Cambodia GeoSpatial Day 2025
18 November 2025
Institute of Technology of Cambodia, Phnom Penh

Integrating MALDI-TOF MS and Machine Learning for Rapid Mosquito Species Identification in Cambodia

Matilin LE BEUX, Ph.D. Student

mbeux@pasteur-kh.org

Medical and Veterinary Entomology Unit, Institut Pasteur du Cambodge

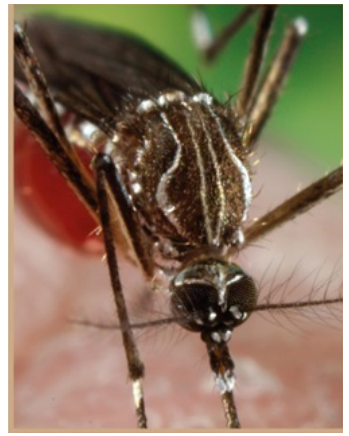
CONTEXT

Institut Pasteur du Cambodge
PASTEUR NETWORK

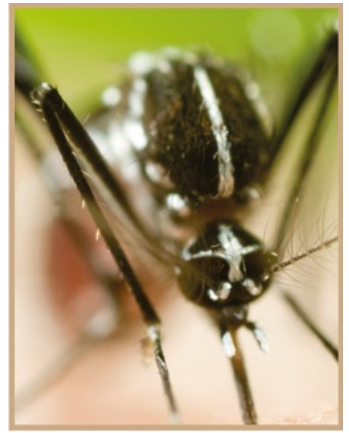
Cambodia

Favorable environment for **arboviruses**

- **Dengue (Den.)** and **Chikungunya (Chik.)** endemic [*Beauté and Sirenda, 2010*]
- **Japanese Encephalitis (JE)** [*Cappelle et al. 2016*]



Aedes aegypti



Aedes albopictus

CONTEXT

Institut Pasteur du Cambodge
PASTEUR NETWORK

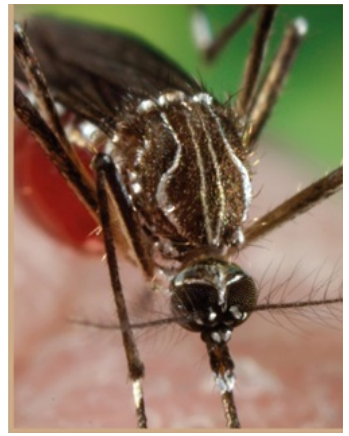
Cambodia

Favorable environment for **arboviruses**

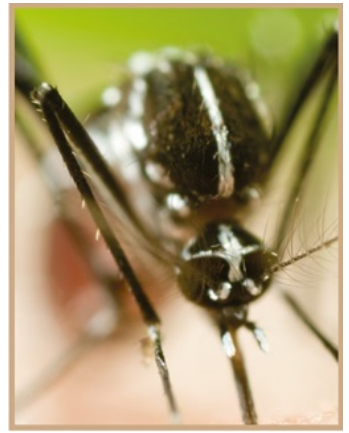
- **Dengue (Den.)** and **Chikungunya (Chik.)** endemic [*Beauté and Sirenda, 2010*]
- **Japanese Encephalitis (JE)** [*Cappelle et al. 2016*]

More than **300 mosquito species** described in Cambodia [*Maquart et al. in press, 2026*]

- Two main vectors of Den. and Chik. viruses: *Aedes aegypti* and *Aedes albopictus*
- Vector of JE virus: *Culex vishnui* group [*Cappelle et al. 2016*]



Aedes aegypti



Aedes albopictus

CONTEXT

Institut Pasteur du Cambodge
PASTEUR NETWORK

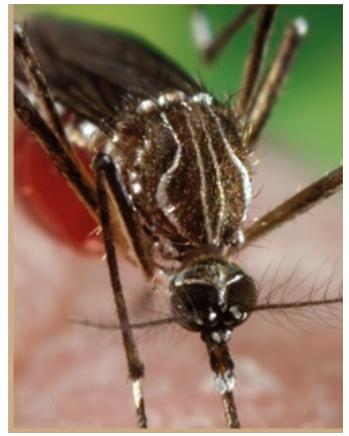
Cambodia

Favorable environment for **arboviruses**

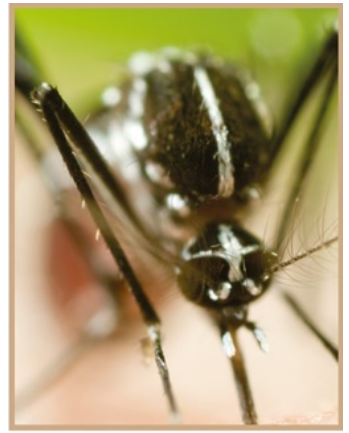
- **Dengue (Den.)** and **Chikungunya (Chik.)** endemic [*Beauté and Sirenda, 2010*]
- **Japanese Encephalitis (JE)** [*Cappelle et al. 2016*]

More than **300 mosquito species** described in Cambodia [*Maquart et al. in press, 2026*]

- Two main vectors of Den. and Chik. viruses: *Aedes aegypti* and *Aedes albopictus*
- Vector of JE virus: *Culex vishnui* group [*Cappelle et al. 2016*]



Aedes aegypti



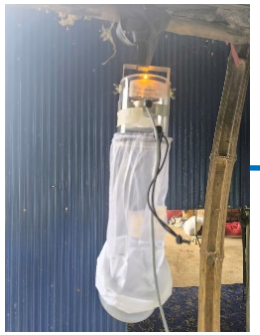
Aedes albopictus

Need to identify mosquito species for vector control and surveillance

BACKGROUND STUDY

Institut Pasteur du Cambodge
PASTEUR NETWORK

Collect & systematically identify vector species from the field



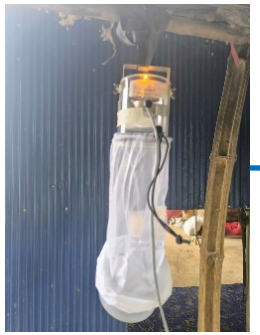
Mosquito collection

Vector species ?

BACKGROUND STUDY

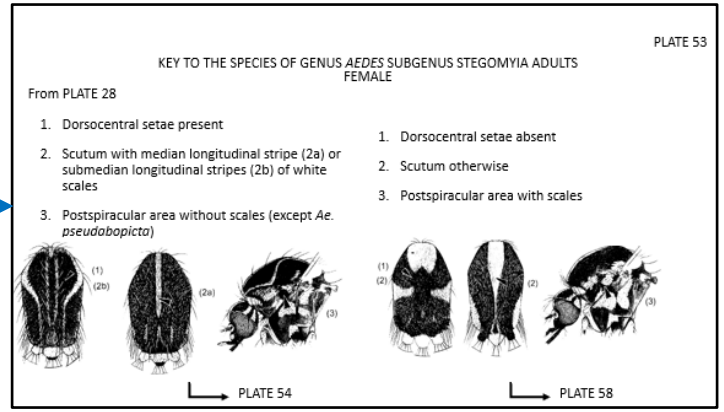
Institut Pasteur du Cambodge
PASTEUR NETWORK

Collect & systematically identify vector species from the field



Mosquito collection

Vector species ?

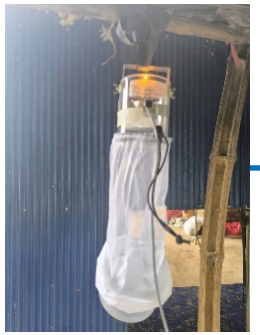


Morphological identification (Gold Standard)

BACKGROUND STUDY

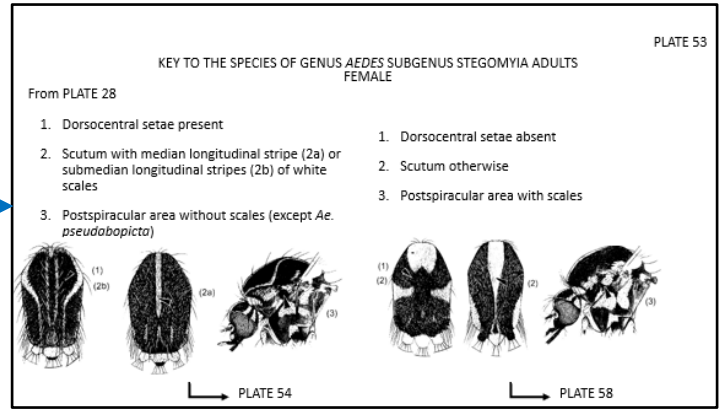
Institut Pasteur du Cambodge
PASTEUR NETWORK

Collect & systematically identify vector species from the field



Mosquito collection

Vector species ?



Morphological identification (Gold Standard)

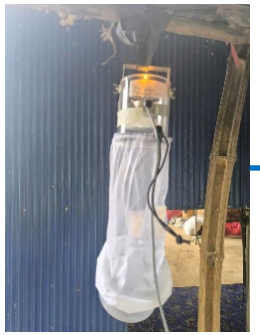
✓ Advantages

- Very low cost
- Fast results (< 1 minute)
- Usable directly in the field (microscope only)

BACKGROUND STUDY

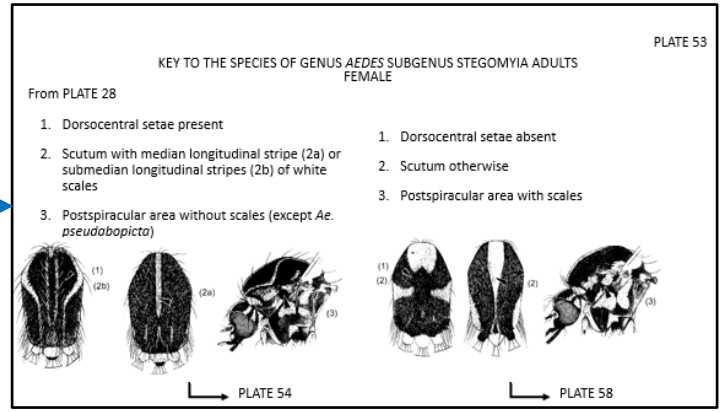
Institut Pasteur du Cambodge
PASTEUR NETWORK

Collect & systematically identify vector species from the field



Mosquito collection

Vector species ?



Morphological identification (Gold Standard)

✓ Advantages

- Very low cost
- Fast results (< 1 minute)
- Usable directly in the field (microscope only)

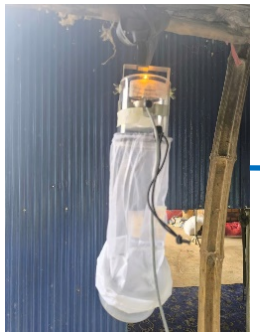
Disadvantages

- Highly skilled entomologist required
- Time/sample handling increases error risk
- **Not adapted for complex group identification**

BACKGROUND STUDY

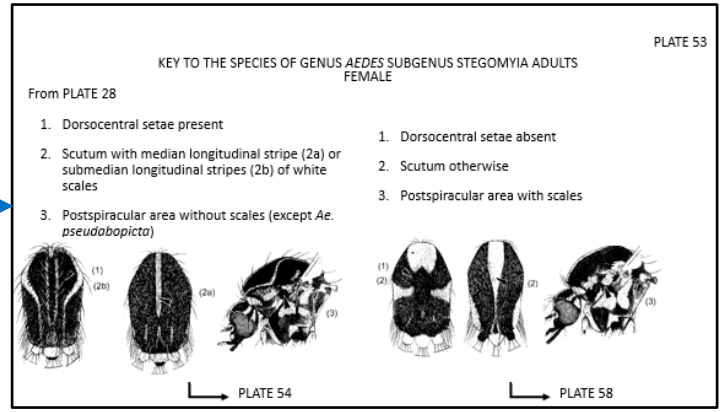
Institut Pasteur du Cambodge
PASTEUR NETWORK

Collect & systematically identify vector species from the field



Mosquito collection

Vector species ?



