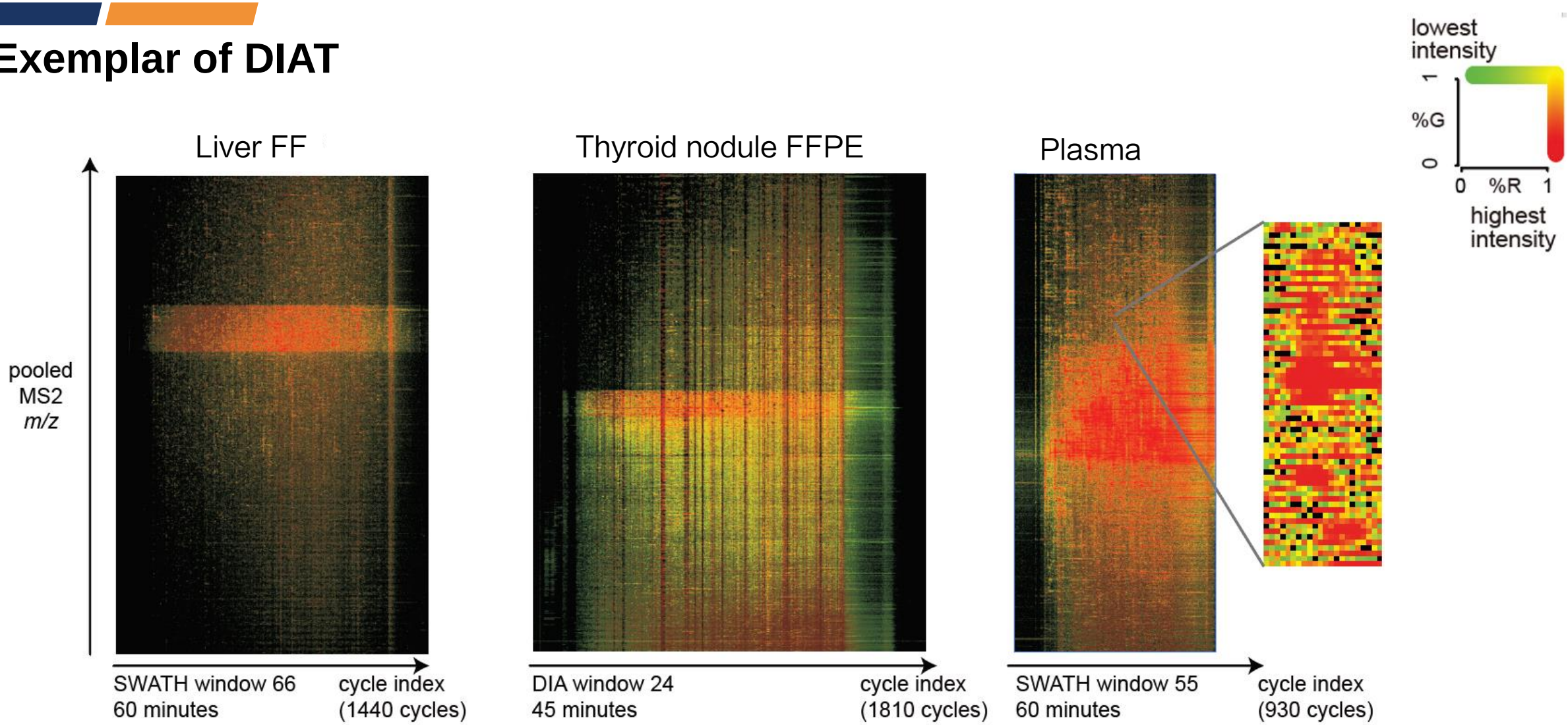




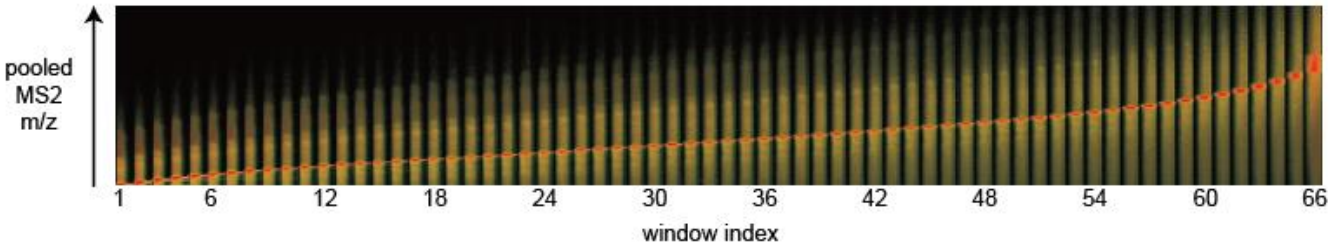
110K -> 2K



Exemplar of DIAT



Characterization of DIAT




```
usage: DIAtensor_v1.1.3_Win64.exe [-h] -i <path> -o <path> -T {image,tensor}
                                  [-t <int>] [-a | -c <int>] [-m <int> <int>]
                                  [-g <float> <float>] [-b <float>]
                                  [--pool_mz] [--pool_rt] [-D {3D,2D}]
                                  [-C {view,gray}] [-F {png,bmp}]

[ Version: 1.1.3 ] DIAtensor is software tool to convert mzXML of data-
independent acquisition (DIA) mass spectrometry (MS) to DIA tensor or image.
