

ГЕНЕТИКА/GENETICS

DOI: https://doi.org/10.60797/jbg.2026.32.1 EDN: ZTKORD

AN AI-ASSISTED PIPELINE TO MODEL AND ANALYZE GENE REGULATORY NETWORKS

Research article

Lopez Prato J.[1,]*, Dávila J.[2], Martínez R.[3], Ramírez Y.[4], Bastidas M.[5]
[1] ORCID : 0000-0003-1973-308X;
[1, 4] University of Táchira, San Cristóbal, Venezuela
[2, 5] University of Los Andes, Mérida, Venezuela
[3] University of Murcia, Murcia, Spain

* Corresponding author (jlopez[at]unet.edu.ve)

Suggested: 09.02.2026; Accepted: 24.06.2026; Published: 26.06.2026

Abstract

This project combines Generative AI and Logic AI to form an integrated process, a pipeline, that retrieves large volumes of scientific information and produces logical models that can be validated and leveraged by human experts. Our goals are to organize and to analyze, with the assistance of AI and other bio-information retrieval tools, the documentation associated with a domain and extract validated knowledge to support planning for problem solving. We have tested our methodology and the pipeline with an experiment on COVID-19 which we also report. The tool helped a team of biologists to review the state of the art and handle the explosion of research papers during pandemia. The main product of the pipeline is a Gene Regulatory Network (GRN); a semantic map which constitutes a representation of causal relations in system biology which can in turn be used to explore different strategies for intervention and treatment. The pipeline (named Biopatternsg) has been structured around three main tasks: a) data and information gathering; b) the use of generative AI and Large Language Models (LLMs) for the automatic extraction of biological entities and their biological relationships, and c) the use of Logical AI to manage the information obtained in a) and b) and combine it with conserved criteria, all represented as knowledge bases in Prolog. The system also keeps track of and allows access to all the original documents used as sources of knowledge. We believe the pipeline is useful for deepening the understanding of COVID-19. We have also organized a systematic evaluation of biopatternsg as a method to produce gene regulatory networks, GRN. The evaluation of names recognition reports an average F1-Score of 0.9069 with a variance of 0.0145 among 30 networks used as the gold standard. This indicates an excellent capacity to recognize names of real, biological objects. The results are not good at other levels of evaluation, e.g. An average F1-Score is of 0.2306 with a variance of 0.0218. It would be worth noticing the big discrepancy between Precision and Recall. The system even exhibits perfect Recall (of 1.0) at some of the networks, but Precision is always below 0.5 driving the F1-score down. We discuss these results and explain why even that partial success with Recall is an encouraging result.

Biopatternsg repository at: https://github.com/biopatternsg/biopatternsg

Supplemental material in: https://github.com/biopatternsg/biopatternsg/tree/feature/evaluation

Keywords: Artificial Intelligence, Generative AI, Logical AI, Gene Regulatory Network, LLM, Semantic Network.

КОНВЕЙЕР ОБРАБОТКИ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ МОДЕЛИРОВАНИЯ И АНАЛИЗА СЕТЕЙ РЕГУЛЯЦИИ ГЕНОВ

Научная статья

Лопез Прато Х.[1,]*, Да́вила Х.[2], Мартинез Р.[3], Рамирас Д.[4], Бастидас М.[5]
[1] ORCID : 0000-0003-1973-308X;
[1, 4] Университет Тачиры, Сан-Кристобаль, Венесуэла (Боливарианская Республика)
[2, 5] Университет Лос-Андэс, Мерида, Венесуэла (Боливарианская Республика)
[3] Университет Мурсии, Мурсия, Испания

* Корреспондирующий автор (jlopez[at]unet.edu.ve)

Предложена: 09.02.2026; Принята: 24.06.2026; Опубликована: 26.06.2026

Аннотация

Этот проект объединяет генеративный ИИ и логический ИИ для создания интегрированного процесса, конвейера, который извлекает большие объемы научной информации и создает логические модели, которые могут быть проверены и использованы экспертами. Наши цели — организовать и проанализировать с помощью ИИ и других инструментов поиска биоинформации документацию, связанную с определенной областью, и извлечь проверенные знания для поддержки планирования решения проблем. Мы протестировали нашу методологию и конвейер на эксперименте с COVID-19, о котором также сообщаем. Инструмент помог команде биологов оценить современное состояние дел и справиться с резким ростом числа научных работ во время пандемии. Основным продуктом конвейера является сеть генной регуляции (GRN); семантическая карта, представляющая собой отображение причинно-следственных связей в системной биологии, которая, в свою очередь, может быть использована для изучения различных стратегий вмешательства и лечения. Конвейер (названный Biopatternsg) был структурирован вокруг трех основных задач: a) сбор данных и информации; b) использование генеративного ИИ и больших языковых моделей (LLM) для автоматического извлечения биологических объектов и их биологических взаимосвязей, и c) использование

логического ИИ для управления информацией, полученной в пунктах a) и b), и ее объединения с консервативными критериями, представленными в виде баз знаний на языке Prolog. Система также отслеживает и обеспечивает доступ ко всем исходным документам, используемым в качестве источников знаний. Мы считаем, что этот конвейер полезен для углубления понимания COVID-19. Мы также организовали систематическую оценку biopatternsg как метода для создания генных регуляторных сетей (GRN). Оценка распознавания имен показывает средний показатель F1-Score 0,9069 с дисперсией 0,0145 среди 30 сетей, используемых в качестве эталонного стандарта. Это указывает на отличную способность распознавать названия реальных биологических объектов. Результаты на других уровнях оценки не очень хороши, например, средний показатель F1-Score составляет 0,2306 с дисперсией 0,0218. Стоит отметить значительную разницу между точностью (Precision) и полнотой (Recall). Система даже демонстрирует идеальную полноту (1,0) в некоторых сетях, но точность всегда ниже 0,5, что снижает показатель F1-меры. Мы обсуждаем эти результаты и объясняем, почему даже частичный успех с полнотой является обнадеживающим.

Репозиторий Biopatternsg находится по адресу: https://github.com/biopatternsg/biopatternsg

Дополнительные материалы находятся в: https://github.com/biopatternsg/biopatternsg/tree/feature/evaluation

Ключевые слова: искусственный интеллект, генеративный ИИ, логический ИИ, генная регуляторная сеть, LLM, семантическая сеть.

Introduction

In his 2011 book, "Thinking, Fast and Slow," Nobel Prize winner Daniel Kahneman [1], explains that human thinking could be seen as having two components: System 1, which is fast and intuitive, and System 2, which is slower and responsible for explainable reasoning. The same phenomenon is observed in Artificial Intelligence: Generative AI is fast and intuitive and can be used to identify patterns in huge volumes of data. Logical AI enables systematic, verifiable, and explainable reasoning on this preprocessed data. This project combines both types of AI to form an integrated process, a pipeline, that retrieves large volumes of scientific information and produces logical models that can be validated and leveraged by human experts.

We propose a connection between generative AI and logic AI to define a framework for knowledge modeling, management, and discovery that enables the automatic construction of knowledge networks for a domain. The main idea here is to retrieve and analyze, with the help of AI, the documentation associated with a domain and extract from it the elements and possible relationships that connect them, thus facilitating its understanding and assisting in the planning of actions for problem solving (See Figure 1).

Figure 1 depicts the general components in our proposal as a pipeline that connects three main tasks:

a) access to various sources for information gathering inherent to a domain and problem of interest;

b) use of generative artificial intelligence to assist in the automatic extraction of knowledge;

c) use of logic artificial intelligence to automatically organize and explore knowledge bases, following constraints that guide such exploration.

The final result in our work are models (graphs or networks) that synthesize knowledge in such a way as to facilitate decision-making related to problem solving. And it is our main goal to answer queries like: given a drug, what kinds of protein-protein interactions result when that drug binds to a known receptor, and which of them lead to the activation (or inhibition) of the transcriptional response of some gene?. We also consider questions such as: for a given pair of proteins, is there a subnet in this modeled GRN that describes interconnected regulatory pathways, in which such proteins stimulate and inhibit their mutual transcription in any way? [2].

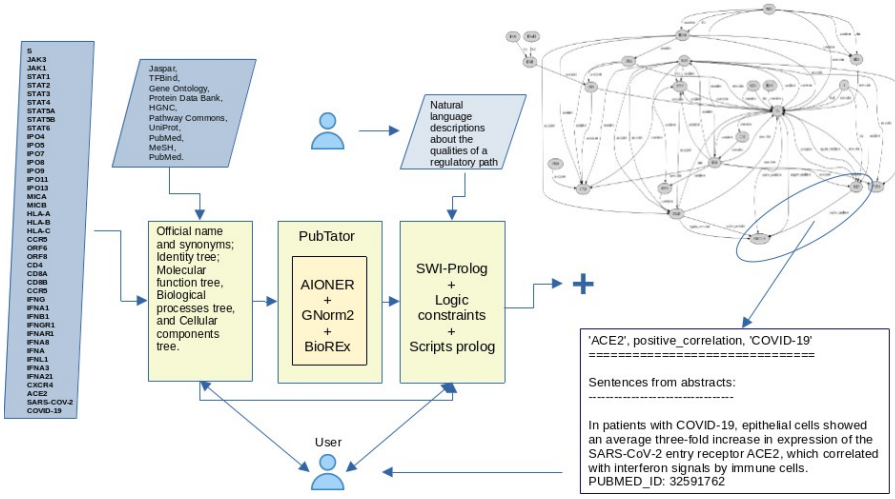

Figure 1 - A schematic view of the modeling and type of results for COVID-19

DOI: https://doi.org/10.60797/jbg.2026.32.1.1

In this work, we aimed to develop a use case for our proposed methodology related to biological systems; in particular, we focused on modeling genetic regulatory networks, specifically using COVID-19 as a testing ground. We sought to use our methodology to organize an updated description of the knowledge domain inherent to the disease, while also seeking to facilitate its analysis and the elaboration of hypotheses about its therapeutic treatment.

1.1 Gene Regulatory Networks (GRNs)

Gene transcription regulatory networks (GRNs) are biological mechanisms responsible for regulating the presence or absence of gene-associated products. In a GRNs, a regulatory event can activate a product that in turn participates in an event, which activates or inhibits another product. It is normal, then, that in such networks the complexity of the interrelationships grows very rapidly [3], [4].

In order to manage the current knowledge inherent to GRNs, informatics strategies have been developed aimed at the description, organization, interrelation and analysis of the elements that constitute them. Among such strategies are ontologies [5], [6], [7], [8] and process diagrams [9], [10].