Morphological identification (Gold Standard)

✓ Advantages

- Very low cost
- Fast results (< 1 minute)
- Usable directly in the field (microscope only)

Disadvantages

- Highly skilled entomologist required
- Time/sample handling increases error risk
- **Not adapted for complex group identification**

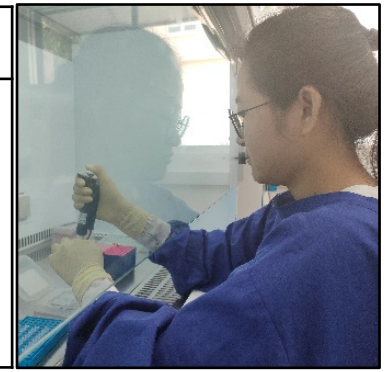
Requires an alternative method: complex species groups

BACKGROUND STUDY

Institut Pasteur du Cambodge
PASTEUR NETWORK

Molecular identification for complex group

Molecular identification	
✓ **Advantages**	**Disadvantages**
More reliable than morpho	Higher cost (machine + reagent) [*Rakotonirina et al.* 2025]
Require less technical expertise	Need a reference Database [*Weis et al. 2020*]
	Time consuming

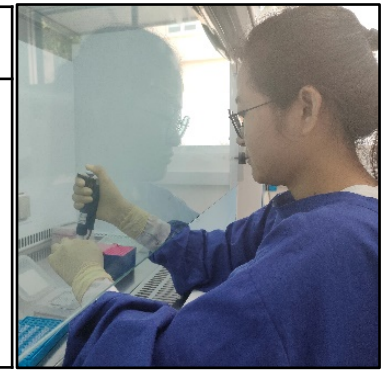


BACKGROUND STUDY

Institut Pasteur du Cambodge
PASTEUR NETWORK

Molecular identification for complex group

Molecular identification	
✓ **Advantages**	**Disadvantages**
More reliable than morpho	Higher cost (machine + reagent) [*Rakotonirina et al.* 2025]
Require less technical expertise	Need a reference Database [*Weis et al. 2020*]
	Time consuming



DNA barcoding (Gold Standard)

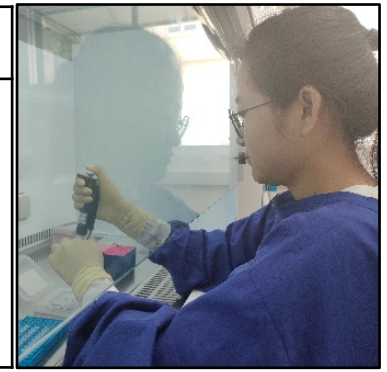
MALDI-TOF Mass Spectrometry

BACKGROUND STUDY

Institut Pasteur du Cambodge
PASTEUR NETWORK

Molecular identification for complex group

Molecular identification	
✓ **Advantages**	**Disadvantages**
More reliable than morpho	Higher cost (machine + reagent) [*Rakotonirina et al.* 2025]
Require less technical expertise	Need a reference Database [*Weis et al. 2020*]
	Time consuming



DNA barcoding (Gold Standard)

- Reliable (Molecular Gold Standard) [*Beebe, 2018*]
- High cost per sample [*Rakotonirina et al.* 2025]
- Long processing time per sample
  - Sample Preparation
  - PCR Amplification
  - Sequencing (South Korea)
  - Analysis

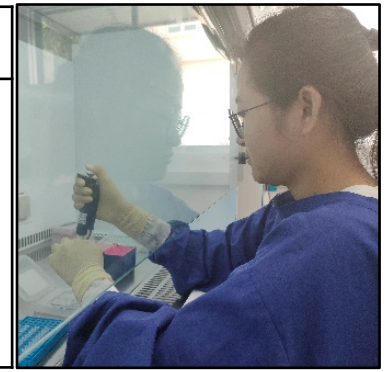
MALDI-TOF Mass Spectrometry

BACKGROUND STUDY

Institut Pasteur du Cambodge
PASTEUR NETWORK

Molecular identification for complex group

Molecular identification	
✓ **Advantages**	**Disadvantages**
More reliable than morpho	Higher cost (machine + reagent) [*Rakotonirina et al.* 2025]
Require less technical expertise	Need a reference Database [*Weis et al. 2020*]
	Time consuming



DNA barcoding (Gold Standard)

- Reliable (Molecular Gold Standard) [*Beebe, 2018*]
- High cost per sample [*Rakotonirina et al.* 2025]
- Long processing time per sample
  - Sample Preparation
  - PCR Amplification
  - Sequencing (South Korea)
  - Analysis

MALDI-TOF Mass Spectrometry

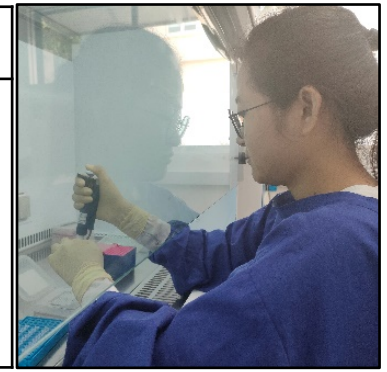
- Low cost and time per sample [*Rakotonirina et al.* 2025]
- Adapted for mosquito identification but not for complex
- Cost of the machine (> 300,000 $)
- Bruker proprietary software [*Weis et al. 2020*]

BACKGROUND STUDY

Institut Pasteur du Cambodge
PASTEUR NETWORK

Molecular identification for complex group

Molecular identification	
✓ **Advantages**	**Disadvantages**
More reliable than morpho	Higher cost (machine + reagent) [*Rakotonirina et al.* 2025]
Require less technical expertise	Need a reference Database [*Weis et al. 2020*]
	Time consuming



DNA barcoding (Gold Standard)

- Reliable (Molecular Gold Standard) [*Beebe, 2018*]
- High cost per sample [*Rakotonirina et al.* 2025]
- Long processing time per sample
  - Sample Preparation
  - PCR Amplification
  - Sequencing (South Korea)
  - Analysis

MALDI-TOF Mass Spectrometry

- Low cost and time per sample [*Rakotonirina et al.* 2025]
- Adapted for mosquito identification but not for complex
- Cost of the machine (> 300,000 $)
- Bruker proprietary software [*Weis et al. 2020*]

Need computer science expertise

BACKGROUND STUDY

Institut Pasteur du Cambodge

PASTEUR NETWORK

How can Machine Learning (ML) contribute to help Entomology ?

BACKGROUND STUDY

Institut Pasteur du Cambodge
PASTEUR NETWORK

How can Machine Learning (ML) contribute to help Entomology ?

First idea: Photos

Exists but limited:

- Specimen fragility affects accuracy
- Not suitable for complex group

Cx. pseudovishnui

Cx. vishnui

Cx. tritaeniorhynchus

BACKGROUND STUDY

Institut Pasteur du Cambodge
PASTEUR NETWORK

How can Machine Learning (ML) contribute to help Entomology ?

First idea: Photos

Exists but limited:

- Specimen fragility affects accuracy
- ~~Not suitable for complex group~~

Cx. pseudovishnui

Cx. vishnui

Cx. tritaeniorhynchus

Second idea: MALDI-TOF Mass Spectrometry

- MALDI-TOF + Bruker Software [*Yssouf et al. 2013*; *Rakotonirina et al. 2020*]
- MALDI-TOF + Machine Learning [*Merchan et al. 2023*]

BACKGROUND STUDY

Institut Pasteur du Cambodge
PASTEUR NETWORK

How can Machine Learning (ML) contribute to help Entomology ?

First idea: Photos

Exists but limited:

- Specimen fragility affects accuracy
- ~~Not suitable for complex group~~

Cx. pseudovishnui

Cx. vishnui

Cx. tritaeniorhynchus

Second idea: MALDI-TOF Mass Spectrometry

- MALDI-TOF + Bruker Software [*Yssouf et al. 2013*; *Rakotonirina et al. 2020*]
- MALDI-TOF + Machine Learning [*Merchan et al. 2023*]

Develop MALDI-TOF identification using machine learning

OBJECTIVES

Institut Pasteur du Cambodge

PASTEUR NETWORK

Integrating MALDI-TOF MS and Machine Learning for Rapid Mosquito Species Identification

Develop a faster, cheaper and more accurate tool for identifying mosquito species:

OBJECTIVES

Institut Pasteur du Cambodge

PASTEUR NETWORK

Integrating MALDI-TOF MS and Machine Learning for Rapid Mosquito Species Identification

Develop a faster, cheaper and more accurate tool for identifying mosquito species:

- Faster: 1 working day vs. several days for the molecular gold standard

OBJECTIVES

Institut Pasteur du Cambodge

PASTEUR NETWORK

Integrating MALDI-TOF MS and Machine Learning for Rapid Mosquito Species Identification

Develop a faster, cheaper and more accurate tool for identifying mosquito species:

- Faster: 1 working day vs. several days for the molecular gold standard
- Cheaper: ~5× lower cost per sample than the molecular gold standard

OBJECTIVES

Institut Pasteur du Cambodge

PASTEUR NETWORK

Integrating MALDI-TOF MS and Machine Learning for Rapid Mosquito Species Identification

Develop a faster, cheaper and more accurate tool for identifying mosquito species:

- Faster: 1 working day vs. several days for the molecular gold standard
- Cheaper: ~5× lower cost per sample than the molecular gold standard
- Reliable: Same reliability as the molecular gold standard

Matrix-Assisted Laser Desorption and Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF)

Institut Pasteur du Cambodge

PASTEUR NETWORK

Used since the early 2000s for microorganism identification [*Lay, 2001*]

Routinely used in clinical microbiology laboratories for diagnosis [*Seng P, et al., 2010*]

Matrix-Assisted Laser Desorption and Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF)

Institut Pasteur du Cambodge
PASTEUR NETWORK

Used since the early 2000s for microorganism identification [*Lay, 2001*]

Routinely used in clinical microbiology laboratories for diagnosis [*Seng P, et al., 2010*]

Mosquito

Legs

Matrix-Assisted Laser Desorption and Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF)

Institut Pasteur du Cambodge
PASTEUR NETWORK

Used since the early 2000s for microorganism identification [*Lay, 2001*]

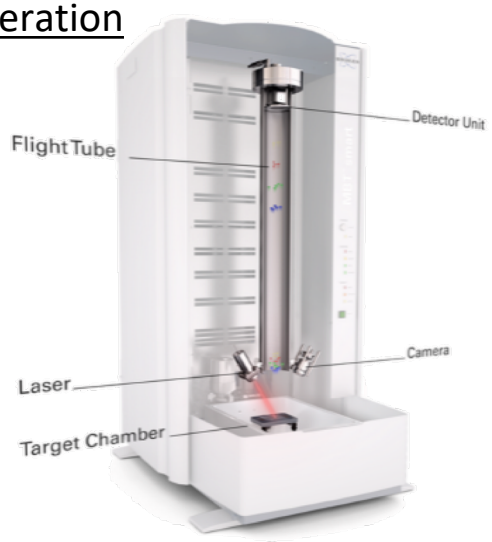
Routinely used in clinical microbiology laboratories for diagnosis [*Seng P, et al., 2010*]

Mosquito

Legs

Spectra generation

Laser shot variability affects reproducibility



Matrix-Assisted Laser Desorption and Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF)

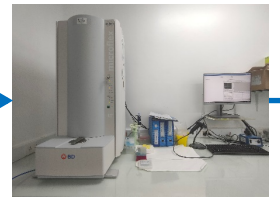
Institut Pasteur du Cambodge
PASTEUR NETWORK

Used since the early 2000s for microorganism identification [*Lay, 2001*]

Routinely used in clinical microbiology laboratories for diagnosis [*Seng P, et al., 2010*]

Mosquito

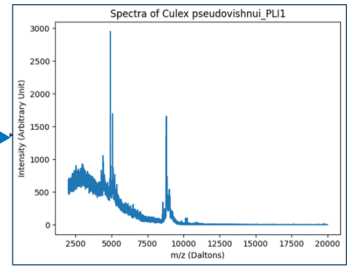
Legs



Spectra generation

1999.89 8447	1999.89 5932	1999.89 8175	1999.89 0	1999.89 5479
2000.31 8655	2000.31 5923	2000.31 8257	2000.31 0	2000.31 5405
2000.73 8655	2000.73 5902	2000.73 8335	2000.73 0	2000.73 5484
2001.15 8568	2001.15 5856	2001.15 8345	2001.15 0	2001.15 5448
2001.57 8509	2001.57 5867	2001.57 8346	2001.57 0	2001.57 5349
2001.98 8471	2001.98 5855	2001.98 8422	2001.98 0	2001.98 5387
2002.40 8550	2002.40 5942	2002.40 8503	2002.40 0	2002.40 5460
2002.82 8528	2002.82 5891	2002.82 8536	2002.82 0	2002.82 5424
2003.24 8484	2003.24 5823	2003.24 8434	2003.24 0	2003.24 5348
2003.66 8535	2003.66 5795	2003.66 8495	2003.66 0	2003.66 5402
2004.08 8474	2004.08 5812	2004.08 8408	2004.08 0	2004.08 5544

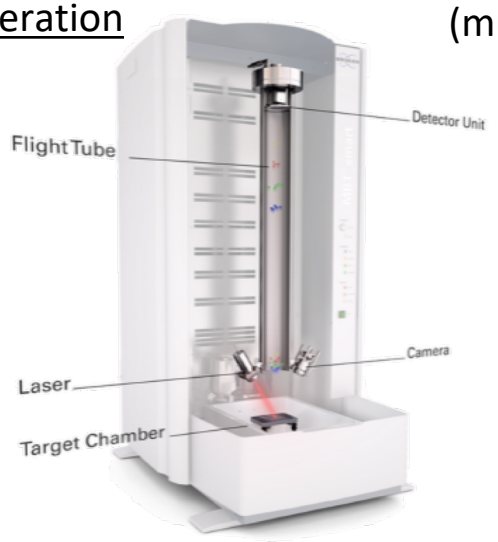
(multiple spectra)



