An end-to-end machine learning workflow for MS-based proteomics



Siegfried Gessulat, Tobias Schmidt, Michael Graber, Samia Ben Fredj, Lizi Mamisashvili, Patroklos Samaras, Florian Seefried, Magnus Rathke-Kuhnert, Daniel P. Zolg, Martin Frejno. MSAID GmbH, Garching b. München, Germany

+ Overview

Machine learning augments various steps in proteomics data analysis, from training models for peptide properties and predicting them, to the usage of target/decoy classifiers as in Percolator¹ for error control. Generating, evaluating, and integrating such models largely remains unautomated manual work. Here we present an end-toend workflow that automates the steps from raw data to production-ready model.

+ Workflow

Data ingest

Download public spectrum and search files from PRIDE.

3 Model training

The model architecture and hyperparameter search generated, trained, and evaluated >2,500 distinct models within the last year. Trainings are logged along with the respective model evaluation.

Figure 3. log of								Metrics >				Parameters >		
rigare s. rog or		↓ Start Time	Duration	Run Name	User	So	urce	Version	Models	s loss	Ir	number of para	batch_size	class_weight
retention time		⊘ 20 days ago	38.3min	retention_ti 4	samfredj	9	hyper py	f407e0	R kors	0.032	1 2800-1	101702	None	None
models trained		⊘ 20 days ago	18.0min	retention_ti	samfredj		Date: 2022-04-	30 18:42:39	Sour	ce: 😐 hyper.py		Git Commit: 41af68c5ea88 dfbfe39602d2	3b52034d27e290d 2da	None
		⊘ 20 days ago	34.6min	retention_ti	samfredj		User: samfredj		Dura	tion: 2.6h		Status: FINISHED		None
with architecture		⊘ 20 days ago	2.7min	retention_ti	samfredj		* 🖿 model		Full Path:/mnt/	storage_machine_learn	.0		L	None
		⊘ 20 days ago	11.9min	retention_ti	samfredj		MLmod	el	Size: 55.71KB				mitt	None
search and hyper-		⊘ 20 days ago	4.9min	retention_ti	samfredj	dj ⊑ Bi conda.yami Bi requirements.txt		- Transfer Learned				None		
parameter		⊘ 20 days ago	2.8min	retention_ti	samfredj	c	 tensorboard_logs base_model_lowesspng 	70	Delta abs. 95%= 1.62		1 M		None	
pararriecer		20 days ago	4 2min	retention ti	samfredi	-	meta_conf	guration.json 60 n=8000	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1			None		

Setting-up an ingest requires meta information that is not always available on PRIDE.	Indexing: infer relationship of spectrum and search files Ingest: convert to standard formats and upload to AWS S3
2 Data preprocessing Preprocessing is purpose- and model- specific.	Filtering: focus on high-quality data, outlier removal Data harmonization: aligned to standards and normalization Dataset preparation: format conversion, shuffle, splitting
3 Model training Model definitions and their evaluation are purpose-specific. Architecture search and hyperparameter search are not.	 Architecture: layers and depth of the models is sampled Optimization: hyperparameters are sampled Training: architecture-hyperparameter combinations are trained Evaluation: purpose-specific scripts evaluate each model
4 Model export	• Optimization for client or server-side

Platform-specific export

model summary.tx optimization B transfer_learned_.png retention ti... 20 30 40 predicted RT Each row is one model training with its evaluation

Retention time prediction

The retention time model is calibrated to a specific dataset via refinement learning. The workflow automatically trains a base model and refines it on 4 external test datasets. Then, the refined model is evaluated on those datasets and compared to baseline models (Figure 3). We identified a model that is substantially smaller than other state-of-the-art models but similarly accurate.

Time for refinement learning on a + Prediction time for a human digest: <2 60-min run: <1 min (CPU) min (GPU)

Figure 4: retention time model size comparison



Fragmentation spectrum prediction

This year we integrated TMT and CID spectra and improved prediction accuracy for HLA peptides. At the same time the model size was reduced by a factor of 4 resulting in a 3-fold speed-up.

Data ingest

Spectra, search results and their relations are ingested to a data lake (AWS S3). Sequences are converted to the Proforma standard. Information is accessible interactively via Jupyter Notebooks or RStudio. All ingests are logged (see below).

