Scientific Research Publishing

Natural Resources, 2025, 16(13), 759-775
https://www.scirp.org/journal/nr
ISSN Online: 2158-7086
ISSN Print: 2158-706X

Relational Learning in Microbial Ecology

Andrei Doncescu

Department of Mathematic and Informatic, University of French West Indies, Pointe-à-Pitre, France
Email: andrei.doncescu@univ-antilles.fr

How to cite this paper: Doncescu, A. (2025) Relational Learning in Microbial Ecology. *Natural Resources*, **16**, 759-775.
https://doi.org/10.4236/nr.2025.1613038

Received: January 22, 2025
Accepted: December 26, 2025
Published: December 29, 2025

Copyright © 2025 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

This paper introduces a methodology that enables the relational learning framework to incorporate quantitative data derived from experimental studies in microbial ecology. The focus of using Default Logic in microbial ecology is to enhance the comprehension of cellular physiological states and the interpretation of interactions among metabolites and signaling networks. To illustrate our approach, a logical model is proposed as model of the glycolysis and pentose phosphate pathways in *E. coli*. This method constructs a symbolic model based on kinetics, utilizing the Michaelis-Menten equation, by discretizing the concentration variations of specific metabolites over time based on relevant levels to be integrated into our Logic Inference framework. Additionally, we generate logical formulas for concentrations of metabolites that are difficult to measure during dynamic states through logical abduction. Given the resulting large set of conclusions/extensions, we employ an expectation maximization algorithm operating on binary decision diagrams for ranking.

Keywords

Logical Model, *E coli*, Default Logic, Binary Decision Diagram, Expectation Maximization, Simulation

1. Introduction

Nowadays, systems ecology represents the key field to explain the functionality of life science. Analyzing a system ecology requires the development of mathematical models that can describe the system's evolution in dynamic contexts or address complex situations where human experience may exceed formal mathematical analysis [1].

The analysis of these systems provides a wealth of data, varied by the diversity of the biological models studied, the conditions of analysis, the levels of observation (macroscopic, microscopic or molecular) and the investigative tools used. However, despite the quantity and quality of the data generated, many questions

remain unanswered, and the discrete nature of the information collected means that it is not yet possible to establish a systemic analysis of biological phenomena [2].

Modern strategies for improving the performance of organisms of industrial interest are based on a global, quantitative approach to understanding the functioning of different levels of biological information integration, from gene flows to metabolic flows (quantitative cell physiology). A recent emergence, metabolomics aims to provide a global analysis of the repercussions on metabolic systems of disturbances of genetic (metabolic engineering) and/or environmental (microbiological engineering) origin. The metabolome represents both the ultimate manifestation of genome expression and the thermodynamic interface with the external environment. The functional analysis of the metabolome of organisms of industrial interest requires the development of two complementary approaches: metabolic profiling, which aims to analyze qualitative/quantitative variations in cellular metabolites, and fluxomic, which aims to analyze metabolic flows within metabolic networks of increasing complexity [3].

In attempts to describe the behavior of living systems, when deductive modeling has failed, the qualitative reasoning approach based on the function of molecules has shown its limits. While we know how to attribute properties to an element of a living system based on its structure, it seems impossible to deduce them either qualitatively or quantitatively. In the same way, although the properties are known, we clearly cannot deduce their function in the living cell, and from the characteristics of living cells, calculate their behavior in each environment. In general, this deductive approach to behavioral description fails because the function of one component of a living system depends on the simultaneous operation of other components. The recurring problem is that, computationally or otherwise, the functional properties of the cell cannot be deduced from the properties of its components alone [4].

In this context, with their ability to describe complexity, mathematical tools offer the prospect of analyzing these structural elements of the living world to propose functionalities that consider their highly non-linear, iterative dynamic interactions. However, this approach cannot free itself from biological reality, in proposing hypothetical model systems, and therefore needs to intimately combine mathematical and biological disciplines.

Various approaches to this description of cell behavior are currently being developed by the international scientific community [5].

The "virtual cell" brings together a range of software programs for describing the metabolic organization of an organism, estimating the distribution of matter in its functioning and extrapolating its production potential. *E-cell* makes it possible to quantify cellular activity based on kinetics that have been perfectly characterized experimentally, both catalytically and quantitatively [2].

Mainly kinetic description obliterates the analysis of real or possible mechanisms that may be involved [6].

Another approach is to seek a precise description of a living cell (or part of it) based on experimentally determined mechanisms and parametric values, *i.e.* data excluding all values estimated on partial models [7].

Over the last decade, methods for analyzing metabolic fluxes using isotopic labelling (stable isotopes such as ^{13}C, ^{15}N) have been widely developed in the context of metabolic engineering. These methods are based on NMR (and, more recently, mass spectrometry) analysis of the isotopic profiles of metabolites accumulating under metabolic and isotopic steady-state conditions. This method results in the creation of flux distribution maps within metabolic networks of increasing complexity [8]. This approach has been widely validated and is an excellent tool for analyzing cell physiology and phenotyping strains. However, it does not allow us to characterize the dynamic behavior of organisms, either to describe their evolution during fermentation or to understand their adaptation in the face of disturbance [9].

Many physical and biological phenomena may be represented in an analytical form using dynamical systems. Our case study is based on wet biology experiment consisting in applying a pulse of glucose in a small bioreactor containing *E. coli* that led to building an ordinary differential equation (ODEs) based simulator [7].

Glucose is a type of sugar that is essential for the metabolism of most living organisms. It is produced by plants, certain bacteria, and protists through the process of photosynthesis. Serving as the primary source of chemical energy, glucose fuels cellular functions in organisms ranging from bacteria and plants to humans [10].