Representation

Laser shot variability affects reproducibility

One mosquito generates **24 spectra (or 8)**



Matrix-Assisted Laser Desorption and Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF)

Institut Pasteur du Cambodge

PASTEUR NETWORK

Two Approaches for Species Identification

Matrix-Assisted Laser Desorption and Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF)

Institut Pasteur du Cambodge
PASTEUR NETWORK

Two Approaches for Species Identification

Bruker software approach: Automated preprocessing + Comparison against the Bruker reference database

Result Overview

Sample Name	Sample ID	Organism (best match)	Score Value	Organism (second-best match)	Score Value
D12 (+++) (A)	aeg2 (Standard)	Aedes aegypti	2.13	Aedes aegypti	2.10
E1 (-) (C)	albo1 (Standard)	**No Organism Identification Possible**	1.06	**No Organism Identification Possible**	0.46
E2 (-) (C)	albo2 (Standard)	**No Organism Identification Possible**	0.90	**No Organism Identification Possible**	0.51
E3 (+++) (A)	cq1 (Standard)	Culex quinquefaciatus	2.11	**No Organism Identification Possible**	0.85
D11 (+++) (A)	aeg1 (Standard)	Aedes aegypti	2.18	Aedes aegypti	2.15

Matrix-Assisted Laser Desorption and Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF)

Institut Pasteur du Cambodge
PASTEUR NETWORK

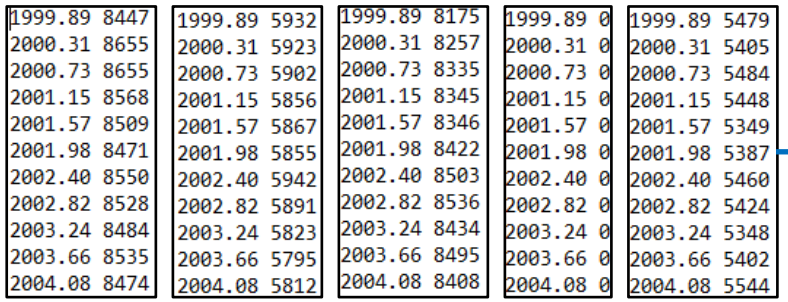
Two Approaches for Species Identification

Bruker software approach: Automated preprocessing + Comparison against the Bruker reference database

Result Overview

Sample Name	Sample ID	Organism (best match)	Score Value	Organism (second-best match)	Score Value
D12 (+++) (A)	aeg2 (Standard)	Aedes aegypti	2.13	Aedes aegypti	2.10
E1 (-) (C)	albo1 (Standard)	No Organism Identification Possible	1.06	No Organism Identification Possible	0.46
E2 (-) (C)	albo2 (Standard)	No Organism Identification Possible	0.90	No Organism Identification Possible	0.51
E3 (+++) (A)	cq1 (Standard)	Culex quinquefaciatus	2.11	No Organism Identification Possible	0.85
D11 (+++) (A)	aeg1 (Standard)	Aedes aegypti	2.18	Aedes aegypti	2.15

Our Custom Workflow: Export MALDI-TOF spectra as text files



1999.89 8447	1999.89 5932	1999.89 8175	1999.89 0	1999.89 5479
2000.31 8655	2000.31 5923	2000.31 8257	2000.31 0	2000.31 5405
2000.73 8655	2000.73 5902	2000.73 8335	2000.73 0	2000.73 5484
2001.15 8568	2001.15 5856	2001.15 8345	2001.15 0	2001.15 5448
2001.57 8509	2001.57 5867	2001.57 8346	2001.57 0	2001.57 5349
2001.98 8471	2001.98 5855	2001.98 8422	2001.98 0	2001.98 5387
2002.40 8550	2002.40 5942	2002.40 8503	2002.40 0	2002.40 5460
2002.82 8528	2002.82 5891	2002.82 8536	2002.82 0	2002.82 5424
2003.24 8484	2003.24 5823	2003.24 8434	2003.24 0	2003.24 5348
2003.66 8535	2003.66 5795	2003.66 8495	2003.66 0	2003.66 5402
2004.08 8474	2004.08 5812	2004.08 8408	2004.08 0	2004.08 5544

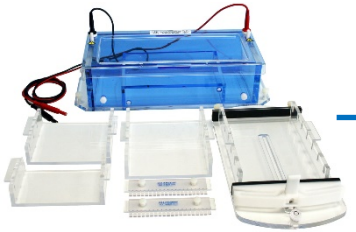
→ **Machine learning analysis**

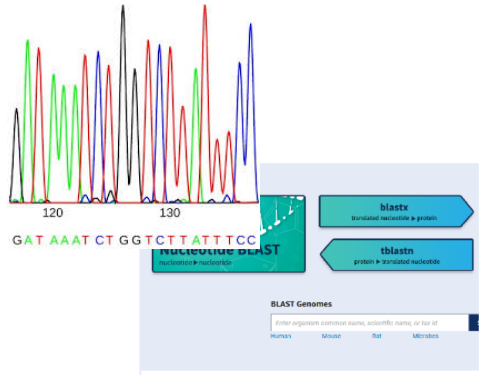
PCR - Label

Institut Pasteur du Cambodge

PASTEUR NETWORK

1st Step: PCR





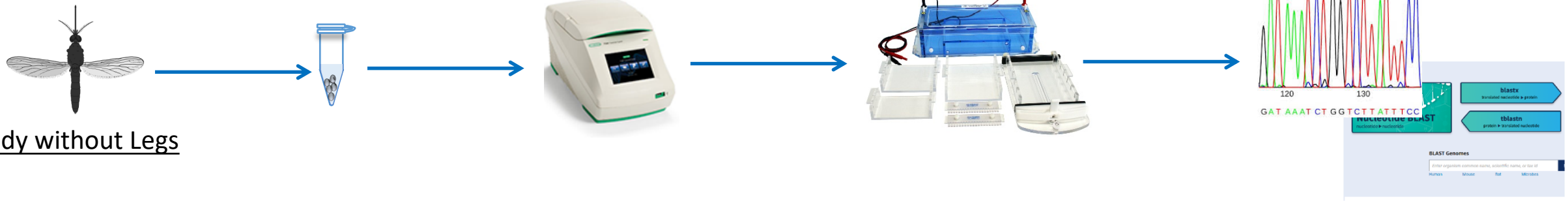
Body without Legs

PCR - Label

Institut Pasteur du Cambodge

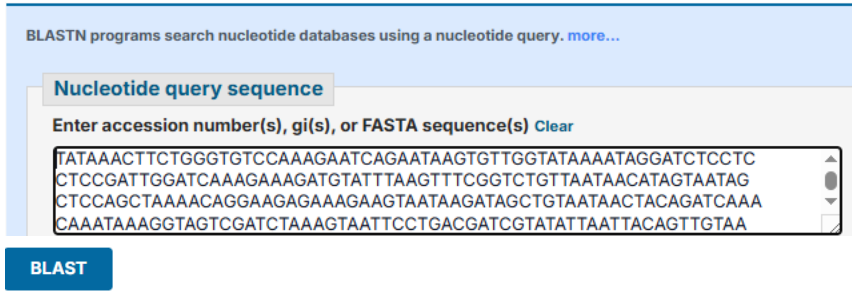
PASTEUR NETWORK

1st Step: PCR



Body without Legs

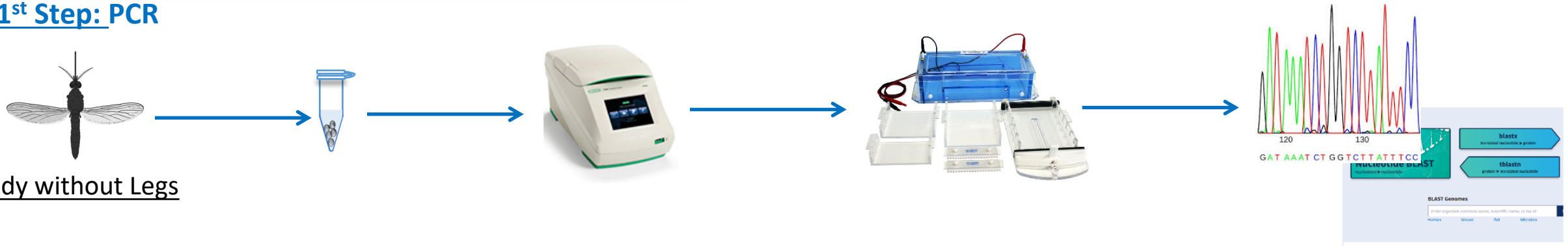
2nd Step: GenBank – Comparison against the reference database



PCR - Label

Institut Pasteur du Cambodge
PASTEUR NETWORK

1st Step: PCR



Body without Legs

2nd Step: GenBank – Comparison against the reference database

BLASTN programs search nucleotide databases using a nucleotide query. more...

Nucleotide query sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
TATAAACTTCTGGGTGTCCAAAGAATCAGAATAAGTGTTGGTATAAAATAGGATCTCCTC
CTCCGATTGGATCAAAGAAAGATGTATTTAAGTTTCGGTCTGTTAATAACATAGTAATAG
CTCCAGCTAAAACAGGAAGAGAAAGAAGTAATAAGATAGCTGTAATAACTACAGATCAAA
CAAATAAAGGTAGTCGATCTAAAGTAATTCCTGACGATCGTATATTAATTACAGTTGTAA
```

BLAST

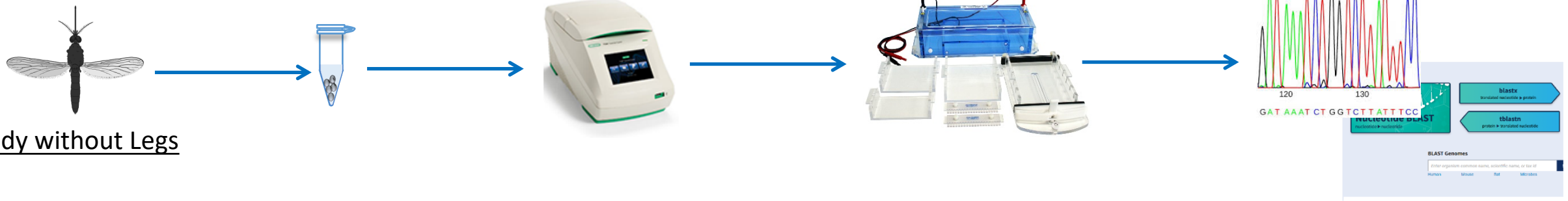
select all 100 sequences selected

	DESCRIPTION	SCIENTIFIC NAME	PER. IDENT
☑	Aedes aegypti mitochondrial partial coi gene for cytochrome oxidase I, from Libreville, Gabon	Aedes aegypti	100.00%
☑	Aedes aegypti isolate COI_Vietn cytochrome oxidase subunit I gene, partial cds; mitochondrial	Aedes aegypti	100.00%
☑	Aedes aegypti isolate Cambodia 3 cytochrome c oxidase subunit I (COI) gene, partial cds; mitochon...	Aedes aegypti	100.00%
☑	Aedes aegypti isolate WMELYOG#1 cytochrome c oxidase subunit I (COX1) gene, partial cds; mitochon...	Aedes aegypti	100.00%
☑	Aedes aegypti isolate WMELYOG#3 cytochrome c oxidase subunit I (COX1) gene, partial cds; mitochon...	Aedes aegypti	100.00%
☑	Aedes aegypti isolate YK_2020 mitochondrion, complete genome	Aedes aegypti	100.00%
☑	Aedes aegypti isolate GV_2020 mitochondrion, complete genome	Aedes aegypti	100.00%
☑	Aedes aegypti isolate wC45_F9 mitochondrion, complete genome	Aedes aegypti	100.00%
☑	Aedes aegypti isolate wC45_F10 mitochondrion, complete genome	Aedes aegypti	100.00%
☑	Aedes aegypti isolate Lab mitochondrion, complete genome	Aedes aegypti	100.00%
☑	Aedes aegypti isolate GV_2013 mitochondrion, complete genome	Aedes aegypti	100.00%
☑	Aedes aegypti isolate GV_2018 mitochondrion, complete genome	Aedes aegypti	100.00%
☑	Aedes aegypti isolate YK_2018 mitochondrion, complete genome	Aedes aegypti	100.00%
☑	Aedes aegypti isolate Aa_F3 mitochondrion, partial genome	Aedes aegypti	100.00%

PCR - Label

Institut Pasteur du Cambodge
PASTEUR NETWORK

1st Step: PCR



Body without Legs

2nd Step: GenBank – Comparison against the reference database

BLASTN programs search nucleotide databases using a nucleotide query. more...

Nucleotide query sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
TATAAACTTCTGGGTGTCCAAAGAATCAGAATAAGTGTTGGTATAAAATAGGATCTCCTC
CTCCGATTGGATCAAAGAAAGATGTATTTAAGTTTCGGTCTGTTAATAACATAGTAATAG
CTCCAGCTAAAACAGGAAGAGAAAGAAGTAATAAGATAGCTGTAATAACTACAGATCAAA
CAAATAAAGGTAGTCGATCTAAAGTAATTCCTGACGATCGTATATTAATTACAGTTGTAA
```

