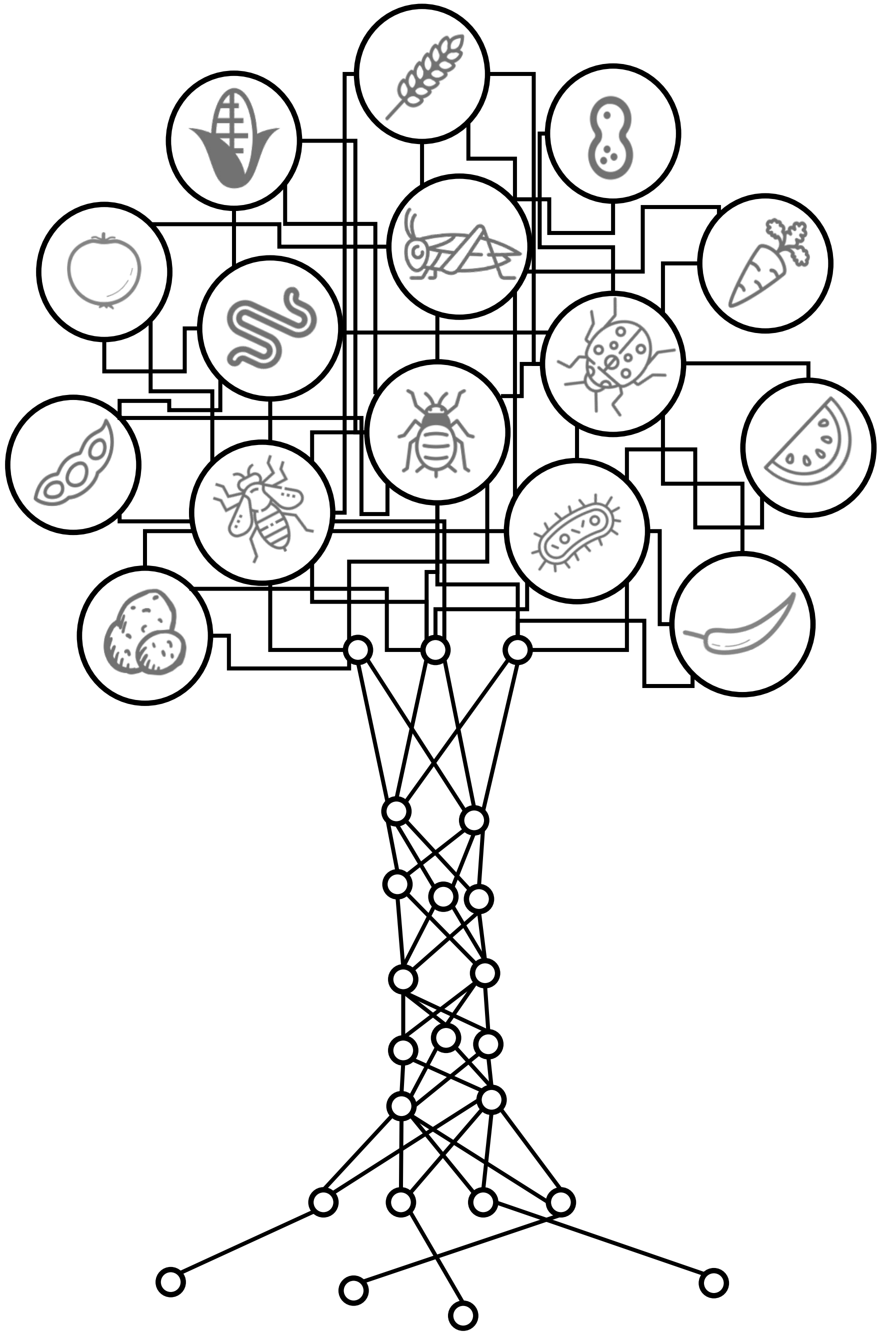


UNIVERSITÉ CÔTE D'AZUR
MAISON DE LA MODÉLISATION, DE LA SIMULATION ET DES INTERACTIONS

INRAE
la science pour la vie, l'humain, la terre

NERD (Nematode EffectoR Discovery)

a tool to predict proteins involved in nematodes' plant parasitism.

Djampa KOZLOWSKI, PhD (MSI/UCA/INRAe)

Silvia BOTTINI, PhD.
Medical Data Laboratory

Etienne DANCHIN, PhD.

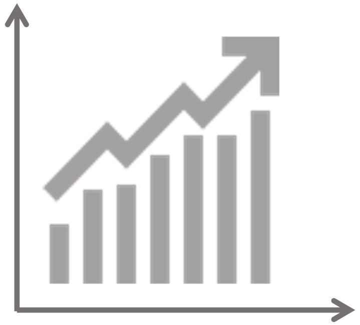
UNIVERSITÉ CÔTE D'AZUR
Centre de calculs

Plant Parasitic Nematodes (PPNs), a threat to global food safety

Population

9.7 bil.
(2064)

8 bil.
(2022) (1)



Food needs + 25-70% in 2050 (2)

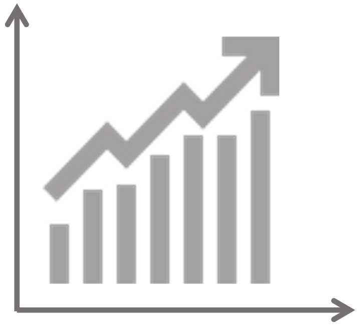
(1) Vollset et al; The Lancet, 2020
(2) Hunter et al; Bioscience, 2017

Plant Parasitic Nematodes (PPNs), a threat to global food safety

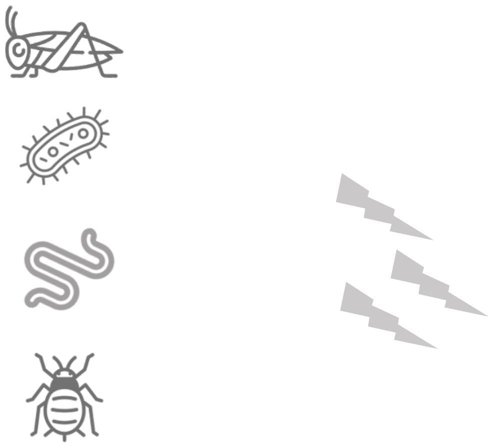
Population

9.7 bil.
(2064)

8 bil.
(2022) (1)



Food needs + 25-70% in 2050 (2)



Crop pest are one of the major threat to food sustainability (20-30 % Mondial Loss) (3)

(1) Vollset et al; The Lancet, 2020
(2) Hunter et al; Bioscience, 2017
(3) Savary et al; Nat. Ecol. & Evol, 2017

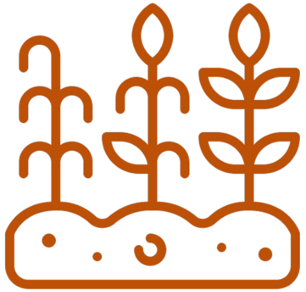
Plant Parasitic Nematodes (PPNs), a threat to global food safety

Population

9.7 bil.
(2064)

8 bil.
(2022) (1)

Food needs + 25-70% in 2050 (2)



Crop pest are one of the major threat to food sustainability (20-30 % Mondial Loss) (3)

Plant Parasitic Nematodes (PPNs) (12% Mondial loss) (3)

(1) Vollset et al; The Lancet, 2020
(2) Hunter et al; Bioscience, 2017
(3) Savary et al; Nat. Ecol. & Evol, 2017

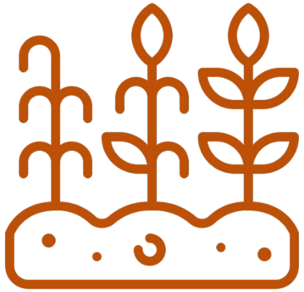
Plant Parasitic Nematodes (PPNs), a threat to global food safety

Population

9.7 bil.
(2064)

8 bil.
(2022) (1)

Food needs + 25-70% in 2050 (2)



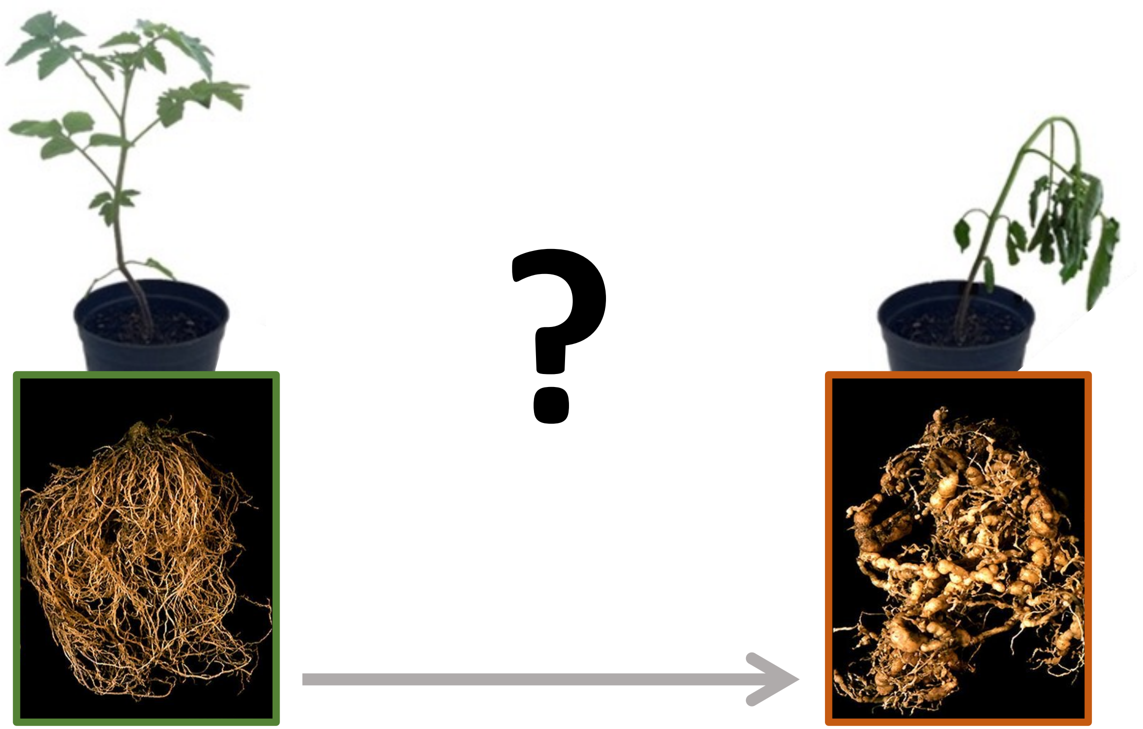
Crop pest are one of the major threat to food sustainability (20-30 % Mondial Loss) (3)

Plant Parasitic Nematodes (PPNs) (12% Mondial loss) (3)

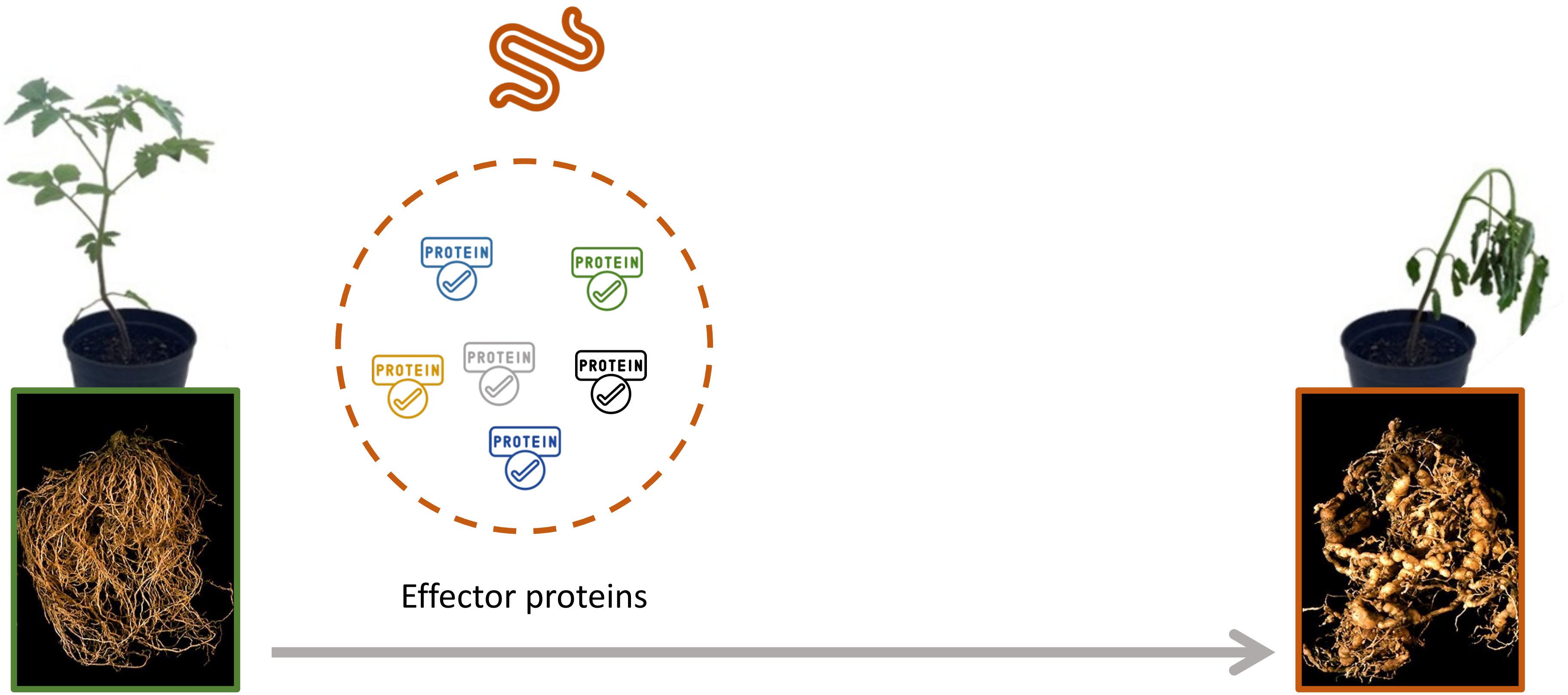
Better understand the biology of those species and how they interact with their environment / host is necessary to develop new control methods.

