Prediction of dynamical properties of biochemical pathways with Graph Neural Networks

Pasquale Bove
Alessio Micheli
Paolo Milazzo
Marco Podda

Department of Computer Science – University of Pisa

milazzo@di.unipi.it

Full text paper

- This presentation is based on the paper

Bove, P.; Micheli, A.; Milazzo, P. and Podda, M. (2020). **Prediction of Dynamical Properties of Biochemical Pathways with Graph Neural Networks**.In *Proc. 13th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 3 BIOINFORMATICS.* pages 32-43. DOI: 10.5220/0008964700320043

- You can download it from https://www.scitepress.org/PublicationsDetail.aspx?ID=x5i8GvSYgwE=&t=1

The BioSystems Modelling Group @UNIPI

- **Web page:** http://www.di.unipi.it/msvbio/
- **People:** R. Barbuti, P. Bove, R. Gori, F. Levi, P. Milazzo, L. Nasti

Activity started in 2004, with the aim of developing formal modeling and analysis techniques for biological systems

Main **areas of expertise**:

- Modeling of biochemical processes, evolution problems and ecosystems
- Differential equations and stochastic simulation
- Formal methods: process algebras, rewriting systems, model checking

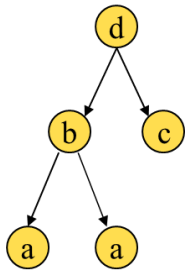
CIML group @UNIPI

Computational Intelligence & Machine Learning Group

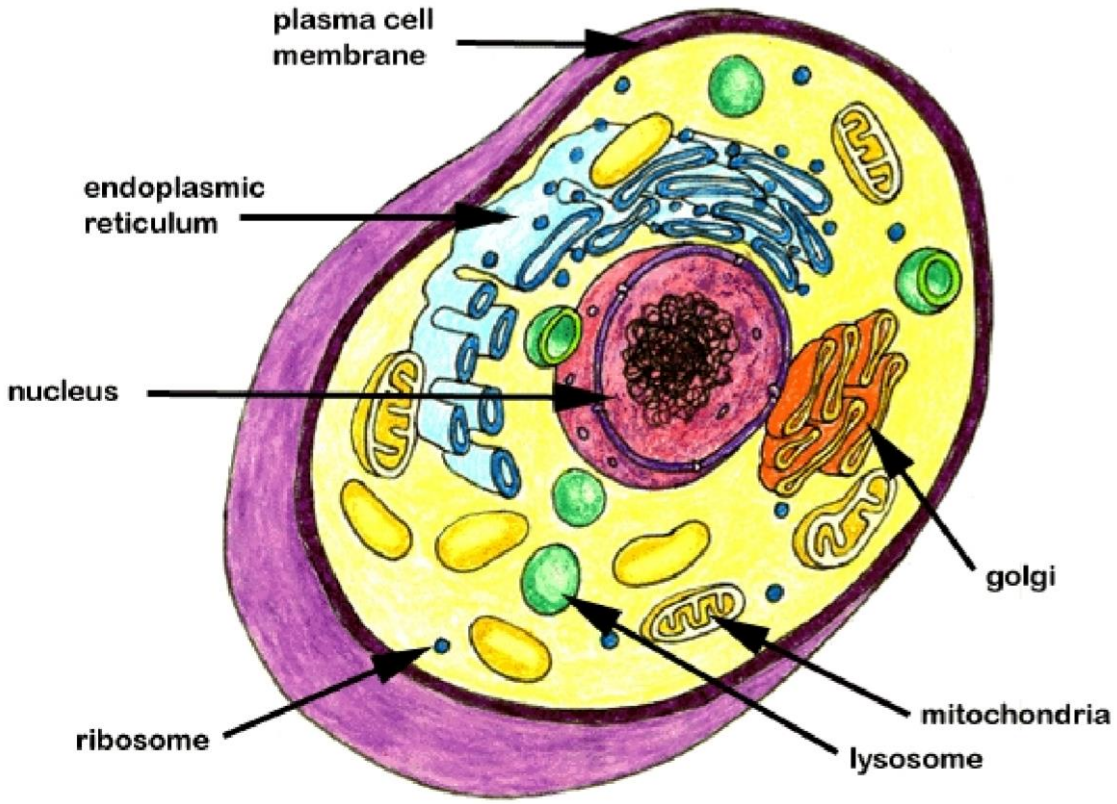
- **Web page:** http://www.di.unipi.it/groups/ciml
- **People:** A. Micheli (coordinator), D. Bacciu, C. Gallicchio, 7 Phd students + 6 post-doc/research associates

Development of basic and applied research on Machine Learning

- Learning in Structured Domains (SD): sequence, trees and graphs
- Neural Networks & Deep learning for SD



