ImmunoGIT, A Repository of Models and Metadata for Human Immune System Responses

M1 BBS - Tutored Project

Yasmine Boukadida
Guillaume Souède
Issa Kerima
Hawa Balde

Supervised by:

Pr. Anna Niaraki
Mr. Nicolas Ricort

2024-2025

I. Introduction

The human immune system is a dynamic and intricate web responsible for protecting the body from pathogens and ensuring homeostasis. Immune responses are unpredictable and challenging to comprehend, representing a major challenge in biomedical research, particularly in the development of therapies specifically aimed at autoimmune diseases, infections, and cancer. The concept of immune digital twins (IDT) has only emerged as a potential solution in this context recently. Immune digital twins are in vivo virtual duplicates of an individual's immune system, built from computational models and biological data, with the goal of simulating and predicting immune responses in various contexts. Immune digital twins hold great promise for enhancing precision medicine, optimizing drug development, and personalizing treatment protocols. [1] [2]

Even though computational immunology has made a lot of progress, developing digital twins of the immune system is still limited by the lack of a dedicated space for immunology models. Right now, these models are scattered across general repositories like BioModels, often without clear organization, proper annotations, or consistent formats. This makes it hard for researchers to find, compare, and reuse them to study the immune system or build digital twins. [1] [3]

ImmunoGIT fills this important gap by creating a centralized, FAIR-compliant repository that focuses entirely on computational models and metadata related to the human immune response. It collects models, improves them with proper annotations using standard formats like SBML and SBML-QUAL, and makes them easily available through a dedicated GitHub repository. The goal is to create a solid base for building immune digital twins. This tool not only supports ongoing research in immunology and systems biology but also encourages international collaboration and helps move digital twin technology forward in precision medicine. [4] [5]

II. Project Context and Hosting Team

This project was carried out within the CoSysBio team, part of the Centre de Biologie Intégrative (CBI) at the University of Toulouse. The CBI is a multidisciplinary research center focusing on the integrative understanding of living organisms. The host team, CoSysBio, specializes in computational systems biology for complex human diseases, combining mathematics, computer science, and bioinformatics to identify new therapeutic targets and model complex diseases.

The project was supervised by Prof. Anna Niarakis and Nicolas Ricort-Teixidor, who are involved in national and international collaborations, including the Research Data Alliance (RDA) and the Working Group Building Immune Digital Twins (WG BIDT).
The ImmunoGIT project directly contributes to these collaborative efforts by providing a structured resource for immune system modeling, which is essential for developing immune digital twins and advancing precision medicine.

III. General Description and Objectives

The main goal of the ImmunoGIT project is to develop a **modular, FAIR-compliant repository** of computational models with their metadata focused on human immune responses.
Concretely, the project aims to:

1. Collect and organize immune-related models from the manually curated section (for standard SBML models) and both curated and non-curated sections (for SBMLQual models) of the BioModels database;
2. Extract standardized metadata for each model
3. Make the models and metadata easily accessible via a dedicated GitHub repository.

IV. Materials and Methods

1. Data Sources and Model Selection

The source of computational models for this project was the **BioModels database,** a well-established repository of biomathematical models in standardized SBML format.
Selection criteria included:

- Curated models were selected for standard SBML models to ensure scientific accuracy and expert validation. For SBML-Qual models, both curated and non-curated models were included due to their limited availability and, possibly, their simpler structure, which might facilitate future curation.
- For SBML models, the initial search was based on the general keyword *“immun*”*, which retrieved some models already tagged, notably under the “Immuno-oncology” category. Untagged SBML models were then further classified using specific keywords such as “immune response,” “immune system,” or “immunotherapy” to better define their thematic focus.
- For SBML-Qual models, the search was limited to the terms “boolean” or “logical.” The resulting models were then filtered to retain only those related to the immune system

2. Tools and Software

Python 3 was used as the main language for handling data and automating tasks.