(1) Vollset et al; The Lancet, 2020
(2) Hunter et al; Bioscience, 2017
(3) Savary et al; Nat. Ecol. & Evol, 2017

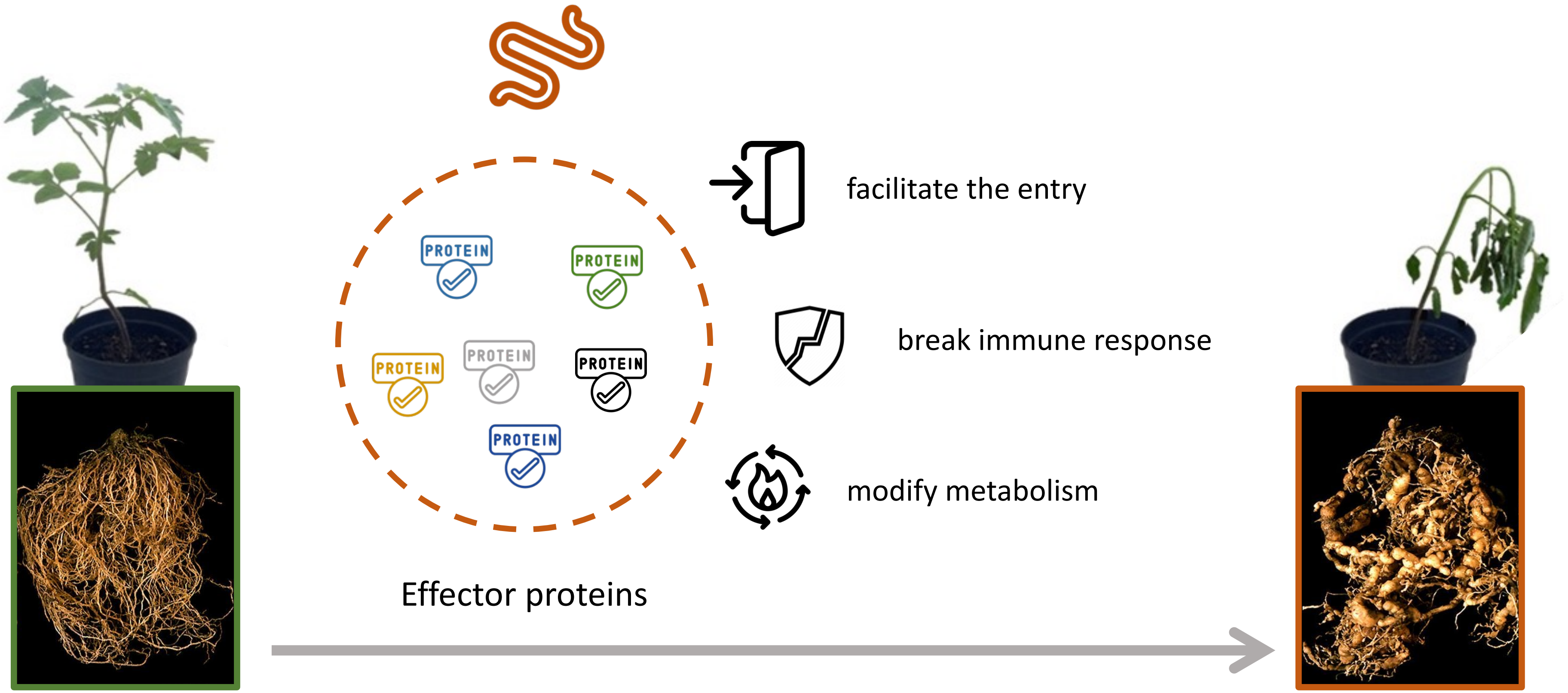
The Plant Parasitic Nematode (PPNs) infectious toolbox : the effector proteins



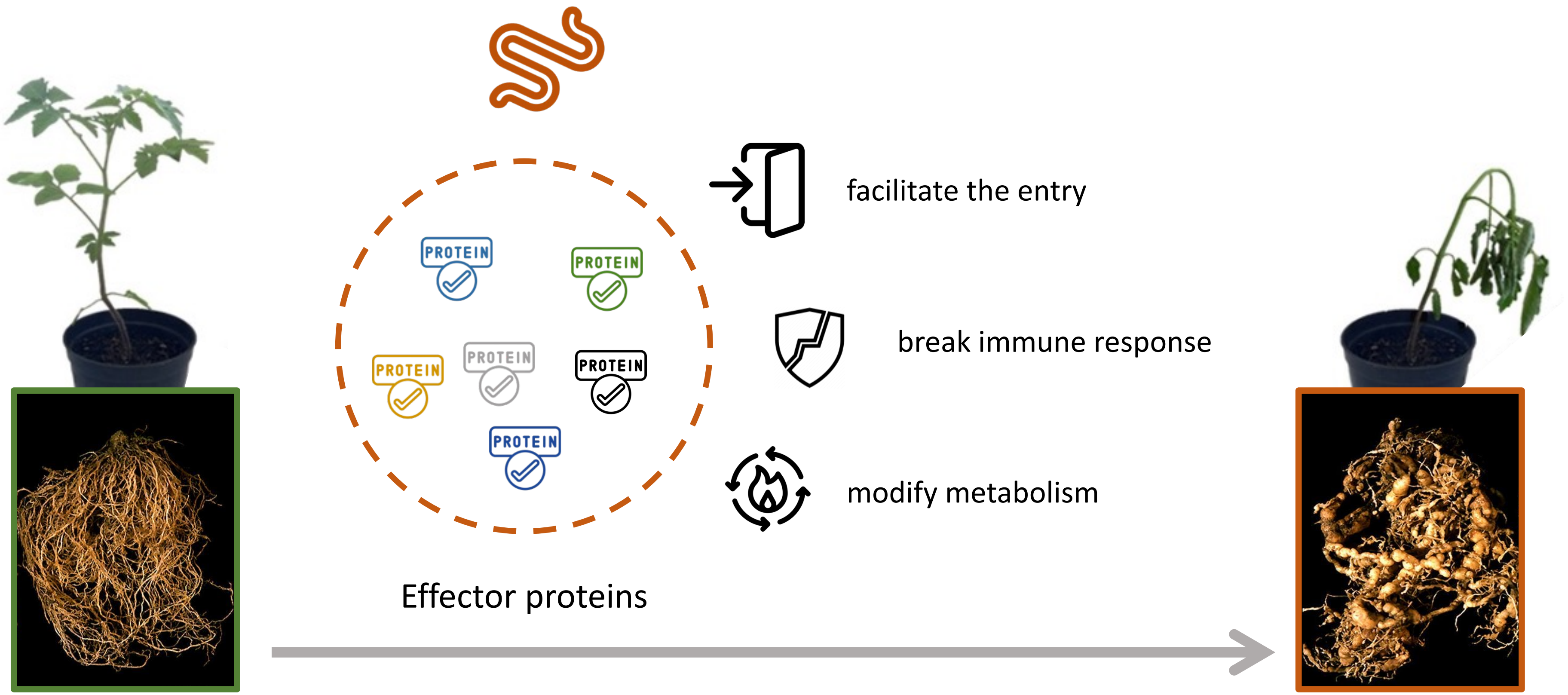
The Plant Parasitic Nematode (PPNs) infectious toolbox : the effector proteins



The Plant Parasitic Nematode (PPNs) infectious toolbox : the effector proteins

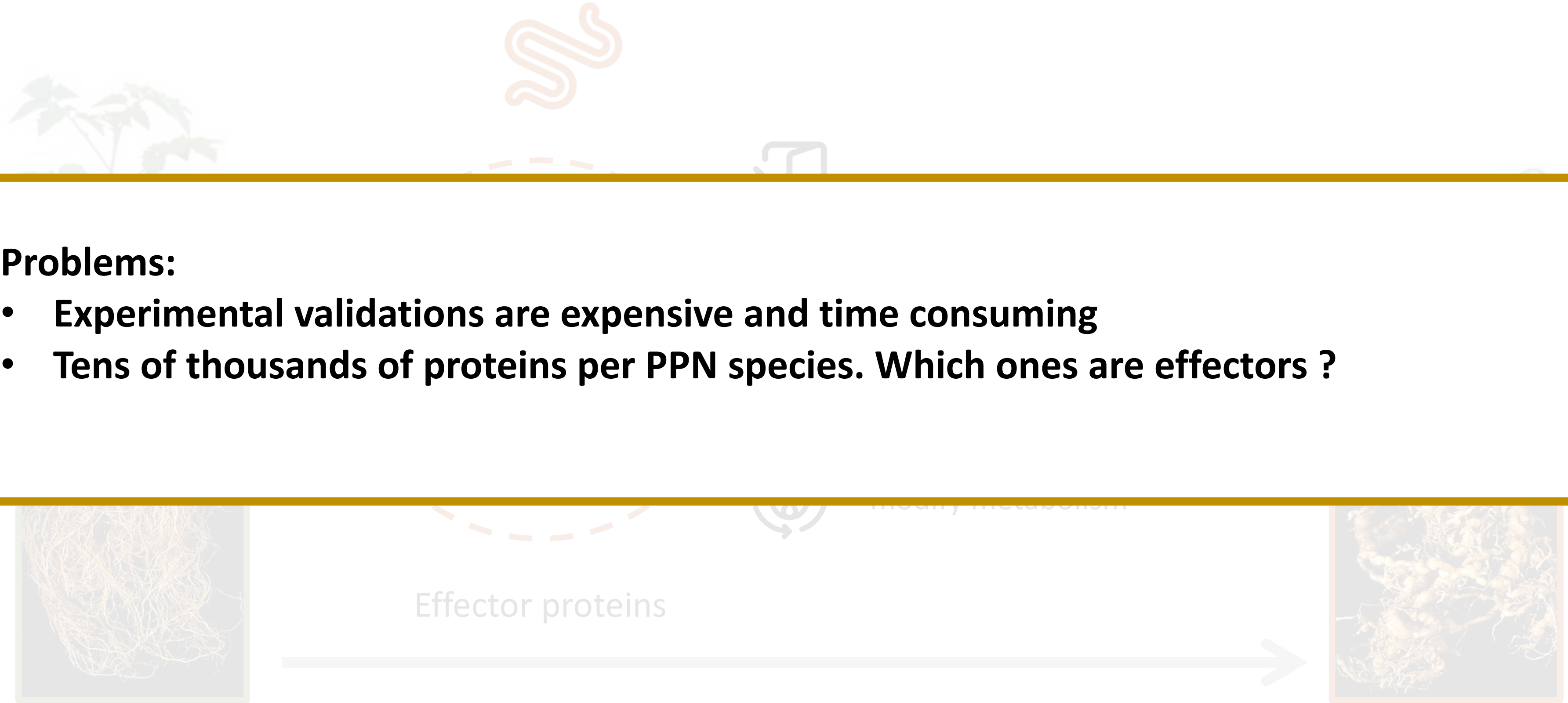


The Plant Parasitic Nematode (PPNs) infectious toolbox : the effector proteins



Effectors proteins are a major key of the host-pathogen interaction and must be studied

The Plant Parasitic Nematode (PPNs) infectious toolbox : the effector proteins



Problems:

- **Experimental validations are expensive and time consuming**
- **Tens of thousands of proteins per PPN species. Which ones are effectors ?**

Effectors proteins are a major key of the host-pathogen interaction and must be studied

The Plant Parasitic Nematode (PPNs) infectious toolbox : the effector proteins

Problems:

- **Experimental validations are expensive and time consuming**
- **Tens of thousands of proteins per PPN species. Which ones are effectors ?**

In-silico approach is necessary to identify effector proteins candidates for experimental validation

Effectors proteins are a major key of the host-pathogen interaction and must be studied

Methodology : datasets creation

11 **Plant Parasitic Nematode(PPN)** species
(5 genera)

11 **Plant Parasitic Nematode(PPN)** species
(5 genera)

Experimentally validated effector proteins from the literature

- Amplification (orthologous sequences from proteomes)
- Redundancy reduction

546 protein sequences

11 **Plant Parasitic Nematode(PPN)** species
(5 genera)

Experimentally validated effector proteins from the literature

- Amplification (orthologous sequences from proteomes)
- Redundancy reduction

546 protein sequences

Proteins universally conserved among nematodes (64 species, PPN and not-PPN)

- Redundancy reduction

3849 protein sequences

Methodology : datasets creation

11 **Plant Parasitic Nematode(PPN)** species
(5 genera)

Experimentally validated effector proteins from the literature

- Amplification (orthologous sequences from proteomes)
- Redundancy reduction

\+ 546 protein sequences

Proteins universally conserved among nematodes (64 species, PPN and not-PPN)

- Redundancy reduction

\- 3849 protein sequences

Data splitting
(same proportion of positive and negative sequences in each dataset)

\+ 382 seqs.
\- 2694 seqs.

