IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 October 2025, accepted 26 December 2025, date of publication 29 December 2025, date of current version 2 January 2026.

Digital Object Identifier 10.1109/ACCESS.2025.3649436

RESEARCH ARTICLE

Graph Alignment Methods for the Development of Interoperable Gene Regulation Knowledge Graphs

JUAN MULERO-HERNÁNDEZ, GINÉS ALMAGRO-HERNÁNDEZ, AND JESUALDO TOMÁS FERNÁNDEZ-BREIS, (Senior Member, IEEE)
Department of Informatics and Systems, University of Murcia, CEIR Campus Mare Nostrum, IMIB-Pascual Parrilla, 30100 Murcia, Spain
Corresponding author: Jesualdo Tomás Fernández-Breis (jfernand@um.es)

This work was supported in part by MICIU/AEI/10.13039/501100011033 under Grant PID2020-113723RB-C22 and Grant PID2024-155257OB-I00; and in part by ERDF, EU (FONDOS FEDER).

ABSTRACT Biological and biomedical knowledge graphs can be disparate in their design when a standard has not been widely accepted by the community, even more when they represent data from a previously underexplored domain. This may result in the formation of disparate knowledge graphs. Entity alignment methodologies examine the similarities and differences between entities across diverse knowledge graphs to ascertain which ones represent the same real-world entity. These alignments facilitate the integration of data and the interoperability between sources, which is a crucial aspect for comprehending the relationships between disparate datasets and uncovering novel insights. However, there is a lack of knowledge regarding the efficacy of these methods across different domains. We evaluate the performance of 20 entity alignment methods on biological knowledge graphs pertaining to gene regulation. Specifically, our investigation concentrates on enhancer sequences, the most extensively researched type of cis-regulatory domains, and their potential targets. The obtained hit values illustrate the potential of this and other methods to align and improve the interoperability between biological and biomedical knowledge graphs. Moreover, GCN Align demonstrates the highest robustness and produces the most accurate results in terms of matching ability and time required. Our study also examines the impact of the entity types included in the data schema. This analysis helps identify which domain areas are more effectively represented, which is crucial for optimizing the methods' efficacy and to identify areas of improvement in the schemas used. Additionally, the workflow is extensible to other data domains and can be adapted to new methods.

INDEX TERMS Cis-regulatory modules, data interoperability, entity alignment, gene regulation, knowledge graphs.

I. INTRODUCTION

A. BACKGROUND

The graph entity alignment (EA) process is designed to identify mappings between the semantic entities of two distinct knowledge graphs (KGs), namely the source and target graphs [1]. These mappings, which indicate that both entities may be equivalent, can be used to integrate data from disparate sources, facilitate interoperability between systems (i.e, communicate, exchange and use data), perform analysis of heterogeneous data, and the discovery of novel insights [2], [3]. This task is a crucial undertaking in biomedical contexts, where it is imperative to comprehend the interconnections between disparate data sets from diverse sources [4].

The landscape of biomedical databases is extensive and continues to expand. The quantity of biomedical data is considerable and presents both computational and storage challenges [5], [6], [7]. Most biomedical databases have been developed for local exploitation and are stored in formats that hinder their interoperability with other databases [8], [9].

The associate editor coordinating the review of this manuscript and approving it for publication was Domenico Rosaci.

Furthermore, the lack of consensus on the use of controlled vocabularies for data representation presents an additional challenge to this task [4], [10], [11].

The principles and tooling of the Semantic Web [12] have been shown to be a more efficient alternative for the representation, storage and querying of heterogeneous biomedical data [8], [13]. Gene Ontology (GO) [14], Sequence Ontology (SO) [15], or SNOMED CT (https://www.snomed.org/) are successful examples of this technology with wide community adoption. We highlight two fundamental technologies, namely, ontologies [16] and KGs [17].

Ontologies are employed to model and formalize a particular domain and its vocabulary. Consequently, they provide the formalization of concepts, relations, and restrictions, primarily through the utilization of formal languages such as OWL (Web Ontology Language), RDF (Resource Description Framework), and RDFS (RDF Schema). The community has developed a number of ontologies with the objective of modeling and formalizing different biomedical domains. The portals of the Open Biomedical Ontologies (OBO) Foundry [16] and BioPortal [18] include a plethora of biological and biomedical ontologies worthy of note. However, some of these ontologies present overlaps, which forces the user to decide which ontology to employ [19].

KGs are collections of relations, or edges, that connect nodes representing entities, which can be conceptualized as real-world concepts or individuals [17]. In this manner, KGs facilitate the representation and correlation of copious amounts of data from disparate sources and domains in a straightforward, computer-processable format [20]. In an ideal scenario, KGs would adhere to the schema provided by an ontology, but this is not always the case.

The community has also developed biomedical KGs that integrate information from multiple sources, and that cover different domains and their relationships [21]. Examples include biomedicine [22], [23], [24], pharmacology [25], anatomy [26], microorganisms [27], multiple omics [28], or systems biology [29]. However, the different KGs are not always interoperable due to the use of disparate vocabularies that are not connected to each other. This can occur for two reasons: either the KG does not adhere to a reference ontology, or the reference ontology does not present relations with equivalent entities in other ontologies that are in use. Moreover, KGs vary in structure, contributing to a high degree of heterogeneity within the field.

In order to address these issues, new initiatives have been proposed, such as the Biolink Model [30]. The objective of this schema is to universalize KGs in the biological and biomedical fields. However, as with other previously proposed schemas, its success will depend on its acceptance and use by the relevant community. Other proposals such as BioCypher [31] and PheKnowLator [32] seek to unify the construction of biomedical KGs.

B. MOTIVATION

In this context of heterogeneity, several EA methods have been developed to identify equivalent entities in different schemes. These methods use different approaches and often include comparisons with other methods in their evaluation reports [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53]. With the rise of large language models (LLMs), these tools have also begun to be included in EA strategies [54], [55], [56].

Because matching equivalent entities enables linking entity graphs that are not directly interoperable, from a biological point of view, these relationships allow better exploitation of available knowledge, enabling communication between isolated systems that can act as federated data sources. In addition to promoting and facilitating the querying and analysis of heterogeneous data, the formulation of hypotheses, and the contracting of results [13], this interoperability, as well as the integration of databases, is capable of providing new knowledge by communicating different data resources and data domains [57], [58]. Moreover, the exploitation of link prediction systems allows the identification of possible relationships between entities that are not directly linked [2]. Consequently, better exploitation of information using EA methods would result in better exploitation of information from available databases and clinical systems.

Specific works carrying out comparative studies on EA methods are also available [1], [51], [59], [60], [61], [62]. However, these studies have used mainly generic KGs, such as DBpedia, Wikipedia, and YAGO. Many of these comparative studies are also focused on multilingual EA. Consequently, there is a dearth of knowledge on the performance and effectiveness of the various EA methods in specific domains and subdomains, such as biological KGs and the gene regulation subdomain. This interest is due to the results of the various Ontology Alignment Evaluation Initiative campaigns (OAEI, https://oaei.ontologymatching.org/) and studies [1], [63] which have demonstrated that the tested methods exhibit varying performance across different domains or tracks.

C. OBJECTIVES AND APPROACH

In this paper, we address biological EA about gene regulation with two main objectives: (1) to analyze whether there are effective EA methods in this area that could contribute to improved interoperability between KGs, and to identify the best performing ones; (2) to analyze the performance by biomedical entity type to identify whether the schema used for modeling has appropriate features that facilitate EA and which aspects can improve efficiency.

In particular, we focus on the domain of human enhancers (Fig. 1), as it represents a biological domain lacking consensus on data representation [64], and for their relevance as modulators of gene expression [65].

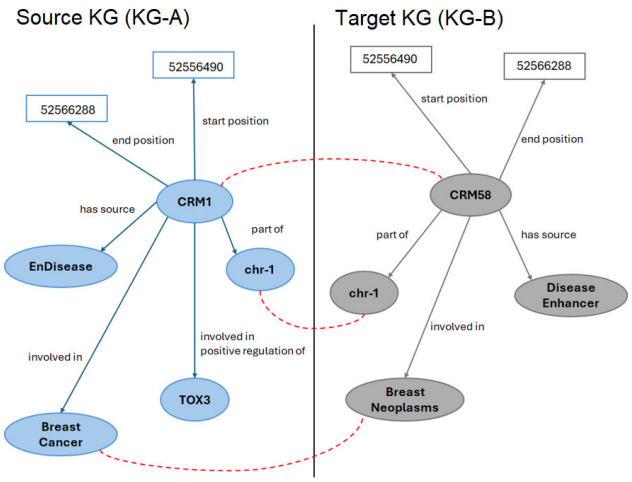

FIGURE 1. Illustration of the problem formulation (fictitious example). Two different graphs (source and target) are aligned to find the entities that represent the same real-world concept (red dotted line).

