SAPIENZA
UNIVERSITÀ DI ROMA

AI-driven analysis of molecular pathways

Facoltà di Ingegneria dell'informazione, informatica e statistica
Informatica

Ionuț Cicio
ID number 2048752

Advisor
Prof. Mancini

Co-Advisor
Prof. Tronci

Academic Year 2024/2025

AI-driven analysis of molecular pathways
Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: cicio.2048752@studenti.uniroma1.it

Abstract

In this report it's presented a tool that uses molecular pathway databases to generate biochemical networks by computing the backwards-reachable reactions from a set of species of interest. The biochemical network is then extended with a system of ODE which describes its dynamics by using mass action kinetics.

Contents

1 Introduction **1**
1.1 Preliminaries . . . 2
1.2 Biochemical networks . . . 2
1.3 Dynamics of biochemical networks . . . 4
1.3.1 The Law of Mass Action . . . 5
1.3.2 Hill equation . . . 6

2 Biological models generation **8**
2.1 Reactome and reachability . . . 8
2.2 Scenarios . . . 8
2.3 Constraint satisfiability problem . . . 9
2.4 Optimization problem . . . 9

3 Blackbox optimization architecture **10**
3.1 Blackbox optimization . . . 10
3.2 OpenBox . . . 10
3.3 Bayesian optimization . . . 10
3.4 Orchestrator-worker infrastructure . . . 10
3.5 Scalability analysis . . . 10

4 Experiments **11**

5 Appendix **12**
.1 Neo4j Graph databse . . . 12
.2 Reactome . . . 12
.3 SBML simulation with Roadrunner . . . 14
.4 HPC Cluster emulation with Docker Compose . . . 14

Chapter 1

Introduction

Biological systems such as organisms, cells, or biomolecules are highly organized in their structure and function, [5]. Such structures can help study the and compare biological functions of different organisms. In order to understand, formalize, abstract and make predictions on these structures a mathematical model of these structures is needed.

Given a set of biochemical species and metabolic pathways of interest, this report presents a tool which generates biological models by using the Reactome pathway database [3] to compute the biochemical network of all the backwards-reachable species and reactions starting from the species of interest and within the metabolic pathways of interest.

The generated biochemical networks are qualitative models and contain information about the backwards-reachable reactions (and reaction-like events), their products, reactants (with the stoichiometries), modifiers and compartments. In order to study the dynamic behaviour of these biochemical networks a system of ODE is produced by using the Law of Mass Action to describe the rates of the reactions within the network, and new reactions are added to produce and consume the species on the boundary of the network.

The product is a SBML (Systems Biology Markup Language) [1] file with the dynamics of the biochemical network and a set of decision variables that determine the speeds of the reactions within the network. This model is then subject to a study of stability by defining a set of constraints for stability and using the OpenBox BBO (blackbox optimization) solver [7] to find *virtual patients* (assignements to the decision variables) which produce stable states for the model.

Validating the stability constraints for a *virtual* patient requries simulating the biological model, which is a very computationally expensive task when done multiple times; in order to reduce steady states search times an orchestrator-worker architecture is introduced by using the OpenBox as a Service [9] feature.

1.1 Preliminaries

The set of non-zero natural numbers is denoted with $\mathbb{N}_+ = \mathbb{N} - \{0\}$.

This report uses the terms *species* and *physical entity* interchangibly to refer to biochemical species, since Reactome [10] uses the term *PhysicalEntity* and SBML [4] uses the term *species*.

Reactome uses the term *ReactionLikeEvent* to refer to reactions [10], since it encompasses different types of events other than simple chemical reactions (such as translocation of molecules from one compartment to another, association of molecules to form complexes and polymers, dissociation of complexes and polymers etc...).

1.2 Biochemical networks

Biochemical networks are one of the main concepts used to model biological systems [6]. Such networks are made up of relations between physical entities (small molecules, proteins, nucleic acids, chemical compounds, complexes, larger macromolecular assemblies, atoms, electrons, and photons [10]) and reactions (such as standard chemical reactions, translocation of molecules from one compartment to another, association of molecules to form complexes and polymers, etc... [10]). More often than not biochemical networks are organized in pathways, sets of causally connected reactions [10]; such pathways, in pathway databases like Reactome (Figure .1) are organized hierarchically: Reactome has top-level pathways, like *Signal transduction* (reactions where extracellular signals elicit changes in cell state and activity), *Muscle contraction* (Figure .2) and the *Neuronal system*, each composed of other smaller pathways or reactions.