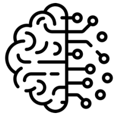
Training set (70%)

\+ 164 seqs.
\- 1155 seqs.

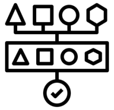
Test set (30%)

Methodology : Predict effector proteins, a challenging task

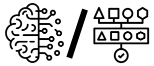
Existing tools



MERCI-1.0 **(1)**
EffectorP-3.0 **(2)**
Deepredeff-0.1.1 **(3)**



EffHunter-1.0 **(4)**

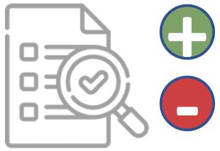


Predector-1.2.7 **(5)**

(1) Vens et al., Bioinformatics, 2011.
(2) Sperschneider & Dodds, MPMI, 2022.
(3) Kristianingsih et al., BMC bioinfo., 2021.
(4) Carreon-Anguiano et al., biomolecules, 2020.
(5) Jones et al., Scientific reports, 2021.

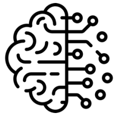
Methodology : Predict effector proteins, a challenging task

Test set

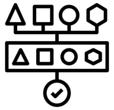


164 seqs.
1155 seqs.

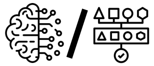
Existing tools



MERCI-1.0 **(1)**
EffectorP-3.0 **(2)**
Deepredeff-0.1.1 **(3)**



EffHunter-1.0 **(4)**



Predector-1.2.7 **(5)**

Performances evaluation

	0 (pred.)	1 (pred.)
0 (obs.)	**TN**	**FP**
1 (obs.)	**FN**	**TP**

(1) Vens et al., Bioinformatics, 2011.
(2) Sperschneider & Dodds, MPMI, 2022.
(3) Kristianingsih et al., BMC bioinfo., 2021.
(4) Carreon-Anguiano et al., biomolecules, 2020.
(5) Jones et al., Scientific reports, 2021.

Methodology : Predict effector proteins, a challenging task

Test set

+ 164 seqs.
- 1155 seqs.

Existing tools

- MERCI-1.0 **(1)**
- EffectorP-3.0 **(2)**
- Deepredeff-0.1.1 **(3)**
- EffHunter-1.0 **(4)**
- Predector-1.2.7 **(5)**

Performances evaluation

	0 (pred.)	1 (pred.)
0 (obs.)	TN	FP
1 (obs.)	FN	TP

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recal} = \frac{TP}{TP + FN}$$

"Ability of the classifier not to label as an effector a protein that is not an effector"

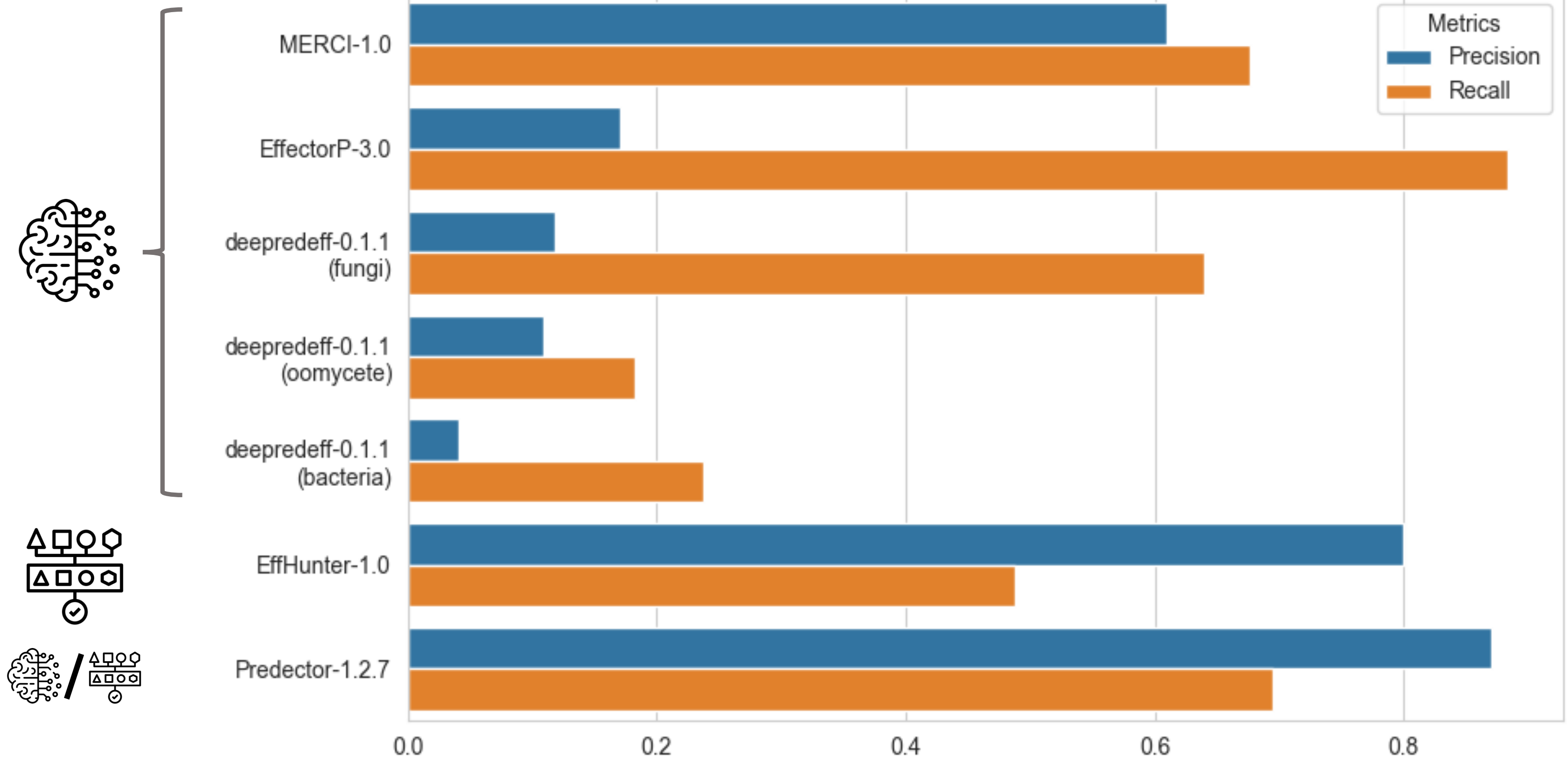
"Ability of the classifier to find all the effector proteins"

(1) Vens et al., Bioinformatics, 2011.
(2) Sperschneider & Dodds, MPMI, 2022.
(3) Kristianingsih et al., BMC bioinfo., 2021.
(4) Carreon-Anguiano et al., biomolecules, 2020.
(5) Jones et al., Scientific reports, 2021.

Methodology : Predict effector proteins, a challenging task

Performances on PPN test set

(164 effectors, 1155 non-effectors; detailed after)



P. = 0.87

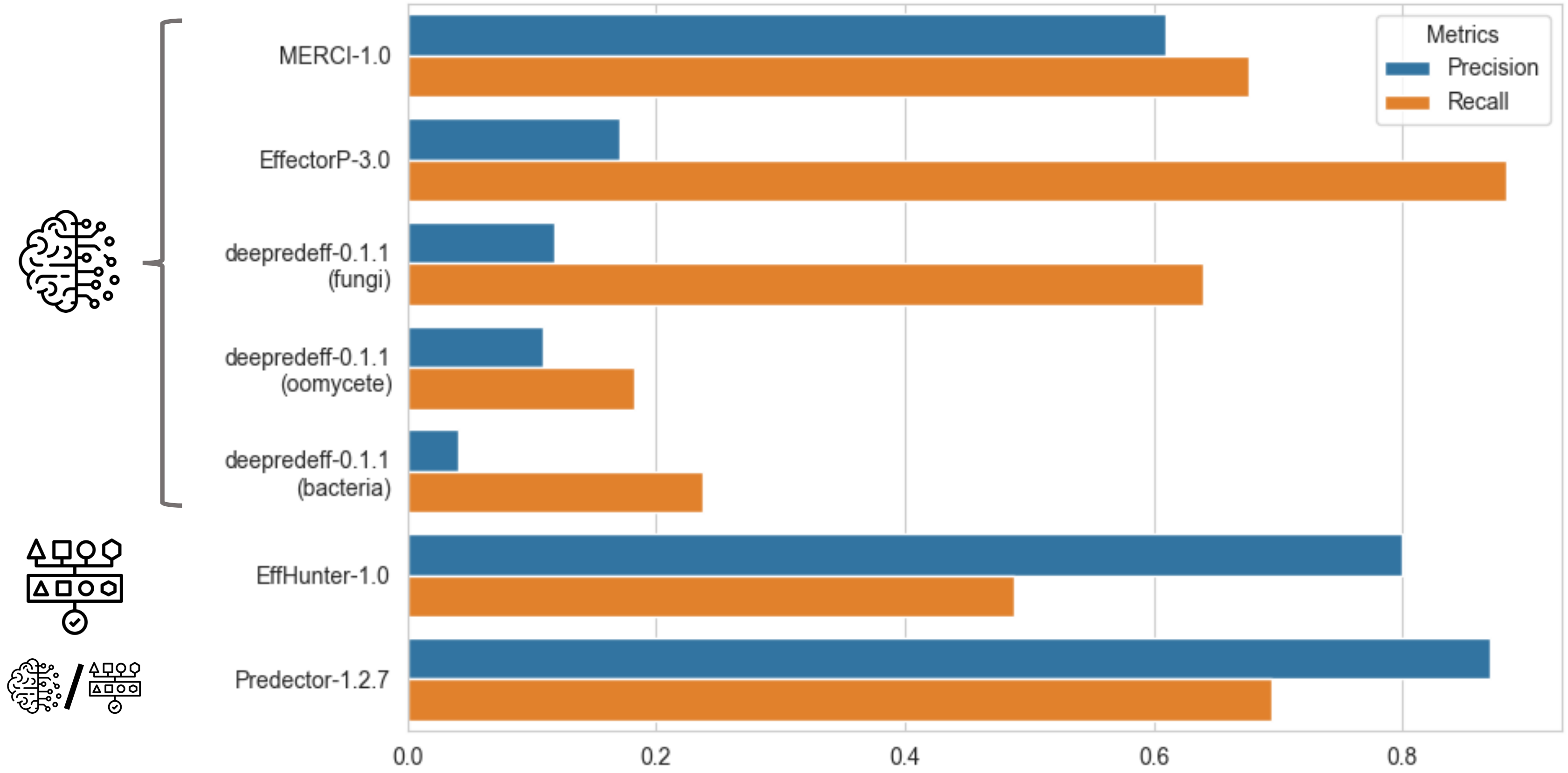
R. = 0.69

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recal} = \frac{TP}{TP + FN}$$

Methodology : Predict effector proteins, a challenging task

Performances on PPN test set
(164 effectors, 1155 non-effectors; detailed after)



P. = 0.87
R. = 0.69

$$\text{Precision} = \frac{TP}{TP + FP}$$

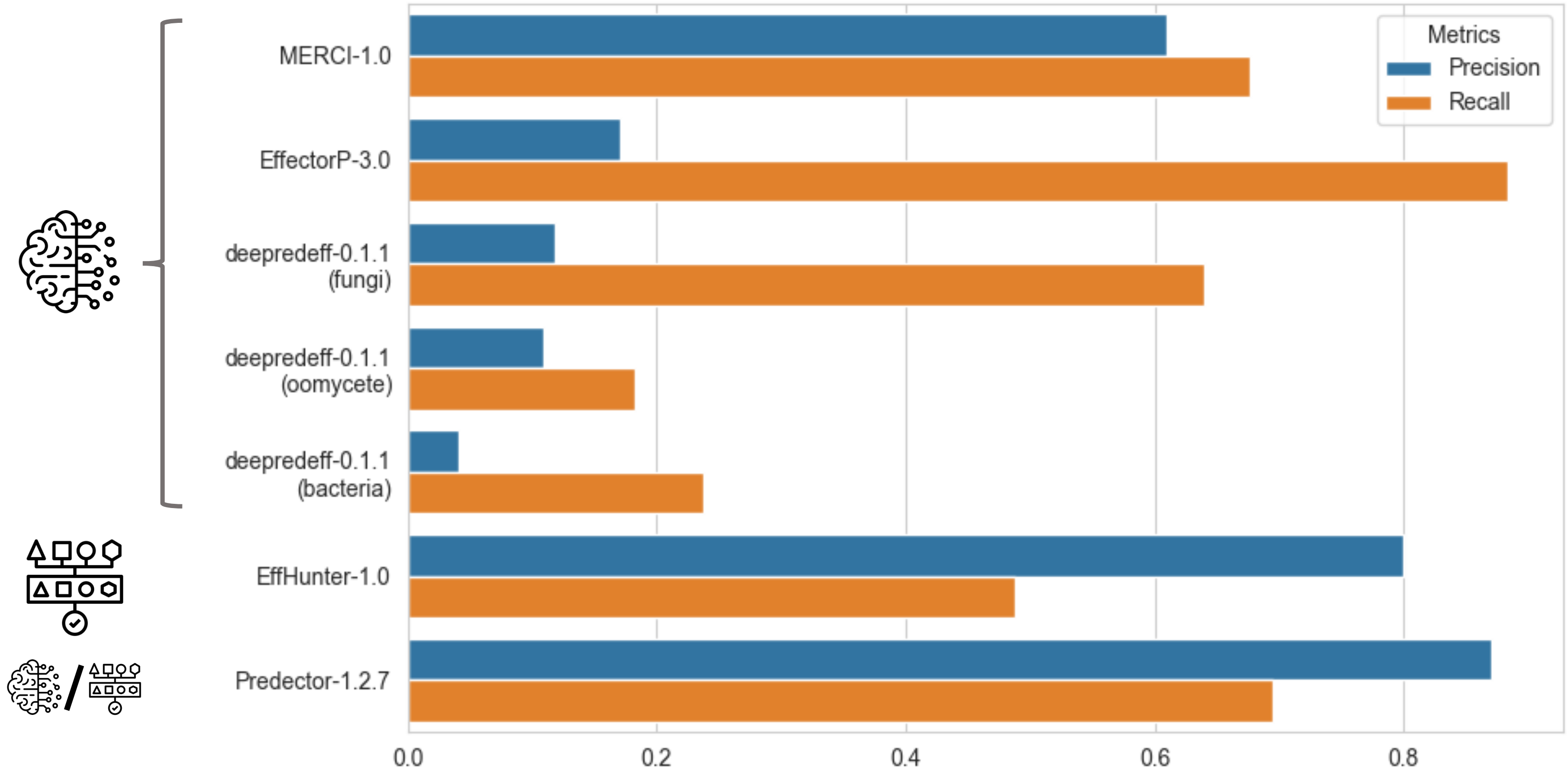
$$\text{Recal} = \frac{TP}{TP + FN}$$

Drawbacks:

- Limited to low performances regarding our question (high number of FP).

