Unraveling the Hepatitis B Cure: A Hybrid AI Approach for Capturing Knowledge about the Immune System's Impact

Shahi Dost*
shahi.dost@tib.eu
TIB Leibniz Information Centre for Science and Technology
Germany

Ariam Rivas*
ariam.rivas@tib.eu
L3S Research Centre and Leibniz University of Hannover
Germany

Hanan Begali*
begali.hanan@mh-hannover.de
Department of Gastroenterology, Hepatology, Infectious Diseases and Endocrinology, Hannover Medical School
Germany

Annett Ziegler
annett.ziegler@twincore.de
TWINCORE, Centre for Experimental and Clinical Infection Research
Germany

Elmira Aliabadi
aliabadi.elmira@mh-hannover.de
Department of Gastroenterology, Hepatology, Infectious Diseases and Endocrinology, Hannover Medical School
Germany

Markus Cornberg
cornberg.markus@mh-hannover.de
Department of Gastroenterology, Hepatology, Infectious Diseases and Endocrinology, Hannover Medical School
Germany

Anke RM Kraft*
Kraft.Anke@mh-hannover.de
Department of Gastroenterology, Hepatology, Infectious Diseases and Endocrinology, Hannover Medical School
Germany

Maria-Esther Vidal*
maria.vidal@tib.eu
TIB Leibniz Information Centre for Science and Technology, L3S Research Centre, and Leibniz University of Hannover
Germany

ABSTRACT

Chronic hepatitis B virus (HBV) infection is still a global health problem, with over 296 million chronically HBV-infected individuals worldwide. The merging data about clinical parameters, immune phenotyping data, and genetic information, together with AI models reliant on this integrated information, holds promise in effectively predicting the likelihood of functional cure in HBV-infected patients. Yet, the limited size of multidimensional datasets and characteristic of HBV cases poses a challenge for machine learning (ML) systems that typically require substantial data for pattern recognition. This paper addresses this challenge by introducing *HyAI*, a hybrid AI framework. *HyAI* employs knowledge graphs (KGs) and inductive learning to unearth meaningful patterns. *HyAI* relies on KG embedding models to learn a numerical representation of the *HyAI* KG in a k-dimensional vector space. Through community detection methods, closely related HBV patients are clustered using similarity metrics formulated from the acquired embeddings. *HyAI* is studied in a population of HBV patients integrated with multidimensional datasets. Our empirical analysis shows that *HyAI* uncovers immune markers that, together with clinical and demographic parameters, correspond to good predictors for forecasting the cure of chronic HBV infection.

CCS CONCEPTS

• **Information systems** → **Data management systems**.

KEYWORDS

Knowledge Graphs, Inductive Learning, Knowledge Graph Embedding, Community Detection, Hepatitis B Virus Infection

ACM Reference Format:
Shahi Dost, Ariam Rivas, Hanan Begali, Annett Ziegler, Elmira Aliabadi, Markus Cornberg, Anke RM Kraft, and Maria-Esther Vidal. 2023. Unraveling the Hepatitis B Cure: A Hybrid AI Approach for Capturing Knowledge about the Immune System's Impact . In *Knowledge Capture Conference 2023 (K-CAP '23), December 05–07, 2023, Pensacola, FL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3587259.3627558

1 INTRODUCTION

Chronic hepatitis B virus (HBV) infection is a global health concern. Worldwide, more than 296 million people are chronically infected with HBV, leading to 820,000 deaths annually [1]. Treatment with nucleos(t)ide analogues (NA) or interferon alpha inhibits HBV replication in chronically infected patients and slows disease progression to hepatocellular carcinoma (HCC). *Functional cure*, defined as hepatitis B surface antigen (HBsAg) loss, is the goal

* Shahi Dost, Ariam Rivas, and Hanan Begali are equally contributing first authors of this paper. Anke Renate Maria Kraft and Maria-Esther Vidal are equally contributing last authors and Maria-Esther Vidal is corresponding author.

K-CAP '23, December 05–07, 2023, Pensacola, FL, USA

ACM ISBN 979-8-4007-0141-2/23/12.
https://doi.org/10.1145/3587259.3627558

[1] https://www.who.int/news-room/fact-sheets/detail/hepatitis-b

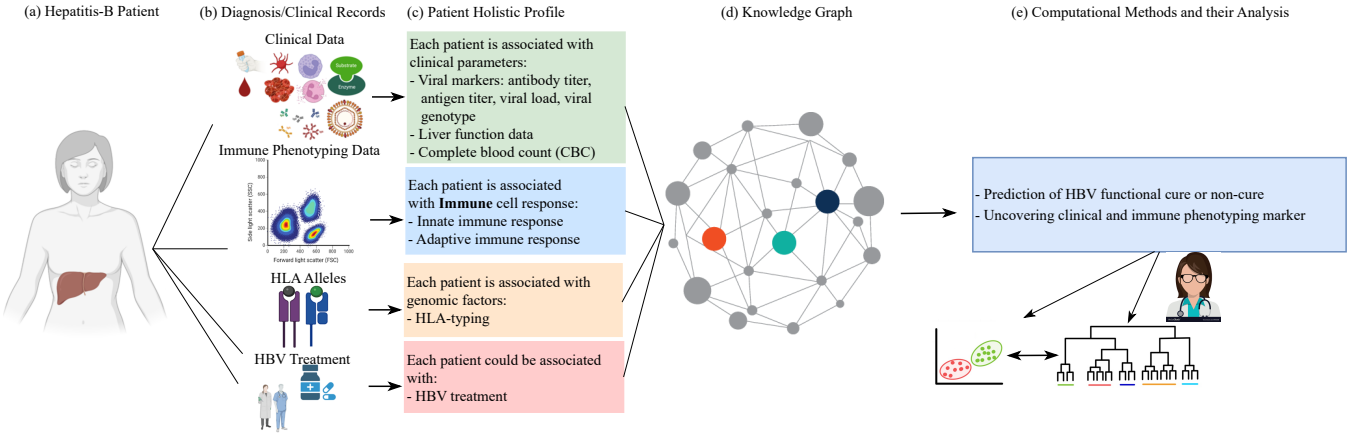


Figure 1: Motivating scenario with data scattered across heterogeneous data sources (clinical data, immune phenotyping data, HLA alleles, HBV treatments), preventing a holistic analysis of HBV patients. A knowledge graph represents factual statements facilitating inductive learning to uncover patterns and to enhance the understanding of parameters impacting HBV outcomes.

of HBV treatment, but is rarely achieved in these patients ($\leq$ 1% per year), and lifelong treatment is often required. Chronic HBV is a very heterogeneous disease, divided in different phases based on various clinical parameters, e.g., HBsAg, hepatitis B envelope antigen (HBeAg), HBV deoxyribonucleic (HBV DNA) and alanine aminotransferase (ALT) levels [9]. It is shown that the immune system is important to control HBV infection [15]. However, it is still not fully understood which parameters and its combinations are associated with achieving HBV functional cure. Thus, a deep understanding of clinical and immune parameters is required.