We used high performance liquid chromatography to measure some metabolites concentrations, and some others had to be estimated, using a simulated annealing algorithm, since no experimental results were available [11]. So, knowing the evolution of metabolites concentrations of this system completely, we applied our approach to show its correctness. For that, we took only steady-state values of metabolites concentrations and ran our model.

The objective of this research is to integrate continuous values and kinetics into the logic-based framework for analyzing metabolic pathways [12].

To achieve this, we propose a loop for learning about a metabolic pathway from experimental data (**Figure 1**):

1) Cluster continuous concentrations of metabolites over time into discrete levels and discrete timesteps.

2) Utilize these clusters in a Logic Model of the pathway, alongside a set of knowledge-generating rules, exemplified here with Michaelis-Menten kinetics.

3) Sorting the resulting abduced facts or inducted rules using our defined metrics.

4) Leverage this ranking to enhance our knowledge base and return to the beginning of the process.

This architecture can be utilized to address an inverse problem: given the measured concentrations of certain metabolites in a steady state, we compute the con-

centrations of those metabolites prior to the dynamic transition to that steady state using kinetic modeling. By implementing this automated system, we aim to enhance the efficiency and effectiveness of research experiments.

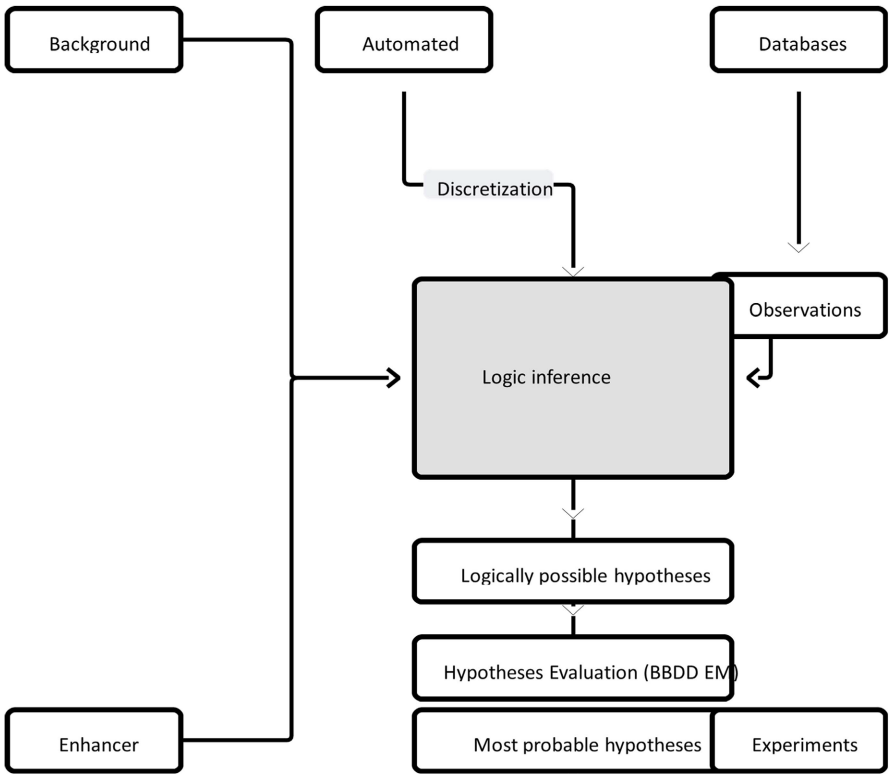


Figure 1. Overview of the complete process.

This framework primarily consists of four tools:

1) The combination of an implementation of continuous HMMs with PY-TSDISC (https://github.com/syhw/py-tsdisc) to discretize experimental values.

2) KEGG2SYMB, using the KEGG API, transforms pathways from KEGG [13] into symbolic models (Figure 2).

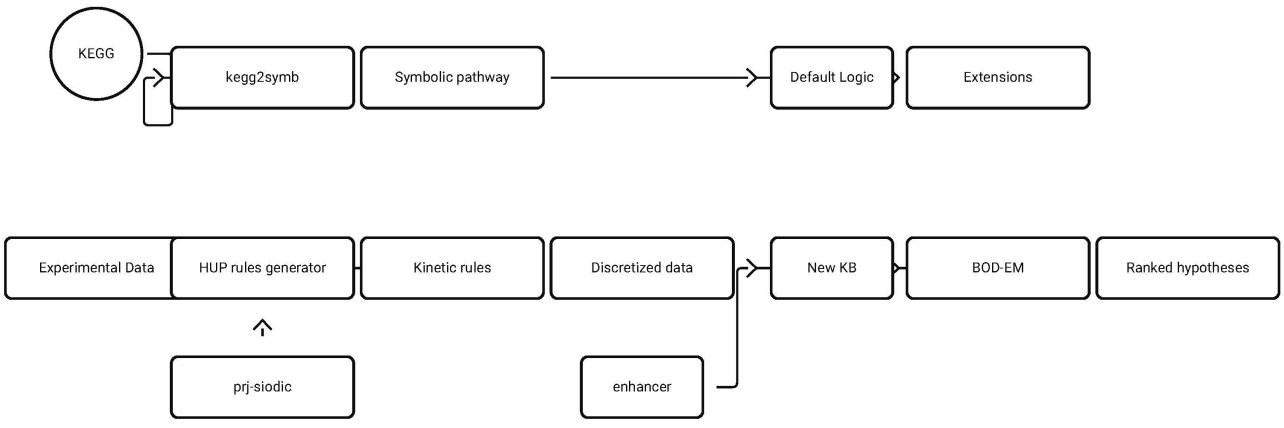


Figure 2. Schema of the process.

3) The Logic Inference is based on Default logic, which is a non-monotonic logic that allows for reasoning with incomplete information. It provides a way to draw conclusions based on certain defaults or assumptions that can be retracted if new information becomes available. In default logic, a set of defaults is defined, and if a default can be applied, it leads to a conclusion unless there is evidence to the contrary. This framework is particularly useful in scenarios where knowledge is uncertain or subject to change, making it a valuable tool in fields such as artificial intelligence and knowledge representation.