Here, we are interested in the modeling and analysis of Gene transcription regulatory networks, based on the inferential processing of knowledge bases [11]; therefore, our work fits in the ontology-based approach and in their logical and analysis facilities. Elaborating on that, we have proposed the automatic construction and integration of artificial intelligence and various knowledge bases (KBs), in order to take advantage of the great effort that the scientific community has made about GRNs, regarding their organization and availability.

1.2 The Pipeline

Figure 1 depicts the flow of information in our system applied to the COVID-19 disease. In this scenario, a biologist or physician explores possible treatment alternatives for the COVID-19, providing some inputs and obtaining a semantic graph that represents a GRN, enriched with a set of knowledge bases and detailed documentation about the regulatory events and pathways that shape that network. Below are some details of the pipeline's inputs, components, and results:

1. Initial list of descriptors (concepts, keywords) provided by the user. For this particular domain, the descriptors are biological or chemical objects; for instance: drugs, small molecules, genes/proteins of various types, and the name of the disease to analyze.

2. Bioinformatics portals accessible through the internet, which describe objects such as those mentioned in the previous item; for instance MeSH, Gene Ontology and PubMed (pubmed.ncbi.nlm.nih.gov), a portal specialized in information related to biology and medicine.

3. Generative artificial intelligence, based on Large Language Models (LLMs), specially trained to handle the information available in PubMed. It allows for:

a) the automatic recognition of biological entities related to COVID-19 (NER Model: Named Entity Recognition);

b) the automatic extraction of the biological relationships that connect them (RE Model: Relation Extraction).

The output obtained from the LLMs consists of entities and relationships that allow the construction of a knowledge base, useful for automatic reasoning with symbolic AI.4. An inference system based on symbolic AI, used for automatic reasoning processes on the knowledge bases organized with the assistance of LLMs, and repositories, like the ones visible in Figure 2, Figure 3 and Figure 4. The user explains the desirable characteristics for the solutions (paths) to be obtained to guide the logical inference system in the deduction of possible paths to solutions. The output in this scenario is a GRN like the one shown in Figure 5. For a detailed view of Figure 5 see VERY-RESTRICTED/covid-19-wide-restricted-list-no-sm.png in supplemental material, in which the user can see how biological objects interact with each other, and how these interactions could eventually lead to the inhibition of COVID-19.

4. An inference system based on symbolic AI, used for automatic reasoning processes on knowledge bases organized with the assistance of LLMs and repositories, like those shown in Figure 2 and Figure 3. The user explains the desirable characteristics for the solutions (paths) to be obtained to guide the logical inference system in deducing possible paths to solutions. The output in this scenario is a GRN like the one shown in Figure 5 (see covid-19-wide-restricted-list-no-sm.svg for a detailed view), in which the user can see how biological objects interact with each other, and how these interactions could eventually lead to the inhibition of COVID-19.


```
base([
...
event('JAK3',association,'STAT'),
event('STAT',positive_correlation,'COVID-19'),
event('CALCIUM',association,'CXCR4'),
event('TYROSINE',association,'CXCR4'),
event('ARGININE',association,'CXCR4'),
event('Aspartic Acid',association,'CXCR4'),
event('Glutamic Acid',association,'CXCR4'),
......
event('CD4',positive_correlation,'COVID-19'),
event('CD4',negative_correlation,'COVID-19'),
event('PIAS3',association,'CD4'),
event('STAT',association,'CD4'),
event('JAK3',association,'CD4'),
event('CD4',positive_correlation,'STAT'),
event('CD4',association,'STAT'),
event('CXCR4',association,'CD4'),
event('CD4',association,'JAK3'),
event('CCR5',bind,'CD4'),
event('ACE2',association,'CD4'),
event('CCR5',association,'CD4'),
….
event('STAT',association,'JAK1'),
event('JAK1',association,'CXCR4'),
event('CXCR4',positive_correlation,'COVID-19'),
event('CXCR4',negative_correlation,'COVID-19'),
...
]).
```

Figure 2 - Regulatory events knowledge base (KB)

DOI: https://doi.org/10.60797/jbg.2026.32.1.2

Note: the KB was modeled automatically and contains more than 20,000 events

Figure 3 - KBs organized following the steps in Fig. 1 and Fig. 6 (Part 1)
DOI: https://doi.org/10.60797/jbg.2026.32.1.3

1) names, synonyms and basic definitions in prolog format (coming from HGNC, Uniprot, PDB and Pathways Commons). alunacedase_alfa('ACE2'). synonyms('ACE2', ['ACE2','angiotensin I converting enzyme 2','angiotensin I converting enzyme (peptidyl-dipeptidase A) 2','Angiotensin-converting enzyme 2','Angiotensin-converting enzyme homolog',...]. tissues('ACE2', ['Heart','Lymphoma','Testis','Lung','Brain','Bile','Liver']). chemokine_cxcl12('CXCR4'). synonyms('CXCR4', ['CXCR4','C-X-C motif chemokine receptor 4','chemokine (C-X-C motif), receptor 4 (fusin)',..,'Lipopolysaccharide-associated protein 3','NPYRL','Stromal cell-derived factor 1 receptor']). tissues('CXCR4',['Lung']). janus_kinase_3('JAK3'). synonyms('JAK3',['JAK3','Janus kinase 3','L-JAK','JAKL','LJAK','JAK3_HUMAN','JAK-3','Tyrosine-protein kinase JAK3','Janus kinase 3','Leukocyte janus kinase']). tissues('JAK3',['Blood']). ...	2) MeSH ontologies for each object in the network object('ACE2'). is_a('ACE2','alunacedase alfa'). is_a('alunacedase alfa','Angiotensin-Converting Enzyme 2'). is_a('Angiotensin-Converting Enzyme 2','Carboxypeptidases'). is_a('Carboxypeptidases','Exopeptidases'). is_a('Exopeptidases','Peptide Hydrolases'). is_a('Peptide Hydrolases','Hydrolases'). is_a('Hydrolases','Enzymes'). is_a('alunacedase alfa','Recombinant Proteins'). is_a('Recombinant Proteins','Proteins'). objeto('CXCR4'). is_a('CXCR4','Chemokine CXCL12'). is_a('Chemokine CXCL12','Chemokines, CXC'). is_a('Chemokines, CXC','Chemokines'). is_a('Chemokines','Cytokines'). is_a('Cytokines','Intercellular Signaling Peptides and Proteins'). is_a('Intercellular Signaling Peptides and Proteins','Peptides'). is_a('Intercellular Signaling Peptides and Proteins','Proteins'). is_a('Intercellular Signaling Peptides and Proteins','Biological Factors'). is_a('Chemokines','Chemotactic Factors'). is_a('Chemotactic Factors','Biological Factors'). is_a('Chemokines','Inflammation Mediators'). is_a('Inflammation Mediators','Biological Factors'). ..

3) GO Ontologies for each gene in the network	4) Logical facts inferred from 1), 2) and 3) and improved by the user.
leaf('ACE2','identical protein binding'). leaf('ACE2','virus receptor activity'). mf('virus receptor activity'). leaf('ACE2','zinc ion binding'). leaf('ACE2','angiotensin maturation'). bp('angiotensin maturation'). .. leaf('CXCR4','C-X-C chemokine receptor activity'). mf('C-X-C chemokine receptor activity'). leaf('CXCR4','C-X-C motif chemokine 12 receptor activity'). mf('C-X-C motif chemokine 12 receptor activity'). .. leaf('CXCR4','ubiquitin protein ligase binding'). leaf('CXCR4','virus receptor activity'). leaf('CXCR4','apoptotic process'). bp('apoptotic process'). ...	disease('COVID-19'). transcription_factor('STAT1'). protein('STAT1'). transcription_factor('STAT'). protein('STAT'). ligand('S'). protein('S'). receptor('CXCR4'). protein('CXCR4'). protein('JAK3'). receptor('CCR5'). protein('CCR5'). protein('ACE2'). enzyme('ACE2'). receptor('ACE2'). receptor('CD4'). protein('CD4'). ...

Figure 4 - KBs organized following the steps in Fig. 1 and Fig. 6 (Part 2)
DOI: https://doi.org/10.60797/jbg.2026.32.1.4

S ---> bind
CD4 ---> association
IFNG ---> association
ACE2 ---> positive_correlation ---> COVID-19

Pathway =
[
event('S',bind,'CD4'),
event('CD4',association,'IFNG'),
event('IFNG',association,'ACE2'),
event('ACE2',positive_correlation,'COVID-19')
] .

T cell assays reveal high frequencies of XBB.1.5 spike-specific CD4+ and CD8+ T cells amongst donors with hybrid immunity, with the CD4+ T cells skewed towards a Th1 cell phenotype and having attenuated effector cytokine secretion as compared to ancestral spike protein-specific cells.
PUBMED_ID: 38424106.
Here, we report that the CD4+ T cell/NK cell axis of gamma-herpesvirus control requires interferon-gamma engagement of CD11c+ dendritic cells
PUBMED_ID: 38578092.
Among these, seven significant cDEGs and proteins - namely, HRAS, IFNG, JUN, CDH1, TLR4, ICAM1, and SCD-were recognized as pivotal host factors linked to ACE2.
PUBMED_ID: 38640424.
In patients with COVID-19, epithelial cells showed an average three-fold increase in expression of the SARS-CoV-2 entry receptor ACE2, which correlated with interferon signals by immune cells.
PUBMED_ID: 32591762.