```

https://github.com/guomics-lab/DIAtensor/releases

Generate tensor (save *.diat)

- Generation of a tensor by 1400 cycle with pooling in mz dimensions.

```
DIAtensor.exe -i E: /HCCSW -o E: /tensor -T tensor -c 1400 -b 0.01 --pool_mz
```

- Generation of a tensor by the m/z range 400~2000 Da, auto aligned cycle based on the gradient of 0-45 minutes,with pooling in mz dimensions.

```
DIAtensor.exe -i E: /HCCSW -o E: /tensor -T tensor -m 400 2000 -a -g 0 45 -b 0.01 --pool_mz
```

Read tensor (read *.diat)

```
import numpy as np
diat = np.load("<file_path>.diat")['diat']
```

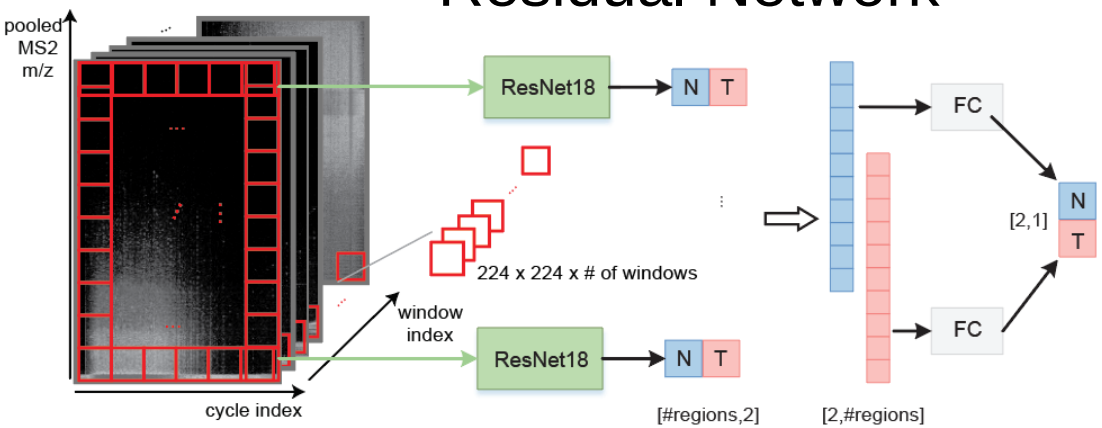
Generate image (save *.png / *.bmp)