4) BDD-EM, an implementation of the expectation-maximization algorithm on binary decision diagrams to rank hypotheses [14].

2. Causal System

Our approach aims to determine (qualitatively and quantitatively) the key points of metabolic adaptation and should ultimately lead to a predictive capability that is particularly sought-after in a metabolic engineering context [15]. The aim is also to reconcile the dynamic data obtained for metabolism with those generated by transcriptome and proteome approaches as part of a systems biology approach [16].

For example, to describe protein/gene interactions in the cell, we start with a classical logic language L (propositional or first-order). In L, the proposition A (resp A) means that A is true (false). For example: *give* (UV), or *screen-glass* -> *give* (UV).

To express interactions (causality) between proteins, we give ourselves two binary relations: *cause* (A, B) and *block* (A, B). Classically, these relations are represented in the metabolite/gene network by A -> B and A -/ B.

If the classical logic inference A -> B, is well described formally, with all the "right" mathematical properties (tautology, non-contradiction, transitivity, contrapositive, modus ponens, etc.), the description of the formal properties of causality is less straightforward. Causality cannot be seen as a classical logical relationship. An elementary example is the expression "If it rains, the lawn is wet". This expression cannot be translated by a formula of classical logic rain -> lawn-wet, which would mean that as soon as it rains, the lawn is automatically wet. In fact, there may be exceptions to this rule (the lawn is under a shed...). You can also change the environment (cover the lawn). These revisable rules with exceptions are well known in Artificial Intelligence. They have given rise to non-monotonic logics and revision theories. On the other hand, and more technically, we find here all the classic problems that arise when we try to formalize and use negation by failure in programming languages such as Prolog or Solar [17].

To give the links between our causal relations cause and block, in a classical language (propositional calculus or first-order logic) we need to do two things:

- *describe the internal properties of the cause and block relations*;
- *describe the links between these relations and classical logic*.

All this while considering the problem of the uncertain and the revisable. For

the first aspect, the minimum and necessary links between the two causal relations will be given explicitly. The links with classical logic will first be described in terms of fault logic. Then, to consider the aspect of discovery (abduction, production field), these links will be given in the logic of hypotheses.

In our context, to give the links between the two causal relations cause and block we go for the simplest, using classical logic. The most elementary is to explicitly give the two axiom patterns:

($C1$) *cause* (A, B) ∧ *cause* (B, C) -> *cause* (A, C)

($C2$) *cause* (A, B) ∧ *block* (B, C) -> *block* (B, C)

We believe this is the minimum, and probably sufficient, axiomatic system for application to the cell. For the moment, there is no formal link between the two relations. It is of course possible to add other axioms to take these links into account, but their relevance is not always obvious in this context.

Causality and Classical Inference

In a first approach, the first laws of causality that we want to give can be expressed naturally, by rules of the type:

1) *If A causes B and A is true*, *then B is true*.

2) *If A blocks B and A is true*, *then B is false*.

Depending on the context, *true* can mean *known*, *certain*, *believed*, or even more technically demonstrated.

These laws could be expressed in classical logic by the axioms:

cause (A, B) ∧ A -> B

block (A, B) ∧ A -> B

Or, weaker, by inference rules close to modus ponens:

cause (A, B) ∧ A / B

block (A, B) ∧ A / B

But these two formulations pose problems, as soon as there's a conflict. If, for example, we have a set F of three pieces of information F = {A, *cause* (A, B), *block* (A, B)}, we will infer from F, B and B in both approaches, which is inconsistent. To resolve these conflicts, we can try to use methods inspired by constraint programming, in particular negation by failure. It is also possible to use revisable reasoning, in particular non-monotonic logics. The first approach poses many theoretical and technical problems if we leave simple cases [18]. Here, we'll be looking at a non-monotonic approach and, more specifically, the use of default logic.

To resolve the conflicts seen above, the intuitive idea is to weaken the formulation of the causality rules into:

(1') If A causes B, and if A is true, and if it *is possible that B*, then B is true.

(2') If A blocks B, and if A is true, and if it *is possible that B is false*, then B is false.

The question then becomes how to formally describe the *possible*. This question began to arise in Artificial Intelligence some thirty years ago, when we wanted to formalize natural human reasoning. In this type of reasoning, we are obliged to

reason with incomplete, uncertain, revisable and sometimes false information. On the other hand, you often have to choose between several possible conclusions. The basic example is: {Penguins are birds, Birds fly, Penguins don't fly}. If Tweety is a bird, we arrive at a contradiction: the system is inconsistent. This inconsistency can be lifted if we manage to handle the exception by saying "Birds usually fly". Non-monotonic logics formally describe modes of reasoning that take these phenomena into account [19].

We'll be using one of the best-known non-monotonic logics here, default logic. In this logic, rules (1') and (2') are expressed intuitively:

(1") If A causes B, and if A is true, and if B *is not contradictory*, then B is true.

(2") If A blocks B, and if A is true, and if B *is not contradictory*, then B is true.

In fault logic, these rules will be represented by *normal faults* and written as:

$d1$: *cause* $(A, B) \wedge A: B / B$

$d2$: *block* $(A, B) \wedge A: B / B$

To simplify, a *default* is a specific inference rule of the type $X: Y/Z$ (X, Y and Z are classical formulas. The formula X is the *prerequisite*, Y the *justification* and Z the *consequent*. A default is normal if the justification is equal to the consequent ($Y = Z$).

A default logic $DL = \{W, D\}$ is given by a set W of classical first-order logic formulas and a set D of defaults. W can be thought of as the set of known (certain, proven) facts.

