UNIVERSITY OF WARSAW | INSTITUTE OF GEOPHYSICS

NATIONAL SCIENCE CENTRE POLAND

NAWA POLISH NATIONAL AGENCY FOR ACADEMIC EXCHANGE

inoe 2000

Methodology of automatic classification of plant pollen using machine learning

Artur Tomczak[1], J. Vasilescu[2], I.S. Stachlewska[1] , Ł. Janicka [1]

[1] Faculty of Physics, University of Warsaw (UW), 02-093 Warsaw, Poland
[2] National Institute of Research and Development for Optoelectronics INOE 2000, 077125 Magurele, Romania

Remote Sensing Laboratory

Contact emails: *artur.tomczak@fuw.edu.pl*
iwona.stachlewska@fuw.edu.pl

Seminar, 14 March 2025, Warsaw University, Warsaw

Motivation and objectives

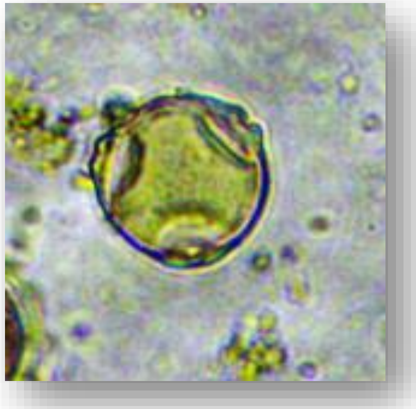
Motivation: *Health impact of allergenic pollen.*

Objective: *Investigate allergenic pollen in a real time over an urban continental site.*

Innovation: *open-source (free) and flexible algorithm that allows to investigate, train and predict pollen taxa.*

Sensor Rapid-E

Betula **pollen grain**



Pine pollen grain

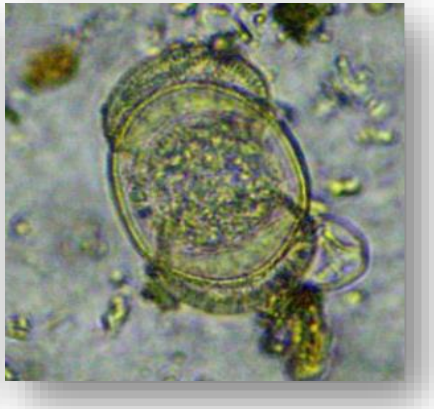
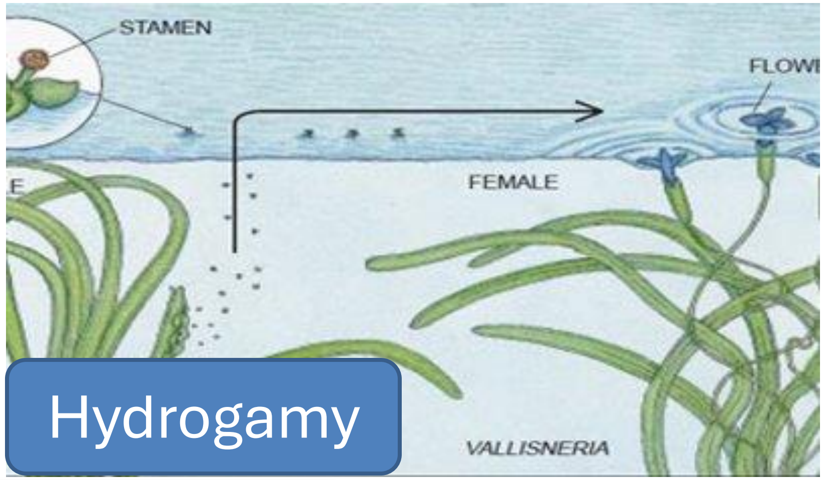


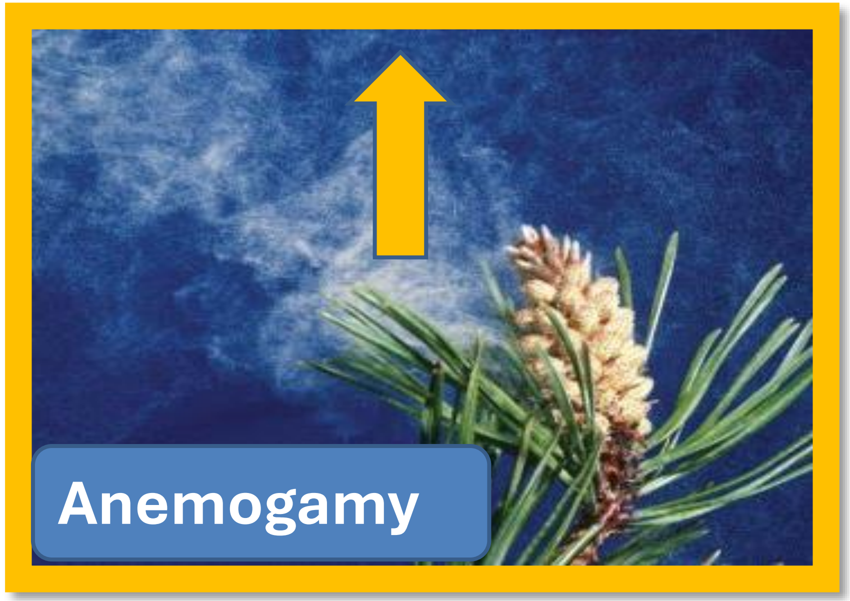
Photo provided by Zuzanna Rykowska

Why it is so challenging

Zoogamy



Hydrogamy

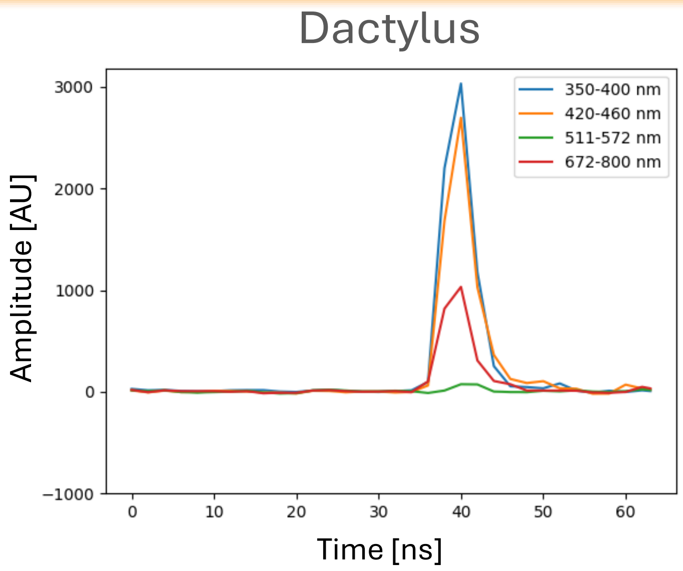


Anemogamy

- Atmospheric dynamics
- Optical properties
- Microphysical properties

Rapid-E sampler

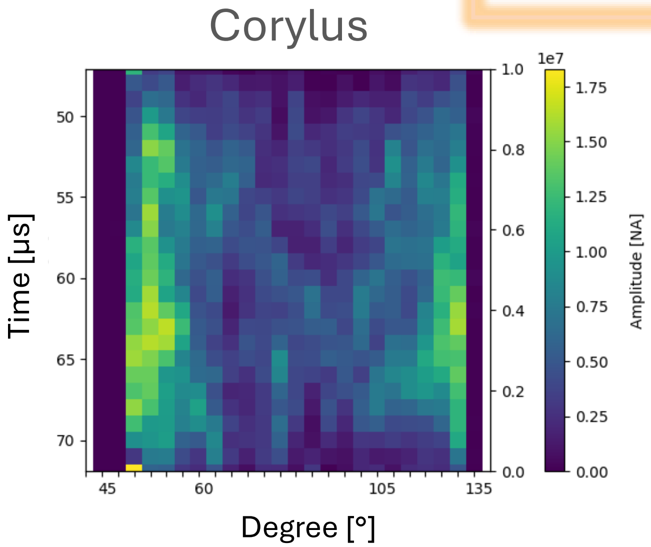
- Ranges:
  - 350-400 nm
  - 420-460 nm
  - 511-572 nm
  - 672-800 nm



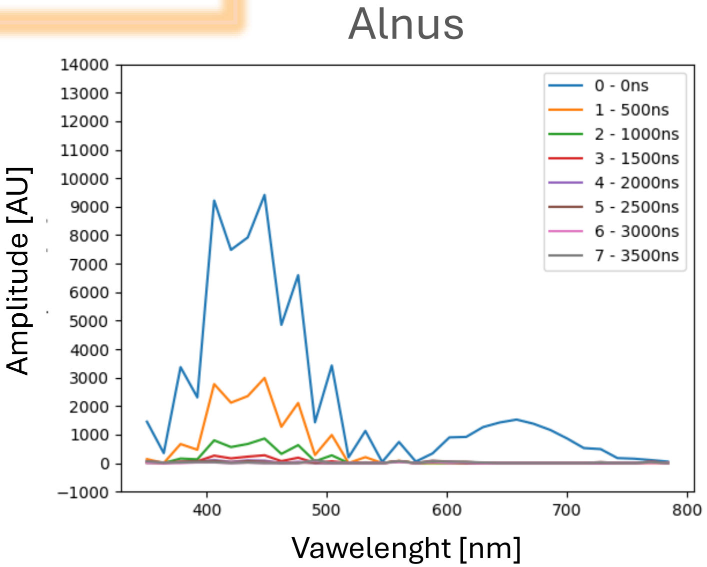
Fluorescence lifetime

Sensor Rapid-E

Scattering

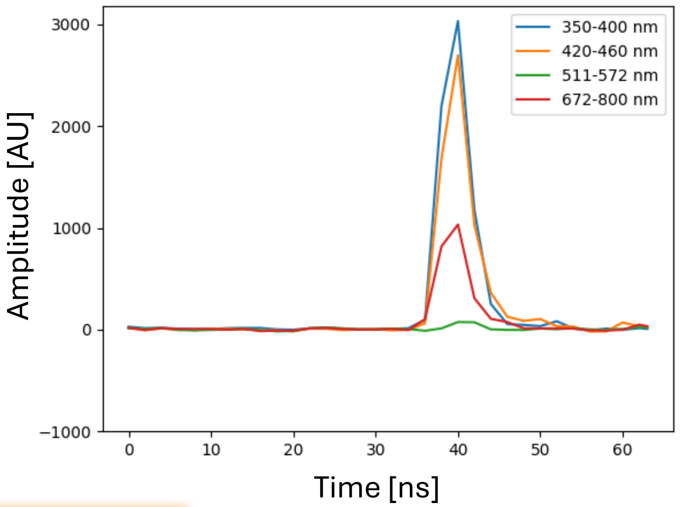


Spectral ranges



Rapid-E sampler

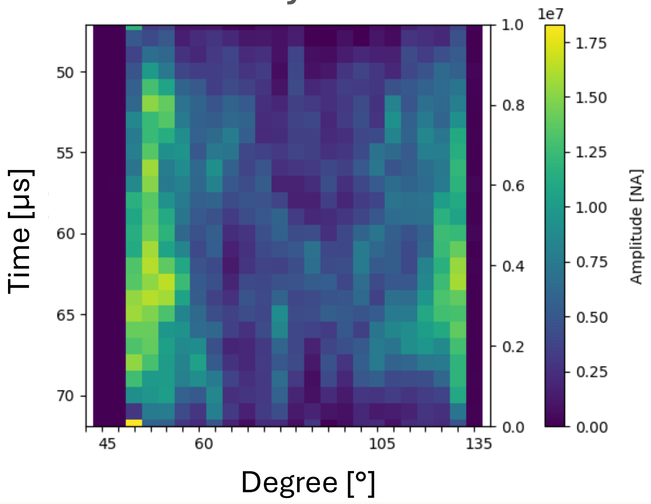
- Ranges:
  - 350-400 nm
  - 420-460 nm
  - 511-572 nm
  - 672-800 nm



Dactylus

Sensor Rapid-E

Scattering



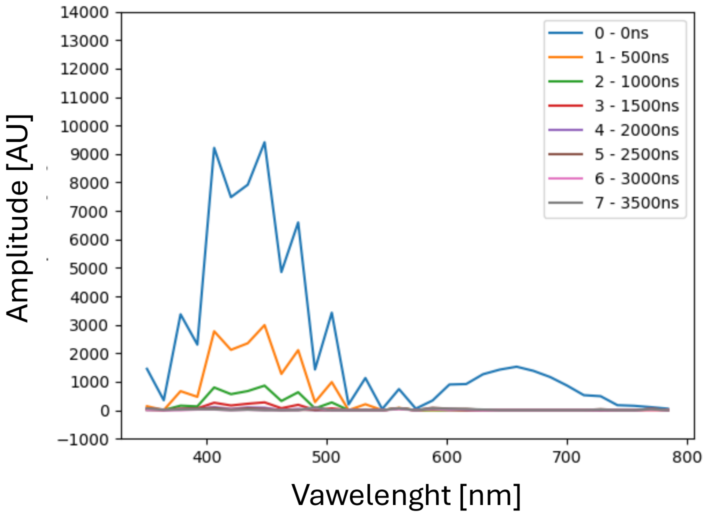
Corylus

Fluorescence lifetime

Degree:

- 45-135 °

Spectral ranges

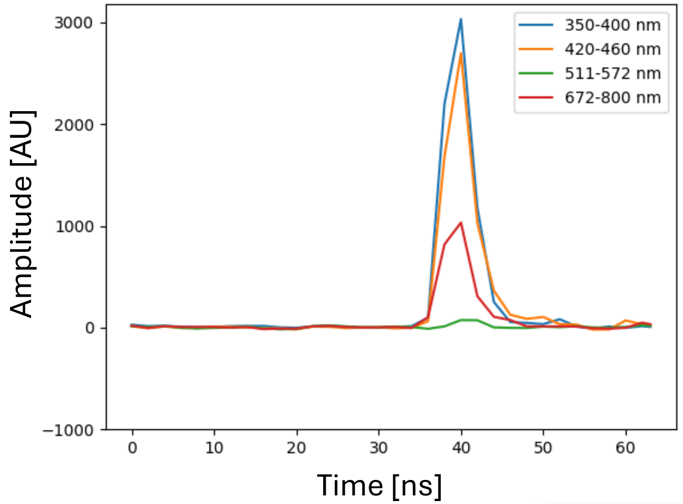


Alnus

Rapid-E sampler

- Ranges:
  - 350-400 nm
  - 420-460 nm
  - 511-572 nm
  - 672-800 nm

Dactylus

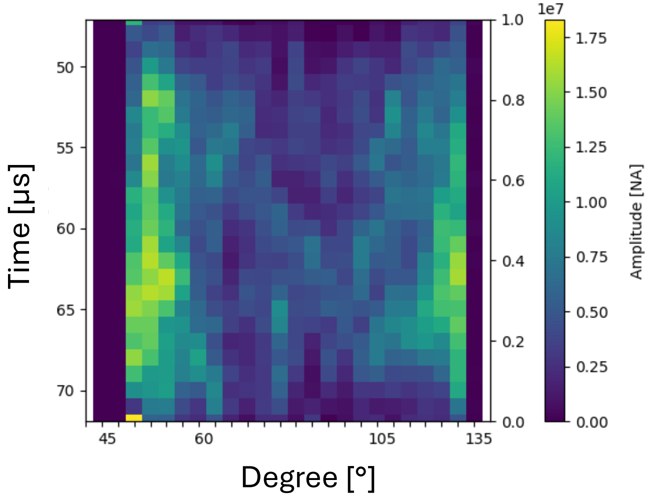


Sensor Rapid-E

Fluorescence lifetime

Scattering

Corylus



Degree:

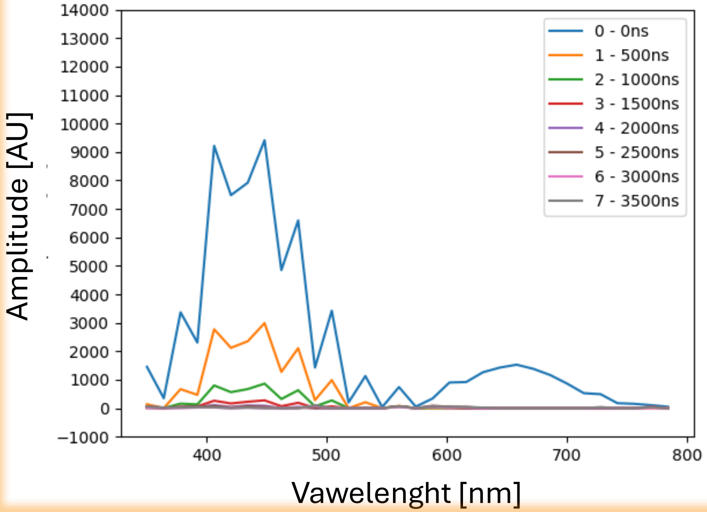
- 45-135 °

Sampling

- 8 x 500 ns

Spectral ranges

Alnus



Best model identification

Pollen type:

1. Alnus 🌳
2. Arrhenatherum 🌾
3. Broussonetia 🌳
4. Corylus 🌳
5. Dactylus 🌾
6. Fraxinus 🌳
7. Juglans 🌳
8. Lolium perenne 🌾
9. Morus 🌳
10. Pinus nigra 🌳
11. Platanus 🌳
12. Populus alba 🌳
13. Quercus 🌳
14. Taxus 🌳

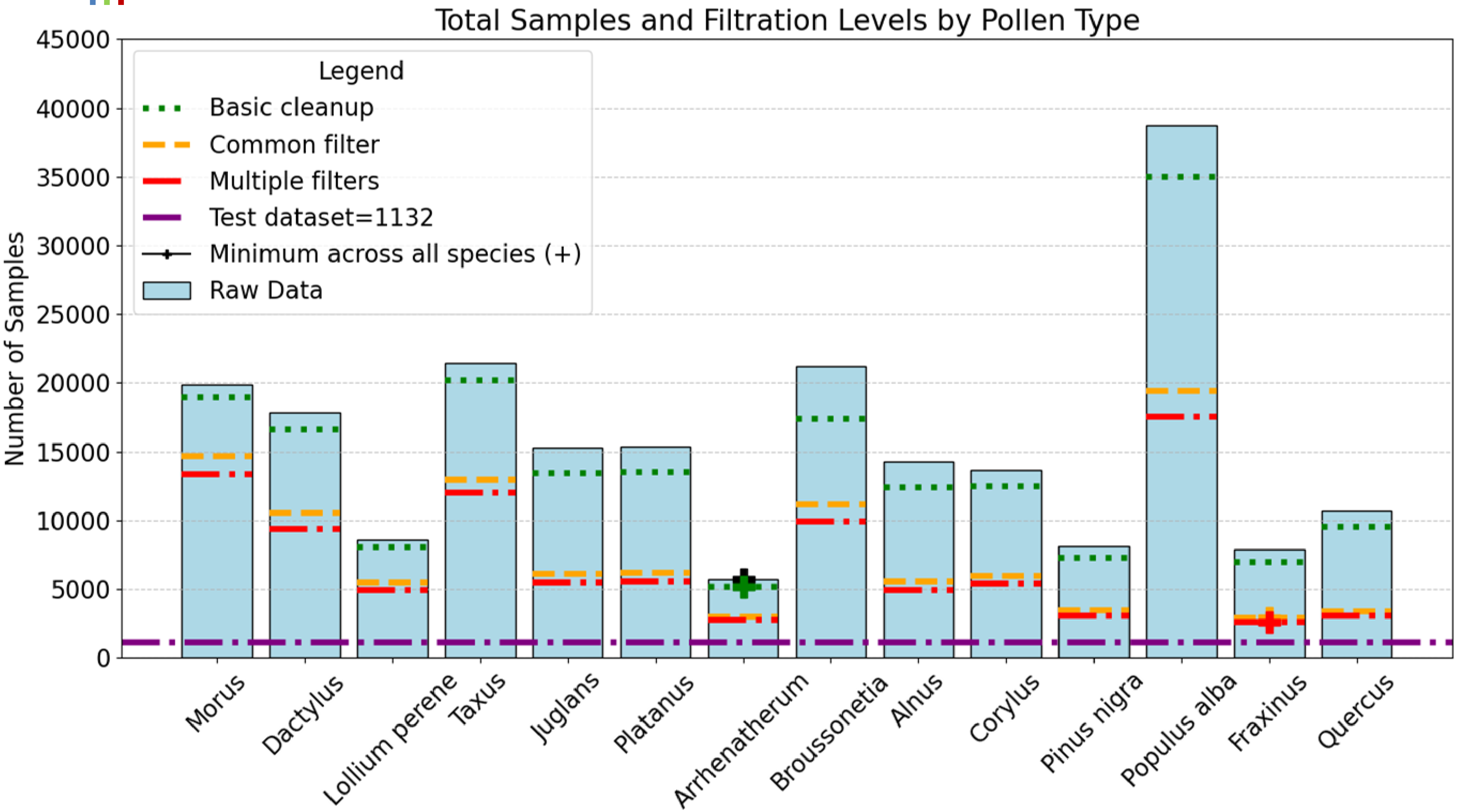
Plants surrounding INOE, collected by Boldeanu M.

