Inductive logic programming for the identification of markers predisposing to hospital acquired pneumonia

Juliette Audemard
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004

Supervised by
Morgan Magnin
Nantes Université, École Centrale Nantes, CNRS, LS2N
Tony Ribeiro
Nantes Université, École Centrale Nantes, CNRS, LS2N

HOMI-LUNG

CENTRALE NANTES

Nantes Université

LS2N Laboratoire des Sciences du Numérique de Nantes

HOMI-LUNG

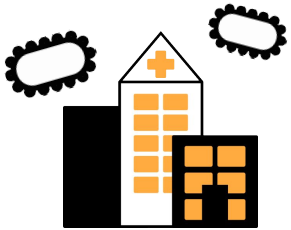
European project* (2024-2028)**, **led by the Pr. Antoine Roquilly**, aiming to *decipher the causal relationship between cardiovascular and respiratory diseases progression, the* ***immune and microbiome alterations *observed* ***during and after pneumonia***.

Funded by the European Union

"Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or ERCEA. Neither the European Union nor the granting authority can be held responsible for them."

Thesis context

Hospital Acquired Pneumonia (HAP)



Nosocomial respiratory infection acquired in hospital settings, more than 48h after hospital admission. Includes ventilator-associated pneumonia.

- **The most common infection acquired in hospitals in Europe**
- **10%–20% mortality rate**
  nearly fatal for vulnerable patients with comorbidities (organ transplant, immunodeficiency)
- **Culture-based diagnostics, systematic uses of antibiotics, commensal bacterium mistook for pathogens…**
  HAP complexity can't be fully assessed without an integrative reasoning

*Lower-left lobe with hypoxemia from a Pneumococcus infection

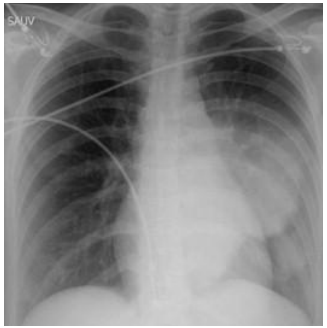
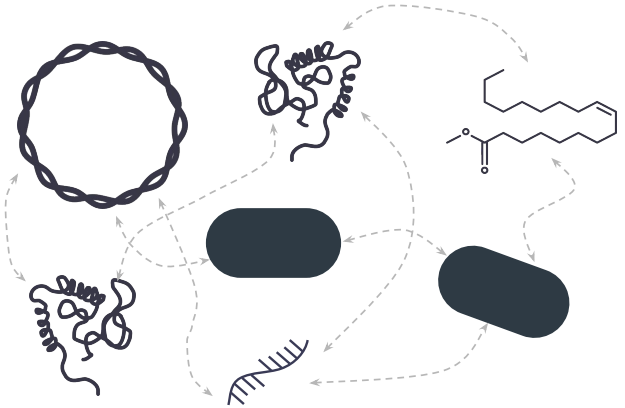


*Figure 1, from C. Beigelman-Aubry et al, Diagnostic and Interventional Imaging, 2012