Default logic enables us to calculate extensions. An extension represents a possible state of the world (a possible result of a logical equation).

If all defaults are normal, to calculate an extension, from a theory

$DL = \{W, D\}$ we start from state $E_0 = W$, choose a fault $d = X: Y/Y$, and check the conditions:

- is X in E_0 (is the prerequisite verified?)
- is Y not in E_0 (justification verified?)

If both conditions are true, we add Y to E_0, and repeat the operation on the result, choosing another default. We stop when all the defaults have been seen; we then have a fixed point. This algorithm is not correct if we leave the field of normal defaults.

For the elementary case above, if A is true, we have:

$W = \{A\}$

$D = \{d_1, d_2\}$

and we obtain two extensions:

E_1 *which contains* B (*by applying* d_1)

E_2 *which contains* B (*applying* d_2)

The conflict is thus resolved, but the problem of which extensions to prefer arises: *is B induced or blocked?* In fact, this really depends on the context. We can, for example, prefer positive interactions to negative ones; or use statistical or probability methods. Another approach is to calculate the probability of each extension according to the properties of the problem.

From an algorithmic point of view, this preference can be evaluated either during the computation of the extension or on the result.

3. Discretization of Time Series from Experiments

The first step in the modeling process is to discretize the concentration levels. Identifying the significant changes in metabolite concentrations is relevant to conclusions/extensions generation through Default Logic. To derive hypotheses with a degree of generality, it is essential to use intervals rather than individual real values. Although an interval constraints approach could have been utilized, we have opted for a discretization method instead.

Our practical problem is that we want to have a statistically relevant (unsupervised) discretization for N metabolites concentrations over time. We also discretize the values of *Km* (*Michaelis-Menten* constants), for each reaction, with the same levels. For that purpose, we use a probabilistic model, used in speech recognition and time series analysis: continuous hidden Markov model (HMMs) [9]. We can therefore compute an appropriate number of levels (that was three for *E. coli*) in regard to a Bayesian score such as Bayesian. This process can be achieved through maximum likelihood estimation or maximum *a posteriori* estimation or through a variational Bayesian method [20].

We utilize continuous (Gaussian) Hidden Markov Models (HMMs). The advantages of using parameter tying in Hidden Markov Models (HMMs) include improved model efficiency and consistency [21]. By sharing parameters across multiple models, we reduce the total number of parameters that need to be estimated, which helps to prevent overfitting, especially when working with limited data (**Figure 3**).

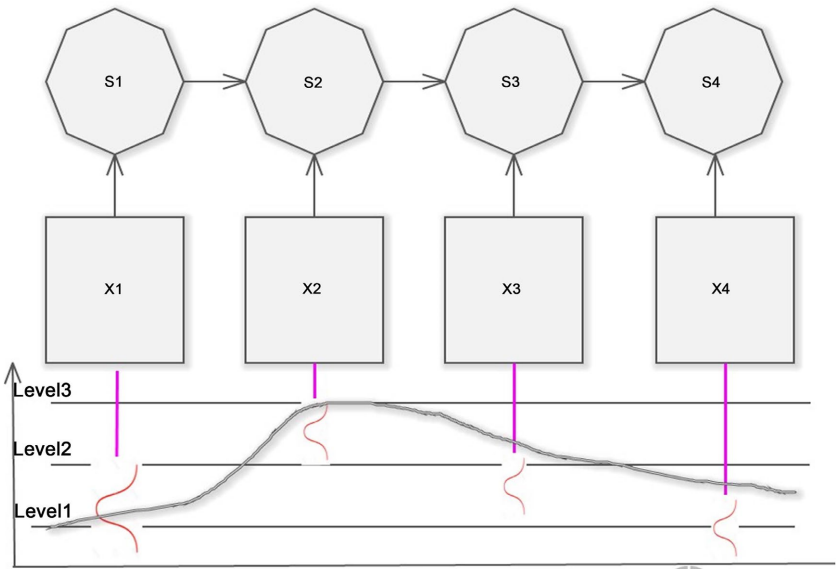


Figure 3. 3-state continuous HMM discretizing one experimental time series, where X_t is the measurement of concentration at time t and S_t is the hidden state that indicates the corresponding discretized level.

This approach also enhances the interpretability of the model, as it maintains a consistent representation of the underlying processes across different metabolites. Additionally, parameter tying allows for better generalization, as the learned parameters can be applied across related states or compounds, leading to more robust predictions. This approach addresses the challenge of maintaining consistent symbolic levels across all logic models, allowing us to assign the concentration level of one compound to another while working with the same underlying real values. Initially, we prepare N continuous HMMs, one for each metabolite, in which each state variable represents a concentration level, and each output variable corresponds to a concentration measurement, following a univariate Gaussian distribution [22]. All HMMs share a common state space and parameters for the output variables (such as means and variances), ensuring that they yield corresponding discrete concentration levels. These relevant discretized concentration levels are determined using the expectation-maximization (EM) algorithm with maximum *a posteriori* (MAP) estimation [23].

4. Modeling of the Pathways of *E. coli*

To gain insights into central metabolism, we developed a logical model based on a kinetic framework that incorporates glycolysis and the pentose phosphate pathway for *Escherichia-coli*. **Figure 4** illustrates the simplified pathway that we modeled logically, incorporating relationships between substrates, enzymes, products, and the Michaelis constant (Km) [24].