4 data filtration types selection

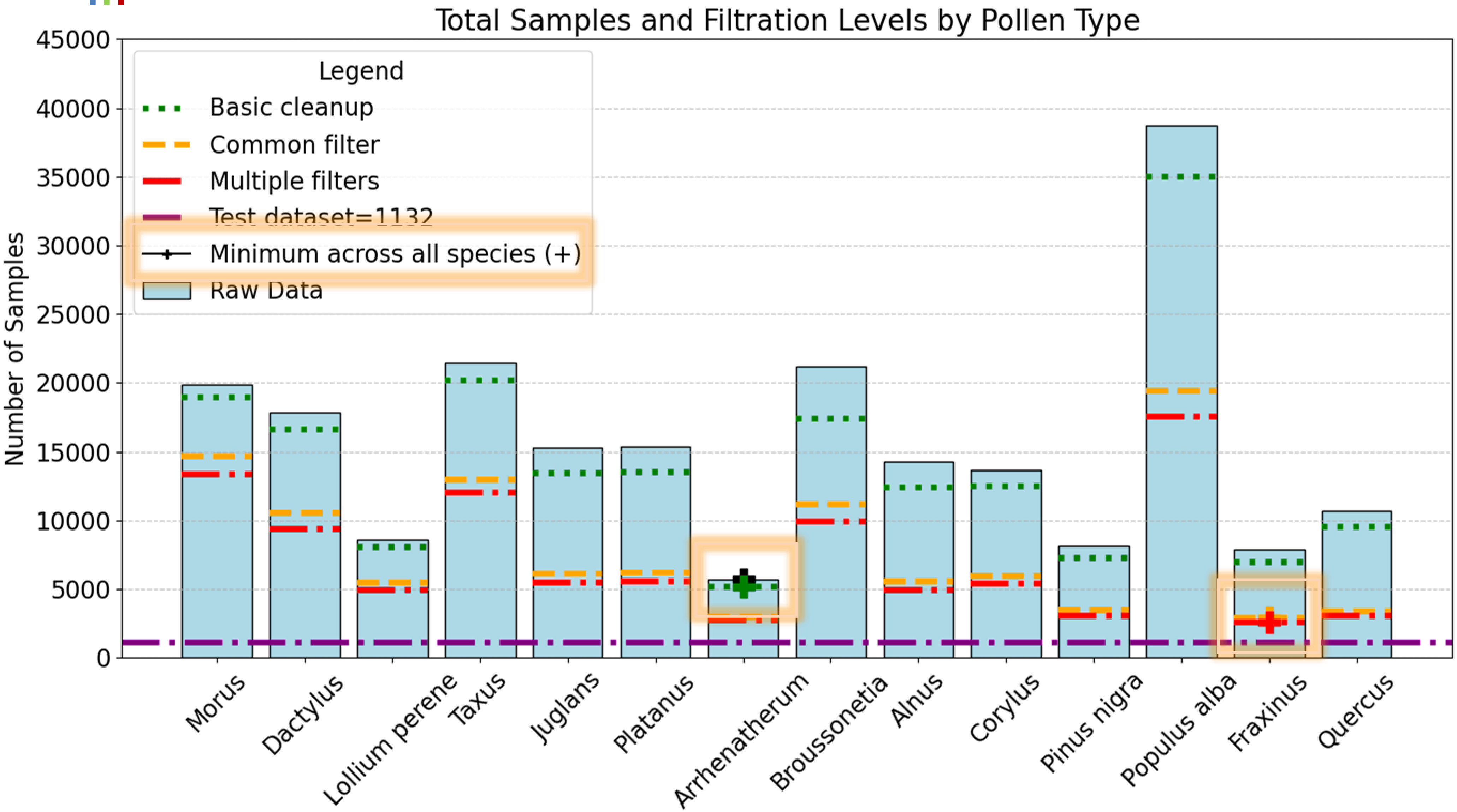
1. Raw data
2. Basic cleanup
3. One common filter
4. Pollen types divided into filtration groups