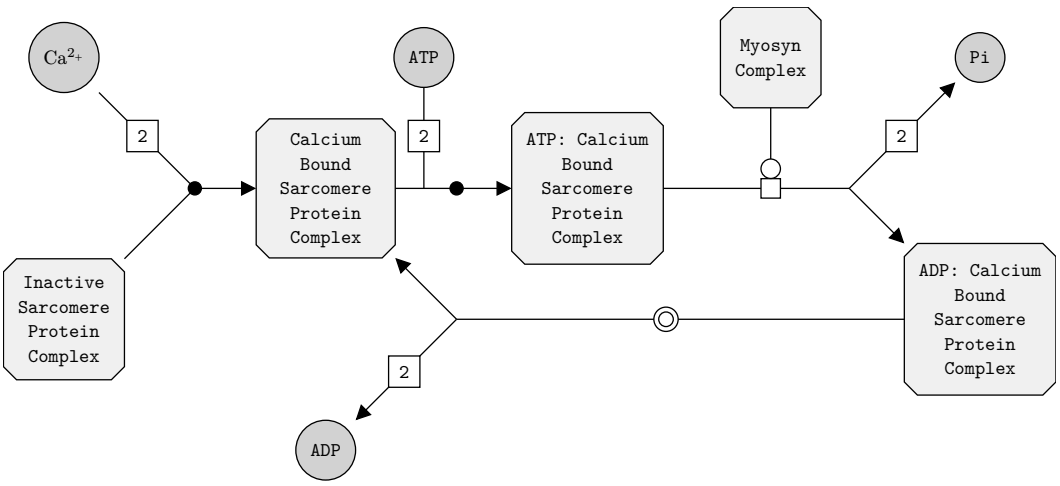


Figure 1.1. *Striated muscle contraction* pathway within the *Muscle contraction* pathway in the Reactome PathwayBrowser [8]

The biochemical network in Figure 1.1 represents the *Striated muscle contraction* pathway in the *Cytosol* compartment, a process where force is generated within striated muscle tissue, resulting in muscle movement [8]. Here the pathway has the following reactions:

1. Ca^{2+} (stoichiometry 2) binds with an *Inactive Sarcomere Protein Complex*
2. ATP (stoichiometry 2) associates with the bound "Ca^{2+} - *Calcium Sarcomere Protein Complex*"
3. The *Myosyn Complex* acts as a catalyst in order to activate the reaction that produces *Inorganic Phosphate* and the "ADP - *Calcium Sarcomere Protein Complex*" bound
4. The ADP dissociates from the "Ca^{2+} - *Sarcomere Protein Complex*" bound, producing two units of ADP (stoichiometry 2)

In particular, Ca^{2+}, ATP, *Inactive Sarcomere Protein Complex* and *Myosyn Complex* are inputs of the biochemical network (no reaction within the network produces these species), while Pi and ADP are outputs (no reaction within the network consumes these species).

It's important to note that since the *Myosyn Complex* is a catalyst (modifier) of the reaction, it is not consumed in the reaction.

Biochemical networks as presented in pathway databases are qualitative models, they describe only the roles of species in reactions and the structure of the network. Other than the stoichiometries there are no informations about typical quantities of species, reactions speeds, compartments sizes etc...

Definition 1 (Biochemical network). A biochemical network G is a tuple (S, R, E, σ) where

- $S = U \cup X \cup Y$ is the set of species of the biochemical network
  - U is the set of input species of the network
  - X is the set of species which are neither inputs or outputs
  - Y is the set of output species of the network
- R is the set of reactions in the biochemical network
- $E = (E_{\text{reactants}}, E_{\text{products}}, E^{+}_{\text{modifiers}}, E^{-}_{\text{modifiers}})$ is a tuple with the sets of relations between species and reactions with $E_i \subseteq S \times R$
- $\sigma : E_{\text{reactants}} \cup E_{\text{products}} \to \mathbb{N}_{+}$ is the stoichiometry of the reactants of products of the reaction

The definition abstracts different reaction types. It also allows for a species to be have multiple roles in a reaction, since in the Reactome pathways database there are some reactions in which, for example, modifiers are also inputs or outputs of the reaction. (TODO: reference cypher query or species)

1.3 Dynamics of biochemical networks

Qualitative models, such as the aforementioned biochemical networks, are not enough to study the behaviour of a pathway, a description of how the network evolves over time when put in a stable environment is needed in order to make predictions and study what happens when the model is subject to perturbations. Systems of ordinary differential equations (ODE) are one of the approaches used in order to describe the dynamic behaviour of biochemical networks [5].