Figure 5 - An inferred pathway correlating ACE2 and COVID-19
DOI: https://doi.org/10.60797/jbg.2026.32.1.5

The Workflow to Build the Knowledge Bases

Table 1 - Subnetworks for the regulation of COVID-19 and CXCR4 (Part 1)
DOI: https://doi.org/10.60797/jbg.2026.32.1.6

a)	**b)**
Pathway 1: event('S',bind,'CD4'), event('CD4',association,'IFNG'), event('IFNG',association,'ACE2'), event('ACE2',positive_correlation,'COVID-19'). **event('S',bind,'CD4'):** T cell assays reveal high frequencies of XBB.1.5 spike-specific CD4+ and CD8+ T cells amongst donors with hybrid immunity, with the CD4+ T cells skewed towards a Th1 cell phenotype and having attenuated effector cytokine secretion as compared to ancestral spike protein-specific cells. PUBMED_ID: 38424106. e**vent('CD4',association,'IFNG'):** Splenic CD4+ and CD8+ T cells had increased expression of C-X-C Motif Chemokine Receptor 3 (CXCR3) and interferon-gamma (IFN-gamma), indicating a T helper 1 (Th1)-like/effector CD8+ T cell-like phenotype. PUBMED_ID: 38504977 e**vent('IFNG',association,'ACE2'):** Among these, seven significant cDEGs and proteins - namely, HRAS, IFNG, JUN, CDH1, TLR4, ICAM1, and SCD-were recognized as pivotal host factors linked to ACE2. PUBMED_ID: 38640424. **event('ACE2',positive_correlation,'COVID-19'):** In patients with COVID-19, epithelial cells showed an average three-fold increase in expression of the SARS-CoV-2 entry receptor ACE2, which correlated with interferon signals by immune cells. PUBMED_ID: 32591762. **Linking event to pathway 2:** event('COVID-19',positive_correlation,'IFNG'). **event('COVID-19',positive_correlation,'IFNG'):** The IL-6, interferon-gamma and endothelial growth factor were significantly higher in COVID-19 infected compared to non-infected individuals. PUBMED_ID: 37843354.	**Pathway 2:** event('IFNG',association,'ACE2'), event('ACE2',association,'CD4'), event('CD4',association,'STAT3'), event('STAT3',positive_correlation,'CXCR4'). **event('IFNG',association,'ACE2'):** Among these, seven significant cDEGs and proteins - namely, HRAS, IFNG, JUN, CDH1, TLR4, ICAM1, and SCD-were recognized as pivotal host factors linked to ACE2. PUBMED_ID: 38640424 **event('ACE2',association,'CD4'):** The expression of ACE2 was strongly positively correlated with the immune infiltration level of CD8+ T cell (r=0.184, p<0.001), CD4+ T cell (r=0.104, p=0.02) and neutrophils (r=0.101, p=0.02). PUBMED_ID: 34413267. **event('CD4',association,'STAT3'):** Rho-kinase inhibitor alleviates CD4+T cell mediated corneal graft rejection by modulating its STAT3 and STAT5 activation. PUBMED_ID: 38479724. We found that Y27632 significantly up-regulated the phosphorylation level of STAT5 in both spleen and lymph nodes, down-regulated the phosphorylation level of STAT3 in the CD4+ T cells in the spleen. PUBMED_ID: 38479724. **event('STAT3',positive_correlation,'CXCR4'):** The prenylflavonoid Icaritin enhances osteoblast proliferation and function by signal transducer and activator of transcription factor 3 (STAT-3) regulation of C-X-C chemokine receptor type 4 (CXCR4) expression. PUBMED_ID: 28863947. **Linking event to pathway 3:** event('CXCR4',bind,'CD4'). **event('CXCR4',bind,'CD4'):** We found that CXCR4 also interacts with the cytoplasmic domain of CD8alpha in a way that is similar to the CD4/CXCR4 interaction. PUBMED_ID: 16808956. **Linking event to pathway 4:** event('CXCR4',association,'JAK3'). **event('CXCR4',association,'JAK3'):** We have previously shown that Jak3 mediates CCR9 and CXCR4 signalling in response to CCL25 and CXCL12 in BM progenitors and thymocytes. PUBMED_ID: 17521370.

Figure 7 details the workflow that we follow to build the knowledge bases (KBs) shown in Figure 2, Figure 3 and Figure 4. Figure 3 and Figure 4 illustrates the following KBs:

1) names, synonyms, biological objects, and related scientific documents from the MeSH [12] service; Gene Ontology [13]; PubMed [14], Protein Data Bank (PDB) [15], [16], HGNC (HUGO Gene Nomenclature Committee) [17], and UniProt [18];

2) object identity, based on the the MeSH service;

3) molecular function, biological processes, and cellular components associated with each network object, from Gene Ontology;

4) definitions, which can be extracted via automatic inferences performed from 1, 2, and 3, establishing that an object satisfies the constraints guiding the analysis of pathways and subnetworks.

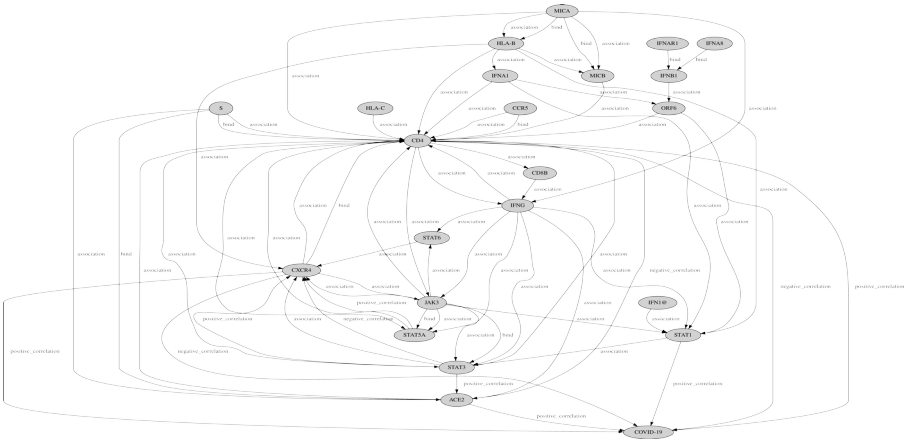

Figure 6 - A restricted GRN for the exploration of regulatory subnetworks for COVID-19 and CXCR4
DOI: https://doi.org/10.60797/jbg.2026.32.1.7

Table 2 - Subnetworks for the regulation of COVID-19 and CXCR4 (Part 2)
DOI: https://doi.org/10.60797/jbg.2026.32.1.8

c)	**d)**
Pathway 3: event('CD4',association,'IFNG'), event('IFNG',association,'STAT6'), event('STAT6',association,'CXCR4'), event('CXCR4',negative_correlation,'COVID-19'). **event('CD4',association,'IFNG'):** Splenic CD4+ and CD8+ T cells had increased expression of C-X-C Motif Chemokine Receptor 3 (CXCR3) and interferon-gamma (IFN-gamma), indicating a T helper 1 (Th1)-like/effector CD8+ T cell-like phenotype. PUBMED_ID: 38504977 **event('IFNG',association,'STAT6'):** Antigen-receptor engagement in B cells induces nuclear expression of STAT5 and STAT6 proteins that bind and transactivate an IFN-gamma activation site. PUBMED_ID: 8683142 **event('STAT6',association,'CXCR4'):** Stromal cell-derived factor (SDF)-1alpha, the ligand of CXCR4, stimulated the activation of JAK2/STAT3 and JAK3/STAT6 signalling in HKC-8 cells. PUBMED_ID: 32119183 **event('CXCR4',negative_correlation,'COVID-19'):** Interestingly, the expression of cell receptors, such as IFNGR1 and CXCR4, was reduced in response to the viral infection and associated with the inhibition of the related signaling pathways and immune functions. These results highlight novel immunoreceptors, selectively expressed in COVID-19 patients, which affect the immune functionality and are correlated with clinical outcomes. PUBMED_ID: 34944610	**Pathway 4:** event('JAK3',association,'STAT1'), event('STAT1',association,'STAT3'), event('STAT3',negative_correlation,'CXCR4'). **event('JAK3',association,'STAT1'):** Furthermore, an association between JAK3 and STAT-1, STAT-3, and STAT-5 activation and cell-cycle progression was demonstrated by both propidium iodide staining and bromodeoxyuridine incorporation in cells of four patients tested. PUBMED_ID: 9391124. **event('STAT1',association,'STAT3'):** STAT3 Regulates the Type I IFN-Mediated Antiviral Response by Interfering with the Nuclear Entry of STAT1. PUBMED_ID: 31575039 With further investigation, we found that importin alpha family member Karyopherin-alpha1 (KPNA1) was employed by both STAT1 and STAT3 for their nuclear import. PUBMED_ID: 31575039 The phosphorylated and un-phosphorylated STAT3 competed with STAT1 for binding to the decreased KPNA1 post infection and repressed downstream ISG expression. PUBMED_ID: 31575039 **event('STAT3',negative_correlation,'CXCR4'):** Mechanistically, the activation of signal transducer and activator of transcription 3(STAT3) signaling was impaired in CXCR4 CAR-T cells, thereof reduced the release of inflammatory factors, such as TNF-alpha, IL-6 and IL-17A. PUBMED_ID: 37735875. Expressing CXCR4 in CAR-T cells Suppresses MDSCs Recruitment via STAT3/NF-kappaB/SDF-1alpha axis to enhance Anti-tumor Efficacy against Pancreatic Cancer. PUBMED_ID: 37735875 **An alternative event to improve pathway 4:** **event('JAK3',association,'STAT3'):** We then show that constitutive activation of the JAK3/STAT3 pathway has a major role in NKCL cell growth and survival and in the invasive phenotype. PUBMED_ID: 23689514

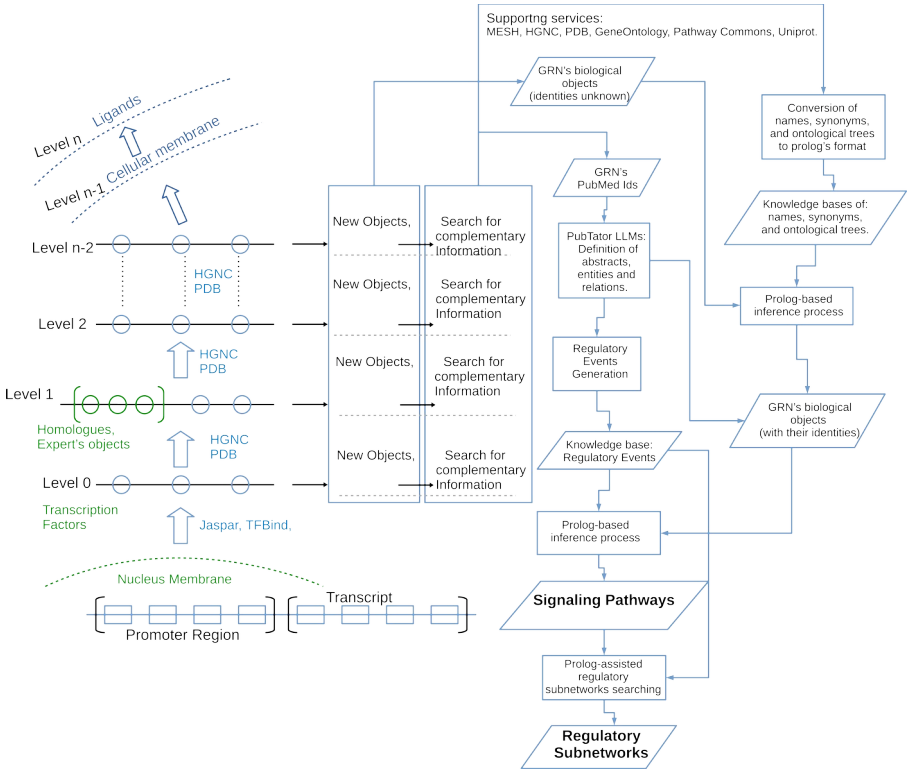

Figure 7 - Work-flow for the semantic modeling and analyzing of a GRN
DOI: https://doi.org/10.60797/jbg.2026.32.1.9

The following steps summarize the procedure for building knowledge bases and executing inference processes in order to model regulatory pathways and regulatory subnetworks:

1. Definition of potential transcription factors linked to the regulatory region of specific proteins of interest, based on TFBIND [19] and JASPAR [20].