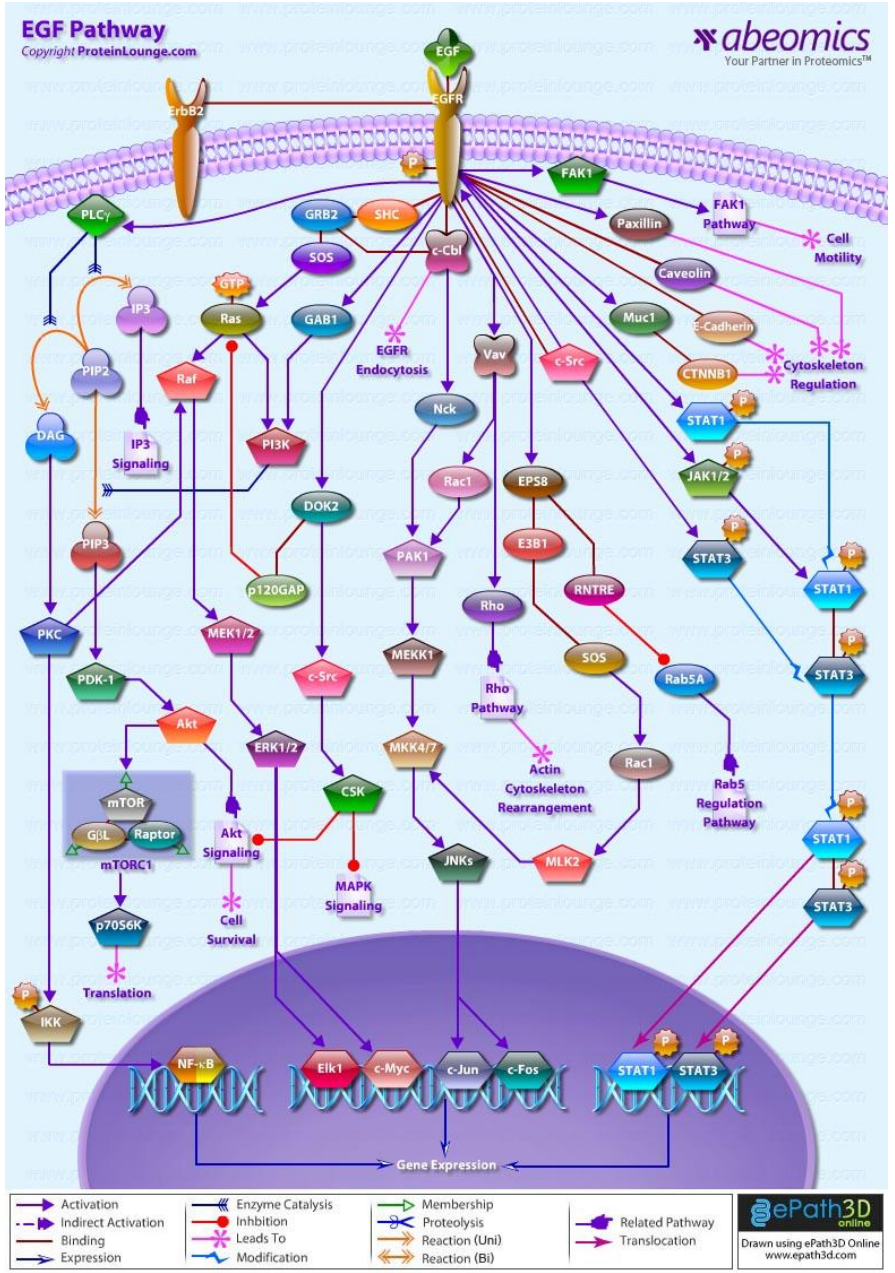
The functioning of living cells



- Cells are complex systems
- Main actors:
  - DNA
  - RNA
  - Proteins
  - Metabolites
  - ……
- Interaction networks:
  - Metabolic pathways
  - Signalling pathways
  - Gene regulatory networks

Biochemical pathways

- A biochemical pathway (metabolic/signaling) is a set of chemical reactions involving biomolecules
- Often denoted as a graph
  - Several notations exist
- Pathways implement cell functionalities



Biochemical pathways in SBML

- A standard language for the description of biochemical pathways is SBML
- A pathway is modeled as a list of reactions
- Each reaction has a list of reactants, products and modifiers
- Rate formulas can be specified

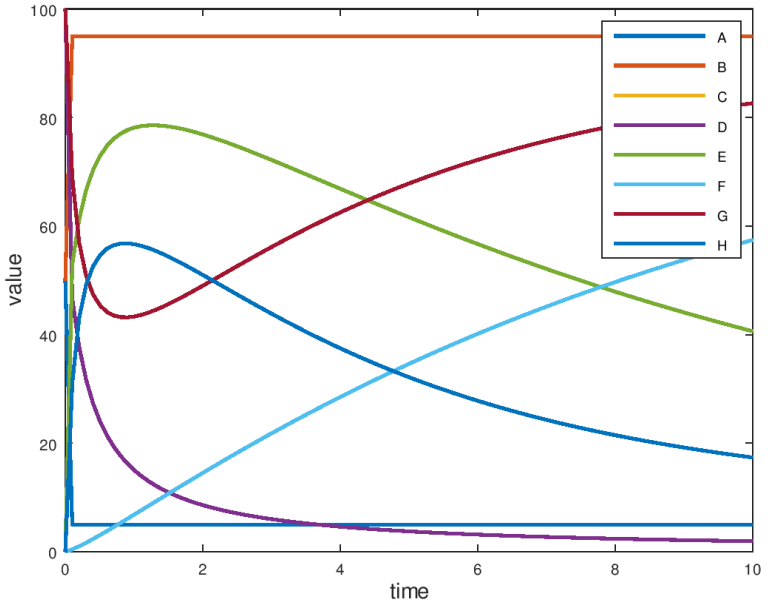
```
...
<reaction id='r1'>
    <listOfReactants>
      ...
    </listOfReactants>
    <listOfProducts>
      ...
    </listOfProducts>
    <listOfModifiers>
      ...
    </listOfModifiers>
</reaction>
...
```

Simulation of pathway dynamics

- Pathway dynamics is how the concentrations of the involved molecules vary over time
- Typical analysis techniques:
  - numerical (ODE-based) and stochastic simulation

Reaction	*Mod*	*Kinetics*
$A + B \rightarrow 2B$		k_1AB
$B \rightarrow A$		k_2B
$C + D \rightarrow E$	A	k_3CDA
$E \rightarrow F$		k_4E
$F \rightarrow E$		k_5F
$G \rightarrow H$	F	$\frac{k_6G}{1+2F}$
$H \rightarrow G$		k_7G

$$\frac{dA}{dt} = -k_1AB + k_2B$$
$$\frac{dB}{dt} = k_1AB - k_2B$$
$$\frac{dC}{dt} = -k_3CDA$$
$$\frac{dD}{dt} = -k_3CDA$$
$$\frac{dE}{dt} = k_3CDA - k_4E + k_5F$$
$$\frac{dF}{dt} = k_4E - k_5F$$
$$\frac{dG}{dt} = -\frac{k_6G}{1+2F} + k_7G$$
$$\frac{dH}{dt} = \frac{k_6G}{1+2F} - k_7G$$