The bioservices Python library was used to interact with the BioModels API, allowing automated searching, filtering, and downloading of models and their associated metadata.

To make the work easier, three Jupyter Notebooks were created. Each one is used for a different type of model.

- **SBMLgetModelsAndMetadata.ipynb**: responsible for collecting all models encoded in **SBML** format, along with their associated metadata.
- **GetLogicalModelAndMetadata.ipynb**: designed to retrieve **logical models** and their metadata, encoded in **SBML-Qual** format.
- **GetBooleanModelAndMetadata.ipynb**: used to collect **Boolean models** and their metadata, which also fall under the **SBML-Qual** category.

These notebooks automate a range of tasks, including:

1. Search BioModels using specific keywords.
2. Filtering models to retain only relevant ones, including curated models for standard SBML and both curated and non-curated models for SBML-Qual;
3. Download models in SBML or SBML-QUAL format.
4. Extract and structure metadata associated with each model.

3. Metadata Extraction

For each selected model, metadata was extracted from the provided JSON file.

The extracted metadata typically included:

- Model format (SBML or SBML-QUAL)
- Unique model identifier
- Model name
- Submission and last modification dates
- Author(s)
- Link to the original BioModels entry

4. Repository Organization

All collected models and their metadata were centralized in a dedicated GitHub repository. The repository was structured for clarity and ease of navigation:

- Separate folders for SBML and SBML-QUAL models.
- Metadata files (JSON format) organized alongside their respective models.
- A README file and additional documentation to guide users.

To further facilitate access, a GitHub Pages site was created, offering a simple-to-use web interface to search and download models and metadata.

5. Documentation and Verification

All scripts and notebooks were clearly documented to make them easy to understand and reuse. To help with this, a user guide was added that explains how to run the notebooks and interpret the results. Additionally, each model was carefully checked to ensure it followed SBML and SBML-QUAL standards before being added to the repository, so everything works well together and can be trusted.

V. Results

1. Overview of Collected Models

A total of **152 computational models** related to the **human immune system** were collected and organized in the ImmunoGIT repository. These models are divided into two main categories based on their format:

- 131 Standard SBML models
- 21 SBML-Qual (logical and Boolean) models

Each model is accompanied by its corresponding metadata file, and all files are structured in a clear folder hierarchy within the GitHub repository. Models are grouped according to biological themes such as *“Immuno-oncology”*, *"Immune System," "Immune Response," "Immune Regulation,"* and others.

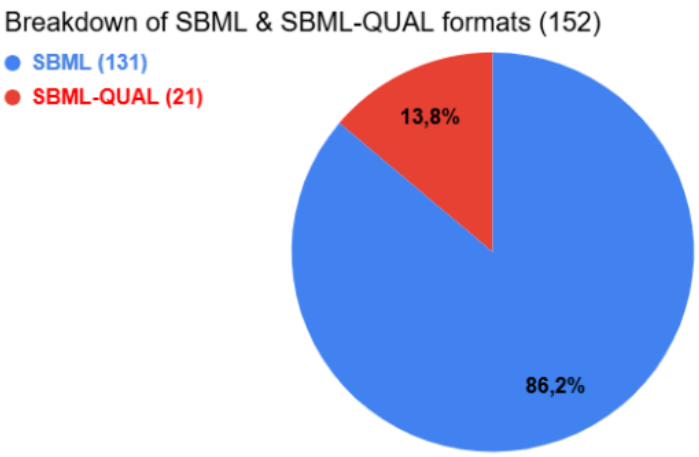


Figure 1. Overview of immune-related computational models in the ImmunoGIT repository.

2. Organization of the Repository

To make the repository easy to use and navigate, it has a modular and simple structure. Models and their metadata are kept in separate folders based on their format—either SBML or SBML-Qual. Inside these folders, models are organized by biological topics. For example, within the SBML directory, there is a subfolder called **biomodels_Infection** that contains all models related to infection processes.