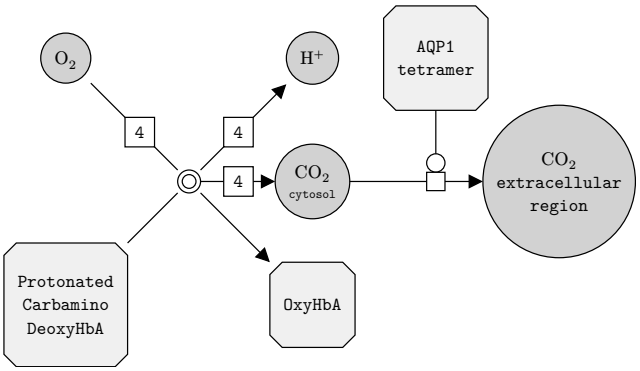


Figure 1.2. *Hemoglobin A* binds O_2, releasing H^+ and CO_2, then AQP1 acts as a catalyst and transports CO_2 from the cytosol to the extracellular region (reaction in the *O_2/CO_2 exchange in erythrocytes* pathway [11])

Let $G = (S, R, E, \sigma)$ be the biochemical network in Figure 1.2

- $S = U \cup X \cup Y$ where
  - $U = \{O_2, Hb_1, A\}$ with Hb_1 and A being respectively *Hemoglobin A* before the dissociation and AQP1
  - $X = \{CO_2^c\}$, with CO_2^c being CO_2 in the cytosol
  - $Y = \{Hb_2, H^+, CO_2^e\}$ with Hb_2 and CO_2^e being respectively *Hemoglobin A* after the dissociation and CO_2 in the extracellular region
- $R = \{r_1, r_2\}$ where r_1 and r_2 are respectively the dissociation reaction and the transport reaction

One possible way to describe the dynamics of G is through the system of ordinary differential equations described in Equations 1.1, 1.2 and 1.3, where the rate of change of a given species $S_i \in S$ depends on the rate of change of the reactions that produce and consume it.

$$\begin{gathered}\frac{dO_2}{dt} = -r_1, \quad \frac{dHb_1}{dt} = -r_1, \quad \frac{dHb_2}{dt} = +r_1, \quad \frac{dH^+}{dt} = +r_1 \\ \frac{dCO_2^c}{dt} = +r_1 - r_2, \quad \frac{dCO_2^e}{dt} = -r_1 + r_2, \quad \frac{dA}{dt} = 0\end{gathered} \tag{1.1}$$

$$r_1 = k_1 \cdot (O_2)^4 \cdot (Hb_1)^1 \tag{1.2}$$

$$r_2 = k_2 \cdot (\mathrm{CO}_2^c)^4 \cdot \frac{A^{n_1}}{K_1 + A^{n_1}} \tag{1.3}$$

The reaction rate of r_1 and r_2 is given by mass action kinetics, where the speed of reaction $r_i \in R$ is determined by the kinetic constant k_i and is proportional to the concentrations of the reactants and modifiers. In particular, since AQP1 acts as a catalyst in Equation 1.3, reaction r_2 can be activated only if A reaches a certain concentration which depends on the constants K_1 and n_1 in the hill function.

1.3.1 The Law of Mass Action

The kinetics in Equantions 1.2 and 1.3 are based on the mass action law, introduced by Guldberg and Waage in 1864 [2]. It states that the reaction rate is proportional to the probability of a collision of the reactants, probability which is in turn proportional to the concentration of reactants to the power of the molecularity (stoichiometry), which is the number in which the molecule species enter the reaction (this is because the higher is the number of molecule species need for the reaction to happen, the lower is the probability the reaction will happen) [5]. Given a biochemical network $G = (S, R, E, \sigma)$, and a reaction $r \in R$, where r is a **reversible** reaction, the rate of the reaction r in its most general form is written as

$$r = k_r^+ \cdot \prod_{\substack{(s,r) \\ \in \\ E_{\text{reactants}}}} s^{\sigma(s,r)} - k_r^- \cdot \prod_{\substack{(s,r) \\ \in \\ E_{\text{products}}}} s^{\sigma(s,r)} \tag{1.4}$$

where

- s indicates the concentration of the species s
- k_r^+ is the kinetic constant of the forward reaction
- k_r^- is the kinetic constant of the backward reaction

The form used for the generation of biological models in this report is the simpler one in Equation 1.5. This is because Reactome models reversible reactions as two different reactions, so it's enough to include both the forward and the backward reaction in order to obtain the equation for a single reversible reaction.

$$r = k_r \cdot \prod_{\substack{(s,r) \\ \in \\ E_{\text{reactants}}}} s^{\sigma(s,r)} \tag{1.5}$$

The law of mass action has the advantage that a single decision variable is associated to each reaction in order to determine the speed of the reaction. This is a gross semplification since the behaviour of reactions is more complex than this, but more detailed kinetics such as convenience kinetics introduce many more decision variables, making the search of *virtual patients* for the model harder.