Knowledge graphs (KGs) and machine learning (ML) techniques have demonstrated significant potential in capturing intricate patterns. These patterns, when harnessed, can greatly enhance our comprehension of various diseases, leading to more precise advancements in diagnosis, personalized treatment recommendations, and accurate outcome predictions [22, 23]. Notably, ML models have demonstrated notable accuracy in addressing HBV-related issues, such as early detection [1], risk assessment for HBV [12], and predicting HBsAg seroclearance [28]. While ML techniques hold significant promise, achieving precise model training and reproducibility often requires substantial population sizes [29]. Transfer learning (TL) has demonstrated effectiveness in scenarios with limited training data. However, its implementation demands intricate configurations and hyperparameter tuning, which may not always be feasible, particularly when dealing with small datasets. This situation was investigated using a small and unique HBV dataset resulting from the integration of clinical, demographic, and immune phenotyping data from HBV-infected patients.

Problem Statement: This paper addresses the problem of partitioning a set of data points, wherein these data points are represented as entities within a knowledge graph (KG). The primary objective is to devise a partitioning strategy that optimizes the cohesion of related entities within the same partition, thereby maximizing their interrelatedness. Simultaneously, the strategy aims to minimize the connections between entities situated in different partitions, consequently reducing their overall inter-partition relatedness.

Proposed Solution: We present *HyAI*, a hybrid AI system that employs inductive learning to identify a partitioning of nodes within a KG. This partitioning addresses the problem of maximizing intra-community similarity while minimizing similarity between entities in separate communities. *HyAI* seamlessly integrates self-supervised and unsupervised learning. The self-supervised methods involve KG embedding models that transform entities and relations from the input KG into k-dimensional vectors, thereby preserving their structural information. These embeddings are then utilized to quantify the relatedness between entities using a similarity metric. These similarity values serve as the foundation for the subsequent unsupervised learning models, which identify the communities within the partitioned entities. Following the methodology proposed by van Bekkum et al. [25] to design hybrid AI systems, *HyAI* is specified using a design pattern that integrates the process of KG creation, with the design patterns of the KGE and community detection models. We demonstrate *HyAI* in the context of HBV to uncover patterns fulfilled by patients experiencing functional cure.

Evaluation: We have empirically studied the performance of *HyAI* over a KG that represents holistic profiles of 87 chroncially HBV infected patients, created from the integration of heterogeneous data sources comprising demographic (e.g., sex and age), clinical (e.g., HBcrAg, HBeAg, and HBsAg) and immune phenotyping parameters. The results put into perspective the benefits of capturing knowledge from different data sources into a KG. Moreover, the analysis of the communities' quality detected by *HyAI* enables uncovering patterns that provide evidence of the importance of parameters like age, HBsAg levels, and immune cells in forecasting HBsAg loss. **Contributions:** This work presents the following contributions: **(1)** *HyAI*: We introduce *HyAI*, a hybrid AI system designed to enhance community detection algorithms by leveraging the knowledge captured within the low-dimensional representation of KG. **(2)** Problem Modeling: We formulate the problem of identifying markers associated with HBV functional cure (specifically, HBsAg loss). **(3)** Empirical Evaluation: Our approach is empirically evaluated using real-world data from 87 HBV-infected patients, incorporating their diverse and multidimensional datasets.

The rest of the paper is structured as follows: Section 2 presents

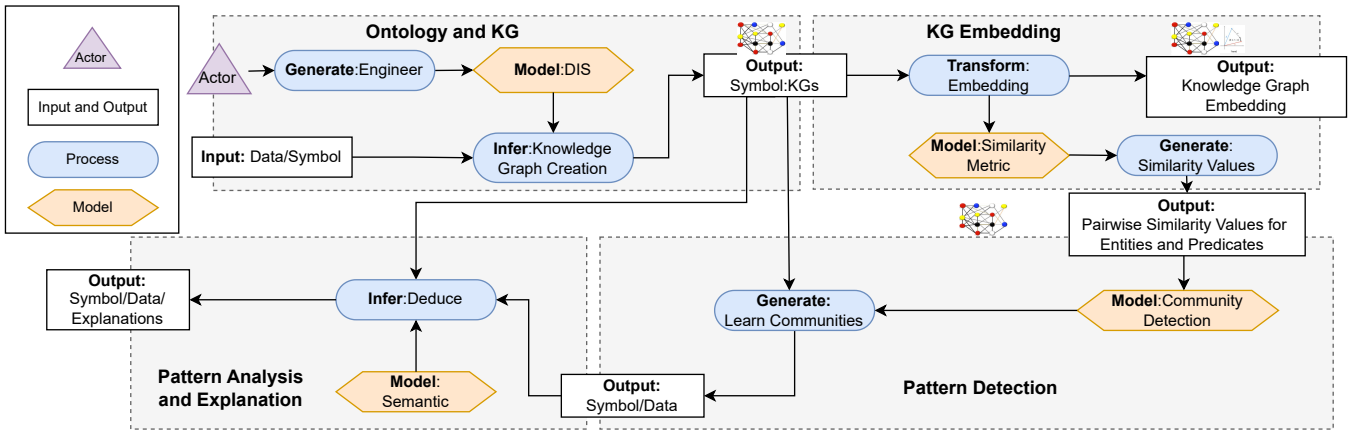


Figure 2: Design patterns describing *HyAI*

basic concepts and a motivating example. Section 3 defines our proposed approach and solution, and illustrates *HyAI* in the context of HBV. Results of the empirical evaluation are reported in Section 4 and the state of the art is briefly analyzed in Section 5. Finally, we close with the conclusion and future work in Section 6.

2 PRELIMINARIES AND MOTIVATION

Knowledge Graphs: A *knowledge graph (KG)* is a directed *edge-labeled graph* $KG = (V, E, L)$, where **i)** nodes in V and labels in L are subsets of a set countable infinite constants; and **ii)** edges in E correspond to the subset of $V \times L \times V$. In the *HyAI* KG, nodes represent patients and the values of the parameters, while edges represent parameters and markers (e.g., *hasAge*, *sufferFromDisorder*). A *KG* can be defined as a data integration system $DIS_{KG} = \langle O, S, M \rangle$, where, O is a unified schema comprising classes and properties, S is a set of data sources, and M corresponds to mapping rules formulated as conjunctive queries over the sources in S. Rules in M can be declaratively specified in mapping languages, e.g., RML[2].

Knowledge Graph Embedding (KGE) Representation: A KGE model learns functions ϵ and ρ, which respectively map nodes in E and edges in V to k-dimensional vector representations. These vectors are constituents of the set $\mathbb{T}$. A plausibility score function ϕ comes into play, acting as a partial function $\mathbb{T} \times \mathbb{T} \times \mathbb{T} \rightarrow \mathbb{R}$; it is used to assess the credibility of a given triple. Given a triple $t = (s, p, o) \in V \times L \times V$, the evaluation of $\phi(\epsilon(s), \rho(p), \epsilon(o))$ computes the plausibility of the triple t. The functions ϵ and ρ have the role of capturing the inherent structural relationships within the knowledge graph, as outlined in the KG literature by [7]. State-of-the-art KGE models include TransE [7], TransH [27], RESCAL [18], and ERMLP [10]. These KGE models leverage distinct scoring functions to effectively encode knowledge graphs into vector representations within the space $\mathbb{T}$.