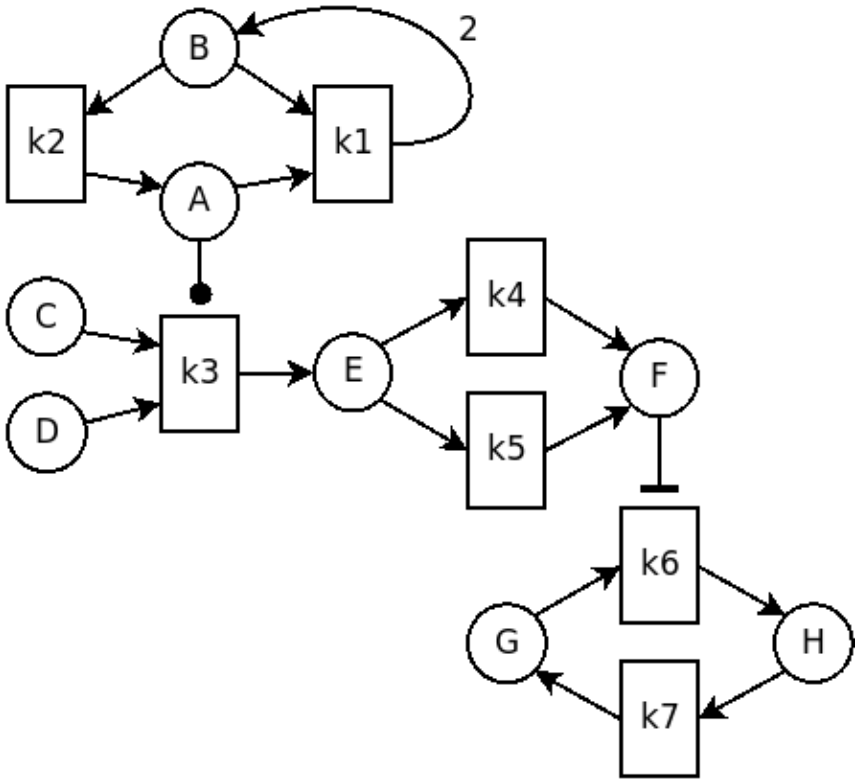
Dynamical Properties

- Simulations aim at assessing dynamical properties:
  - Steady states
  - Oscillatory behaviours
  - Sensitivity
  - Robustness

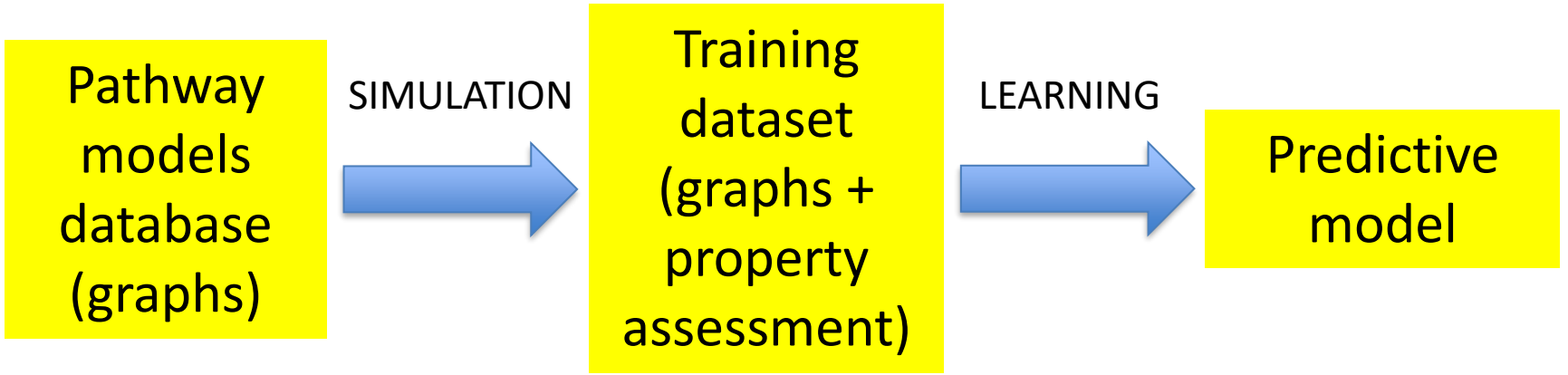
- Property assessment through simulation is often expensive:
  - Stiffness/scalability problems
  - Huge number of simulations to vary parameters/initial values

The Idea…

- Biochemical pathway can be represented as graphs (e.g. Petri nets)
- **Assumption:** Dynamical properties of pathways could be correlated with topological properties of their graphs
- Let's infer such topological properties through Machine Learning (ML) on graphs
- The ML model could then predict the dynamical property by avoiding the burden of expensive numerical simulations