Methodology : Predict effector proteins, a challenging task

Performances on PPN test set
(164 effectors, 1155 non-effectors; detailed after)



$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recal} = \frac{TP}{TP + FN}$$

P. = 0.87
R. = 0.69

Drawbacks:

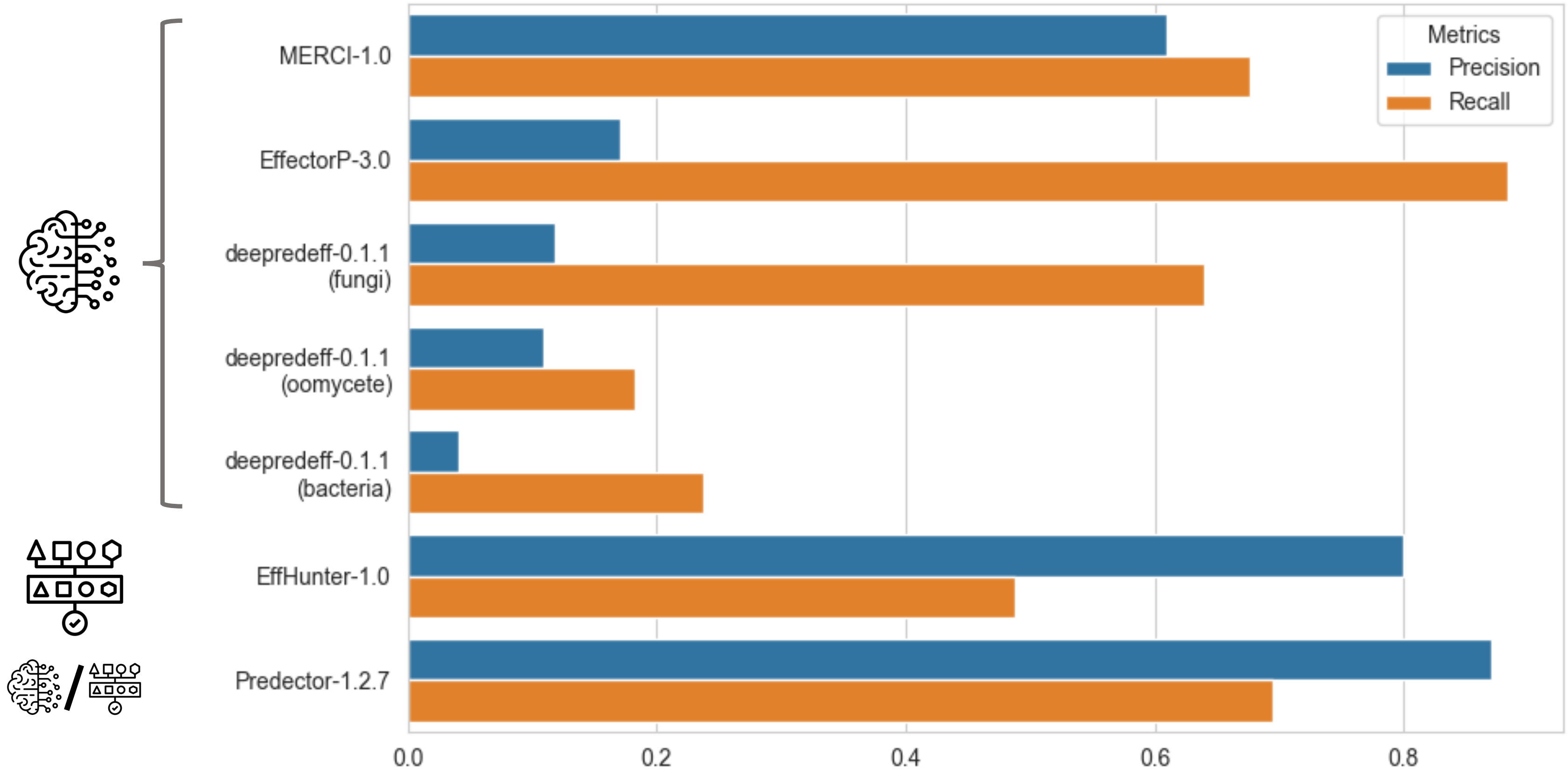
- Limited to low performances regarding our question (high number of FP).

Hypothesis:

- effectors proteins are poorly conserved. Hence, no simple discriminative characteristics.
- Tools were not trained on PPNs ( ).

Methodology : Predict effector proteins, a challenging task

Performances on PPN test set
(164 effectors, 1155 non-effectors; detailed after)



P. = 0.87
R. = 0.69

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recal} = \frac{TP}{TP + FN}$$

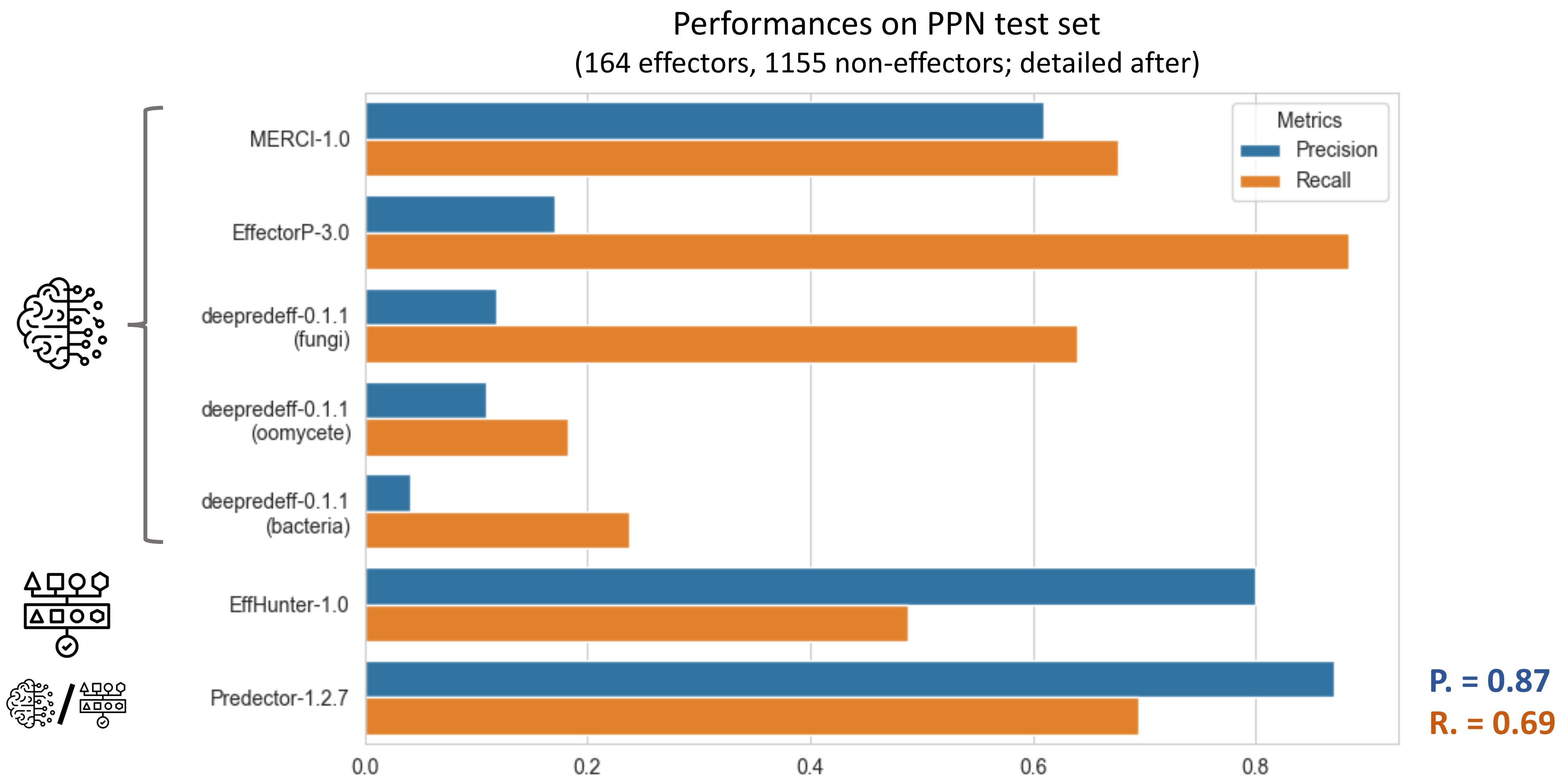
Drawbacks:

- Limited to low performances regarding our question (high number of FP).
- No availability of the code (except MERCI) to re-train the models ( ).

Hypothesis:

- effectors proteins are poorly conserved. Hence, no simple discriminative characteristics.
- Tools were not trained on PPNs ( ).

Methodology : Predict effector proteins, a challenging task



$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recal} = \frac{TP}{TP + FN}$$

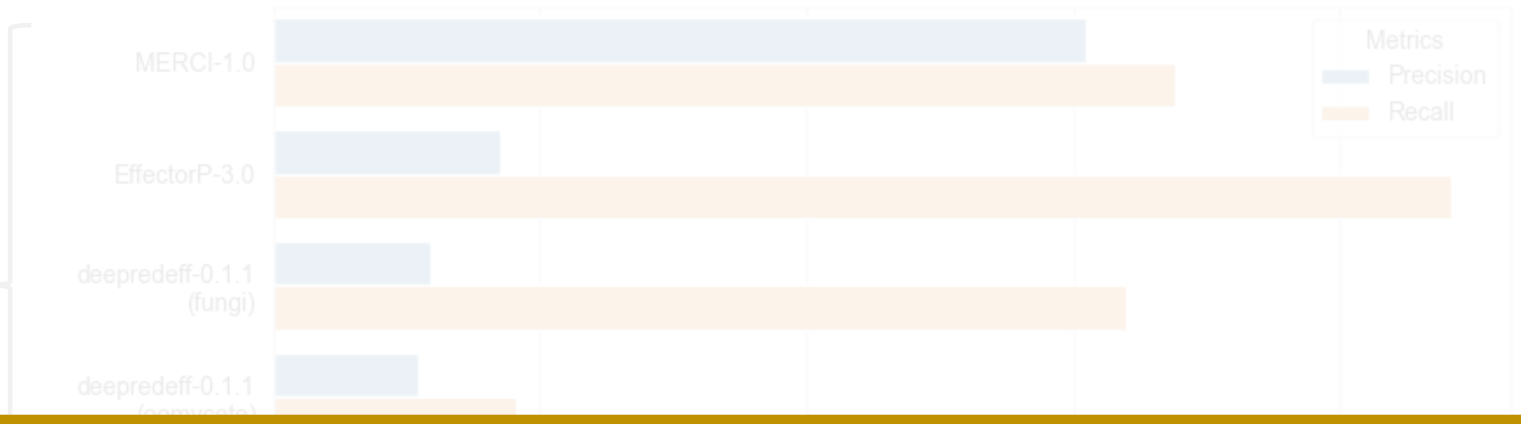
Drawbacks:

- Limited to low performances regarding our question (high number of FP).
- No availability of the code (except MERCI) to re-train the models ( ).
- Only Predector & EffectorP provide a score to rank the predictions and make easier the candidate selection for potential experiments.

Hypothesis:

- effectors proteins are poorly conserved. Hence, no simple discriminative characteristics.
- Tools were not trained on PPNs ( ).

Performances on PPN test set
(164 effectors, 1155 non-effectors; detailed after)



$$\text{Precision} = \frac{TP}{TP + FP}$$

Existing tools are not suited for PPNs.

It is necessary to develop a new tool to efficiently predict and rank protein effector candidates for experimental validation

Drawbacks:

- Limited t... ...rved. Hence,
- No availa... ...stics.
- Only Predector & EffectorP provide a score to rank the predictions and make easier the candidate selection for potential experiments.
- Tools were not trained on PPNs ( ).