Enhancers are distal DNA sequences (cis-regulatory modules or CRMs) capable of increasing the transcription rate of their target genes, regardless of their orientation and over long distances [65], [66]. Moreover, these regulatory sequences have shown relevance in processes such as cell identity and disease development, and have been proposed as potential therapeutic targets in emerging treatment strategies [67], [68], [69], [70], [71], [72], [73], [74]. Indeed, the term ''enhanceropathies'' has been proposed to denote pathologies associated with these sequences [75]. For these reasons, enhancers are widely studied regulatory sequences, and with a large presence in specialized and generic databases [64]. In this context, KGs can play a pivotal role in standardizing data and facilitating interoperability between disparate sources or databases [13], [64].

To accomplish the objectives, a variety of EA techniques are employed to examine a multitude of potential scenarios, including the EA at both graph and subgraph levels. To assess the outcomes, we utilized the results pertaining to the general alignment metrics, with a particular emphasis on hits and time values. Additionally, we employ the results of an entity count, which facilitate a comprehensive examination of the performance, quality, and potential enhancements of the various modeled entities.

D. CONTRIBUTIONS BEYOND THE STATE-OF-THE-ART

This work provides a novel contribution in terms of the alignment strategies employed, the focus of the evaluations and the field of study. By evaluating the performance of different methods on KGs with different characteristics in an unexplored field, such as gene regulation this work will contribute to a more nuanced understanding of the efficacy of EA methods in specific domains, facilitating the identification of those with superior performance. This will facilitate the use of these tools to discern analogous entities across disparate data sources, thereby enhancing interoperability between heterogeneous data.

On the other hand, evaluating the performance of methods against entities with different characteristics, and leveraging the acquired knowledge to identify improvements in entity modeling, will also enable the adoption of improvements in semantic schemas. This will result in improved EA, thereby promoting the use of these methods as tools to improve the interoperability between entities of different KGs.

II. METHODS

This section outlines the data used for the generation of KGs, the EA methods employed, and the pipeline to conduct the experiments. All material and scripts used have been included in the cisregEA repository (https://github.com/juan-mulero/cisregEA), with additional details and supplementary material that supports and complements the main results presented in this manuscript.

A. THE DATASETS

To perform EA, a sample of five manually curated, open-access databases with available information about human enhancer sequences was utilized. The aforementioned databases were then transformed into KGs: ENdb [76], EnDisease [77], DiseaseEnhancer [78], VISTA Enhancer Browser [79], and RefSeq [80]. In addition to information regarding on enhancer sequences, they contain relations between these entities and other biological entities of interest [64], such as target genes, diseases or DNA-binding transcription factors (Table 1). These data are included in the KGs.

ENdb, EnDisease and DiseaseEnhancer contain data collected from literature reviews. In addition to annotations on enhancer sequences, relationships between these elements and phenotypes and target genes are provided, because these are more specialized databases. ENdb also contains data on transcription factors that interact with these sequences. On the other hand, VISTA and RefSeq are more general databases. VISTA collects experimentally validated enhancers from in vivo reporter gene expression studies, sequences previously selected for their evolutionary conservation. This database also includes data on potential target genes by proximity. Finally, RefSeq hosts sequences described in the literature and validated experimentally, and it provides the functional elements that are annotated in the reference genome.

The selection of these open-access databases was based on two considerations. Firstly, as they are manually curated they contain a larger number of common entities than databases with non-curated information. Secondly, they are databases with different types of annotations in addition to CRM sequences. As EA methods employ the attributes and relations of entities for their characterization and subsequent comparison [1], [62], these databases are a suitable sample.

B. ENTITY ALIGNMENT METHODS

Pairwise alignments were conducted using 20 different EA methods provided by the OpenEA package [81], and which are briefly described in Table 2.

TABLE 1. Databases utilized for KG construction and subsequent EA. Each dataset contains a variety of annotations or content, indicated by an "X", that we include in the KGs.

Dataset	Assembly	Coordinates	Papers	Methods	Samples	Target genes	Transcription factors	Diseases
ENdb	hg19	X	X	X	X	X	X	X
EnDisease	hg38	X	X		X	X		X
Disease Enhancer	hg19	X	X			X		X
VISTA	hg19	X	X	X	X	X		
RefSeq	hg38	X	X	X				

TABLE 2. Short description of the EA methods used.

Method	Short Description	Reference
AlignE	Embedding-based model for EA across different KGs.	[35]
AliNet	It uses graph neural networks (GNNs) and aims to mitigate the non-isomorphism of neighborhood structures in an end-to-end manner. It introduces distant neighbors and employs an attention mechanism to highlight helpful distant neighbors and reduce noises. Then, it controls the aggregation of both direct and distant neighborhood information using a gating mechanism.	[33]
AttrE	It incorporates attribute information in addition to graph structure to enhance EA accuracy. For this, it uses a transitivity rule to further enrich the number of attributes of an entity to enhance the attribute character embedding	[34]
BootEA	Model with bootstrapping approach for EA based on embedding. It iteratively labels likely EA as training data for learning alignment-oriented KG embeddings. Furthermore, it employs an alignment editing method to reduce error accumulation during iterations.	[35]
BootEA RotatE	A variant of BootEA that uses RotatE, a model based on rotations in the complex space to represent relations.	[81]
GCN Align	It applies GCN to capture entity features and improve the alignment. The embeddings can be learned from the structural and attribute information of the entities, and the results of structural embedding and attribute embedding are combined to obtain accurate alignments.	[36]
HolE	Holographic embedding model to learn compositional vector space representations and to capture interactions between entities using circular convolutions.	[37]
IMUSE	Unsupervised iterative model using both attribute and relationship triples. It also uses a bivariate regression model to determine the respective weights of the similarities.	[38]
IPTransE	Extension of TransE that introduces an iterative propagation function to enhance information transfer between graphs. The model encodes both entities and relations of various KGs into a unified low-dimensional semantic space according to a small seed set of aligned entities.	[39]
JAPE	Hybrid model combining attribute-based and structure-based alignment through joint learning techniques.	[40]
MTransE	Extension of TransE that projects different graphs into separate representation spaces and learns transformations between them.	[41]
ProjE	Deep neural network-based model that maps entities and relations into an embedding space for alignment.	[42]
RDGCN	Relation-aware Dual-Graph Convolutional Network (RDGCN) to incorporate relation information via attentive interactions. It uses graph convolutional networks with relation-based attention mechanisms to improve EA.	[43]
RSN4EA	Use of residual sequential networks (RSN), which employ a skipping mechanism to bridge the gaps between entities, to better capture structural information and enhance EA. RSNs integrate recurrent neural networks (RNNs) with residual learning to efficiently capture long-term relational dependencies within and between KGs.	[44]
RotatE	Model representing relations as rotations in a complex space. It is able to model and infer various relation patterns including: symmetry/antisymmetry, inversion, and composition. Specifically, the RotatE model defines each relation as a rotation from the source entity to the target entity in the complex vector space.	[45]
SEA	Semi-supervised alignment model integrating structural and semantic signals to improve entity correspondence.	[46]
SimplE	Tensor decomposition-based model (canonical polyadic decomposition) that represents relations more expressively.	[47]
TransD	Variant of TransE that models dynamic projections of entities and relations into different representation spaces.	[48]
TransH	Extension of TransE that introduces hyperplanes to improve the modeling of relationships in EA.	[49]
TransR	Model that builds embeddings of entities and relationships in separate entity spaces and relationship spaces.	[50]

These methods were chosen for two main reasons. First, they are implemented within the OpenEA package, a toolkit that facilitates the standardized execution of EA, an important task to compare different methods fairly. Second, the selected methods differ in the strategy for predicting matched entities. Consequently, this sample of methods is useful to study the applicability of different approaches to improve interoperability between KGs, and to study the performance of these methods against entities with different semantic schemas.

C. THE GRAPH ALIGNMENT PIPELINE

The execution of EA was conducted in accordance with a standardized pipeline described in a previous work [82]. This pipeline shows the flow of tasks necessary to carry out a complete EA, starting from data files in a typically tabular format to the subsequent evaluation of results (Fig. 2). The cisreg package (https://github.com/juan-mulero/cisreg) was employed for database preprocessing and RDF generation [13], while the OpenEA library (https://github.com/nju-websoft/OpenEA) was used as core to perform the EA. The process is also described in pseudocode in Algorithm 1.

Using as input two KGs, a seed alignment, pre-trained word embeddings, and a configuration file, OpenEA returns a set of aligned entities and a report with values of EA metrics [1]. These metrics are later detailed in the evaluation subsection. Therefore, OpenEA facilitates the systematic and standardized execution of EA between two KGs through the implementation of diverse experimental methods.

