15.572 Analytics Lab
A Collaboration With Sanofi: "Patient Like Me"

Nidhish Nerur
MIT
nerur325@mit.edu

Naiqi Zhang
MIT
naiqiz79@mit.edu

Jaeyoon Wang
MIT
jwang416@mit.edu

Hunter Sporn
MIT
hsporn@mit.edu

1 Executive Summary

Problem Statement: The project aims to find similar patient segments who are most likely to benefit from specific medications, especially those targeting the TNF-α protein. The TNF-α inhibition drugs often treat inflammatory illnesses including rheumatoid arthritis and Crohn's disease, yet their efficacy differs across patients. Consequently, we leverage the latest data science methods to identify exceptional responders to particular treatments. These findings would optimize Sanofi's treatment allocation and resource distribution strategies, with the goal of improving patient outcomes and improving the drug development process. Then, the main goal is to cluster patient groups that would perhaps experience similar responses to a given drug, enhancing personalized healthcare services.

Overall Approach: Our team combined the publicly available Prime Knowledge Graph (PrimeKG) and Medical Information Mart for Intensive Care (MIMIC-IV) datasets to create an interconnected knowledge graph with drug, disease, and patient information. PrimeKG offers drug and disease descriptions while MIMIC-IV contains real patient data such as demographics, lab tests, and past diagnoses. To connect these databases, we added/connected MIMIC-IV patient and hospital admission nodes to PrimeKG. With the resulting "super-graph" of information, we sought to compute similarity scores between patient segments and the TNF-α inhibition drug. We further looked at characteristics of patients who were more likely to be administered the TNF-α drug. This process allowed us to develop actionable insights to assist with personalized treatment allocation for patients.

Technical Insights: The super-graph revealed there are 38 drugs and 225 diseases related to the TNF-α protein, and we wanted to delve deeper into patients associated with these specific medications or illnesses. Specifically, we discovered the hemolytic-uremic syndrome, acute coronary syndrome, and phosphorus metabolism diseases have the highest similarity scores to TNF-α. Biomedical literature suggests these diseases are strongly linked to the TNF-α protein, corroborating our results. We also identified therapeutically aligned drugs that show strong connections to TNF-α, such as PR-104, OMS-103HP, and Onerecept.

We further perform a random walk across the super-graph, which is a stochastic process that traverses the graph across various nodes in the network. Random walks help to create node embeddings that capture the semantic meaning and relationships between nodes in the network graph. Additionally, the t-SNE visualization of node embeddings highlights the clusters for diseases, drugs, and other node types, emphasizing the rich semantic relationships captured through the embeddings. This information enabled us to create clusters to visualize node embeddings for patients, drugs, and diseases. Our findings can significantly improve Sanofi's approach to drug development, identifying similar patients who would likely respond well to novel treatments.

Sanofi Impact: Sanofi can utilize our technical insights to enhance drug-patient matching algorithms, improve the drug development process, and prioritize patient outcomes. As we map relationships between drugs and patients, we can predict the level of a patient's response to TNF-α inhibition drugs, enabling synthetic construction of theorized patient "super-responder" cohorts. Consequently, Sanofi can distribute resources more effectively and design targeted randomized clinical trials for patients who are most likely to benefit from new medications. Sanofi will realize reduced costs associated with the experimentation of treatment assignment, increasing operational efficiency. Even saving marginal amounts of time in clinical trials can expedite the process for drugs to reach the market and directly impact patients. As a result, Sanofi will contribute to precision medicine, or more personalized healthcare for patients. Sanofi can further maintain its competitive position in the pharmaceutical industry with data-driven decisions that better the human condition.

2 Introduction

Our project leverages data science methods to transform the drug development process. We focuses on identifying similar patients who would likely respond well to drugs with varying mechanisms of action, or pathways through which the drug affects the body. The primary task is to develop data-driven solutions to better understand characteristics of "super-responders," which are patients who show significant benefits from drug treatments. Consequently, these user profiles can inform future drug development and provide insights into which patients should receive a specific drug.

We utilized the Prime Knowledge Graph (PrimeKG) network database and the Medical Information Mart for Intensive Care (MIMIC-IV) dataset to assess patients' response to different drug mechanisms. By coupling an integrated biomedical knowledge-graph with real electronic health record data, we sought to develop a rich understanding of patient-drug interactions. In particular, our team analyzed cohorts receiving the TNF-α inhibition drug, often used to treat inflammatory conditions like rheumatoid arthritis, psoriasis, and more [5]. As a result, we focused on modeling connections between drugs targeting TNF-α, patients, and diseases, and identifying "super-responders" who would likely respond well to drugs with varying mechanisms of action, or pathways through which the drug affects the body. This work enables Sanofi to enhance its lead in artificial intelligence for research, development, and drug development [8].