Same architecture for all models

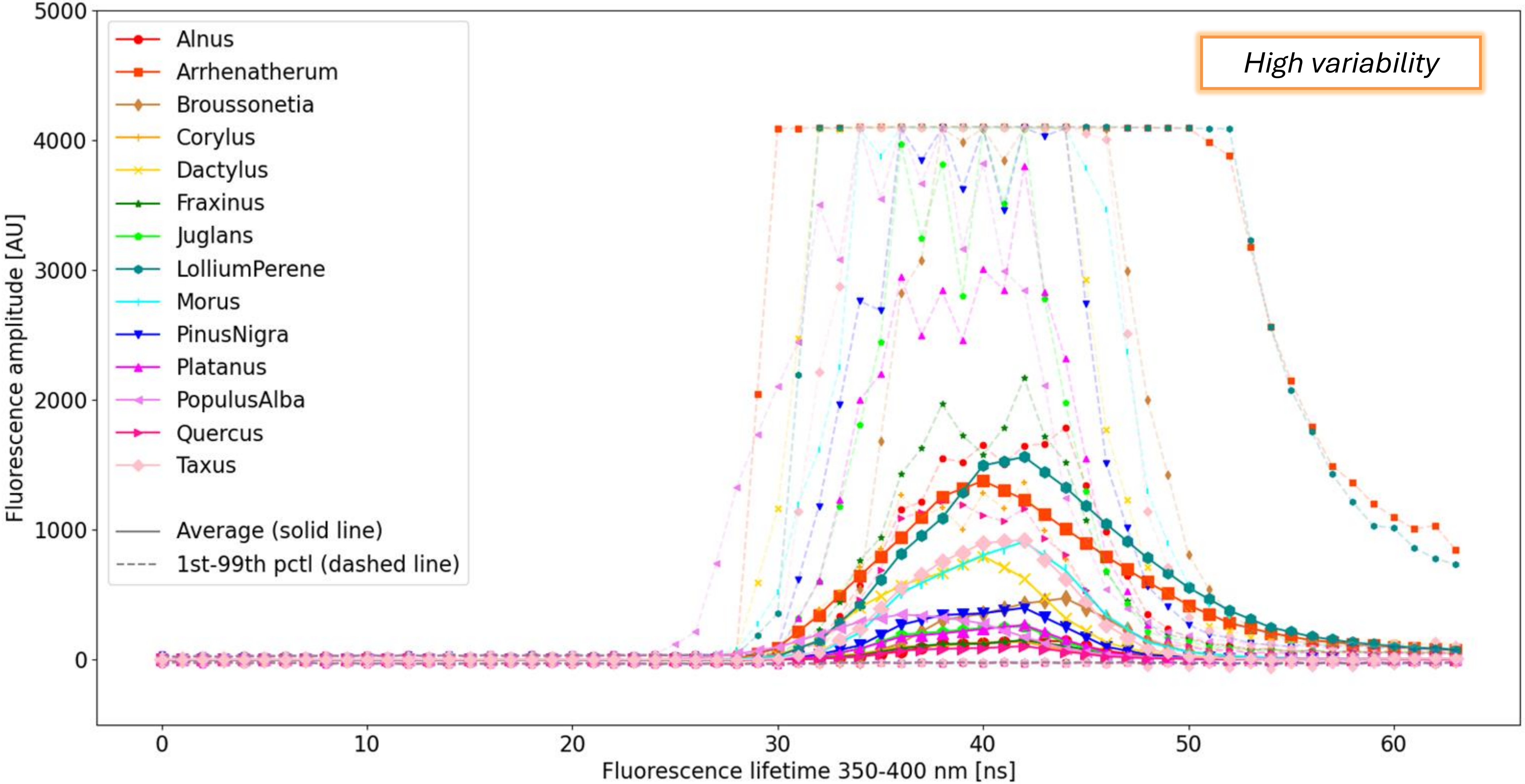
Available dataset



Available dataset

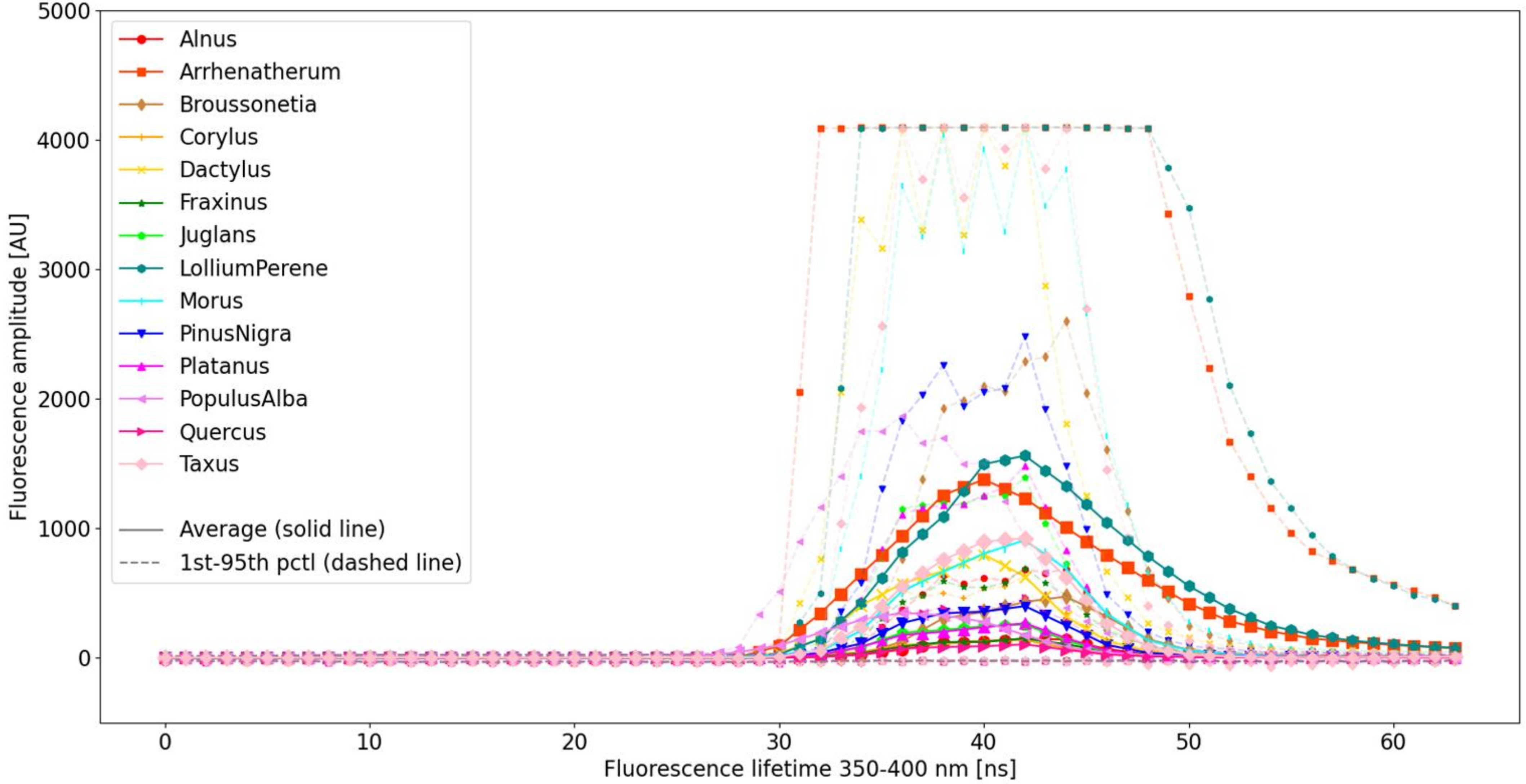


F. lifetime – average and boundaries

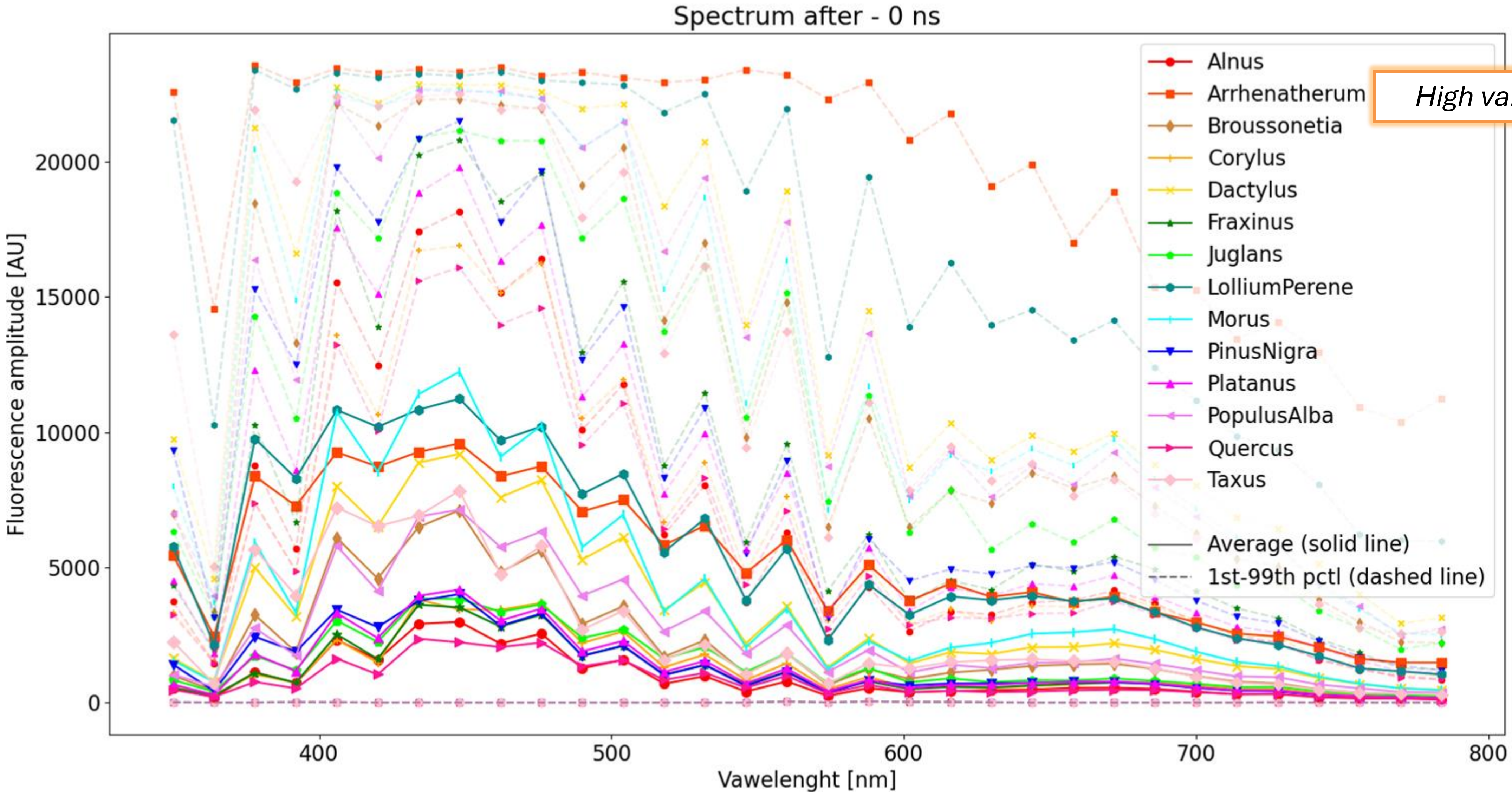


High variability

F. lifetime – average and boundaries

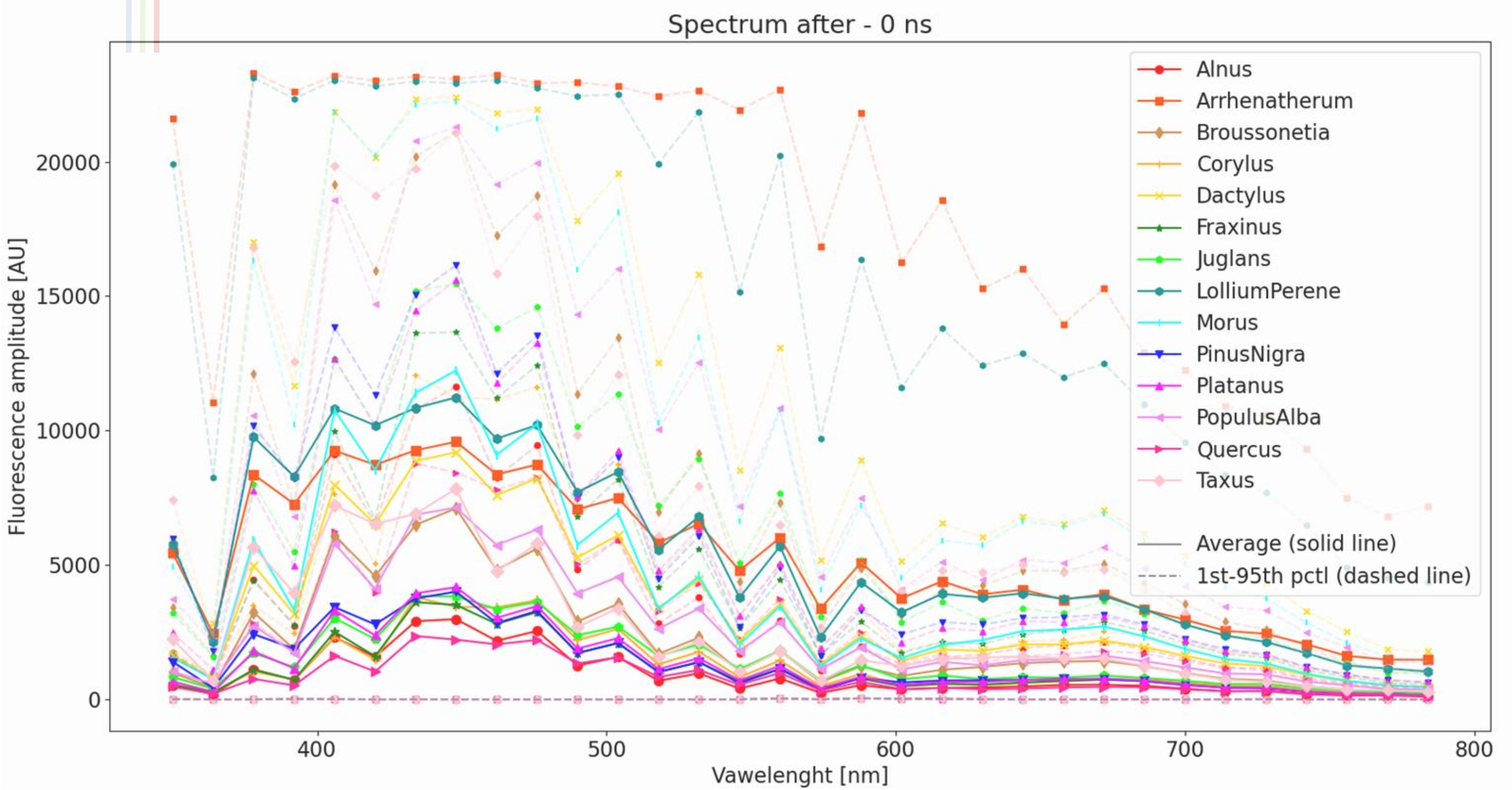


F. spectra – average and boundaries

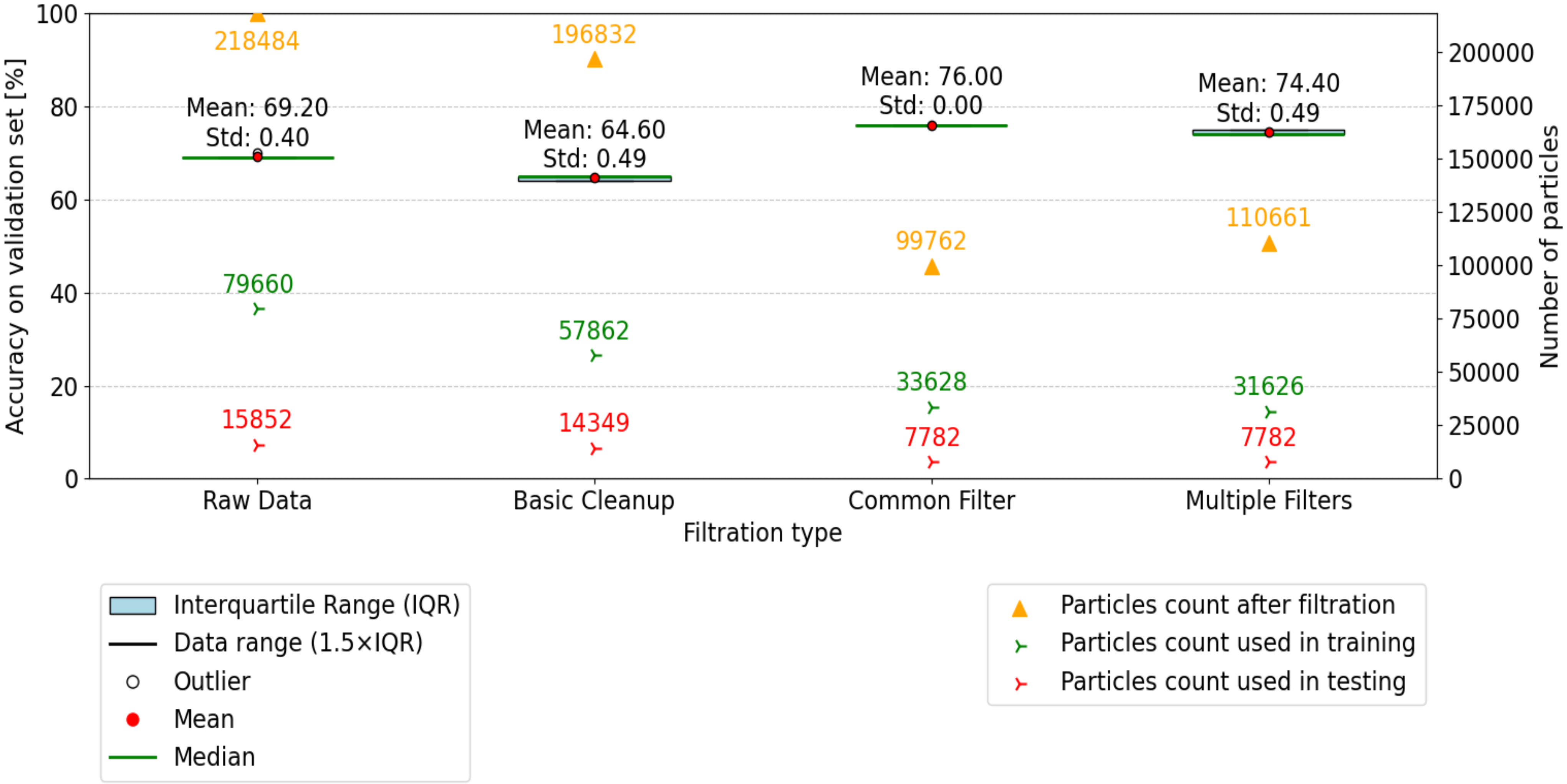


High variability

F. spectra – average and boundaries

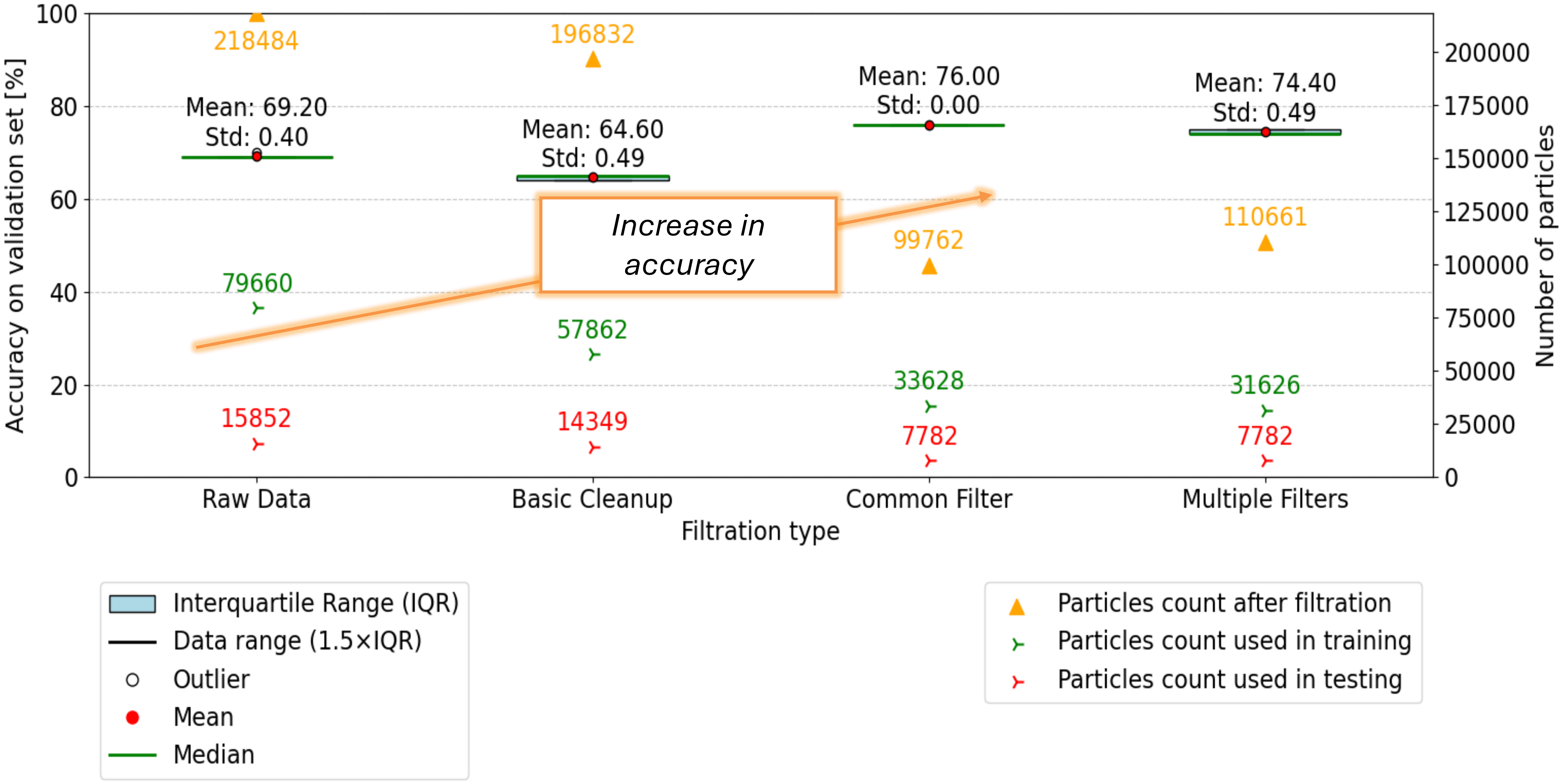


Model summary after filtration



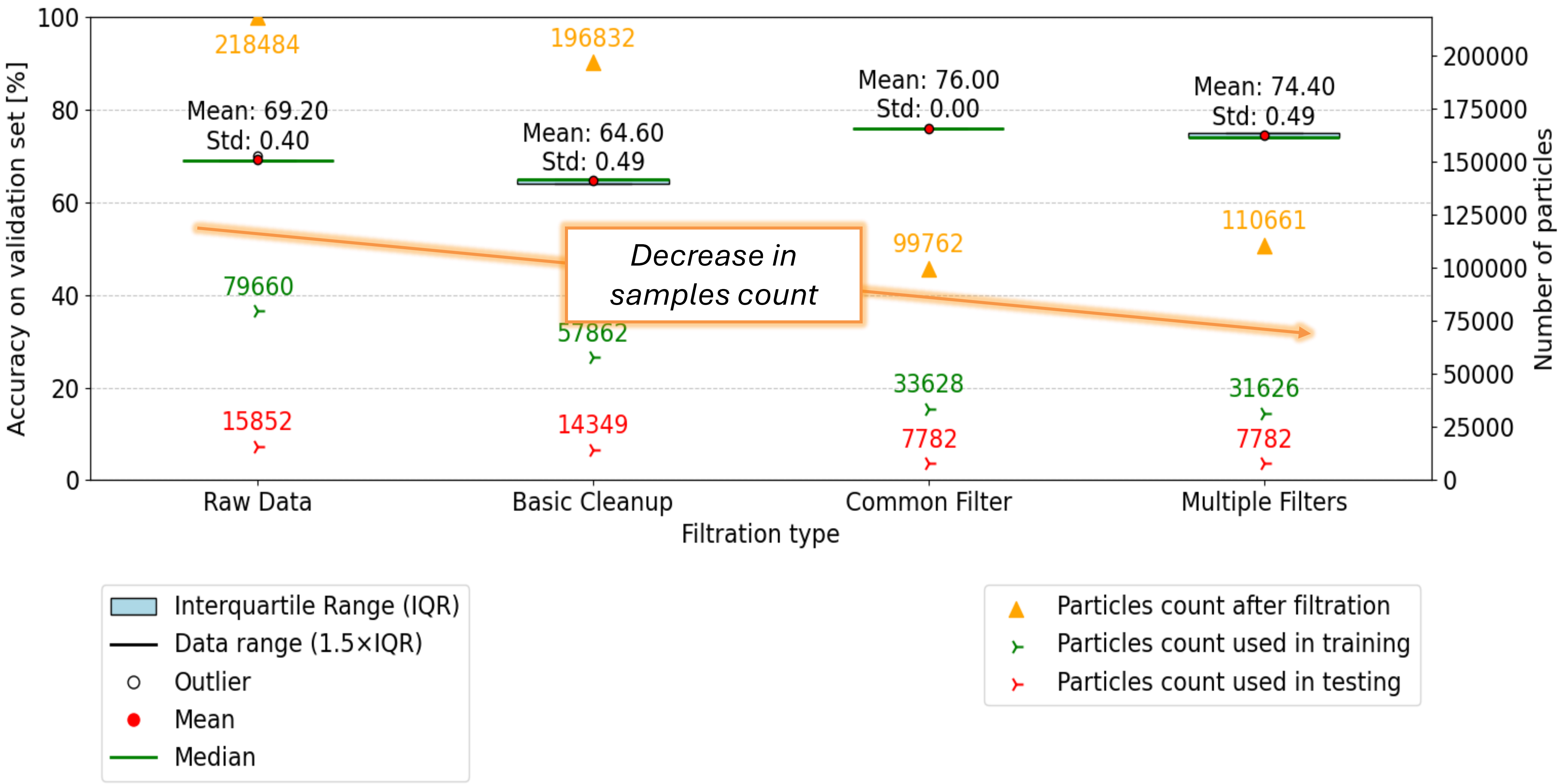
Model accuracy based on different filtration types. An average after 5 trials. Raw data represents the total number of samples

Model summary after filtration



Model accuracy based on different filtration types. An average after 5 trials. Raw data represents the total number of samples

Model summary after filtration



Decrease in samples count

Model accuracy based on different filtration types. An average after 5 trials. Raw data represents the total number of samples