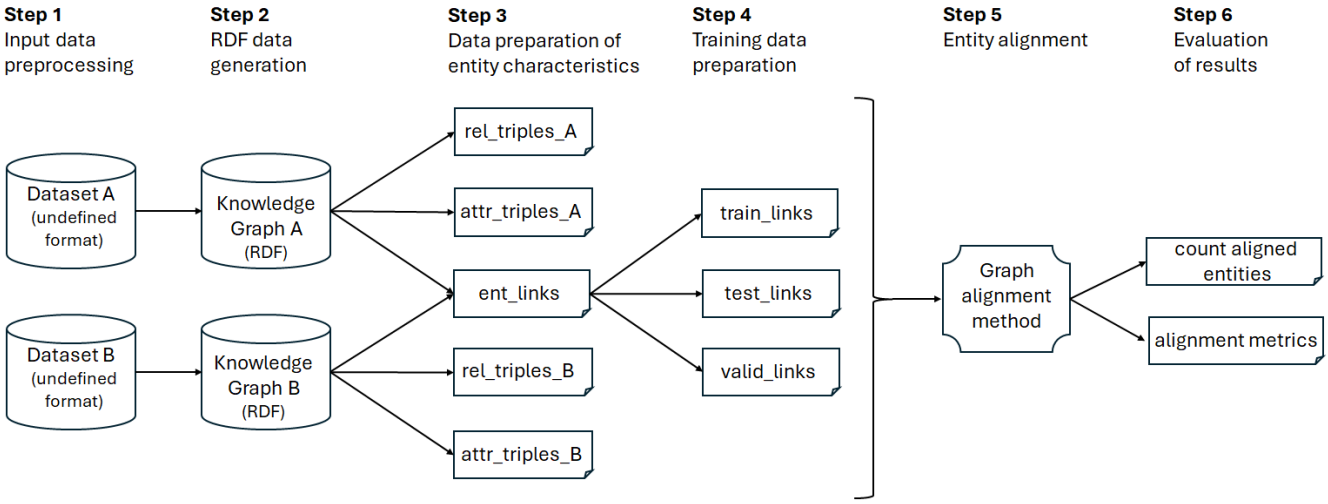

FIGURE 2. The pipeline applied in the experiments. Initially, the data undergo preprocessing to enable the automated alignment of RDF files. Two stages of data preparation are necessary. First, the RDF files are divided into three distinct sets: entities, attributes, and relations. Conversely, the set of entities is divided into three subsets, namely training, testing, and validation sets. Subsequently, the EA method is executed, followed by a final step of results evaluation.

Algorithm 1 Workflow

Step 1 - Input data preprocessing (cisreg package)
preprocessedA ← PreprocessData(DatasetA)
preprocessedB ← PreprocessData(DatasetB)

Step 2 - RDF data generation (cisreg package)
rdfA ← ConvertToRDF(preprocessedA)
rdfB ← ConvertToRDF(preprocessedB)

Step 3 - Data preparation of entity characteristics
attrTriplesA, relTriplesA, entTriplesA ← SplitRDF(rdfA)
attrTriplesB, relTriplesB, entTriplesB ← SplitRDF(rdfB)
entLinks ← GenerateEntityLinks(entTriplesA, entTriplesB)

Step 4 - Training data preparation
trainLinks, testLinks, validLinks ← SplitEntityLinks(entLinks, [0.7, 0.2, 0.1])

for method in methods **do** // Run 20 alignment methods
Step 5 - Entity alignments (OpenEA package)
alignedEntities, alignmentMetrics ← EA(relTriplesA, relTriplesB, attrTriplesA, attrTriplesB, entLinks, trainLinks, testLinks, validLinks, configFile[method])

Step 6 - Evaluation of results
countAlignedEntities ← EvaluateResults(alignedEntities, testLinks)
end for

The objective of each experiment is to find the optimal alignment between two KGs, which requires the implementation of distinct procedures, as outlined in the following sections.

1) STEP 1: INPUT DATA PREPROCESSING

The objective is to transform the data to produce standardized input files that will facilitate the subsequent automated generation of RDF files. This is necessary because the original data sources use different, sometimes incompatible, file formats and data structures. The cisreg package was employed for this purpose [13].

2) STEP 2: RDF DATA GENERATION

The objective is the generation of RDF files in N3 format. These RDF files represent the elements that compose the KGs, and the information stored in the original databases. We used the cisreg package that provides scripts to convert the tabular files of the sample databases to RDF [13]. In the event that those scripts had not been available, the use of mapping-based tools such as SWIT [83] or Morph-KGC [84] would have been the best option to support the generation of RDF data.

Each KG incorporates data from a distinct database, yet each database encompasses disparate biological subdomains. Consequently, two distinct types of KGs are generated: KGs comprising all data from each source (referred to as ''all'' graphs) and subgraphs focusing on specific subdomains. These subgraphs include enhancer sequences (referred to as ''crm''), relations between enhancers and their target genes (referred to as ''crm2gene''), relations with phenotypes (referred to as ''crm2phen''), and relations with transcription factors (referred to as ''crm2tfac'').

These subgraphs are automatically generated by the cisreg package, being the ''all'' graphs the result of combining the different subgraphs that result from a single dataset.

This subdivision is employed for the purpose of comparing the performance of alignment methods that utilize complete graphs and subgraphs.

The schema used for RDF modeling is summarized in Fig. 3, and extended in the cisregEA repository. The main vocabularies used are also summarized in Table 3, which are vocabularies broadly extended in the semantic web, such as rdf, rdfs, owl and skos, and biological ontologies of relevance, mainly belonging to OBO Foundry [16].

3) STEP 3: DATA PREPARATION OF ENTITY CHARACTERISTICS

In order to perform the EA between two input graphs, the methods must first extract the relevant characteristics of each entity, including attributes and relations. Subsequently, the aforementioned features are compared in order to ascertain the degree of similarity or equivalence between the entities in question, with a view to ultimately determining the corresponding alignments.

The OpenEA package requires the preprocessing of the graphs to be aligned. This preprocessing step consists of classifying the triples of each graph according to whether they correspond to attribute definitions or to relationships between entities. This task is carried out by generating dedicated files. A file with the related entities between graphs must also be generated, as it will be used in the next step to generate the training, test and validation sets. Therefore, in this step three kinds of files are created for each pairwise alignment:

- **Entity set (ent_links)**: The entities of each graph are linked between them. The analysis encompassed both classes and instances. This task takes into account that enhancer sequences are modeled based on their genomic coordinates, hence two sequences are considered equivalent if their coordinates overlap.
- **Attribute set (attr_triples)**: The triples defining attributes for each entity, which correspond to datatype properties.
- **Relation set (rel_triples)**: The defining relationships between two different entities, which correspond to object properties.

Table 4 shows the triples generated for each database for the complete graphs (all), and the subgraphs that compose them (crm, crm2gene, crm2phen, crm2tfac).

4) STEP 4: TRAINING DATA PREPARATION

In addition to the graphs to be aligned, the methods also need a training set or seed alignment. For its generation, the subset of related entities (rel_triples in Fig. 2) was randomly divided into 3 sets: training (70%), testing (20%) and validation (10%).

5) STEP 5: ENTITY ALIGNMENT (EA)

Given two KGs, KG_1 and KG_2, with entity sets E_1 and E_2, respectively, the EA aims to identify pairs of entities (e_1, e_2), where $e_1 \in E_1$, $e_2 \in E_2$, and e_1 and e_2 represent the same real-world entity. That is, e_1 and e_2 are aligned entities. Each EA method uses different strategies and functions that try to maximize the similarity between matched entities, i.e, $S(E_1, E_2) \in [0, 1]$. Consequently, the objective is to find the results closest to 1 in $S_{KG_1,KG_2} = \{(e_1, e_2) \in E_1 \times E_2 \mid e_1 \sim e_2\}$, where $\sim$ denotes an equivalence relation [81].

Pairwise alignments were conducted using the 20 EA methods aforementioned (Table 2). The default configuration of the methods was employed in all executions to ensure a fair comparison and to provide comparable results with other state-of-the-art studies and possible future works. Each EA was also conducted twice to enhance the statistical significance of the findings. Regarding computing resources, a server with Intel(R) Xeon(R) CPU E5-2697A v4 @ 2.60GHz model was used, while a limit of 250 GB of memory was established for each EA. Furthermore, a designated cut-off time of 48 hours was set to initiate the EA processes following the launch.