Thesis context

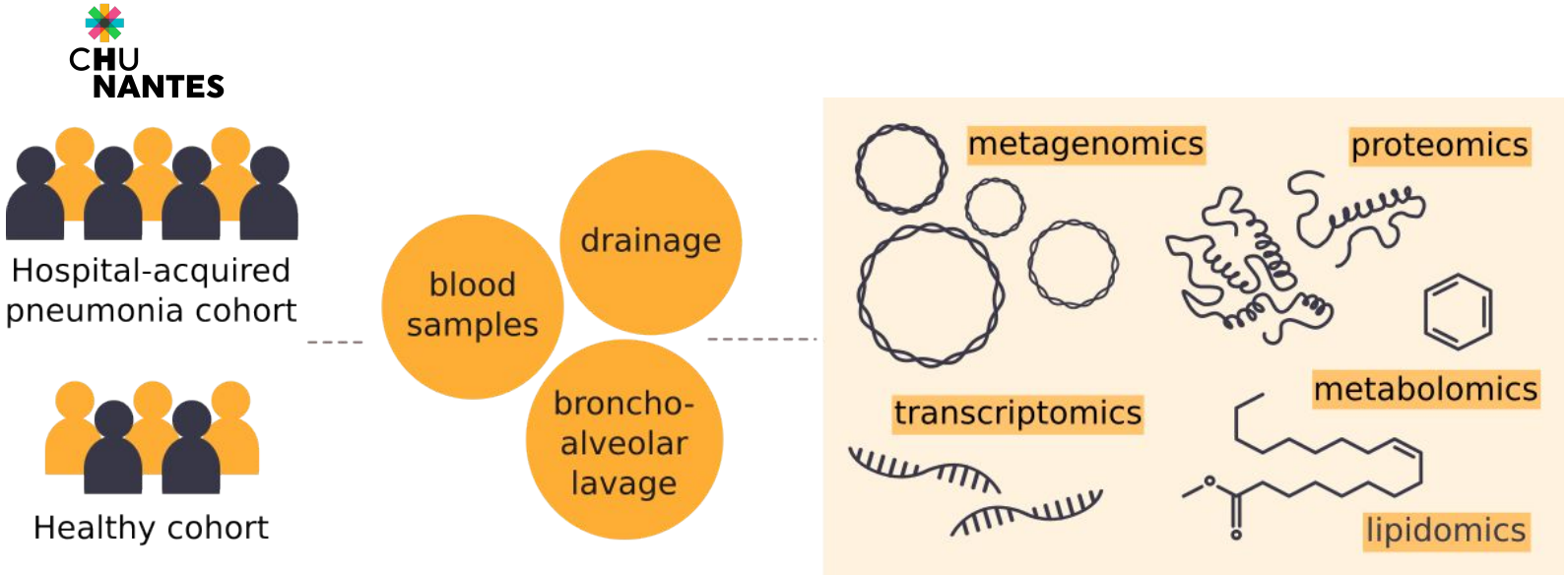
HomiLung Project ambitions

By integrating ecology, immunity, and time, HAP management can move from a reactive chase of pathogens to a proactive, personalised strategy, one that not only improves survival but also preserves resilience in the critically ill.
Martin-Loaches I., J Crit Care 2026



- **New paradigm focused on ecology and risk assessment**
- **Concept of *lungs interactome***
  Molecular, cellular and microbial interactions network of the host lung cells, virus, commensal and pathogenic bacteria
- **<u>Characterise key interactions in HAP patients lungs interactome</u>**

From HAP and healthy cohorts to multi-omics



Thesis context

From multi-omics to lungs interactions modelling

What are the relations between

- **Bronchoalveolar and plasma proteins?**
- **Gender and bacteria communities during lung infection?**
- **Pneumonia phenotypes and proteins expression?**

From multi-omics to lungs interactions modelling

What are the relations between

- **Bronchoalveolar and plasma proteins?**
- **Gender and bacteria communities during lung infection?**
- **Pneumonia phenotypes and proteins expression?**

Logical models could answer that !

Thesis context

Learning from interpretation transition (LFIT)