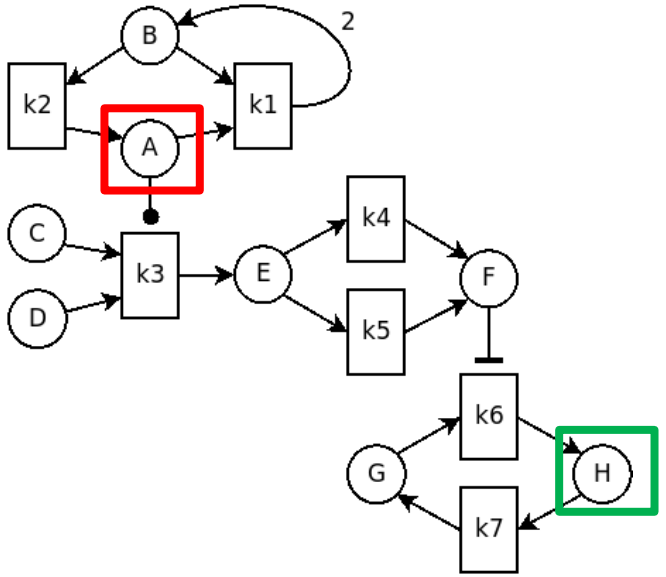


The approach



Essay: prediction of concentration robustness

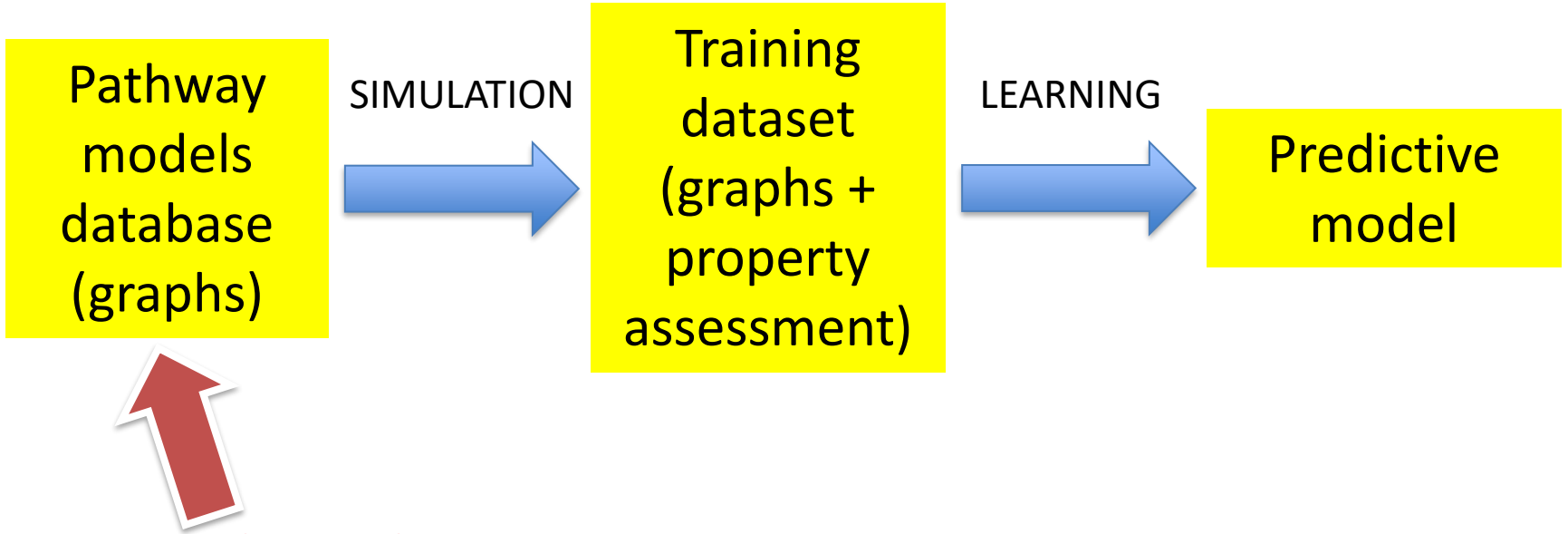
- Concentration robustness:
  - Preservation of steady state concentrations despite perturbations on initial conditions



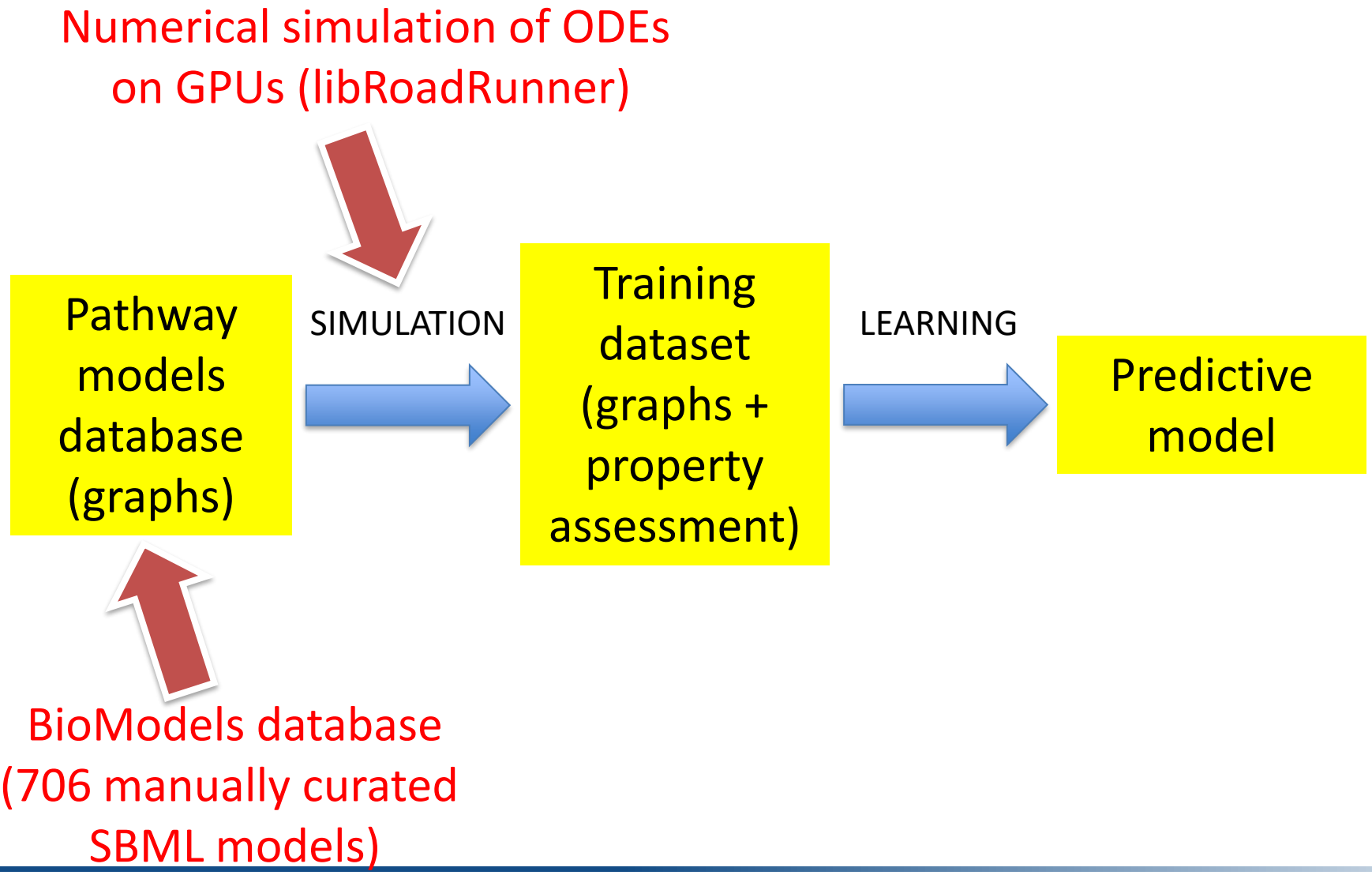
- More precisely:
  - Relative α-robustness
  - Given an input species I and an output species O it is as follows:

$$1 - \frac{\textit{size of the steady state concentration interval of O}}{\textit{size of the initial concentration interval of I}}$$