In order to evaluate the efficacy of the EA methods in a variety of scenarios, multiple pairwise alignments were conducted, with the input sets varying in each instance. These alignments can be classified into the following experimental groups (Table 5):

- Alignment of each KG with itself (AvsA). We assume that the best possible alignment for one entity is against itself [82], so these alignments will provide the best results. Therefore, these alignments serve the purpose of identifying the upper limit for the results of the methods, as well as to distinguish between entities that have been appropriately modeled and those that require further modeling to improve the results.
- Alignments of subgraphs (crm, crm2gene, crm2phen, crm2tfac) vs. complete graphs (all), with the aim of identifying whether subdivision into subgraphs or subdomains offers better results by reducing the number of entities to align and the number of comparisons [81].
- Alignment of different KGs to identify whether entities with different features can be linked even though they constitute the same real-world entity. Moreover, we tested different scenarios (Table 5). These included different entities at the attribute level (attr-AvsB), different entities at the relation level (rel-AvsB), and a combination of both (AvsB).

The alignments of different KGs at the attribute level (attr-AvsB) exhibit identical relations but differ in their attributes. In order to generate these KGs, the sequence identifiers were modified, as they are utilized for the naming and definition of sequences. Other attributes, such as sequence coordinates, were not modified, as this would result in the generation of different sequences and, consequently, different entities. Due to persistent difficulties encountered with the various EA methods employed for the RefSeq dataset, it has been excluded from the present analysis.

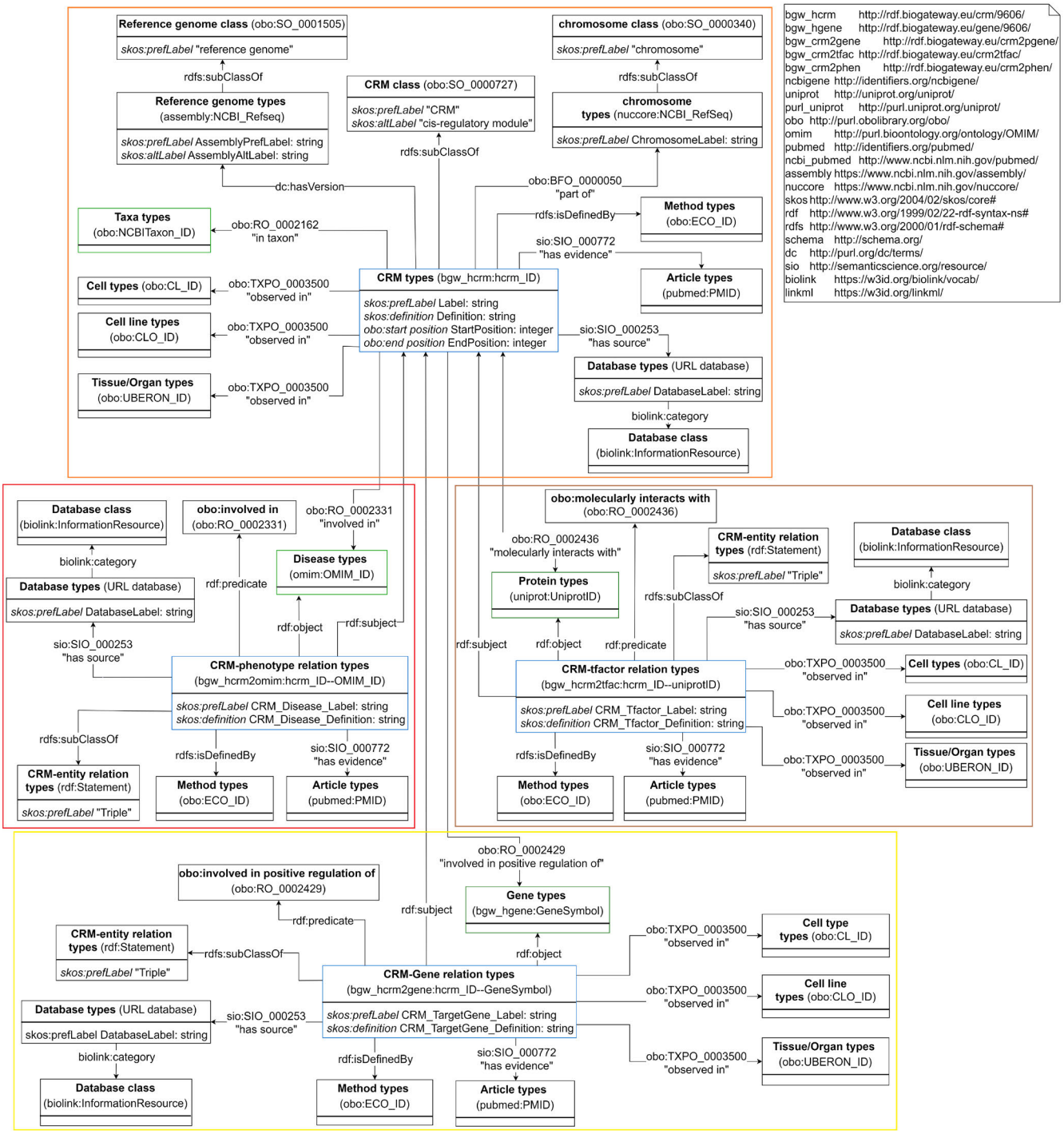

FIGURE 3. Underlying summarized schema of the KGs. The schema models the structure and semantics of CRM sequences and their relations with other entities. The colored boxes specify the different subgraphs or subdomains: CRM sequences (crm graph - orange), and their relations with target genes (crm2gene graph - yellow), transcription factors (crm2tfac graph - brown), and phenotypes (crm2phen graph - red). The blue classes constitute the central entities of each graph, while the green classes are biological classes that were not modeled in detail because they are already modeled in the BioGateway knowledge network, with a schema interoperable with this one [85]. The schema shown includes the relations of the core entities, i.e. CRM entities and entities of relations between these sequences and other entities. Therefore, the relations of the entities acting as objects were omitted to reduce the size of the image. A complete schema can be found in the cisregEA repository.

In the alignment of different KGs at the relations level (relAvsB), the entities exhibit identical attributes, yet differ in their relations with other entities. Consequently, the outcomes align CRM sequences and their associated entities from disparate databases, encompassing diverse types of entity relations (see Table 1).

TABLE 3. Main vocabularies and ontologies used in the scheme of the KGs generated and employed for the EA experiments.

Vocabulary/Ontology	Prefix	Namespace
Resource Description Framework	rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
RDF Schema	rdfs	http://www.w3.org/2000/01/rdf-schema#
Web Ontology Language	owl	https://www.w3.org/2002/07/owl#
Simple Knowledge Organization System	skos	http://www.w3.org/2004/02/skos/core#
Schema.org	schema	https://schema.org/
Dublin Core terms	dc	http://purl.org/dc/terms/
Gene Ontology (GO)	obo	http://purl.obolibrary.org/obo/
Sequence types and features ontology (SO)	obo	http://purl.obolibrary.org/obo/
Basic Formal Ontology (BFO)	obo	http://purl.obolibrary.org/obo/
OBO Relations Ontology (RO)	obo	http://purl.obolibrary.org/obo/
Cell Ontology (CL)	obo	http://purl.obolibrary.org/obo/
Cell Line Ontology (CLO)	obo	http://purl.obolibrary.org/obo/
Uber-anatomy ontology (UBERON)	obo	http://purl.obolibrary.org/obo/
Human Disease Ontology (DO)	obo	http://purl.obolibrary.org/obo/
Genotype Ontology (GENO)	obo	http://purl.obolibrary.org/obo/
TOXic Process Ontology (TXPO)	obo	http://purl.obolibrary.org/obo/
NCBI organismal classification (NCBITaxon)	obo	http://purl.obolibrary.org/obo/
Evidence & Conclusion Ontology (ECO)	obo	http://purl.obolibrary.org/obo/
Molecular Interactions Controlled Vocabulary (MI)	obo	http://purl.obolibrary.org/obo/
Ontology for Biomedical Investigations (OBI)	obo	http://purl.obolibrary.org/obo/
NCI Thesaurus OBO Edition (NCIT)	obo	http://purl.obolibrary.org/obo/
Experimental Factor Ontology (EFO)	efo	http://www.ebi.ac.uk/efo/
BioAssay Ontology (BAO)	bao	http://www.bioassayontology.org/bao#
Semanticscience Integrated Ontology (SIO)	sio	http://semanticscience.org/resource/
Medical Subject Headings (MeSH)	mesh	https://www.ncbi.nlm.nih.gov/mesh/
UniProt Knowledgebase	uniprot	http://uniprot.org/uniprot/
UniProt Knowledgebase	purl_uniprot	http://purl.uniprot.org/uniprot/
Online Mendelian Inheritance in Man (OMIM)	omim	http://purl.bioontology.org/ontology/OMIM/
PubMed	pubmed	http://identifiers.org/pubmed/
PubMed	ncbi_pubmed	https://www.ncbi.nlm.nih.gov/pubmed/
RefSeq assembly accession	assembly	https://www.ncbi.nlm.nih.gov/assembly/
NCBI Reference Sequence	nuccore	https://www.ncbi.nlm.nih.gov/nuccore/
NCBI Gene	ncbigene	http://identifiers.org/ncbigene/
Biolink Model	biolink	https://w3id.org/biolink/vocab/
LinkML Model	linkml	https://w3id.org/linkml/
BioGateway	bgw_hgene	http://rdf.biogateway.eu/gene/9606/
BioGateway	bgw_hcrm	http://rdf.biogateway.eu/crm/9606/
BioGateway	bgw_crm2gene	http://rdf.biogateway.eu/crm2pgene/
BioGateway	bgw_crm2tfac	http://rdf.biogateway.eu/crm2tfac/
BioGateway	bgw_crm2phen	http://rdf.biogateway.eu/crm2phen/

TABLE 4. Distribution of triples for the KGs generated for each dataset.