BLAST

select all 100 sequences selected

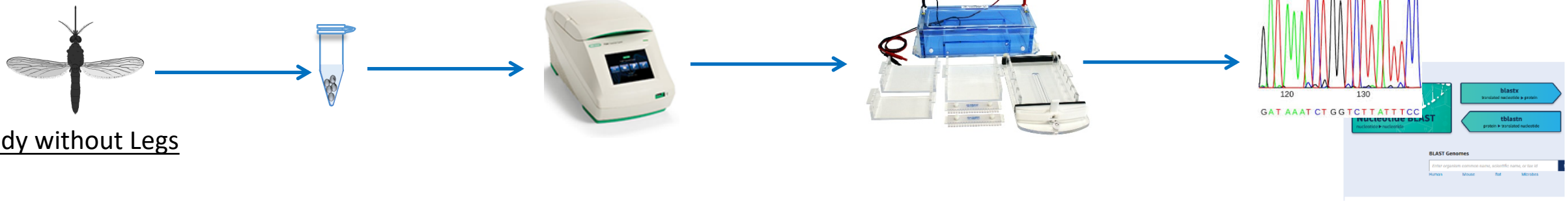
DESCRIPTION	SCIENTIFIC NAME	PER. IDENT
Aedes aegypti mitochondrial partial coi gene for cytochrome oxidase I, from Libreville, Gabon	Aedes aegypti	100.00%
Aedes aegypti isolate COI_Vietn cytochrome oxidase subunit I gene, partial cds; mitochondrial	Aedes aegypti	100.00%
Aedes aegypti isolate Cambodia 3 cytochrome c oxidase subunit I (COI) gene, partial cds; mitochon...	Aedes aegypti	100.00%
Aedes aegypti isolate WMELYOG#1 cytochrome c oxidase subunit I (COX1) gene, partial cds; mitochon...	Aedes aegypti	100.00%
Aedes aegypti isolate WMELYOG#3 cytochrome c oxidase subunit I (COX1) gene, partial cds; mitochon...	Aedes aegypti	100.00%
Aedes aegypti isolate YK_2020 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate GV_2020 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate wC45_F9 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate wC45_F10 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate Lab mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate GV_2013 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate GV_2018 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate YK_2018 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate Aa_F3 mitochondrion, partial genome	Aedes aegypti	100.00%

Label = *Aedes aegypti*

PCR - Label

Institut Pasteur du Cambodge
PASTEUR NETWORK

1st Step: PCR



Body without Legs

2nd Step: GenBank – Comparison against the reference database

BLASTN programs search nucleotide databases using a nucleotide query. more...

Nucleotide query sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
TATAAACTTCTGGGTGTCCAAAGAATCAGAATAAGTGTTGGTATAAAATAGGATCTCCTC
CTCCGATTGGATCAAAGAAAGATGTATTTAAGTTTCGGTCTGTTAATAACATAGTAATAG
CTCCAGCTAAAACAGGAAGAGAAAGAAGTAATAAGATAGCTGTAATAACTACAGATCAAA
CAAATAAAGGTAGTCGATCTAAAGTAATTCCTGACGATCGTATATTAATTACAGTTGTAA
```

BLAST

select all 100 sequences selected

DESCRIPTION	SCIENTIFIC NAME	PER. IDENT
Aedes aegypti mitochondrial partial coi gene for cytochrome oxidase I, from Libreville, Gabon	Aedes aegypti	100.00%
Aedes aegypti isolate COI_Vietn cytochrome oxidase subunit I gene, partial cds; mitochondrial	Aedes aegypti	100.00%
Aedes aegypti isolate Cambodia 3 cytochrome c oxidase subunit I (COI) gene, partial cds; mitochon...	Aedes aegypti	100.00%
Aedes aegypti isolate WMELYOG#1 cytochrome c oxidase subunit I (COX1) gene, partial cds; mitochon...	Aedes aegypti	100.00%
Aedes aegypti isolate WMELYOG#3 cytochrome c oxidase subunit I (COX1) gene, partial cds; mitochon...	Aedes aegypti	100.00%
Aedes aegypti isolate YK_2020 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate GV_2020 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate wC45_F9 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate wC45_F10 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate Lab mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate GV_2013 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate GV_2018 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate YK_2018 mitochondrion, complete genome	Aedes aegypti	100.00%
Aedes aegypti isolate Aa_F3 mitochondrion, partial genome	Aedes aegypti	100.00%

We have MALDI-TOF spectra + associated species

Label = *Aedes aegypti*

DATA

Institut Pasteur du Cambodge
PASTEUR NETWORK

Number of specimen per species

```
Number of individuals per label:
label
Aedes aegypti                35
Aedes albopictus             35
Anopheles peditaeniatus       5
Anopheles sinensis            4
Anopheles vagus               9
Armigeres subalbatus         32
Culex brevipalpis            31
Culex epidesmus               8
Culex gelidus                 4
Culex nigropunctatus          4
Culex pseudovishnui          34
Culex quinquefasciatus       30
Culex sitiens                 5
Culex tritaeniorhynchus      63
Culex vishnui                62
Lutzia fuscana               30
Mansonia indiana              6
Mansonia uniformis            3
Mimomyia luzonensis           4
Name: individual, dtype: int64
```

Data for mosquitoes

- **404** individuals
- **19** species
- 7 genus
- **97%** DNA confirmed

DATA

Institut Pasteur du Cambodge
PASTEUR NETWORK

Number of specimen per species

```
Number of individuals per label:
label
Aedes aegypti                35
Aedes albopictus             35
Anopheles peditaeniatus       5
Anopheles sinensis            4
Anopheles vagus               9
Armigeres subalbatus         32
Culex brevipalpis            31
Culex epidesmus               8
Culex gelidus                 4
Culex nigropunctatus          4
Culex pseudovishnui          34
Culex quinquefasciatus       30
Culex sitiens                 5
Culex tritaeniorhynchus      63
Culex vishnui                62
Lutzia fuscana               30
Mansonia indiana              6
Mansonia uniformis            3
Mimomyia luzonensis           4
Name: individual, dtype: int64
```

Data for mosquitoes

- **404** individuals
- **19** species
- 7 genus
- **97%** DNA confirmed

⚠ **Few data compare to classical Machine learning**

DATA

Institut Pasteur du Cambodge
PASTEUR NETWORK

Number of specimen per species

```
Number of individuals per label:
label
Aedes aegypti                35
Aedes albopictus             35
Anopheles peditaeniatus       5
Anopheles sinensis            4
Anopheles vagus               9
Armigeres subalbatus         32
Culex brevipalpis            31
Culex epidesmus               8
Culex gelidus                 4
Culex nigropunctatus          4
Culex pseudovishnui          34
Culex quinquefasciatus       30
Culex sitiens                 5
Culex tritaeniorhynchus      63
Culex vishnui                62
Lutzia fuscana               30
Mansonia indiana              6
Mansonia uniformis            3
Mimomyia luzonensis           4
Name: individual, dtype: int64
```

Data for mosquitoes

- **404** individuals
- **19** species
- 7 genus
- **97%** DNA confirmed

⚠ **Few data compare to classical Machine learning**

$$Accuracy = \frac{\text{Number of correct identification}}{\text{Total number of identification}}$$

DATA

Institut Pasteur du Cambodge
PASTEUR NETWORK

Number of specimen per species

```
Number of individuals per label:
label
Aedes aegypti                35
Aedes albopictus             35
Anopheles peditaeniatus       5
Anopheles sinensis            4
Anopheles vagus               9
Armigeres subalbatus         32
Culex brevipalpis            31
Culex epidesmus               8
Culex gelidus                 4
Culex nigropunctatus          4
Culex pseudovishnui          34
Culex quinquefasciatus       30
Culex sitiens                 5
Culex tritaeniorhynchus      63
Culex vishnui                62
Lutzia fuscana               30
Mansonia indiana              6
Mansonia uniformis            3
Mimomyia luzonensis           4
Name: individual, dtype: int64
```

Data for mosquitoes

- **404** individuals
- **19** species
- 7 genus
- **97%** DNA confirmed

⚠ **Few data compare to classical Machine learning**

$$Accuracy = \frac{\text{Number of correct identification}}{\text{Total number of identification}}$$

But … class imbalance problem

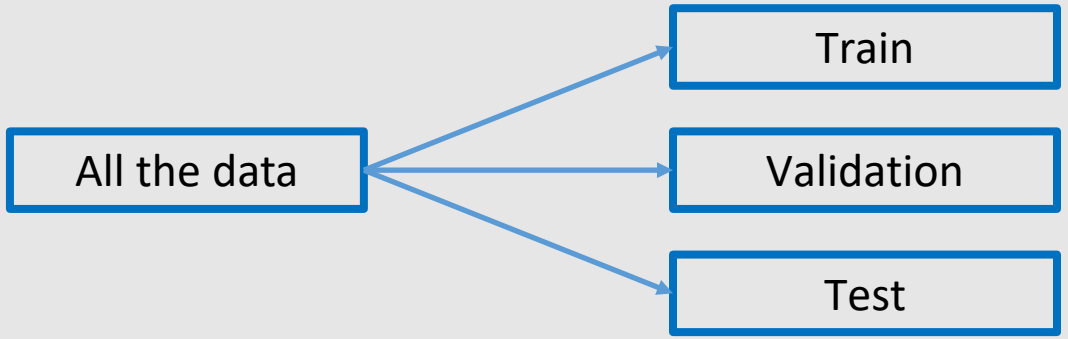
$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

MODEL CLASSIFICATION

Institut Pasteur du Cambodge
PASTEUR NETWORK

Train a Supervised machine learning

Step 1: Split randomly the data in three set



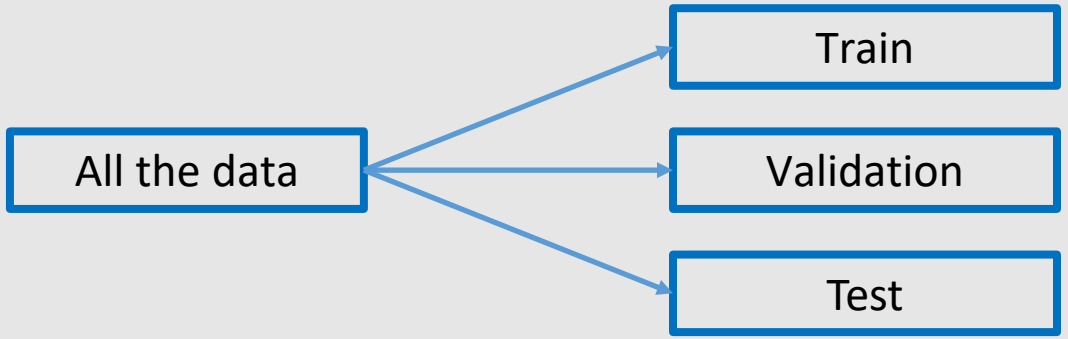
Split based on the number of individuals

MODEL CLASSIFICATION

Institut Pasteur du Cambodge
PASTEUR NETWORK

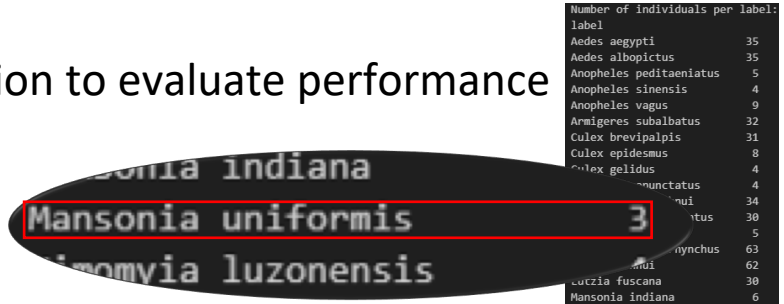
Train a Supervised machine learning

Step 1: Split randomly the data in three set



Split based on the number of individuals

Use 3-fold cross-validation to evaluate performance



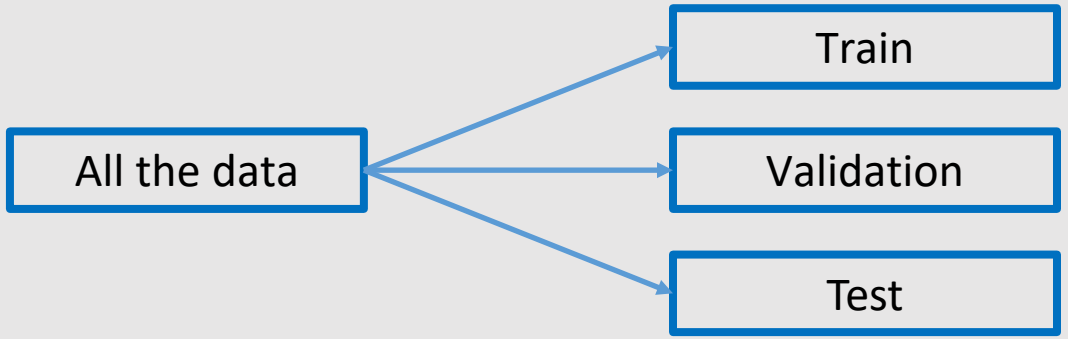