2. Search for potential biological objects (e.g., ligands and proteins) linked to both the system-proposed transcription factors and other biological objects provided by the user, using the Protein Data Bank [15], [16] and UniProt.

3. Generation of knowledge bases describing the identity and functionality of network objects (including ligands, proteins, enzymes, and receptors), based on HGNC [17], UniProt, MeSH [12], and Gene Ontology [13].

4. Generation of keyword combinations among related objects in the network, considering not only the names defined by the modeler but also any synonyms defined for them.

5. Retrieval of PubMed IDs to subsequently obtain abstracts addressing the objects defined for the network, via PubMed [14].

6. Automatic generation of a corpus of abstracts and annotations using the NCBI PubTator service [21], utilizing the PubMed IDs defined in the previous step.

7. Automatic extraction of regulatory events associated with the Gene Regulatory Network (GRN) objects from the corpus of abstracts and annotations collected from PubTator. Generation of Prolog KBs.

8. Inference processes on Prolog code to discover regulatory pathways and subnetworks from the KBs modeled for the network.

Once the modeling levels depicted in Figure 7 are complete, we searched PubMed [14] for a set of abstracts related to each possible interaction between the objects in the initial GRN. To do this, the user-provided objects were organized into pairs, and PubMed was asked for a specified number of abstracts for each pair. In this case, we used a user-configurable threshold of 200 abstracts, as this number is usually sufficient to cover the most relevant and recent publications on each interaction in recent years. However, we could repeat the modeling process with a different threshold to obtain more information. Then, we use the services provided by the NCBI PubTator [21] service, to download the abstracts, accompanied by annotations, which include the recognition of the biological objects and the extraction of their relations. PubTator identifies all the objects in the abstracts, which normally includes objects not defined in the previous steps of the modeling process.

2.1 About Generative AI in our pipeline

There is a huge amount of biological and medical information and knowledge, available today through different bioinformatics resources (some of them mentioned above). On the other hand, other resources have recently emerged, based on Generative Artificial Intelligence (GenAI) and Large Language Models, LLMs, to extract biological entities and their relationships from any publication available in portals such as PubMed [14]. In this paper we use one of them, specifically PubTator [21], and explore how independent knowledge representations, associated with the different resources mentioned above, can assist and complement each other. In other words, we show how the entities and relations extracted from publications, using a collection of LLMs, can be connected to other KBs to facilitate the analyzing process of the knowledge gathered and modeled. The NCBI's PubTator service relies on a set of LLMs to implement a pipeline with three main tasks:

a) biological entities recognition, implemented with an LLM named AIONER [22];

b) a set of LLMs and tools, to normalize the names of biological entities, implemented as a tool named GNorm2 [23];

c) an LLM dedicated to the task of relation extraction from PubMed abstracts, implemented through the BioREx LLM [24].

2.2. Modeling COVID-19 Knowledge Domain

Figure 1 shows that the user participates directly in the first and last pipeline's stages. On the first stage, for instance, the user has access to a KB that she can check and improve the names and synonyms for each object. Those names and synonyms are critical later on when the pipeline defines the PubMed IDs, that the generative AI uses to assist the construction of the KB in Figure 2. On the other hand, at the pipeline's last stage, the user has access to all the prolog KBs that the system automatically produces. For instance, the user can add or modify the identity descriptors of an object; the system can infer (using prolog) that an object is just a protein, but the user could improve the description by writing down that the object is a receptor and an enzyme too. This kind of human intervention has an effect on the quality of the pipeline's results. The human-in-the-loop approach is a fundamental feature in our proposal; and we reinforced such quality, by offering the user the facility of re-executing the pipeline, from some point onwards, once she has modified a particular KB.

Inferring Regulatory Pathways and Regulatory Subnetworks

In our work, a set of constraints define the characteristics of the objects that will shape regulatory pathways. For instance, the regulatory pathways could end with certain types of transcription factors, or to begin with certain types of drugs, or to include only objects with specific molecular functions and biological processes. Using the constraints, the system can explore the knowledge bases described in Figure 3 and Figure 4, searching for the objects that satisfy the desired characteristics, and then, shape the objects' definitions that will be used later when inferring the pathways (see section 4, Figure 4). In this scenario, a regulatory pathway is a causal chain of regulatory events such that their biological objects satisfy certain kinds of constraints regarding their biological identities, molecular functions, biological processes, and the ways they are related (bind, stimulation, inhibition, and so on).

Our system keeps track of and allows access to all the original documents used as sources of knowledge. It is possible, for instance, to choose an inferred regulatory pathway and retrieve the abstracts that support it (see Figure 5). A GRN, in our work, is the collection of regulatory pathways that it is possible to infer from the KBs modeled, and that fulfill the set of user's constraints for the biology system (and problem) on consideration. Figure 6 is an instance of a GRN that satisfies the description formulated before. In the pathways shown in Figure 6, each pathway begins either with a ligand, receptor, or transcription factor, that binds to or associates with a protein. Subsequently, a chain of protein interactions shapes the pathways until an ending regulatory event is inferred, in which a receptor, or transcription factor, interacts either to an ending protein (ACE2, CXCR4) or to the COVID-19 disease. In Biopatterngs's pipeline the logic AI stage offers ways to configure the set of constraints that guide the way the pathways are searched.

Figure 6 shows that there is no interest (at that moment) on small molecules or drugs working in the network; but that can be easily changed by incorporating them and their roles at inference time (see RESTRICTED/covid-19-restricted-list-and-small-molecules-from-pubtator.png in supplemental material). Figure 6 also shows that the objects visible in it are limited to the list of the proteins initially delivered to the system (see Figure 1). However, this can also be configured to allow the incorporation during inference time of other gene/gene products, small molecules, or diseases, provided by the generative AI stage of the pipeline (instances of these in RESTRICTED/covid-19-genes-from-pubtator.png and RESTRICTED/covid-19-diseases-from-pubtator.png in supplemental material).

Figure 6 allows one to have a first look at biological interactions and possible pathways, and then choose some of particular interest. Suppose that we choose from Figure 6 the following set of objects: the Spike (S) protein and CD4 receptor, the interferon protein (IFNG), the receptors STAT3, STAT5A and ACE2, the enzyme JAK3 and, of course, COVID-19. We can also see in Figure 6 a negative correlation between the CXCR4 receptor and the CODID-19 disease. Therefore, we can decide to search for subnetworks showing CXCR4, mediating a shift in the regulatory process of the disease. Then, the following question can be formulated: does the regulatory network in Figure 6, includes any subnetworks branching off from the regulatory pathway initially chosen (the one in Figure 5), such that the COVID-19 regulation shifts from up-regulated to down-regulated, mediated by the receptor CXCR4?. The process to answer the question above goes as follows:

1. The searching process is restricted to a list of objects: Spike (S), CD4, IFNG, STAT3, STAT5A, ACE2, JAK3, COVID-19, and CXCR4; therefore, we ask to the Biopatternsg system to restrict the knowledge base of regulatory events to those objects, and then to execute the inference of possible pathways related with them. This step produces the new network depicted in Figure 8 (for a detailed view see VERY-RESTRICTED/subnetworks/covid-19-CXCR4-wide-subnetwork.png in supplemental material).

2. We provide to the system the pathway that must guide the searching of regulatory subnetworks for COVID-19 (the one in Figure 5).

3. Biopatternsg consults the knowledge base of regulatory events, looking for events that might serve as regulatory links, between the initial pathway and possible others that could change the regulatory state of the disease.

4. Once the system has produced a collection of candidate subnetworks and we make a revision of them, a three pathways subnetwork is selected for COVID-19; for instance, the one described in Table 1 (a and b) and Table 2 (c) .

5. Normally we want to know a way to regulate the protein that mediates the regulatory shifting in the initial regulatory process. Therefore, we run another search for a regulatory subnetwork for CXCR4, and after a revision, we select the subnetwork described in Table 1 (b) and Table 2 (d). Note that the second pathway mediating the regulation shift for COVID-19 visible in Table 1 (b), is used as the pathway that guides the searching for a regulatory subnetwork for CXCR4.

6. Finally, the subnetworks shown in Table 1 and Table 2 are documented using the documentation that the system provides for the knowledge base of regulatory events. Note that several sentences can explain a regulatory event and that not all of them correctly model the related event, but normally one of them does.

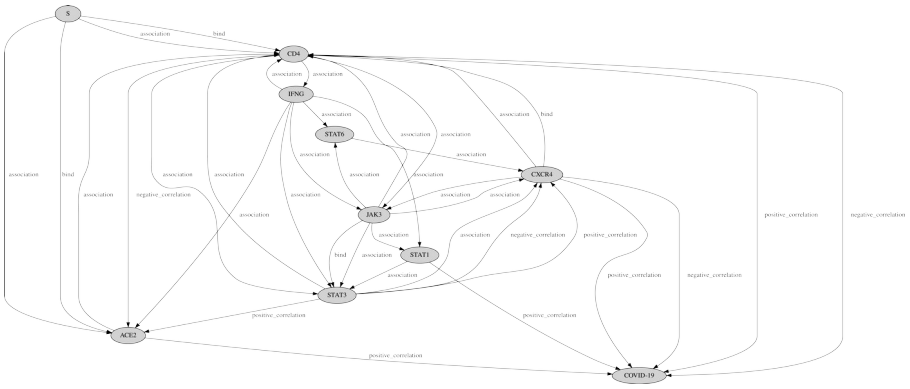

Figure 8 - Restricted GRN to explore subnetworks for COVID-19 and CXCR4
DOI: https://doi.org/10.60797/jbg.2026.32.1.10

The procedure above leads to hypothetical subnetworks but it is possible to go further. Table 2(d) shows the Pathway 4, which regulates the regulatory shift in the regulation of the receptor CXCR4. Pathway 4 includes the interaction event('STAT1', association, 'STAT3'), accompanied by several sentences that support it. Upon studying the nature of the aforementioned interaction, it can be noted that it does not actually contribute with a significant step, in the context of the CXCR4 receptor regulation. One might then ask: does a regulatory event exist in the knowledge base kBase.pl, that may directly link the JAK3 enzyme with the transcription factor STAT3?. And the answer is yes. In fact, when we go to the documented knowledge base we found out that a whole pathway exists, relating the interactions of JAK3 and STAT3. Therefore, the interaction event('STAT1', association, 'STAT3') can be omitted in Pathway 4, and the user can replace it by the interaction event('JAK3', association, 'STAT3') (see bottom side of Table 2(d)).