Dataset	KG	Triples	Attributes	Relations	Entities
ENdb	all	42,707	7,571	35,136	4,511
	crm	13,424	2,559	10,865	1,491
	crm2gene	12,343	1,731	10,612	1,721
	crm2phen	5,764	1,428	4,336	949
	crm2tfac	12,933	1,861	11,072	1,532
EnDisease	all	19,035	5,875	13,160	3,149
	crm	6,696	2,505	4,191	1,124
	crm2gene	6,876	1,739	5,137	1,371
	crm2phen	5,764	1,428	4,336	949
Disease Enhancer	all	63,194	19,195	43,999	9,983
	crm	11,672	4,785	6,887	1,818
	crm2gene	16,344	4,471	11,873	3,088
	crm2phen	35,984	9,944	26,040	5,478
VISTA	all	93,917	23,036	70,881	11,032
	crm	42,969	11,709	31,260	3,969
	crm2gene	50,948	11,327	39,621	7,108
RefSeq	all	1,422,118	448,521	973,597	149,850
	crm	1,422,118	448,521	973,597	149,850

Finally, we examined the alignments of different KGs at the relation and attribute levels (AvsB). To address the first condition, datasets from disparate databases were utilized, as they exhibit disparate types of annotations (see Table 1).

TABLE 5. Different scenarios evaluated in the EA experiments carried out.

Type of EA	Description
AvsA	EA of identical graphs
attr-AvsB	EA of different graphs at attributes level
rel-AvsB	EA of different graphs at relations/edges level
AvsB	EA of different graphs at both levels (attributes and relations)

To satisfy the second requirement (attributes), the identifiers of the sequences were modified to produce entities with different labels. The RefSeq dataset was also excluded from the analysis due to its incompatibility with several methods.

However, not all databases possess a sufficient number of common entities to allow the entity splitting process and to generate an appropriate training set (Step 4). Only two graph combinations had sufficient common entities to perform alignments with different KGs: EnDisease-DiseaseEnhancer (1284 entities) and RefSeq-VISTA (1028 entities). We will refer to these alignments as typical AvsB or typ-AvsB in

the text. To address this issue, a variation of the standard workflow was implemented to increase the number of common entities. For this experiment, the two datasets (A and B) were unified in order to generate a sufficient number of common entities for optimal training (set union: $A \cup B$ vs $B \cup A$). The entities which are not common between different KGs were employed for the purposes of training and validation (set difference: $A \setminus B$), while the entities common to both datasets were used for testing (set intersection: $A \cap B$). With this modification, the following alignments could be performed: ENdb-DiseaseEnhancer, ENdb-EnDisease, EnDisease-Disease-EnDiseaseEnhancer, VISTA-DiseaseEnhancer, VISTA-EnDisease, and RefSeq-VISTA. Furthermore, these alignments were performed and evaluated at the level of relations (modified rel-AvsB), and attributes and relations (modified AvsB or mod-AvsB).

6) STEP 6: EVALUATION OF RESULTS

To evaluate the results of the alignments, we used the following metrics related to the performance of the EA methods: hits@1, hits@5, hits@10, Mean ranking (mr), Mean Reciprocal Ranking (mrr) and Execution time, which are introduced next [1]. These metrics are used by the OpenEA package to report the results, and they are the standard metrics used in the state-of-the-art [1], [51], [59], [60], [61], [62], [81], [86], [87].

- The hits@k metrics indicates the percentage of cases where the correct entity is in the first k predictions generated by the model. These metrics are important to understand how often the model puts the right entity in the top ranking positions.

$$Hits@k = \frac{\left|\{q \in Q : q \leq k\}\right|}{|Q|}$$

  q denotes the prediction element, and Q the prediction items given by the algorithm. The value of Hits@k ranges from 0 to 1, being better the algorithm when the value is higher.
- The Mean Ranking (MR) is the mean of the positions of the correct entities in the ranking generated by the model. A lower MR score indicates a better performance of the model, as the correct entities are closer to the top of the ranking.

$$MR = \frac{1}{|Q|} \sum_{q \in Q} q$$

- Mean Reciprocal Ranking (MRR) is the mean of the reciprocals of the positions of the correct entities in the ranking. A higher MRR score indicates better performance, since it implies that the correct entities tend to be ranked higher.

$$MRR = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{q}$$

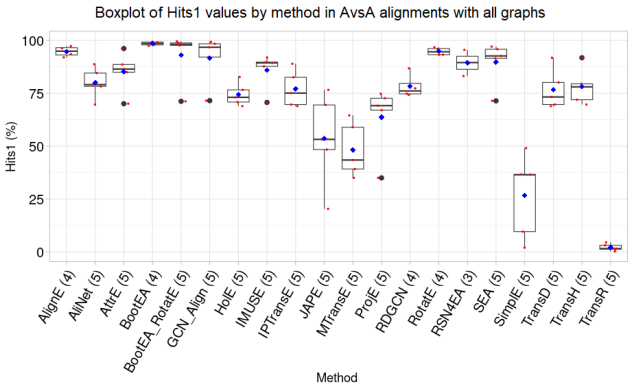

FIGURE 4. Boxplot of hits@1 values by method in alignments of same KGs (AvsA). All domains and datasets were used (ENdb, EnDisease, DiseaseEnhancer, VISTA and RefSeq).

Since each alignment was performed twice, we also calculated the variability values between replicates for each main metric. In the results section, we mainly include the coefficient of variation (CV), but the variance and standard deviation values are available in the supplementary material.

For this work, the most relevant metrics are hits@1 (how many true predictions correspond to the best given prediction), and the execution time, since they are related to automation and scalability, respectively. The relevance of those metrics is justified because the main objective of this work aims to study the feasibility of these methods as a tool to contribute to the interoperability between KG entities. Consequently, the automation of alignments is a critical task, which makes hits@1 and time stand out as the most important metrics. However, all the metrics presented were used and reported.

Additionally, the percentage of aligned entities in accordance with the anticipated values within the testing set (test_links) was evaluated in order to assess the efficacy of the schema in modeling each entity.

III. RESULTS

This section presents the main findings of the present study. The complete results, accompanied by supplementary material, are accessible in the repository cisregEA (https://github.com/juan-mulero/cisregEA). This includes data files, code used, tables, and figures that support the main results reported in this section.

A. ALIGNMENT OF THE SAME GRAPHS

The results are structured according to the two subtypes of alignments performed: alignments of complete datasets (all domains) and alignments of subgraphs (crm, crm2gene, crm2phen and crm2tfac).

1) ALIGNMENT OF COMPLETE DATASETS

The mean hits@1 obtained for each EA method and dataset are reported in Table 6, and represented in the boxplot of

TABLE 6. Hits@1 values (%) per method in AvsA alignments using all data domains. When the method was unable to be utilized, NA was recorded.

Method	DiseaseEnhancer	ENdb	EnDisease	RefSeq	VISTA	Mean hits@1	Mean CV
AlignE	91.96	96.29	93.49	NA	97.30	94.76	0.47
AliNet	88.73	84.53	79.05	69.61	78.22	80.03	0.64
AttrE	84.95	88.58	86.43	70.05	96.15	85.23	4.69
BootEA	97.42	98.00	99.13	NA	99.30	98.46	0.32
BootEA RotatE	98.77	98.17	99.52	71.16	97.62	93.05	0.28
GCN Align	92.11	98.39	96.75	71.49	99.30	91.61	0.37
HolE	70.82	73.06	76.59	68.90	82.73	74.42	1.30
IMUSE	87.70	91.91	89.68	70.65	89.35	85.86	0.98
IPTransE	75.05	82.54	68.89	69.65	88.94	77.01	1.96
JAPE	48.32	69.46	76.59	20.30	53.17	53.57	5.84
MTransE	39.15	58.93	64.52	34.96	43.40	48.19	1.61
ProjE	72.65	69.12	66.98	34.95	74.73	63.69	1.11
RDGCN	74.25	86.81	77.30	NA	74.84	78.30	0.45
RotatE	96.72	93.18	93.25	NA	95.92	94.77	1.01
RSN4EA	83.11	95.49	89.49	NA	NA	89.36	3.14
SEA	92.64	91.52	95.79	71.39	97.05	89.68	0.20
SimplE	9.54	49.06	36.59	1.96	36.42	26.71	12.92
TransD	73.22	80.10	68.89	69.67	91.75	76.73	2.08
TransH	78.01	79.43	71.90	69.66	91.80	78.16	2.38
TransR	3.08	4.43	1.51	0.17	1.20	2.08	36.34
Mean	72.91	79.45	76.62	52.97	77.31	74.08	3.90

Fig. 4. The mean hits@1 value considering all datasets is represented as a blue dot.