2.1 PrimeKG

This graph database showcases relationships among biological entities including genes, diseases, drugs, and protein pathways. We can traverse the knowledge graph edges connecting nodes to acquire detailed information about how drugs target parts of the body to cure patients from diseases. Specifically, we extracted drug and disease feature datasets, which contain descriptions about the interactions between treatment options and diseases [6]. This provided us with context to subsequently find patient similarity scores for those assigned to receive particular medications. Hence, we need MIMIC-IV's patient feature data in order to delve deeper into patient-drug connections.

2.2 MIMIC-IV

The publicly available MIMIC-IV contains anonymized patient data from the Beth Israel Deaconess Medical Center in Boston, MA. It provides patient demographic, diagnoses, treatment, and outcome fields. We sought to link MIMIC-IV into the PrimeKG database to create a deeper, interconnected graph that allows us to identify features of the super-responders to drugs targeting TNF-α [1]. After integrating the two datasets, our team could calculate similarity scores between patients taking a TNF-α inhibition drug. In particular, we employed the network random walk algorithm to automatically traverse the super-graph and compute patients' degree of connectedness. The random walk algorithm is discussed in more detail during the technical analysis portion of the paper.

3 Technical Analysis

For our analysis, we created a technical architecture framework, discovered insights from PrimeKG and MIMIC-IV, and computed patient similarity scores. We used Python and several Python libraries (including pandas, node2vec, torch, numpy, and transformers) to implement our methodology and highlight our primary findings in the subsequent sections.

3.1 Technical Architecture

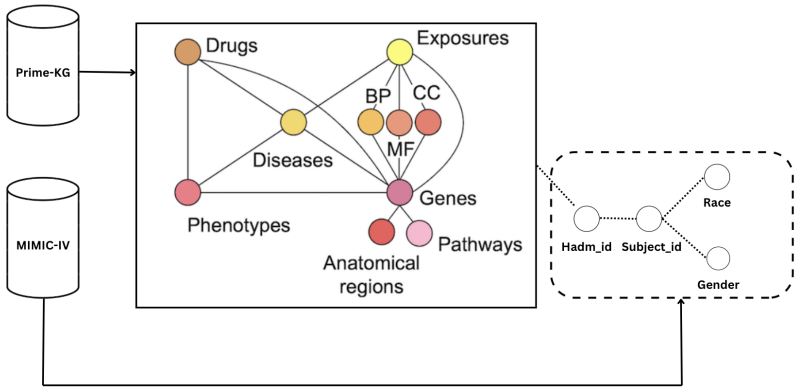

Figure 1: Super-Graph Technical Architecture

Figure 1 highlights the process of combining the drug and disease datasets from PrimeKG along with the patient demographic information in MIMIC-IV (i.e., gender, race). Specifically, we added a set of patient nodes into PrimeKG using MIMIC-IV consisting of people who were taking drugs related to TNF-α or had diseases affecting TNF-α. Then, we could directly access connections between drugs, diseases, and patients within the graph. The output of our Super-Graph enabled us to compute similarity scores between patients as shown in Figure 2 below:

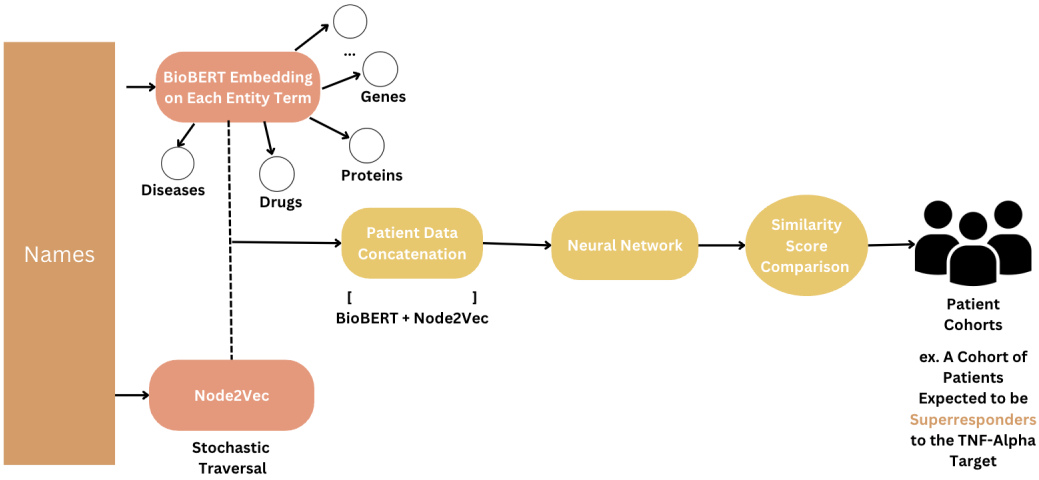

Figure 2: Patient Similarity Score Architecture

For this architecture, we pass in all our biomedical entities (for which BioBERT would have existing context), including (non-exhaustive) disease, drug, pathway, and gene/protein into BioBERT to extract numeric embeddings representing the biomedical positioning of the entity. BioBERT is pre-trained on a vast corpus of medical documents, so we believe it is well-suited to analyze our biomedical super-graph. We further apply Node2Vec, translating the nodes of the super knowledge graph into numeric vector representations. Subsequently, we concatenate the embeddings representing patient connections to particular diseases and drugs then use the Random Walk stochastic process to calculate similarity scores. This flow process is discussed step by step in the next sections.

3.2 Preliminary PrimeKG Insights

To analyze PrimeKG, we primarily sought to understand the types of drugs and diseases related to TNF-α. We found a list of 38 drugs and 225 diseases linked to TNF-α, and we have provided the full list of these drugs and diseases in the Appendix section. The drugs seem to focus on TNF-α inhibition, which reduces inflammation in patients. However, there is a wide range of diseases that are likely unrelated to inflammation but still connected to TNF-α within the knowledge graph. For instance, we have mental health disorders (i.e., schizophrenia, Alzheimer disease), metabolic conditions (i.e., hypoglycemia, Gilbert syndrome), and forms of cancer (i.e., thyroid, colorectal), which are often not categorized as inflammatory diseases. In contrast, we also have established health conditions directly affecting TNF-α and inflammation levels, such as rheumatoid arthritis, psoriatic arthritis, Crohn's disease, and others [2]. Given our PrimeKG findings, we focused our MIMIC-IV analysis on patients receiving TNF-α related drugs or diagnosed with TNF-α associated diseases.

3.3 Preliminary MIMIC-IV Insights

For our MIMIC-IV analysis, we focused on building a connection to PrimeKG using data from the hosp module, specifically diagnosis details, prescription information, and patient demographic data. Diagnoses, sourced from the diagnoses_icd file, provided insights on patient conditions using ICD codes and their descriptive titles. Prescription data from the prescriptions file captured information on medications administered during hospital stays, including drug names and dosages. Patient demographic information from the admissions file contained critical details such as admission age (calculated using anchor age and anchor year) and other attributes like gender and ethnicity. These datasets were combined into high-dimensional vectors representing each patient and hospital admission, which we encoded using BioBERT to generate comprehensive patient representations. Focusing on this approach, we specifically analyzed patients receiving TNF-α-related drugs or diagnosed with diseases associated with TNF-α. This allowed us to align patient-level clinical data with PrimeKG's biomedical knowledge and explore connections between clinical treatments and disease mechanisms.

3.4 Super-Graph Development

The process of supergraph development serves as a critical bridge between biomedical knowledge graphs like PrimeKG and electronic health record (EHR) datasets such as MIMIC-IV, creating a unified dataset that connects diverse and fragmented healthcare information. This supergraph is constructed as a CSV file that integrates data across multiple dimensions. Specifically, it connects drugs and diseases to their related information from PrimeKG, hospital admissions to diseases and prescriptions as derived from MIMIC, hospital admissions to patients, and patients to crucial demographic information such as race and gender. By linking these entities as nodes and edges, the supergraph captures relationships and pathways that are otherwise siloed in separate datasets, enabling the generation of a holistic view of patient health, drug interactions, and disease progression.

The implications of this development are far-reaching. The supergraph supports advanced predictive modeling and hypothesis generation by allowing researchers to identify patterns and relationships across disparate data modalities. For instance, it enables identifying patient subpopulations likely to respond to specific drugs by combining preclinical knowledge (e.g., drug mechanisms and genetic pathways) with real-world clinical data (e.g., patient outcomes and demographics). Furthermore, this integration paves the way for innovative approaches in precision medicine, such as constructing patient embeddings that represent individual health journeys in a way that reflects their biological, demographic, and clinical contexts. Ultimately, supergraph development not only enhances the ability to match patients to suitable treatments but also provides a robust platform for exploring novel mechanisms of action and validating clinical hypotheses. This is vital for improving clinical trial design, drug efficacy evaluation, and patient stratification, directly contributing to the advancement of personalized healthcare solutions.