Biopatternsg also offer resources to improve the list of objects initially provided as input to the system. For instance, once the first two stages of the Biopatternsg's pipeline has finished, a report (named aligned.pl), is produced indicating which of the initially provided objects are already part of the knowledge base of regulatory events (KBase.pl), and which ones of them could be part of it, but represented using synonyms. Another report is also available (named synonyms.pl), regarding the main name of an object in the knowledge base and its synonyms; that report is produced using the metadata that accompanies the NER predictions made by PubTator's LLMs. The report aligned.pl also shows which of the initial objects are not initially part of the knowledge base, and using both reports, aligned.pl and synonyms.pl, and the documentation of the knowledge base (named kBaseDoc.txt), it is possible also to explore alternative names for them (see examples of the files mentioned above in EVALUATION/CREB-phosphorylation). The result of all this is a better list of objects aligned to the names that the generative AI actually uses, for the objects that the researcher provided initially. Once a better aligned list has been elaborated, by means of the knowledge of who models the GRN, and the resources that Biopatternsg provides, then a new inference process can be run and an improved set of pathways and subnetworks can be obtained. This defines an iterative process of modeling and analysing that eventually could lead to important findings (see the researcher's guide in DOC and the Biopatternsg's repository wiki for a deeper description of its functionalities).

A framework for the evaluation of the method to produce GRNs

We have organized a systematic evaluation of biopatternsg as a method to produce gene regulatory networks, GRN. By a GRN we exactly mean the collection of objects and events reported in kBase.pl and it is the one we use here to evaluate Biopatternsg with the standard analysis of precision and recall (the F1 measures [25]).

In order to achieve this goal we retrieved, from Pathways Commons [26], a set of knowledge bases that serve as golden references to compare with. Pathways Commons offers facilities to export pathways in formats like sif (simple interaction format) and that allows us to create knowledge bases as the one in Table 7. We have compared the golden standard, or reference, from Pathways Common with the output of our system in three levels:

Level 1

At the first level, we checked whether a set of biological objects of interest, named in the sif files of a given GRN from Pathways Commons, can be identified by the pipeline's tools. The key comparison is whether each name from each network used as reference is recognised by the AIONER LLM from Pubtator, the tool used by the pipeline.

We define as true positive, TP, as those names that are recognized by the tool and, of course, false negative, FN, those that fail to be recognized. The tool also suggests, in some cases, synonyms for given names which can be later verified to correspond or not to the original name in the GRN. So, those synonyms are a pessimistic estimate of false positives, FP, when they end up not corresponding to the original name. The usual definitions of Precision (TP/TP+FP) and Recall (TP/TP+FN) then hold. Table 3 lists 30 GRN provided by Pathways Commons and summarizes the results that we obtain for them.

The evaluation reports an average F1-Score of 0.9069 with a variance of 0.0145 among the 30 networks. This indicates an excellent capacity to recognize names of real, biological objects. It is worth noticing, also, that this F1 score is a minimum, as it could be higher when synonyms offered by the pipeline can be confirmed to be real names too.

We want to call the attention upon the column to the left of Table 3, sm/obs, which indicates the fraction of the number of names for small molecules (column sm, in the middle) divided by the amount of all the names in the GRN (column objects) provided as input to the pipeline. Notice that, as this fraction gets smaller, the F1 Score increases. We are actually using that column as an ordering key for the rest of the information in Table 3, a strategy that will prove to be useful to explain the results below at the other levels.

```
base([
event(MAPKAPK2, controls-phosphorylation-of, CREB1),
event(MAPKAPK2, controls-state-change-of, CREB1)
event(RPS6KA1, controls-phosphorylation-of, CREB1),
event(RPS6KA1, controls-state-change-of, CREB1)
event(RPS6KA2, controls-phosphorylation-of, CREB1),
event(RPS6KA2, controls-state-change-of, CREB1)
event(RPS6KA3, controls-phosphorylation-of, CREB1),
event(RPS6KA3, controls-state-change-of, CREB1)
event(RPS6KA5, controls-phosphorylation-of, ATF1),
event(RPS6KA5, controls-state-change-of, ATF1)
event(RPS6KA5, controls-phosphorylation-of, CREB1),
event(RPS6KA5, controls-state-change-of, CREB1)
]).
```

Figure 9 - Pathways Commons sif file of the reactome's pathway named creb phosphorylation (prolog format)
DOI: https://doi.org/10.60797/jbg.2026.32.1.11

Level 2

At the second level, the system is provided with pairs (object1, object2) as references, indicating relations in the GRN without actually naming the relations (a process confirmed to be noisy, as shown below).

We selected the referential GRN from Pathways Commons, of any size and with and without small molecules, to test PubTator's NER [22] and Normalization [23] functions in those conditions. The procedure to build a knowledge base of regulatory events for each network/pathway in our experiments is as follows:

1) the sif (for comparison later) and the extended version of the sif file are downloaded from Pathways Commons;

2) using the extended sif file we collect the objects interacting in a network and, using the pipeline of Biopatternsg, a knowledge base of regulatory events is modeled for it;

3) the network's sif file is compared with the corresponding knowledge base produced by Biopatternsg. Step (3) is automated by a Prolog script available from the repo (see EVALUATION/README.txt and graph-comparison.pl in supplemental material). The data used for the comparison is also available in the repository (see EVALUATION).

Table 2 - Evaluation of names recognition in biopatternsg (level 1)
DOI: https://doi.org/10.60797/jbg.2026.32.1.12

Pathway	objects	sm	TP	FP	FN	Precision	Recall	F1	sm/objs
Pentose-phosphate-cycle-(Pentose-phosphate-cycle)	31	19	14	0	17	1,0000	0,4516	0,6222	0,6129
Citric-acid-cycle-(TCA-cycle)	43	26	19	2	22	0,9048	0,4634	0,6129	0,6047
Mitochondrial-fatty-acid-beta-oxidation	35	16	19	1	15	0,9500	0,5588	0,7037	0,4571
Glycolysis	53	22	32	3	18	0,9143	0,6400	0,7529	0,4151
SARS-COV-2-Maturation-of-spike-protein	28	10	18	0	10	1,0000	0,6429	0,7826	0,3571
SARS-COV-2-ATTACHMENT-ENTRY	35	12	24	0	11	1,0000	0,6857	0,8136	0,3429
superpathway-of-D-myo-inositol-1-4-5-trisphosphate-metabolism	30	10	19	1	10	0,9500	0,6552	0,7755	0,3333
Interleukin-2-signaling	17	5	12	0	5	1,0000	0,7059	0,8276	0,2941
NFE2L2-regulates-pentose-phosphate-pathway-genes	7	2	5	0	2	1,0000	0,7143	0,8333	0,2857
SARS-CoV-1-activates-modulates-innate-immune-responses	22	3	22	0	0	1,0000	1,0000	1,0000	0,1364
Paradoxical-activation-of-RAF-signaling-by-kinase-inactive-BRAF	48	5	42	1	5	0,9767	0,8936	0,9333	0,1042
Interferon-alpha-beta-signaling	46	4	36	0	10	1,0000	0,7826	0,8780	0,0870
Signaling-downstream-of-RAS-mutants	47	4	42	0	5	1,0000	0,8936	0,9438	0,0851
HDACs-deacetylate-histones	101	7	75	1	25	0,9868	0,7500	0,8523	0,0693
Selenocysteine-synthesis	103	7	95	1	7	0,9896	0,9314	0,9596	0,0680
Negative-regulation-of-FGFR1-signaling	27	1	26	1	0	0,9630	1,0000	0,9811	0,0370
SARS-COV-2-modulates-host-translation	52	1	51	0	1	1,0000	0,9808	0,9903	0,0192
Regulation_of_telomerase	71	1	70	1	0	0,9859	1,0000	0,9929	0,0141
CREB-phosphorylation	7	0	7	0	0	1,0000	1,0000	1,0000	0,0000
ERKs-are-inactivated	13	0	13	0	0	1,0000	1,0000	1,0000	0,0000
IFNG-signaling-activates-MAPKs	8	0	8	0	0	1,0000	1,0000	1,0000	0,0000
Negative-regulation-of-MAPK	37	0	36	0	1	1,0000	0,9730	0,9863	0,0000
Regulation-of-NFE2L2-gene-expression	8	0	8	0	0	1,0000	1,0000	1,0000	0,0000
SARS-CoV-1-modulates-host-translation-machinery	35	0	35	0	0	1,0000	1,0000	1,0000	0,0000
SARS-COV-2-AUTOPHAGY	10	0	10	0	0	1,0000	1,0000	1,0000	0,0000
SARS-COV-2-Translation-of-Replicase-and-Assembly-of-the-Replication-Tra	13	0	13	0	0	1,0000	1,0000	1,0000	0,0000
Signalling-to-RAS	10	0	10	0	0	1,0000	1,0000	1,0000	0,0000
Spry-regulation-of-FGF-signaling	16	0	16	0	0	1,0000	1,0000	1,0000	0,0000
TRAF3-dependent-IRF-activation-pathway	15	0	14	1	0	0,9333	1,0000	0,9655	0,0000
Trans-Golgi	7	0	7	0	0	1,0000	1,0000	1,0000	0,0000
						0,9851	0,8574	0,9069	
						0,0007	0,0329	0,0145	

The statistics are defined as follows: if a pair (object1, object2) also appears in the output of the pipeline, it is counted as a true positive (TP). If the pair appears in the reference, but not in our output, is a false negative (FN). And if it does not appear

in the reference, but our system produces it, it is a false positive (FP). Table 4 lists the 30 GRN provided by Pathways Commons and summarizes the results of comparing pairs (object1, object2) that we obtain for them.

The results in Table 4 are discussed in the following section. The Average F1-Score is 0.3041 with a variance of 0.04 (Table 4 also reports an alternative F1-Score of 0.3652 obtained by adding up all TP, FP and FN from the 30 networks). The discrepancy may be due to the fact that the networks are intrinsically different and cannot be considered as one. For the CREB phosphorylation pathway, the aligned.pl file (see EVALUATION/CREB-phosphorylation) shows that all the objects visible in the reference are present in the knowledge base. It is a case of perfect recall.

Table 4 and Table 5 show the size of the knowledge bases for both reference and model, highlighting their high variability. The tables mentioned above highlight the cases where the reference size is larger and vice versa, and we can see that the cases are quite balanced.