The results varied according to the method and the dataset used. Ten methods showed hits@1 mean values higher than 80% (BootEA>RotatE>AlignE>BootEA-RotatE> GCN Align >SEA>RSN4EA>IMUSE>AttrE> AliNet), and the mean for six of them was equal to or higher than 90% if we round to the unit. The clustering of the methods by hits@1 performance also facilitated the categorization and distinction of these groups.

However, not all methods yielded successful results in all alignments. Consequently, the number of alignments that could be conducted is indicated adjacent to the designation of the method in Fig. 4.

RSN4EA was unable to achieve an alignment using RefSeq and VISTA within the 48-hour timeframe following its launch, which was established as the cut-off point. The AlignE, BootEA, and RDGCN methods were unable to complete the alignment for RefSeq due to the requirement of more than 250 GB of RAM, which exceeded the specified memory limit. The RotatE method was unsuccessful due to errors that occurred during the training process, also with the RefSeq dataset. AliNet completed the task, but required a higher amount of RAM (∼100 GB) than the other methods (∼10 GB).

Moreover, most of the methods demonstrated homogeneous hits@1 between replicates, with a CV lower than 5% (mean CV = 3.90). A robust inverse correlation was also identified between hits@1 and their CV (cor = −0.84, df = 18, p-value = 4.2e-06, Pearson test), indicating that higher hits@1 values are associated with reduced variability between replicates.

We also analyzed the execution time of alignments. The alignment of KGs with a higher number of triples and entities required more time. However, the time required for this process was not uniform across all methods. The clustering by time also identified groups of methods with short time requirements (e.g. GCN Align<BootEA<RDGCN), and methods with long time ones (e.g. HolE>BootEA-RotatE>ProjE).

The variability between replicates for time was also explored, being higher than for hits@1 (mean CV values between 10-37%). Furthermore, no correlation was identified between the time values obtained and their associated variability (cor = −0.33, df = 18, p-value = 0.15) or between the hits@1 values and the variability in execution time (cor = 0.02, df = 18, p-value = 0.92).

Next, we clustered the methods by hits@1 and time, and plotted the relationship between the variables (see Fig. 5A). We highlighted 5 clusters. Cluster 1 was the most relevant because it corresponds to the group of methods with the best alignment performance and generally short execution times. The methods included in this cluster were GCN Align, BootEA, AlignE, RotatE, AttrE, IMUSE, RSN4EA, and SEA. Notwithstanding the favorable hits@1 values of BootEA-RotatE, this method was not included in the aforementioned group due to its elevated execution time.

RefSeq was the dataset that yielded the poorest overall hit results and required longer tun times due to the higher volume of triples. Since BootEA, AlignE, RDGCN, RotatE, and RSN4EA were not feasible to execute on RefSeq, an analysis without this dataset was also performed.

Fig. 5B shows the clustering by hits@1 and time with this approach. In this case, the best results were observed in the methods of cluster 4 (GCN Align, SEA, IMUSE), and BootEA, AttrE and BootEA-RotatE as the best methods of clusters 1, 2 and 3, respectively. These results are justified because clustering of methods by hits@1 values without RefSeq did not yield significant differences, but clustering by time revealed notable discrepancies, with BootEA, AlignE,

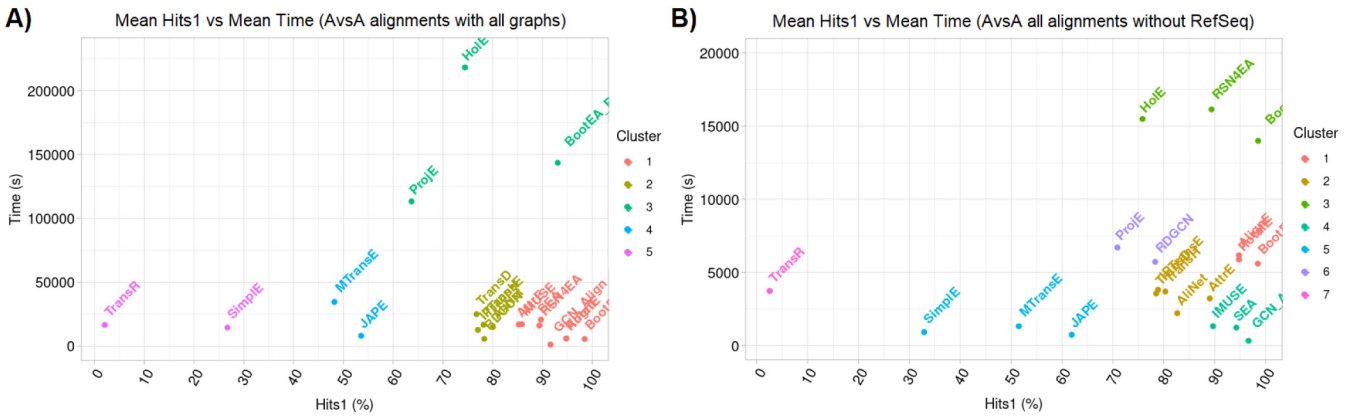

FIGURE 5. Hits@1 values against time in EA of equal KGs (AvsA) and considering all the data of each dataset (all). A) Results considering the five databases (ENdb, EnDisease, DiseaseEnhancer, VISTA and RefSeq). B) Results excluding RefSeq.

RDGCN, RotatE, and RSN4EA requiring comparatively longer processing times.

Since each dataset contains different types of annotations (Table 1), the hits@1 values also differed between datasets (Table 6). To evaluate the databases, we ranked each method according to its hits@1 value. Then, a metric was designed and used to assign a score to each method:

$$weight = \sum_{i=1}^{n} x_i \cdot \left(1 - \frac{i-1}{n}\right)$$

where the variable x_i denotes the number of times the dataset occupied position i, or ranking, which is determined by the number of datasets used (n).

This approach enabled the identification of datasets with a better set of annotations, which can facilitate the EA. In this case, VISTA > EnDisease > ENdb > DiseaseEnhancer > RefSeq.

The hit count per entity type, i.e. the percentage of correctly aligned biological concept type, is also a crucial element in the analysis, since it indicates what type of entities yields disparate results. Table 7 presents the mean of this entity count. In this case, the core entities of the graphs demonstrated favorable outcomes and minimal variability (low CV values). Specifically, CRM sequences and entities pertaining to the relationships between these CRMs and other entities. In contrast, entities that primarily function as object nodes, such as databases or assemblies, demonstrated inferior performance and higher CV values. Similarly, other entities that also act as object nodes, such as proteins or diseases, exhibited inferior performance and higher CV values due to their distinctive modeling, which is incorporated into other graphs that were not included in the analysis. A similar phenomenon was observed for external entities drawn from reference ontologies, the associated data for which are accessible through federated queries. Examples of such entities include cell lines and articles.

2) ALIGNMENTS OF SUBDOMAINS OR SUBGRAPHS

This section presents the findings pertaining to the domain-specific alignments of the datasets, or subgraphs (crm, crm2gene, crm2phen, crm2tfac).

Table 8 illustrates the difference in hits@1 values obtained when subsets of data are considered, as opposed to the results obtained when all data domains are included. Overall, the difference was negative, indicating that the subgraphs exhibited inferior hits@1 outcomes compared to when the entire network was considered.