ML architecture

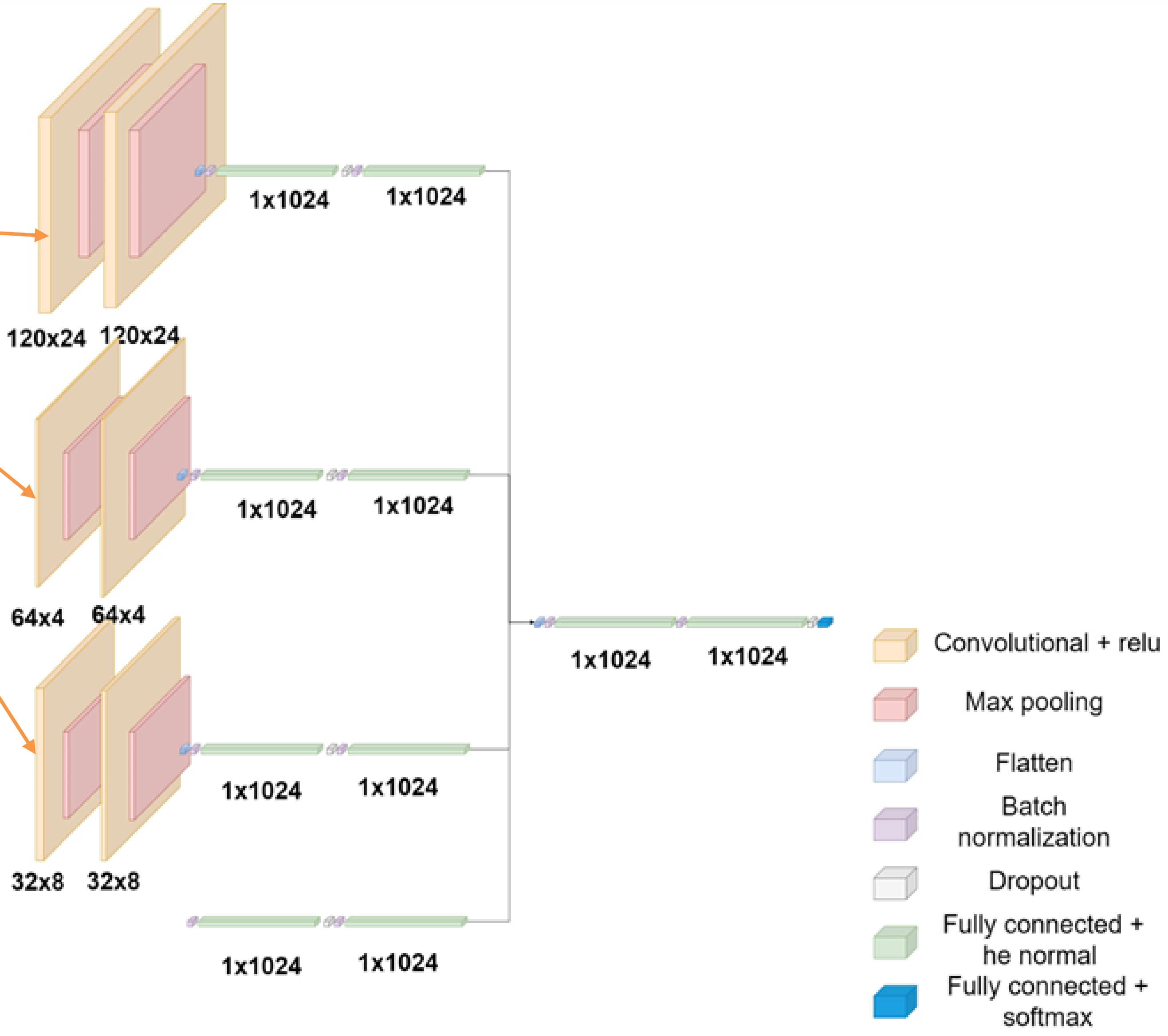
Convolutional layer

Scattering

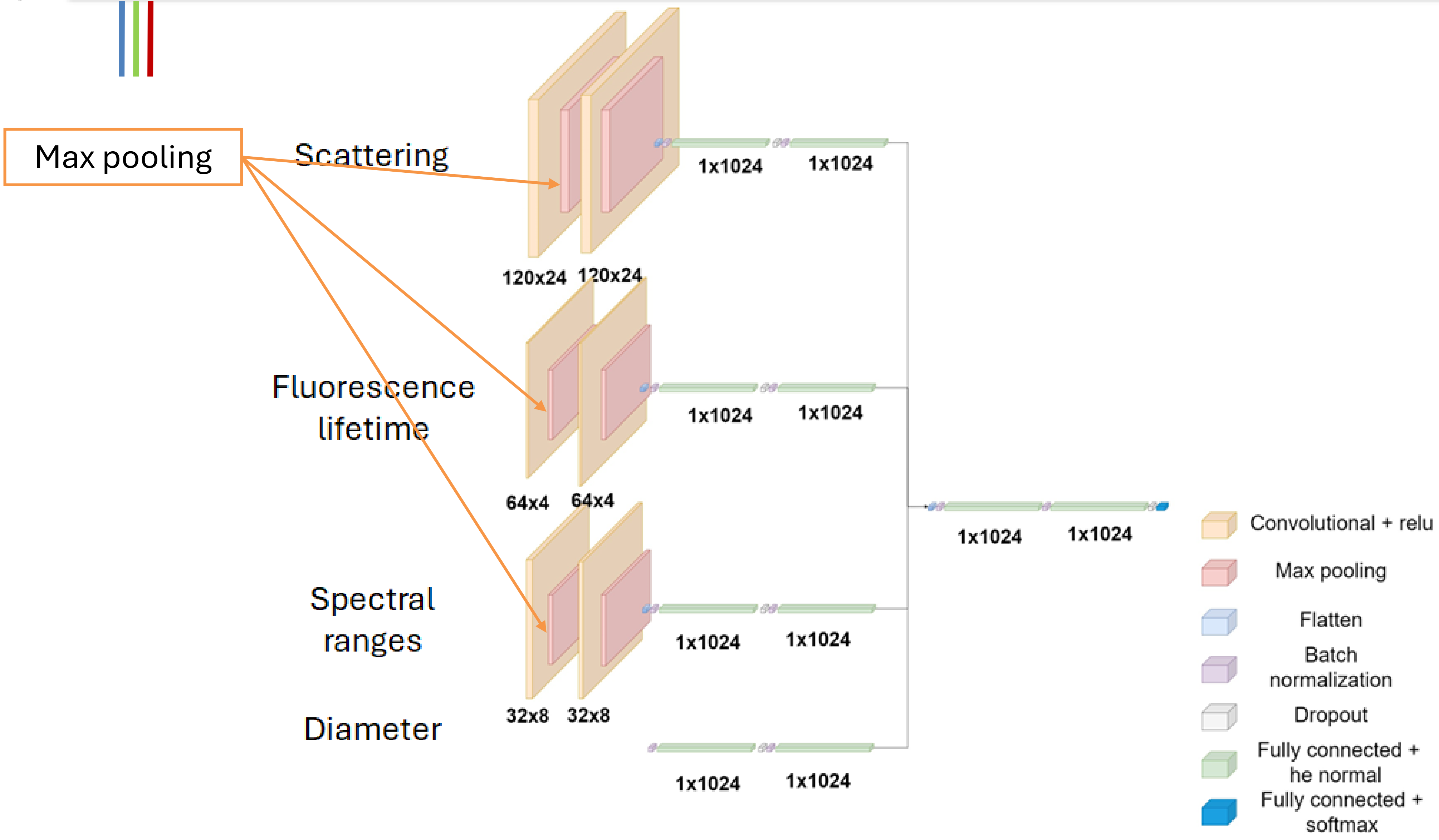
Fluorescence lifetime

Spectral ranges

Diameter



Architecture



Max pooling

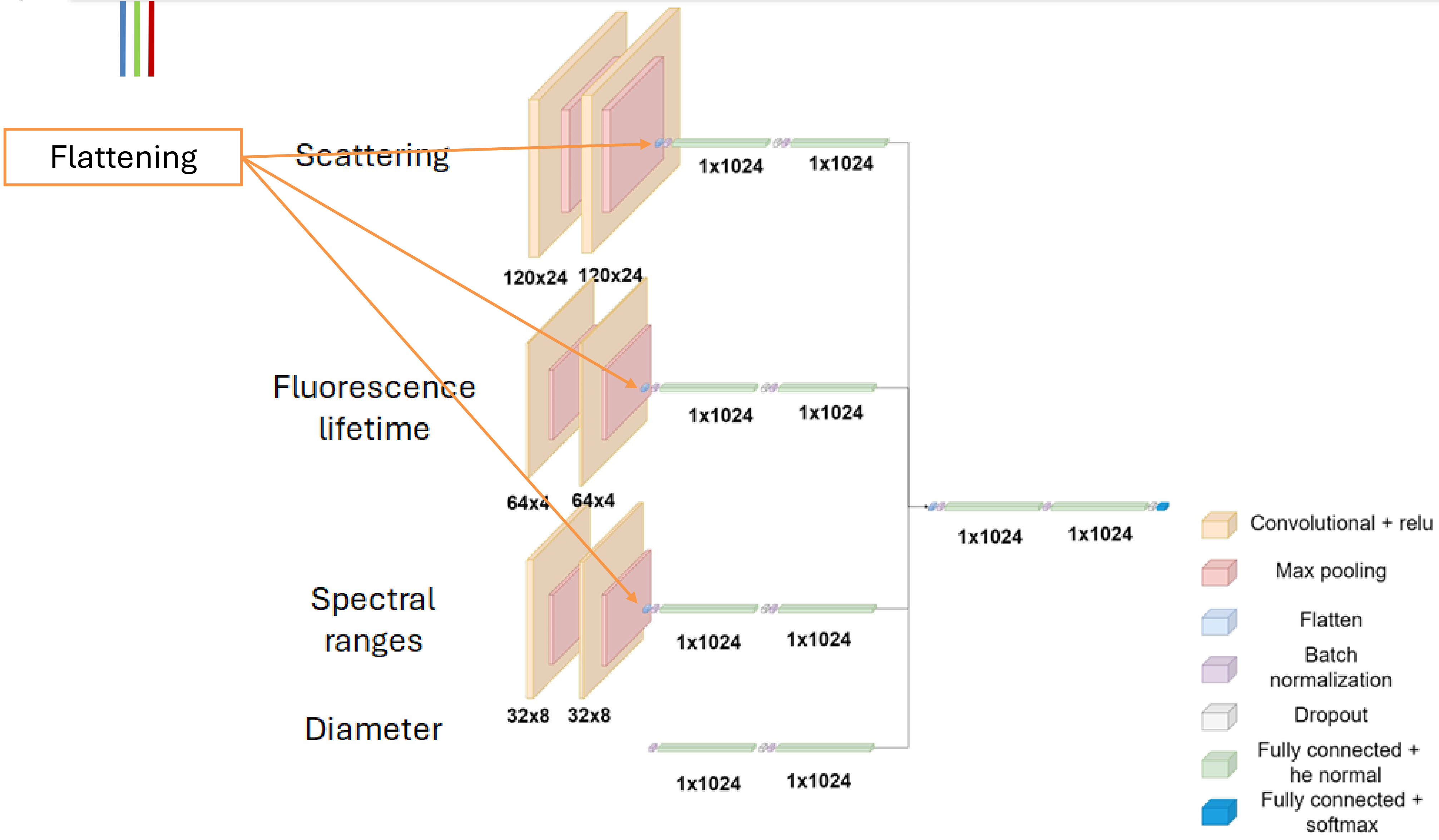
Scattering

Fluorescence lifetime

Spectral ranges

Diameter

Architecture



Flattening

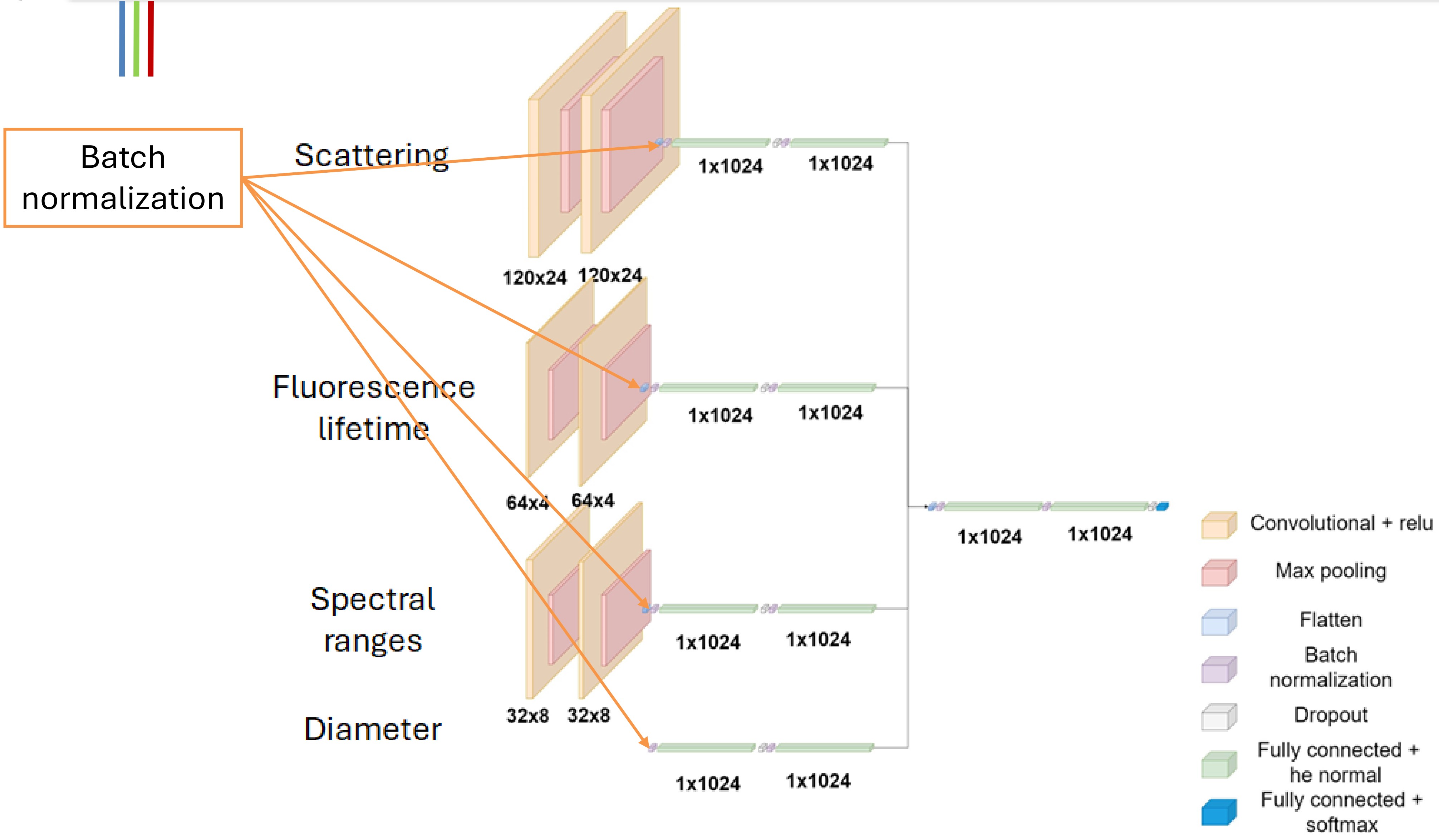
Scattering

Fluorescence lifetime

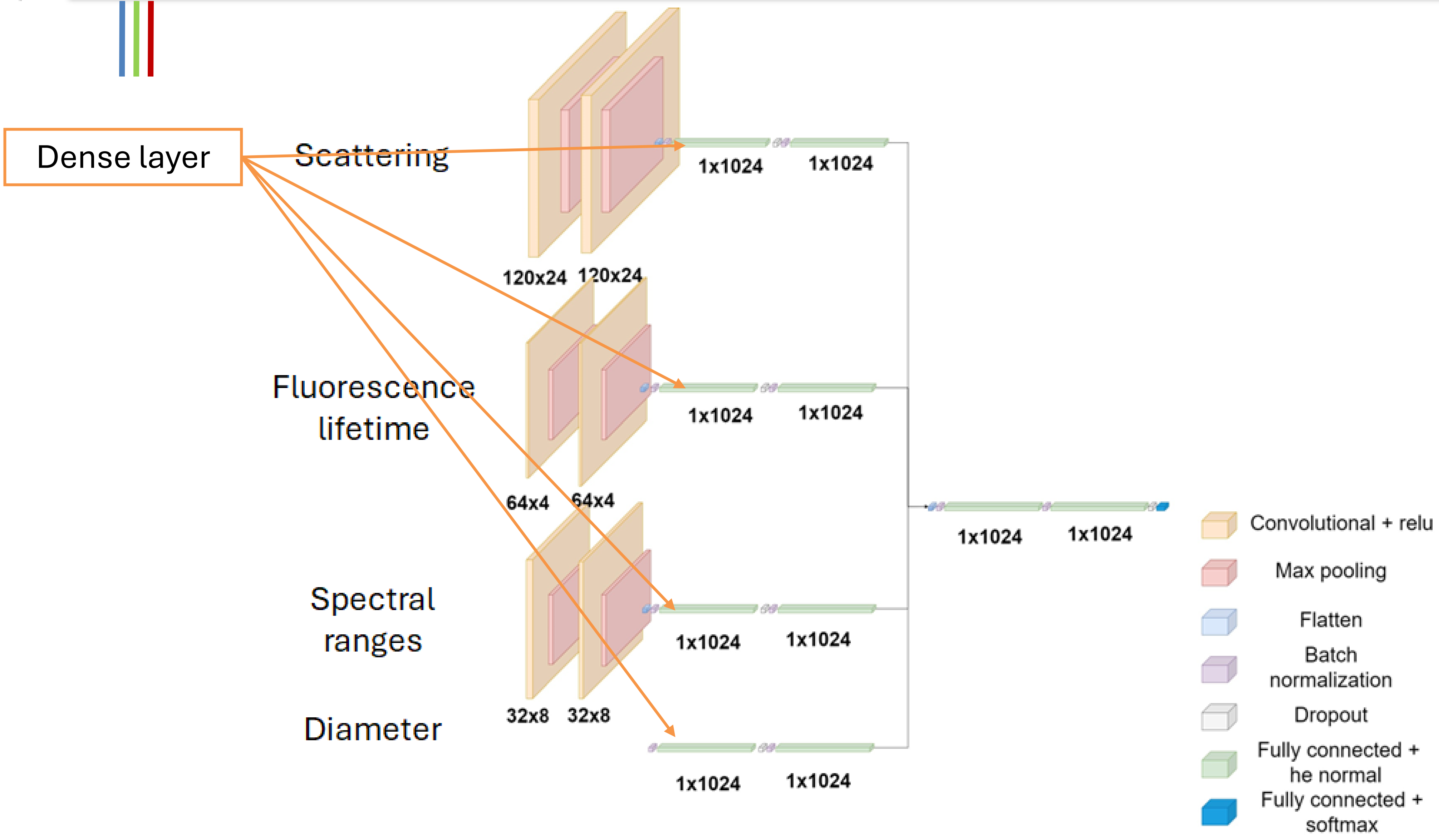
Spectral ranges

Diameter

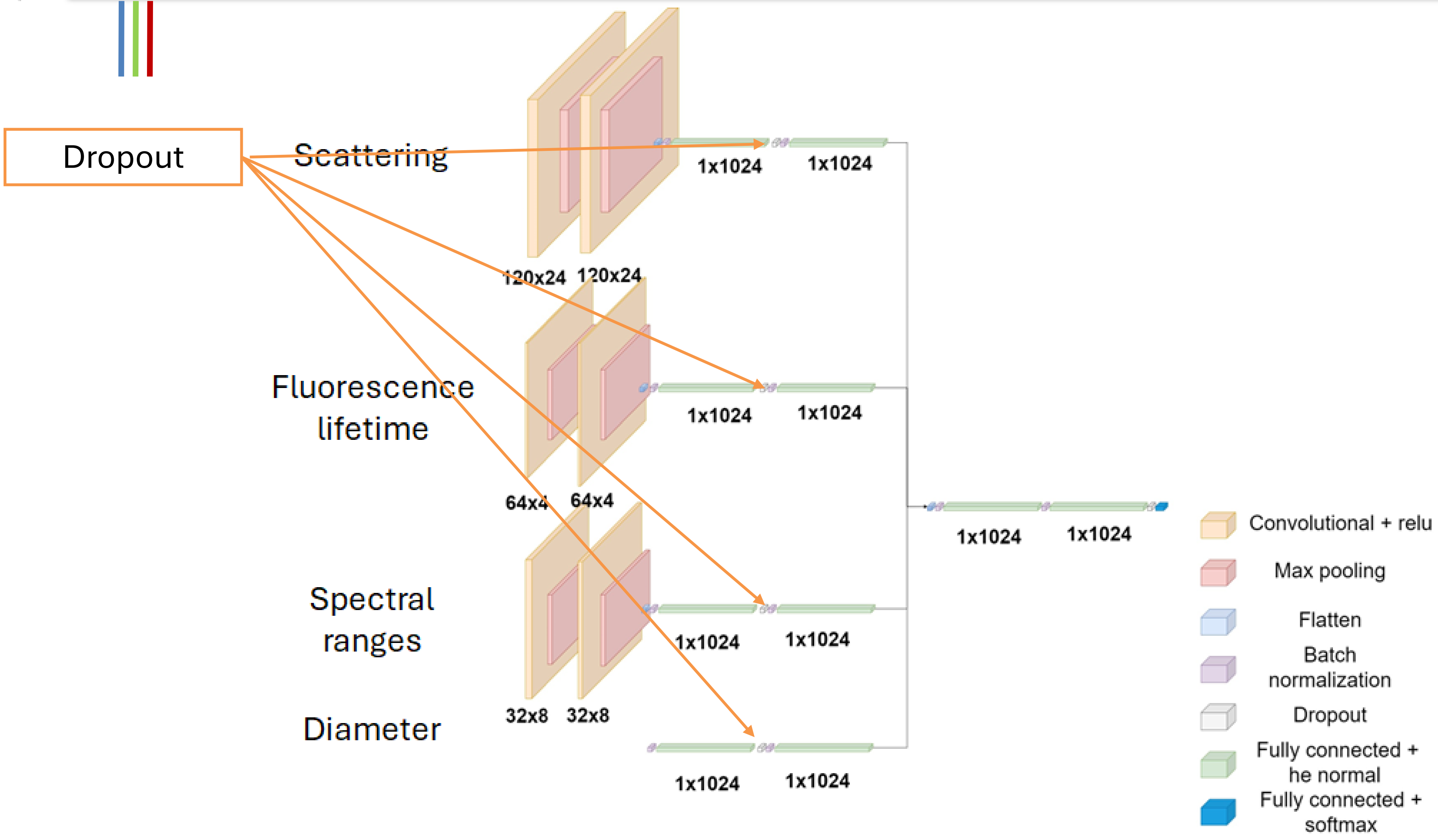
Architecture



Architecture

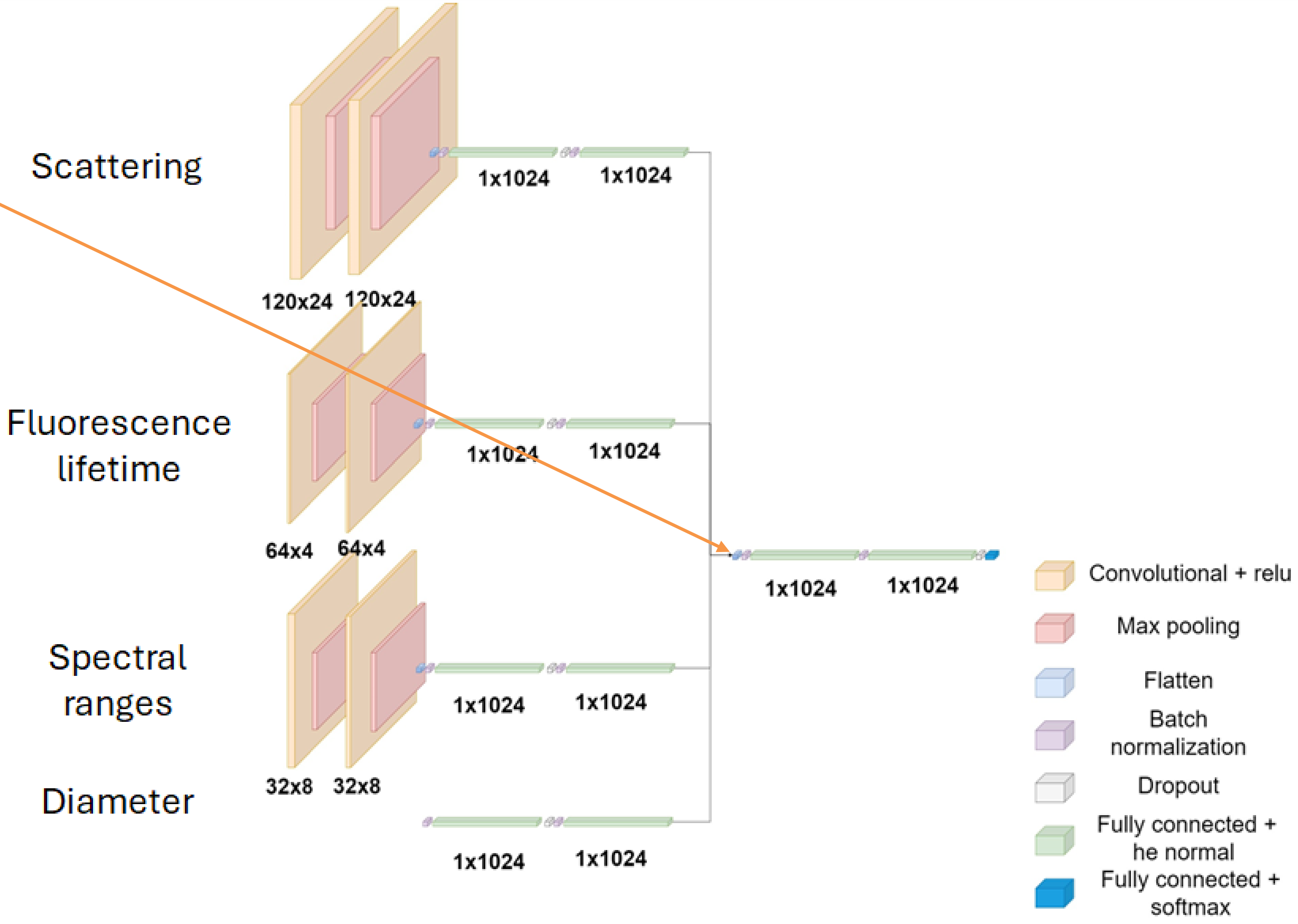


Architecture

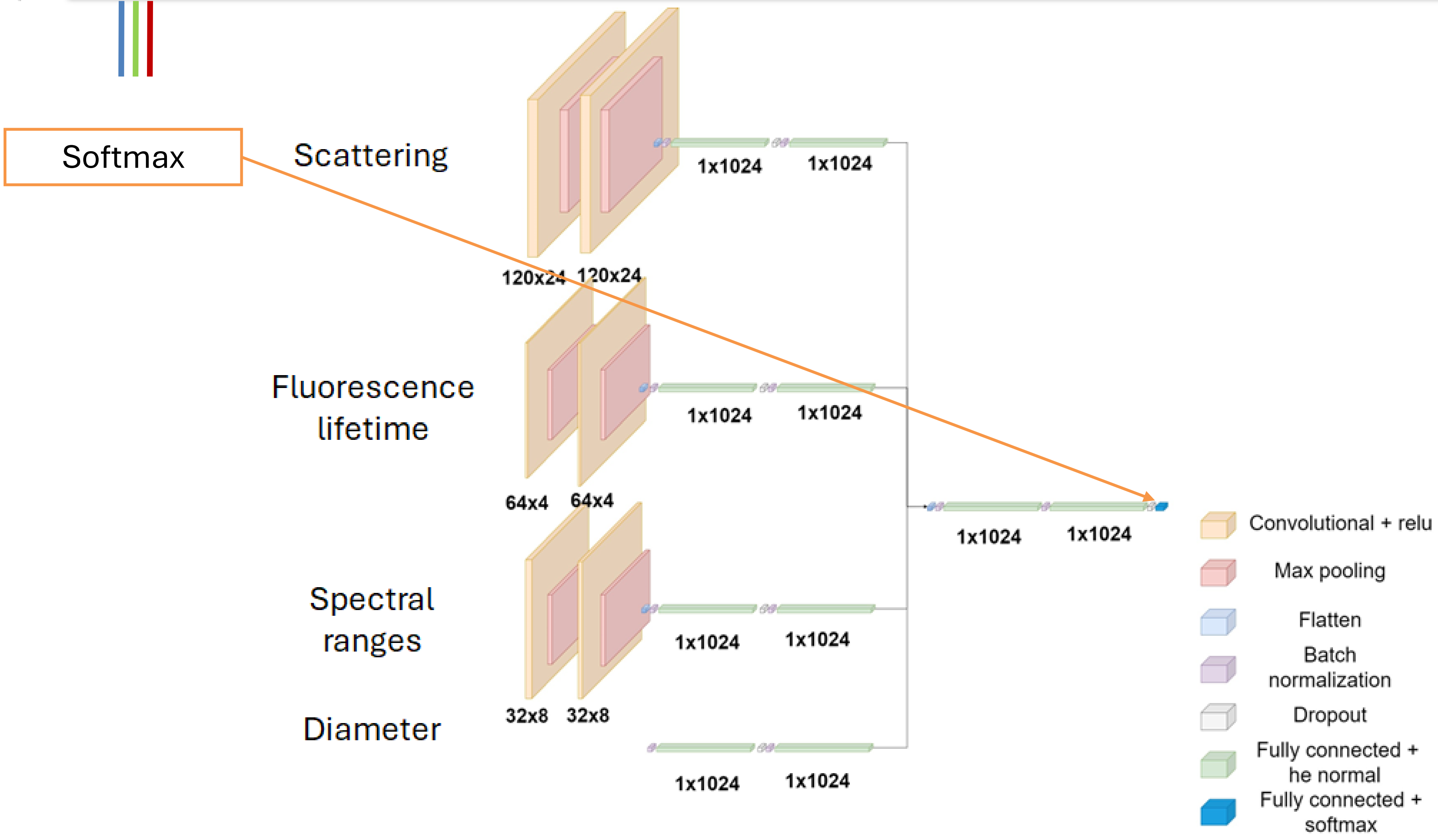


Architecture

Layers concatenation



Architecture



Model identification

Comparison of:

1. Precision and recall vs threshold
2. Total accuracy
3. Identification count vs threshold
4. ROC curve
5. Accuracy vs threshold
6. Confusion matrix

Model identification

Comparison of:

1. Precision and recall vs threshold
2. Total accuracy
3. Identification count vs threshold
4. ROC curve
5. Accuracy vs threshold
6. Confusion matrix

$T_p - true\ positived$
$T_n - true\ negatives$
$F_p - false\ positives$
$F_n - false\ negatives$
$\mathrm{N} - samples\ count$
$p_o - threshold$

$$f(p_o, precission, recall)$$

$$\text{precission} = \frac{T_p}{T_P + F_n}$$

$$recall = \frac{T_p}{T_p + F_p}$$

Model identification

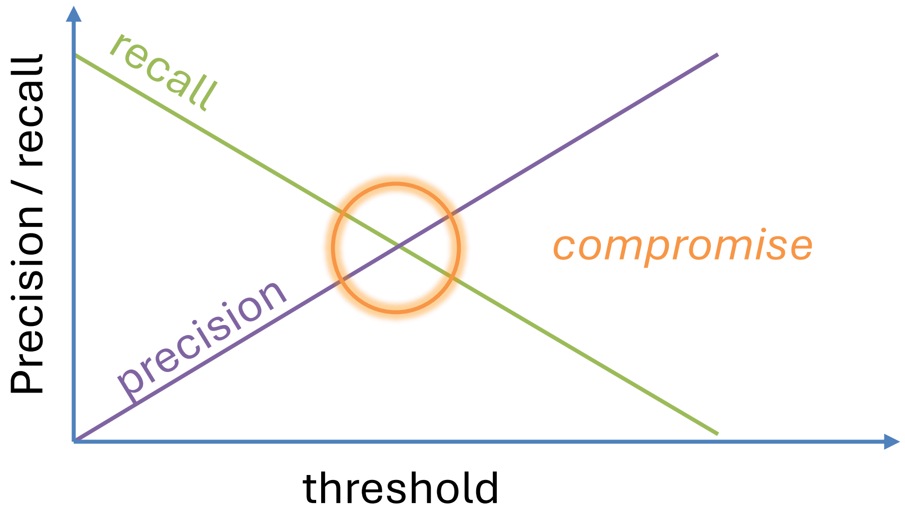
Comparison of:

1. Precision and recall vs threshold
2. Total accuracy
3. Identification count vs threshold
4. ROC curve
5. Accuracy vs threshold
6. Confusion matrix

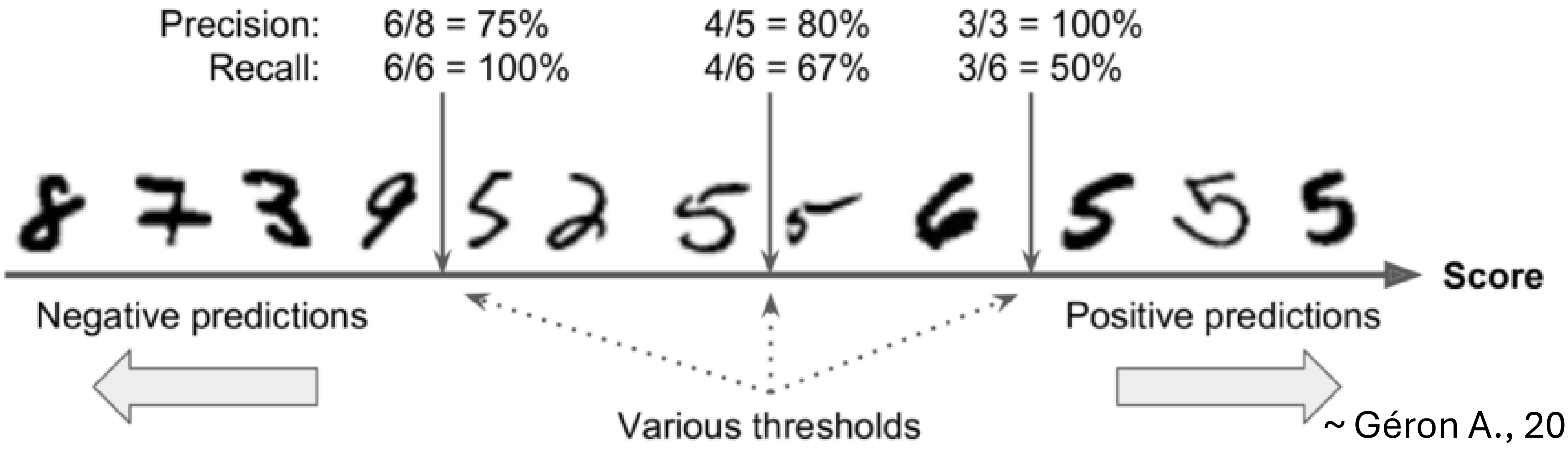
$T_l - true\ label$
$p_l - predicted\ label$

$f(pl, Tp)$

Precision-recall - interpretation



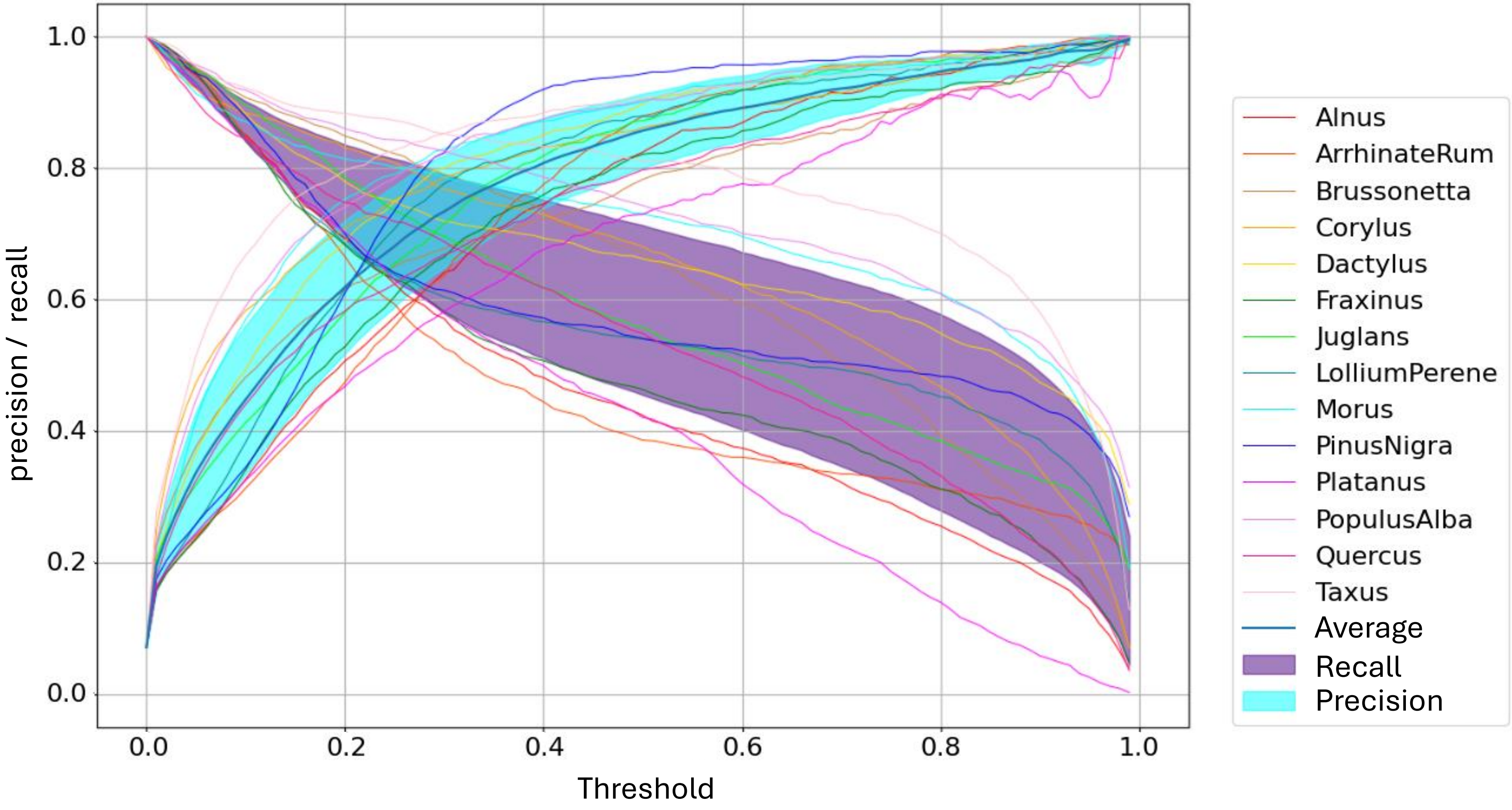
The higher precision and recall with lower threshold the better