Table 3 - Evaluation of pairs-like relations generation in biopatternsg (level 2)
DOI: https://doi.org/10.60797/jbg.2026.32.1.13

Ruta	TP	FP	FN	Precision	Recall	F1	sm/objs	sm	objects	ref.size	model size
Pentose-phosphate-cycle-(Pentose-phosphate-cycle)	3	38	115	0.0732	0.0254	0.0377	0.6129	19	31	51	27
Citric-acid-cycle-(TCA-cycle)	16	78	138	0.1702	0.1039	0.1290	0.6047	26	43	145	67
Mitochondrial-fatty-acid-beta-oxidation	1	47	103	0.0208	0.0096	0.0132	0.4571	16	35	104	29
Glycolysis	16	290	187	0.0523	0.0788	0.0629	0.4151	22	53	202	191
SARS-COV-2-Maturation-of-spike-protein	15	22	66	0.4054	0.1852	0.2542	0.3571	10	28	74	23
SARS-COV-2-ATTACHMENT-ENTRY	30	62	311	0.3261	0.0880	0.1386	0.3429	12	35	329	52
superpathway-of-D-myo-inositol-1-4-5-trisphosphate-metabolism	5	66	133	0.0704	0.0362	0.0478	0.3333	10	30	137	54
Interleukin-2-signaling	188	142	16	0.5697	0.9216	0.7041	0.2941	5	17	86	167
NFE2L2-regulates-pentose-phosphate-pathway-genes	10	22	4	0.3125	0.7143	0.4348	0.2857	2	7	9	23
SARS-CoV-1-activates-modulates-innate-immune-responses	33	422	5	0.0725	0.8684	0.1339	0.1364	3	22	21	349
Paradoxical-activation-of-RAF-signaling-by-kinase-inactive-BRAF	509	586	558	0.4648	0.4770	0.4709	0.1042	5	48	834	564
Interferon-alpha-beta-signaling	369	448	636	0.4517	0.3672	0.4050	0.0870	4	46	787	516
Signaling-downstream-of-RAS-mutants	415	448	584	0.4809	0.4154	0.4458	0.0851	4	47	830	468
HDACs-deacetylate-histones	155	448	3978	0.2570	0.0375	0.0655	0.0693	7	101	4089	385
Selenocysteine-synthesis	172	263	4113	0.3954	0.0401	0.0729	0.0680	7	103	4277	227
Negative-regulation-of-FGFR1-signaling	217	681	133	0.2416	0.6200	0.3478	0.0370	1	27	217	565
SARS-COV-2-modulates-host-translation	72	105	667	0.4068	0.0974	0.1572	0.0192	1	52	726	96
Regulation_of_telomerase	273	2766	95	0.0898	0.7418	0.1603	0.0141	1	71	229	2246
CREB-phosphorylation	18	41	2	0.3051	0.9000	0.4557	0.0000	0	7	12	39
ERKs-are-inactivated	65	94	36	0.4088	0.6436	0.5000	0.0000	0	13	69	110
IFNG-signaling-activates-MAPKs	49	69	15	0.4153	0.7656	0.5385	0.0000	0	8	44	57
Negative-regulation-of-MAPK	84	195	165	0.3011	0.3373	0.3182	0.0000	0	37	213	191
Regulation-of-NFE2L2-gene-expression	35	156	1	0.1832	0.9722	0.3084	0.0000	0	8	13	159
SARS-CoV-1-modulates-host-translation-machinery	25	34	567	0.4237	0.0422	0.0768	0.0000	0	35	591	29
SARS-COV-2-AUTOPHAGY	18	28	15	0.3913	0.5455	0.4557	0.0000	0	10	29	21
SARS-COV-2-Translation-of-Replicase-and-Assembly-of-the-Replication-Transcription	40	39	30	0.5063	0.5714	0.5369	0.0000	0	13	54	53
Signaling-to-RAS	39	110	13	0.2617	0.7500	0.3881	0.0000	0	10	35	102
Spry-regulation-of-FGF-signaling	31	107	27	0.2246	0.5345	0.3163	0.0000	0	16	44	102
TRAF3-dependent-IRF-activation-pathway	138	184	29	0.4286	0.8263	0.5644	0.0000	0	15	86	233
Trans-Golgi	32	46	0	0.4103	1.0000	0.5818	0.0000	0	7	15	34
	3073	8037	12742	0.3040	0.4572	0.3041					
				0.0238	0.1176	0.3652					
						0.0399					

Our scripts for the automation of the evaluation process depends on report_alignments.txt, the file that offers the metrics described in Table 3; report.txt, the file that offers the metrics described in Table 4 (see EVALUATION/CREB-phosphorylation); and report_simple.txt, the file that offers the metrics described in Table 5 (see Figure 10 below for an example of report.txt).

```
Report on KBase.sif vs  kBase.pl
True Positives (the ref has them and the system predicts them): 18
False Positives (the ref does not have them but the system predicts them): 41
False Negatives (the ref has them but the system does not predict them): 2
Precision (TP/(TP+FP)), from all the predictions, how many are correct: 0.3050847457627119
Recall (TP/(TP+FN)): from all the correct ones, how many are predicted: 0.9
F1 Score (2*Precision*Recall)/(Precision+Recall): 0.45569620253164556

18    41    2    0.3050847457627119    0.9    0.45569620253164556
```

Figure 10 - An instance of report.txt for CREB-phosphorylation
DOI: https://doi.org/10.60797/jbg.2026.32.1.14

Level 3

At level 3 of the evaluation, we combine all the information produced by the pipeline to describe the GRN. At the level, the evaluation checks whether a triple (object 1, relation, object 2) appears in the reference. Unfortunately, the naming of relations differs between the golden standard and our system, so that we have to appeal to a series of synonyms, also produced by our system. Therefore, on the pipeline's output we have something like (object 1, another name for the relation, object 2).

In any case, if a triple also appears in our output, it is counted as a true positive (TP). If the triple appears in the reference, but not in our output, is a false negative (FN). And if it does not appear in the reference, but our system produces it, it is a false positive (FP). Precision and Recall are calculated accordingly and as before. Table 5 lists the 30 GRN provided by Pathways Commons and summarizes the results of extracting the relation between pairs of biological objects.

The results overall are worse than at level 2. The F1-Score is 0.2306 with a variance of 0.0218, sensibly lower than in Table 4. The system exhibits perfect Recall (at 1) at more networks (4 in Table 5 vs. 1 in Table 4), but Precision drops even more and the distance between the two measures increases (and that reflects in that lower F1 Score).

We discuss these results in the following section and explain why even that partial success with Recall is an encouraging result.

Table 4 - Evaluation of relations extraction in biopatternsg (level 3)
DOI: https://doi.org/10.60797/jbg.2026.32.1.15

Ruta	TP	FP	FN	Precision	Recall	F1	sm/objs	sm	objects	ref.size	model size
Pentose-phosphate-cycle-(Pentose-phosphate-cycle)	5	94	273	0.0505	0.0180	0.0265	0.6129	19	31	51	27
Citric-acid-cycle-(TCA-cycle)	14	261	300	0.0509	0.0446	0.0478	0.6047	26	43	145	67
Mitochondrial-fatty-acid-beta-oxidation	3	158	248	0.0186	0.0120	0.0146	0.4571	16	35	104	29
Glycolysis	42	1822	350	0.0225	0.1071	0.0372	0.4151	22	53	202	191
SARS-COV-2-Maturation-of-spike-protein	13	23	115	0.3611	0.1016	0.1585	0.3571	10	28	74	23
SARS-COV-2-ATTACHMENT-ENTRY	22	96	270	0.1864	0.0753	0.1073	0.3429	12	35	329	52
superpathway-of-D-myo-inositol-1-4-5-trisphosphate-metabolism	18	179	201	0.0914	0.0822	0.0865	0.3333	10	30	137	54
Interleukin-2-signaling	51	105	20	0.3269	0.7183	0.4493	0.2941	5	17	86	167
NFE2L2-regulates-pentose-phosphate-pathway-genes	5	15	8	0.2500	0.3846	0.3030	0.2857	2	7	9	23
SARS-CoV-1-activates-modulates-innate-immune-responses	17	789	6	0.0211	0.7391	0.0410	0.1364	3	22	21	349
Paradoxical-activation-of-RAF-signaling-by-kinase-inactive-BRAF	476	1638	324	0.2252	0.5950	0.3267	0.1042	5	48	834	564
Interferon-alpha-beta-signaling	148	522	466	0.2209	0.2410	0.2305	0.0870	4	46	787	516
Signaling-downstream-of-RAS-mutants	433	1306	305	0.2490	0.5867	0.3496	0.0851	4	47	830	468
HDACs-deacetylate-histones	430	1712	4189	0.2007	0.0931	0.1272	0.0693	7	101	4089	385
Selenocysteine-synthesis	798	1260	3590	0.3878	0.1819	0.2476	0.0680	7	103	4277	227
Negative-regulation-of-FGFR1-signaling	161	2060	75	0.0725	0.6822	0.1311	0.0370	1	27	217	565
SARS-COV-2-modulates-host-translation	170	255	572	0.4000	0.2291	0.2913	0.0192	1	52	726	96
Regulation_of_telomerase	190	10230	23	0.0182	0.8920	0.0357	0.0141	1	71	229	2246
CREB-phosphorylation	6	50	0	0.1071	1.0000	0.1935	0.0000	0	7	12	39
ERKs-are-inactivated	34	120	5	0.2208	0.8718	0.3523	0.0000	0	13	69	110
IFNG-signaling-activates-MAPKs	27	81	0	0.2500	1.0000	0.4000	0.0000	0	8	44	57
Negative-regulation-of-MAPK	65	399	107	0.1401	0.3779	0.2044	0.0000	0	37	213	191
Regulation-of-NFE2L2-gene-expression	10	80	0	0.1111	1.0000	0.2000	0.0000	0	8	13	159
SARS-CoV-1-modulates-host-translation-machinery	83	125	508	0.3990	0.1404	0.2078	0.0000	0	35	591	29
SARS-COV-2-AUTOPHAGY	22	50	7	0.3056	0.7586	0.4356	0.0000	0	10	29	21
SARS-COV-2-Translation-of-Replicase-and-Assembly-of-the-Replication-Transcription	31	39	11	0.4429	0.7381	0.5536	0.0000	0	13	54	53
Signalling-to-RAS	29	103	0	0.2197	1.0000	0.3602	0.0000	0	10	35	102
Spry-regulation-of-FGF-signaling	23	180	13	0.1133	0.6389	0.1925	0.0000	0	16	44	102
TRAF3-dependent-IRF-activation-pathway	63	131	14	0.3247	0.8182	0.4649	0.0000	0	15	86	233
Trans-Golgi	8	31	0	0.2051	1.0000	0.3404	0.0000	0	7	15	34
	3397	23914	12000	0.1998	0.5043	0.2306					
				0.0161	0.1284	0.0218					

Results and Discussion

Tables 4 and 5 show low averages for both the metrics (precision and recall). It is clear that, under the operational constraint of 200 abstracts retrieved for the model, the predictive capacity of Biopatternsg for relations is, on average, very limited. Precision is very low and gets worse when the name of the relation is actually compared (at level 3). However, Recall is perfect in more cases and increases between level 2 and level 3. In fact, in 10 out of 30 cases is a good Recall (above .70).