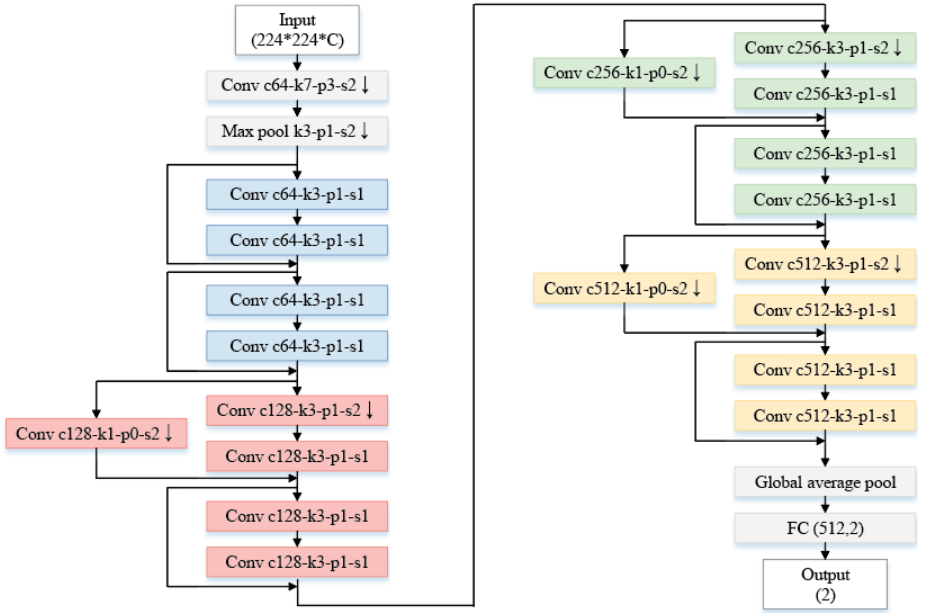
- Generation of an aligned image by 1400 cycle with pooling in both the m/z and RT dimensions.

```
DIAtensor.exe -i E: /HCCSW -o E: /img -T image -c 1400 -b 0.01 --pool_mz --pool_rt -D 2D -C view -F png
```

Deep learning for DIAT

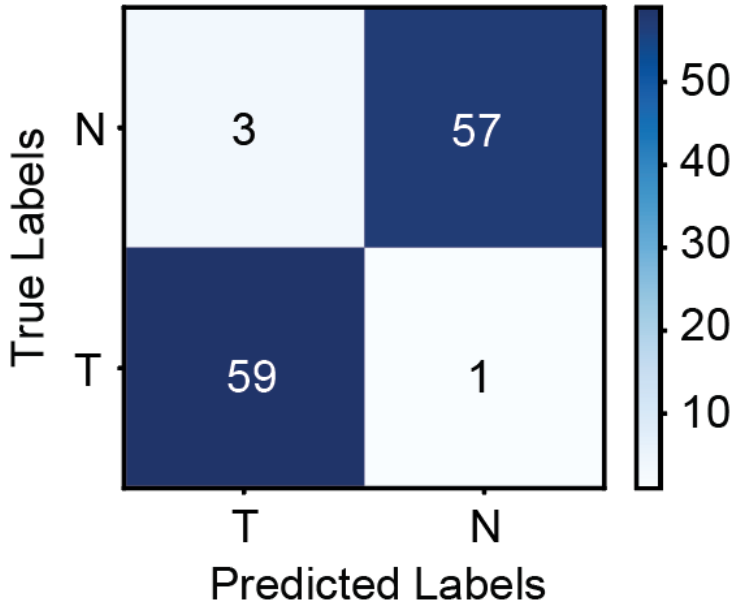
Residual Network

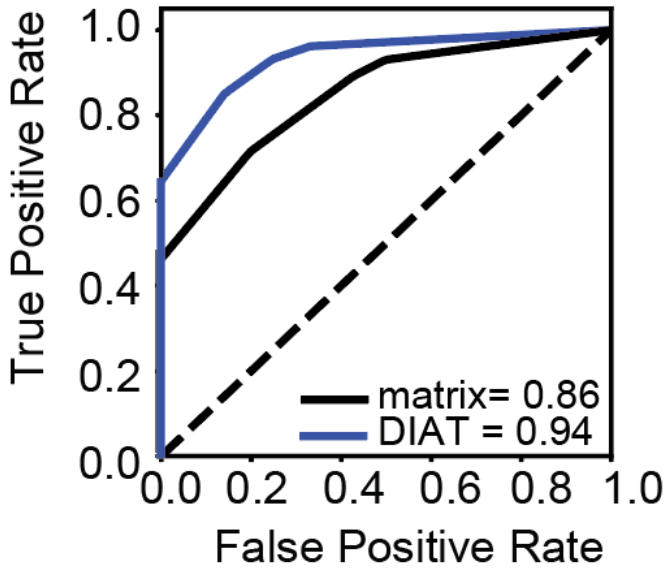




Phenotype Prediction with DIAT

Hepatocellular Carcinoma Diagnosis
Fresh Frozen Tissue
51Tumor (T) 51Normal (N)
102 PCT-SWATH samples