1.3.2 Hill equation

In nature the speed of reactions is heavily influenced by modifiers (enzymes which speed-up the reaction and inhibitors which slow down or stop the reaction). The characteristic of modifiers is that they are neither consumed or produced by the reaction. Given a biochemical network $G = (S, R, E, \sigma)$, and a pair $(s, r) \in E^+_{\text{modifiers}} \cup E^-_{\text{modifiers}}$ (meaning that the species s is a modifier for reaction r, either as a catalyzer or an inhibitor), the hill function can be defined as

$$H^s_r = \begin{cases} \frac{(s)^{n^s_r}}{K^s_r + (s)^{n^s_r}}, & \text{if } (s, r) \in E^+_{\text{modifiers}} \\ \frac{K^s_r}{K^s_r + (s)^{n^s_r}}, & \text{if } (s, r) \in E^-_{\text{modifiers}} \end{cases} \tag{1.6}$$

where

- K^s_r is the apparent dissociation constant of modifier s in reaction r
- n^s_r the hill coefficient of modifier s (often in literature it's indicated with h)

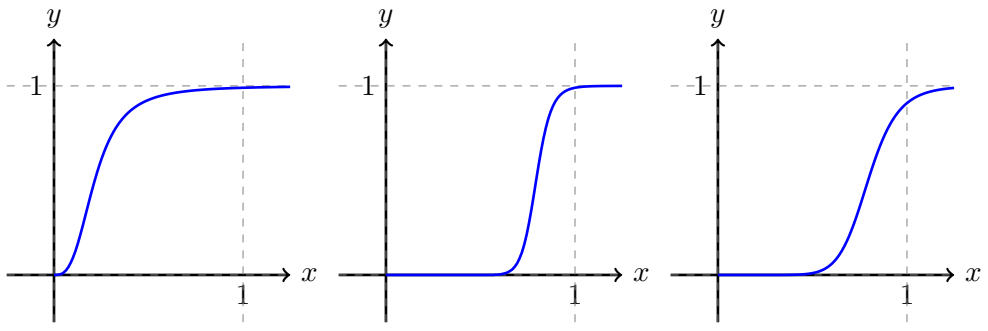


Figure 1.3. hill function for a catalyst with parameters K^s_r, n^s_r respectively $(0.01, 3), (0.01, 10), (0.1, 10)$

With this information the kinetics of Equation 1.3, which describes the transport of CO_2 from the cytosol to the extracellular region, can be explained like this:

- k_1 is the kinetic constant of the forward reaction
- $(O_2)^4$ is the concentration of O_2, raised to the power of 4 (the stoichiometry), given by the law of mass action
- A acts as a catalyst for the transportation, so a hill function is used to moderate the speed of the reaction based on the concentration of A

In order to define the dynamics of a given biochemical network, the network needs to be expanded to handle boundary species. In particular, for all species that that are only inputs of the network a reaction needs to be added in order to constantly produce the species (as the network needs to be simulated in a stable environment). Contrarywise, all species which are only outputs of the network need to be consumed. Some species are both inputs and outputs of the network, these species are modifiers (otherwise they would either be consumed in a reaction or produced by a reaction), and can be set to a fixed concentration.

Definition 2 (Biological model). Given a biochemical network $G = (S, R, E, \sigma)$ let $B = (G', \mathcal{K})$ be the biological model derived from G with added mass action kinetics, with $G' = (S, R', E', s')$ where

- $R' = R \cup R_{\text{inputs}} \cup R_{\text{outputs}}$ where
  - $R_{\text{inputs}} = \{r_s \mid s \in U - Y\}$
  - $R_{\text{outputs}} = \{r_s \mid s \in Y - U\}$
- $E' = (E'_{\text{reactants}}, E'_{\text{products}}, E^+_{\text{modifiers}}, E^-_{\text{modifiers}})$, where
  - $E'_{\text{reactants}} = E_{\text{reactants}} \cup \{(s, r) \mid r_s \in R_{\text{inputs}}\}$
  - $E'_{\text{products}} = E_{\text{products}} \cup \{(s, r) \mid r_s \in R_{\text{outputs}}\}$

(1.7)