Low precision, however, requires special attention. The pipeline produces too many instances of events/relations that are not explicitly listed in the reference. A first hypothesis is that the reason for this may be the LLM (BioRex) behind PubTator has access to a wealth of documental references considerably bigger than the one used to sustain the GRNs reported by Pathways Commons. Our system may be reporting relations from other publications not considered by the Pathways Commons studies. Another hypothesis is that this low precision may be an effect of the different provenances for the data involved, which could also be affecting the Recall. The references come from actual GRN reported in Pathways Commons, whereas the data to train the model comes from a biomedical relation extraction dataset (BioRED) with relation pairs (e.g. gene–disease; chemical–chemical) at the document level, on a set of 600 PubMed abstracts, in which each relation has been labelled as describing either a novel finding or previously known background knowledge. BioRED has been assessed by benchmarking several existing state-of-the-art methods and results show that there is much room for improvement for the relation extraction task (*F*-score of 47.7%).

Regarding the high variability in the sizes of the knowledge bases, we want to highlight situations like this: when the reference size is larger than the size of the model, many of its events may be covered by events in the model with a relationship (such as association), that covers up a broad spectrum of possible interactions in the reference (like in-complex-with, controls-production-of, used-to-produce, among others) (see EVALUATION/Selenocysteine-synthesis). In such a situation, an event with a general relationship such as association will only be counted once, leaving other implicated events as false negatives, which negatively impacts the recall.

Nevertheless, achieving high recall under those conditions is particularly encouraging because it indicates that the system is effectively recovering, from a possibly considerably bigger collection of documents, information that coincides with well-curated, state of the art, reports of GRN. Pathways commons does not contain a manually adjusted set of positive examples, prepared to train an LLM, but human-oriented reports of conserved truths in biology. Therefore, we believe that these levels of recall represent a deeper confirmation of the effectiveness of the AI tools.

BioREx was trained with the summaries of papers, and that means that when PubTator predicts a relation between two biological objects, the LLM takes the whole abstract of the paper as the text that explains the relation. BioREx depends on PubMedBERT [27], a PreTrained Language Model (PLM) specialized on the PubMed abstracts. BERT [28] based LLMs, are very good at learning the context that relates the entities in a sentence. We believe the BioREx's false positive rate could be reduced by training the BioREx LLM at the sentence level, instead of the abstract level. The problem with such a change is the need for a training corpus properly tagged. We are taking steps in that direction regarding the NER and Normalization tasks. PubTator offers a very handy service as it offers in one place the normalization, ner and relation extractions tasks.

The PubTator's normalization of names does not always match the names of the objects provided as input to the system and that is normal; but this is more frequent in the case of small molecules than for genes and proteins as the metrics reveal. Table 6, Table 7, Table 8 and Table 9, show the metrics values for the experiments presented in Table 4 and 5, differentiated by whether or not they include small molecules, and they show a small difference when the reference does not include the last ones. We can say, however, that regardless of the case, the names of the objects of interest are generally not fully detected by PubTator and this, as expected, negatively impacts the quality of our models.

Table 5 - Evaluation of pairs-like relations generation in Biopatternsg with small molecules
DOI: https://doi.org/10.60797/jbg.2026.32.1.16

Ruta	TP	FP	FN	Precision	Recall	F1	sm/objs
Pentose-phosphate-cycle-(Pentose-phosphate-cycle)	3	38	115	0.0732	0.0254	0.0377	0.6129
Citric-acid-cycle-(TCA-cycle)	16	78	138	0.1702	0.1039	0.1290	0.6047
Mitochondrial-fatty-acid-beta-oxidation	1	47	103	0.0208	0.0096	0.0132	0.4571
Glycolysis	16	290	187	0.0523	0.0788	0.0629	0.4151
SARS-COV-2-Maturation-of-spike-protein	15	22	66	0.4054	0.1852	0.2542	0.3571
SARS-COV-2-ATTACHMENT-ENTRY	30	62	311	0.3261	0.0880	0.1386	0.3429
superpathway-of-D-myo-inositol-1-4-5-trisphosphate-metabolism	5	66	133	0.0704	0.0362	0.0478	0.3333
Interleukin-2-signaling	188	142	16	0.5697	0.9216	0.7041	0.2941
NFE2L2-regulates-pentose-phosphate-pathway-genes	10	22	4	0.3125	0.7143	0.4348	0.2857
SARS-CoV-1-activates-modulates-innate-immune-responses	33	422	5	0.0725	0.8684	0.1339	0.1364
Paradoxical-activation-of-RAF-signaling-by-kinase-inactive-BRAF	509	586	558	0.4648	0.4770	0.4709	0.1042
Interferon-alpha-beta-signaling	369	448	636	0.4517	0.3672	0.4050	0.0870
Signaling-downstream-of-RAS-mutants	415	448	584	0.4809	0.4154	0.4458	0.0851
HDACs-deacetylate-histones	155	448	3978	0.2570	0.0375	0.0655	0.0693
Selenocysteine-synthesis	172	263	4113	0.3954	0.0401	0.0729	0.0680
Negative-regulation-of-FGFR1-signaling	217	681	133	0.2416	0.6200	0.3478	0.0370
SARS-COV-2-modulates-host-translation	72	105	667	0.4068	0.0974	0.1572	0.0192
Regulation_of_telomerase	273	2766	95	0.0898	0.7418	0.1603	0.0141
				0.2701	0.3238	0.2267	
				0.0314	0.1041	0.0383	

Table 6 - Evaluation of pairs-like relations generation in Biopatternsg without small molecules
DOI: https://doi.org/10.60797/jbg.2026.32.1.17

Ruta	TP	FP	FN	Precision	Recall	F1	sm/objs
CREB-phosphorylation	18	41	2	0.3051	0.9000	0.4557	0.0000
ERKs-are-inactivated	65	94	36	0.4088	0.6436	0.5000	0.0000
IFNG-signaling-activates-MAPKs	49	69	15	0.4153	0.7656	0.5385	0.0000
Negative-regulation-of-MAPK	84	195	165	0.3011	0.3373	0.3182	0.0000
Regulation-of-NFE2L2-gene-expression	35	156	1	0.1832	0.9722	0.3084	0.0000
SARS-CoV-1-modulates-host-translation-machinery	25	34	567	0.4237	0.0422	0.0768	0.0000
SARS-COV-2-AUTOPHAGY	18	28	15	0.3913	0.5455	0.4557	0.0000
SARS-COV-2-Translation-of-Replicase-and-Assembly-of-the-Replication-Transcription	40	39	30	0.5063	0.5714	0.5369	0.0000
Signalling-to-RAS	39	110	13	0.2617	0.7500	0.3881	0.0000
Spry-regulation-of-FGF-signaling	31	107	27	0.2246	0.5345	0.3163	0.0000
TRAF3-dependent-IRF-activation-pathway	138	184	29	0.4286	0.8263	0.5644	0.0000
Trans-Golgi	32	46	0	0.4103	1.0000	0.5818	0.0000
				0.3550	0.6574	0.4201	
				0.0095	0.0764	0.0215	

PubTator's AIONER LLM currently proposes identities for the objects too, but those identities are still very general; only categories such as small molecules, gene/gene product, and disease are available. Therefore, services such as MeSH must still be used to define the identity of a biological object (enzyme, receptor, transcription factor, and so on), and since the identity tree provided by MeSH is expressed as an ontology that describes the different roles that an object can have, it results critical for the tasks related with the three ways of constraints that we use in the logic AI stage of our pipeline.

Table 7 - Evaluation of relations extraction in Biopatternsg with small molecules
DOI: https://doi.org/10.60797/jbg.2026.32.1.18

Ruta	TP	FP	FN	Precision	Recall	F1	sm/objs
Pentose-phosphate-cycle-(Pentose-phosphate-cycle)	5	94	273	0.0505	0.0180	0.0265	0.6129
Citric-acid-cycle-(TCA-cycle)	14	261	300	0.0509	0.0446	0.0475	0.6047
Mitochondrial-fatty-acid-beta-oxidation	3	158	248	0.0186	0.0120	0.0146	0.4571
Glycolysis	42	1822	350	0.0225	0.1071	0.0372	0.4151
SARS-COV-2-Maturation-of-spike-protein	13	23	115	0.3611	0.1016	0.1585	0.3571
SARS-COV-2-ATTACHMENT-ENTRY	22	96	270	0.1864	0.0753	0.1073	0.3429
superpathway-of-D-myo-inositol-1-4-5-trisphosphate-metabolism	18	179	201	0.0914	0.0822	0.0865	0.3333
Interleukin-2-signaling	51	105	20	0.3269	0.7183	0.4493	0.2941
NFE2L2-regulates-pentose-phosphate-pathway-genes	5	15	8	0.2500	0.3846	0.3030	0.2857
SARS-CoV-1-activates-modulates-innate-immune-responses	17	789	6	0.0211	0.7391	0.0410	0.1364
Paradoxical-activation-of-RAF-signaling-by-kinase-inactive-BRAF	476	1638	324	0.2252	0.5950	0.3267	0.1042
Interferon-alpha-beta-signaling	148	522	466	0.2209	0.2410	0.2305	0.0870
Signaling-downstream-of-RAS-mutants	433	1306	305	0.2490	0.5867	0.3496	0.0851
HDACs-deacetylate-histones	430	1712	4189	0.2007	0.0931	0.1272	0.0693
Selenocysteine-synthesis	798	1260	3590	0.3878	0.1819	0.2476	0.0680
Negative-regulation-of-FGFR1-signaling	161	2060	75	0.0725	0.6822	0.1311	0.0370
SARS-COV-2-modulates-host-translation	170	255	572	0.4000	0.2291	0.2913	0.0192
Regulation_of_telomerase	190	10230	23	0.0182	0.8920	0.0357	0.0141
				0.1752	0.3213	0.1673	
				0.0185	0.0882	0.0178	