3.5 Similarity Score Algorithm

We employed the random walk method to compute similarity scores across the vast biomedical knowledge super-graph. This algorithm simulates a stochastic traversal of the graph structure. Specifically, it will start from a particular drug, disease, or patient node then utilize sophisticated probability calculations to move to adjacent nodes. Then, the random walk looks at the number of times other nodes are visited, which could be indicative of relevance to the starting node. In our context, we apply the random walk to find patients close to the TNF-α related drugs or diseases in the graph, enabling us to quantify similarity scores. We show an illustration of one step in the random walk below:

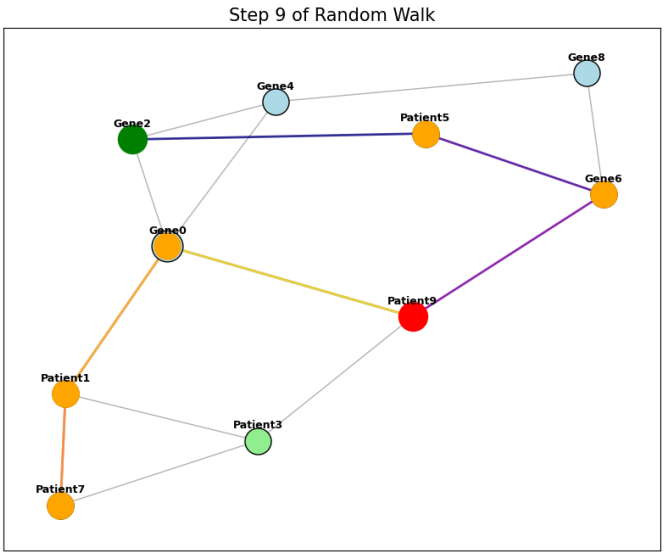

Figure 3: Random Walk

We note the random walk shows the direct and second-order connections between particular genes and patients in this example. Consequently, we can apply this stochastic process to find edges connecting the patients, drugs, and diseases of interest.

3.6 Embedding Visualization

In addition to the random walk, our team used BioBERT to extract relevant embedding scores from the entire knowledge graph. The node embeddings for various drugs, diseases, patients, and more can be seen here:

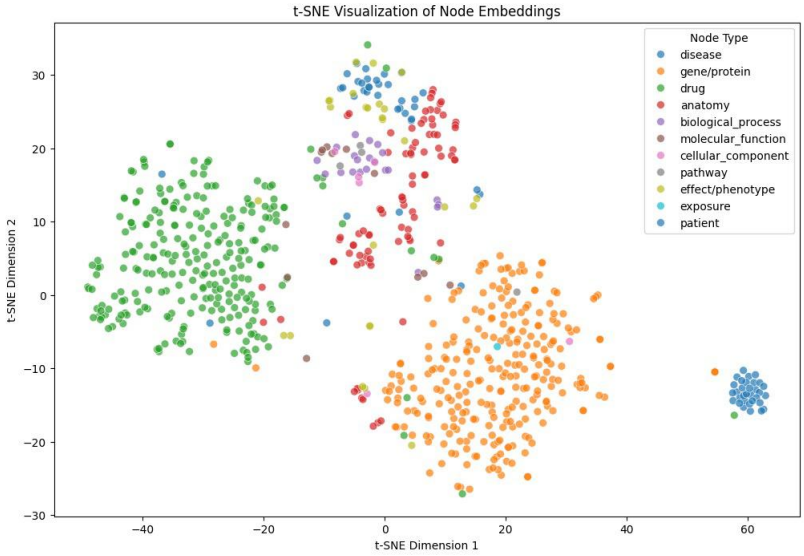

Figure 4: Node Embeddings

The t-SNE visualization highlights the node embeddings from using BioBERT on our super-graph. The t-SNE algorithm reduces the high-dimensionality of the BioBERT embeddings into a two-dimensional space, enabling us to represent the relationships between node types more easily. The clusters indicate that BioBERT is able to capture the structural similarities between nodes and form logical groupings of related entities. For instance, we see that the green drug points are generally in

the same region, while patients are more tightly clustered in the far right of the graph. The distinct separations between most of the clusters suggests that our model can parse biomedical information.

3.7 Identification of Theorized Responders

Building on the embeddings and similarity scores generated through our super-graph, we developed predictive models to classify patients based on their likelihood of being administered TNF-α inhibitors. While TNF-α served as an example case study for this project, the framework we developed is broadly generalizable and can be applied to other drugs, mechanisms of action, and patient populations. Our approach sought to identify patterns in responder and untreated patient profiles, with a particular focus on patients who were not treated with TNF-α inhibitors but shared characteristics with known responders. This analysis aimed to uncover overlooked or emerging responder profiles and refine our understanding of broader patient populations.

3.7.1 Complementary Approaches

We pursued two complementary approaches to achieve this. First, we trained a Random Forest model using the concatenated embeddings generated through BioBERT and Node2Vec. This method allowed us to leverage graph and text-based embeddings to classify treated vs. untreated patients with high performance. Second, we implemented a more advanced heterogeneous neural network (HNN) to capture the intricate relationships between patients, drugs, diseases, and clinical features in the super-graph. The HNN utilized the graph structure to propagate information across connected nodes, learning richer representations and yielding results consistent with the Random Forest model. The alignment of insights across these two distinct methodologies underscores the robustness of our framework.