Python package
pylfit
github.com/Tony-sama/pylfit

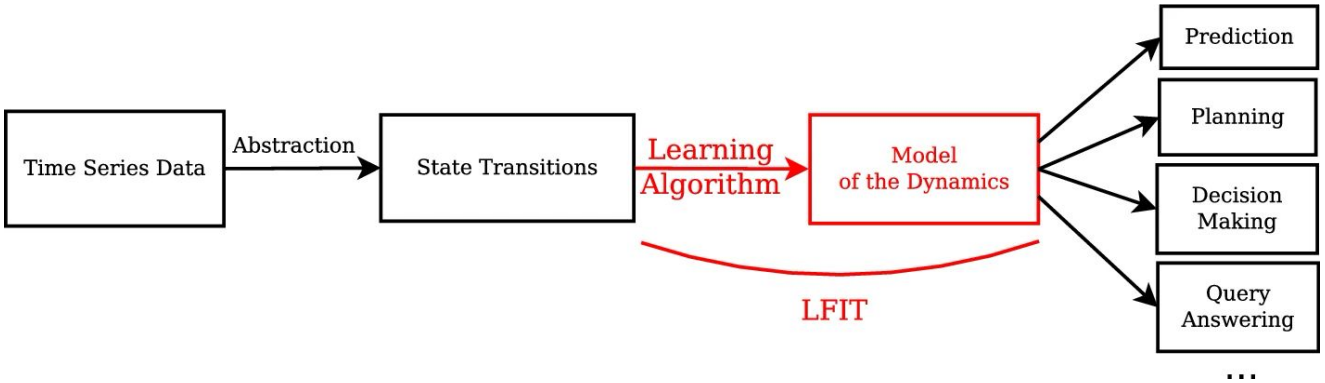


Figure 1 from Ribeiro, T., Folschette, M., Magnin, M., & Inoue, K. (2021). *Learning any memory-less discrete semantics for dynamical systems represented by logic programs.* Machine Learning

- **Developed since 2014**
- **Generate explainable models of a system dynamics**
- **Takes heterogenous data as an input, as long as they are discrete**

Learning from interpretation transition (LFIT)

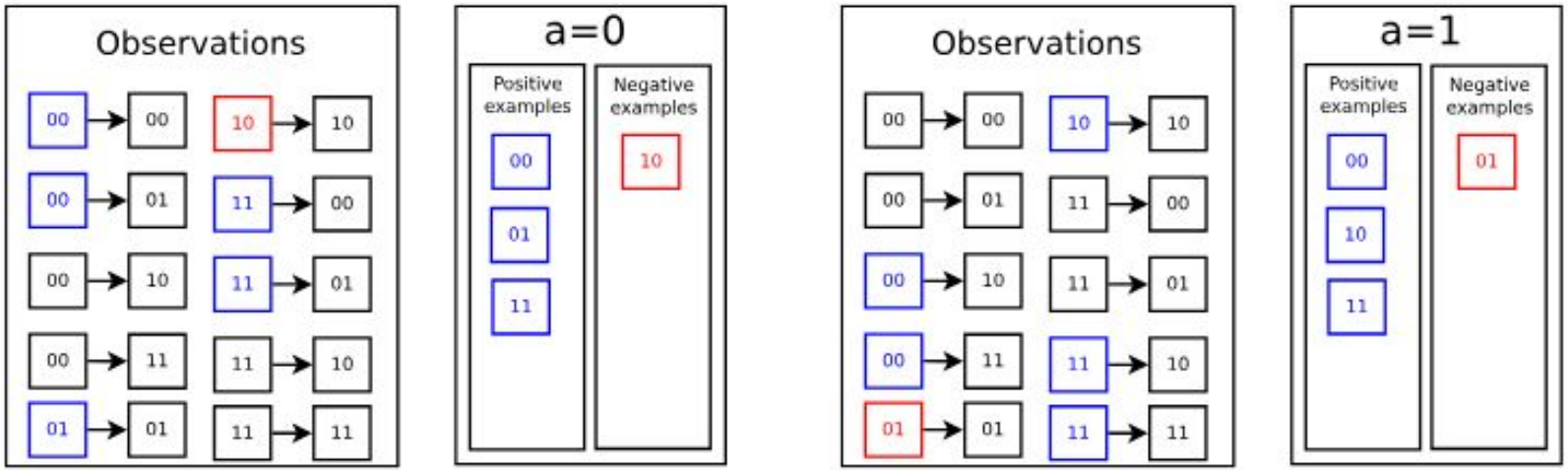
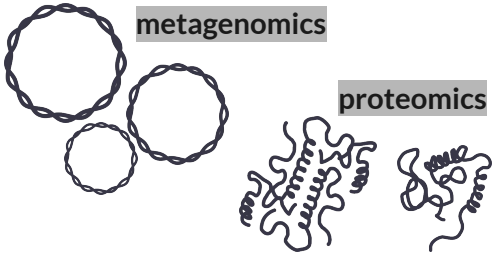