~ Géron A., 2022

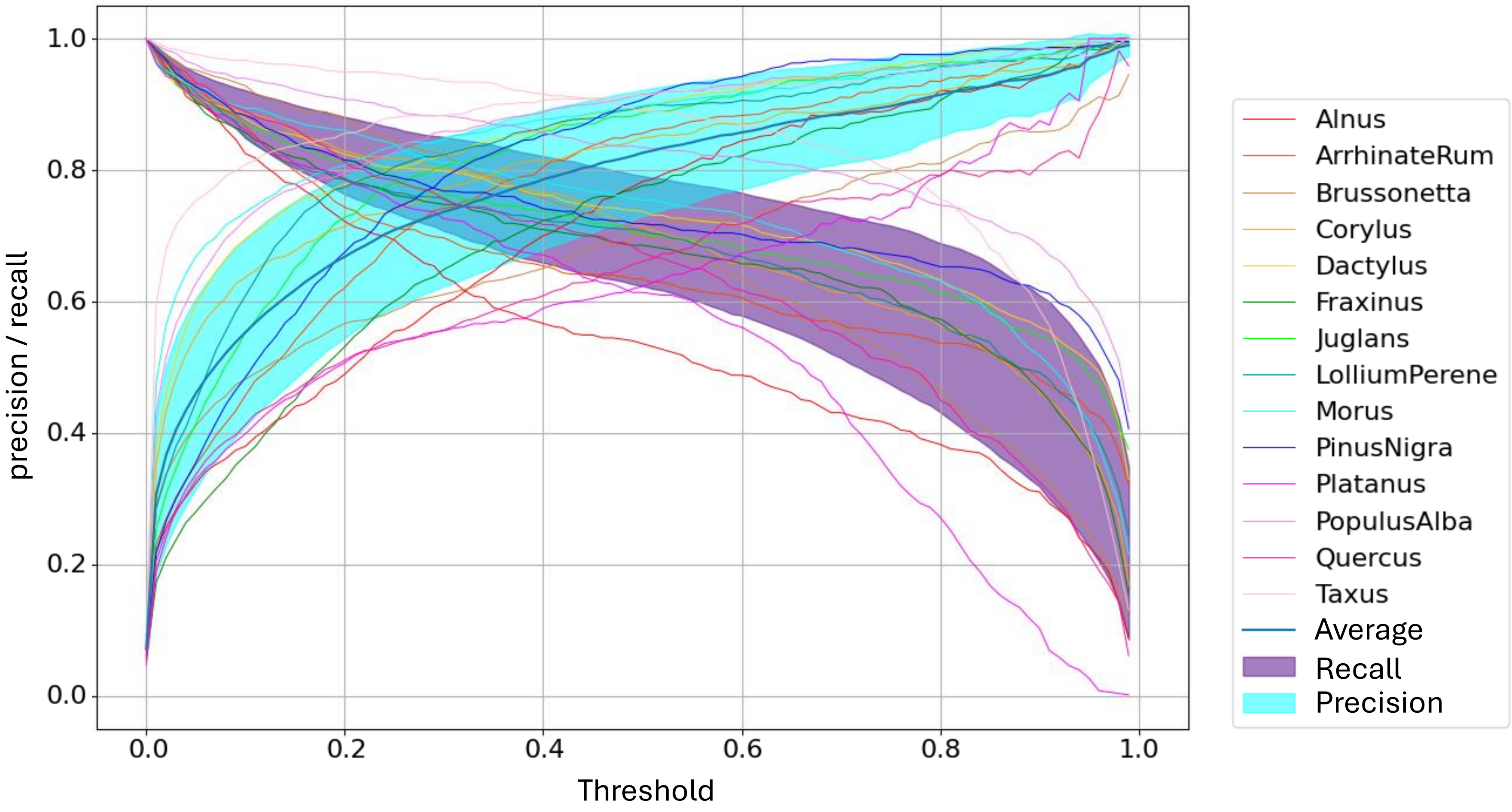
Precision and recall vs threshold

*Precision and recall – **raw data***



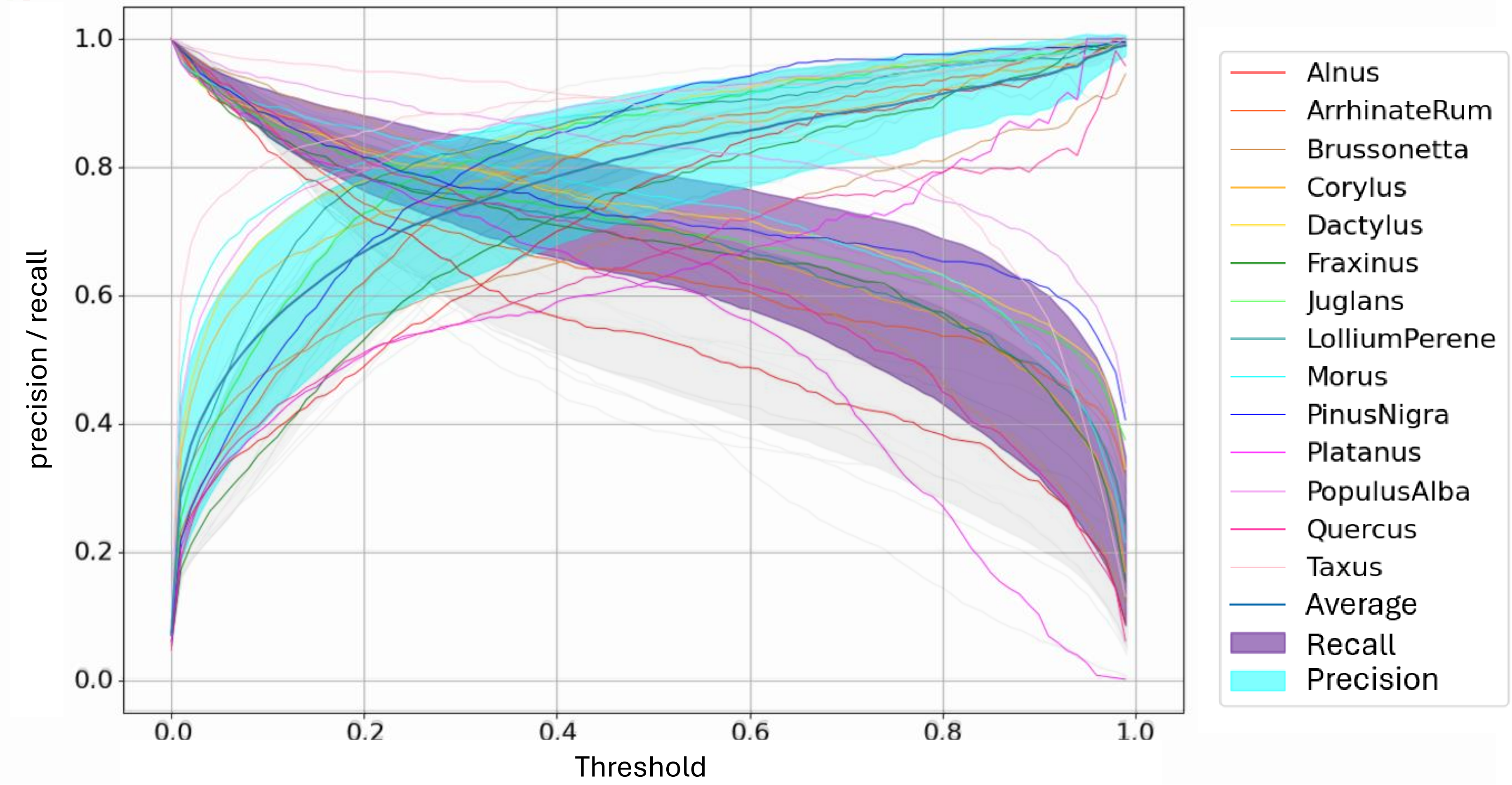
Precision and recall vs threshold

*Precision and recall – **common filtered data***



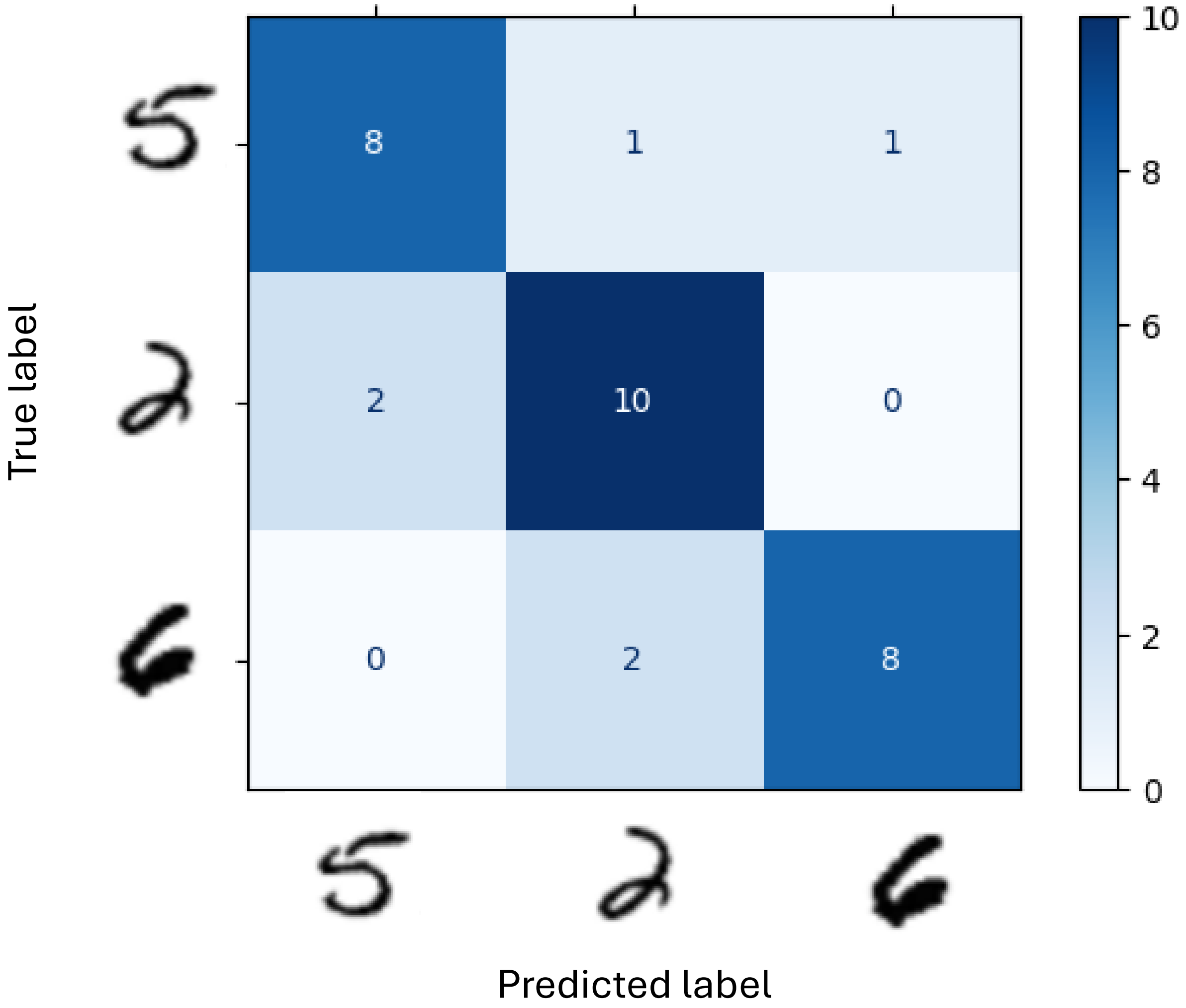
Precision and recall vs threshold

*Precision and recall – **raw data** (grey) vs **common filtered data** (coloured)*



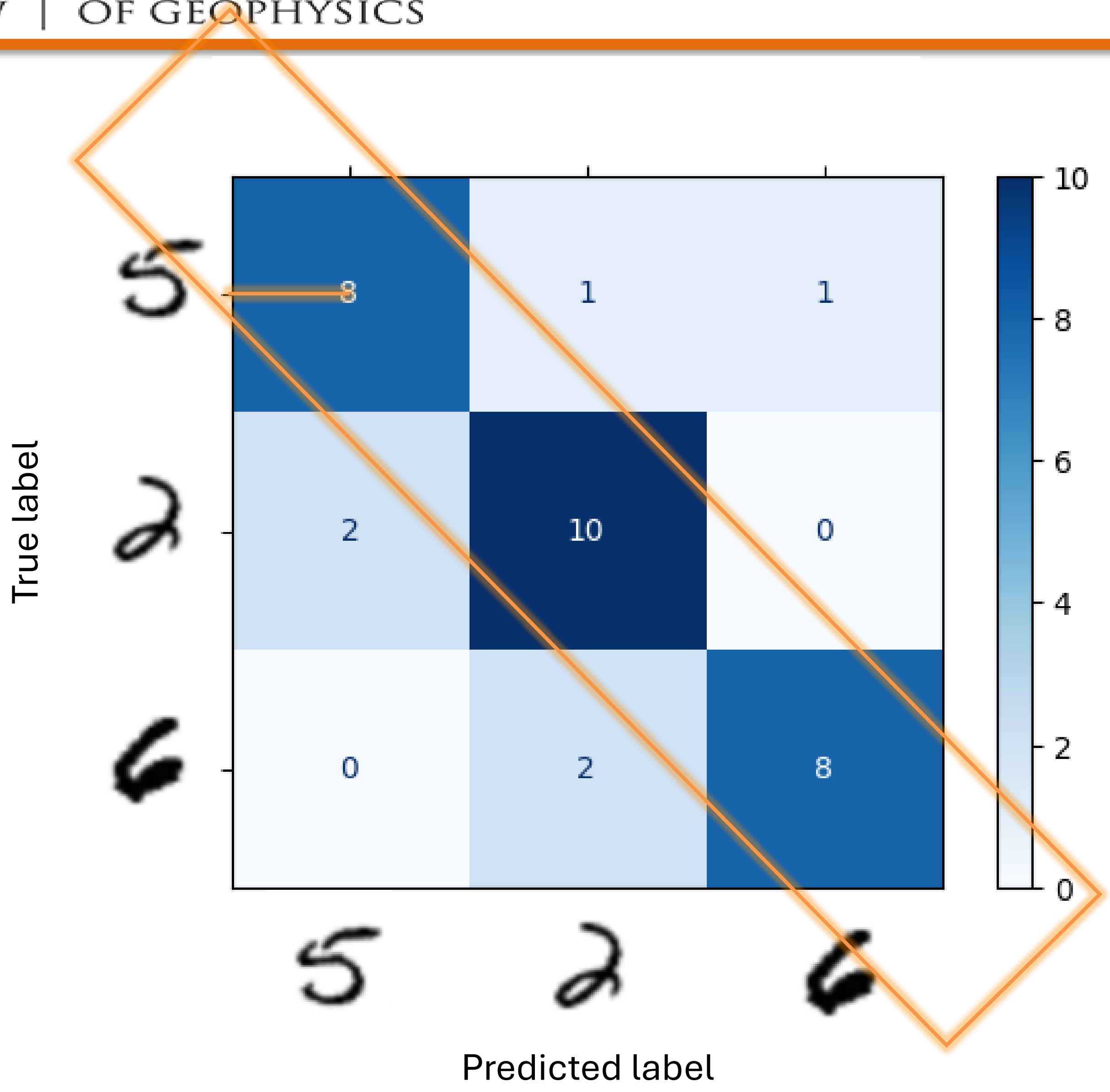
Confusion matrix - interpretation

The higher diagonal the better



Confusion matrix - interpretation

The higher diagonal the better

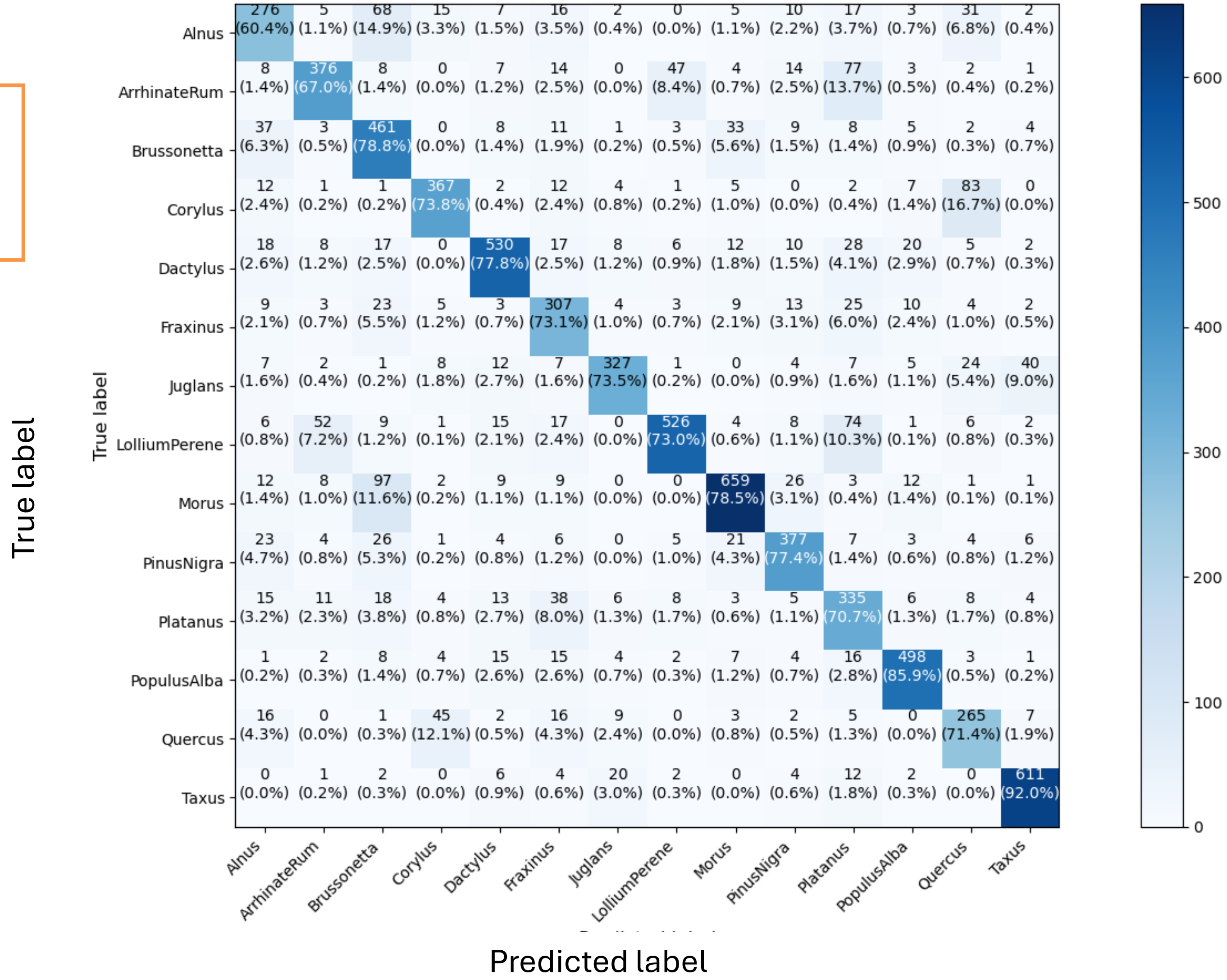


	5	2	6
5	8	1	1
2	2	10	0
6	0	2	8

Confusion matrix – many filters

The higher diagonal the better

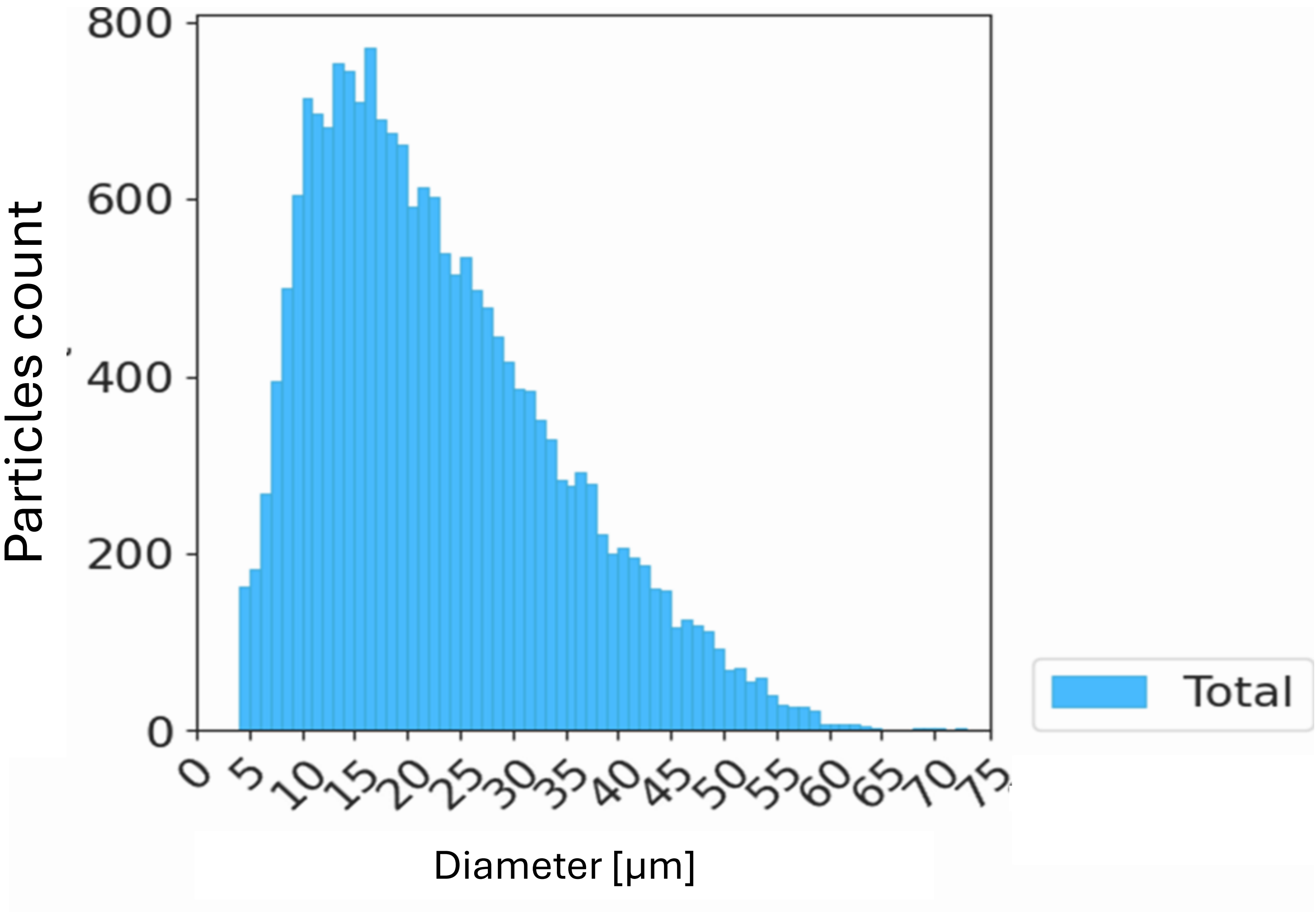
Common filter	76%
Multiple filters	75%
E-Rapid Clean	65%
Raw data	69%