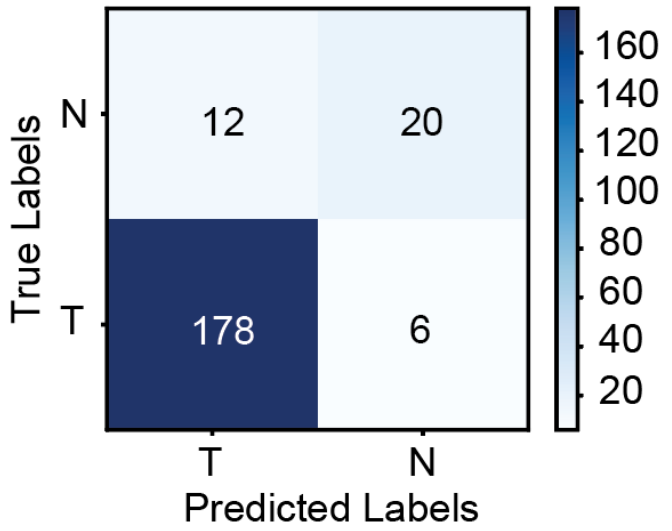


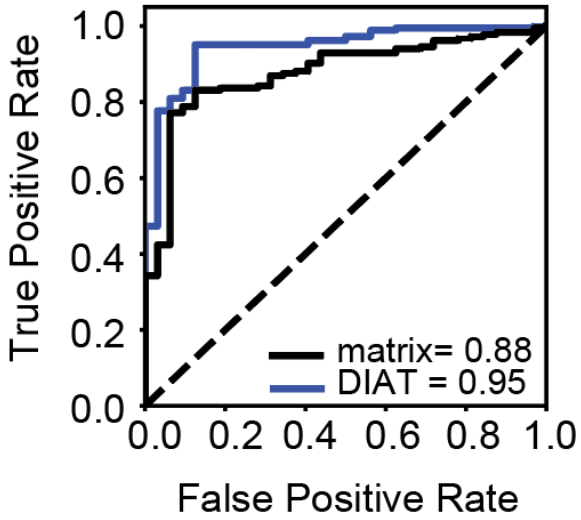
	HCC	
	DIAT	OpenSWATH Matrix
Accuracy	0.968	0.933
Precision	0.952	0.919
Recall	0.983	0.950
F1-score	0.967	0.934

Phenotype Prediction with DIAT

Thyroid Nodule Diagnosis

Discovery cohort Single center FFPE tissue 366 papillary thyroid carcinoma 126 normal thyroid 492 PCT-DIA samples
Test cohort Multicenter FFPE tissue 184 papillary thyroid carcinoma 32 normal thyroid 216 PCT-DIA samples





	Thyroid Cancer	
	DIAT	OpenSWATH Matrix
Accuracy	0.917	0.884
Precision	0.937	0.930
Recall	0.967	0.935
F1-score	0.952	0.932

Summary

Current role as storage and prediction entity for large scale DIA proteomics deep learning phenotype prediction.

Next,

- Direct analysis of super short gradient file (<5min)
- Interpretable model for feature selection
    - fragment ion identification
    - peptide/protein inference
- Pooled DIAT back to mzXML, incorporated with targeted analysis approaches
- Include another dimension for data structures as diaPASEF

Thanks

- Guo’s group for MS-DIA data support
- Luan’s group for coding support
- Li’s group for deep learning support

BioRxived: March 5th, 2020 10.1101/2020.03.05.978635v1

Phenotype Prediction using a Tensor Representation and Deep Learning from Data Independent Acquisition Mass Spectrometry.

Fangfei Zhang #, Shaoyang Yu #, Lirong Wu #, Zelin Zang, Xiao Yi, Jiang Zhu, Cong Lu, Ping Sun, Yaoting Sun, Sathiyamoorthy Selvarajan, Lirong Chen, Xiaodong Teng, Yongfu Zhao, Guangzhi Wang, Junhong Xiao, Shiang Huang, Oi Lian Kon, Gopalakrishna N. Iyer, Stan Z. Li *, Zhongzhi Luan *, Tiannan Guo *.

www.guomics.com