```
Number of individuals per label:
label
Aedes aegypti                35
Aedes albopictus             35
Anopheles peditaeniatus       5
Anopheles sinensis            4
Anopheles vagus               9
Armigeres subalbatus         32
Culex brevipalpis            31
Culex epidesmus               8
Culex gelidus                 4
                              4
                             34
                             30
                              5
                             63
                             62
Lutzia fuscana               30
Mansonia indiana              6
Mansonia uniformis            3
Mimomyia luzonensis           4
Name: individual, dtype: int64
```

MODEL CLASSIFICATION

Institut Pasteur du Cambodge
PASTEUR NETWORK

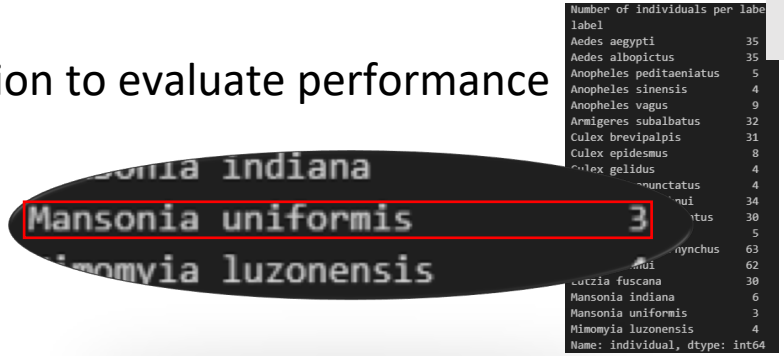
Train a Supervised machine learning

Step 1: Split randomly the data in three set

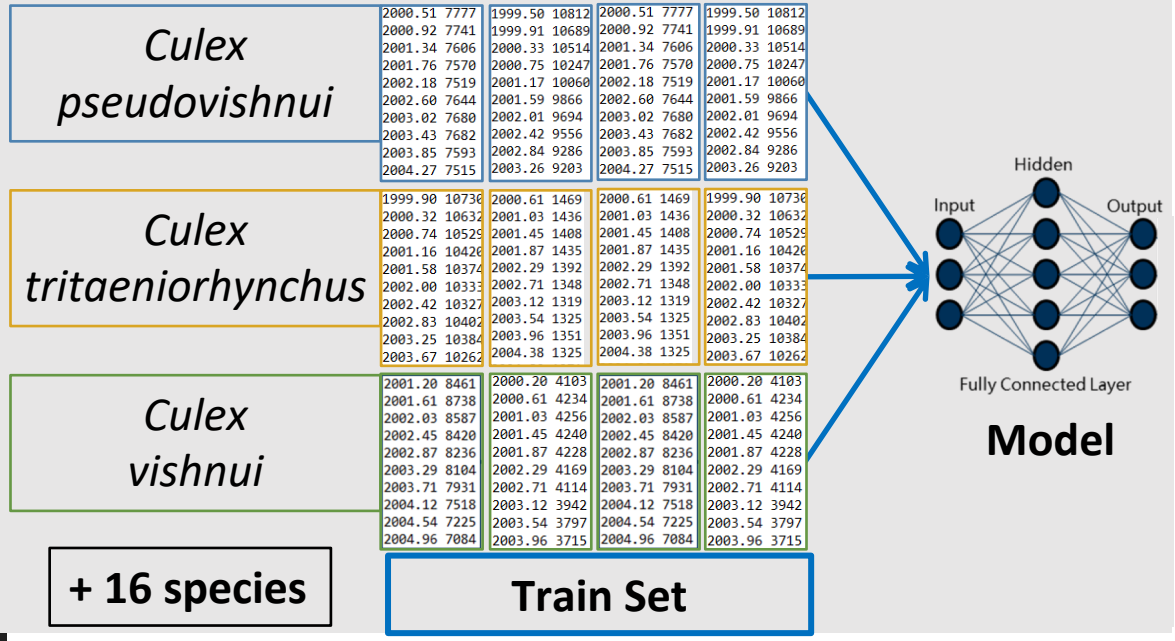


Split based on the number of individuals

Use 3-fold cross-validation to evaluate performance



Step 2: Train the model to learn to identify species

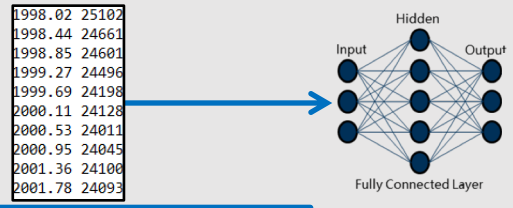


MODEL CLASSIFICATION

Institut Pasteur du Cambodge
PASTEUR NETWORK

Step 3: Evaluate the model on unseen data

Model Classification: Determine the species



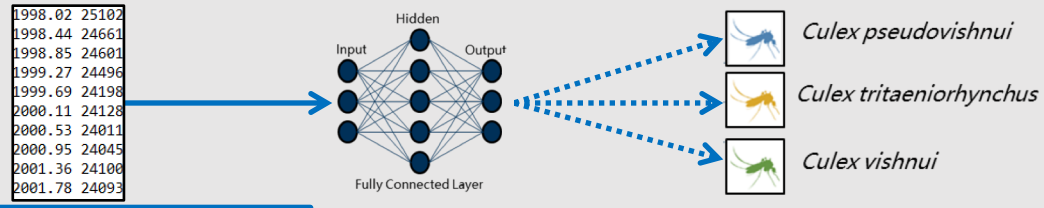
Test Set

MODEL CLASSIFICATION

Institut Pasteur du Cambodge
PASTEUR NETWORK

Step 3: Evaluate the model on unseen data

Model Classification: Determine the species



Test Set

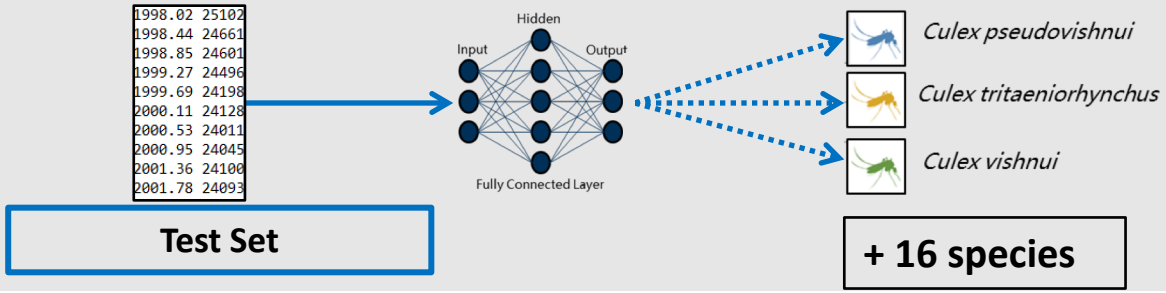
+ 16 species

MODEL CLASSIFICATION

Institut Pasteur du Cambodge
PASTEUR NETWORK

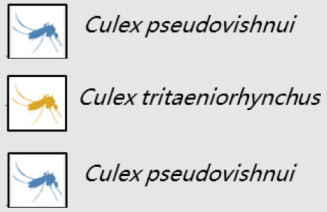
Step 3: Evaluate the model on unseen data

Model Classification: Determine the species



Evaluation: Compare results to the molecular classification

Model Classification

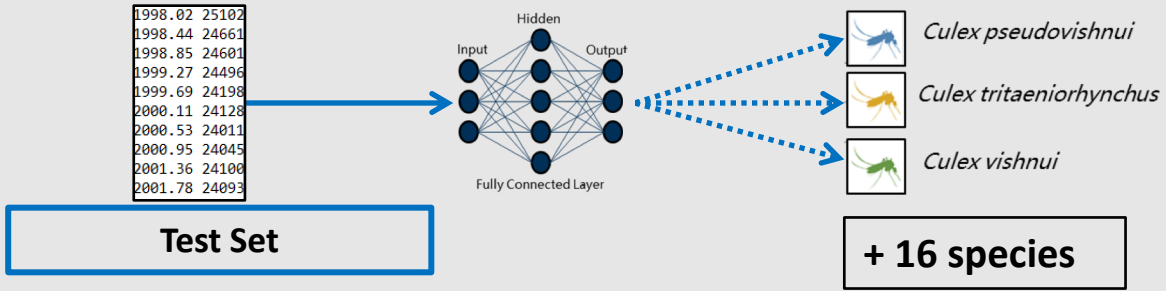


MODEL CLASSIFICATION

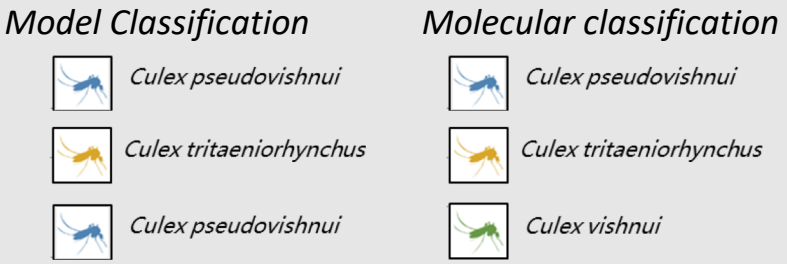
Institut Pasteur du Cambodge
PASTEUR NETWORK

Step 3: Evaluate the model on unseen data

Model Classification: Determine the species



Evaluation: Compare results to the molecular classification

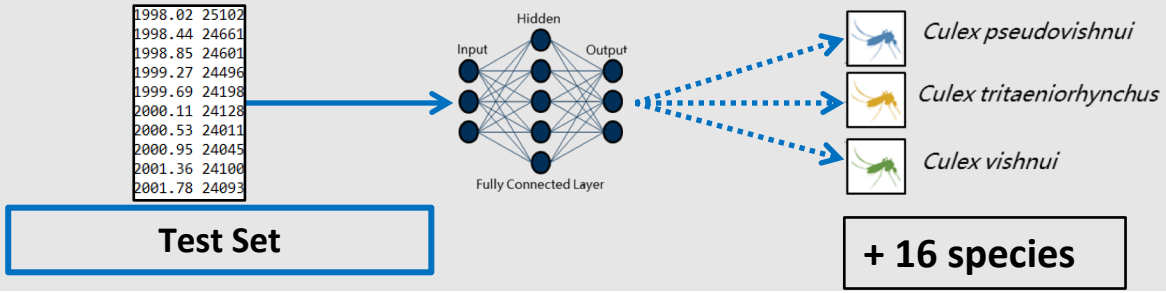


MODEL CLASSIFICATION

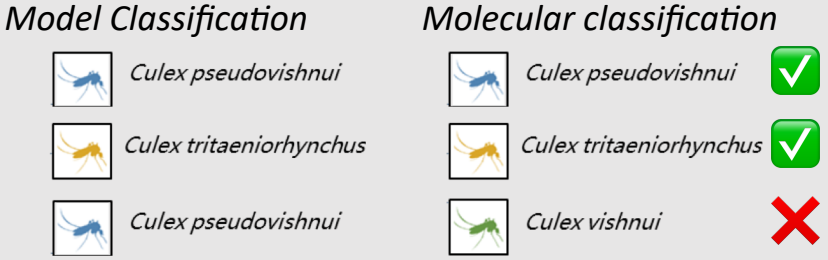
Institut Pasteur du Cambodge
PASTEUR NETWORK

Step 3: Evaluate the model on unseen data

Model Classification: Determine the species



Evaluation: Compare results to the molecular classification



Metrics

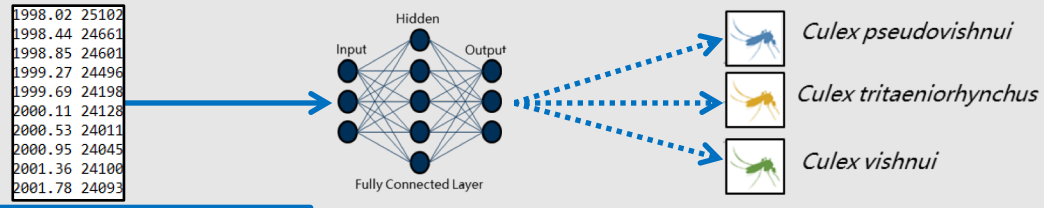
Score: 2/3 = 66.6%

MODEL CLASSIFICATION

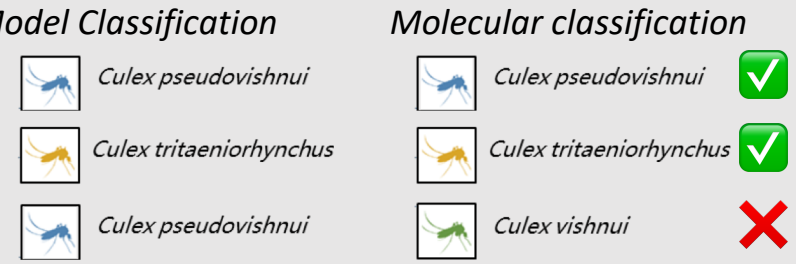
Institut Pasteur du Cambodge
PASTEUR NETWORK

Step 3: Evaluate the model on unseen data

Model Classification: Determine the species



Evaluation: Compare results to the molecular classification

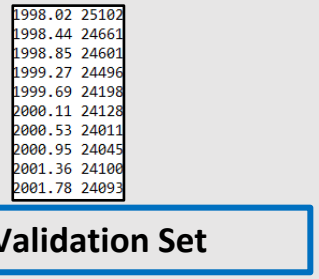


Metrics

Score: 2/3 = 66.6%

Step 2.5: Determine the best hyperparameters

On multiple data configuration



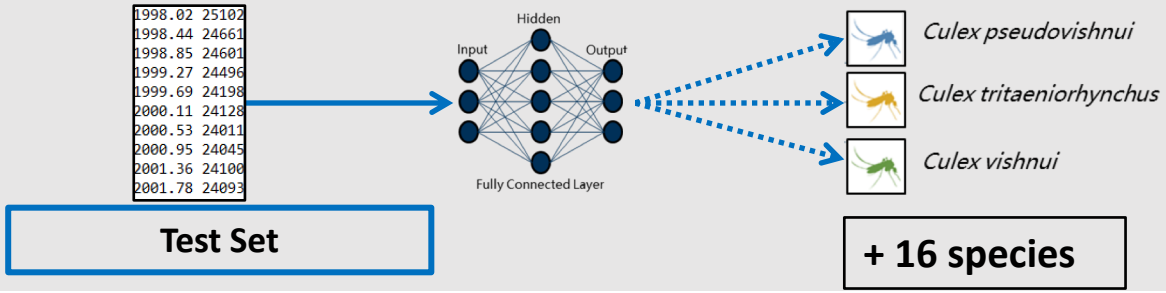
OPTUNA

MODEL CLASSIFICATION

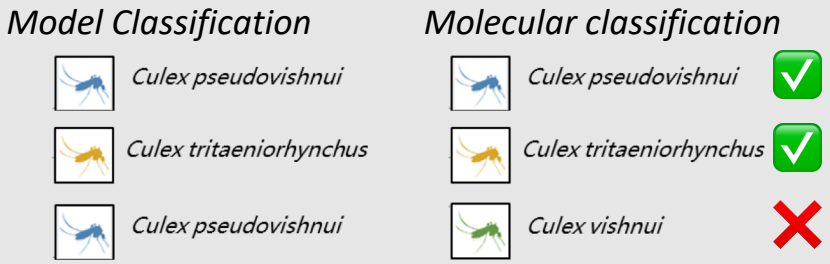
Institut Pasteur du Cambodge
PASTEUR NETWORK

Step 3: Evaluate the model on unseen data

Model Classification: Determine the species



Evaluation: Compare results to the molecular classification

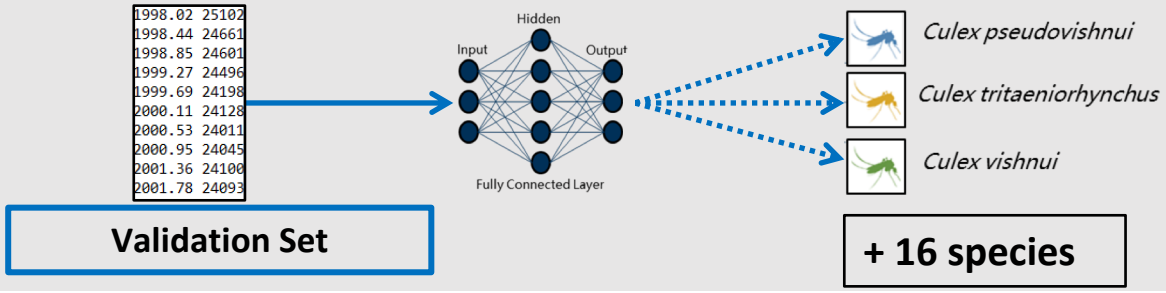


Metrics

Score: 2/3 = 66.6%

Step 2.5: Determine the best hyperparameters

On multiple data configuration



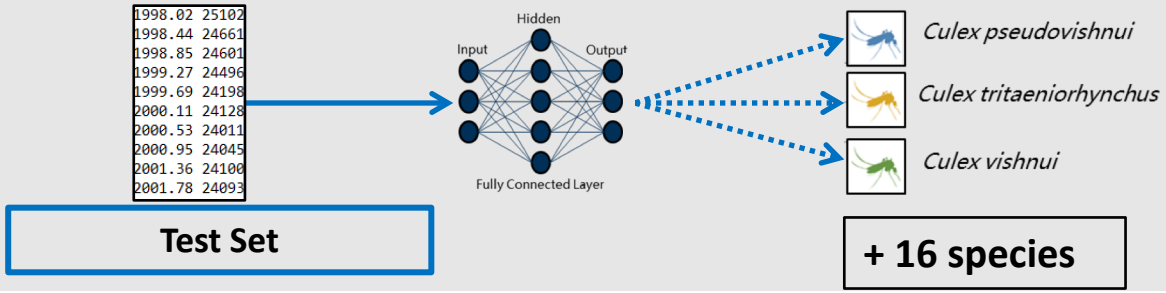
OPTUNA

MODEL CLASSIFICATION

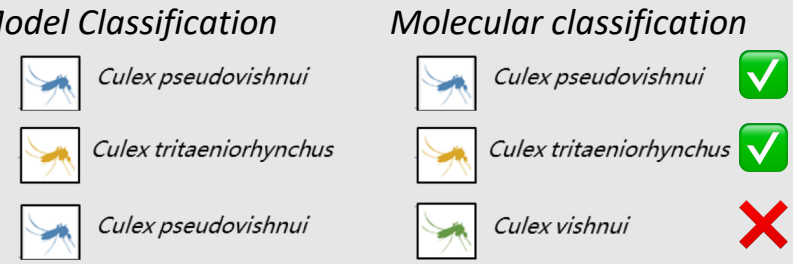
Institut Pasteur du Cambodge
PASTEUR NETWORK