Community Detection: Methods aimed at community detection involve the partitioning of a KG into subgraphs comprised of densely connected and similar nodes. Leading-edge techniques in this field encompass SemEP [19], METIS [14], and KMeans [6]. The evaluation of detected communities relies on assessment metrics such as Conductance [11], Coverage [11], and Total Cut [8], are used to check the quality of detected communities.

Motivating Scenario: Clinical research indicates that the progression of HBV infection is influenced by both the duration of infection and the level of HBsAg [9]. Recent assessments have also highlighted the insufficiency of relying on isolated factors for accurately predicting HBV functional cure. Therefore, there is a compelling need to gather comprehensive information regarding the clinical and immunological parameters that characterize an HBV patient, as this is pivotal for predicting HBV functional cure. However, this disease related data often exists across multiple sources (depicted in Figure 1a), encompassing demographics, clinical records, immune phenotyping data, HBV treatments, HLA alleles, and other relevant datasets from diverse healthcare facilities. Consequently, there arises a requirement for creating holistic profiles of HBV patients, as depicted in Figure 1b) and 1c). HBV patient profiles might consist of differing parameters, preventing, thus, the adoption of a fixed schema, where all patients are characterized by uniform features. This semi-structured nature of HBV patient profiles results in a substantial number of missing values, if represented as a universal relational table. A more effective representation can be achieved by portraying them as factual statements within a KG framework, as shown in Figure 1d). This shift would enable inductive learning methods to encode individual profiles and unveil patterns facilitating the identification of viral markers, cell markers, and predictions regarding HBV functional cured or non-cured patients, as depicted in Figure 1e). This paper introduces an AI framework designed to seamlessly combine data from diverse sources, while presenting the merged information in two distinct forms: factual statements and a low-dimensional continuous vector space. Symbolic and numerical representations serve as fundamental components for modeling relationships between entities within the KG. They are building blocks for unraveling patterns through advanced community detection techniques. The goal is to leverage these patterns to shed light on features that elucidate the conditions of HBV cured patients.

[2]https://rml.io/

Table 1: Statistics of Data Integration System

Data Integration System	Records
Mapping Assertions	559
Classes	66
Object Properties	185
Data Properties	177

Table 2: Statistics of *HyAI* Knowledge Graph

Benchmark	# Triples	# Entities	# Relations
HyAI KG	32964	5208	82

Table 3: Summary of *HyAI* Knowledge Graph

Knowledge Graph for HyAI	Records
Patients	87
Demographic Parameters	2
HBV Treatment	6
Number of Clinical Parameters	18
Number of Immune Phenotyping Parameters	45

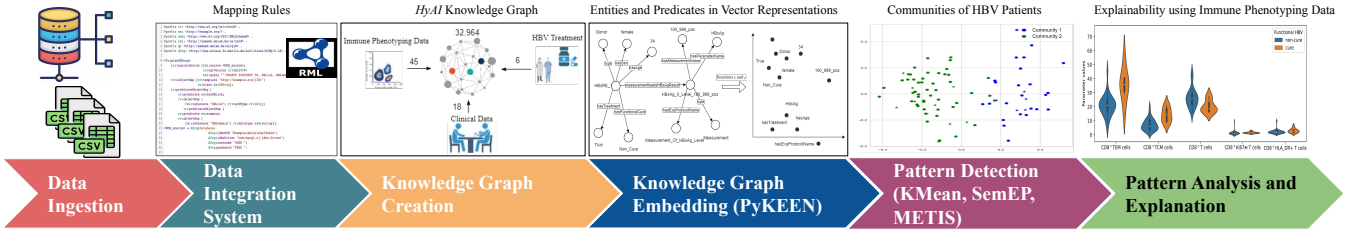


Figure 3: The pipeline of *HyAI* framework, tackles the problem of predicting the impact of the immune system in the functional cure and non-cure of chronic HBV patients. HyAI seamlessly integrates selfsupervised learning involves knowledge graph embedding models and unsupervised learning, which identifies the communities based on similarity values.

3 *HYAI*- OUR APPROACH

In this section, we formalize the problem tackled in this work and present the architecture of our proposed solution.

Problem Statement: We tackle the problem of detecting communities of nodes in a knowledge graph KG, such that the nodes inside the same community are very related, while nodes in different communities are not. Let $KG = (V, E, L)$ be a knowledge graph and let $\mathcal{S}$ be the space of all the partitions into communities of the nodes in V. Let $Quality{:}\mathcal{S} \rightarrow \mathbb{R}$ be a utility function that captures both the cohesion within communities and the separation between communities of the input partition of the nodes in V. The goal is to find the partitioning $C^* \in \mathcal{S}$ that maximizes $Quality(.)$.

$$C^* = argmax_{C \in \mathcal{S}} Quality(C)$$

Proposed Solution: We propose *HyAI*, a hybrid AI system that solves the *problem of detecting communities* on the knowledge graph $KG = (V, E, L)$ that integrates heterogeneous data sources. In addition to the symbolic representation of the entities in V and their relationships in E provided by the factual statements of KG, a KG embedding model generates their numerical representation in a k-dimensional vector space $\mathbb{T}$ in terms of a score function ϕ. Using a similarity metric v, these numerical representations are utilized to determine relatedness between entities in V. Values of similarity provide the basis for creating a partition C in the space $\mathcal{S}$ that corresponds to a solution of *problem of detecting communities*. Communities in C are explored to identify intra- or inter-community properties, and uncover shared patterns.

HyAI is conceptualized following the design principles outlined by Bekkhum et al. [25]; a basic vocabulary allows for representing the components of these patterns: *actor* (indigo triangle), *input and output* (white rectangle), *process* (blue oval rectangle), and *models* (yellow hexagon). Figure 2 depicts *HyAI*; it comprises four design patterns describing the sub-systems that implement the tasks mentioned above towards solving *problem of detecting communities*.