An analysis of the resulting values of EA metrics and their variability between replicates revealed the following aspects:

- The profile of hits@1 and time values in subgraphs exhibited a similarity to that obtained with the complete set. Therefore, the methods were also clustered in analogous groups. BootEA, BootEA-RotatE, and GCN-Align methods demonstrated the most promising results in terms of hits@1 and exhibited consistent performance in the branch of dendrograms with optimal outcomes.
- Hits@1 and time values were generally lower due to inferior alignment performance and reduced data volume. Furthermore, there was an increase in the variability of hits@1 between replicates (average CV: 3.90% in all domains, 5.31% in crm subgraph, 8.16% in crm2gene, 7.58% in crm2phen, and 3.64% in crm2tfac) (Table 8).
- AlignE, BootEA, RDGCN, RotatE and RSN4EA methods presented the same execution issues with the crm subgraph of RefSeq, predominantly linked to high RAM requirements. However, ProjE and JAPE methods demonstrated sensitivity in subgraphs with a smaller volume of triples (crm2gene, crm2phen, and crm2tfac in ENdb, EnDisease, DiseaseEnhancer).
- The clusters generated using hits@1 and time values also exhibit a tendency toward similarity, with some fluctuations in the methods between neighboring branches. GCN Align, IMUSE, and SEA methods consistently

TABLE 7. Entity counting by biological type in AvsA alignments. All data domains and datasets were used. The figures display the expected and obtained mean values, along with the CV between replicates. In cases where CV could not be calculated due to the absence of the entity in both replicates, this is indicated by NA. This is a consequence of the random distribution of triples employed for training, testing, and validation purposes.

Type of entity or biological concept	Mean of expected entities	Mean of obtained entities	Percentage of obtained entities	Mean CV of matches
Chromosome class	1.00	0.90	89.87	NA
Relation class	1.00	0.89	89.47	NA
CRM class	1.00	0.89	89.47	NA
Method types	2.85	2.35	87.73	19.46
Chromosome types	4.26	3.69	87.27	12.76
CRM-phenotype relation types	73.50	61.88	84.21	8.10
CRM types	2540.85	1423.05	81.50	4.93
CRM-Entity relation types	422.58	340.52	81.43	3.96
Cell types	13.81	10.92	80.81	22.84
Tissue/Organ types	10.61	8.28	77.84	20.79
Instances of CRM-Entity relations	427.52	322.47	77.50	5.28
Instances of CRM types	2521.52	1350.02	77.32	4.47
CRM-phenotype relation class	1.00	0.75	75.00	NA
Cell line types	24.50	18.15	73.51	23.03
Disease types	37.81	26.42	66.62	18.59
Database class	1.00	0.65	65.00	78.57
Transcription factor class	1.00	0.65	65.00	NA
Database types	1.00	0.60	60.00	NA
Protein types	27.50	14.20	51.6	21.34
Article types	88.54	45.84	49.28	28.6
Gene types	127.37	64.42	47.51	31.56
Reference genome types	1.00	0.43	42.86	NA
Taxa types	1.00	0.30	30.19	NA

TABLE 8. Difference in mean hits@1 values per method in AvsA alignments when using different subdomains. The five datasets were used. When the method was unable to be utilized, NA was recorded.

Method	hits@1 all (%)	diff crm	diff crm2gene	diff crm2phen	diff crm2tfac
AlignE	94.76	-10.95	-8.65	-9.47	-7.34
AliNet	80.03	-8.76	-24.98	-10.20	-9.11
AttrE	85.23	-8.99	-16.77	-10.14	-13.50
BootEA	98.46	-11.96	-7.36	-9.81	-6.63
BootEA_R	93.05	-6.23	0.89	2.19	3.03
GCN Align	91.61	-6.25	0.84	-1.02	3.81
HolE	74.42	-9.74	-23.42	-14.48	-20.83
IMUSE	85.86	-5.52	-4.84	3.30	4.66
IPTransE	77.01	-9.58	-21.52	-12.64	-13.61
JAPE	53.57	-20.96	NA	NA	NA
MTransE	48.19	-14.59	7.18	2.40	-4.56
ProjE	63.69	-26.44	-28.55	8.94	NA
RDGCN	78.30	-30.58	-14.07	-15.88	-10.98
RotatE	94.77	-15.56	-14.62	-15.35	-12.42
RSN4EA	89.36	-8.80	-23.28	-64.25	-7.59
SEA	89.68	-6.72	0.04	-1.54	-4.06
SimplE	26.71	-10.72	-3.96	3.24	5.81
TransD	76.73	-10.59	-22.41	-13.49	-16.27
TransH	78.16	-11.06	-22.48	-14.58	-14.43
TransR	2.08	-0.82	-0.47	-0.63	1.02
Mean hits	74.08	-11.74	-12.02	-9.13	-6.83
Mean CV	3.90	5.31	8.16	7.58	3.64

demonstrated the most favorable results in terms of time efficiency, whereas the BootEA, BootEA-RotatE, and AttrE methods exhibited the most optimal performance within their respective clusters. While both BootEA and BootEA-RotatE exhibited high hits@1 values, they were associated with longer runtimes. In contrast, AttrE required shorter execution times but yields lower hits@1 values.

- Given the existence of disparate types of annotations across different databases, the performance of the hits@1 metric varied according to the domain in question (crm: ENdb>EnDisease>VISTA>DiseaseEnhancer >RefSeq; crm2gene: VISTA>ENdb>DiseaseEnhancer >EnDisease; crm2phen: ENdb>Disease Enhancer> EnDisease; crm2tfac: ENdb).

The difference in hits@1 values by subdomain was also reflected in the discrepancies in the aligned biological concepts (Table 8). In consideration of the fundamental entities, the classes pertaining to CRMs exhibited a reduction in hits@1 of 14.42%, and 9.10% in their instances (crm domain). On the other hand, classes that model relations between CRMs and other entities demonstrated a reduction of 12.38%, 11.85%, and 10.34% in the domains crm2gene, crm2phen, and crm2tfac, respectively. Furthermore, the difference was less pronounced in the case of instances. The other entities, such as genes, proteins, or diseases, also had lower matches compared to the EA using all the graphs.

B. ALIGNMENT OF DIFFERENT GRAPHS

The distinction of various subsections is necessary for the implementation of different alignment types, which were employed to test a range of real-world scenarios. These included the EA at the level of relations, at the attribute level, and in both situations (Table 5). EA with all graphs and by subgraphs were also performed and reported.

1) RELATIONS LEVEL

Only two pairs of datasets had enough entities to generate training and test sets: EnDisease-DiseaseEnhancer and RefSeq-VISTA pairs. Therefore, only two pairwise alignments could be performed with the typical or standard workflow. The results of these EA revealed the following noteworthy aspects:

- The standard workflow was unable to address EA for all KGs of interest.
- The hits@1 values were lower than those obtained in AvsA. This is an expected result due to the introduction of differences between entities, but also by a smaller

TABLE 9. Hits@1 values per method in typical AvsB alignments with differences at relations level (typical rel-AvsB), and using all domains or subgraphs. EnDisease and DiseaseEnhancer were abbreviated in the caption as EnDis and DisEnh, respectively.

	EnDis - DisEnh		RefSeq - VISTA		Means	
Method	**H@1 (%)**	**CV (%)**	**H@1 (%)**	**CV (%)**	**H@1 (%)**	**CV (%)**
AlignE	80.93	2.72	12.14	0.00	46.53	1.36
AliNet	64.20	4.29	6.07	84.86	35.14	44.57
AttrE	48.44	10.79	14.56	9.43	31.50	10.11
BootEA	91.83	1.20	11.89	2.89	51.86	2.04
BootEA_R	90.86	2.12	14.56	18.86	52.71	10.49
GCN Align	90.08	2.14	2.18	47.16	46.13	24.65
HolE	31.71	6.07	4.85	28.29	18.28	17.18
IMUSE	65.18	2.11	34.22	13.04	49.70	7.58
IPTransE	34.44	19.98	1.21	28.26	17.82	24.12
JAPE	32.69	5.05	1.70	20.23	17.19	12.64
MTransE	29.57	7.44	1.21	28.26	15.39	17.85
ProjE	54.67	21.64	8.25	8.32	31.46	14.98
RDGCN	53.50	2.57	2.43	56.58	27.96	29.58
RotatE	81.13	0.99	12.86	32.30	47.00	13.92
RSN4EA	79.13	9.16	3.16	18.68	41.14	16.65
SEA	83.85	1.64	12.14	5.66	47.99	3.65
SimplE	5.06	10.88	0.97	70.75	3.01	40.81
TransD	26.26	38.76	1.46	47.14	13.86	42.95
TransH	38.13	17.32	1.46	47.14	19.79	32.23
TransR	1.36	101.03	0.73	47.21	1.05	74.12
Mean	54.15	13.39	7.40	30.75	30.78	22.07

training set due to the existence of fewer common entities.