Each model has its own dedicated subfolder, which includes:

- The model file (in `.xml` format)
- Its related metadata (in `.json` format)

This clear organization makes it easier for researchers to browse manually or access the data programmatically, improving the overall user experience.

3. Analysis of SBML Models

A more detailed analysis was performed on the **131 SBML models**. Among them:

- **68 models** were directly related to **immuno-oncology**, illustrating a strong representation of models focused on cancer immunology.
- **63 models** were initially **untagged**, necessitating further analysis and classification.

These untagged models were found to cover a diverse array of biological processes, including *immune cell dynamics*, *immune responses to infection*, *antibody production*, *complement system function*, and *immunotherapy mechanisms*.

To better address this heterogeneity, an **automated keyword-based tagging pipeline** was applied, allowing for a more consistent and structured classification of the 63 untagged SBML models. The process involved two levels of analysis: an initial broad categorization (Figure 2a), followed by a more detailed breakdown of the dominant "Others" group (Figure 2b).

Figure 2a shows that 77.8% of the models fell into the general "Others" category, highlighting the need for further refinement.
Figure 2b provides this refinement, revealing that the "Others" group includes models related to **generic immune mechanisms (25.4%)**, **immune cell models (19%)**, and **infection-related responses (7.9%)**, among others.

This two-step classification offers a clearer understanding of the range and complexity of immune-related models beyond those explicitly focused on immuno-oncology.

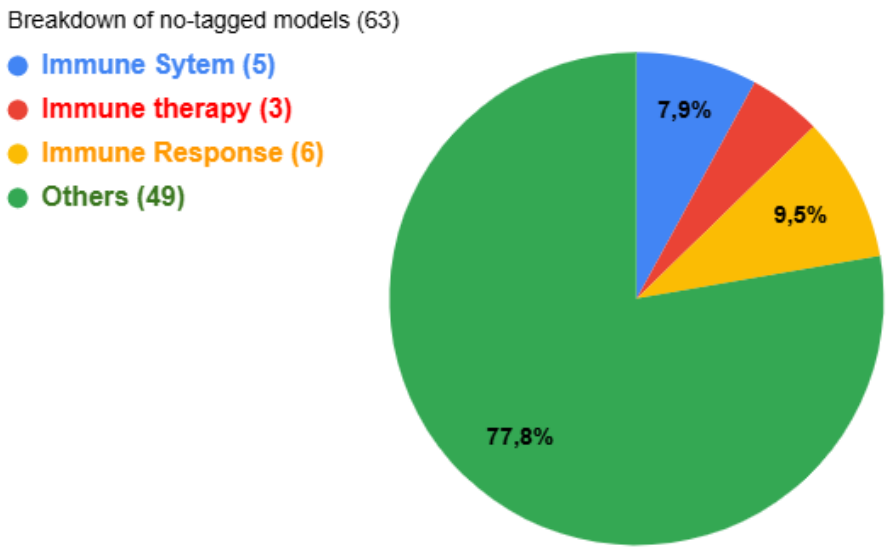


Figure 2a. Broad categorization of the initially untagged SBML models.