More generally, the heterogeneous neural network approach can be represented as follows:

$$h_v^{(l+1)} = \sigma\left(\sum_{(u,v,r)\in\mathcal{E}_r} \mathbf{W}_r \cdot \mathrm{AGG}_r\left(h_u^{(l)}\right)\right) \tag{1}$$

- $h_v^{(l+1)}$: The updated feature vector of node v at layer $l+1$.
- $\mathcal{E}_r$: The set of edges of type r in the heterogeneous graph.
- $h_u^{(l)}$: The feature vector of the neighboring node u connected to v via edge type r at layer l.
- AGG_r: The aggregation function (e.g., mean, sum) specific to edge type r, applied to the neighbors' features.
- $\mathbf{W}_r$: A learnable weight matrix specific to edge type r, transforming the aggregated message.
- σ: A non-linear activation function (e.g., ReLU or sigmoid), applied to the combined messages.

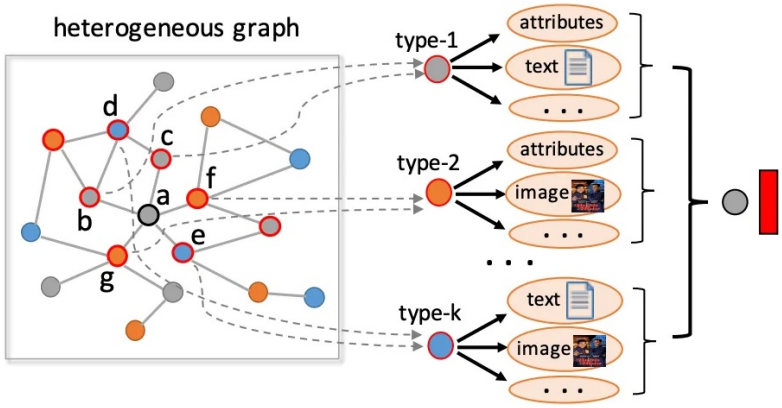

Figure 5: Illustrative heterogeneous neural network [4]

Using the model outputs, we analyzed patients who had not actually been prescribed a TNF-α inhibitor but who were predicted by the models to have non-zero probabilities of having been prescribed such a drug (i.e., identify the non-treated patients who exhibited strong biomedical and clinical similarities with known responders). These patients represent potential opportunities to expand the treatment cohort, as they share similarities with known responders despite not being treated.

3.7.2 Mitigation of Data Leakage Risk

To ensure the validity of our predictions, we implemented rigorous safeguards during feature engineering to prevent data leakage. For example, we imputed BioBERT embeddings for patient and hospital admission (HADM) nodes as the weighted average of BioBERT embeddings from their immediate neighbors and second-degree neighbors in the graph. This approach allowed us to provide rich contextual embeddings while limiting the risk of models overfitting to non-actionable patient insights and avoiding direct reliance on patient-specific or drug-specific features. By assigning higher weights to immediate neighbors, the embeddings emphasized local relationships while still capturing broader graph context.

Additionally, we masked all direct connections between patients or HADMs and the TNF-α inhibitors in the graph. This step ensured that the model could not exploit direct associations, forcing it to learn meaningful patterns from the broader graph structure and indirect relationships. These precautions minimized the risk of data leakage and ensured that predictions were grounded in generalizable insights rather than spurious correlations.

3.8 Mantis

We drew further inspiration from the Mantis tool, developed by MIT Professor Kellis' Research Group and the MIT CSAIL. Similar to our t-SNE graph, Mantis leverages AI tools to transform complex high-dimensional data into an interactive, visual map. We illustrate an example of how we used Mantis below:

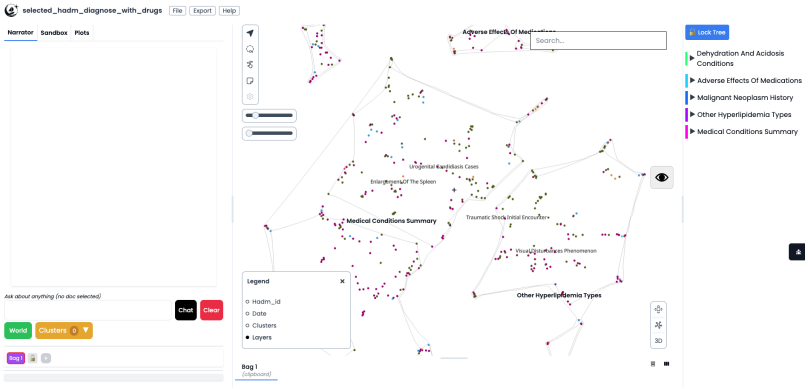

Figure 6: Node Embeddings

In Figure 4, we utilized Mantis to visualize patient data embeddings from the MIMIC-IV dataset. These embeddings connect prescribed drugs and diagnoses to each unique hospital admission ID. The Mantis map highlights clusters of related medical conditions and treatments, providing an intuitive way to explore patterns and correlations within the data. For instance, it reveals groupings of hyperlipidemia cases, adverse medication effects, and other specific conditions, showing deeper insights into patient care and treatment relationships. By interacting with the map, we can delve into individual nodes, uncovering specific diagnoses or drug associations tied to particular clusters or admissions.