Figure 6 from Ribeiro, T., Folschette, M., Magnin, M., & Inoue, K. (2021). Learning any memory-less discrete semantics for dynamical systems represented by logic programs. Machine Learning

From biological data to a logic program

Using LFIT to identify relations between proteins detection and bacterial communities



B_1	B_2	B_3	...	B_121	B_122
0	0	1		0	0
0	1	0		0	0
1	1	1		1	0

P_1	P_2	P_3	...	P_654	P_655
0	1	0		1	1
1	1	0		0	0
1	0	1		0	1

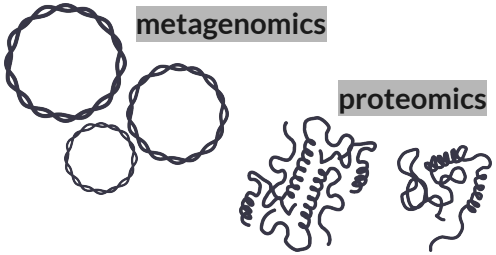
LFIT

```
...
P_44(1) :- B_6(1), B_33(0), B_58(1), B_112(0)
P_44(1) :- B_6(0), B_84(1)
P_44(1) :- B_8(1)
P_44(0) :- B_9(1), B_12(1)
P_44(0) :- B_27(0), B_53(1)
....
```

- **Do we have enough observations for the model to be generalizable?**
- **Will the amount of produced rules be interpretable?**
- **Does the rules match litterature knowledge?**

From biological data to a logic program

Using LFIT to identify relations between proteins detection and bacterial communities



B_1	B_2	B_3	...	B_121	B_122
0	0	1		0	0
0	1	0		0	0
1	1	1		1	0

P_1	P_2	P_3	...	P_654	P_655
0	1	0		1	1
1	1	0		0	0
1	0	1		0	1

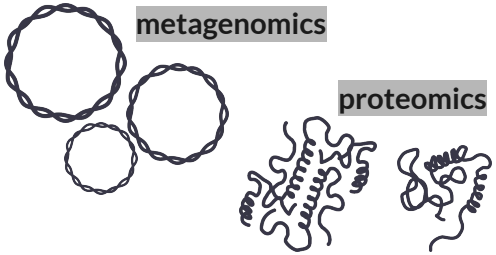
LFIT

```
...
P_44(1) :- B_6(1), B_33(0), B_58(1), B_112(0)
P_44(1) :- B_6(0), B_84(1)
P_44(1) :- B_8(1)
P_44(0) :- B_9(1), B_12(1)
P_44(0) :- B_27(0), B_53(1)
....
```

- **Do we have enough observations for the model to be generalizable?**
  Most likely no ; there is so much features, we won't cover all possible transitions
- **Will the amount of produced rules be interpretable?**
- **Does the rules match litterature knowledge?**

From biological data to a logic program

Using LFIT to identify relations between proteins detection and bacterial communities



B_1	B_2	B_3	...	B_121	B_122
0	0	1		0	0
0	1	0		0	0
1	1	1		1	0

P_1	P_2	P_3	...	P_654	P_655
0	1	0		1	1
1	1	0		0	0
1	0	1		0	1

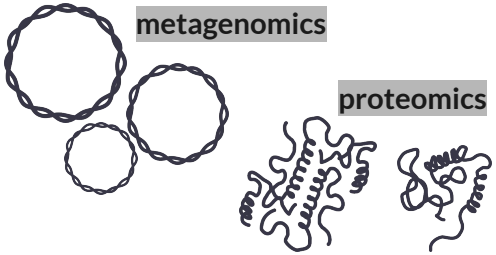
LFIT

```
...
P_44(1) :- B_6(1), B_33(0), B_58(1), B_112(0)
P_44(1) :- B_6(0), B_84(1)
P_44(1) :- B_8(1)
P_44(0) :- B_9(1), B_12(1)
P_44(0) :- B_27(0), B_53(1)
....
```