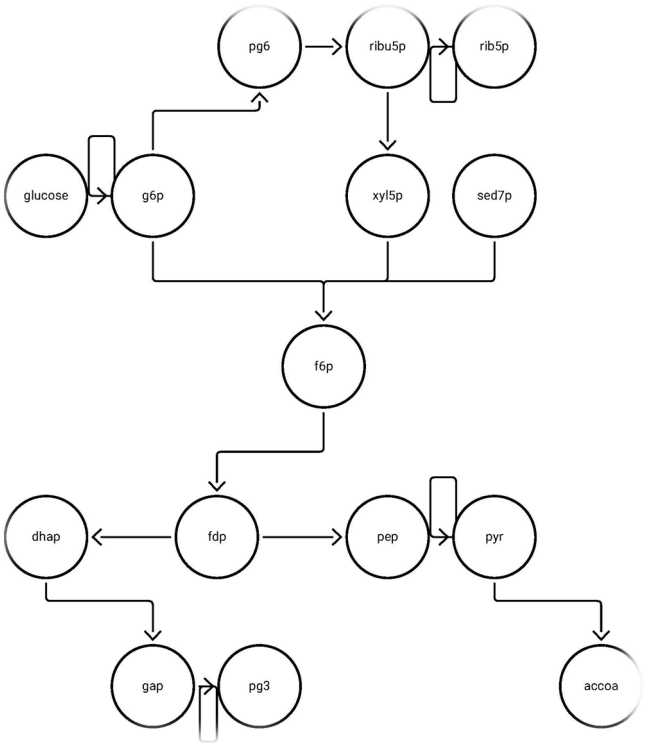


Figure 4. Simplified glycolysis and pentose phosphate pathways for *E. coli*.

The dynamics of metabolic networks are primarily influenced by classical kinetics, particularly Michaelis-Menten, Hill, and allosteric kinetics. By focusing our modeling efforts on these specific kinetic types, we can greatly streamline the mathematical analysis, which is the strategy we employed.

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} E + P$$

$$\text{Michaelis-Menten eq.:} \frac{\mathrm{d}[P]}{\mathrm{d}t} = V_m \frac{[S]}{[S] + K_m} \tag{1}$$

The implications of using Michaelis-Menten kinetics for modeling metabolic processes are significant. This framework improves our understanding of enzyme behavior and reaction rates by creating a clear connection between substrate concentration and reaction velocity [25]. By applying Michaelis-Menten kinetics, researchers can predict how changes in substrate availability will affect metabolic flux, aiding in the analysis and interpretation of experimental data. Moreover, this approach simplifies mathematical modeling, which is especially useful in complex metabolic networks. However, it is important to acknowledge that while Michaelis-Menten kinetics provides a useful approximation, it may not completely encompass the complexities of allosteric regulation or cooperative binding in enzymes, leading to potential oversimplifications in certain contexts.

When both the substrate (S) and the product (P) are present, neither can fully saturate the enzyme. For any specific concentration of S, the fraction of S that binds to the enzyme decreases with an increase in the concentration of P, and the same applies in reverse. For any given concentration of P, an increase in S will reduce the fraction of P bound to the enzyme. We will examine a time discretization of the chemical rate equation for the reaction involving the substrate and product, characterized by their respective stoichiometric coefficients s and p.

$$\begin{aligned} s.S \rightarrow p.P : \text{rate} &= \frac{1}{p} \times \frac{\mathrm{d}[P]}{\mathrm{d}t} \rightarrow_{\text{dise\_time}} \frac{1}{p} \times \frac{\Delta[P]}{\Delta T} \\ &\Rightarrow p \times \text{rate} = V_m \frac{[S]_T}{[S]_T + K_m} \\ &\approx \frac{[P]_{T+\text{ timestep}} - [P]_T}{(T + \text{timestep}) - T} \end{aligned} \tag{2}$$

We chose to work with a constant timestep:

$$\Rightarrow [P]_{T+1} = V_m \frac{[S]_T}{[S]_T + K_m} + [P]_T \tag{3}$$

It is important to note that the Michaelis-Menten constants (Km) are equivalent to a concentration unit. In our modeling, we can represent them as:

conc (*Km*, *Level*, *Time*), where "*conc*" signifies concentration.

The experimental observations of intracellular metabolites in response to a glucose pulse were conducted in a continuous culture using automatic stopped flow and manual fast sampling techniques, capturing data in the time frame of seconds

and milliseconds post-glucose stimulus. We measured extracellular glucose and intracellular metabolites, including glucose-6-phosphate (G6P) and fructose-6-phosphate (F6P), fructose1-6bisphosphate (fdp), glyceralde-hyde3phosphate (gap), phospho-enol pyruvate (pep), pyruvate (pyr), 6phosphate-gluconate (6 pg), glucose-1-phosphate (g1p) as well as the cometabolites: atp, adp, amp, nad, nadh, nadp, nadph were measured using enzymatic methods or high performance liquid chromatography. All the steady-state concentrations measurements of the *E. coli* experiment and their corresponding discrete levels are summarized in Table 1.

Table 1. Concentrations (mM/L) of the Metabolites and their discretized levels for steady states.

#	Metab.	Conc.	Lvl	#	Metab.	Conc.	Lvl
1	glucose	0.055	0	2	g6p	3.480	2
3	f6p	0.600	0	4	fdp	0.272	0
5	gap	0.218	0	6	pep	2.670	2
7	pyr	2.670	2	8	6pg	0.808	1
9	glp	0.653	0	10	amp	0.955	1
11	adp	0.595	0	12	atp	4.270	2
13	nadp	0.195	0	14	nadph	0.062	0
15	nad	1.470	1	16	nadh	0.100	0

Our logical framework is founded on the simplified Michaelis-Menten equation, which is articulated here through three background clauses utilizing the: *conc* (*Compound*, *Level*, *Time*), predicate.

5. Ranking Extensions

BDD-EM stands for Binary Decision Diagram-Expectation Maximization. It is an algorithm that combines the principles of binary decision diagrams with the expectation maximization technique. This approach allows for efficient handling of boolean functions, particularly in the context of ranking hypotheses and managing probabilities in uncertain environments. BDD-EM is particularly useful in applications such as machine learning, reasoning under uncertainty, and abductive reasoning. Inoue applied the BDD-EM algorithm to rank hypotheses generated through abduction [14].