Step 3: Evaluate the model on unseen data

Model Classification: Determine the species



Evaluation: Compare results to the molecular classification

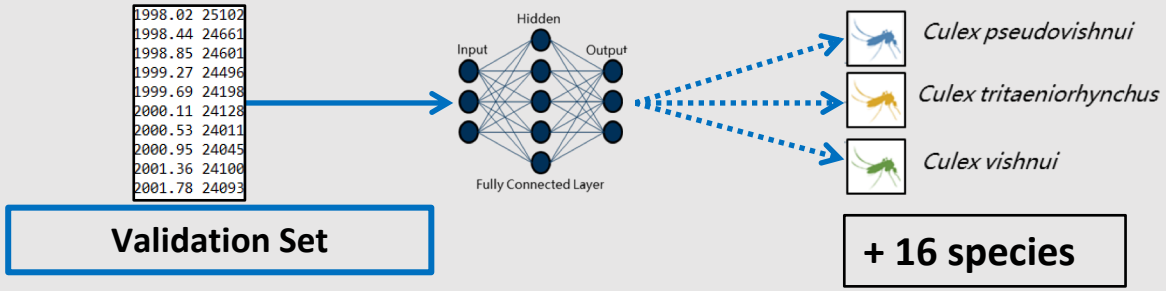


Metrics

Score: 2/3 = 66.6%

Step 2.5: Determine the best hyperparameters

On multiple data configuration



Evaluation: Compare results to the molecular classification

Save the hyperparameters corresponding to the best score
Retrain the model one last time using the optimal hyperparameters

OPTUNA

MODEL & PREPROCESSING

Institut Pasteur du Cambodge
PASTEUR NETWORK

- 6 different models

	1. SVM	2. Multi-Layer Perceptron	3. Logistic Regression	4. Random Forest	5. XGBoost	6. 1D CNN

MODEL & PREPROCESSING

Institut Pasteur du Cambodge
PASTEUR NETWORK

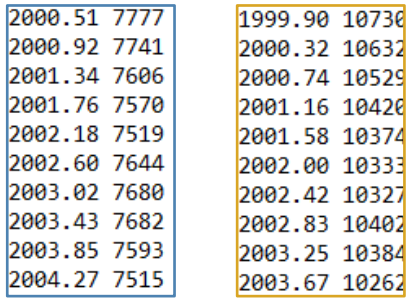
- 6 different models
- 7 different preprocessing methods:

	1. SVM	2. Multi-Layer Perceptron	3. Logistic Regression	4. Random Forest	5. XGBoost	6. 1D CNN
A) Interpolation						
B) Z-score						
C) TIC						
D) 90th quantile						
E) Baseline						
F) Smoothing						
G) Binning						

MODEL & PREPROCESSING

Institut Pasteur du Cambodge
PASTEUR NETWORK

- 6 different models
- 7 different preprocessing methods:
  - Interpolation (Mandatory) -> challenge: variable m/z value

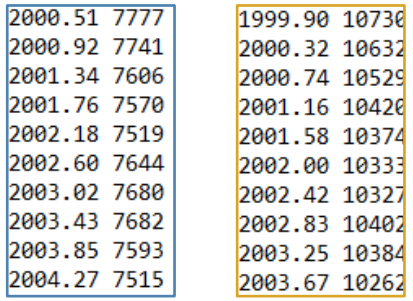


	1. SVM	2. Multi-Layer Perceptron	3. Logistic Regression	4. Random Forest	5. XGBoost	6. 1D CNN
A) Interpolation						
B) Z-score						
C) TIC						
D) 90th quantile						
E) Baseline						
F) Smoothing						
G) Binning						

MODEL & PREPROCESSING

Institut Pasteur du Cambodge
PASTEUR NETWORK

- 6 different models
- 7 different preprocessing methods:
  - Interpolation (Mandatory) -> challenge: variable m/z value



	1. SVM	2. Multi-Layer Perceptron	3. Logistic Regression	4. Random Forest	5. XGBoost	6. 1D CNN
A) Interpolation						
B) Z-score						
C) TIC						
D) 90th quantile						
E) Baseline						
F) Smoothing						
G) Binning						

Focus on the best-performing model + preprocessing combination

MODEL & PREPROCESSING

Institut Pasteur du Cambodge
PASTEUR NETWORK

1D Convolutional Neural Network

- Learns local patterns by sliding a 1D filter across the sequence
- Fast and efficient
- Well-adapted to time-series or sequential spectral data

MODEL & PREPROCESSING

Institut Pasteur du Cambodge
PASTEUR NETWORK

1D Convolutional Neural Network

- Learns local patterns by sliding a 1D filter across the sequence
- Fast and efficient
- Well-adapted to time-series or sequential spectral data

Binning

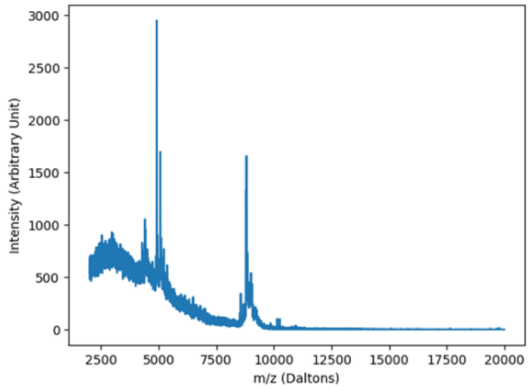
- Groups data into fixed-size windows
- Reduces noise and dimensionality

MODEL & PREPROCESSING

Institut Pasteur du Cambodge
PASTEUR NETWORK

1D Convolutional Neural Network

- Learns local patterns by sliding a 1D filter across the sequence
- Fast and efficient
- Well-adapted to time-series or sequential spectral data



Binning

- Groups data into fixed-size windows
- Reduces noise and dimensionality

20 000 peaks → 4 000 peaks (bin)

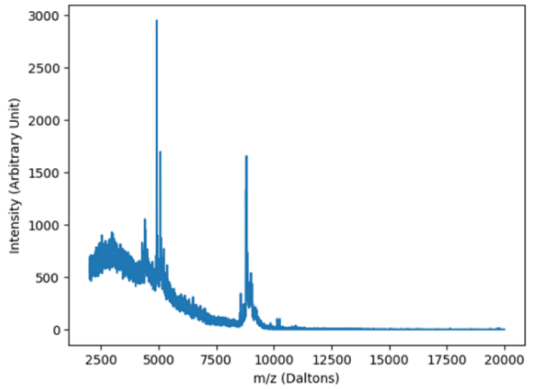
- With a window size of 5
- Aggregation methods: average, maximum …

MODEL & PREPROCESSING

Institut Pasteur du Cambodge
PASTEUR NETWORK

1D Convolutional Neural Network

- Learns local patterns by sliding a 1D filter across the sequence
- Fast and efficient
- Well-adapted to time-series or sequential spectral data



20 000 peaks → 4 000 peaks (bin)

Binning

- Groups data into fixed-size windows
- Reduces noise and dimensionality

- With a window size of 5
- Aggregation methods: average, maximum …

