A white text on an orange background

Description automatically generated

A blue and black logo

AI-generated content may be incorrect.

A close-up of a logo

AI-generated content may be incorrect.

**VPE**

A blue medical background with a person holding a tablet

AI-generated content may be incorrect.

Development of an AI-driven tool for Talk2BioModels

Internship Experience Report at Team Virtual Patient Engine (VPE), BioMed X Institute

Supervisor: Dr. Douglas McCloskey([mccloskey@bio.mx](mailto:mccloskey@bio.mx))

Sponsored by: Sanofi

Rakesh Hadne Sreenath

Matriculation number: 11038138

ADSA Winter 2024

SRH Hochschule Heidelberg, Germany

A white text on an orange background

Description automatically generated**1. Introduction and Objectives**

In recent years, the application of artificial intelligence in the life sciences has opened new doors in biomedical research and clinical decision-making. One significant area of advancement is in the development of computational models of disease, particularly in systems biology and pharmacology. These models help simulate the complex interactions within biological systems, such as how a drug might affect the different organs or processes in the human body. However, these models are often technical, complex.

As part of my ADSA Winter 2024 internship at the BioMed X Institute in Heidelberg, Germany, I worked within the Virtual Patient Engine (VPE) team to tackle this challenge. My primary project involved contributing to the Talk2BioModels (T2B) platform, an AI-powered system that allows users to query and interact with biological models using natural language.

My specific task was to design, implement, and integrate a tool called the **Get Annotations Tool**, which automatically retrieves human-readable biological descriptions for species identifiers found in SBML (Systems Biology Markup Language) models. These models, used extensively in systems biology, often contain cryptic identifiers that are difficult to interpret without referencing external resources.

The tool needed to be accurate, fast, and compatible with the LangGraph framework used by T2B’s agent-based AI system. It would allow the AI agent to explain the components of a model in natural language, making it possible for researchers, students, and clinicians to understand and interact with models in a conversational manner.

By enabling automatic annotation and explanation, the tool aimed to contribute to the broader goals of transparency, reproducibility, and inclusivity in biomedical modeling and data science. It also had to meet the technical and collaborative standards of a research-grade open-source project, including documentation, testing, and integration with other modules of the platform.

A white text on an orange background

Description automatically generated**2. Company and Department Overview**

**BioMed X Institute** is an independent research organization that operates at the intersection of academia and industry. With headquarters in Heidelberg, Germany, and additional sites in New Haven and Ridgefield, Connecticut, BioMed X focuses on advancing biomedical research across a range of fields including oncology, immunology, neuroscience, platform technologies, and artificial intelligence.

The institute’s unique model combines global talent sourcing with local incubation. Early-career scientists from top universities around the world are recruited to tackle some of the most pressing challenges in biomedical science. Each research project is co-designed and sponsored by leading pharmaceutical companies, ensuring both scientific and real-world relevance.

The main campus of BioMed X is located within the Heidelberg Technology Park, which is part of the University of Heidelberg’s Neuenheimer Feld campus. This location places the institute in one of Europe’s largest and most prestigious biomedical research hubs. Nearby institutions include the German Cancer Research Center (DKFZ), the European Molecular Biology Laboratory (EMBL), and numerous biotech companies, all contributing to a rich, collaborative research ecosystem.

**The Virtual Patient Engine (VPE)** team stands at the forefront of artificial intelligence and digital health. VPE’s core mission is to develop a next-generation computational platform that predicts the efficacy of drug candidates in virtual patient populations with exceptional accuracy. This platform is designed to address one of the pharmaceutical industry’s greatest challenges: the high failure rate of new drug candidates during clinical trials, which currently hovers around 90%.

The VPE project is supported by **Sanofi**, a global healthcare leader, and leverages artificial intelligence to model complex diseases. The initial focus of the platform is on **chronic immune-mediated diseases**, such as atopic dermatitis (AD) and inflammatory bowel disease (IBD). These diseases are highly heterogeneous across patient populations, and current treatments often fail due to variability in individual response. VPE aims to close this gap by modeling virtual populations and predicting drug performance across diverse patient profiles.