True label \ Predicted label	Alnus	ArrhinateRum	Brussonetta	Corylus	Dactylus	Fraxinus	Juglans	LolliumPerene	Morus	PinusNigra	Platanus	PopulusAlba	Quercus	Taxus
Alnus	276 (60.4%)	5 (1.1%)	68 (14.9%)	15 (3.3%)	7 (1.5%)	16 (3.5%)	2 (0.4%)	0 (0.0%)	5 (1.1%)	10 (2.2%)	17 (3.7%)	3 (0.7%)	31 (6.8%)	2 (0.4%)
ArrhinateRum	8 (1.4%)	376 (67.0%)	8 (1.4%)	0 (0.0%)	7 (1.2%)	14 (2.5%)	0 (0.0%)	47 (8.4%)	4 (0.7%)	14 (2.5%)	77 (13.7%)	3 (0.5%)	2 (0.4%)	1 (0.2%)
Brussonetta	37 (6.3%)	3 (0.5%)	461 (78.8%)	0 (0.0%)	8 (1.4%)	11 (1.9%)	1 (0.2%)	3 (0.5%)	33 (5.6%)	9 (1.5%)	8 (1.4%)	5 (0.9%)	2 (0.3%)	4 (0.7%)
Corylus	12 (2.4%)	1 (0.2%)	1 (0.2%)	367 (73.8%)	2 (0.4%)	12 (2.4%)	4 (0.8%)	1 (0.2%)	5 (1.0%)	0 (0.0%)	2 (0.4%)	7 (1.4%)	83 (16.7%)	0 (0.0%)
Dactylus	18 (2.6%)	8 (1.2%)	17 (2.5%)	0 (0.0%)	530 (77.8%)	17 (2.5%)	8 (1.2%)	6 (0.9%)	12 (1.8%)	10 (1.5%)	28 (4.1%)	20 (2.9%)	5 (0.7%)	2 (0.3%)
Fraxinus	9 (2.1%)	3 (0.7%)	23 (5.5%)	5 (1.2%)	3 (0.7%)	307 (73.1%)	4 (1.0%)	3 (0.7%)	9 (2.1%)	13 (3.1%)	25 (6.0%)	10 (2.4%)	4 (1.0%)	2 (0.5%)
Juglans	7 (1.6%)	2 (0.4%)	1 (0.2%)	8 (1.8%)	12 (2.7%)	7 (1.6%)	327 (73.5%)	1 (0.2%)	0 (0.0%)	4 (0.9%)	7 (1.6%)	5 (1.1%)	24 (5.4%)	40 (9.0%)
LolliumPerene	6 (0.8%)	52 (7.2%)	9 (1.2%)	1 (0.1%)	15 (2.1%)	17 (2.4%)	0 (0.0%)	526 (73.0%)	4 (0.6%)	8 (1.1%)	74 (10.3%)	1 (0.1%)	6 (0.8%)	2 (0.3%)
Morus	12 (1.4%)	8 (1.0%)	97 (11.6%)	2 (0.2%)	9 (1.1%)	9 (1.1%)	0 (0.0%)	0 (0.0%)	659 (78.5%)	26 (3.1%)	3 (0.4%)	12 (1.4%)	1 (0.1%)	1 (0.1%)
PinusNigra	23 (4.7%)	4 (0.8%)	26 (5.3%)	1 (0.2%)	4 (0.8%)	6 (1.2%)	0 (0.0%)	5 (1.0%)	21 (4.3%)	377 (77.4%)	7 (1.4%)	3 (0.6%)	4 (0.8%)	6 (1.2%)
Platanus	15 (3.2%)	11 (2.3%)	18 (3.8%)	4 (0.8%)	13 (2.7%)	38 (8.0%)	6 (1.3%)	8 (1.7%)	3 (0.6%)	5 (1.1%)	335 (70.7%)	6 (1.3%)	8 (1.7%)	4 (0.8%)
PopulusAlba	1 (0.2%)	2 (0.3%)	8 (1.4%)	4 (0.7%)	15 (2.6%)	15 (2.6%)	4 (0.7%)	2 (0.3%)	7 (1.2%)	4 (0.7%)	16 (2.8%)	498 (85.9%)	3 (0.5%)	1 (0.2%)
Quercus	16 (4.3%)	0 (0.0%)	1 (0.3%)	45 (12.1%)	2 (0.5%)	16 (4.3%)	9 (2.4%)	0 (0.0%)	3 (0.8%)	2 (0.5%)	5 (1.3%)	0 (0.0%)	265 (71.4%)	7 (1.9%)
Taxus	0 (0.0%)	1 (0.2%)	2 (0.3%)	0 (0.0%)	6 (0.9%)	4 (0.6%)	20 (3.0%)	2 (0.3%)	0 (0.0%)	4 (0.6%)	12 (1.8%)	2 (0.3%)	0 (0.0%)	611 (92.0%)

Histogram – particle size

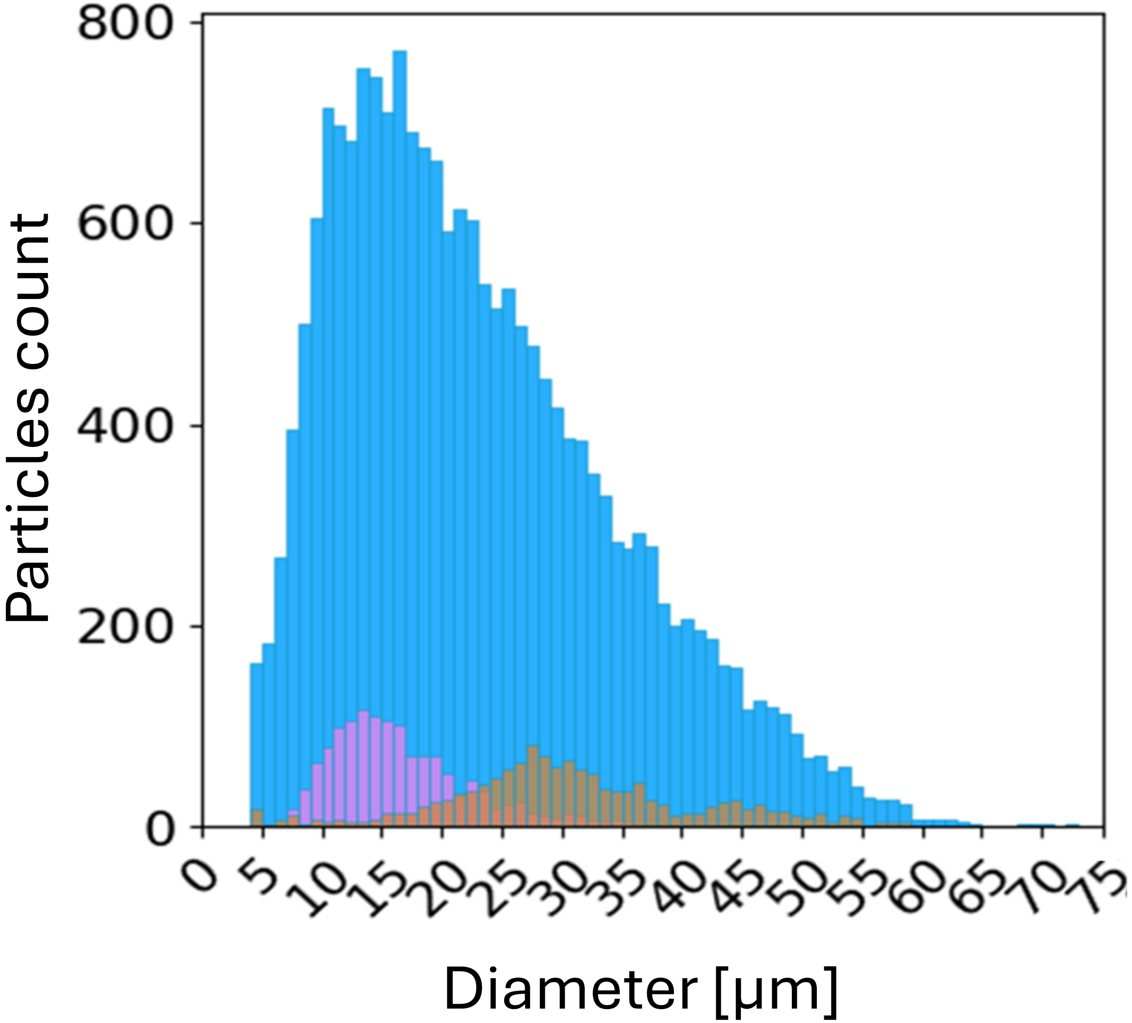
Particle size histogram for training data – common filter

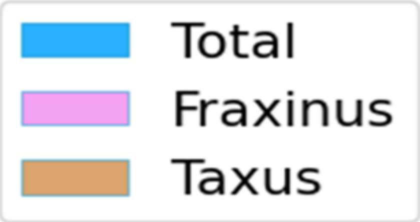


Histogram – particle size

Overlapping sizes

Particle size histogram with Fraxinus and Taxus for training data – common filter

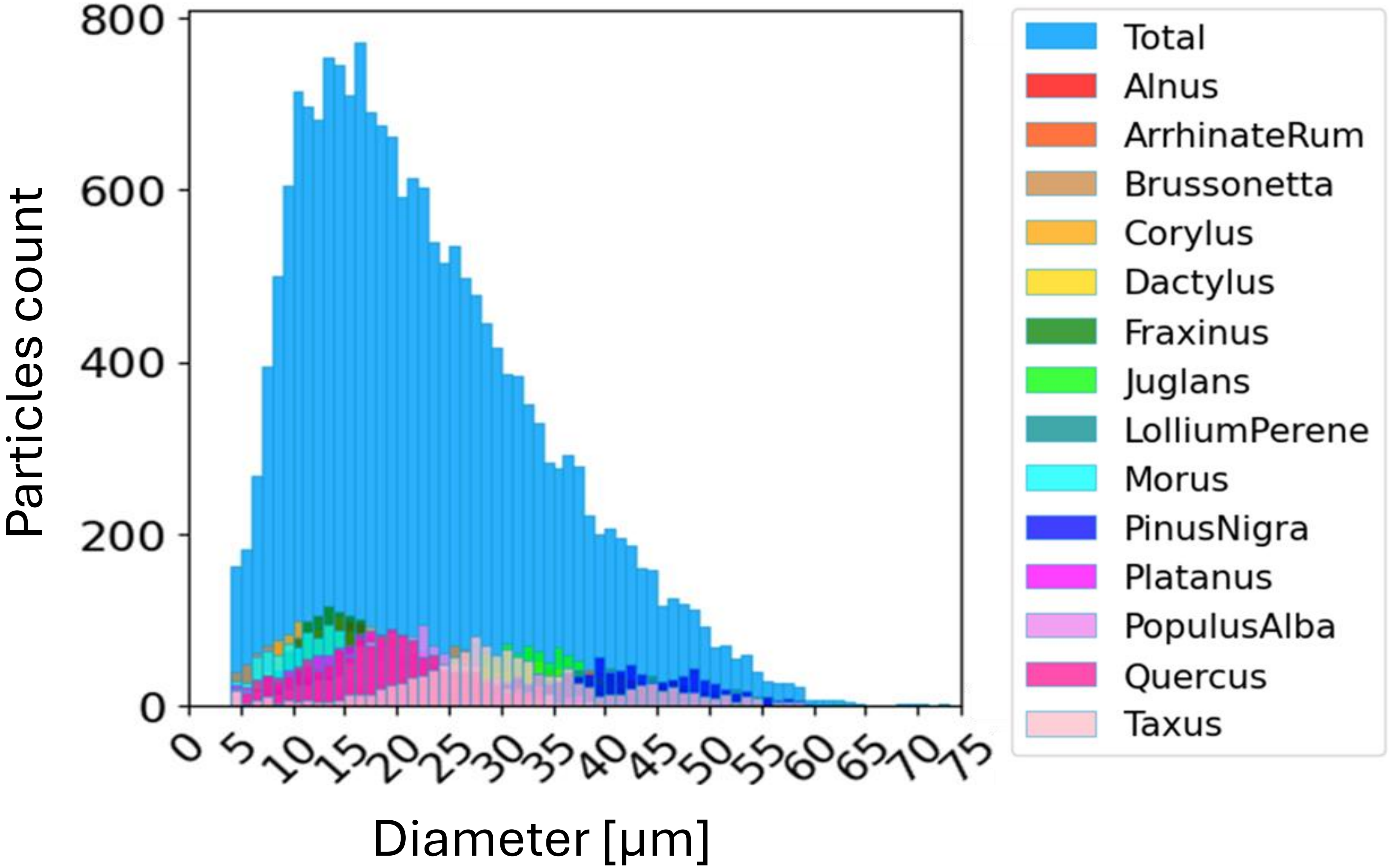




Histogram – particle size

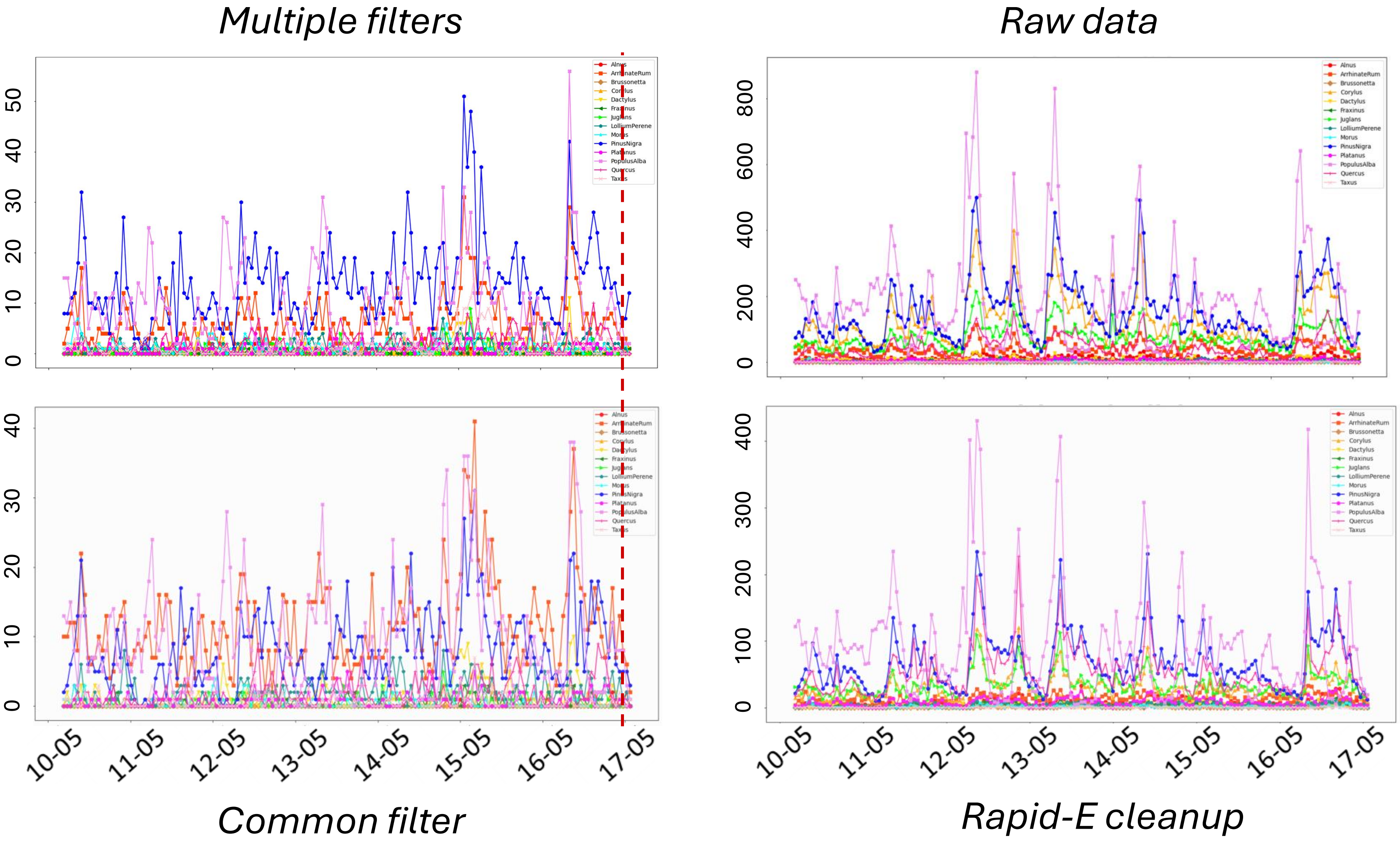
Particle size histogram for all pollen types for training data – common filter