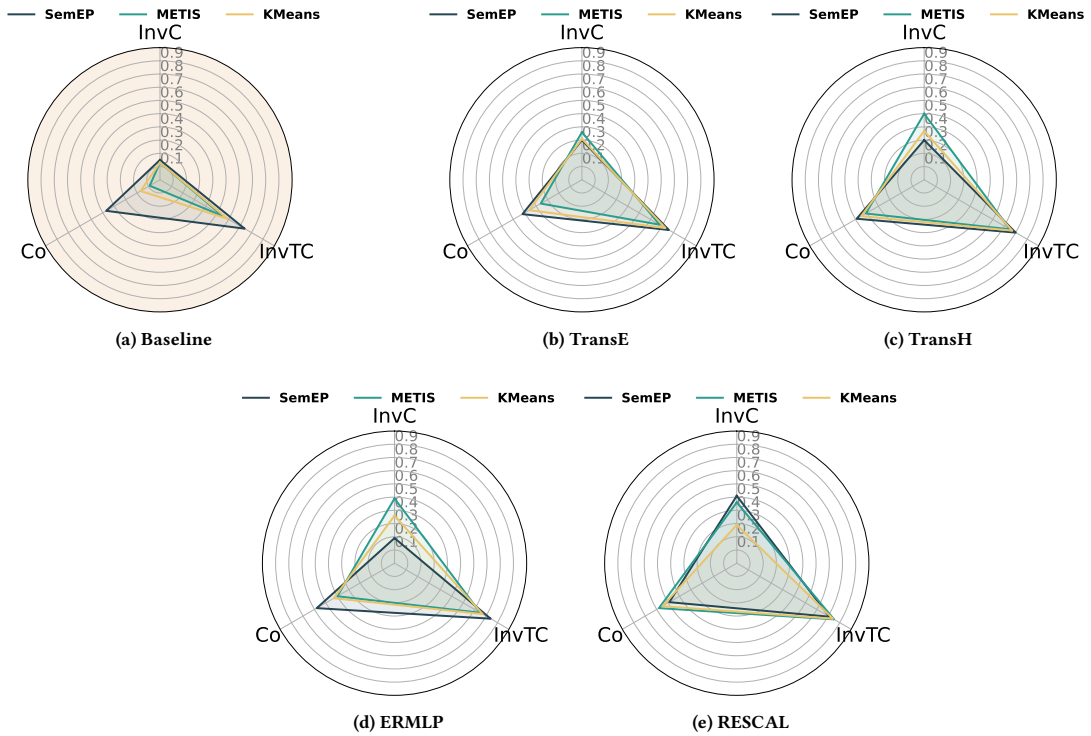
Ontology and KG: In this pattern, actors (e.g., domain experts or knowledge engineers) design a data integration system $DIS_{KG} = \langle O, S, M \rangle$ composed by a unified schema O, input data sources S, and mapping rules M; it is given as input to the process of *Knowledge Graph Creation* which performs a bottom-up evaluation of the mapping rules in M on the sources in S to generate the *HyAI* KG.

KG Embedding: This pattern represents the system that learns the k-dimensional vector representations of entities in V (i.e., $\epsilon(.)$) and labels of properties in E (i.e., $\rho(.)$). Using a particular KGE model, a score function ϕ is followed to learn the encoding of triples $t = (s, p, o)$ where $\phi(\epsilon(s), \rho(p), \epsilon(o))$ is maximized. The k-dimensional vectors are utilized to compute values of relatedness for entities in V and labels of properties in E. A similarity metric v is computed pair-widely in E and in labels of V. This system outputs both the embedding representations and the similarity values.

Pattern Detection: This system follows a community detection algorithm to partition V into a set of communities C, in a way that the values of a utility function $Quality(C)$ are maximized. Entities in each community C_i in C are described in terms of the subgraphs of the *HyAI* KG reachable from each of them; they provide the basis for the profiling and analysis of the entities in V.

Pattern Analysis and Explanation: This pattern designs a system for explaining knowledge captured by the detected communities using statistical and symbolic statements. A semantic based model on top of SPARQL engines enables the traversal of the *HyAI* KG.

Use Case (Understanding HBV Patients): *HyAI* is used to uncover parameters (clinical, demographic, or immune phenotyping data) that may characterize HBV patients with functional cure. Figure 3 shows the implementation of *HyAI* using state-of-the-art tools and techniques. Data acquisition captured heterogeneous data from 87 chronic HBV patients, including age, sex, 18 clinical observational parameters, 45 immune phenotyping parameters, and HBV treatment (depicted in Figure 3). The *Ontology and KG* system receives $DIS_{KG} = \langle O, S, M \rangle$ composed of a unified schema O with 66 classes and 185 properties, five data sources in S, and 559 RML mapping assertions in M. As a result of executing DIS_{KG}, the *HyAI* KG is created; it comprises 32,964 RDF triples, 5,208 entities, 82 labels (Tables in Figure 3). A *KG Embedding* model transforms 87 holistic profiles of HBV patients into vector representations of 154 dimensions; a vector-based similarity metric (e.g., cosine similarity, inverse of Euclidean or Manhattan distances) enables the computation of the relatedness between HBV patients. The *Pattern Detection* system utilizes community detection algorithms to identify groups of closely interconnected HBV patients. This process helps in distinguishing between cured and non-cured HBV patients by partitioning them into distinct categories based on their



(a) Baseline (b) TransE (c) TransH

(d) ERMLP (e) RESCAL

Figure 4: Quality of the generated communities. Communities are evaluated in terms of three quality metrics: Inverse Conductance (InvC), Inverse Total Cut (InvTC), and Coverage (Co), using the SemEP, METIS, and KMeans algorithms. In this case, higher values are better. Figures 4b, 4c, 4d, and 4e, assess *HyAI*, and Figure 4a shows the behavior of the baseline. We can observe that the communities' quality generated by *HyAI* performs better than the baseline.

relatedness. Finally, the *HyAI* KG and the computed communities are analyzed to uncover patterns among demographic (e.g., sex and age), clinical (e.g., HBsAg), and immune phenotypic parameters (e.g., $CD8^+$ T cells, $CD8^+$ TCM cells, and $CD4^+$ TCM cells) that may characterize cured and non-cured HBV patients.

4 EXPERIMENTAL STUDY

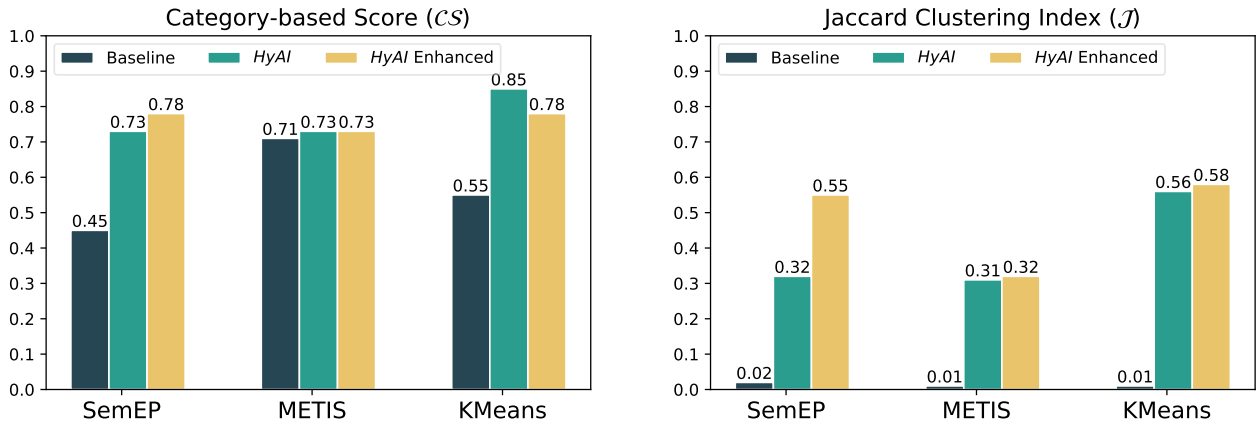
We assess the effectiveness of *HyAI* to capture knowledge encoded in chronic HBV infected patients. In particular, we aim to answer the following research questions: **RQ1)** What is the impact of capturing knowledge from the data integration system and knowledge graph creation on predicting functional HBV cure? **RQ2)** What is the effect of combining distinct embedding techniques and community detection algorithms in detecting functional cured and non-cured HBV patients? **RQ3)** Can *HyAI* capture meaningful patterns for the HBV-domain experts or reported in the literature? The following experimental configuration is set up to answer these questions.