Methodology : NERD, a NLP approach to effector prediction (FRAMEWORK)

NERD

NLP based encoder (pre-trained)
Elnaggar *et al.*, IEEE, 2021

Binary classifier (NN)

Methodology : NERD, a NLP approach to effector prediction (FRAMEWORK)

NERD

>seq
MKSKLVAILMALMAFY

Pre-trained ProtT5-XL-UniRef50 (encoder only) [1]

Language model (auto-encoder T5)
trained on Uniref50
(45 million proteins, whole tree of life)

words

M K S K L V A I L M A L M A F Y

sentence

Embedded sequences

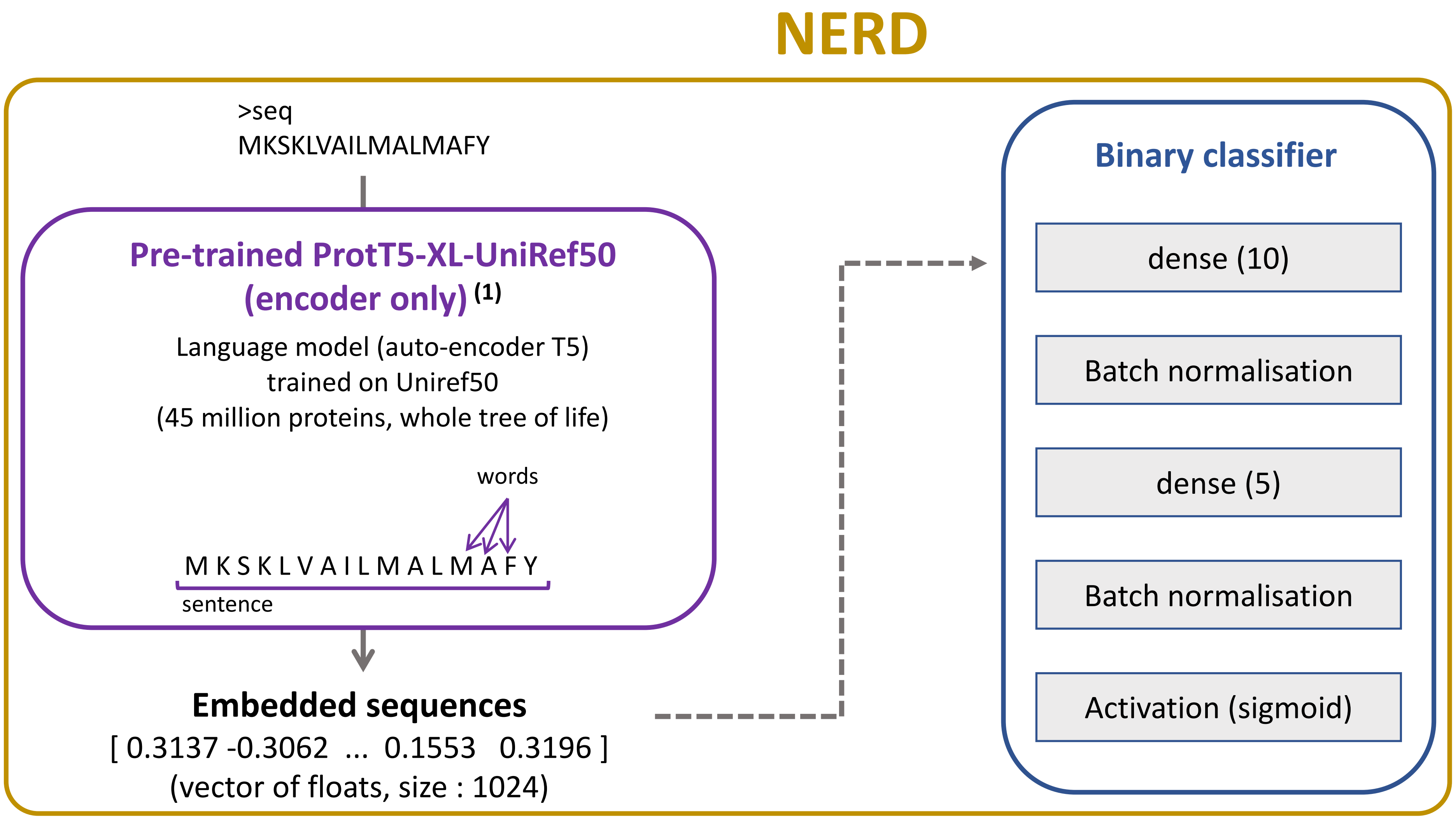
[0.3137 -0.3062 ... 0.1553 0.3196]

(vector of floats, size : 1024)

Binary classifier (NN)

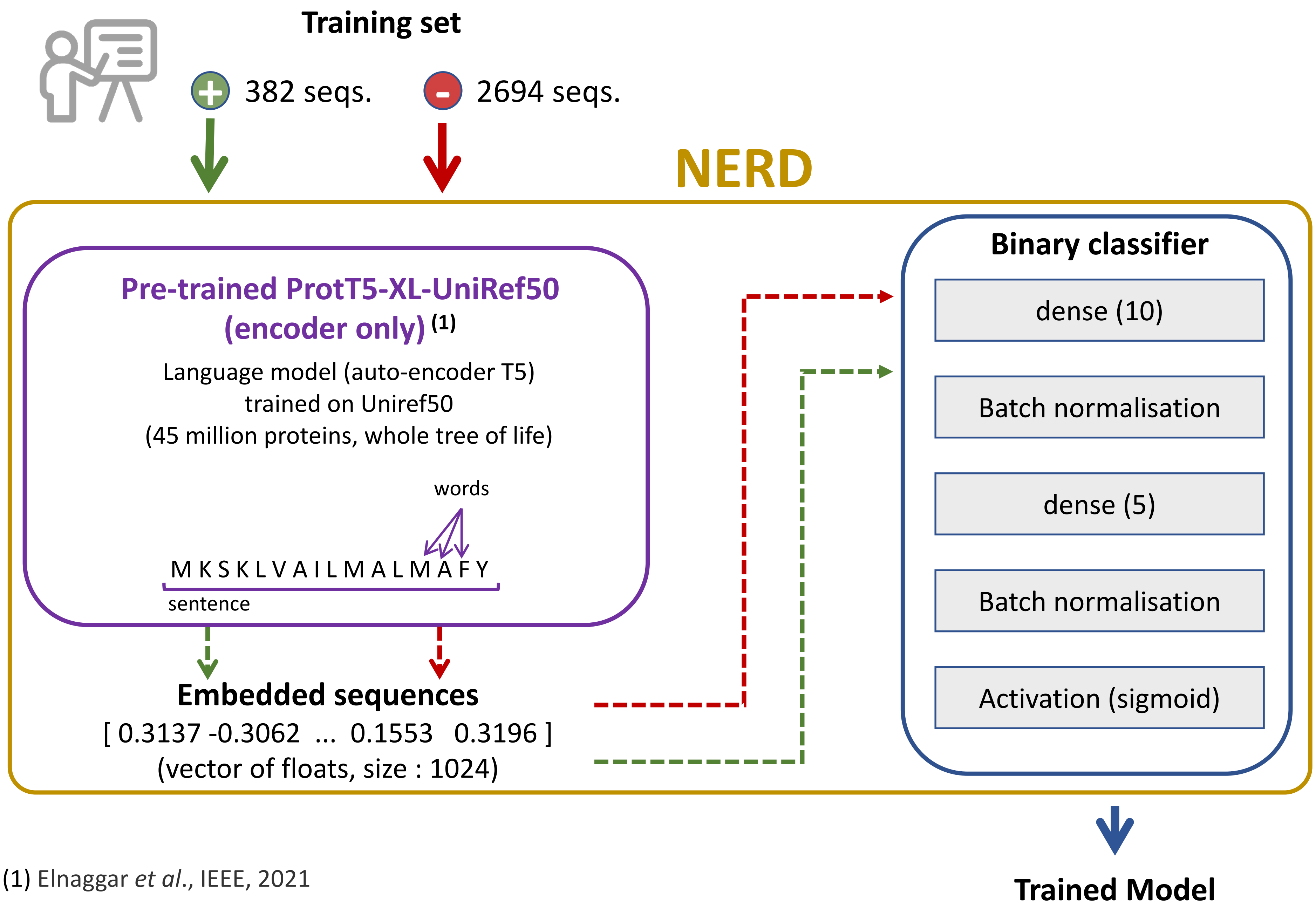
(1) Elnaggar *et al*., IEEE, 2021

Methodology : NERD, a NLP approach to effector prediction (FRAMEWORK)



(1) Elnaggar *et al.*, IEEE, 2021

Methodology : NERD, a NLP approach to effector prediction (TRAINING)



Training parameters:
Loss : Binary cross-entropy
Optimizer : Adam
Max epochs : 500
Batch size : 32
Early stopping criterion : best weights
Patience (5)

Ressources (AZZURA) :
1 GPU (Testla V4)
8 CPU
32 Go ram

(1) Elnaggar *et al.*, IEEE, 2021

Methodology : NERD, a NLP approach to effector prediction (PREDICTION)

Protein sequences

> seq_1
MKSKLVAILMALMAFYA
> seq_2
MTETMLDCSDKVTESKE

NERD

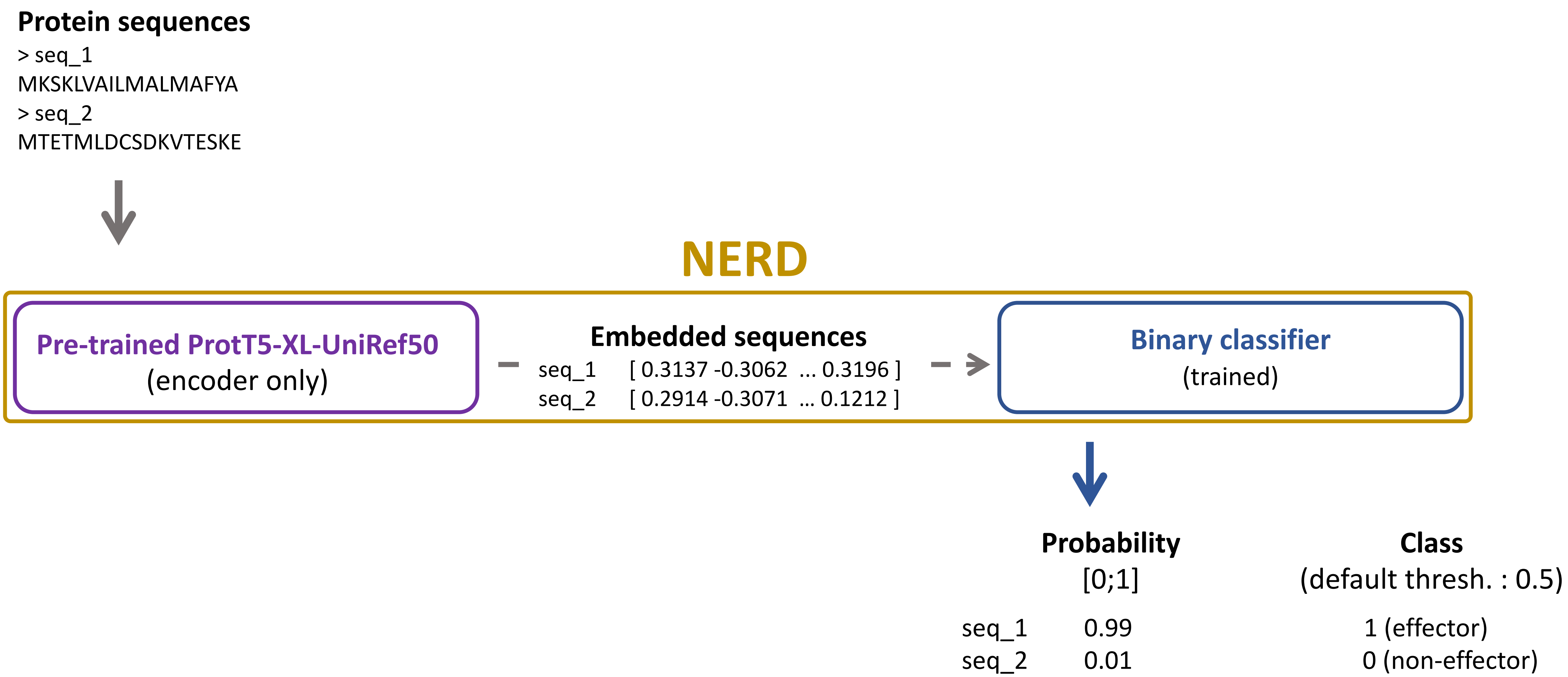
Pre-trained ProtT5-XL-UniRef50
(encoder only)

Embedded sequences

seq_1 [0.3137 -0.3062 ... 0.3196]
seq_2 [0.2914 -0.3071 ... 0.1212]

Binary classifier
(trained)

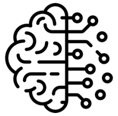
Methodology : NERD, a NLP approach to effector prediction (PREDICTION)



NERD performances evaluation and comparison with existing methods (test set)

Tools

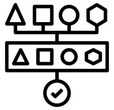
NERD (trained)



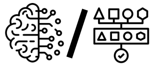
[(1)] MERCI-1.0

[(2)] EffectorP-3.0

[(3)] Deepredeff-0.1.1



[(4)] EffHunter-1.0



[(5)] Predector-1.2.7

(1) Vens *et al*., Bioinformatics, 2011.