Overlapping sizes



Case study

One week of measurements comparison: 2022-05-10 – 2022-05-17



Multiple filters

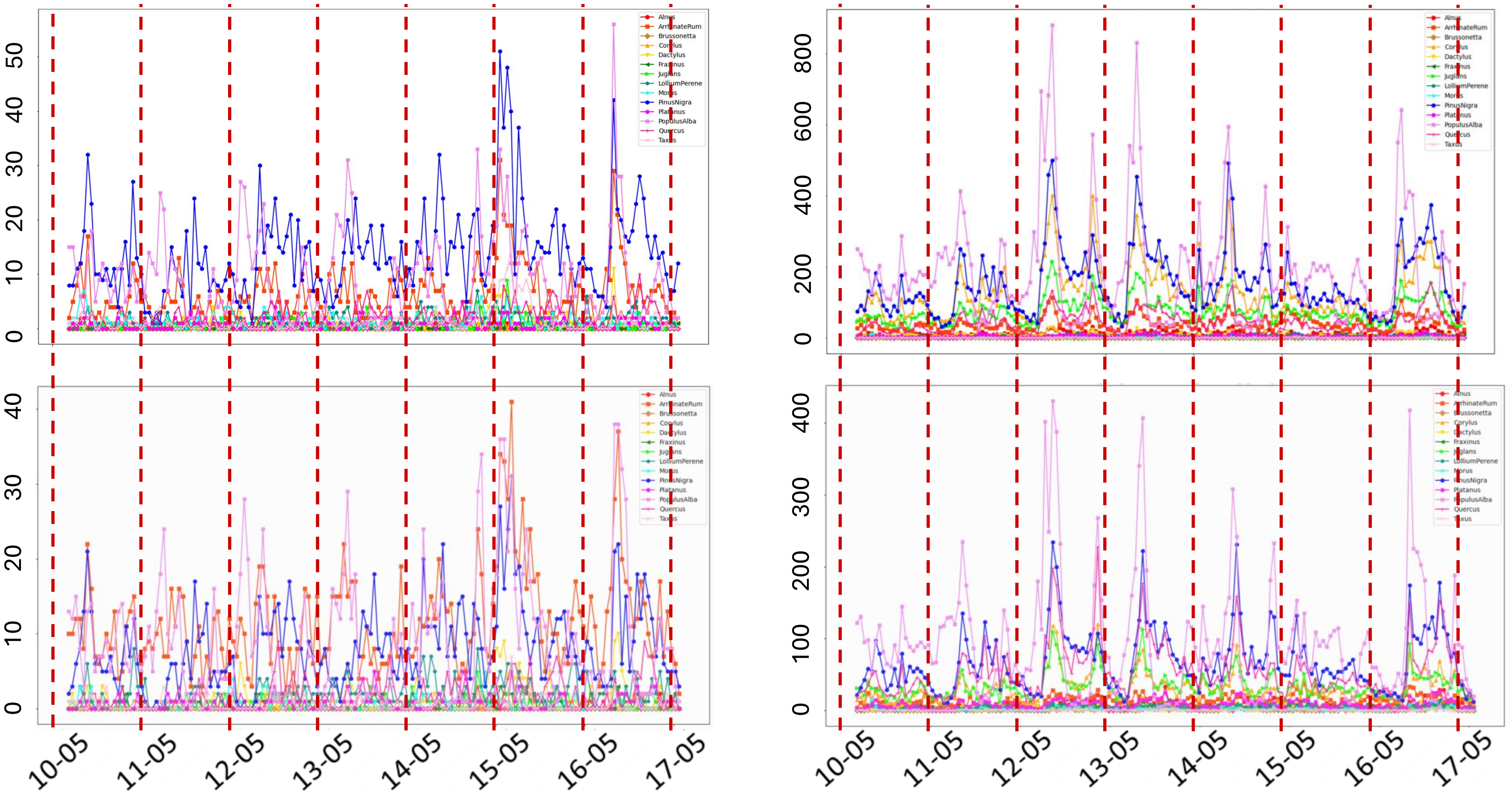
Raw data

Common filter

Rapid-E cleanup

Case study

One week of measurements comparison: 2022-05-10 – 2022-05-17



Multiple filters

Raw data

Common filter

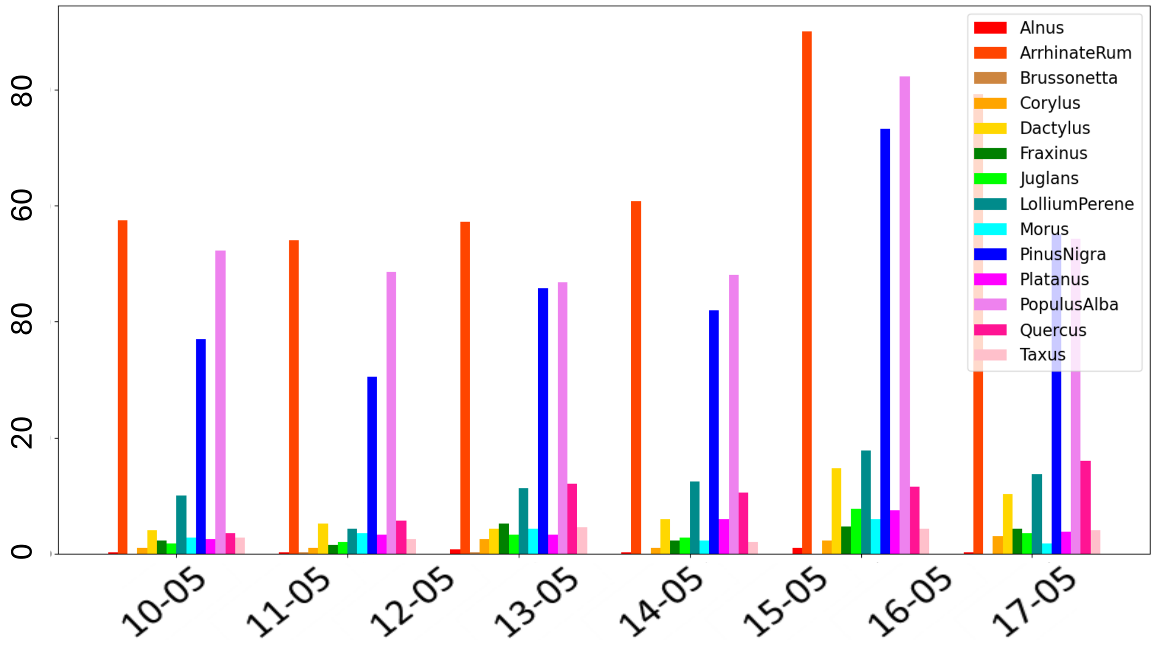
Rapid-E cleanup

Daily fluctuations

Case study

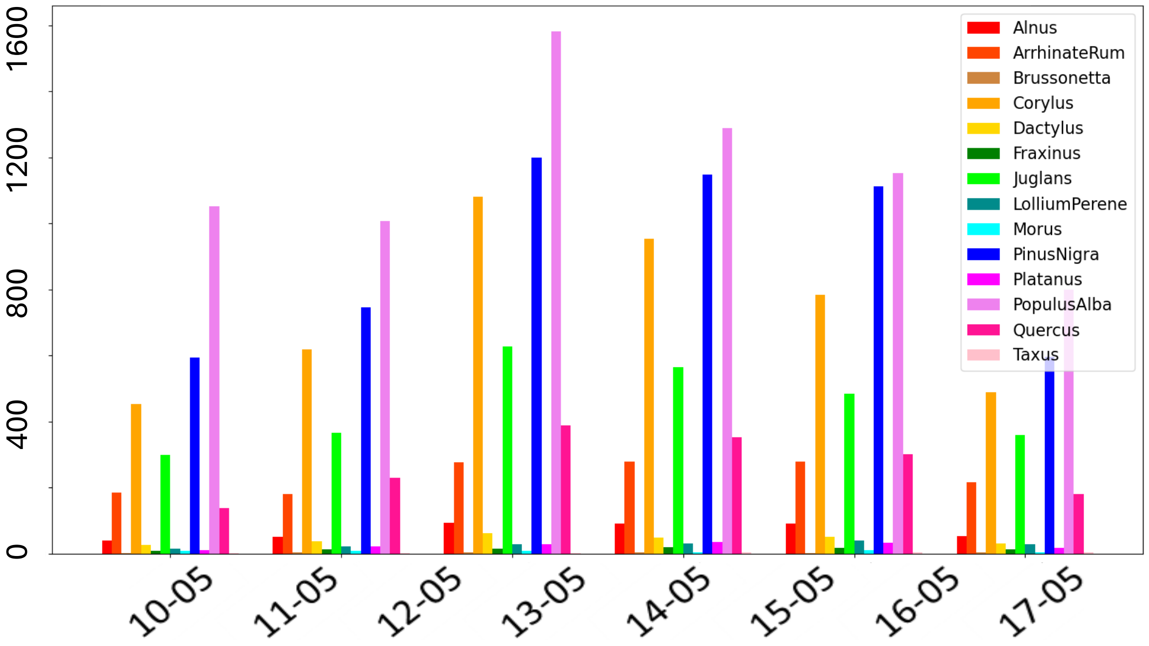
One week of measurements comparison: 2022-05-10 – 2022-05-17

Common filter



Acceptable: 13/14

Raw data



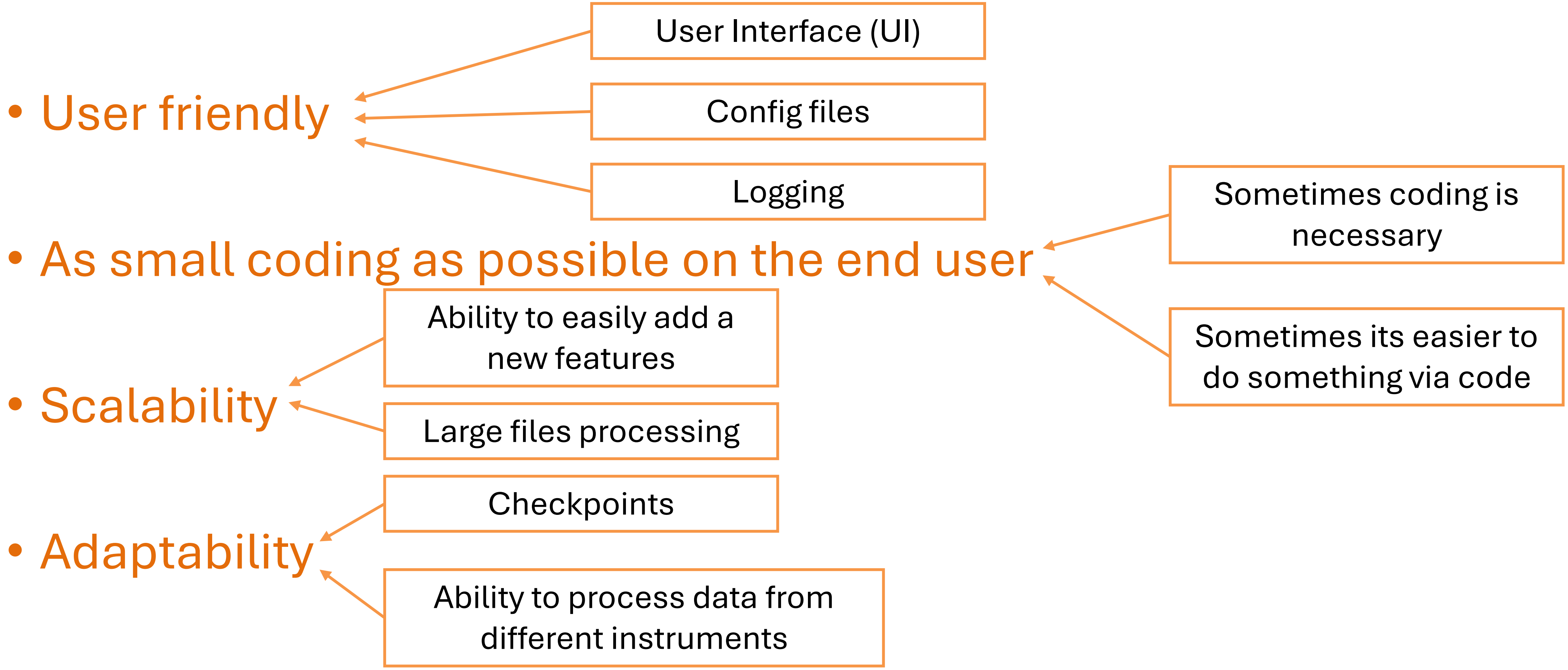
Acceptable: 8/14

Framework

~7k lines of code

- Ready to use tool for:
  - Data filtering and inspecting - *DataViewer*
  - Training and validating models - *ModelBuilder*
  - Mapping results - *ModelRunner*
  - Visualising results - *PredictionsMapper*
- Ability to run in near real-time and map historical data

Design assumptions



- User friendly
- As small coding as possible on the end user
- Scalability
- Adaptability

User Interface (UI)

Config files

Logging

Sometimes coding is necessary

Sometimes its easier to do something via code

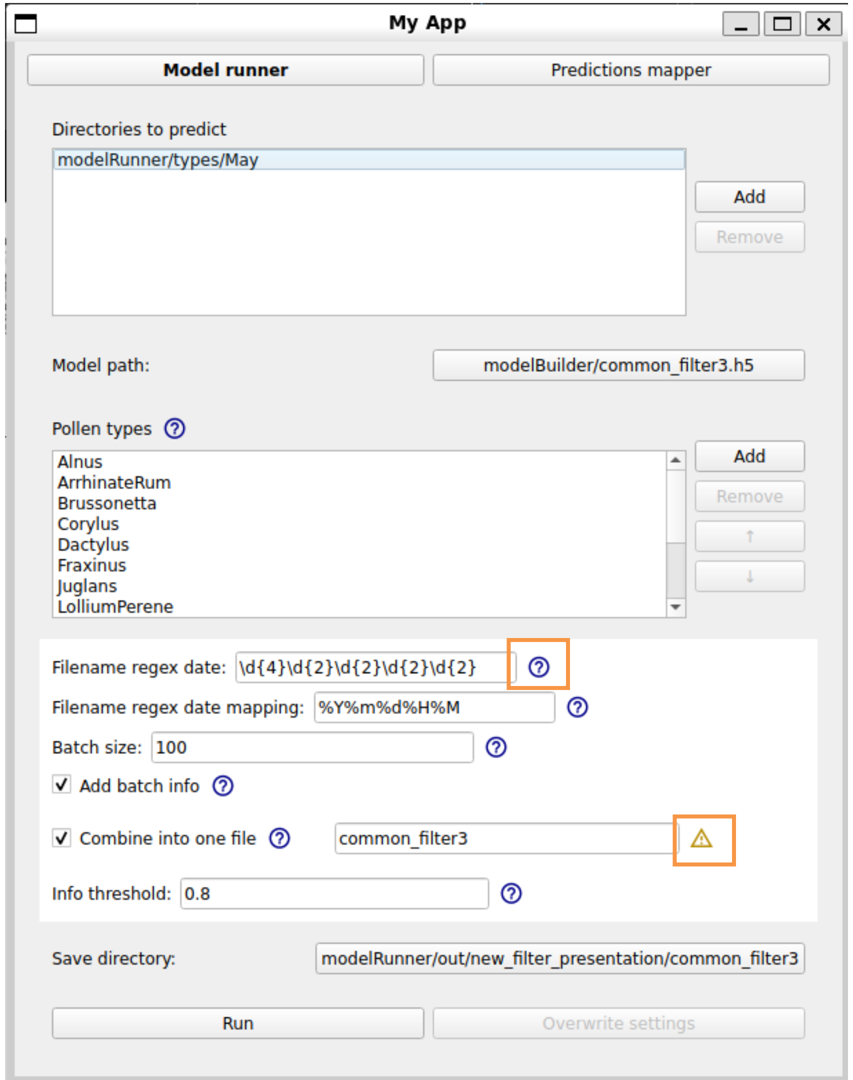
Ability to easily add a new features

Large files processing

Checkpoints

Ability to process data from different instruments

User Interface



My App

Model runner | Predictions mapper

Directories to predict

modelRunner/types/May

Add
Remove

Model path: modelBuilder/common_filter3.h5

Pollen types ?

Alnus
ArrhinateRum
Brussonetta
Corylus
Dactylus
Fraxinus
Juglans
LolliumPerene

Add
Remove
↑
↓

Filename regex date: \d{4}\d{2}\d{2}\d{2}\d{2} ?

Filename regex date mapping: %Y%m%d%H%M ?

Batch size: 100 ?

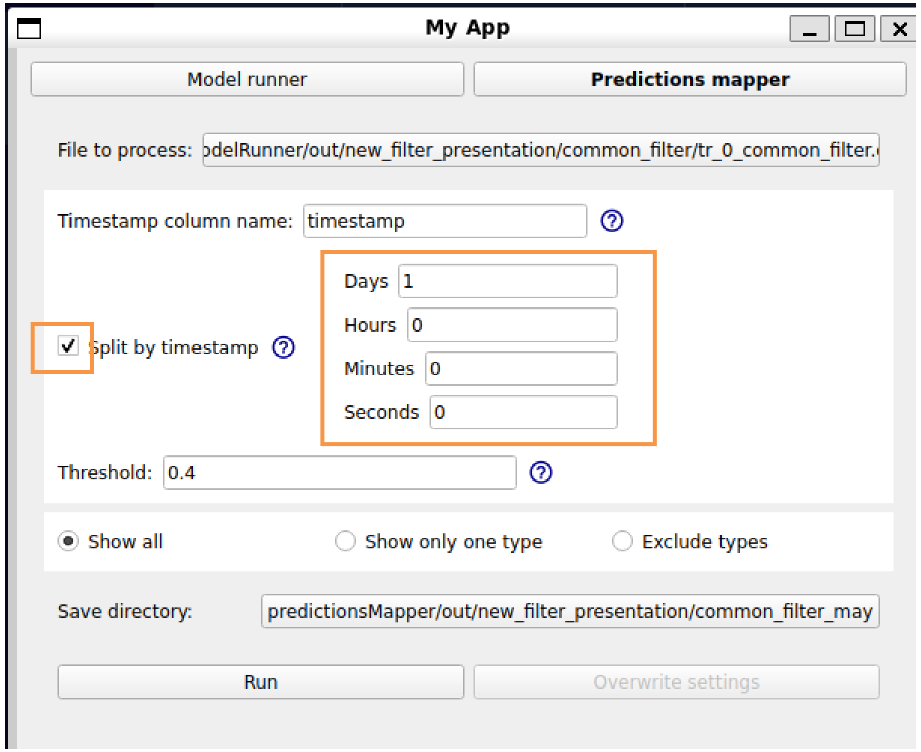
☑ Add batch info ?

☑ Combine into one file ? common_filter3 ⚠

Info threshold: 0.8 ?

Save directory: modelRunner/out/new_filter_presentation/common_filter3

Run | Overwrite settings



My App

Model runner | Predictions mapper

File to process: odelRunner/out/new_filter_presentation/common_filter/tr_0_common_filter.

Timestamp column name: timestamp ?

☑ Split by timestamp ?

Days 1
Hours 0
Minutes 0
Seconds 0

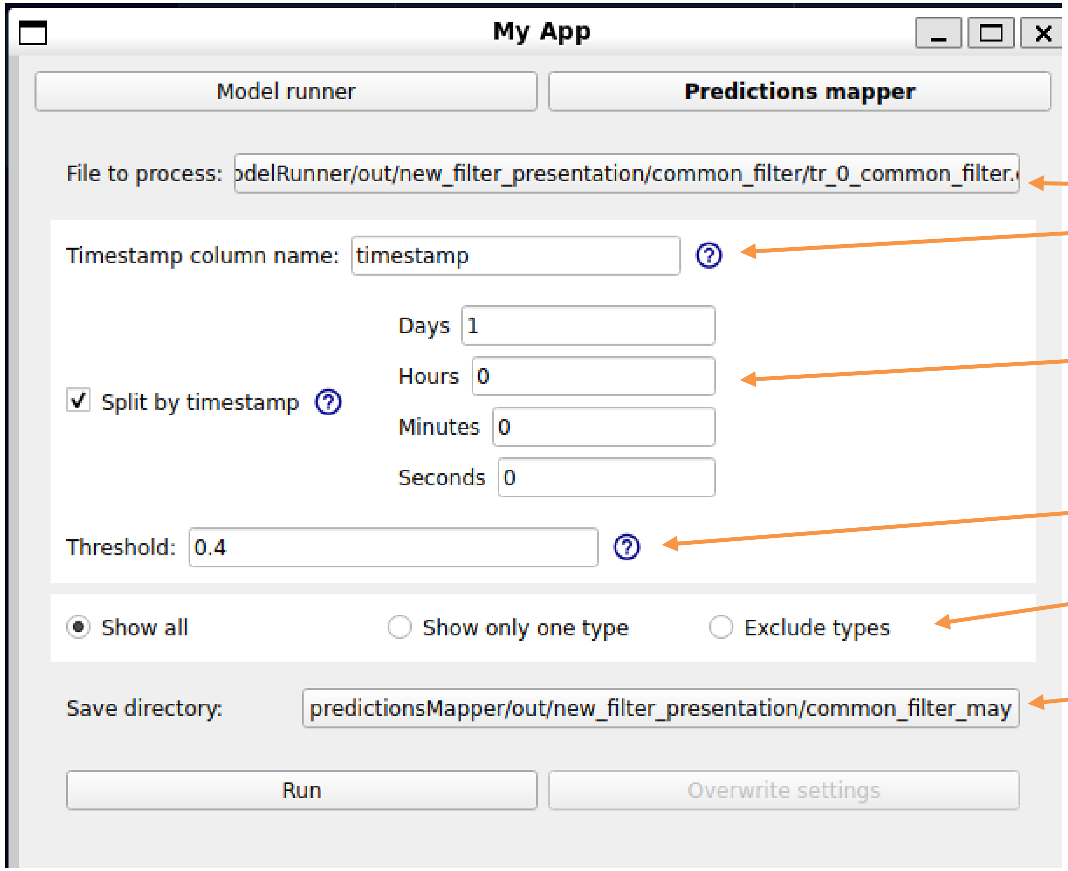
Threshold: 0.4 ?

● Show all ○ Show only one type ○ Exclude types

Save directory: predictionsMapper/out/new_filter_presentation/common_filter_may

Run | Overwrite settings

Config files



File to process: odelRunner/out/new_filter_presentation/common_filter/tr_0_common_filter.

Timestamp column name: timestamp

☑ Split by timestamp

Days 1

Hours 0

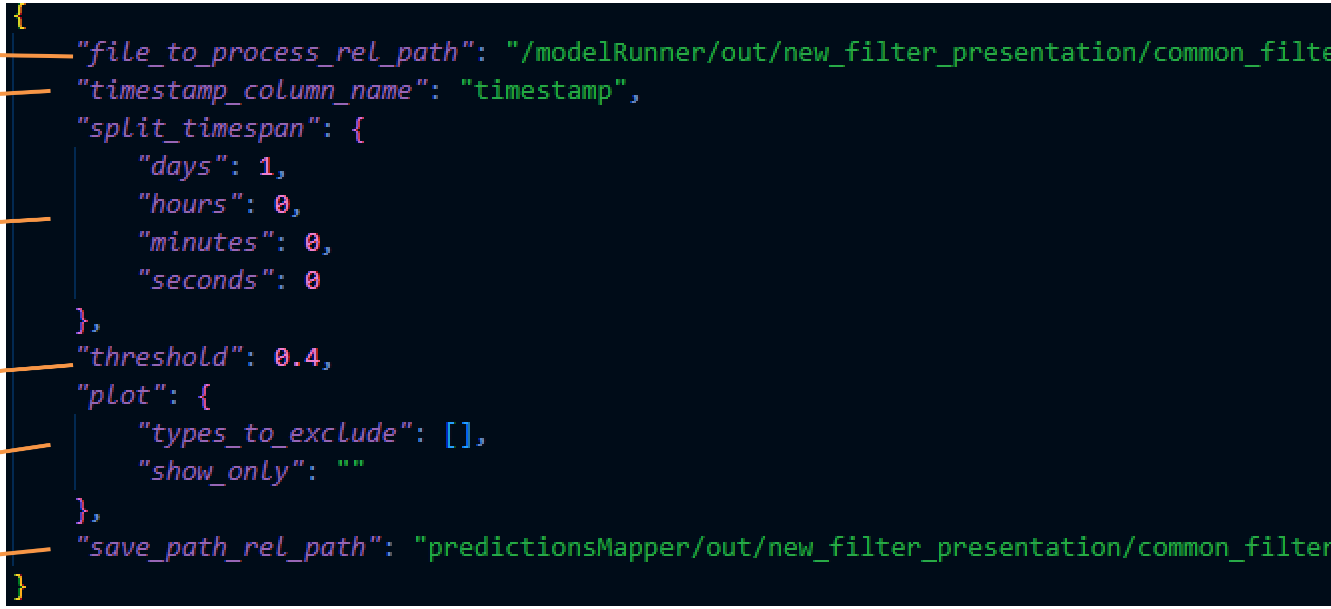
Minutes 0

Seconds 0

Threshold: 0.4

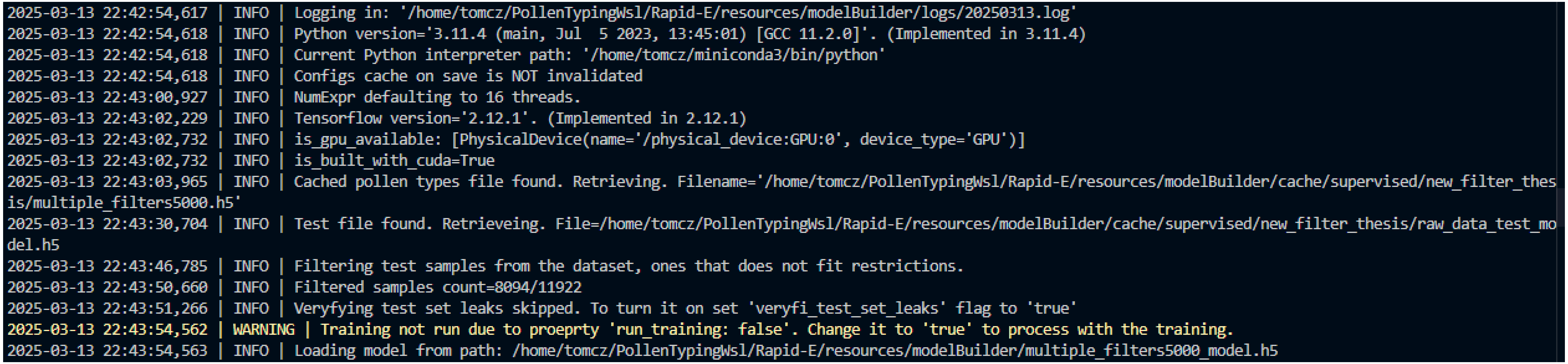
● Show all ○ Show only one type ○ Exclude types