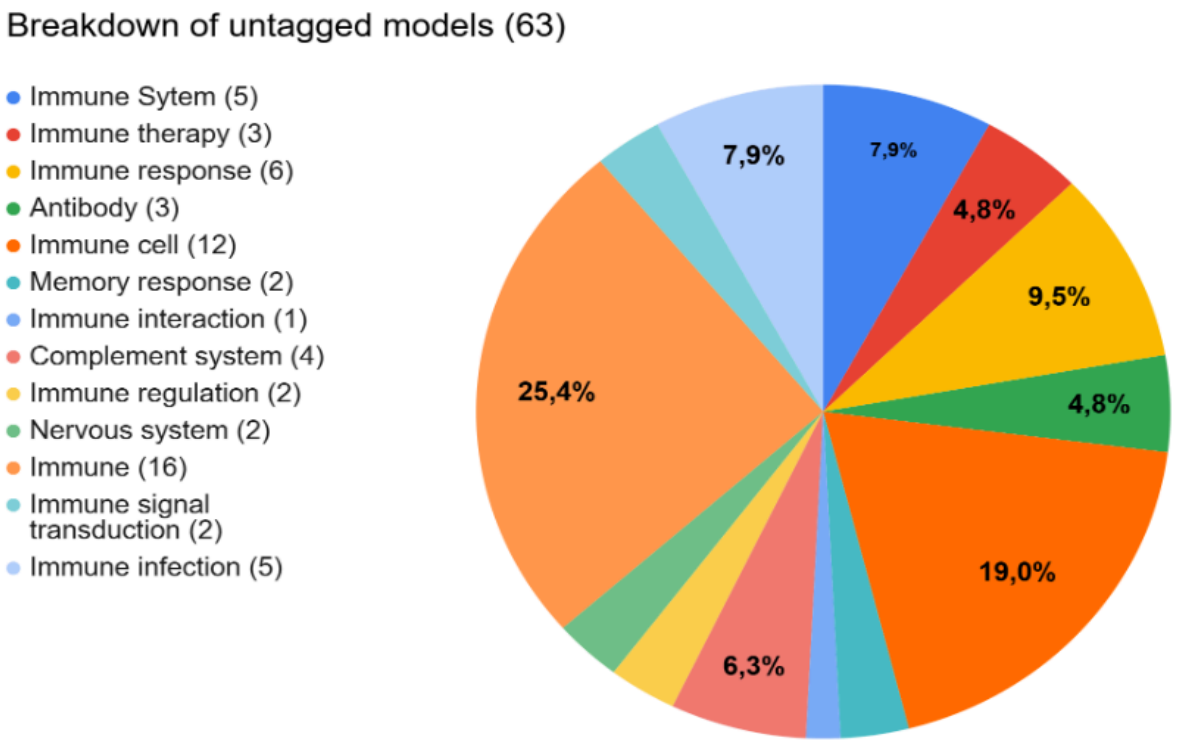


Figure 2b. Detailed breakdown of the "Others" category using automated tagging.

4. Analysis of SBML-Qual Models

The subset of **21 SBML-Qual models** presented a distinct situation:

- None of these models had **predefined biological tags**, making initial classification more challenging.

To address this, each model was individually reviewed and categorized based on:

- Their curation status (curated vs. non-curated)
- Their biological focus, although, unlike standard SBML models, they were mainly tagged under a single category: "immune cells"

The analysis revealed that:

- **10 models** relate to immune cells but are not curated
- **3 models** are curated but are not tagged
- **7 models** are neither curated nor tagged
- **1 model** is both curated and related to immune cells

These classifications are illustrated in **Figure 3**, which includes:

- a Venn diagram representing the overlap between curation and immune relevance

- 2 pie charts showing the breakdown of **logical** and **boolean models** according to these criteria.

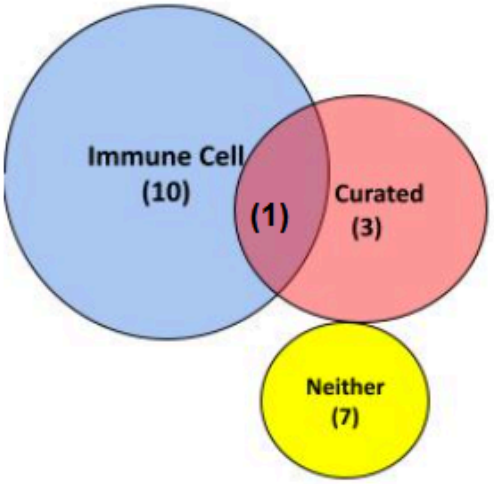


Figure 3a. Overlap between immune cell annotation and curation status among SBML-Qual models (Venn Diagram).