Then $\mathcal{K}$ can be defined on G'

$$\begin{aligned}\mathcal{K} =& \{k_r \mid r \in R'\} \cup \\ & \{K_r^s \mid r \in R' \wedge (s, r) \in E^+_{\text{modifiers}} \cup E^-_{\text{modifiers}}\} \cup \\ & \{n_r^s \mid r \in R' \wedge (s, r) \in E^+_{\text{modifiers}} \cup E^-_{\text{modifiers}}\} \cup\end{aligned} \tag{1.8}$$

where

- k_r is the kinetic constant of reaction r
- K_r^s is the apparent dissociation constant of modifier s in reaction r
- n_r^s the hill coefficient of modifier s in reaction r

Chapter 2

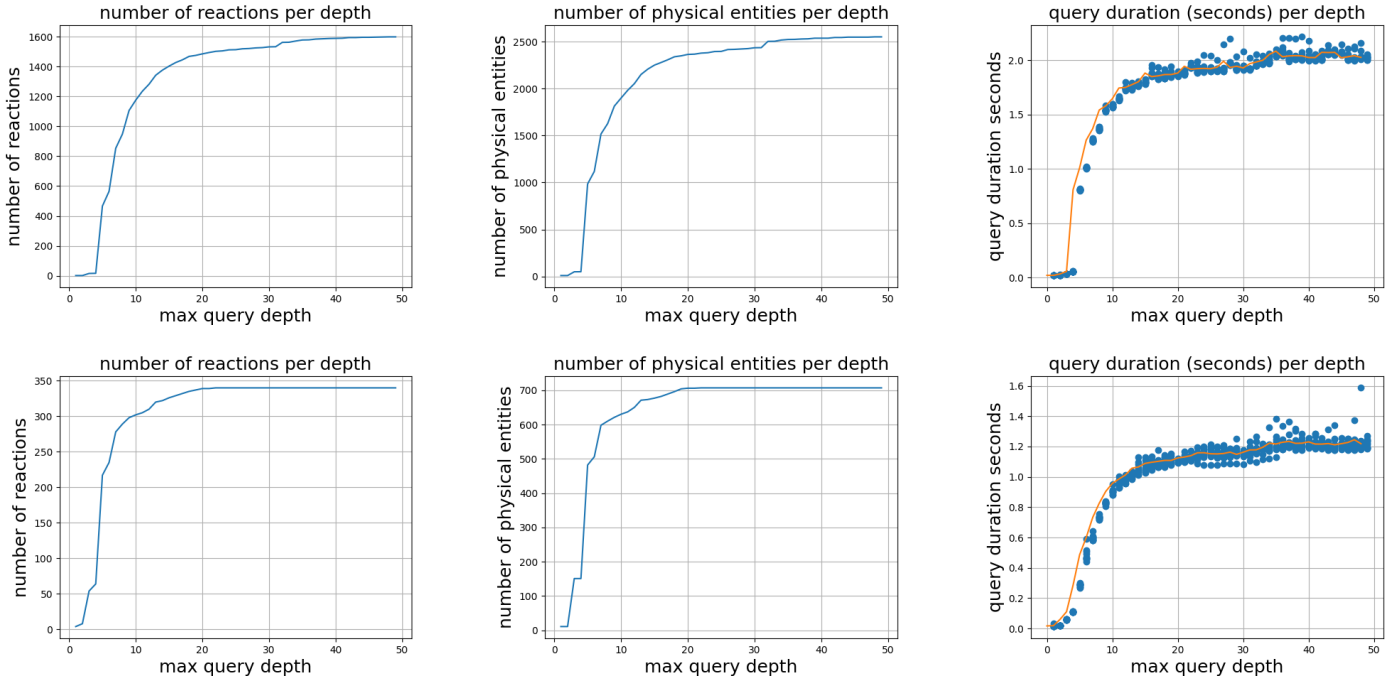
Biological models generation

2.1 Reactome and reachability

REACTOME is an open-source, open access, manually curated and peer-reviewed pathway database. (TODO: cite Reactome)

Given a biochemical species S one possibile way to

2.2 Scenarios



2.3 Constraint satisfiability problem

Definition 3 (Biological model satisfiability problem). Given a biological model $B = (G, \mathcal{K})$ let $\mathcal{S}$ be the set of concentrations of species of the species in the network, $\mathcal{S} = \{s \mid s \in S\}$ and $\mathcal{S}_{\text{avg}} = \{s_{\text{avg}} \mid s \in S\}$ the set of average concentrations of the species, $T \in \mathbb{R}$ the time horizon, the following constraints must hold:

$$\forall s \in \mathcal{S} \quad 0 \leq s \leq 1 \tag{2.1}$$

$$\begin{aligned} &\forall k_{r_1}, k_{r_2} \in \mathcal{K},\ s \in S \\ &\quad (s, r_1) \in E_{\text{products}} \wedge (s, r_2) \in E_{\text{modifiers}} \wedge r_1 \neq r_1 \Rightarrow k_{r_1} < k_{r_2} \end{aligned} \tag{2.2}$$

$$\forall s \in \mathcal{S}_{\text{avg}} \quad s(\phi \cdot T) - s(T) = 0 \tag{2.3}$$