Save directory: predictionsMapper/out/new_filter_presentation/common_filter_may

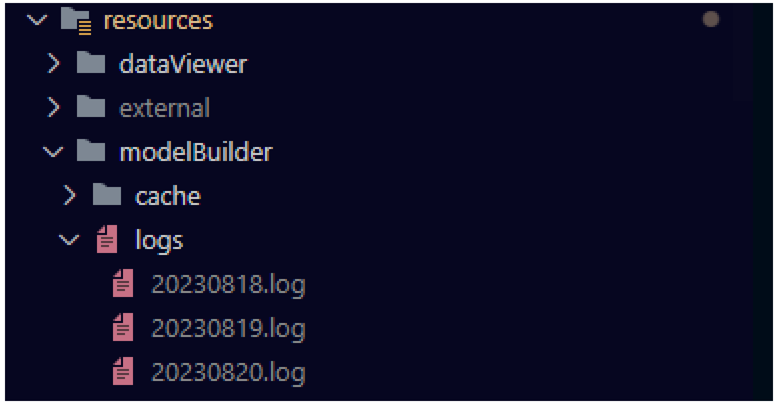


```
{
    "file_to_process_rel_path": "/modelRunner/out/new_filter_presentation/common_filte
    "timestamp_column_name": "timestamp",
    "split_timespan": {
        "days": 1,
        "hours": 0,
        "minutes": 0,
        "seconds": 0
    },
    "threshold": 0.4,
    "plot": {
        "types_to_exclude": [],
        "show_only": ""
    },
    "save_path_rel_path": "predictionsMapper/out/new_filter_presentation/common_filter
}
```

Logging

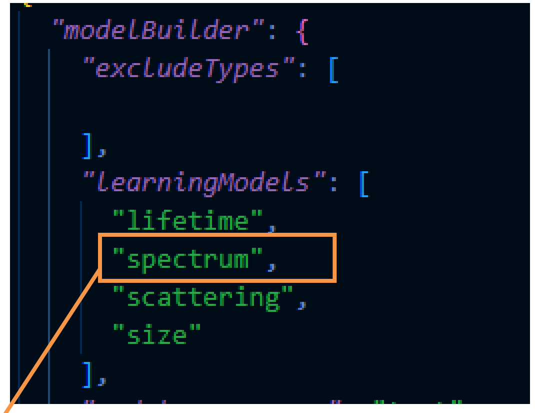


```
2025-03-13 22:42:54,617 | INFO | Logging in: '/home/tomcz/PollenTypingWsl/Rapid-E/resources/modelBuilder/logs/20250313.log'
2025-03-13 22:42:54,618 | INFO | Python version='3.11.4 (main, Jul  5 2023, 13:45:01) [GCC 11.2.0]'. (Implemented in 3.11.4)
2025-03-13 22:42:54,618 | INFO | Current Python interpreter path: '/home/tomcz/miniconda3/bin/python'
2025-03-13 22:42:54,618 | INFO | Configs cache on save is NOT invalidated
2025-03-13 22:43:00,927 | INFO | NumExpr defaulting to 16 threads.
2025-03-13 22:43:02,229 | INFO | Tensorflow version='2.12.1'. (Implemented in 2.12.1)
2025-03-13 22:43:02,732 | INFO | is_gpu_available: [PhysicalDevice(name='/physical_device:GPU:0', device_type='GPU')]
2025-03-13 22:43:02,732 | INFO | is_built_with_cuda=True
2025-03-13 22:43:03,965 | INFO | Cached pollen types file found. Retrieving. Filename='/home/tomcz/PollenTypingWsl/Rapid-E/resources/modelBuilder/cache/supervised/new_filter_thes
is/multiple_filters5000.h5'
2025-03-13 22:43:30,704 | INFO | Test file found. Retrieveing. File=/home/tomcz/PollenTypingWsl/Rapid-E/resources/modelBuilder/cache/supervised/new_filter_thesis/raw_data_test_mo
del.h5
2025-03-13 22:43:46,785 | INFO | Filtering test samples from the dataset, ones that does not fit restrictions.
2025-03-13 22:43:50,660 | INFO | Filtered samples count=8094/11922
2025-03-13 22:43:51,266 | INFO | Veryfying test set leaks skipped. To turn it on set 'veryfi_test_set_leaks' flag to 'true'
2025-03-13 22:43:54,562 | WARNING | Training not run due to proeprty 'run_training: false'. Change it to 'true' to process with the training.
2025-03-13 22:43:54,563 | INFO | Loading model from path: /home/tomcz/PollenTypingWsl/Rapid-E/resources/modelBuilder/multiple_filters5000_model.h5
```



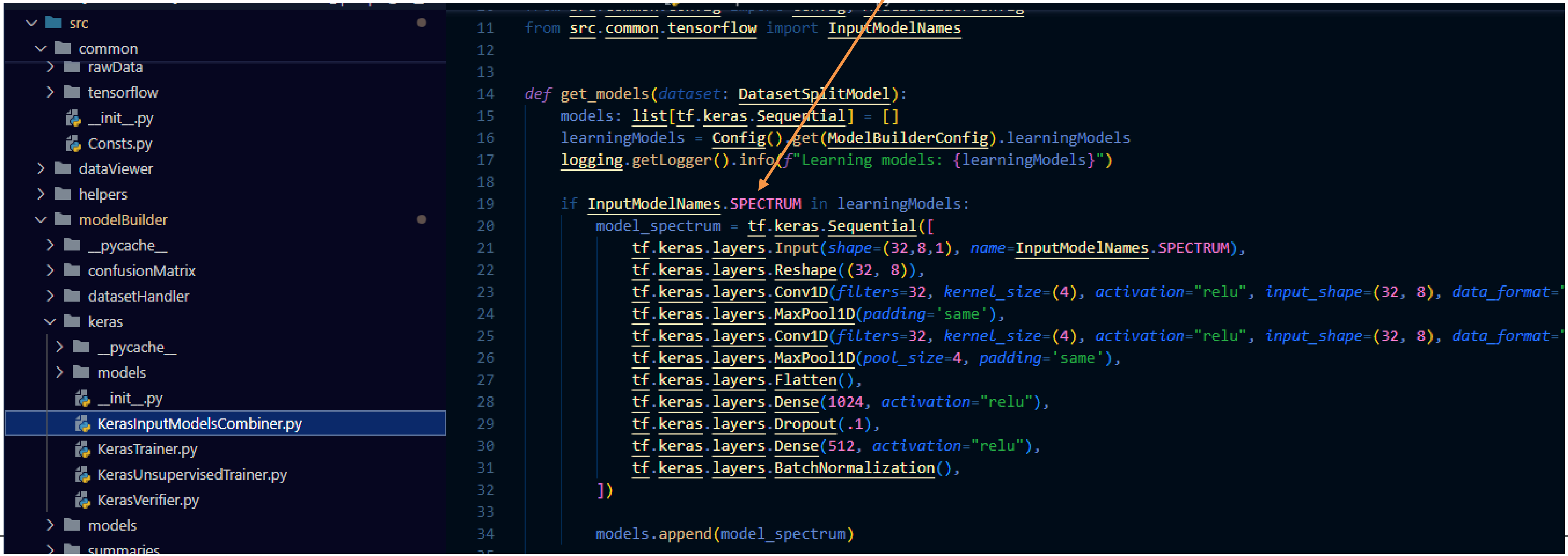
Coding parts

- Filters
- Changes in ML model



```
"modelBuilder": {
  "excludeTypes": [

  ],
  "learningModels": [
    "lifetime",
    "spectrum",
    "scattering",
    "size"
  ],
```



```
src
  common
    rawData
    tensorflow
    __init__.py
    Consts.py
  dataViewer
  helpers
  modelBuilder
    __pycache__
    confusionMatrix
    datasetHandler
    keras
      __pycache__
      models
      __init__.py
      KerasInputModelsCombiner.py
      KerasTrainer.py
      KerasUnsupervisedTrainer.py
      KerasVerifier.py
    models
    summaries
```

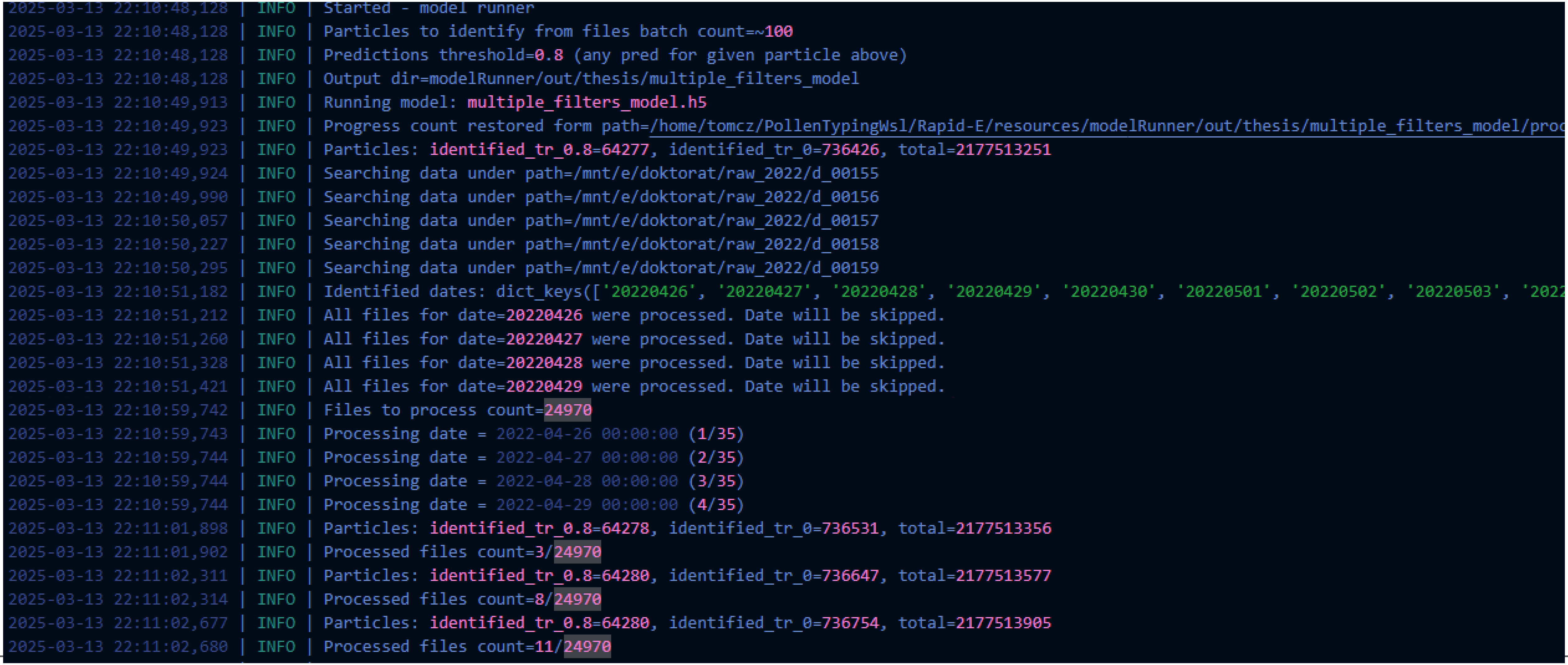
```python
from src.common.tensorflow import InputModelNames


def get_models(dataset: DatasetSplitModel):
    models: list[tf.keras.Sequential] = []
    learningModels = Config().get(ModelBuilderConfig).learningModels
    logging.getLogger().info(f"Learning models: {learningModels}")

    if InputModelNames.SPECTRUM in learningModels:
        model_spectrum = tf.keras.Sequential([
            tf.keras.layers.Input(shape=(32,8,1), name=InputModelNames.SPECTRUM),
            tf.keras.layers.Reshape((32, 8)),
            tf.keras.layers.Conv1D(filters=32, kernel_size=(4), activation="relu", input_shape=(32, 8), data_format='
            tf.keras.layers.MaxPool1D(padding='same'),
            tf.keras.layers.Conv1D(filters=32, kernel_size=(4), activation="relu", input_shape=(32, 8), data_format='
            tf.keras.layers.MaxPool1D(pool_size=4, padding='same'),
            tf.keras.layers.Flatten(),
            tf.keras.layers.Dense(1024, activation="relu"),
            tf.keras.layers.Dropout(.1),
            tf.keras.layers.Dense(512, activation="relu"),
            tf.keras.layers.BatchNormalization(),
        ])

        models.append(model_spectrum)
```

Large files processing & checkpoints

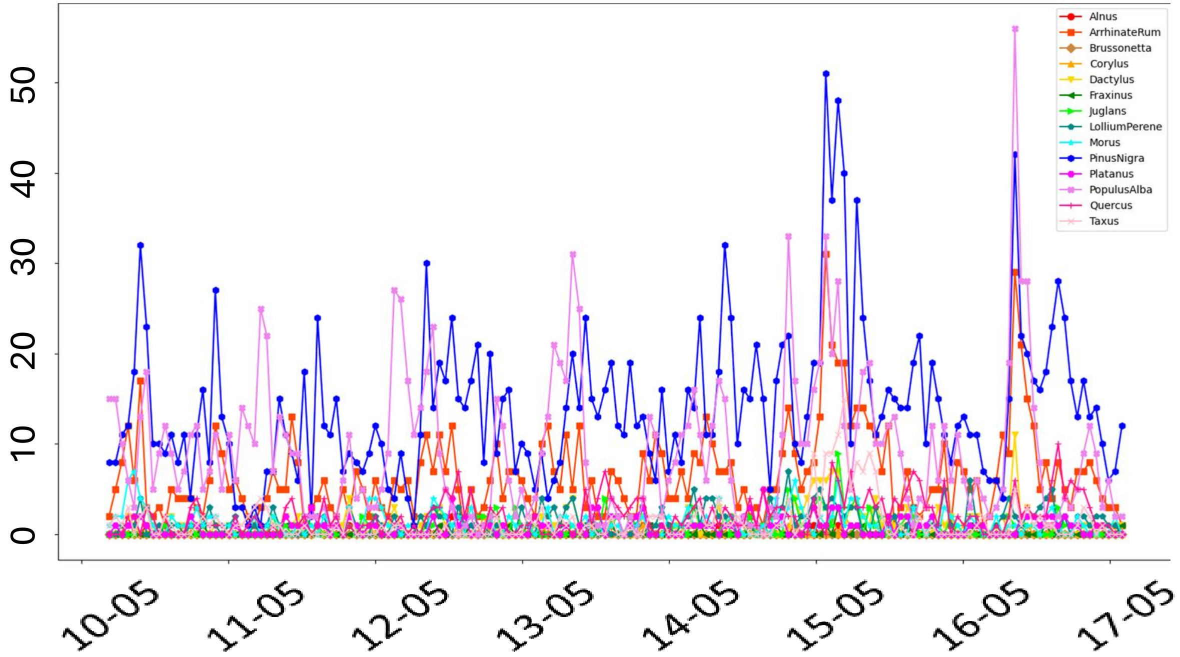


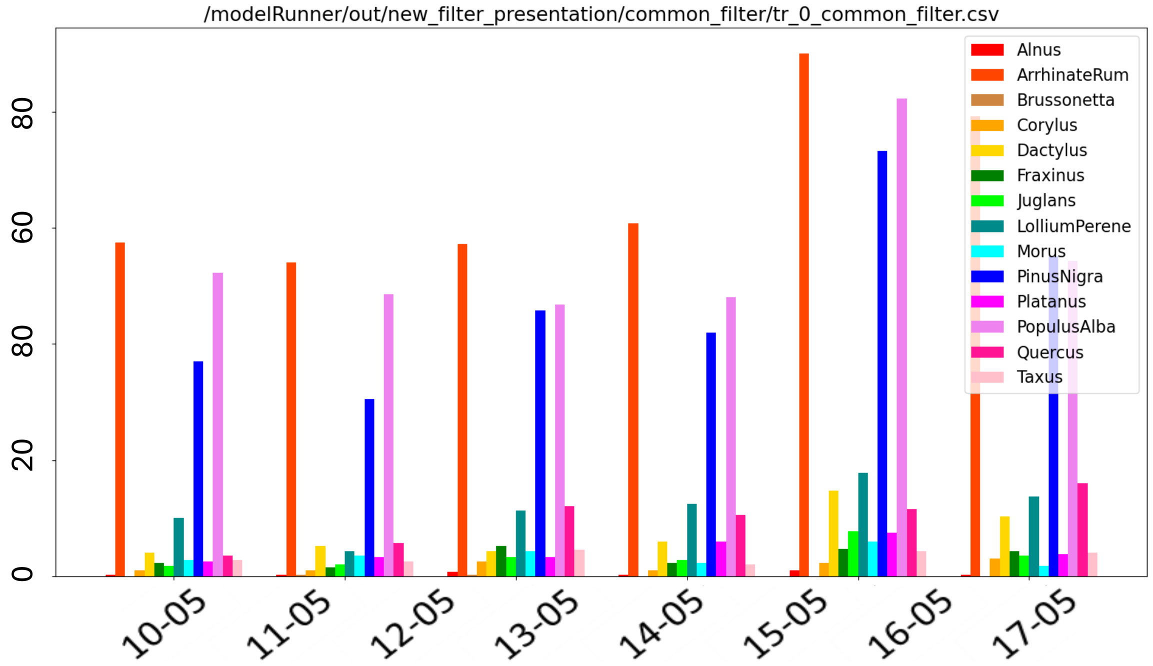
```
2025-03-13 22:10:48,128 | INFO | Started - model runner
2025-03-13 22:10:48,128 | INFO | Particles to identify from files batch count=~100
2025-03-13 22:10:48,128 | INFO | Predictions threshold=0.8 (any pred for given particle above)
2025-03-13 22:10:48,128 | INFO | Output dir=modelRunner/out/thesis/multiple_filters_model
2025-03-13 22:10:49,913 | INFO | Running model: multiple_filters_model.h5
2025-03-13 22:10:49,923 | INFO | Progress count restored form path=/home/tomcz/PollenTypingWsl/Rapid-E/resources/modelRunner/out/thesis/multiple_filters_model/proc
2025-03-13 22:10:49,923 | INFO | Particles: identified_tr_0.8=64277, identified_tr_0=736426, total=2177513251
2025-03-13 22:10:49,924 | INFO | Searching data under path=/mnt/e/doktorat/raw_2022/d_00155
2025-03-13 22:10:49,990 | INFO | Searching data under path=/mnt/e/doktorat/raw_2022/d_00156
2025-03-13 22:10:50,057 | INFO | Searching data under path=/mnt/e/doktorat/raw_2022/d_00157
2025-03-13 22:10:50,227 | INFO | Searching data under path=/mnt/e/doktorat/raw_2022/d_00158
2025-03-13 22:10:50,295 | INFO | Searching data under path=/mnt/e/doktorat/raw_2022/d_00159
2025-03-13 22:10:51,182 | INFO | Identified dates: dict_keys(['20220426', '20220427', '20220428', '20220429', '20220430', '20220501', '20220502', '20220503', '2022
2025-03-13 22:10:51,212 | INFO | All files for date=20220426 were processed. Date will be skipped.
2025-03-13 22:10:51,260 | INFO | All files for date=20220427 were processed. Date will be skipped.
2025-03-13 22:10:51,328 | INFO | All files for date=20220428 were processed. Date will be skipped.
2025-03-13 22:10:51,421 | INFO | All files for date=20220429 were processed. Date will be skipped.
2025-03-13 22:10:59,742 | INFO | Files to process count=24970
2025-03-13 22:10:59,743 | INFO | Processing date = 2022-04-26 00:00:00 (1/35)
2025-03-13 22:10:59,744 | INFO | Processing date = 2022-04-27 00:00:00 (2/35)
2025-03-13 22:10:59,744 | INFO | Processing date = 2022-04-28 00:00:00 (3/35)
2025-03-13 22:10:59,744 | INFO | Processing date = 2022-04-29 00:00:00 (4/35)
2025-03-13 22:11:01,898 | INFO | Particles: identified_tr_0.8=64278, identified_tr_0=736531, total=2177513356
2025-03-13 22:11:01,902 | INFO | Processed files count=3/24970
2025-03-13 22:11:02,311 | INFO | Particles: identified_tr_0.8=64280, identified_tr_0=736647, total=2177513577
2025-03-13 22:11:02,314 | INFO | Processed files count=8/24970
2025-03-13 22:11:02,677 | INFO | Particles: identified_tr_0.8=64280, identified_tr_0=736754, total=2177513905
2025-03-13 22:11:02,680 | INFO | Processed files count=11/24970
```

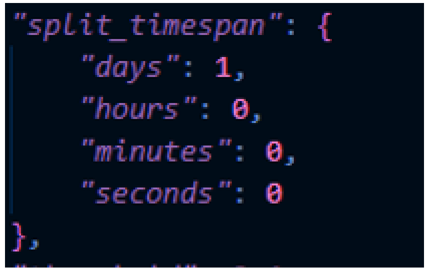
Instrument independent

- Current status:
  - ModelRunner and PredictionsMapper – fully independent
  - DataViewer and ModelBuilder – partially independent – requires some adaptations
- Ability to read:
  - Plair binary files – raw data files from instrument
  - .json files

Visualising the data







```
"split_timespan": {
    "days": 1,
    "hours": 0,
    "minutes": 0,
    "seconds": 0
},
```

Key takeaways

- Promising results – validation required
- Ready to use tool for:
  - Data filtering and inspecting
  - Training models
  - Validating models
  - Mapping results
  - Visualising results
- Ability to run in near real-time and map historical data

Thank you

The research was carried out in cooperation between the Remote Sensing Laboratory (RS-Lab) of the Institute of Geophysics, Faculty of Physics, University of Warsaw (https://www.igf.fuw.edu.pl) and the National Institute for Research and Development in Optoelectronics INOE 2000 (https://www.inoe.ro).
We would like to thank Andrei Dandocsi from INOE for general pollen characterization.

The research is financed by the Polish National Agency for Academic Exchange (NAWA) within the NAWA Preludium Bis 2 internship grant, PN/PRE/2022/1/00024

The research is financed by the Poland National Science Centre within the Preludium BIS-2 project, 2020/39/O/ST10/03586

Contact emails: *artur.tomczak@fuw.edu.pl*
iwona.stachlewska@fuw.edu.pl