To rank our extension E_1, ..., E_n by probability, we consider the finite set of ground atoms A, which includes all possible values for *conc* (*Compound*, *Level*, *Time*) and *reaction* (*Substrate*, *Product*, *Km*).

Each element of A is treated as a boolean variable. Among these, there is a subset of abducibles Γ consisting of all potential values of *conc* (*Compounds*, *Level*, 0). By denoting $\theta_i = P(A_i)$ for $A_i \in A$, we aim to maximize the probability of the disjunction of extensions: $F = (E_1 \vee \cdots \vee E_n)$, to determine the optimal θ parameters using the BDD-EM algorithm. An **extension** is a complete set of conclusions that

can be drawn from a knowledge base when applying default rules consistently. Therefore, an extension captures all the inferences that can be derived while maintaining the consistency of the beliefs. Since F can be excessively large to maintain as a BDD, an optimized version F' is created using the minimal proofs for B and each extension E_i. Subsequently, the BDD-EM algorithm calculates the probabilities of the ground atoms in A that maximize the probability of F'. Finally, the probabilities for each extension used in the ranking are computed as the product of the probabilities of the literals within each E_i. Ishihata *et al.*, proposed the BDD-EM algorithm, which implements the expectation maximization algorithm using binary decision diagrams, enabling it to handle Boolean functions effectively. Inoue *et al.* [13] applied the BDD-EM algorithm to rank hypotheses generated through abduction. To rank our extensions ($E_1,\cdots,E_n$) based on probability, we consider the finite set of ground atoms (A), which includes the following reaction:

$$reaction(S,P,Km)\wedge conc(S,L,T)\wedge conc(Km,L,T)$$
$$\wedge conc(P,L2,T)\rightarrow conc(P,L2,T+1)$$

The change in concentration of the product between time (T) and $(T+1)$ is not significant enough to transition from one level to another. This is an approximation resulting from our discretization (using a logarithmic scale on real values).

$$\left([S]\gg Km\Rightarrow \Delta[P]=Vm\Rightarrow [P]_{T+1}=Vm+[P]_T\right)$$
$$reaction(S,P,Km)\wedge conc(S,2,T)\wedge conc(Km,0,T)$$
$$\wedge conc(P,L,T)\rightarrow conc(P,2,T+1)$$

If the reaction occurs very rapidly, it will convert all the substrate into product within a single time step. If we had more than three levels, we would need additional rules (which can be generated automatically) or a general procedure for managing our kinetic model. Another approach being explored to handle more levels involves the automated generation of kinetic rules concerning discretization. Additionally, we simplified the pathways to utilize only Michaelis-Menten kinetics; another research avenue is to expand our model to include reactions governed by other types of kinetics.

We also imposed constraints regarding the uniqueness of levels at a given time to reduce the number of hypotheses while maintaining consistency:

$$\left(conc(S,0,T)\vee conc(S,1,T)\right)$$
$$\left(conc(S,0,T)\vee conc(S,2,T)\right)$$
$$\left(conc(S,1,T)\vee conc(S,2,T)\right)$$

Now, we set the observations for the six metabolites (#2 - #7) from **Table 1**, which have potentially been influenced by glucose stimulation, and defined the abducibles as literals of the form $\left(conc(,,0)\right)$. Using Default Logic, we obtained

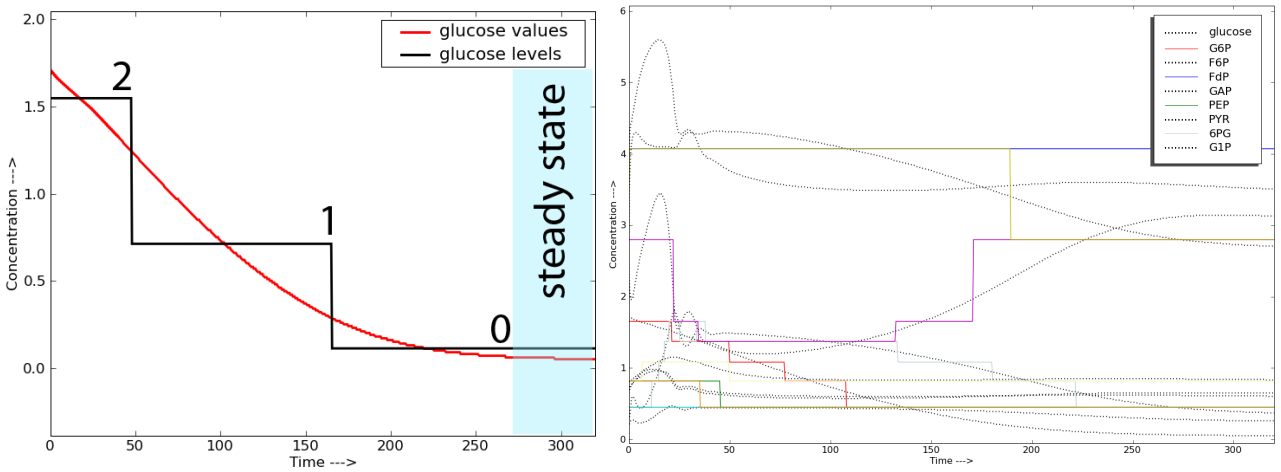
48 extensions, for example:

$$E26 = conc(g6p,2,0) \wedge conc(adp,2,0) \wedge conc(gap,0,0) \wedge conc(glucose,2,0) \wedge conc(pg3,2,0) \wedge conc(pep,2,0) \wedge conc(atp,0,0) \wedge conc(pyr,2,0)$$

These conclusions/extensions correspond to our biological knowledge that pyruvate is a bottleneck [21] and that the glucose that is totally consumed (**Figure 5** from simulation) was in high concentration at the beginning of the experiment (pulse). It goes along with the very general reaction of glycolysis:

$$glucose + 2ADP + 2P + 2NAD^{+} \rightarrow 2pyruvate + 2ATP + 2(NADH, H^{+}) + 2H_2O.$$

Also, for some metabolites, such as fructose-6-phosphate, the levels found through abduction are corresponding to the output of the simulation (**Figure 5(b)**) with the same low level (0) before and after the dynamic transition.