2.4 Optimization problem

Chapter 3

Blackbox optimization architecture

3.1 Blackbox optimization

Given a function $f : X \to Y$, which is expensive to compute, and an optimization problem of the type $\text{argmin}_{x \in X} f(x)$ is a blackbox optimization problem if no information about the derivative of f

3.2 OpenBox

OpenBox is an efficient open-source system designed for solving gener alized black-box optimization (BBO) problems. It can be used either as a Standalone python package or Online BBO service [7].

3.3 Bayesian optimization

3.4 Orchestrator-worker infrastructure

OpenBox is an efficient open-source system designed for solving generalized black-box optimization (BBO) problems. It can be used either as a Standalone python package or Online BBO service @open-box.

OpenBox has a great support for bayesian optimization, so that will be the main subject of the analysis @open-box-automatic-algorithm-selection.

3.5 Scalability analysis

Chapter 4

Experiments

Chapter 5

Appendix

.1 Neo4j Graph databse

Neo4j is a native graph database, which means that it implements a true graph model all the way down to the storage level. Instead of using a "graph abstraction" on top of another technology, the data is stored in Neo4j in the same way you may whiteboard your ideas.

https://neo4j.com/docs/getting-started/whats-neo4j/

.2 Reactome

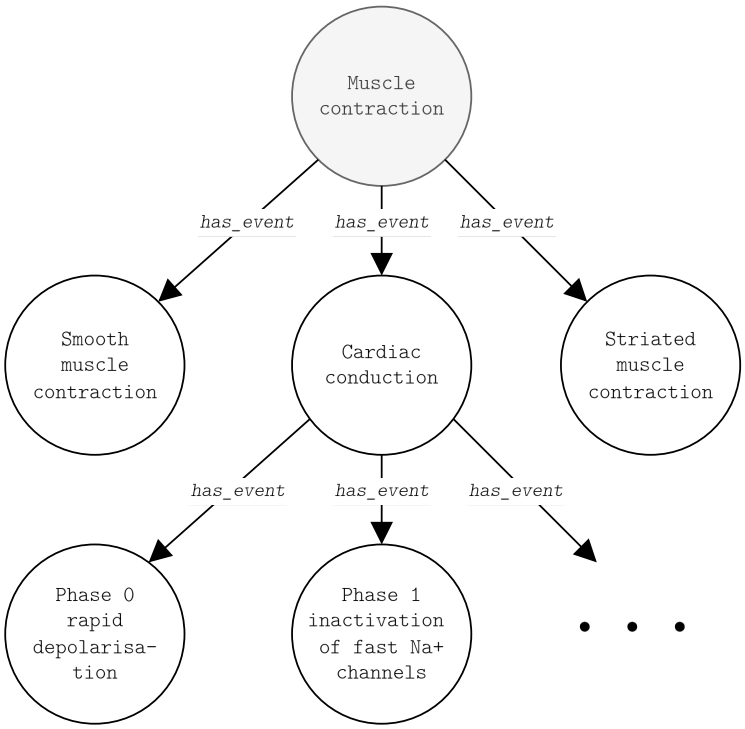


Figure .2. Muscle contraction pathway hierarchy [8]