(2) Sperschneider & Dodds, MPMI, 2022.

(3) Kristianingsih *et al*., BMC bioinfo., 2021.

(4) Carreon-Anguiano *et al*., biomolecules, 2020.

(5) Jones *et al*., Scientific reports, 2021.

NERD performances evaluation and comparison with existing methods (test set)

Test set

164 seqs.

1155 seqs.

Tools

NERD (trained)

(1) MERCI-1.0

(2) EffectorP-3.0

(3) Deepredeff-0.1.1

(4) EffHunter-1.0

(5) Predector-1.2.7

Performances evaluation

$$\text{Precision} = \frac{TP}{TP + FP}$$

"Ability of the classifier not to label as an effector a protein that is not an effector"

$$\text{Recal} = \frac{TP}{TP + FN}$$

"Ability of the classifier to find all the effector proteins"

(1) Vens *et al.*, Bioinformatics, 2011.

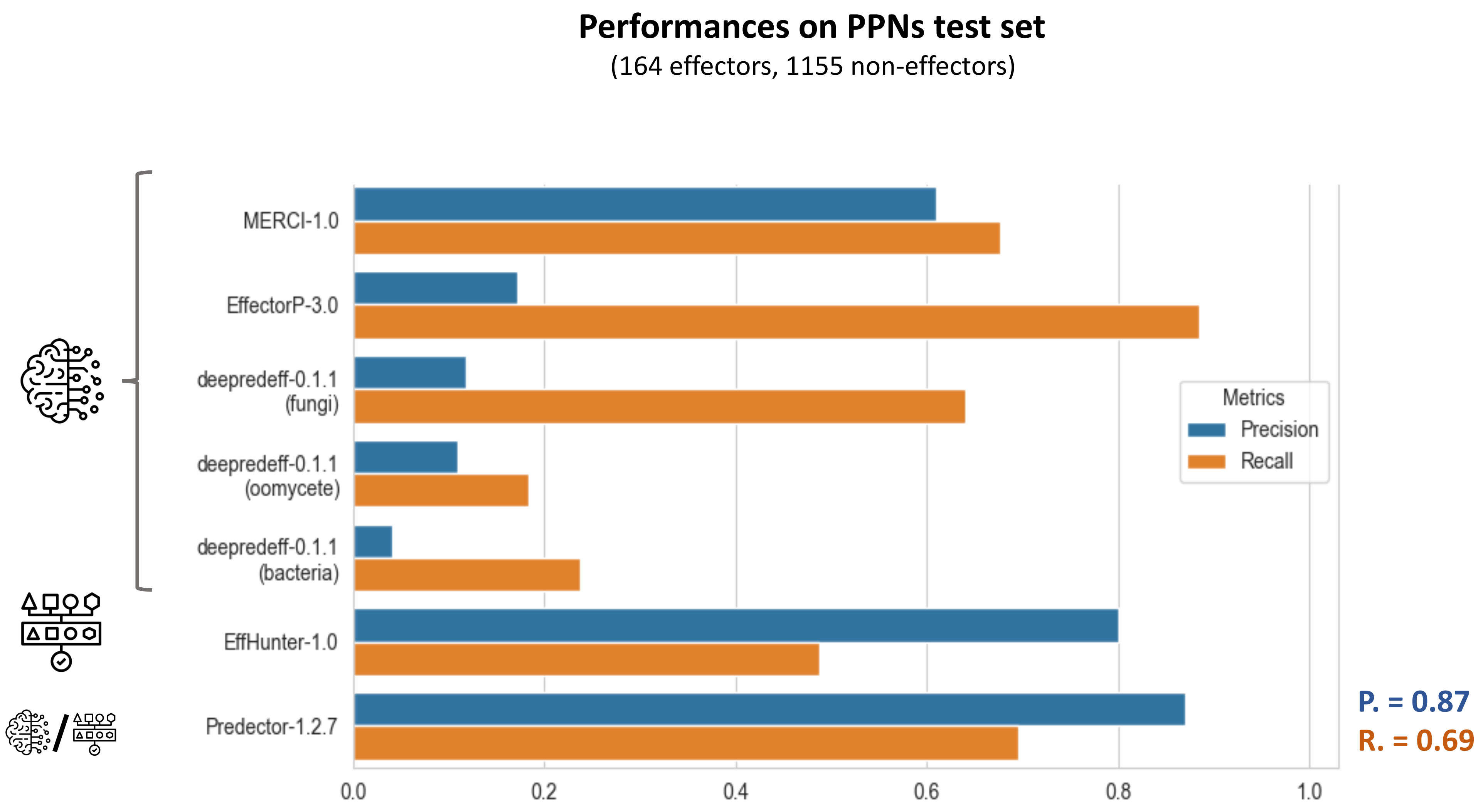
(2) Sperschneider & Dodds, MPMI, 2022.

(3) Kristianingsih *et al.*, BMC bioinfo., 2021.

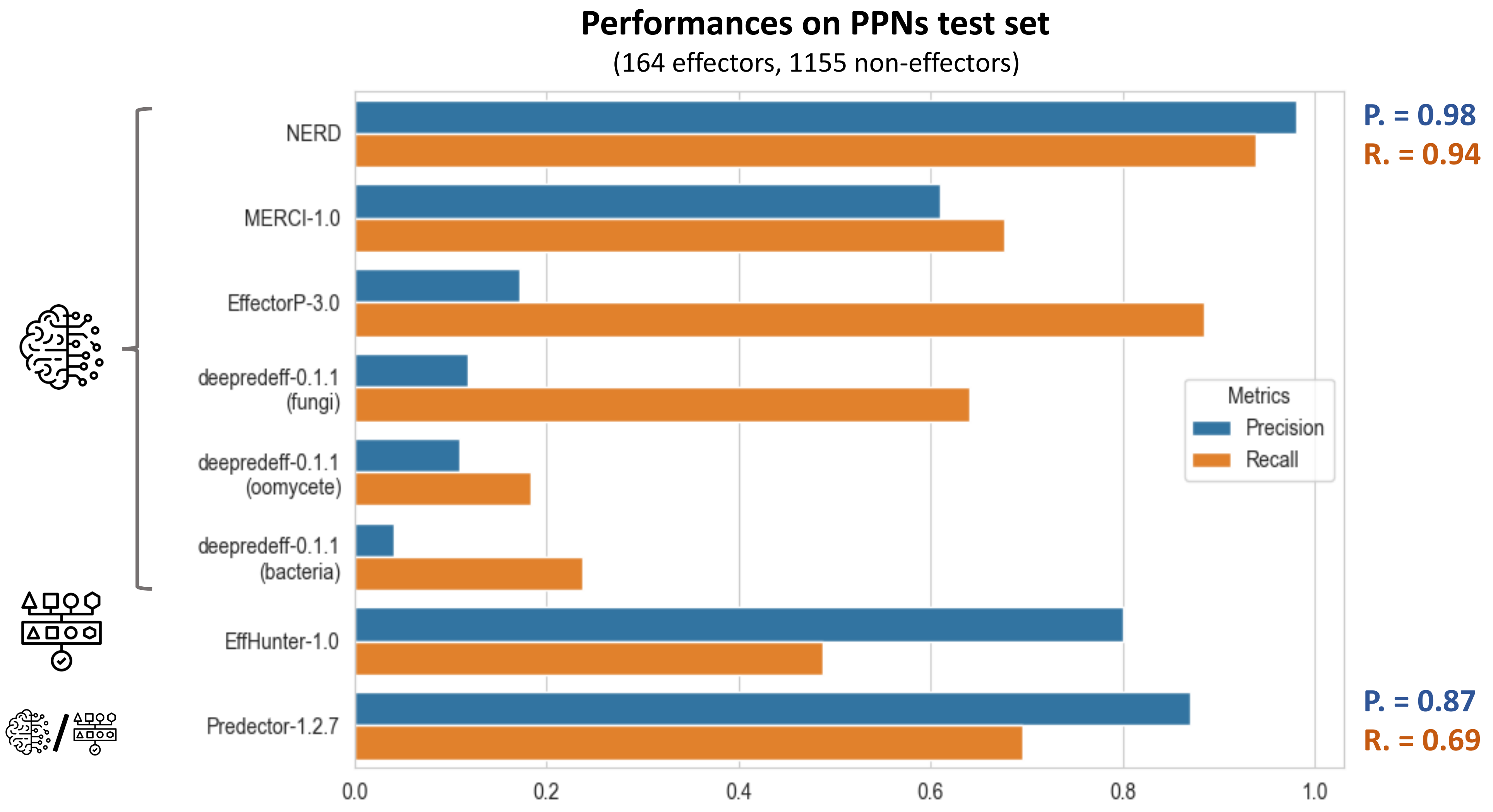
(4) Carreon-Anguiano *et al.*, biomolecules, 2020.

(5) Jones *et al.*, Scientific reports, 2021.

NERD performances evaluation and comparison with existing methods (test set)

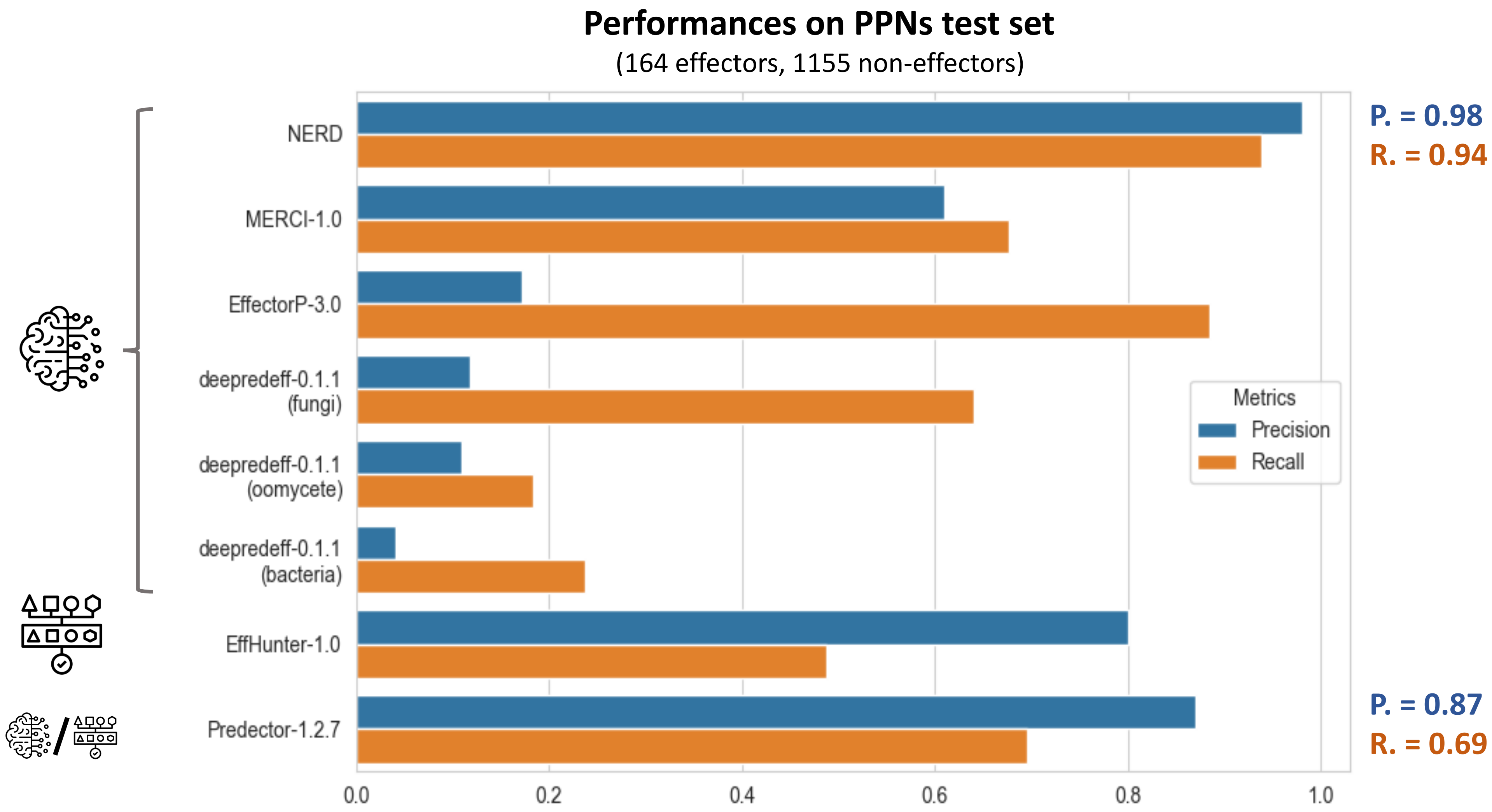


NERD performances evaluation and comparison with existing methods (test set)



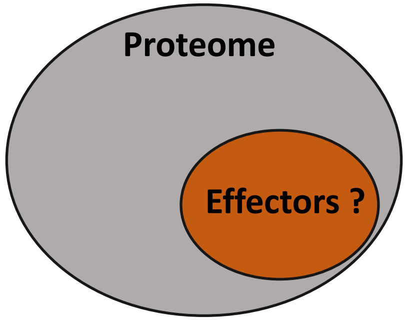
- NERD outperforms all the existing methods on the test set.