4 Findings

4.1 TNF-α case study

To further interpret the factors driving our model's predictions for would-be responders, we applied a decision tree to demographic features and clinical data on the untreated patients, specifically bringing insights to life on the characteristics dis-

tinguishing our models' theorized responders vs. non-responders. The decision tree provided an interpretable framework to uncover key characteristics influencing the predictions.

The decision tree analysis revealed several actionable insights. Age and comorbidities emerged as significant predictors, with older patients and those with specific underlying conditions more likely to resemble responder profiles. Gender and ethnicity also played a critical role, highlighting potential disparities in treatment patterns that warrant further investigation. Additionally, hospitalization context, such as ICU admissions and length of stay, was a strong determinant of predicted probabilities, underscoring the importance of clinical context in evaluating treatment suitability. These findings suggest that expanding treatments such as TNF-α inhibitors to previously overlooked patient populations could improve patient outcomes and refine clinical trial inclusion criteria.

This analysis bridges advanced predictive modeling and actionable insights, offering a pathway to extend treatments to new patient segments. While TNF-α served as a focused case study, the methods and frameworks developed here are broadly generalizable and can be applied to other drugs and mechanisms of action. By integrating this approach into its broader decision-making framework, Sanofi can better target emerging responder profiles, improve resource allocation, and reduce missed opportunities in patient care.

Analyzing the model outputs, we identified 167 not treated with TNF-α inhibitors—who were predicted by the models to have non-zero probabilities of receiving these drugs. These patients represent potential opportunities to expand the treatment cohort (including potential label expansion), as they share similarities with known responders despite not being treated. To interpret the factors driving these predictions, we applied a decision tree to demographic features of untreated patients, providing an interpretable framework to uncover key characteristics influencing the predictions.

The decision tree analysis revealed several actionable insights. The presence of ICD Code 556.9 (Ulcerative Colitis) was associated with a higher probability of being a potential responder, aligning with the clinical use of TNF-α inhibitors in treating inflammatory conditions like ulcerative colitis. Similarly, patients with ICD Code 730.08 (Acute Osteomyelitis) were more likely to be identified as potential responders. This finding is consistent with evidence exploring the use of TNF-α inhibitors in the management of osteomyelitis, particularly in cases where conventional treatments are insufficient.

Age emerged as another significant factor, with older patients—particularly those aged 65 and above—more likely to be classified as potential responders. This reflects the increased prevalence of conditions such as rheumatoid arthritis in older populations, where TNF-α inhibitors are commonly prescribed [7][3]. Additionally, certain clinical conditions such as pyonephrosis also influenced the classification. Pyonephrosis, a severe kidney infection characterized by pus accumulation in the renal pelvis, often results from urinary tract obstruction or pyelonephritis. While this condition's association with potential responders may suggest expanded use of TNF-α inhibitors, it also underscores the need for caution due to the potential risks involved. Careful monitoring and management are critical in such cases [9].

These findings suggest that expanding the use of TNF-α inhibitors to previously overlooked patient populations could improve outcomes and refine clinical trial inclusion criteria.

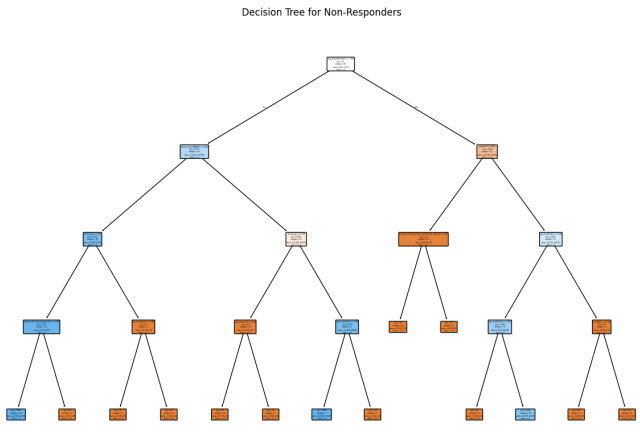

Figure 7: Decision tree: Predicting theorized responders among untreated patients (*Illustrative*)