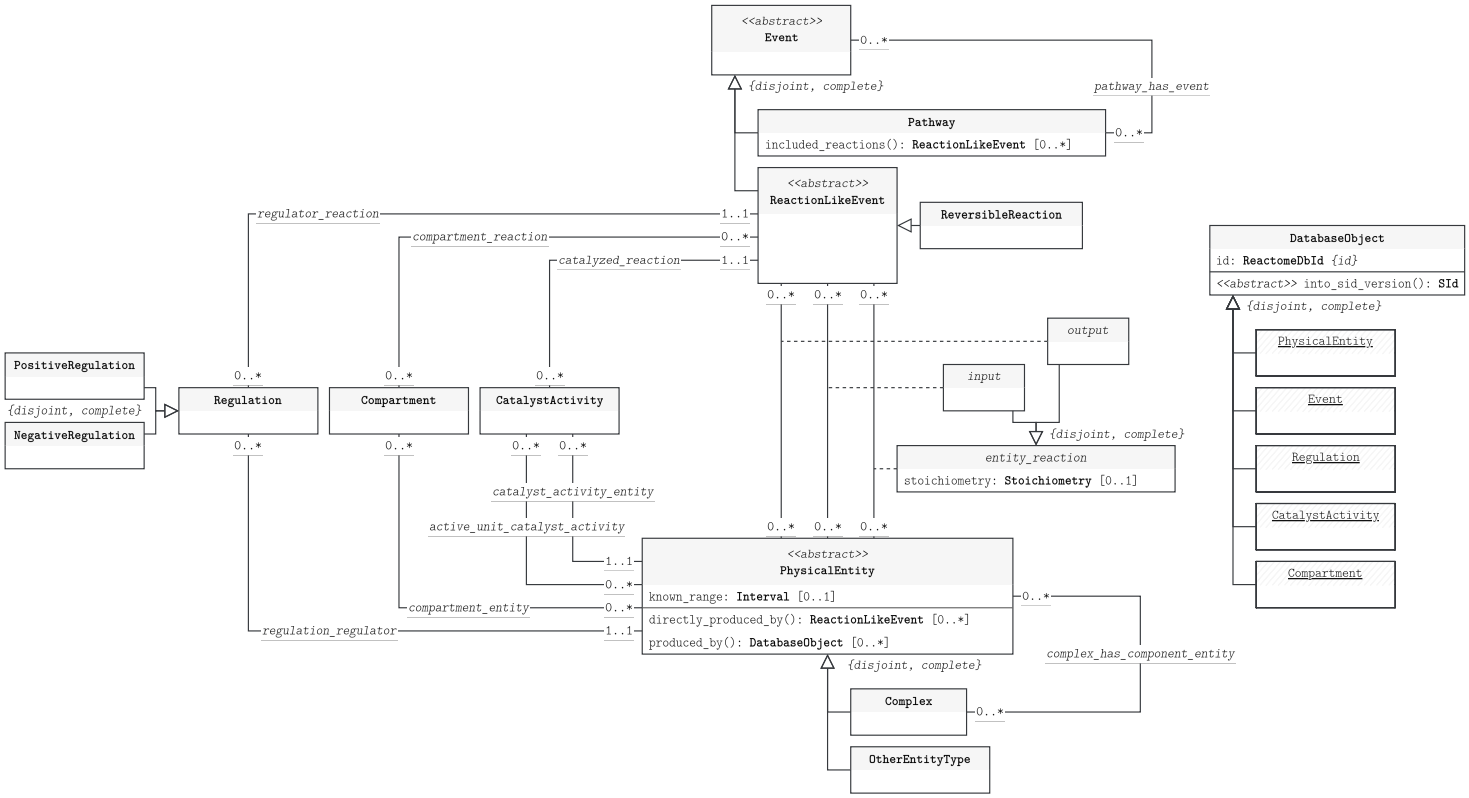
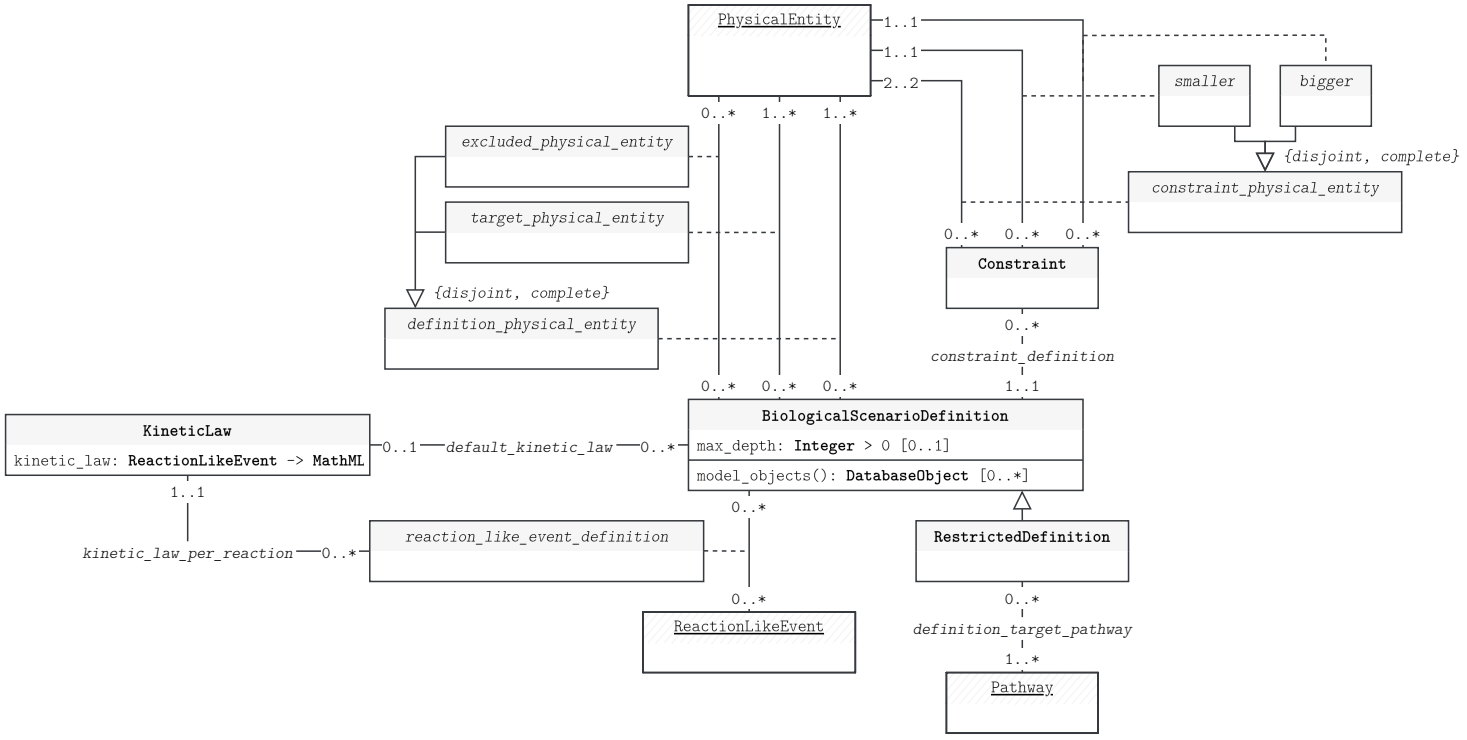