Benchmark: We conduct our evaluation over the *HyAI* KG. This KG consists of observational data describing 87 real-world HBV patients in everyday routine care. The experiments are carried out with observational data where all patients have measured clinical and immune phenotyping parameters for a specific time point called *Day0* [3]. All patients in that observational data cohort are non-cured at the time of data collection. The data cohort integrated into the *HyAI* KG involves demographic parameters, sex and age with categorical values *female* and *male*, and $Young < 40$, $40 \leq Middle \leq 49$, and $Old > 49$, respectively. In addition, the *HyAI* KG describes the HBV patients in terms of clinical parameters with categorical values: **HBsAg** with four categories values ≤ 99, $100 \leq$ HBsAg levels ≤ 999, $1000 \leq$ HBsAg levels ≤ 9999, and HBsAg levels ≥ 10000 (IU/mL). **Others** including albumin, ALT, AST, bilirubin, CRP, GGT, leukocyte, lymphocyte, neutrophil and thrombocyte counts, Quick-Test, INR, HBV-DNA, HBcrAg, HBeAg, Anti-HCV and Fibroscan as well as patient's treatment information. Furthermore, the *HyAI* KG comprises 45 immune phenotyping parameters with continuous values generated by the Department for Gastroenterology, Hepatology, Infectious Diseases, and Endocrinology at Hannover Medical School to analyze and characterize immune cells with the emphasis on HBV-specific T cell responses. We aim to identify communities of patients who can achieve a functional HBV cure and validate whether *HyAI* identifies predictors of functional HBV cure.

Gold Standard: The goal standard (G) partitioning corresponds to the partition of the 87 HBV patients into two groups of patients: 14 HBV cured and 73 HBV non-cured patients [4]. The categorization of HBV patients into these two groups is determined by their final registered status as indicated in the clinical records. We can appreciate the imbalance between the two groups of HBV patients, which is

[3] *Day0* is the observational data point in which patients have been measured for both clinical and immune phenotyping parameters in a specific time and date.

[4] These 14 HBV cured and 73 HBV non-cured patients are considered at the last observational data point.



(a) Category-based Score $\mathcal{CS}$ **(b) Jaccard $\mathcal{J}$**

Figure 5: Quality of the communities is based on the gold standard. Metrics $\mathcal{CS}$ and $\mathcal{J}$ assess the baseline and *HyAI*. The reported results of *HyAI* are obtained with the RESCAL embedding model. We can observe that *HyAI* for the three community detection algorithms, Figure 5a and Figure 5b, outperforms the baseline for both $\mathcal{CS}$ and $\mathcal{J}$.

challenging to obtain communities close to the gold standard.

Metrics: We resort to two types of metrics. The first group measures the quality of a partition's communities. While the second group quantifies the quality of partitioning concerning a gold standard. All the metrics are normalized in the range [0,1]; the inverse scores are reported to ensure that all values are higher is better.

Quantifying the Quality of the Communities. **a) Conductance**: measures relatedness of entities in a community, and how different they are to entities outside the community [11]. Inverse (InvC) is reported. **b) Total Cut**: sums up all similarities among entities in different communities [8]. Inverse (InvTC) is reported. **c) Coverage (Co)**: compares the fraction of intra-community similarities between entities to the sum of all similarities between entities [11].

Quantifying the Quality of a Partition. The utility function *Quality(.)* is implemented with two metrics: Jaccard Clustering Index ($\mathcal{J}$) and Average Category-based Score ($\mathcal{CS}$). $\mathcal{J}(C_1, C_2)$ is computed by the number of pair that are in both C_1 and C_2 divided by the same numerator plus the pair that are in C_1 or C_2, and not in both. $\mathcal{J}$ is defined as follows, where the numerator is computed by the number of two combinations of a set with cardinality n:

$$\mathcal{J}(C_1, C_2) = \frac{\binom{|C_1 \cap C_2|}{2}}{\binom{|C_1|}{2} + \binom{|C_2|}{2} - \binom{|C_1 \cap C_2|}{2}}$$

$\mathcal{J}(C_1, C_2) = 1.0$ if the pairs of patients that appear in C_1 are the same as the pairs of community C_2 and $\mathcal{J}(C_1, C_2) = 0.0$ if there are not pairs of patients that appear together in both communities. $\mathcal{CS}$ compares the gold standard, with the communities generated by the community detection algorithms. Given a community C_1, the Average Category-based Score, $\mathcal{CS}(C_1)$, corresponds to the average of the 'Category-based' measure for each pair of patients in the community C_1. A value equal to 0.0 indicates that there is no intersection between the pairs of patients in community C_1 and the gold standard, whereas a value close to 1.0 represents that almost all the pairs of patients in each community share exactly the same response that the gold standard. $\mathcal{CS}(C_1)$ is defined as follows:

$$\mathcal{CS}(C_1) = \frac{\sum_i^{|G|} \binom{|C_1 \cap G_i|}{2}}{\binom{|C_1|}{2}}$$

Baseline: In our evaluation, we establish a baseline using the HBV data in its relational form. This baseline dataset consists of 87 rows, representing individual HBV patients, along with 154 columns that collectively define their attributes within the same observational data cohort as the *HyAI* KG. Representing heterogeneous data as a relational structure leads to the presence of null values, contributing to a null value rate of 35.1% within the baseline dataset. Our objective is to ascertain whether the inclusion of Ontology & KG, along with KG embeddings, can lead to enhancements in the accuracy of the resolution of the *problem of community detection*.