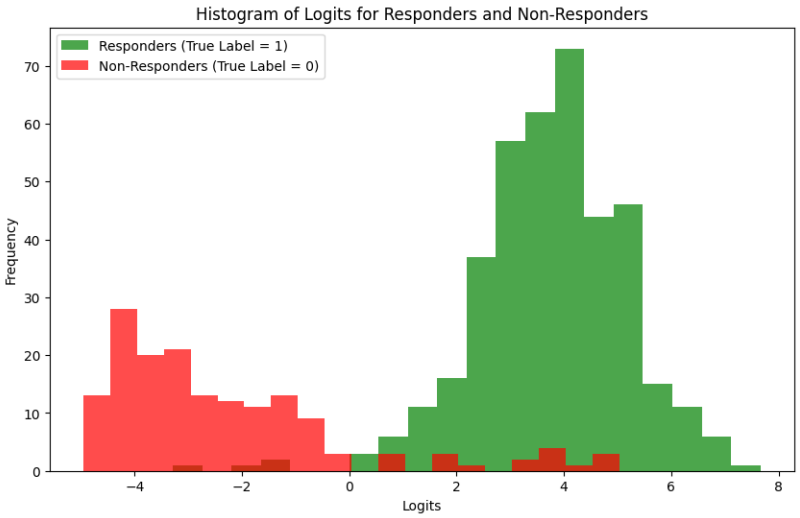

Figure 8: Predicted probability of being a known responder (HNN-derived)

4.2 Broader Implications

This analysis bridges advanced predictive modeling and actionable insights, offering a pathway to extend treatments to new patient segments. While TNF-α inhibitors served as a focused case study, the methods and frameworks developed here are broadly generalizable and can be applied to other drugs and mechanisms of action. By integrating this approach into its broader decision-making framework, Sanofi can better target emerging responder profiles, improve resource allocation, and reduce missed opportunities in patient care.

In summary, our careful feature engineering, dual modeling approaches, and alignment with biomedical knowledge have provided a robust framework for identifying and understanding untreated patient profiles. This framework holds significant potential for broad application across various therapeutic areas, ultimately enhancing patient outcomes and informing strategic decision-making.

5 Business and Patient Impact

Our team's results enable Sanofi to identify patient super-responders, expedite the drug development process, and improve health outcomes for different patient segments. We explore each of these areas in greater detail.

5.1 Patient Super-Responders

The patient similarity scores allow Sanofi to identify super-responder user profiles who are likely to significantly benefit from specific drug mechanisms of action. Sanofi can make more informed decisions about treatment allocation and optimization, with the goal of improving patient outcomes. Additionally, Sanofi could make informed predictions (while still leveraging established, compliant methodologies) on which patient segments are most likely to exhibit adverse reactions to particular TNF-α inhibition drugs, helping improve trial inclusion/exclusion criteria. Overall, our insights enable Sanofi to provide more personalized healthcare plans for patients and design future drugs with specific patient characteristics in mind.

5.2 Drug Development

Our team's work accelerates the drug development process because Sanofi can interpret relations between drugs, diseases, and patient features. Given the discussion surrounding super-responders, Sanofi can adjust its randomized clinical trials to include patients most likely to benefit from novel medications. Consequently, there is greater probability of successful trial outcomes, enabling Sanofi to save significant amounts of time and resources. The super-graph will also highlight complex connections

between various diseases and underlying patient biological conditions, which could assist medical professionals in determining which genes, proteins, or specific parts of the body are being compromised.

5.3 Social Implications

Sanofi's potential to provide personalized treatment plans has profound social implications, particularly in improving patient outcomes and reducing mortality risk. Rather than arbitrarily assigning a medication that appears to work best for the average patient, Sanofi can provide tailored drugs for patients based on their biological markers and characteristics. This helps with Sanofi's resource utilization and distribution, and our work can reduce disparities in the healthcare system.

An important aspect of this work is identifying patient profiles within a highly heterogeneous disease population that respond to a particular mechanism of action (MoA). While known MoAs can often be identified using real-world data, such as electronic health records, and further explored through predictive modeling and causal inference, identifying novel MoAs presents a greater challenge. By leveraging biomedical knowledge graphs and MIMIC-IV datasets, we can hypothesize what an ideal patient—one likely to be a super-responder for a drug with a novel MoA—may look like. From this, we aim to identify "patients like me," individuals in the real world who match the ideal super-responder profile for such drugs.

Our analysis and recommendations support a precision medicine approach, as we leverage data-driven results to enhance Sanofi's decision-making process for each patient. By focusing on both known and novel MoAs, this approach facilitates meaningful advancements in patient care, clinical trial design, and resource allocation. Sanofi should continue to iterate on our findings to validate results and uncover actionable connections within the super-graph. This will ultimately enable more effective treatments for diverse patient populations and a more equitable healthcare system.