As a collaborative, interdisciplinary team, VPE includes computational biologists, AI researchers, and software engineers. The team is led by **Dr. Gurdeep Singh**, with key contributors such as Dr. Lilija Wehling, Dr. Ahmad Wisnu Mulyadi, and others. My role as an intern focused on the AI development side—contributing directly to the Talk2BioModels platform, which forms an integral part of the VPE ecosystem.

**3. Project Background**

Before building the annotation tool, I started by learning about the key technologies needed for the project. One of the most important was **SBML (Systems Biology Markup Language)**. SBML is a format used to describe biological models in a structured way, like listing components such as species, reactions, and parameters. However, the names used for these species are often short and unclear, making them hard to understand without expert knowledge.

To address this issue, I identified external databases capable of supplying semantic annotations and descriptive information. These included:

* **UniProt**: A comprehensive protein sequence and annotation database, widely regarded as the gold standard in protein biology.
* **KEGG**: The Kyoto Encyclopedia of Genes and Genomes, providing insights into biochemical pathways, diseases, and drugs.
* **OLS (Ontology Lookup Service)**: A service for accessing standard biomedical ontologies such as GO (Gene Ontology), ChEBI, and more.

Each of these databases offers API access, which was critical for real-time annotation retrieval. I studied the structure and access protocols of these APIs to ensure efficient integration and proper response handling.

**6. Tools and Technologies Used**

* **Programming & Scripting:** Python, Markdown
* **Frameworks:** Streamlit, LangChain, LangGraph
* **AI Models:** OpenAI GPT, NVIDIA models
* **Databases:** UniProt, OLS, KEGG
* **Documentation:** MkDocs, mkdocstrings
* **Validation & Testing:** Pydantic, pytest, pylint
* **Version Control & CI/CD:** Git, GitHub Actions

**7. Methodology**

A diagram of a diagram

AI-generated content may be incorrect.

***Figure 1: High-level pipeline showing the flow from SBML upload to species annotation and description retrieval.***

The development of the Get Annotations Tool followed a structured approach, grounded in the principles of modular software design and integration with AI-driven agents. Based on the tasks outlined in the project and inspired by the system architecture presented in the internship presentation, the methodology involved the following phases:

1. **Tool Design and LLM Integration:**

The tool's goal was to assist AI agents by retrieving readable descriptions for biological species. Using LangGraph’s agent orchestration, the tool incorporated a structured output interface powered by a large language model (LLM). The LLM received prompts customized to extract relevant species names from a user question, using dynamic headers generated based on the model species list.

1. **SBML Model Parsing and Species Extraction:**

The system utilized the basico library to interface with COPASI models. Species data were extracted from the SBML file using COPASI’s API, providing an indexed list of species used in simulations. The LLM, invoked with a structured prompt, then selected the subset of species relevant to the user's query.

1. **Miriam Annotation Collection:**

**A screenshot of a computer program

AI-generated content may be incorrect.**

***Figure 2: Process for identifying and selecting relevant species names using an LLM interface.***

The tool first attempted to retrieve MIRIAM annotations (Minimal Information Required In the Annotation of Models) for each species via basico.get\_miriam\_annotation(). These annotations, often pointing to authoritative external databases, were parsed for links, qualifiers, and identifiers. If a species lacked a MIRIAM entry, it was flagged and reported in the final output.

1. **Description Fetching via External APIs:**

**A screen shot of a computer program

AI-generated content may be incorrect.**

***Figure 4: How different API are handled, and Descriptions are fetched***

For species with valid MIRIAM links, the tool extracted the ID and the source database (e.g., UniProt, OLS, KEGG). API wrappers for each database were then used to fetch descriptive labels or summaries:

* UniProt: Protein function and biological context.
* OLS: Ontology terms like GO or ChEBI for compounds or processes.
* KEGG: Biochemical pathways and compound names.

1. **Data Processing and Structuring:**

****

***Figure 4: Example of how annotations are processed and formatted for user-friendly output.***

Fetched results were assembled into a uniform data structure including:

* Species Name
* Description
* Database
* ID
* Link
* Qualifier

1. **Tool Execution and LangGraph Implementation**

[**https://github.com/VirtualPatientEngine/AIAgents4Pharma/blob/main/aiagents4pharma/talk2biomodels/agents/t2b\_agent.py**](https://github.com/VirtualPatientEngine/AIAgents4Pharma/blob/main/aiagents4pharma/talk2biomodels/agents/t2b_agent.py)

(The above link has the flow of LangGraph)

The completed tool, implemented as a LangChain-compatible BaseTool class, was registered within the LangGraph agent runtime. When called, it executed the full pipeline:

**model parsing → species extraction → annotation fetching → state updating → user response.**

1. **Testing, Logging, and CI Integration**

Every component—from prompt design to API logic—was independently tested using pytest. GitHub Actions ensured new commits were tested and verified automatically. The system's state object was continuously updated to preserve experiment metadata and outputs.