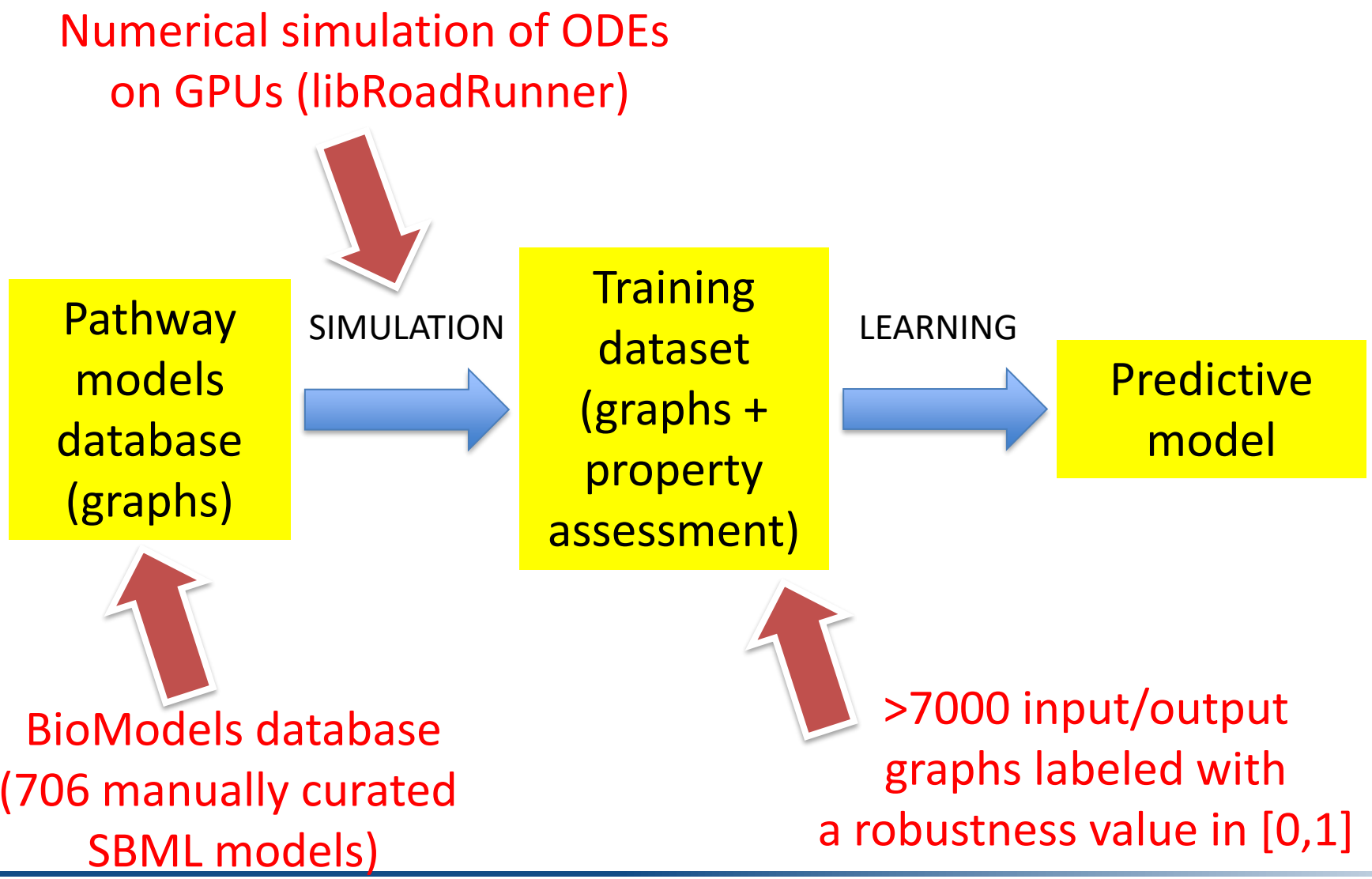
Methodology



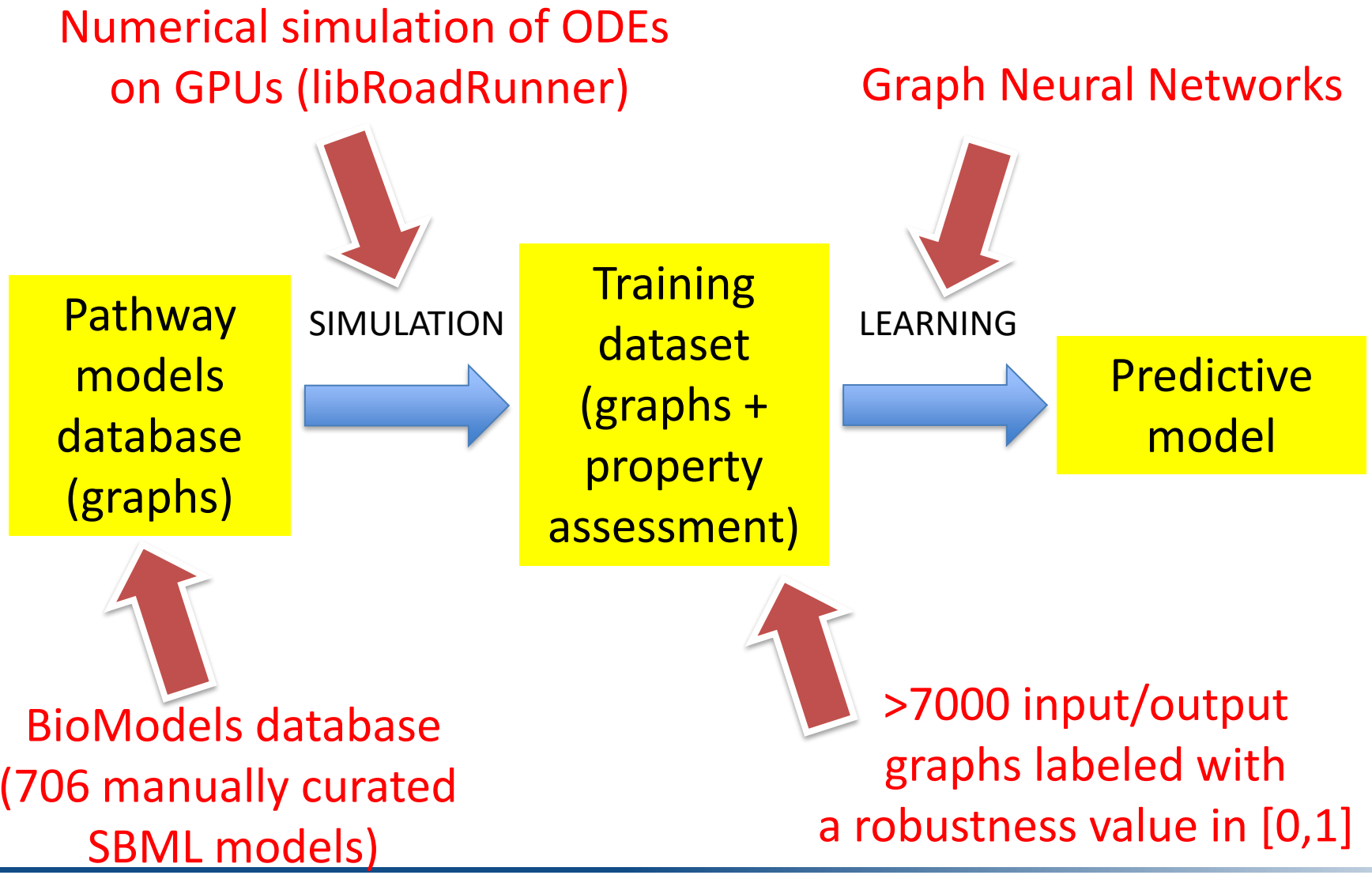
Methodology



Methodology



Methodology



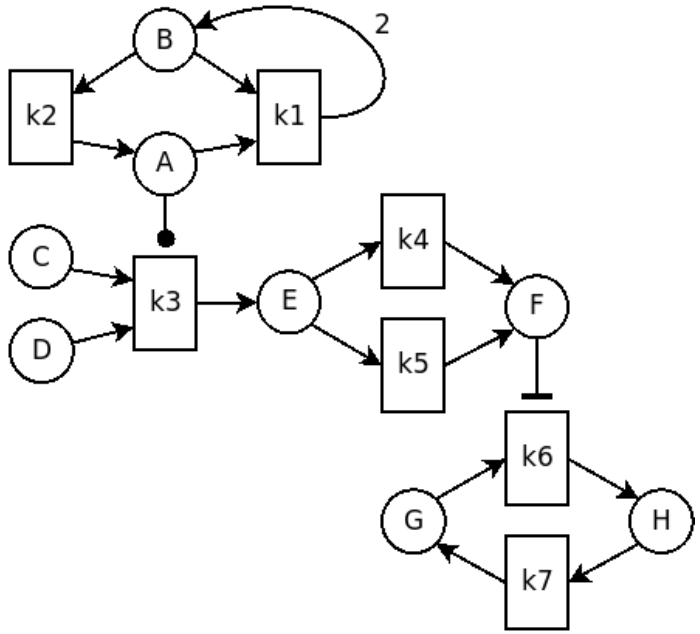
Construction of the dataset: more details

- BioModels database of pathways in SBML format:

https://www.ebi.ac.uk/biomodels-main/

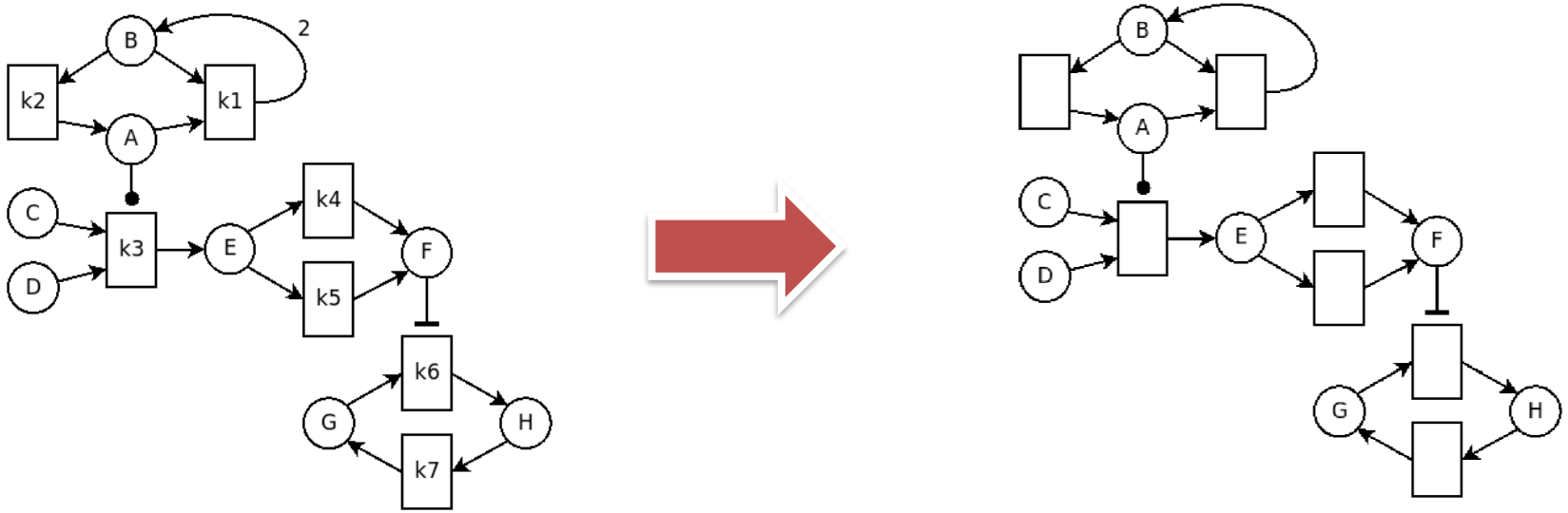
```
...
<reaction id='r1'>
    <listOfReactants>
       ...
    </listOfReactants>
    <listOfProducts>
       ...
    </listOfProducts>
    <listOfModifiers>
       ...
    </listOfModifiers>
</reaction>
...
```

Reaction	*Mod*	*Kinetics*
$A + B \rightarrow 2B$		k_1AB
$B \rightarrow A$		k_2B
$C + D \rightarrow E$	A	k_3CDA
$E \rightarrow F$		k_4E
$F \rightarrow E$		k_5F
$G \rightarrow H$	F	$\frac{k_6G}{1+2F}$
$H \rightarrow G$		k_7G