- **Do we have enough observations for the model to be generalizable?**
  Most likely no ; there is so much features, we won't cover all possible transitions
- **Will the amount of produced rules be interpretable?**
  Without a guided search, it might be hard to identify relevant information
- **Does the rules match litterature knowledge?**

From biological data to a logic program

Using LFIT to identify relations between proteins detection and bacterial communities



B_1	B_2	B_3	...	B_121	B_122
0	0	1		0	0
0	1	0		0	0
1	1	1		1	0

P_1	P_2	P_3	...	P_654	P_655
0	1	0		1	1
1	1	0		0	0
1	0	1		0	1

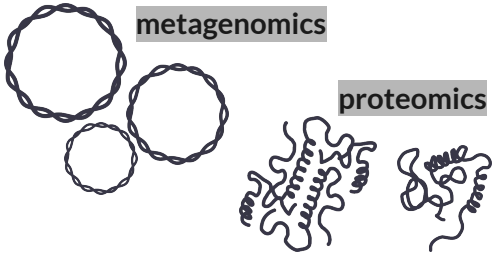
LFIT

```
...
P_44(1) :- B_6(1), B_33(0), B_58(1), B_112(0)
P_44(1) :- B_6(0), B_84(1)
P_44(1) :- B_8(1)
P_44(0) :- B_9(1), B_12(1)
P_44(0) :- B_27(0), B_53(1)
....
```

- **Do we have enough observations for the model to be generalizable?**
  Most likely no ; there is so much features, we won't cover all possible transitions
- **Will the amount of produced rules be interpretable?**
  Without a guided search, it might be hard to identify relevant information
- **Does the rules match litterature knowledge?**
  It can be checked

From biological data to a logic program

Using LFIT to identify relations between proteins detection and bacterial communities



B_1	B_2	B_3	...	B_121	B_122
0	0	1		0	0
0	1	0		0	0
1	1	1		1	0

P_1	P_2	P_3	...	P_654	P_655
0	1	0		1	1
1	1	0		0	0
1	0	1		0	1

LFIT

```
...
P_44(1) :- B_6(1), B_33(0), B_58(1), B_112(0)
P_44(1) :- B_6(0), B_84(1)
P_44(1) :- B_8(1)
P_44(0) :- B_9(1), B_12(1)
P_44(0) :- B_27(0), B_53(1)
....
```

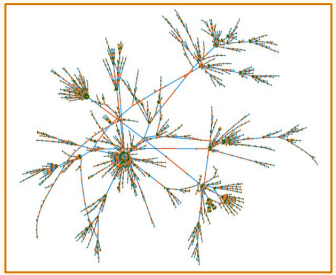
- **Do we have enough observations for the model to be generalizable?**
  Most likely no ; there is so much features, we won't cover all possible transitions
- **Will the amount of produced rules be interpretable?**
  Without a guided search, it might be hard to identify relevant information
- **Does the rules match litterature knowledge?**
  It can be checked

We should use background knowledge to address those problematics

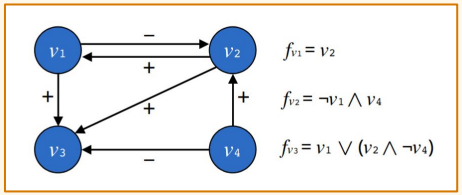
BioGRID STRING KEGG Kyoto Encyclopedia of Genes and Genomes RegulonDB BiGG BioModels

Prior knowledge network

Structured collection of interactions between different atoms
(e.g. enzymes, small molecules, genes, transcription factors)



BiGG metabolic model for
E. coli 042-iEC042_1314



A Boolean logical model
Aleixo F. et al, 2023

- **Logical models**
- **Regulatory graph**
- **Literature**

Over the last decades, more and more knowledge on molecular interactions and regulation has been available

Prior knowledge networks

Prior knowledge integration and ILP

Early works on ILP already consider integration of prior knowledge, but relies on the accuracy of it with respect to the data

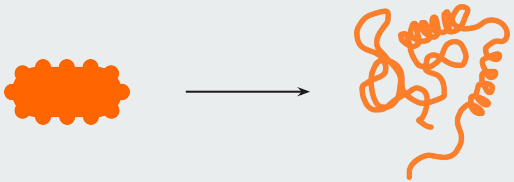
- **Is the background knowledge reliable enough?**
  is it knowledge or suppositions?