Table 8 - Evaluation of relations extraction in Biopatternsg without small molecules
DOI: https://doi.org/10.60797/jbg.2026.32.1.19

Pathway	TP	FP	FN	Precision	Recall	F1	sm/objs
CREB-phosphorylation	6	50	0	0.1071	1.0000	0.1935	0.0000
ERKs-are-inactivated	34	120	5	0.2208	0.8718	0.3523	0.0000
IFNG-signaling-activates-MAPKs	27	81	0	0.2500	1.0000	0.4000	0.0000
Negative-regulation-of-MAPK	65	399	107	0.1401	0.3779	0.2044	0.0000
Regulation-of-NFE2L2-gene-expression	10	80	0	0.1111	1.0000	0.2000	0.0000
SARS-CoV-1-modulates-host-translation-machinery	83	125	508	0.3990	0.1404	0.2078	0.0000
SARS-COV-2-AUTOPHAGY	22	50	7	0.3056	0.7586	0.4356	0.0000
SARS-COV-2-Translation-of-Replicase-and-Assembly-of-the-Replication-Transcription	31	39	11	0.4429	0.7381	0.5536	0.0000
Signalling-to-RAS	29	103	0	0.2197	1.0000	0.3602	0.0000
Spry-regulation-of-FGF-signaling	23	180	13	0.1133	0.6389	0.1925	0.0000
TRAF3-dependent-IRF-activation-pathway	63	131	14	0.3247	0.8182	0.4649	0.0000
Trans-Golgi	8	31	0	0.2051	1.0000	0.3404	0.0000
				0.2366	0.7787	0.3254	
				0.0127	0.0768	0.0155	

Conclusion

We have tested the feasibility of a methodology that combines generative AI and logical AI. We are leveraging the capacity of generative AI to address the task of entity and relations extraction. With respect to logical AI, we developed prolog scripts to infer biological object's identities from knowledge bases automatically organized, in addition to handling of constraints, which guide the inference of regulatory pathways and subnetworks.

The experiment also suggests the need to model the task of relation extraction at the sentence level, as opposed to the current modeling approach, developed at the abstract level; that granularity is important to improve the biological understanding about the knowledge domain on consideration. The automatic modeling of the knowledge bases with object's names and synonyms, and their identities, requires improvements. And this requires devising ways to map identifiers from a service like Uniprot to another like MeSH. So far, we only inferred regulatory pathways using the object's identities coming from MeSH. In the near future, we will perform experiments with regulatory pathways restricted by the knowledge in the ontology trees provided by gene ontology and the metadata provided by Uniprot and PDB; we have not yet built GRNs with restrictions shaped by those services.

The results shown in Table 4 and Table 5 establish that an initial exploratory phase conducted by the researcher could help; the idea is to review manually a better alignment of the objects of interest with the names assigned to them by the PubTator's LLMs. Therefore, future experimental tasks regarding Table 4 and Table 5 involve running experiments with a better-aligned list of objects, and measuring from there the impact that this may have on the results that we get. We did not proceed in that manner in this work, as our objective was to see how far our pipeline could go in modeling knowledge bases, without human intervention.

Results such as those described in Table 5 and Table 6 are possible even with the metrics shown in Table 4 and Table 5, given the wide variety of interactions that generative AI provides and that logical AI can leverage to yield valuable findings. Stage 1 of the pipeline guides the gathering of abstracts from PubMed, then the stage 2 shapes from them a corpus of information closely related to a research problem; the COVID-19 experiment presented here demonstrates the type of functionalities that we have implemented to leverage that kind of corpora. The COVID-19 experiment also shows that our pipeline defines a knowledge management system useful to formulate hypothetical scenarios; in addition to facilitating the modeling of the specific domain of biology in which the researcher is working.

Благодарности

Мы благодарим учреждения, к которым мы принадлежим, за поддержку, оказанную в ходе выполнения этой работы.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Acknowledgement

We thank the institutions to which we belong for the support provided in the development of this work.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы на английском языке / References in English

1. Kahneman D. Thinking, Fast and Slow / D. Kahneman. — Farrar, Straus And Giroux: 1st Edition, 2011. — 499 p.

2. López J. A logical and ontological framework for knowledge discovery on gene regulatory networks. case study: bile acid and xenobiotic system (BAXS) / J. López, Y. Ramírez, J. Dávila et al. // Journal of Bioinformatics and Genomics. — 2020. — 2(14). — DOI: 10.18454/jbg.2020.2.14.1

3. Cary M.P. Pathway information for systems biology / M.P. Cary, G.D. Bader, C. Sander // FEBS Lett. — 2004. — 579(8). — P. 1815–20. — DOI: 10.1016/j.febslet.2005.02.005

4. Papin J.A. Reconstruction of cellular signalling networks and analysis of their properties / J.A. Papin, T. Hunter, B.O. Palsson et al. // Nat Rev Mol Cell Biol. — 2005. — 6(2). — P. 99–111. — DOI: 10.1038/nrm1570

5. Croft D. Reactome: a database of reactions, pathways and biological processes / D. Croft, G. O'Kelly, G. Wu et al. // Nucleic Acids Res. — 2011. — 39 (Database issue). — DOI: 10.1093/nar/gkq1018

6. Demir E. The BioPAX community standard for pathway data sharing / E. Demir, M.P. Cary, S. Paley et al. // Nat Biotechnol. — 2010. — 28(9). — P. 935–42. — DOI: 10.1038/nbt.1666

7. Khamparia A. Comprehensive analysis of semantic web reasoners and tools: a survey / A. Khamparia, B. Pandey // Education and Information Technologies. — 2017. — 22. — P. 3121–3145. — DOI: 10.1007/s10639-017-9574-5

8. Muñoz-Torres M. Get GO! Retrieving GO Data Using AmiGO, QuickGO, API, Files, and Tools / M. Muñoz-Torres, S. Carbon // Methods Mol Biol. — 2017. — 1446. — P. 149–160. — DOI: 10.1007/978-1-4939-3743-1_11

9. Kitano H. Accelerating systems biology research and its real world deployment / H. Kitano // NPJ Syst Biol Appl. — 2015. — 1. — P. 15009. — DOI: 10.1038/npjsba.2015.9

10. Kitano H. Nobel Turing Challenge: creating the engine for scientific discovery / H. Kitano // NPJ Syst Biol Appl. — 2021. — 7(1). — P. 29. — DOI: 10.1038/s41540-021-00189-3

11. Rougny A. A logic-based method to build signaling networks and propose experimental plans / A. Rougny, P. Gloaguen, N. Langonné et al. // Sci Rep. — 2018. — 8(1). — P. 7830. — DOI: 10.1038/s41598-018-26006-2

12. Baumann N. How to use the medical subject headings (MeSH) / N. Baumann // Int J Clin Pract. — 2016. — 70(2). — P. 171–4. — DOI: 10.1111/ijcp.12767

13. Thomas P.D. The Gene Ontology and the Meaning of Biological Function / P.D. Thomas // Methods Mol Biol. — 2017. — 1446. — P. 15–24. — DOI: 10.1007/978-1-4939-3743-1_2

14. Fiorini N. Towards PubMed 2.0 / N. Fiorini, D.J. Lipman, Z. Lu // Elife. — 2017. — 6. — P. e28801. — DOI: 10.7554/eLife.28801

15. Rose P.W. The RCSB protein data bank: integrative view of protein, gene and 3D structural information / P.W. Rose, A. Prlić, A. Altunkaya et al. // Nucleic Acids Res. — 2017. — 45(D1). — P. D271–D281. — DOI: 10.1093/nar/gkw1000

16. Berman H.M. The Protein Data Bank archive as an open data resource / H.M. Berman, G.J. Kleywegt, H. Nakamura et al. // J Comput Aided Mol Des. — 2014. — 28(10). — P. 1009–14. — DOI: 10.1007/s10822-014-9770-y

17. Gray K.A. A review of the new HGNC gene family resource. / K.A. Gray, R.L. Seal, S. Tweedie et al. // Hum Genomics. — 2016. — 10. — P. 6. — DOI: 10.1186/s40246-016-0062-6

18. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025 / Consortium. UniProt // Nucleic Acids Res. — 2025. — 53(D1). — P. D609–D617. — DOI: 10.1093/nar/gkae1010

19. Tsunoda T. Estimating transcription factor bindability on DNA. / T. Tsunoda, T. Takagi // Bioinformatics. — 1999. — 15(7-8). — P. 622–30. — DOI: 10.1093/bioinformatics/15.7.622

20. Khan A. Jaspar restful API: accessing Jaspar data from any programming language / A. Khan, A. Mathelier // Bioinformatics. — 2018. — 34(9). — P. 1612–1614. — DOI: 10.1093/bioinformatics/btx804

21. Wei C.H. PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge / C.H. Wei, A. Allot, P.T. Lai et al. // Nucleic Acids Res. — 2024. — 52(W1). — P. W540–W546. — DOI: 10.1093/nar/gkae235

22. Luo L. AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning / L. Luo, C.H. Wei, P.T. Lai et al. // Bioinformatics. — 2023. — 39(5). — DOI: 10.1093/bioinformatics/btad310

23. Wei C.H. GNorm2: an improved gene name recognition and normalization system / C.H. Wei, L. Luo, R. Islamaj et al. // Bioinformatics. — 2023. — 39(10). — DOI: 10.1093/bioinformatics/btad599

24. Lai P.T. BioREx: Improving biomedical relation extraction by leveraging heterogeneous datasets / P.T. Lai, C.H. Wei, L. Luo et al. // J Biomed Inform. — 2023. — 146. — P. 104487. — DOI: 10.1016/j.jbi.2023.104487

25. Witten I. Data Mining Practical Machine Learning Tools and Techniques / I. Witten, E. Frank, M. Hall. — Third Edition: Morgan Kaufmann Publishers, 2011. — 665 p.

26. Rodchenkov I. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data / I. Rodchenkov, O. Babur, A. Luna et al. // Nucleic Acids Res. — 2020. — 48(D1). — P. D489–D497. — DOI: 10.1093/nar/gkz946

27. Gu Y. Domain-specific language model pretraining for biomedical natural language processing / Y. Gu, R. Tinn, H. Cheng et al. // ACM Transactions on Computing for Healthcare (HEALTH). — 2021. — 3(1). — P. 1–23. — DOI: 10.1145/3458754

28. Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding / J. Devlin, M.W. Chang, K. Lee et al. // In Proceedings of the 2019 Conference of the North American chapter of the association for computational linguistics: human language technologies. — 2019. — 1. — P. 4171–4186. — DOI: 10.48550/arXiv.1810.04805