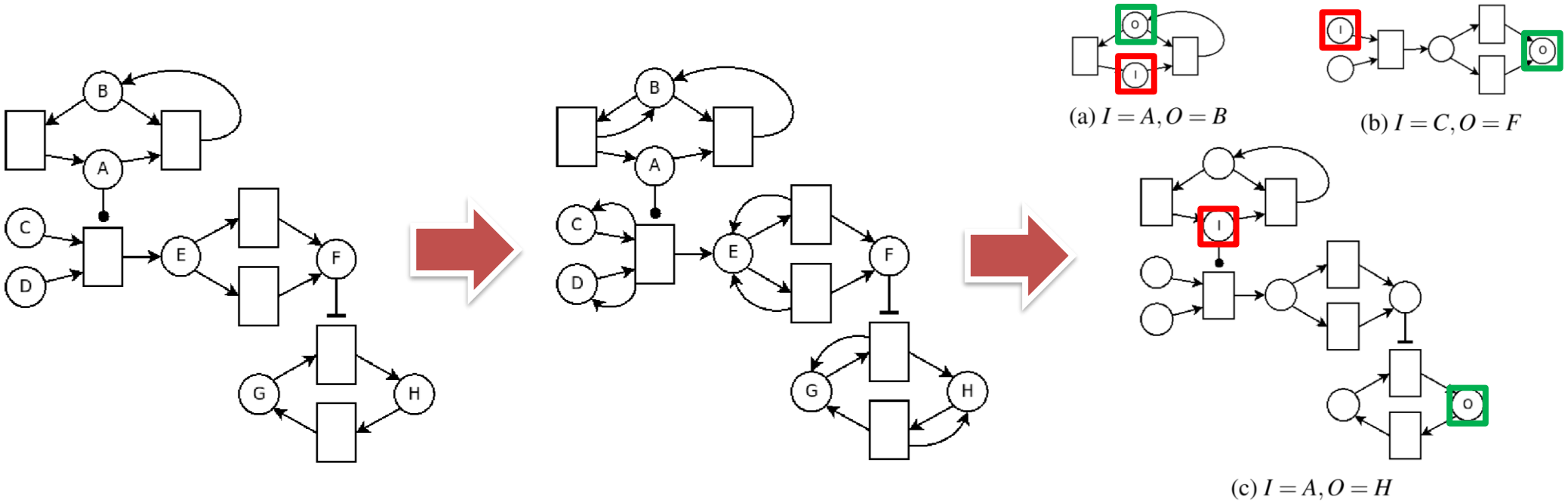
Construction of the dataset: more details

- Graph preprocessing
  1. Removal of quantitave information (focus on topology)



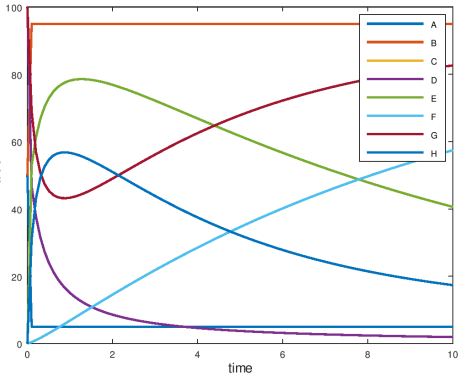
Construction of the dataset: more details

- Graph preprocessing
  1. Removal of quantitave information (focus on topology)
  2. Extraction of input/output induced subtasks



Construction of the dataset: more details

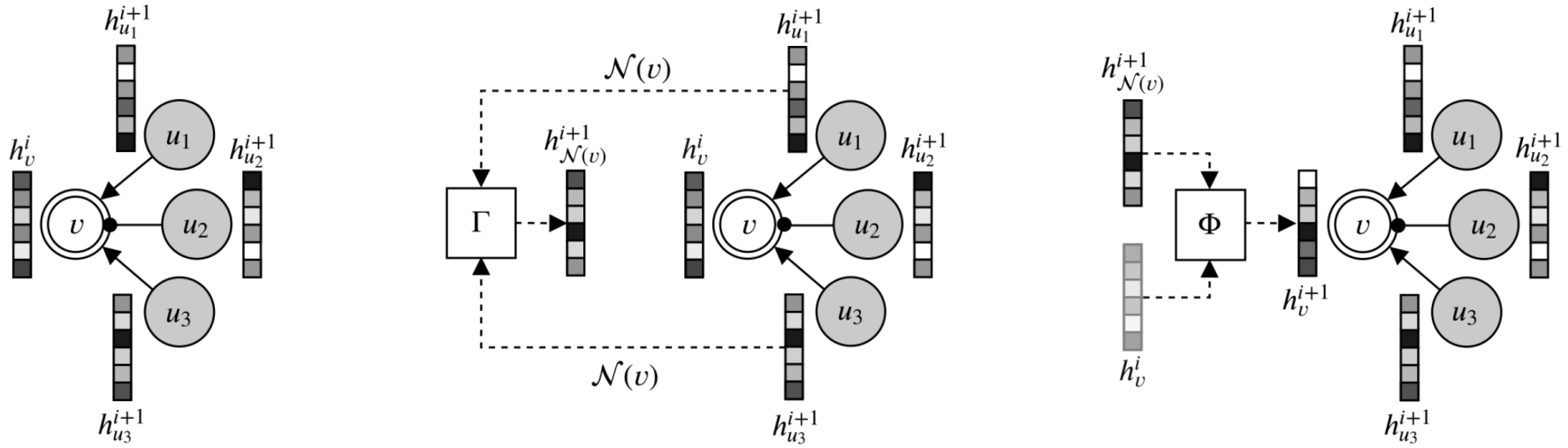
- The dataset consists of >7000 induced subgraphs
  - Obtained from the 706 complete graphs
  - Up to 40 nodes
- Each subgraph is associated to a robustness classification label (1 if robustness > 0.5 -- 0 otherwise)
  - Obtained by performing extensive simulations of the 706 graphs
  - Initial concentration of each (input) molecule perturbed in the interval [-20%,+20%]
  - Simulations gave the interval of (output) steady state concentrations for the computation of robustness



Machine Learning: more details

- Machine Learning on graphs:
  - Traditional ML modelling assumes continuous fixed-size vectors as input data
  - Graphs are discrete variable-size objects
- There is no a universally effective way of mapping graphs into fixed-size vectors
- Graph Neural Networks (GNNs) are able to learn meaningful graph-to-vector mappings adaptively from data