NERD performances evaluation and comparison with existing methods (test set)



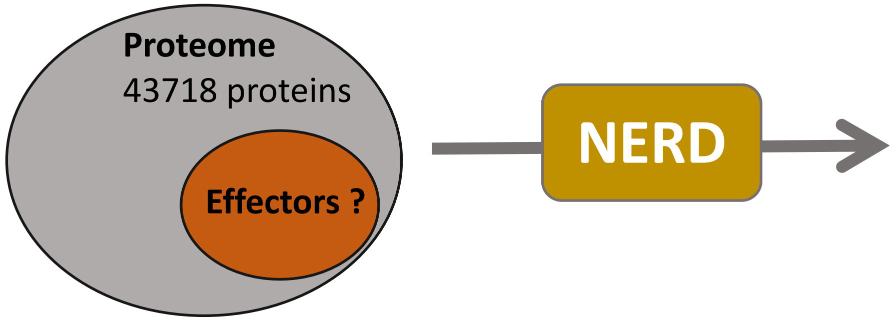
- NERD outperforms all the existing methods on the test set.
- NERD associates a probability to each sequence. Useful to rank the candidates for potential experimentation.

Real world predictions



Real world predictions

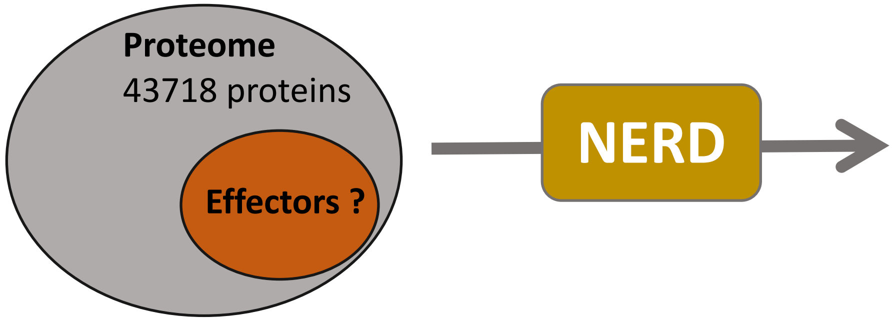
Meloidogyne incognita
Plant-parasitic Nematode



6985 candidates (15.9 %)

Real world predictions

Meloidogyne incognita
Plant-parasitic Nematode

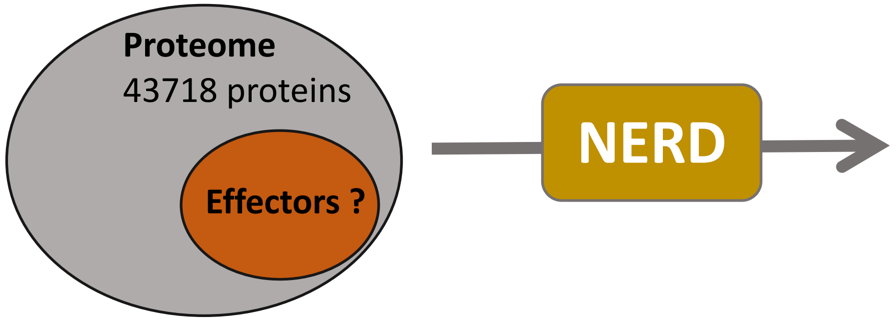


6985 candidates (15.9 %)

- High number of candidates but consistent with organism life style (multiple protein needed to parasitise).

Real world predictions

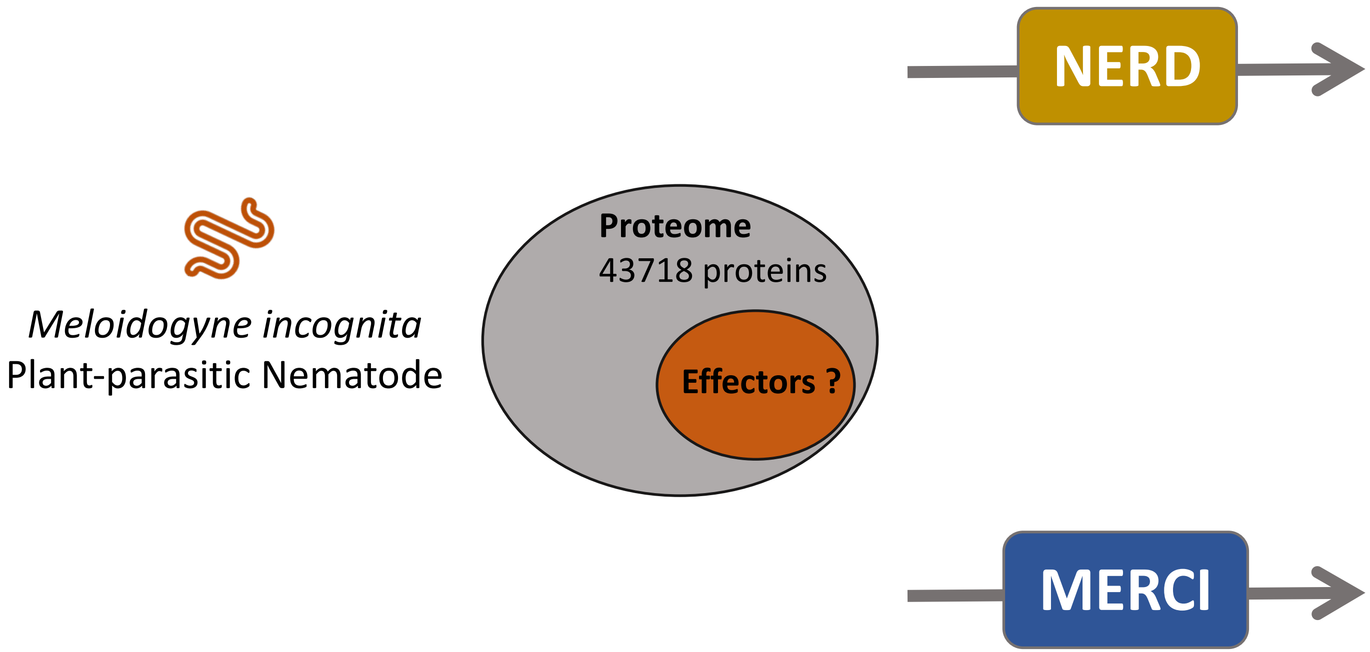
Meloidogyne incognita
Plant-parasitic Nematode



6985 candidates (15.9 %)

- High number of candidates but consistent with organism life style (multiple protein needed to parasitise).
- Finds 100% of *M. incognita* known effectors from positive test set (33 seqs.) with a prob. >= 0.999199

Real world predictions



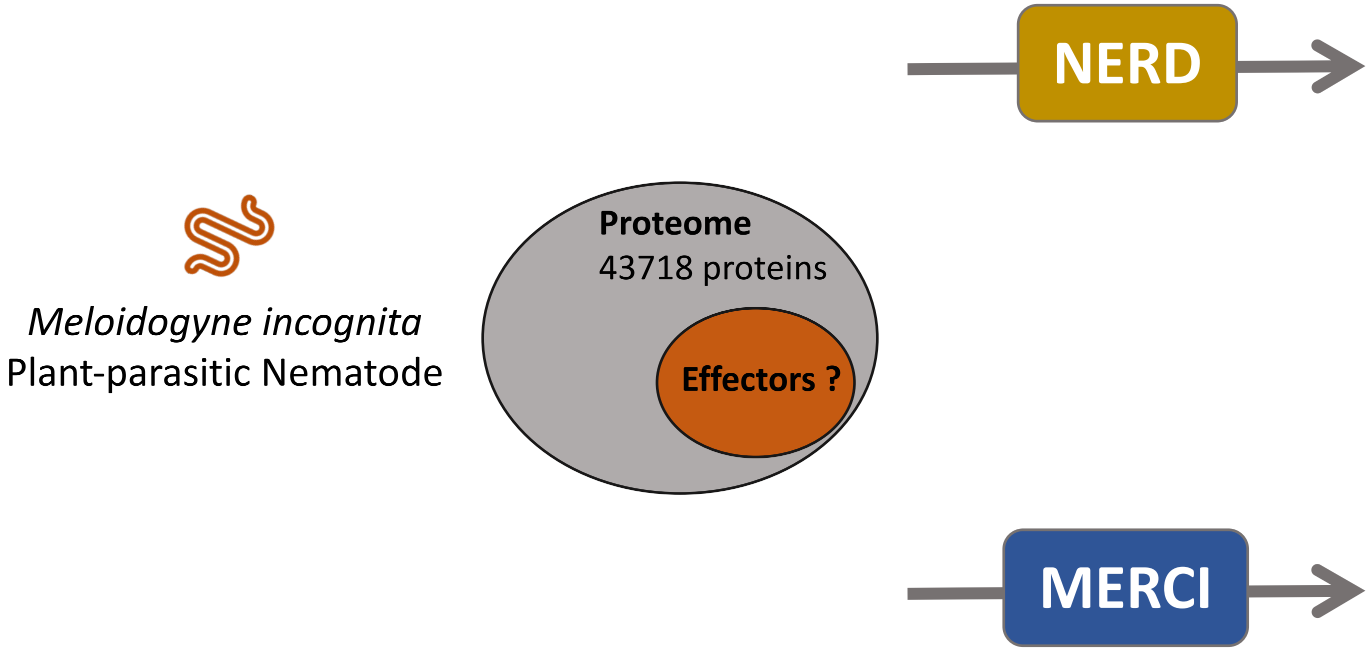
6985 candidates (15.9 %)

- High number of candidates but consistent with organism life style (multiple protein needed to parasitise).
- Finds 100% of *M. incognita* known effectors from positive test set (33 seqs.) with a prob. >= 0.999199

5217 candidates (11.9 %)

Vens *et al.*, Bioinformatics, 2011.

Real world predictions



Vens *et al.*, Bioinformatics, 2011.

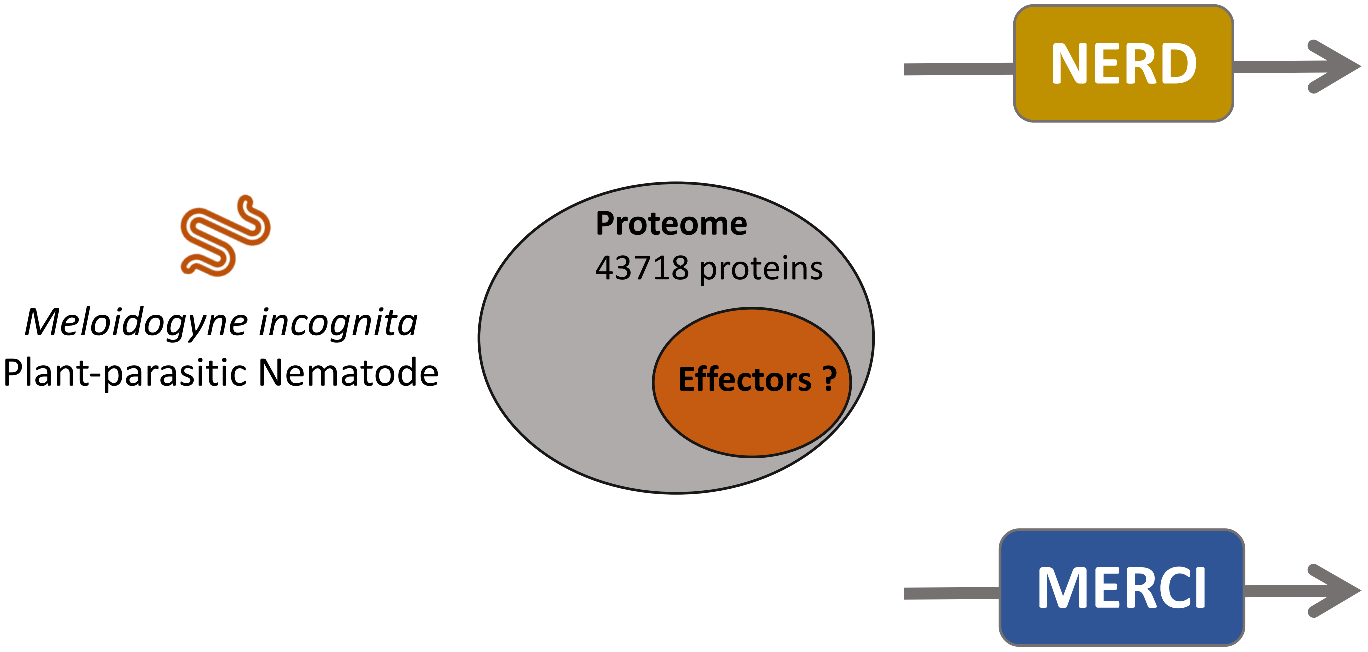
6985 candidates (15.9 %)

- High number of candidates but consistent with organism life style (multiple protein needed to parasitise).
- Finds 100% of *M. incognita* known effectors from positive test set (33 seqs.) with a prob. >= 0.999199

5217 candidates (11.9 %)

- Only finds 17/33 know effectors from the test set.

Real world predictions



Vens *et al.*, Bioinformatics, 2011.

6985 candidates (15.9 %)

- High number of candidates but consistent with organism life style (multiple protein needed to parasitise).
- Finds 100% of *M. incognita* known effectors from positive test set (33 seqs.) with a prob. >= 0.999199

5217 candidates (11.9 %)

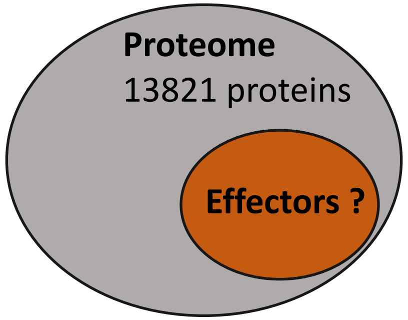
- Only finds 17/33 know effectors from the test set.