Prior knowledge integration and ILP

Early works on ILP already consider integration of prior knowledge, but relies on the accuracy of it with respect to the data

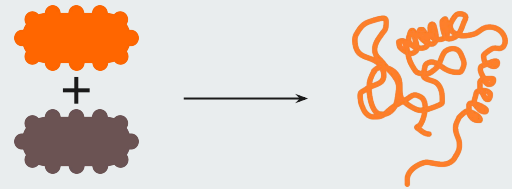
- **Is the background knowledge reliable enough?**
  is it knowledge or suppositions?
  - **Most databases also gather uncurated knowledge**
    Specially for less studied biological processes (*the lungs interactome*)
  - ***Ready-to-use* models can include unreliable information**
    From automated reconstruction, or past work

Prior knowledge rule R:



Protein_Or(1) :- Bacteria_Or(1)

Data-driven rule R:



Protein_Or(1) :- Bacteria_Or(1), Bacteria_Br(1)

The influence of the brown bacteria wasn't known

Prior knowledge integration and ILP

Early works on ILP already consider integration of prior knowledge, but relies on the accuracy of it with respect to the data

- **Is the background knowledge reliable enough?**
  is it knowledge or suppositions?

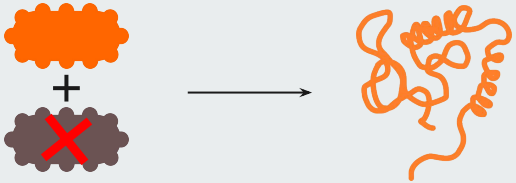
- **Even if it is, could the context differences between its curation and the experiment leads to different behaviour of the system**?

Prior knowledge integration and ILP

Early works on ILP already consider integration of prior knowledge, but relies on the accuracy of it with respect to the data

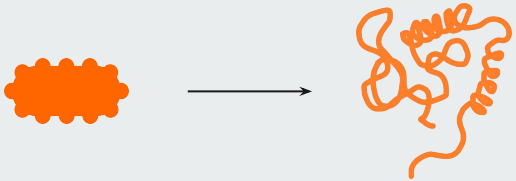
- **Is the background knowledge reliable enough?**
  is it knowledge or suppositions?
- **Even if it is, could the context differences between its curation and the experiment leads to different behaviour of the system?**

Prior knowledge rule R:



Protein_Or(1) :- Bacteria_Or(1), Bacteria_Br(0)

Data-driven rule R:



Protein_Or(1) :- Bacteria_Or(1)

Prior knowledge is right, but in the data, the orange bacteria is never observed together with the brown bacteria

Prior knowledge integration into an ILP framework

To guide learning without constraining

1) **The state covered by the prior knowledge rule R is never observed in the pair of states (s,s')**
2) **The prior knowledge rule R is too general in our context**
3) **The prior knowledge rule R is inconsistent with the data**
4) **The prior knowledge rule R includes feature(s) that are not seen in our data**

Prior knowledge integration into an ILP framework

To guide learning without constraining *(using unknown values)*

Learning From Interpretation Transitions with Unknowns

Ribeiro, T., Folschette, M., Magnin, M., Okazaki, K., Lo, K.-Y., Roquilly, A., Poschmann, J., & Inoue, K. (2024). Learning from Interpretation Transition with Unknowns.