To better understand the 21 SBML-Qual models, we separated them into 2 groups based on their modeling approach: **logical models** (n = 16) and **Boolean models** (n = 5).

Logical models use multi-level logic to represent graded biological activity, while Boolean models simplify behavior into binary states (on/off). We analyzed these 2 groups separately due to their structural differences.

For the **Boolean models** (Figure 3a):

- 40% (2 models) were immune-related but not curated,

- 40% (2 models) had neither tag,

- and 20% (1 model) was both curated and immune-related.

In the **Logical models** (Figure 3b):

- 50% (8 models) were immune-related but not curated,
  14
- 31.3% (5 models) were neither curated nor tagged,

- 18.8% (3 models) were curated only.

This highlights the value of manual review to clarify model content when automated tags are missing.

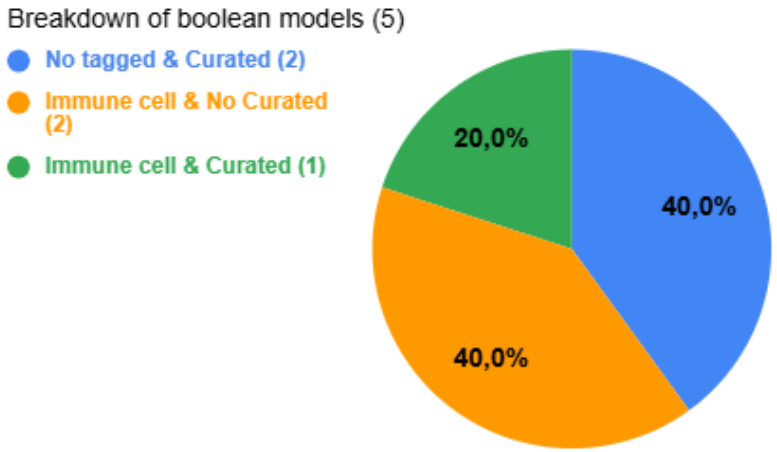


Figure 3b. Curation and immune annotation status among boolean models (n = 5).

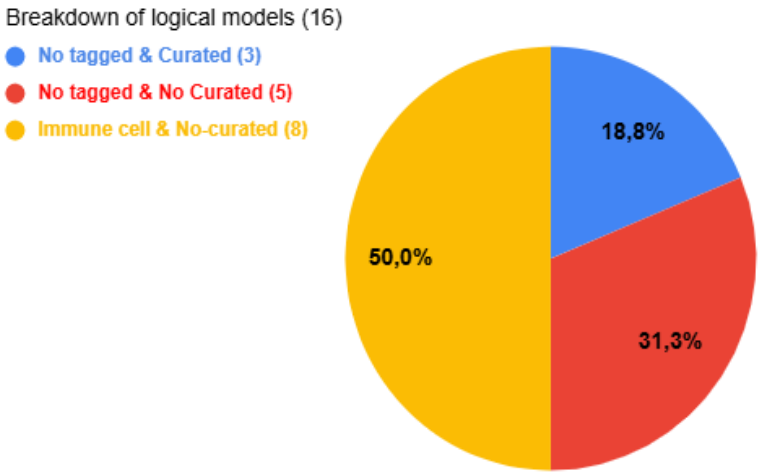


Figure 3c. Curation and immune annotation status among logical models (n =16).

5. Metadata Tagging

To further improve the usability and interoperability of the collected models, an **automated keyword-based tagging process** was implemented. This involved scanning model content and annotations for relevant biological terms, enabling the assignment of **standardized tags** aligned with key immunological concepts.

6. Accessibility

All models and their associated metadata are publicly available in the ImmunoGIT GitHub repository, where they are organized to ensure intuitive access and usability. The repository structure, combined with the standardized tagging system, facilitates efficient navigation and allows researchers to quickly locate models relevant to specific immune processes, cell types, or disease contexts. This open-access approach promotes transparency, collaboration, and reusability within the immunology modeling community.