(a) (b)

Figure 5. (a) Discretization of the concentration of glucose in the Glycolysis Pathway of *E. coli* after an initial pulse; (b) Simulated evolution of the concentration of all metabolites during the whole experiment.

The Preferred Knowledge Base

Enhancing our understanding of a system is viewed as an iterative endeavor. Initially, we establish a knowledge base that merges our existing background knowledge with observations [26]. From this foundation, we generate hypotheses and apply algorithms to update our knowledge base with certain identified hypotheses, referred to as abducibles. The goal is to continually revisit the extension generation process until no new insights emerge. This iterative approach is crucial, especially when dealing with complex chained reactions and multiple time steps, as it fosters a more profound understanding. The concept of revising the knowledge base is also discussed in the work of Baral [4], who adopt a nonmonotonic strategy, although their method remains limited qualitative modeling and does not take quantitative aspects into account.

It is essential to select extensions that align with background knowledge. For instance, when employing a greedy algorithm (like **Algorithm 1**) that prioritizes

hypotheses based on decreasing probability, we ensure that the chosen hypotheses contribute to our knowledge while maintaining consistency. This approach limits us to abducing only the discoverable found within extension E26.

Algorithm 1. Algorithm to enhance the knowledge base: most probable first.

```
knowledge ← knowledge base
sorted extensions ← sort (extensions)
while length (discoverable) > 0 && length (sorted hypotheses) > 0
    do
    tmp ← sorted extensions.pop()
    if contains (tmp, discoverable) && consistent (tmp,
    knowledge) then knowledge.enhance (tmp) discoverable.remove (tmp)
    end if
end while
```

With the explicit functions *length*, *pop* (destructive), and:

- *sort* sorts the extensions by decreasing probability.
- *contains* is a function that returns statements of first argument contained in the second.
- *enhance* adds statements that are not yet present in the considered ("self", "this") knowledge.
- *remove* deletes statements from argument present in the considered ("self", "this") object (could make use of *contains*).

6. Conclusions

Our findings at time $T = 0$ (*steps*) concerning pyruvate and the concentration of pyruvate at $T = 38$ (*steps*) are consistent with established biological knowledge and our ODE-based simulator. This paper introduces a method for addressing the kinetics of metabolic pathways through a symbolic model, as demonstrated in Figure 1. We elaborated on the process of discretizing biological experiments into relevant levels that can be applied using Default Logic and logic programs. Additionally, by discretizing concentration into levels, we described our approach to transforming the Michaelis-Menten kinetics equation into logic rules, a direction we believe has not been previously explored. The uniqueness of our work lies in the capability of a logical model to capture the dynamic response of microorganisms to glucose pulses, thereby improving the accuracy of metabolic flux analysis. Expanding this framework to encompass reactions involving two substrates and/or two products would aid in developing more comprehensive models [27].

Like the approach taken by King *et al.* [12], our method examines the behavior of multiple ordinary differential equations while utilizing the strengths of a symbolic model [8], especially in the statistical evaluation of hypotheses. This evaluation process, facilitated by BDD-EM, effectively extracts relevant insights from large data sets. The practical validity of our entire methodology, including the discretization process, is highlighted by the results presented in this paper, which operate within a well-established theoretical framework. We strongly believe that

integrating time series discretization with kinetic modeling will lead to significant enhancements in Default Logic's ability to tackle ODEs. Furthermore, we propose that knowledge discovery should be regarded as an iterative process, in which one continuously updates their knowledge base based on new findings (as illustrated in **Figure 2** with the introduction of "New KB"). Nonetheless, there is potential for improvement in our modeling approach, particularly in refining time and concentration discretization. Future experiments will explore more than three levels and various time steps, focusing on the Glycolysis and Pentose Phosphate pathways in the bacterium Saccharomyces cerevisiae [28], using both real experimental data and simulated data.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G.K., Bryant, C.H., Muggleton, S.H., *et al.* (2004) Functional Genomic Hypothesis Generation and Experimentation by a Robot Scientist. *Nature*, **427**, 247-252. https://doi.org/10.1038/nature02236

[2] Kitano, H. (2002) Systems Biology: Toward System-Level Understanding of Biological Systems. *Science*, **295**, 1662-1664.

[3] Sriyudthsak, K., Shiraishi, F. and Hirai, M.Y. (2016) Mathematical Modeling and Dynamic Simulation of Metabolic Reaction Systems Using Metabolome Time Series Data. *Frontiers in Molecular Biosciences*, **3**, Article 15. https://doi.org/10.3389/fmolb.2016.00015

[4] Baral, C., Chancellor, K., Tran, N., Tran, N.L., Joy, A. and Berens, M. (2004) A Knowledge Based Approach for Representing and Reasoning about Signaling Networks. *Bioinformatics*, **20**, i15-i22. https://doi.org/10.1093/bioinformatics/bth918

[5] Moraru, I.I., Schaff, J.C., Slepchenko, B.M., Blinov, M.L., Morgan, F., Lakshminarayana, A., *et al.* (2008) Virtual Cell Modelling and Simulation Software Environment. *IET Systems Biology*, **2**, 352-362. https://doi.org/10.1049/iet-syb:20080102