Figure .1. UML model of a portion of the Reactome database



```
MATCH
path =
      (reactionLikeEvent:ReactionLikeEvent)-[:catalystActivity]->
      (:CatalystActivity)-[:physicalEntity]->
      (physicalEntity:PhysicalEntity)
WHERE
EXISTS {
     MATCH (reactionLikeEvent)-[:input]->(physicalEntity)
}
RETURN COUNT(DISTINCT path)
```

.3 SBML simulation with Roadrunner

.4 HPC Cluster emulation with Docker Compose

Bibliography

[1] URL: https://raw.githubusercontent.com/combine-org/combine-specifications/main/specifications/files/sbml.level-3.version-2.core.release-2.pdf.

[2] URL: https://www3.nd.edu/~powers/ame.50531/guldberg.waage.1864.pdf.

[3] Antonio Fabregat et al. "Reactome graph database: Efficient access to complex pathway data". In: *PLoS Computational Biology* 14.1 (2018), e1005968. DOI: 10.1371/journal.pcbi.1005968.

[4] Michael Hucka et al. "The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core Release 2". In: *Journal of Integrative Bioinformatics* 16.2 (2019), p. 20190021. ISSN: 1613-4516. DOI: 10.1515/jib-2019-0021. URL: https://www.degruyter.com/view/j/jib.ahead-of-print/jib-2019-0021/jib-2019-0021.xml.

[5] Edda Klipp et al. *Systems Biology: A Textbook*. 2nd. Wiley-Blackwell, 2016, p. 504. ISBN: 9783527336364.

[6] M Koutrouli et al. "Erratum: A Guide to Conquer the Biological Network Era Using Graph Theory". In: *Frontiers in Bioengineering and Biotechnology* 11 (Mar. 2023), p. 1182500. DOI: 10.3389/fbioe.2023.1182500. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7004966/.

[7] Yang Li et al. "Openbox: A generalized black-box optimization service". In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2021, pp. 3209–3219.

[8] Rush MG Marc E. Gillespie. *Striated Muscle Contraction*. Accessed: 2025-11-25. 2025. URL: https://reactome.org/PathwayBrowser/#/R-HSA-390522&PATH=R-HSA-397014.

[9] *OpenBox Documentation – Service Introduction*. URL: https://open-box.readthedocs.io/en/latest/openbox_as_service/service_introduction.html (visited on 11/26/2025).

[10] Reactome. *Data Model Glossary*. 2025. URL: https://download.reactome.org/documentation/DataModelGlossary_V90.pdf.

[11] *Reactome Pathway Browser: R-HSA-1247673, R-HSA-1247668, R-HSA-382551, R-HSA-1480926*. URL: https://reactome.org/PathwayBrowser/#/R-HSA-1247673&SEL=R-HSA-1247668&PATH=R-HSA-382551,R-HSA-1480926.