```
1998.02 24101
1998.44 23879
1998.85 23601
1999.27 23326
1999.69 23269
2000.11 23399
2000.53 23256
2000.95 23129
2001.36 23204
2001.78 23083
```

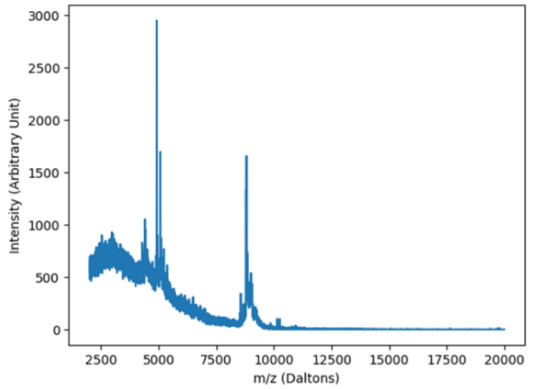
First 10 values of the spectrum

MODEL & PREPROCESSING

Institut Pasteur du Cambodge
PASTEUR NETWORK

1D Convolutional Neural Network

- Learns local patterns by sliding a 1D filter across the sequence
- Fast and efficient
- Well-adapted to time-series or sequential spectral data



20 000 peaks → 4 000 peaks (bin)

Binning

- Groups data into fixed-size windows
- Reduces noise and dimensionality

- With a window size of 5
- Aggregation methods: average, maximum …

1998.02 24101
1998.44 23879
1998.85 23601
1999.27 23326
1999.69 23269
2000.11 23399
2000.53 23256
2000.95 23129
2001.36 23204
2001.78 23083

First 10 values of the spectrum

Bin 1

1998.02 24101
1998.44 23879
1998.85 23601
1999.27 23326
1999.69 23269

average = 23635.2
maximum = 24101

Bin 2

2000.11 23399
2000.53 23256
2000.95 23129
2001.36 23204
2001.78 23083

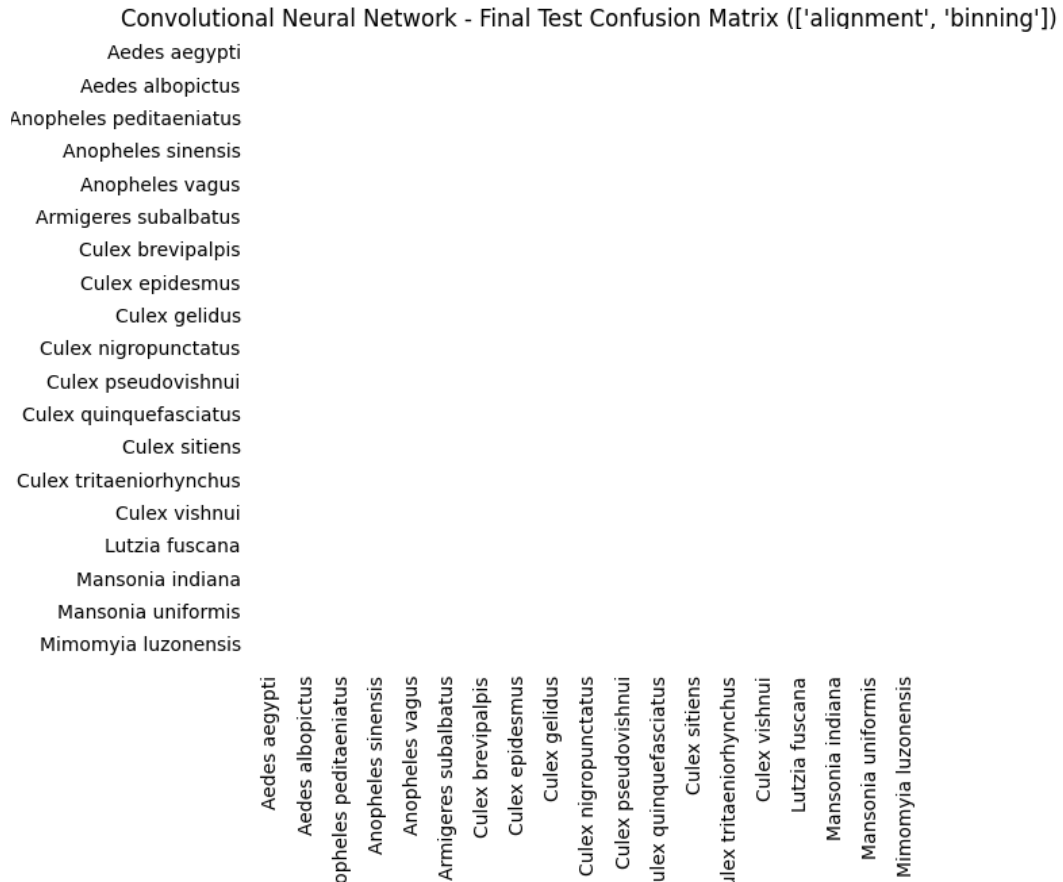
average = 23214.2
maximum = 23399

PRELIMINARY RESULTS

Institut Pasteur du Cambodge

PASTEUR NETWORK

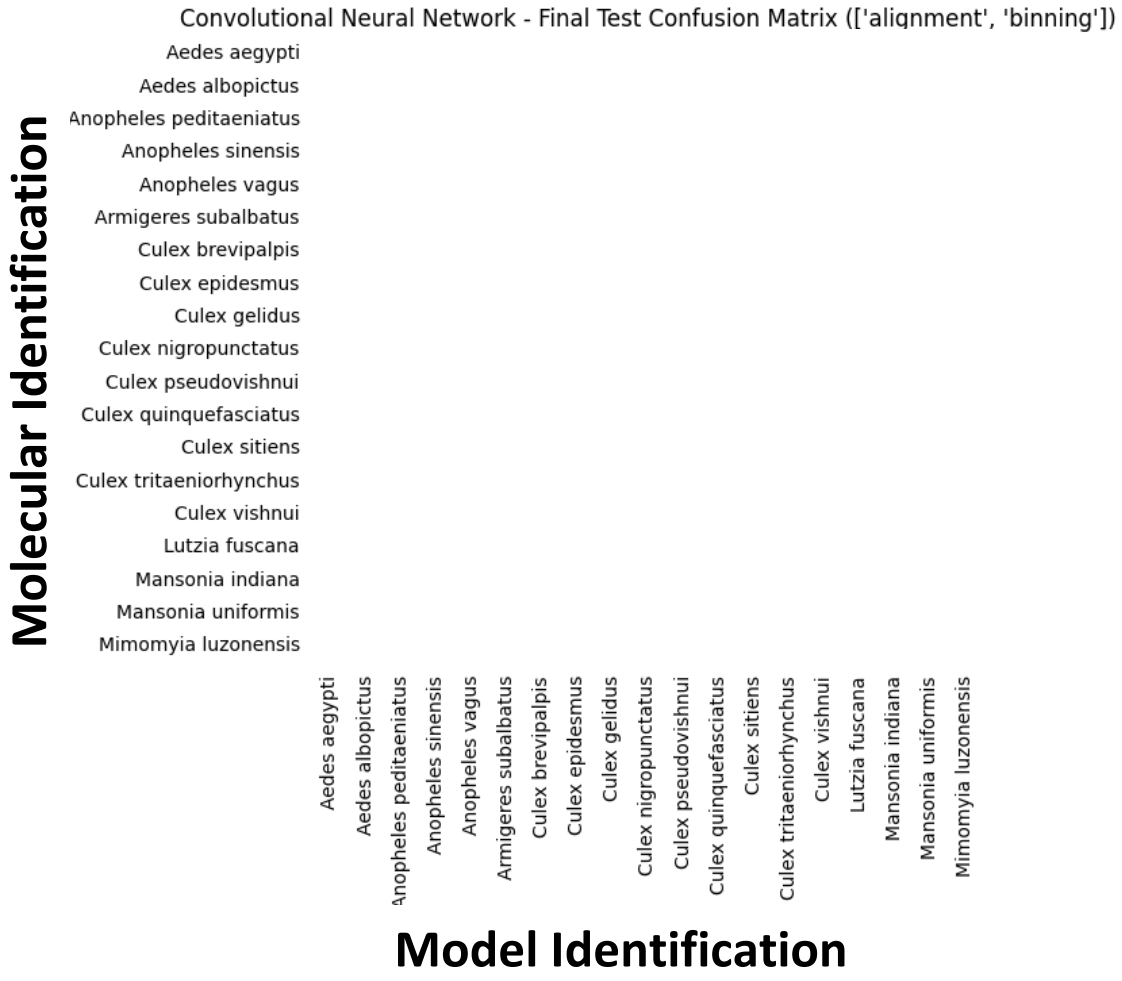
Molecular Identification



PRELIMINARY RESULTS

Institut Pasteur du Cambodge

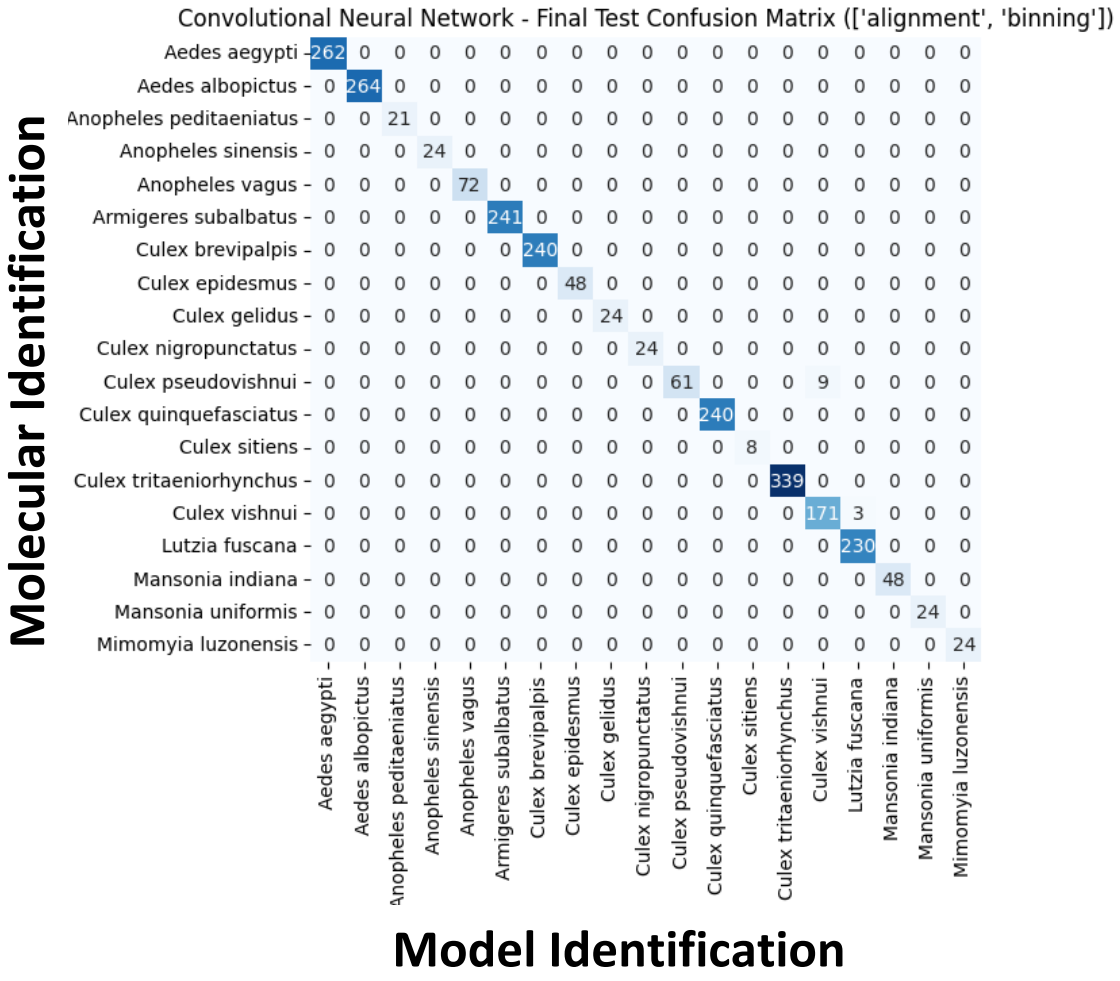
PASTEUR NETWORK



PRELIMINARY RESULTS

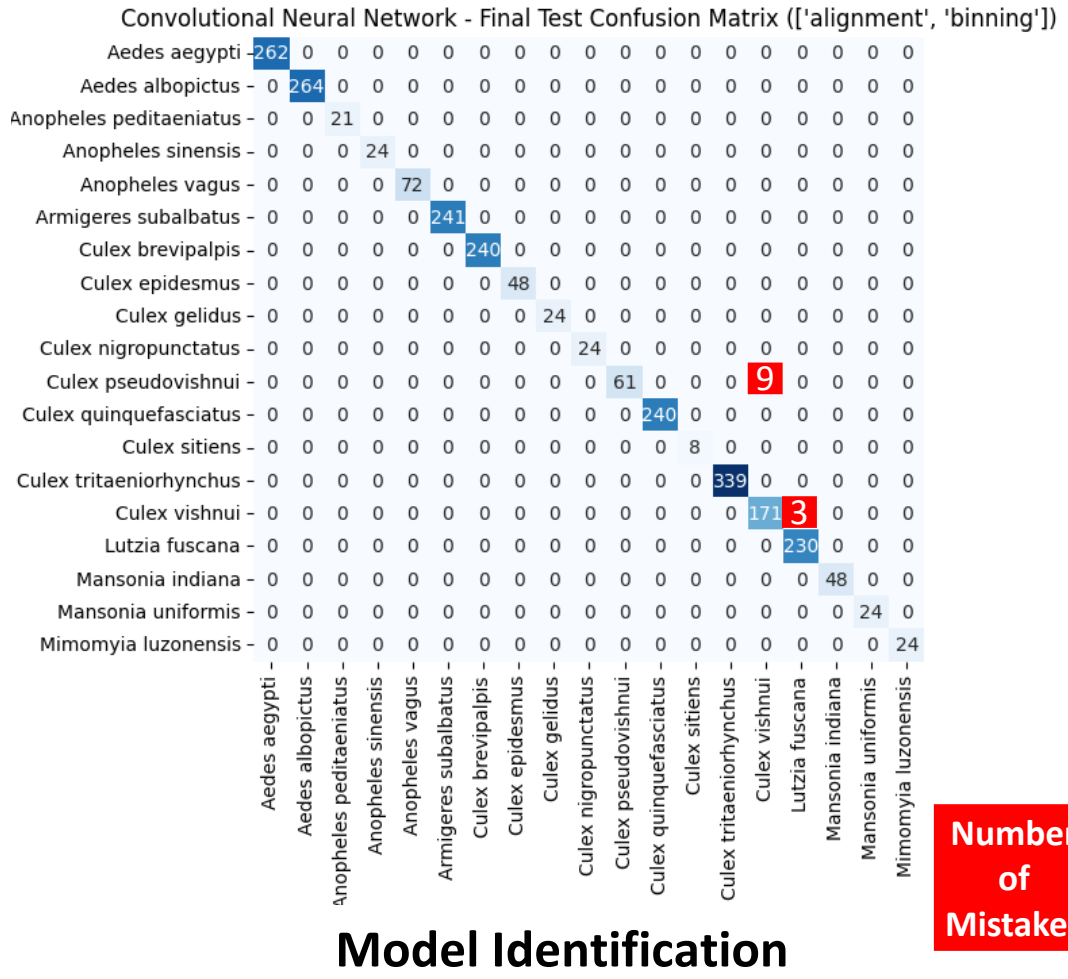
Institut Pasteur du Cambodge

PASTEUR NETWORK



PRELIMINARY RESULTS

Institut Pasteur du Cambodge
PASTEUR NETWORK

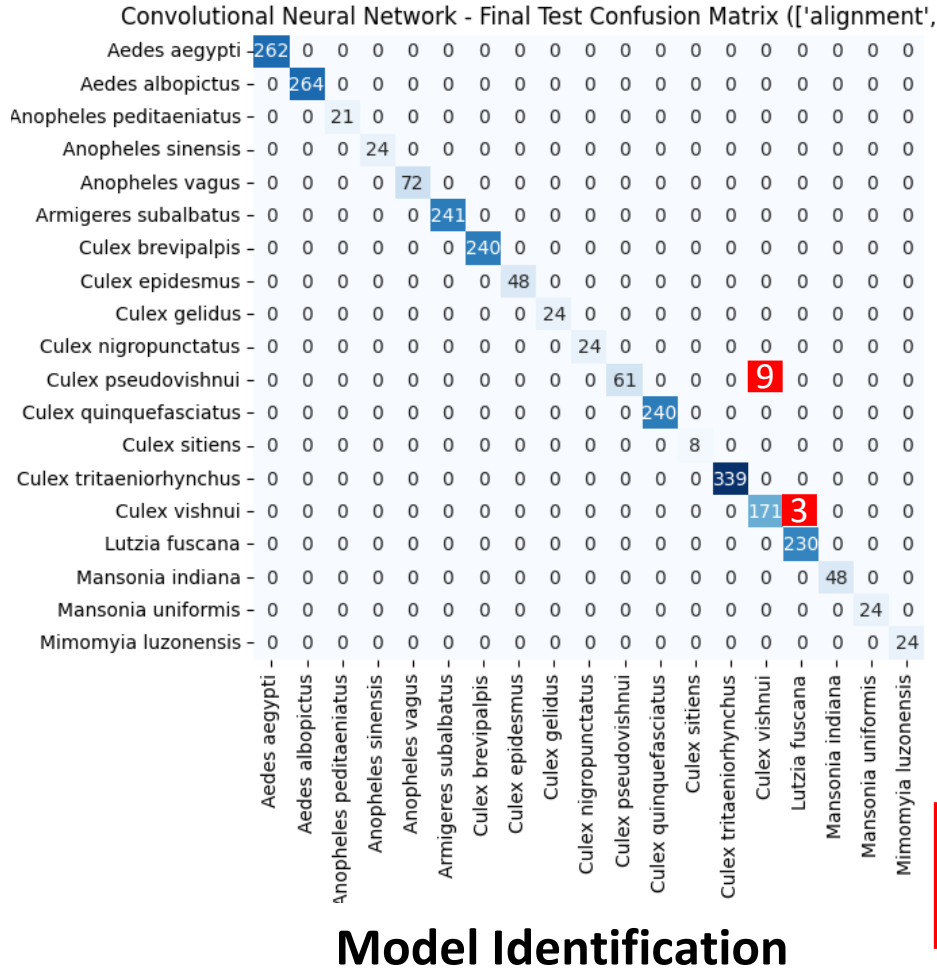


Number of Mistakes