[6] Schaffer, L.V. and Ideker, T. (2021) Mapping the Multiscale Structure of Biological Systems. *Cell Systems*, **12**, 622-635. https://doi.org/10.1016/j.cels.2021.05.012

[7] Bandara, S., Schlöder, J.P., Eils, R., Bock, H.G. and Meyer, T. (2009) Optimal Experimental Design for Parameter Estimation of a Cell Signaling Model. *PLOS Computational Biology*, **5**, e1000558. https://doi.org/10.1371/journal.pcbi.1000558

[8] Reisz, J.A. and D'Alessandro, A. (2017) Measurement of Metabolic Fluxes Using Stable Isotope Tracers in Whole Animals and Human Patients. *Current Opinion in Clinical Nutrition & Metabolic Care*, **20**, 366-374. https://doi.org/10.1097/mco.0000000000000393

[9] Geiger, D. (2021) Correction To: Plant Glucose Transporter Structure and Function. *Pflügers Archiv—European Journal of Physiology*, **473**, 1687-1687. https://doi.org/10.1007/s00424-021-02603-5

[10] Chassagnole, C., Rodrigues, J., Doncescu, A. and Yang, L.T. (2006) Differential Evolutionary Algorithms for *in Vivo* Dynamic Analysis of Glycolysis and Pentose Phosphate Pathway in *Escherichia coli*. A. Zomaya.

[11] Emwas, A., Szczepski, K., Al-Younis, I., Lachowicz, J.I. and Jaremko, M. (2022) Flux-

omics—New Metabolomics Approaches to Monitor Metabolic Pathways. *Frontiers in Pharmacology*, **13**, Article 805782. https://doi.org/10.3389/fphar.2022.805782

[12] King, R.D., Garrett, S.M. and Coghill, G.M. (2005) On the Use of Qualitative Reasoning to Simulate and Identify Metabolic Pathways. *Bioinformatics*, **21**, 2017-2026. https://doi.org/10.1093/bioinformatics/bti255

[13] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., *et al.* (2007) KEGG for Linking Genomes to Life and the Environment. *Nucleic Acids Research*, **36**, D480-D484. https://doi.org/10.1093/nar/gkm882

[14] Inoue, K., Sato, T., Ishihata, M., Kameya, Y. and Nabeshima, H. (2009) Evaluating Abductive Hypotheses Using EM Algorithm on BDDs. *Proceeding of IJCAI*-09, Pasadena, 17-18 July 2009, 820-815.

[15] Doncescu, A., Yamamoto, Y. and Inoue, K. (2007) Biological Systems Analysis Using Inductive Logic Programming. 21*st International Conference on Advanced Information Networking and Applications Workshops* (*AINAW*'07), Niagara Falls, 21-23 May 2007, 690-695. https://doi.org/10.1109/ainaw.2007.112

[16] Dworschak, S., Grell, S., Nikiforova, V.J., Schaub, T. and Selbig, J. (2008) Modeling Biological Networks by Action Languages via Answer Set Programming. *Constraints*, **13**, 21-65. https://doi.org/10.1007/s10601-007-9031-y

[17] Tiwari, A., Talcott, C., Knapp, M., Lincoln, P. and Laderoute, K. (2007) Analyzing Pathways Using Sat-Based Approaches. In: Anai, H., Horimoto, K. and Kutsia, T., Eds., *Algebraic Biology*, Springer, 155-169. https://doi.org/10.1007/978-3-540-73433-8_12

[18] De Raedt, L. (2008) Logical and Relational Learning. Springer.

[19] Benhamou, F. (1995) Interval Constraint Logic Programming. In: Podelski, A., Ed., *Constraint Programming*: *Basics and Trends*, Springer, 1-21. https://doi.org/10.1007/3-540-59155-9_1

[20] Ji, S.H., Krishnapuram, B. and Carin, L. (2006) Variational Bayes for Continuous Hidden Markov Models and Its Application to Active Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, 522-532. https://doi.org/10.1109/tpami.2006.85

[21] Gauvain, J. and Lee, C.H. (1994) Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, **2**, 291-298. https://doi.org/10.1109/89.279278

[22] Holmes, I. and Rubin, G.M. (2002) An Expectation Maximization Algorithm for Training Hidden Substitution Models. *Journal of Molecular Biology*, **317**, 753-764. https://doi.org/10.1006/jmbi.2002.5405

[23] Rabiner, L.R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, **77**, 257-286. https://doi.org/10.1109/5.18626

[24] Peters-Wendisch, P.G., Schiel, B., Wendisch, V.F., Katsoulidis, E., Möckel, B., Sahm, H. and Eikmanns, B.J. (2001) Pyruvate Carboxylase Is a Major Bottleneck for Glutamate and Lysine Production by *Corynebacterium glutamicum*. *Journal of Molecular Microbiology and Biotechnology*, **3**, 295-300.

[25] Seo, J., Shin, J., Leijten, J., Jeon, O., Camci-Unal, G., Dikina, A.D., *et al.* (2018) High-Throughput Approaches for Screening and Analysis of Cell Behaviors. *Biomaterials*, **153**, 85-101. https://doi.org/10.1016/j.biomaterials.2017.06.022

[26] Fages, F. and Soliman, S. (2008) Model Revision from Temporal Logic Properties in Computational Systems Biology. In: De Raedt, L., Frasconi, P., Kersting, K. AND

Muggleton, S., Eds., *Probabilistic Inductive Logic Programming*, Springer, 287-304. https://doi.org/10.1007/978-3-540-78652-8_11

[27] Fife, S.T. and Gossner, J.D. (2024) Deductive Qualitative Analysis: Evaluating, Expanding, and Refining Theory. *International Journal of Qualitative Methods*, **23**, 1-12. https://doi.org/10.1177/16094069241244856

[28] Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **28**, 27-30. https://doi.org/10.1093/nar/28.1.27