Implementation: *HyAI* is implemented in Python 3.9 and executed on a GPU NVIDIA GeForce RTX 3060 and Intel Xeon CPUs. The implementation combines self-supervised and unsupervised learning following the hybrid design pattern in Figure 2. *HyAI* resorts to SDM-RDFIzer [13] for transforming HBV data sources into RDF triples executing RML mapping rules in the set M of the data integration system $DIS_{KG} = \langle O, S, M \rangle$. A KGE model computes the k-dimensional vector representations of the *HyAI* KG's entities and properties. The following KGE models–from PyKEEN library [4]–are included in the current version of *HyAI*: TransE [7], TransH [27], RESCAL [18], and ERMLP [10]. The cosine similarity is computed on the vectors created by the KGE models. Moreover, the state-of-the-art community detection solvers SemEP, METIS, and KMeans are used to implement the system described by the pattern *Pattern Detection*. Finally, the communities detected are analyzed and explained, making the process transparent and understandable to the user. The KGE models are configured to produce vectors of 154 dimensions, adhering to the default hyperparameters proposed by

the PyKEEN library [5]. METIS and KMeans are set up to create two clusters, mirroring the number of cluster found by SemEP and the ones in the goal standard. *HyAI* is available at [6].

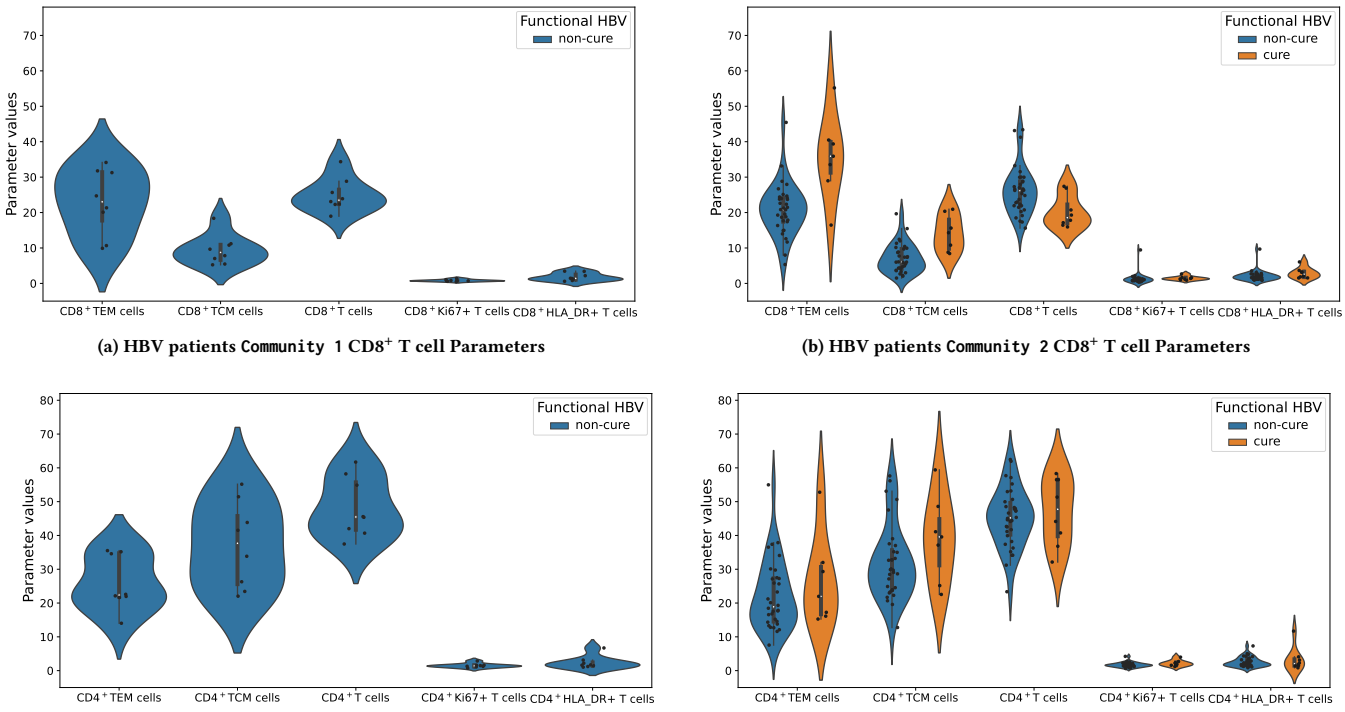
4.1 Effectiveness of *HyAI*

Quality of the Detected Communities: We assess the effectiveness of *HyAI* based on the quality of the communities created on top of the *HyAI* KG based on the knowledge captured by the KGE models and the applied similarity metrics. The quality of the generated partitions is quantified using the three metrics: Inverse of Conductance, Total Cut, and Coverage (Co); their values were computed for the partitions generated by SemEP, METIS, and KMeans over the relational data of the baseline dataset, and the *HyAI* KG. Figure 4a depicts the computed values in a radar plot. We can observe that the *HyAI* results outperform the baseline, when the cosine similarity metric is computed using the embeddings learned by KGE models. This improvement concerning the baseline suggests that these embeddings can encode the structural characteristics of the integrated data, particularly in the case of missing values– which correspond to 35.1% of the values of integrated parameters. They also indicate that the system specified by *Ontology and KG* pattern (Figure 2) allows *HyAI* to capture knowledge about the relatedness of the HBV patients based on the representation of the integrated clinical, demographic, and immune phenotyping parameters (**RQ1**). The HBV dataset includes observational data collected at the data point named *Day0*. Thus, the HBV patients have one value per parameter and clinical protocol used to measure the parameter. Moreover, all the relationships are modeled as binary relations, and reverse triples are not modeled. As a result, the KG includes *n-to-1* relations modeling that several HBV patients have the same value for a given parameter. As shown by Akrami et al. [2], KGE models like TransE. TransH extends TransE, and represents each relation in a hyperplane, performing better than TransE in KGs comprising *n-to-1* [3]. Further, models like RESCAL represents each KG edge as a weight matrix whose entries specify the interaction of latent features, i.e., relationships learned by the KGE model through a training process. This weight matrix representation might perform well for *n-to-1*, as each entry in the matrix models the interaction between latent features of entities [21]. Similarly, ERMLP relies on a positive definite matrix that characterizes the geometry of the relationship space. ERMLP employs metric learning techniques to ensure the learned matrices reflect the desired relationship topology [3]. These characteristics of the KGE models support the results reported in Figure 4 where RESCAL, TransH, and ERMLP improve the quality of communities computed by METIS, SemEP, and KMeans. METIS and SemEP are known for their effectiveness in partitioning graphs and exploiting a KG connectivity [16, 24]. Contrary, KMeans is designed for numerical data and not for graph structures found in KGs. These features of the studied community detection solvers justify the observed behavior of METIS and SemEP, particularly when computed on the embedding learned by RESCAL (**RQ2**). **Quality of Learned Partitions:** We evaluate the generated communities against the gold standard using two metrics: the Average Category-based Score ($\mathcal{CS}$) and the Jaccard Clustering Index ($\mathcal{J}$). Additionally, we expand the *HyAI* KG with four immune system parameters: $CD8^+$ TEM cells, $CD8^+$ TCM cells, $CD4^+$ TCM cells, and $CD8^+$ T cells. These parameters are assigned two categorical values each: *HighRange* and *LowRange*. To establish category thresholds, we rely on the observation that HBV-cured patients typically exhibit high values for $CD8^+$ TEM cells, $CD8^+$ TCM cells, and $CD4^+$ TCM cells, while demonstrating low values for $CD8^+$ T cells. Incorporating these new properties aims to determine if they enhance the models' ability to create partitions that closely resemble the gold standard. Figure 5a and Figure 5b report on the results on the two metrics in the three settings when the cosine similarity is computed using the embeddings learned by RESCAL. As shown in Figure 4, solving the problem of community detection over the *HyAI* KG empowers the community detection solvers with knowledge about the HBV patients that more accurately captures their relatedness. As a result, *HyAI* outperforms the baseline for both Jaccard Clustering Index, Figure 5b, and Average Category-based Score, Figure 5a. Moreover, the *HyAI enhanced*–including the four categorical properties– enables the learning of embeddings that facilitate METIS and SemEP to distinctly categorize non-cured patients within a singular cluster. This outcome consequently brings the generated partitions closer to the established gold standard, as illustrated in Figure 5. This alignment suggests that the categorical representation of $CD8^+$ TEM cells, $CD8^+$ TCM cells, $CD4^+$ TCM cells, and $CD8^+$ T cells represent pivotal immune system parameters, enhancing the capacity to learn embeddings characterized by more accurate values of the plausibility function ϕ (**RQ2** and **RQ3**).
Analyzing Patterns of HBV Patients: RESCAL and SemEP divides the entities of HBV patients into `Community 1` and `Community 2`. The former only comprises entities representing HBV non-cured patients, while the latter includes both cured and non-cured. Figure 6 describes the entities grouped into these two communities based on the values of immune phenotyping parameters $CD8^+$ (Figures 6a and 6b) and $CD4^+$ (Figures 6c and 6d). The frequency of high and low values for these parameters differ in non-cured and cured patients. Specifically, HBV cured patients tend to have higher values of the parameters $CD8^+$ TEM cells, $CD8^+$ TCM cells, and $CD4^+$ TCM cells, and lower values of $CD8^+$ T cells. This observation represents a relevant finding that even requires further clinical study, contributes to the enhancement of the understanding of the role of the immune system in the functional HBV cure (**RQ3**).