6 Future Direction

With additional time and computational resources, the inclusion of multiple target proteins distinct from TNF-α in the super-graph would significantly enhance the scope and robustness of this project. Expanding the analysis to other proteins and mechanisms of action (MoAs) would enable a comparative study of how patient clusters differ across treatments. Proteins related to TNF-α could act as secondary connections, linking more patient cohorts and creating a richer super-graph. This expansion would also allow for rigorous validation of the similarity scoring algorithm by testing its sensitivity and consistency across diverse biological contexts.

Another key direction is to integrate real-world outcome data, such as patient responses to treatments and longitudinal health outcomes, into the framework. By comparing actual outcomes against model predictions, we could refine the feature importance rankings and improve the precision of responder predictions. This would also help evaluate and validate the real-world effectiveness of TNF-α inhibitors, ensuring that predictions align with clinical realities.

Moreover, the framework has significant potential for improving decision-making in situations where patients are misdiagnosed or lack precise diagnoses. In such cases, the integration of demographic, clinical, and biomedical data in the super-graph can still provide meaningful insights into the likely treatments or conditions affecting a patient. For example, even if the exact disease is not identified, the embeddings derived from connected clinical features (e.g., ICD codes, admission details) can guide clinicians toward the most relevant interventions. This resilience against incomplete or inaccurate data underscores the utility of the framework in real-world, complex clinical settings.

Integrating additional patient data from the MIMIC-IV database and other biomedical datasets, such as lab results, imaging reports, and clinical notes, would further enrich the super-graph. These unstructured data types could add more edges between patients, diseases, and prescriptions, providing additional anchors for integration with PrimeKG. A more comprehensive graph would yield better-defined patient clusters and more accurate similarity scores.

Building on similarity scores derived for TNF-α-related patients, the next step involves evaluating treatment effects in specific subpopulations, such as those with rheumatoid arthritis. Comparing outcomes for patients treated with TNF-α inhibitors against untreated controls would highlight treatment effects and identify super-responders. These insights could refine patient selection for clinical trials, improving both treatment strategies and trial design.

Finally, this approach provides a framework for advancing precision medicine. By identifying patient clusters and expanding the analysis to include misdiagnosed or ambiguous cases, we can enhance the reach and efficacy of treatments. This will help Sanofi improve patient outcomes, optimize clinical trials, and make data-driven decisions to address unmet medical needs.

7 Conclusion

Our project with Sanofi Pharmaceuticals reflects the value of utilizing cutting-edge data science algorithms and methodology to optimize treatment allocation and assist with the drug development process. By creating a super-graph with PrimeKG and MIMIC-IV data, we bridge knowledge between drug mechanisms, diseases, and patient outcomes, with a focus on TNF-α inhibition drugs. We derived meaningful insights into patient-drug similarity scores and observed patient outcomes, advancing Sanofi's mission to leverage artificial intelligence in drug research and development. In the future, we recommend Sanofi compares treatment effects with randomized clinical trial data to achieve its goal of finding super-responders to TNF-α drugs. We hope and expect Sanofi will maintain its lead in developing data-driven healthcare solutions that positively impact its consumers and enable the firm to outperform peers.

References

[1] Alistair Johnson et al. Medical information mart for intensive care (mimic)-iv. *PhysioNet*, 2024.

[2] Dan in Jang et al. The role of tumor necrosis factor alpha (tnf-) in autoimmune disease and current tnf- inhibitors in therapeutic. *National Library of Medicine*, 2021.

[3] Jeffrey R. Curtis Joshua F. Baker, Michael D. George. Achieving equity in screening for latent tuberculosis infection in rheumatology practice. *The Journal of Rheumatology*, 85(27):27–33, 2023.

[4] Arthur Lee. Kdd '19: Heterogeneous graph neural network. *Medium*, 2019.

[5] American College of Rheumatology. Tumor necrosis factor (tnf) inhibitors. *Rheumatoid Research Foundation*, 2024.

[6] Marinka Zitnik Payal Chandak, Kexin Huang. Precision medicine oriented knowledge graph. *Harvard Zitnik Lab*, 2023.

[7] Osman Atalay Gülsah Köksal Ahmet Karaaslan Payal Desai, Ferah Bozkurt. Ochronotic arthropathy: A case of brown spine with a review of the literature. *Archives of Rheumatology*, 28(3):248–252, 2013.

[8] Sanofi Pharmaceuticals. Artificial intelligence in research and development. *Sanofi Pharmaceuticals Post*, 2023.

[9] Anne Barton Sarah Meade, Suzanne Anderson. Recent advances in the genetics of rheumatoid arthritis. *Rheumatology International*, 42:1013–1025, 2022.