VI. Discussion

a. Existing Resources

The building of the ImmunoGIT collection answers a rising need in the computational biology group: bettering the access, understanding, and re-use of current immune-related models. Several places to store and rules have helped a lot with sharing models in this area, especially BioModels [6], MIRIAM [7], and Reactome [8]. While these sites offer key help for keeping models and notes about them they vary in their goals, setup and quality of metadata. However, there continues to be little interoperability between repositories and no universal standard for metadata. The majority of the models in ImmunoGIT are encoded in SBML (Systems Biology Markup Language) [9], a leading format for quantitative models. SBML allows standardized representation of mathematical modeling of biological systems and is supported by wide tools in the area of systems biology. However, there remains a lack of interoperability between repositories and a universal standard for metadata, making model integration across platforms challenging. ImmunoGIT addresses this by introducing a unified folder structure, metadata format, and thematic tagging system to improve organization and cross-comparability.

Among the 152 curated immune-related models, most were in standard SBML, and more than half of these standard SBML models are related to cancer immunotherapy. This indicates a good amount of research and funding interest in that area; quantitative modeling is highly employed to simulate the dynamics between immune and tumor cells, predict treatment outcome, and even help in therapeutic design [10]. On the other hand, logical and boolean models available in SBML-Qual [11] were fewer and not so well curated. They suit exploring qualitative issues like immune signaling, differentiation, and gene regulation but lesser visibility limits adoption.

b. User Guide and Documentation Overview

Clear and minimal documentation provides support for users while interacting with ImmunoGIT. The repository contains a structured README file which details its purpose and organization while providing navigation instructions. The document serves as a rapid guide for users to help non-expert users start working with the repository.

The usability of the project would improve with structured documentation that Sphinx [12] creates from Markdown or reStructuredText as searchable web documentation. This concept remains unimplemented but presents a smooth progression from the existing README while improving scalability through community input.

The project used automated keyword-based tagging to address metadata limitations found in the initial model files. Reclassification of more than 60 previously untagged SBML models became possible through this approach while bringing attention to underrepresented subjects like the complement system and host–pathogen interactions. Although this method has its flaws, it shows what scalable metadata enrichment can accomplish for extensive model collections. Long-term maintenance of these improvements requires well-maintained documentation.

c. Sustainability

ImmunoGIT lacks long-term maintenance strategies as its main limitation. At present, there exists no automated system to facilitate the import of new models together with metadata updates and file validation processes. All curation tasks were performed manually.

The repository architecture supports upcoming handover processes and subsequent extensions. The repository's sustainability benefits from its standardized structure and annotation system. The project’s README and contribution notes which reflect internal task tracking along with documentation serve as essential preparation for new developers or researchers to join the project.

Multiple student teams managing similar repositories incorporated sections dedicated to task distribution to maintain transparency and accountability. We assigned clear responsibilities to individuals for curation, automation and documentation to simplify the identification of knowledge holders during necessary future updates.

VII. Conclusion, Limitations and Perspectives

The ImmunoGIT project created a central, easy-to-use collection of 152 computer models related to the human immune system. By organizing the models and using clear tags and standard formats (following the FAIR principles), the project helps researchers find and use immune-related models more easily and work together more effectively.

But there are still some challenges. Since the process of adding and updating models is done manually, it can be hard to keep everything up to date in the long run. In the future, it would be helpful to create tools that can add new models and update information automatically. Better documentation and input from the research community could also improve the platform.

Next steps could include adding more models from different sources, building better search and filter options, and working with other projects in computational immunology. In the long term, ImmunoGIT could help build “digital twins” of the immune system and support progress in precision medicine and computational systems biology.