```
=== Final Misclassified Spectra (Grouped) ===
individual      spectrum_indices                #wrong/total    true_species          pred_species
KS7             [3932, 3933, 3934, 3935, 3936, 3937, 3938] 7/7              Culex pseudovishnui  Culex vishnui
PVD3            [3991, 3996]                    2/7             Culex pseudovishnui  Culex vishnui
vis158          [6069, 6070, 6087]              3/24            Culex vishnui        Lutzia fuscana
```

PRELIMINARY RESULTS

Institut Pasteur du Cambodge
PASTEUR NETWORK



Main observations:

- **F1-score: 99.5%**

- **Culex vishnui complex**: 583/595 correctly classified (97.98%)

⚠️ Confusion One individual Culex vishnui and Lutzia fuscana
  - 3 misclassified spectra / 24
  - **87.5% (21/24)** still correctly classified

Number of Mistakes

```
=== Final Misclassified Spectra (Grouped) ===
individual      spectrum_indices                  #wrong/total    true_species          pred_species
KS7             [3932, 3933, 3934, 3935, 3936, 3937, 3938] 7/7                Culex pseudovishnui   Culex vishnui
PVD3            [3991, 3996]                      2/7             Culex pseudovishnui   Culex vishnui
vis158          [6069, 6070, 6087]                3/24            Culex vishnui         Lutzia fuscana
```

CONCLUSION

Institut Pasteur du Cambodge

PASTEUR NETWORK

Key Results

High performance (F1 = 99.5%)

Strong classification of the Culex vishnui complex (98%)

CONCLUSION

Institut Pasteur du Cambodge
PASTEUR NETWORK

Key Results

High performance (F1 = 99.5%)

Strong classification of the Culex vishnui complex (98%)

Limitations

Label

- GenBank-based labels (community data; US-funded)

CONCLUSION

Institut Pasteur du Cambodge
PASTEUR NETWORK

Key Results

High performance (F1 = 99.5%)

Strong classification of the Culex vishnui complex (98%)

Limitations

Label

- GenBank-based labels (community data; US-funded)

Model

- Model limited to 19 species classes (>300 species present in Cambodia)
- No threshold for result acceptance

CONCLUSION

Institut Pasteur du Cambodge

PASTEUR NETWORK

Key Results

High performance (F1 = 99.5%)

Strong classification of the Culex vishnui complex (98%)

Limitations

Label

- GenBank-based labels (community data; US-funded)

Model

- Model limited to 19 species classes (>300 species present in Cambodia)
- No threshold for result acceptance

Go Further

- Extend identification to other arthropods (e.g., ticks)
- Detect viruses in mosquitoes and humans

Machine learning application in Medical Entomology : determining potential biomarkers in mosquito and human for Dengue and Chikungunya viruses FR

Auteur / Autrice :	Matilin Le beux
Direction :	Sebastien Boyer, Nicolas Courty
Type :	Projet de thèse
Discipline(s) :	Sciences de la vie et de la sante
Date :	Inscription en doctorat le 29/11/2024
Etablissement(s) :	Université Paris Cité
Ecole(s) doctorale(s) :	562 - BIO SORBONNE PARIS CITÉ (BIOSPC)
Partenaire(s) de recherche :	**Equipe de recherche :** UMR 2000 - GEMS - Génomique évolutive, modélisation et santé

ACKNOWLEDGEMENTS

Medical and Veterinary Entomology
Sébastien Boyer
Kimly Heng
Kimsear Nov

Unité d'Entomologie Médicale
Antsa Rakotonirina

IRISA - OBELIX
Nicolas Courty

Thank you for your attention!
Do you have any questions?

REFERENCES

Institut Pasteur du Cambodge
PASTEUR NETWORK

Beauté, Julien, and Sirenda Vong. "Cost and disease burden of dengue in Cambodia." *BMC public health* 10.1 (2010): 521.

Cappelle, Julien, et al. "Intensive circulation of Japanese encephalitis virus in peri-urban sentinel pigs near Phnom Penh, Cambodia." *PLoS neglected tropical diseases* 10.12 (2016): e0005149.

Rakotonirina, A., Dusadeepong, R., Hide, M., Vuth, L., Chea, R., Heng, K., ... & Boyer, S. (2025). Emerging approaches to mosquito species identification: an overview with emphasis on nanopore sequencing technology. *Journal of Medical Entomology*, tjaf044.

Weis, Caroline V., Catherine R. Jutzeler, and Karsten Borgwardt. "Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review." *Clinical Microbiology and Infection* 26.10 (2020): 1310-1317.

Beebe, Nigel W. "DNA barcoding mosquitoes: advice for potential prospectors." *Parasitology* 145.5 (2018): 622-633.

Yssouf, Amina, et al. "Matrix-assisted laser desorption ionization-time of flight mass spectrometry: an emerging tool for the rapid identification of mosquito vectors." *PLoS One* 8.8 (2013): e72380.

Rakotonirina, Antsa, et al. "MALDI-TOF MS: optimization for future uses in entomological surveillance and identification of mosquitoes from New Caledonia." *Parasites & vectors* 13.1 (2020): 359.

Merchan, Fernando, et al. "Deep Metric Learning for the Classification of MALDI-TOF Spectral Signatures from Multiple Species of Neotropical Disease Vectors." *Artificial Intelligence in the Life Sciences*, vol. 3, 2023, p. 100071. Elsevier,

Lay Jr, J. O. (2001). MALDI-TOF mass spectrometry of bacteria. *Mass spectrometry reviews*, *20*(4), 172-194.

Seng, P., Rolain, J. M., Fournier, P. E., La Scola, B., Drancourt, M., & Raoult, D. (2010). MALDI-TOF-mass spectrometry applications in clinical microbiology. *Future microbiology*, *5*(11), 1733-1754.