5 RELATED WORK

Traditional ML algorithms have been applied to leverage a data-driven approach for predicting tasks related to hepatitis B and C-infected patients using clinical information [26]. Tian X, et al. [28] used different ML algorithms for predicting *HBsAg seroclearance* using demographic and laboratory data of 2, 235 chronic HBV patients. Busayo I, et al. [1] proposed predictive models using ML algorithms for identifying early detection of HBV infections on an interrogate patients dataset of 916 individuals that consists of hematology blood tests, results of HBsAg, and routine clinical test. These approaches rely only on data-driven systems, missing the essential part of capturing knowledge from HBV patients' data including demographic, clinical, and immune phenotyping data.
Knowledge-driven approaches targeting HBV patients' data resources with meaning (metadata) and external knowledge can be

[5] https://pykeen.readthedocs.io/en/stable/tutorial/running_hpo.html
[6] https://github.com/SDM-TIB/HyAI



(a) HBV patients `Community 1` $CD8^+$ T cell Parameters

(b) HBV patients `Community 2` $CD8^+$ T cell Parameters

(c) HBV patients `Community 1` $CD4^+$ T cell Parameters

(d) HBV patients `Community 2` $CD4^+$ T cell Parameters

Figure 6: Immune phenotying values for HBV patients clustered by *HyAI* with RESCAL and SemEP into communities C1 and C2. HBV- cure and non-cure patients have a different frequency distribution of $CD8^+$ and $CD4^+$ T cell subsets; cure patients have higher values of $CD8^+$ TEM cells, $CD8^+$ TCM cells, and $CD4^+$ TCM cells, and lower values of $CD8^+$ T cells.

represented in structured graph models called KGs. These KGs can offer several advantages and make them a powerful tool to organize, harmonize, integrate, and represent complex HBV data with their knowledge. Yin Y, et al. [30] developed a questions-answering system on top of a KG using real medical records and datasets about Chinese medicine diagnosis and treatment of HBV patients. The answers provided by systems used data and knowledge with respect to HBV disease diagnosis, treatment, and patient self-care. Moreover, KGE has gained rapid advances in the design of analytical and predictive tasks in the areas of biomedical and health sciences [5]. Mohamed, S. K, et al. [17] show the capabilities of KGE models in the context of biological KGs and their predictive and analytical capabilities in two use cases of (i) prediction of drug-target interactions and (ii) polypharmacy side effects. A similar hybrid AI system based on data and knowledge-driven approaches proposed by Rivas, A. et, al. [20], which integrate symbolic and sub-symbolic systems represented as deductive databases for link prediction tasks in the use case of KG for lung cancer patients treatment effectiveness. *HyAI* also implements a hybrid AI system and uniquely provides a solution to the problem of discovering parameters that may play a relevant role in the functional HBV cure.

6 CONCLUSIONS AND FUTURE WORK

HyAI implements hybrid AI framework combining KG, KGE, and inductive learning to discover communities of related entities. *HyAI* used KG embedding models to transform entities and relations of KG into k-dimensional vector space by preserving their structure. We used unsupervised learning methods to find similar relationships between entities with the help of similarity metrics. Hence, the community detection algorithms (SemEP, METIS, KMeans) cluster closely related chronic HBV patients via similarity metrics formulated from the acquired embeddings. Our proposed *HyAI* framework has been used in the multidimensional datasets of 87 chronic HBV patients consisting of demographic, clinical, immune phenotyping, and HBV treatment [9, 15]. The experimental analysis shows that the addition of KG and KG embeddings leads to enhancement in the accuracy of detected communities, using KGE models ERMLP, TransE, TransH, and RESCAL, (depicted in Figures 4 and 5). In the future, we are interested in the analysis and effectiveness of *HyAI* in the chronic HBV patients' multidimensional datasets at all time points. We are also interested in using *HyAI* in different use cases associated with human immune system studies.

ACKNOWLEDGMENTS

This work is part of the ImProVIT project funded by Niedersachsen Vorab (project ZN3438) by the Lower Saxony Ministry of Research and Culture and the Volkswagen Foundation. Maria-Esther Vidal is partially supported by Leibniz Association, program "Leibniz Best Minds: Programme for Women Professors", project TrustKG-Transforming Data in Trustable Insights; Grant P99/2020.

REFERENCES

[1] Busayo I Ajuwon, Alice Richardson, Katrina Roper, Meru Sheel, Rosemary Audu, Babatunde L Salako, Matthew O Bojuwoye, Ibraheem A Katibi, and Brett A Lidbury. 2023. The development of a machine learning algorithm for early

detection of viral hepatitis B infection in Nigerian patients. *Scientific Reports* 13 (2023), 100755. Issue 1. https://doi.org/10.1038/s41598-023-30440-2

[2] Farahnaz Akrami, Lingbing Guo, Wei Hu, and Chengkai Li. 2018. Re-evaluating Embedding-Based Knowledge Graph Completion Methods. In *CIKM*. https://doi.org/10.1145/3269206.3269266

[3] Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Mikhail Galkin, Sahand Sharifzadeh, Asja Fischer, Volker Tresp, and Jens Lehmann. 2022. Bringing Light Into the Dark: A Large-Scale Evaluation of Knowledge Graph Embedding Models Under a Unified Framework. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 12 (2022). https://doi.org/10.1109/TPAMI.2021.3124805

[4] Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. 2021. PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research* 22, 82 (2021).