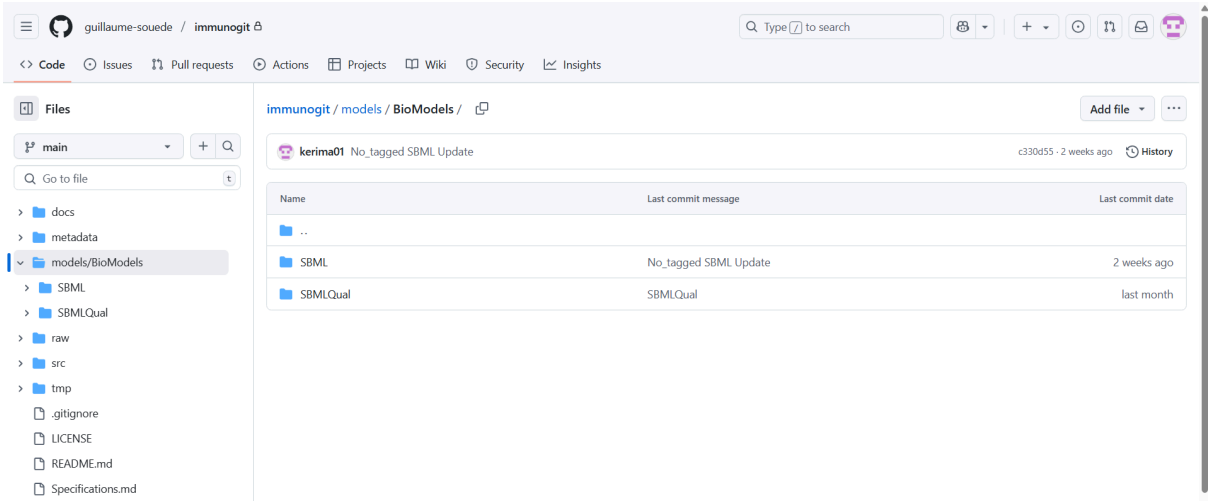
VIII. References

[1] Letort et al., *Immune digital twins for complex human pathologies*, Nature Communications Biology, 2024.
[2] Cockrell & An, *Building digital twins of the human immune system: toward a roadmap*, NPJ Digital Medicine, 2022.
[3] Letort et al., *Immune digital twins for complex human pathologies*, PubMed, 2024.
[4] Research Data Alliance, *Building Immune Digital Twins WG*.
[5] Helikar et al., *Toward mechanistic medical digital twins: some use cases in immunology*, Frontiers in Digital Health, 2024.
[6] Malik-Sheriff, R. S., et al. (2020). BioModels—15 years of sharing computational models in life science. *Nucleic Acids Research*, 48(D1), D407–D415. https://doi.org/10.1093/nar/gkz1055

[7] Le Novère, N., et al. (2005). MIRIAM: a standard for annotating biochemical models. *Nature Biotechnology*, 23(12), 1509–1515. https://doi.org/10.1038/nbt1156
[8] Jassal, B., et al. (2020). The Reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1), D498–D503. https://doi.org/10.1093/nar/gkz1031
[9] Hucka, M., et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4), 524–531.
[10] Peskov, K., et al. (2019). Quantitative modeling of immune checkpoint inhibition and its application to optimization of clinical dosing regimens. *CPT: Pharmacometrics & Systems Pharmacology*, 8(4), 247–259.
[11] Chaouiya, C., et al. (2013). SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. *BMC Systems Biology*, 7(1), 135.
[12] Brandl, G. (2008). *Sphinx: Python documentation generator*. https://www.sphinx-doc.org/

IX. Annexes

1. Online resources: https://github.com/guillaume-souede/immunogit
2. Git Repository Structure:



X. Acknowledgments

We would like to thank Prof. Anna Niarakis and Nicolas Ricort-Teixidor for their supervision and guidance throughout this project. We also acknowledge the support of Rahuman Sheriff Malik Sheriff, BioModels Senior Project Leader, and of Hasan Balcı Postdoctoral Researcher at National Library of Medicine (NLM).