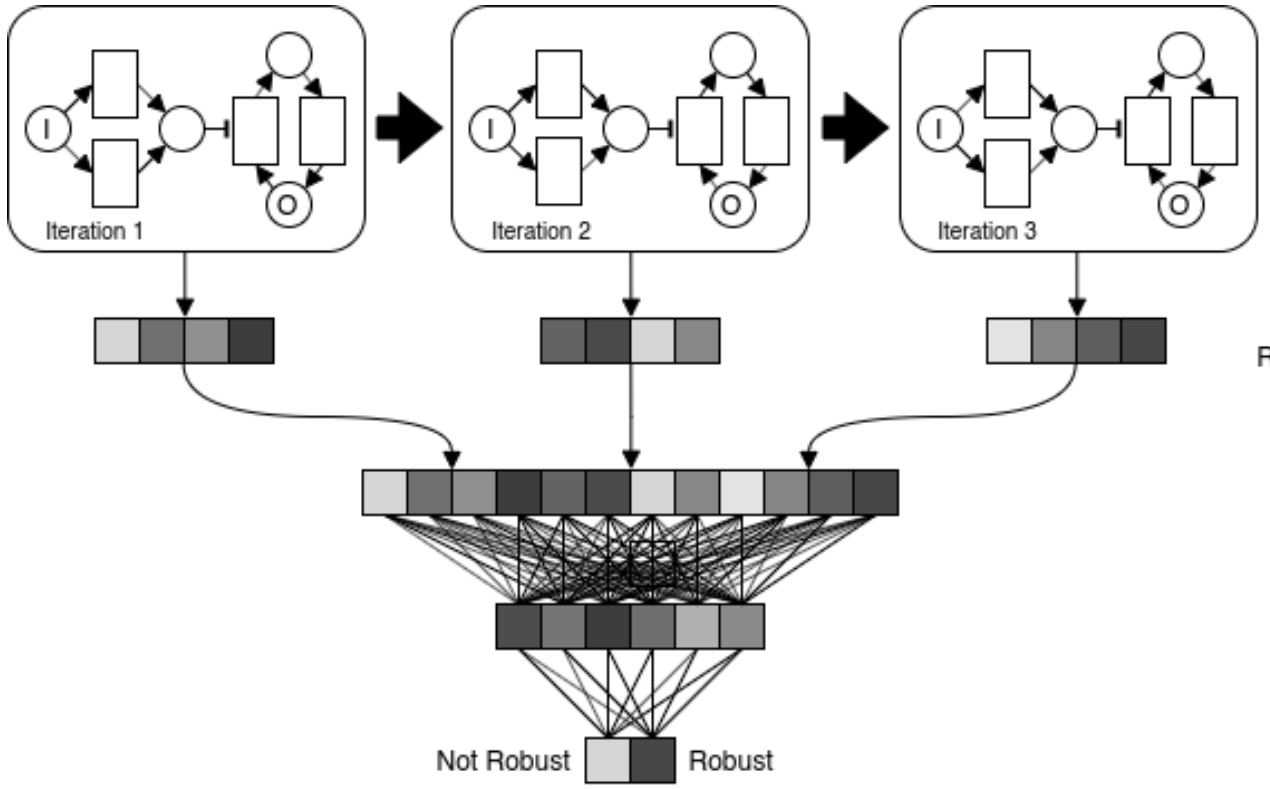
Machine Learning: more details



(a) Initial graph (b) Neighbor aggregation (c) State update

- GNNs are based on node embedding and neighborhood aggregation
- Iterative process: at the k-th step each node receive information from nodes at distance k (layering)

Machine Learning: more details

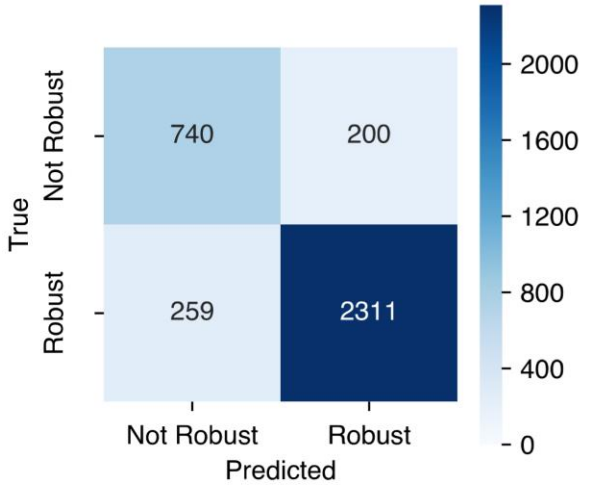


- Node embeddings are then aggregated to get graph embeddings (one for each layer)
- Graph embeddings are concatenated into a single fixed-size vector suitable for multilayer perceptron classification

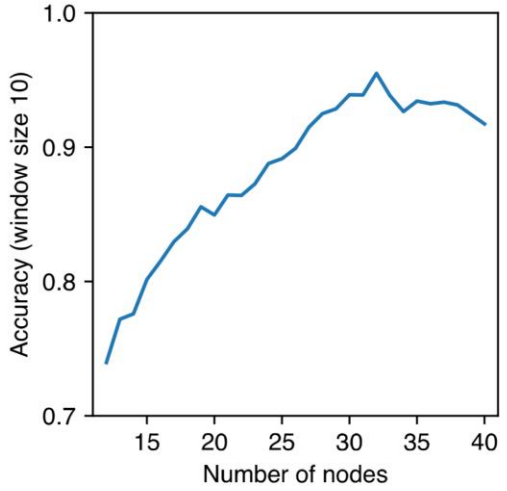
Results: accuracy

# Nodes (# subgraphs)	Model	Null
1-10 (685)	**0.7456** $\pm$ 0.0455	0.6385 $\pm$ 0.0213
11-20 (2679)	**0.8625** $\pm$ 0.0091	0.7327 $\pm$ 0.0207
21-30 (1690)	**0.9350** $\pm$ 0.0136	0.8384 $\pm$ 0.0267
31-40 (1959)	**0.8645** $\pm$ 0.0303	0.6683 $\pm$ 0.0105
Overall (7013)	**0.8692** $\pm$ 0.0140	0.7322 $\pm$ 0.0000

(a) Overall and stratified accuracies.



(b) Confusion Matrix.



(c) Accuracy improvement plot.

- Dataset slightly imbalanced in favor of robustness
- Better accuracy compared to Null model (always says “Robust”)
- Accuracy increases with number of nodes

Conclusions

- Our experiments suggest that it is possibile to learn something about dynamical properties of pathways by looking only at their structure/topology
- The approach works better for bigger (sub)graphs
  - In small graphs quantitative parameters are more relevant
  - In big graphs it is the structure that matters
- Next steps:
  - Try to recover quantitative parameters, properly normalized/generalized
  - Apply to other dynamical properties
  - Explainability: evaluate the contribution of each edge by performing selective «knock-outs» of edges