[5] Mona Alshahrani, Maha A Thafar, and Magbubah Essack. 2021. Application and evaluation of knowledge graph embeddings in biomedical data. *PeerJ Computer Science* 7 (2021), e341.

[6] David Arthur and Sergei Vassilvitskii. 2007. K-means++: The Advantages of Careful Seeding. In *ACM-SIAM Symposium on Discrete Algorithms*.

[7] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, Vol. 26. Curran Associates, Inc.

[8] Aydin Buluç, Henning Meyerhenke, Ilya Safro, Peter Sanders, and Christian Schulz. 2016. Recent Advances in Graph Partitioning. In *Algorithm Engineering - Selected Results and Surveys*. https://doi.org/10.1007/978-3-319-49487-6_4

[9] Markus Cornberg, Anna Suk-Fong Lok, Norah A Terrault, Fabien Zoulim, Thomas Berg, Maurizia R Brunetto, Stephanie Buchholz, Maria Buti, Henry LY Chan, Kyong-Mi Chang, et al. 2020. Guidance for design and endpoints of clinical trials in chronic hepatitis B-Report from the 2019 EASL-AASLD HBV Treatment Endpoints Conference. *Journal of hepatology* 72, 3 (2020), 539–557.

[10] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In *Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2623330.2623623

[11] Marco Gaertler and Thomas Erlebach. 2005. *Clustering*. Springer Berlin Heidelberg, Berlin, Heidelberg, 178–215. https://doi.org/10.1007/978-3-540-31955-9_8

[12] V Harabor, R Mogos, A Nechita, AM Adam, G Adam, AS Melinte-Popescu, M Melinte-Popescu, M Stuparu-Cretu, IA Vasilache, E Mihalceanu, A Carauleanu, A Bivoleanu, and Harabor A. 2023. Machine Learning Approaches for the Prediction of Hepatitis B and C Seropositivity. *Int J Environ Res Public Health* 20 (2023), 2380. Issue 3. https://doi.org/10.3390/ijerph20032380

[13] Enrique Iglesias, Samaneh Jozashoori, David Chaves-Fraga, Diego Collarana, and Maria-Esther Vidal. 2020. SDM-RDFizer: An RML interpreter for the efficient creation of RDF knowledge graphs. In *CIKM*.

[14] George Karypis and Vipin Kumar. 1998. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. *SIAM J. Sci. Comput.* 20, 1 (1998). https://doi.org/10.1137/S1064827595287997

[15] Arshi Khanam, Joel V. Chua, and Shyam Kottilil. 2021. Immunopathology of Chronic Hepatitis B Infection: Role of Innate and Adaptive Immune Response in Disease Progression. *International Journal of Molecular Sciences* 22, 11 (2021). https://doi.org/10.3390/ijms22115497

[16] Jure Leskovec, Kevin J. Lang, and Michael W. Mahoney. 2010. Empirical comparison of algorithms for network community detection. In *International Conference on World Wide Web, WWW*. https://doi.org/10.1145/1772690.1772755

[17] Sameh K Mohamed, Aayah Nounu, and Vít Nováček. 2021. Biological applications of knowledge graph embedding models. *Briefings in bioinformatics* 22, 2 (2021), 1679–1693.

[18] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A Three-Way Model for Collective Learning on Multi-Relational Data. In *International Conference on Machine Learning*.

[19] Guillermo Palma, Maria-Esther Vidal, and Louiqa Raschid. 2014. Drug-Target Interaction Prediction Using Semantic Similarity and Edge Partitioning. In *ISWC*. https://doi.org/10.1007/978-3-319-11964-9_9

[20] Ariam Rivas, Diego Collarana, Maria Torrente, and Maria-Esther Vidal. 2022. A neuro-symbolic system over knowledge graphs for link prediction. *Semantic Web* Preprint (2022), 1–25.

[21] Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. 2020. You CAN Teach an Old Dog New Tricks! On Training Knowledge Graph Embeddings. In *8th International Conference on Learning Representations, ICLR 2020*.

[22] Ahmad Sakor, Samaneh Jozashoori, Emetis Niazmand, Ariam Rivas, Konstantinos Bougiatiotis, Fotis Aisopos, Enrique Iglesias, Philipp D. Rohde, Trupti Padiya, Anastasia Krithara, Georgios Paliouras, and Maria-Esther Vidal. 2023. Knowledge4COVID-19: A semantic-based approach for constructing a COVID-19 related knowledge graph from various sources and analyzing treatments' toxicities. *J. Web Semant.* 75 (2023). https://doi.org/10.1016/j.websem.2022.100760

[23] Jenni A.M. Sidey-Gibbons and Chris J. Sidey-Gibbons. 2019. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology* 19 (2019). Issue 1. https://doi.org/10.1186/s12874-019-0681-4

[24] Sahar Vahdati, Guillermo Palma, Rahul Jyoti Nath, Christoph Lange, Sören Auer, and Maria-Esther Vidal. 2018. Unveiling Scholarly Communities over Knowledge Graphs. In *International Conference on Theory and Practice of Digital Libraries, TPDL*. https://doi.org/10.1007/978-3-030-00066-0_9

[25] Michael van Bekkum, Maaike de Boer, Frank van Harmelen, André Meyer-Vitali, and Annette ten Teije. 2021. Modular design patterns for hybrid learning and reasoning systems: a taxonomy, patterns and use cases. *Applied Intelligence* 51, 9 (2021), 6528–6546.

[26] C Vijayalakshmi and S Pakkir Mohideen. 2022. Predicting Hepatitis B to be acute or chronic in an infected person using machine learning algorithm. *Advances in Engineering Software* 172 (2022), 103179.

[27] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. *Conference on Artificial Intelligence (AAAI)* (2014).

[28] Tian X, Chong Y, Huang Y, Guo P, Li M, Zhang W, Du Z, and Hao Y Li X. 2019. Using Machine Learning Algorithms to Predict Hepatitis B Surface Antigen Seroclearance. *Comput Math Methods Med* (2019). https://doi.org/10.1155/2019/6915850

[29] J. Yang, A.A.S. Soltan, and D.A. Clifton. 2022. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. *Digit. Med* 5 (2022). Issue 69. https://doi.org/10.1038/s41746-022-00614-9

[30] Yating Yin, Lei Zhang, Yiguo Wang, Mingqiang Wang, Qiming Zhang, Guozheng Li, et al. 2022. Question answering system based on knowledge graph in traditional Chinese medicine diagnosis and treatment of viral hepatitis B. *BioMed Research International* 2022 (2022).