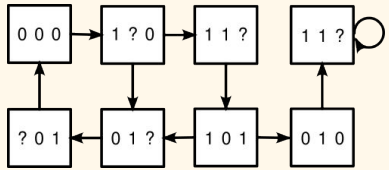
- **More realistic for biological data handling**
- **Learn a model from incomplete data, through an over approximation of the dynamics**

features

B_1	B_2	B_3	…	B_121	B_122
0	0	1	…	?	0
0	1	?	…	0	0
…	…	…	…	…	…

targets

P_1	P_2	P_3	…	P_654	P_655
?	1	0	…	1	0
1	1	0	…	?	1
…	…	…	…	…	…



To guide learning without constraining *(using unknown values)*

1) The state covered by the prior knowledge rule R is never observed in the pair of states (s,s')

$\forall (s, s') \in S^2, R \not\subseteq s$

Prior knowledge:

Protein_Or(1) :- Bacteria_Or(1), Bacteria_Br(0)

But the data-driven rule is:

Protein_Or(1) :- Bacteria_Or(1)

Mismatch

The brown and orange bacteria are never observed together in our dataset

features

B_1	B_2	B_3	…	B_121	B_122
0	0	1	…	0	0
0	1	0	…	0	0
…	…	…	…	…	…

targets

P_1	P_2	P_3	…	P_654	P_655
0	1	0	…	1	0
1	1	0	…	0	1
…	…	…	…	…	…

To guide learning without constraining *(using unknown values)*

1) The state covered by the prior knowledge rule R is never observed in the pair of states (s,s')

$\forall (s, s') \in S^2, R \not\subseteq s$

- **The prior knowledge is treated as a prior observation, where the variables not in the rule body take unknown values**
- The prior knowledge rule will be a subset or equal to the one inferred by LFIT

1) Transform prior knowledge rule into partial observation

P_655(1) :- B_2(1), B_3(0) → [['?','1','0',...,'?','?'],['?','?','?',...,'?','1']]

partial features *partial targets*

2) If the state covered by the rule is not in the data, add it

features

B_1	B_2	B_3	...	B_121	B_122
0	0	1	...	0	0
0	1	0	...	0	0
?	1	1	...	?	?
...	...	...	...	...	...

+

targets

P_1	P_2	P_3	...	P_654	P_655
0	1	0	...	1	0
1	1	0	...	0	1
?	?	?	...	?	1
...	...	...	...	...	...

+

3) Process with LFIT learning phase

To guide learning without constraining *(using unknown values)*

4) The prior knowledge rule R includes feature(s) that are not seen in our data

Prior knowledge rule:

Protein_Or(1) :- Bacteria_Or(1), Bacteria_Gr(1)

Not seen in experiment, but may be later

But the data-driven rule is:

Protein_Or(1) :- Bacteria_Or(1)

To guide learning without constraining *(using unknown values)*

4) The prior knowledge rule R includes feature(s) that are not seen in our data

- **The prior knowledge is treated as a prior observation, and the new atom is treated as an unknown value in the experiment's observations**
- **The new atom must have at least 2 different possible values**

1) Transform prior knowledge rule into partial observation

P_655(1) :- B_2(1), B_Gr(1) ⟶ [['?','1','?',...,'?','1'],['?','?','?',...,'?','1']]

partial features *partial targets*

2) Add the new feature in the dataset

features

B_1	B_2	B_3	...	B_121	B_122	B_Gr
0	0	1	...	0	0	?
0	1	0	...	0	0	?
?	1	?	...	?	?	1
...	...	...	...	...	...	...

+ +

targets

P_1	P_2	P_3	...	P_654	P_655
0	1	0	...	1	0
1	1	0	...	0	1
?	?	?	...	?	1
...	...	...	...	...	...

+

3) Process with LFIT learning phase

Conclusion and perspectives

- **Leverage prior knowledge to address the partiality of observations**
- **Use of prior knowledge as a guide rather than a constraint**
  - ★ Consider the data and prior knowledge on an equal level
- **Following work : test of the established scenario on synthetic networks**
  - ★ Accuracy
  - ★ Speed
- **Inconsistencies between prior knowledge and data still need to be handled**
  - ★ Putative solution : weight between incomplete prior knowledge and noisy data