This methodology provided an intelligent, robust, and user-focused approach to bridging LLMs and systems biology models, while ensuring scientific reliability and maintainability. that integrated seamlessly into the T2B ecosystem, aligning with the project’s open-source and user-centric goals.

1. **Code Documentation of AIAgents4pharma**

Designed and implemented a documentation website featuring automated code documentation powered by MkDocs.

A screenshot of a computer

AI-generated content may be incorrect.

***Figure 4: Code Documentation website.***

***Link to website:*** [***https://virtualpatientengine.github.io/AIAgents4Pharma/***](https://virtualpatientengine.github.io/AIAgents4Pharma/)

1. **Results and Achievements**

**Results Achieved**

The Get Annotations Tool successfully met its core objectives of enhancing the interpretability of SBML models. Once integrated into the Talk2BioModels (T2B) platform, the tool could process user-uploaded or pre-curated models and return accurate, concise annotations for the species within those models.

In a real-world demonstration involving a model of IL6-mediated inflammation (related to Crohn’s disease), the tool was able to:

* Identify species such as IL6, JAK2, and CCR6 from a user’s natural language query.
* Retrieve readable descriptions from UniProt and KEGG APIs.
* Present annotation results directly within a conversational agent interface.

The tool was successfully integrated into LangGraph and tested via the T2B Streamlit frontend. Additionally, its contribution was recognized in two scientific publications— an ICLR MLGenX 2025 paper and a bioRxiv preprint—both highlighting the tool’s role in increasing model accessibility.

**Evaluation of Results**

To evaluate the effectiveness of the tool below criteria were used:

**Functional Metrics:**

* Accuracy: Over 95% success rate in retrieving valid annotations.
* Stability: Passed all unit tests and continuous integration checks.

**User Feedback:**

* Output was consistently aligned with user intent as inferred from structured prompts.

1. **Skills Acquired and Challenges Faced**

This internship taught me how to bridge the theoretical aspects of AI with practical biomedical challenges. I learned the importance of modular design, clear documentation, and collaborative coding. Most importantly, I gained confidence in my ability to solve real-world problems in a high-impact, interdisciplinary environment.

**Technical Skills:**

* MkDocs documentation
* LangChain & LangGraph integration
* Streamlit development
* Unit testing with pytest
* Code validation with pylint

**Soft Skills:**

* Open-source collaboration
* Technical communication
* Agile development practices
* Time and project management

**Challenges Faced**

* Initial unfamiliarity with MkDocs and its plugin ecosystem.
* Difficulty integrating LangChain workflows with Streamlit UIs.
* API inconsistencies and data formatting from third-party sources.
* Writing unit tests and enforcing linting standards across the codebase.
* Managing pull requests and peer review cycles on GitHub.

*“My internship at BioMed X has been a transformative experience. I had the opportunity to work on a cutting-edge project at the intersection of AI and systems biology. I’m grateful for the mentorship provided by Dr. Douglas McCloskey and the VPE team. This experience has reinforced my ambition to pursue a career in biomedical research and AI-driven healthcare solutions.”*

**15. References**

1. Singh, G., Wehling, L., Mulyadi, A. W., Sreenath, R. H., et al. (2025). *Talk2Biomodels and Talk2KnowledgeGraph: AI agent-based application for prediction of patient biomarkers and reasoning over biomedical knowledge graphs.* MLGenX Workshop at ICLR 2025. OpenReview. <https://openreview.net/forum?id=av4QhBNeZo>
2. Wehling, L., Singh, G., Mulyadi, A. W., Sreenath, R. H., et al. (2025). *Talk2Biomodels: AI agent-based open-source LLM initiative for kinetic biological models.* bioRxiv. <https://doi.org/10.1101/2025.03.11.642548>