NERD seems suited for effectors prediction in PPN species.

Limitations

Drosophila melanogaster

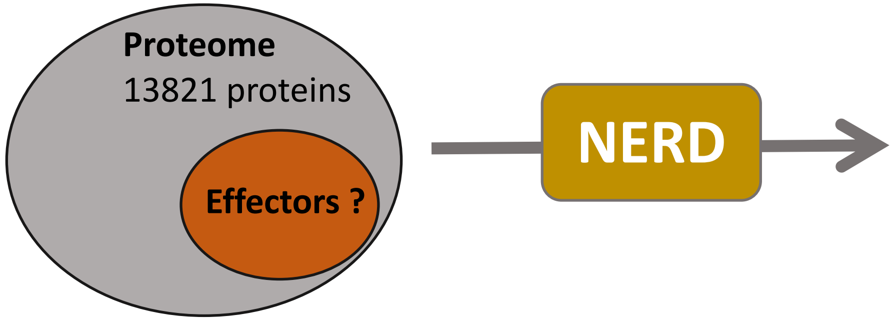
Not a nematode, not a plant parasite



Limitations

Drosophila melanogaster

Not a nematode, not a plant parasite



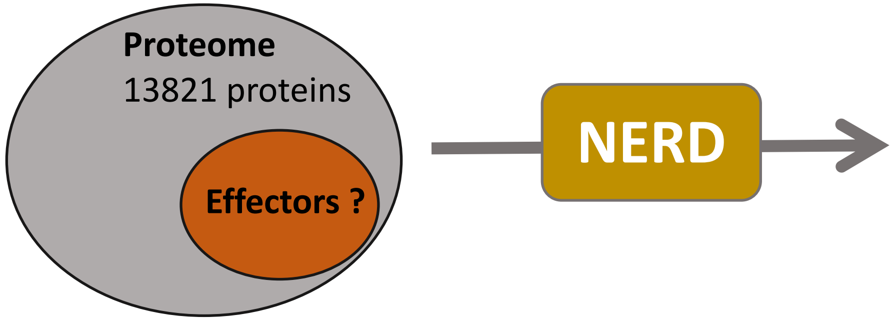
2169 candidates (15.7 %)

- High number of candidates. Not expected at all

Limitations

Drosophila melanogaster

Not a nematode, not a plant parasite



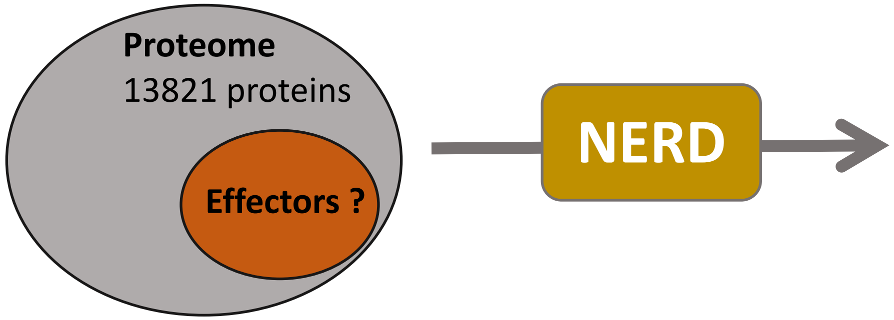
2169 candidates (15.7 %)

- High number of candidates. Not expected at all
    - ➢ Hyp. 1: Predicted proteins are truly similar to PPN effectors but have different functions.
    - ➢ Hyp. 2:Predictions are erroneous and the dataset / model need to be reworked.

Limitations

Drosophila melanogaster

Not a nematode, not a plant parasite



2169 candidates (15.7 %)

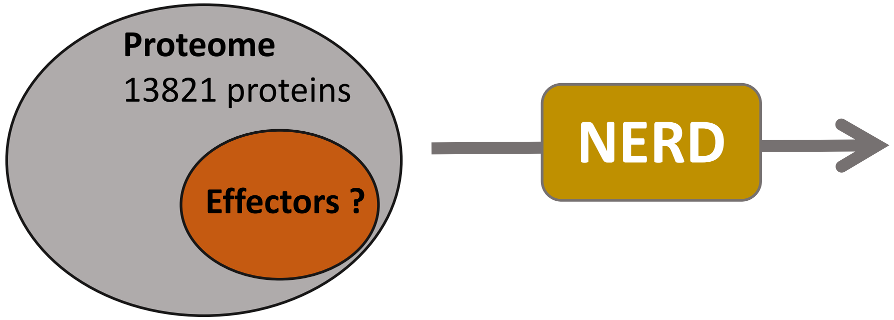
- High number of candidates. Not expected at all
  - Hyp. 1: Predicted proteins are truly similar to PPN effectors but have different functions.
  - Hyp. 2:Predictions are erroneous and the dataset / model need to be reworked.

Predicted effector in non plant parasitic species and parasitic species need to be compared

Limitations

Drosophila melanogaster

Not a nematode, not a plant parasite



2169 candidates (15.7 %)

- High number of candidates. Not expected at all
  - ➢ Hyp. 1: Predicted proteins are truly similar to PPN effectors but have different functions.
  - ➢ Hyp. 2:Predictions are erroneous and the dataset / model need to be reworked.

Predicted effector in non plant parasitic species and parasitic species need to be compared

BUT:

Molecular Plant Pathology

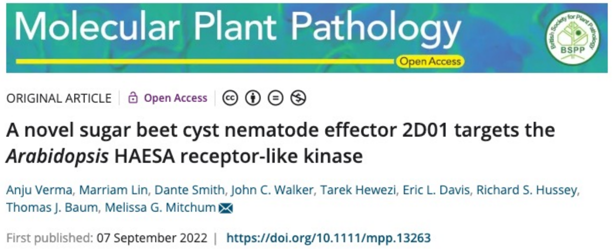
ORIGINAL ARTICLE | Open Access

A novel pine wood nematode effector, BxSCD1, suppresses plant immunity and interacts with an ethylene-forming enzyme in pine

Tong-Yue Wen, Xiao-Qin Wu, Long-Jiao Hu, Yi-Jun Qiu, Lin Rui, Yan Zhang, Xiao-Lei Ding, Jian-Ren Ye

First published: 16 August 2021 | https://doi.org/10.1111/mpp.13121 | Citations: 1

Predicted proba. : 0.9999



Molecular Plant Pathology

ORIGINAL ARTICLE | Open Access

A novel sugar beet cyst nematode effector 2D01 targets the *Arabidopsis* HAESA receptor-like kinase

Anju Verma, Marriam Lin, Dante Smith, John C. Walker, Tarek Hewezi, Eric L. Davis, Richard S. Hussey, Thomas J. Baum, Melissa G. Mitchum

First published: 07 September 2022 | https://doi.org/10.1111/mpp.13263

Predicted proba. : 0.9984

NERD successfully predicted as highly probable candidates two experimentally validated PPNs effectors (not present in the dataset)

Drosophila melanogaster

Not a nematode, not a plant parasite

Proteome
13821 proteins

Effectors ?

NERD

2169 candidates (15.7 %)

- High number of candidates. Not expected at all
  - Hyp. 1: Predicted proteins are truly similar to PPN effectors but have different functions.
  - Hyp. 2:Predictions are erroneous and the dataset / model need to be reworked.

BUT:

NERD seems suited for effectors prediction in PPN species but shows generalisation issues that need to be further investigated.

immunity and interacts with an ethylene-forming enzyme in pine

Tong-Yue Wen, Xiao-Qin Wu, Long-Jiao Hu, Yi-Jun Qiu, Lin Rui, Yan Zhang, Xiao-Lei Ding, Jian-Ren Ye

First published: 16 August 2021 | https://doi.org/10.1111/mpp.13121 | Citations: 1

Predicted proba. : 0.9999

Arabidopsis HAESA receptor-like kinase

Anju Verma, Marriam Lin, Dante Smith, John C. Walker, Tarek Hewezi, Eric L. Davis, Richard S. Hussey, Thomas J. Baum, Melissa G. Mitchum

First published: 07 September 2022 | https://doi.org/10.1111/mpp.13263

Predicted proba. : 0.9984

NERD successfully predicted as highly probable candidates two experimentally validated PPNs effectors (not present in the dataset)

TAKE HOME MESSAGES

NERD:

- outperforms existing tools on PPNs data.

- is useful to efficiently predict effectors in PPNs and will help experimental biologist in their candidate choice.

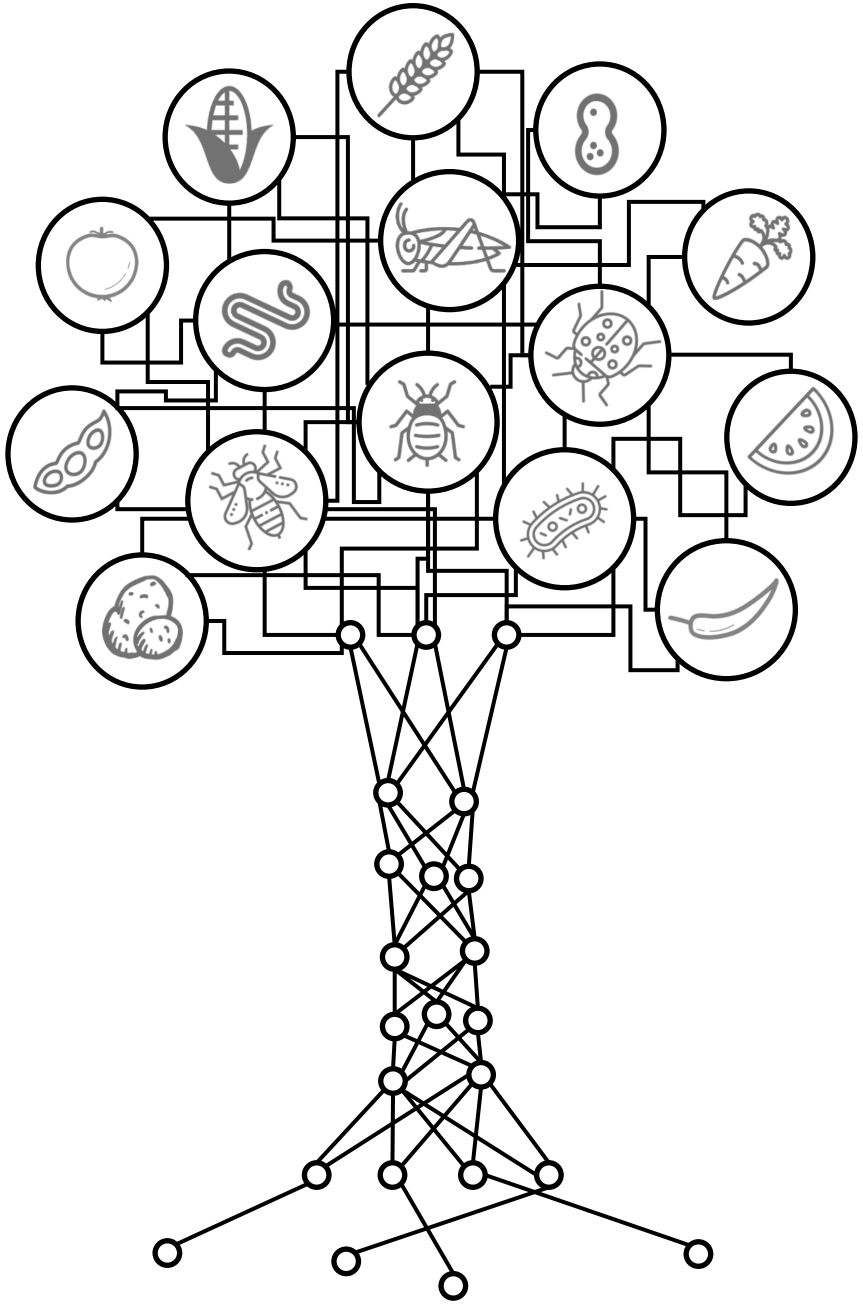
- gives good results despite small amount of positive sequences, a considerable plus for biology where sample can be scarce.

NERD:

- outperforms existing tools on PPNs data.
- is useful to efficiently predict effectors in PPNs and will help experimental biologist in their candidate choice.
- gives good results despite small amount of positive sequences, a considerable plus for biology where sample can be scarce.

Future improvements:

- NERD is still a work in progress : we noticed generalisation issues outside the PPNs which need to be further investigated.
- Extend NERD (new models) to other plant parasites which produce effectors proteins (mycetes, oomycetes)
- Study NERD predictions to identify underlying protein characteristics related to parasitism (Explanatory Learning)
- Other suggestions ?



NERD (Nematode EffectoR Discovery) : a tool to predict proteins involved in nematodes' plant parasitism.

Djampa KOZLOWSKI (MSI/UCA/INRAe)
djampa.kozlowski@univ-cotedazur.fr

Silvia BOTTINI, PhD.

UNIVERSITÉ CÔTE D'AZUR
MAISON DE LA MODÉLISATION, DE LA SIMULATION ET DES INTERACTIONS

Medical Data Laboratory

Etienne DANCHIN, PhD.

INRAe
la science pour la vie, l'humain, la terre

GAME

Maeva ANTOINE, PhD.

UNIVERSITÉ CÔTE D'AZUR
Centre de calculs

Thank you for your attention !