Example: **ProteomeTools**² (synthetic peptide dataset)

Finished at Duration 05-04-2022 13:48:23 23h 49m 27s

Fit to screen - +

23h 49m 28s

Group by

Each row is one

RAW-search file.

is stored and

After completion it

accessible on AWS

- Original size: ~9TB
- Total ingest time: 23h +
- ~110M unique PSMs +

31.0s = 31.0s

3.0s a 3.0s

2m 52s 2m 52s 23.9s **2**3.9s

52.7s = 52.7s

49.7s = 49.7s

1m 50s = 1m 50s

25.7s 25.7s

1m 21s = 1m 21s

15.0s = 15.0s

2m 3s 🔳 2m 3s

15.2s 🔳 15.2s

51.2s = 51.2s

1m 27s = 1m 27s

42.0s = 42.0s

1m 7s 🔳 1m 7s

0.0s

start (1)

35038

35053

35067

35056

35055

35054

35065

35059

35066

35057

35068

35062

35060

35061

get_raw_search_mer.

-	Compressed size: ~890GB
-	AWS S3 cost: ~20\$ / month

~9M precursors

Figure 1: log of ProteomeTools search and spectra ingest to the data lake.

- + 15,852 RAW-search files
- + 1,5 min processing time per RAWsearch file.
- Parallelization in our cluster with 16 workers brings down processing time to <24h
- + > 700 RAW files per hour

Data preprocessing 2

Started at 05-03-2022 13:58:56

Started at 🤟 🔻

Figure 5: fragmentation model evaluation: a) Speed b) C-termini c) TMT



Model export 4

Models are exported for usage within client software (via C++ or Python-bindings) or remotely via tensorflow serving (gRPC). Models can be encrypted and are optimized for NVIDIA GPUs, Intel, AMD, or ARM CPUs.

+ Integrations

Example: fragmentation spectrum prediction

For fragmentation we annotate y-, b-, immonium, parent and the most frequent combinations of neutral loss ions with charges 1-3. Data is written to threcords files. The data is split in train, test, and validation while preventing duplication of peptides across splits.

v	start (1)	11.0s	Attempt 0					
	181364	6.3s	Attempt U					
~	merged_df_to_tfreco	9h 38m 56s						
	181365	15.5s	Task ID 181377	Step merged_df_to_tfrecords	Status Completed	Started at 05-16-2022 19:40:24	Finished at 05-16-2022 19:40:48	Duration 24.4s
	181367	11.9s	New York Control of Co					110050000
	181366	15.5s	Task details					\sim
	181370	16.1s						
	181368	8.4s	stdout					
	181378	23.2s						
	181371	12.1s	Search	Q				<u>دم</u> ج
	181375	5.9s	0 19:40:28 7	ask('IndestDataFlow/8	/det raw search me	arged/53184')		
	181372	10.3s	1 19:40:43 F	awfile: 6749 Search: {	3144	5190a, 55104 /		
	181373	8.9s	2 19:40:44 r	number of psms: #3667				
	181374	8.3s	3 19:40:49 f	ragmentation type: HCI)			
	181369	6.9s	4 19:40:49 n	mass analyzer: FTMS				
	181377	24.4s	5 19:40:49 c	collision energy: 35				
	181379	27.4s	6 19:40:49 t	frecords file: /mnt/da	ita/training/fragme	entation/6749_8144_FTMS	S_HCD_ce35.tfrecords	

Figure 2: log of the fragmentation data preprocessing. Running on a server with 32 workers takes ~5h for the 15,852 RAW-search files of ProteomeTools

Fragmentation and retention time models generated by this workflow are integrated into the intelligent search algorithm CHIMERYS[®], a software node in Thermo Scientific™ Proteome Discoverer™ (PD). A fragmentation model is integrated in INFERYS[®] spectral library generation (in PD) and INFERYS Rescoring (PD Node).

+ Related Content

Poster IM-PA-019: A unifying, spectrum-centric approach for the analysis of peptide tandem mass spectra (Abstract #534) Poster Session A – Monday 29th and Tuesday 30th August 2022

https://www.msaid.de/conferences/imsc2022

References

¹The et al 2016. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0 ²Zolg, Wilhelm et al 2017. Building ProteomeTools based on a complete synthetic human proteome ³Gessulat, Schmidt et al 2019. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning

CHIMERYS®, INFERYS® and MSAID® are registered trademarks of MSAID GmbH. Thermo Scientific™ Proteome Discoverer™ software is a trademark of Thermo Fisher Scientific Inc.