- Only two methods achieved hits@1 values above 50% (BootEA-RotatE and BootEA) (Table 9). In contrast, in AvsA alignments, 10 methods yielded hits@1 values exceeding 80% (6 of which at least 90%). Nevertheless, a considerable difference was evident between pairwise alignments. In the RefSeq-VISTA pair, no values higher than 35% were obtained, while in EnDisease-DiseaseEnhancer five of the six methods produced results higher than 80% (three of them exceeding 90%). These methods matched with the six methods that demonstrated the highest hits@1 values in AvsA alignments, namely (BootEA>BootEA-RotatE>GCN Align>SEA>RotatE>AlignE).
- Partitioning of data into distinct subgraphs also yielded lower hits@1 values than the complete KGs. As in the previous case, the ProjE, JAPE and AliNet methods also presented execution problems associated with the size of the graphs.
- The time requirements of EA methods exhibited comparable patterns, resulting in analogous groups during the clustering process.
- The clusters of methods with the highest hits@1-time ratio were similar to those obtained in AvsA (Fig. 5B), namely the cluster with GCN Align, IMUSE, and SEA methods; and the cluster with BootEA, RotatE, and AlignE (Fig. 6). Furthermore, the BootEA-RotatE method also showed a notable hits@1 performance, yet it is also characterized by a considerable runtime demand. In the crm subgraph, AttrE also emerges as a method with favorable outcomes.

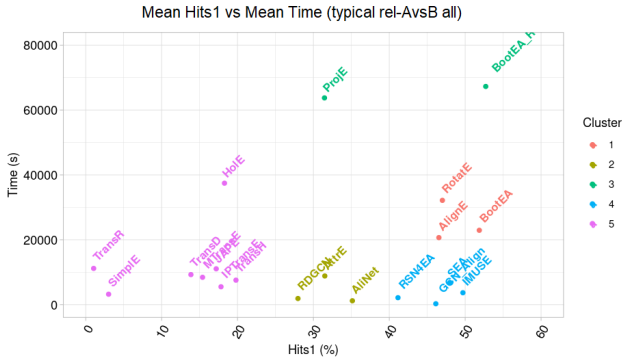

FIGURE 6. Plot of hits@1 values versus time values, with clustered methods according to its values of hist@1 and time. Results corresponding to 2 pairwise alignments of different KGs at relations level (rel-AvsB), with all data and the typical methodology.

According to these results, and as also indicated in the methodology, a modified version of the standard workflow was performed. This version was able to carry out EA even when the number of common entities between them is reduced. We refer to these EA as modified (mod-) alignments.

In these EA, the hits@1 results were positive across a wide range of methods, with nine of them demonstrating mean hits@1 results above 80% (GCN Align >BootEA RotatE > RDGCN > RotatE > AlignE > BootEA > AttrE > IMUSE > SEA) (Fig. 7 and Table 10). Excluding RDGCN, they also stood out in AvsA alignments. This resemblance gives rise to analogous clusters as evidenced by the hits@1 metric.

These methods are also highlighted in EA by subgraphs, with BootEA, BootEA-RotatE, GCN Align, and SEA being the constant methods in the best hits@1 branch when a clustering is performed (BootEA, BootEA-RotatE, and GCN Align were also constant in the AvsA alignments by subgraphs). The hits@1 values also exhibited lower scores than when all graphs were used.

As with the AvsA alignments, low CV values for hits@1 were obtained in most methods, particularly in those with the most favorable outcomes (Table 10). Consequently, the reported hits@1 values demonstrate consistency between replicates.

Additionally, a correlation was observed between hits@1 values and their CV ($cor = -0.74$, $df = 18$, p-value < 0.05), but no correlation was found between time and their CV.

The time requirements also varied according to the method applied. The methods that required the lowest and highest time requirements, respectively, were GCN Align and BootEA-RotatE. However, both methods yielded good hits@1 results.

We also identified methods with execution issues, particularly with the RefSeq-VISTA pair. AlignE, BootEA, RDGCN, RSN4EA, HolE, and TransR exhibited the same issues as previously described, namely, memory constraints, prolonged runtime, and internal errors. Additionally, ProjE and JAPE exhibited deficiencies in subgraph alignments, with errors attributable to sample size and internal inconsistencies.

TABLE 10. Hits@1 values (%) per method in AvsB alignments with differences at relations level. All graphs and the modified methodology were used (modified rel-AvsB all). When the method was unable to be utilized, NA was recorded.

Method	ENdb DisEnh	ENdb EnDis	EnDis DisEnh	RefSeq VISTA	VISTA DisEnh	VISTA EnDis	Mean hits@1	Mean CV
AlignE	97.24	86.76	96.42	NA	94.33	92.78	93.50	0.86
AliNet	49.72	47.57	68.11	99.66	13.40	22.22	50.11	7.80
AttrE	97.24	87.03	80.76	99.85	96.91	95.56	92.89	0.10
BootEA	95.03	85.14	97.70	NA	95.36	91.67	92.98	1.30
BootEA RotatE	93.65	89.19	97.63	98.15	92.27	96.11	94.50	2.20
GCN Align	98.34	89.19	96.81	99.81	96.91	96.67	96.29	0.23
HolE	30.39	48.11	63.43	NA	51.03	36.67	45.93	19.85
IMUSE	98.34	86.76	73.72	89.45	98.45	98.33	90.84	1.75
IPTransE	74.31	54.05	76.25	98.15	62.89	71.67	72.89	11.95
JAPE	41.44	45.68	63.20	89.79	35.05	46.67	53.64	3.94
MTransE	32.60	31.62	55.61	89.06	22.16	27.22	43.04	11.53
ProjE	4.70	10.27	47.12	51.31	5.16	11.11	21.61	7.47
RDGCN	96.13	92.70	84.15	NA	98.45	97.78	93.84	0.65
RotatE	93.09	87.03	93.69	94.12	97.42	97.22	93.76	1.99
RSN4EA	NA	71.30	NA	NA	NA	NA	71.30	1.72
SEA	81.49	82.43	91.82	98.69	85.57	90.00	88.33	3.30
SimplE	25.14	19.73	18.69	4.86	48.97	32.78	25.03	19.79
TransD	67.68	58.38	76.05	98.98	52.58	72.78	71.07	18.84
TransH	56.35	51.35	80.10	97.81	87.11	92.22	77.49	7.09
TransR	6.91	9.73	8.41	NA	6.70	10.56	8.46	17.10
Mean	65.25	61.70	72.09	86.41	65.30	62.68	68.88	6.97

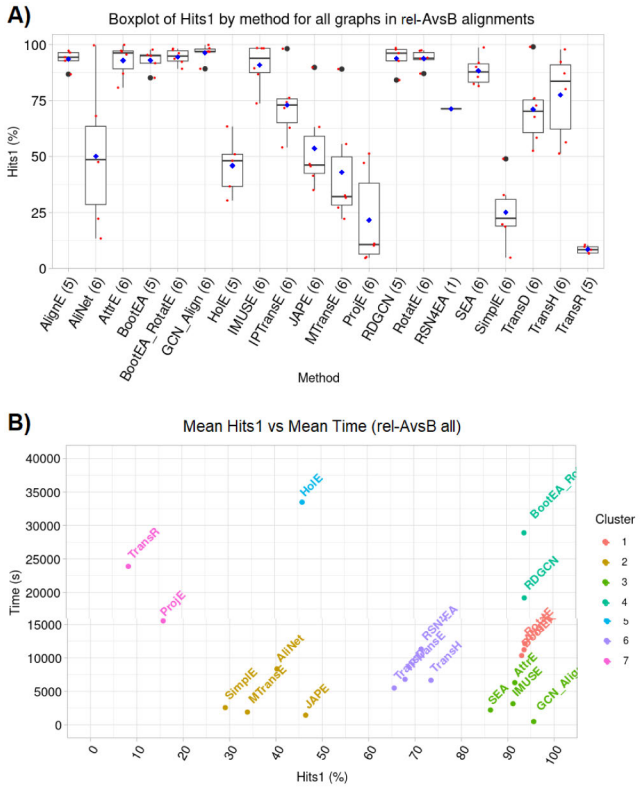

FIGURE 7. Results associated with EA of different KGs at relations level. All graphs and the modified methodology were used (modified rel-AvsB all). A) The boxplot depicts the distribution of hits@1 values by method. B) Plot of hits@1 versus time, with clustered methods also by hits@1 and time values. The RefSeq dataset was omitted due to the absence of values for several tested methods.

The clustering of methods by hits@1 and time, both with all graphs and with different subgraphs, also revealed a cluster comprising the methods that yielded the best hits@1 results. GCN Align, IMUSE and SEA methods were identified as the most optimal and persistent results across different KGs. When considering only EA where most methods were operational (excluding the RefSeq-VISTA pair), these methods and AttrE were also highlighted (Fig. 7B).

Regarding entity count, the entities corresponding to CRMs exhibited an average match rate of 90.20%, while the entities pertaining to the relations between CRMs and other entities demonstrated an average match rate of 89.24%. Furthermore, the results demonstrated consistency, as evidenced by the low variability observed. CV was found to be 4.33% and 4.45%, respectively. These results correspond to the EA conducted using all KGs. The results pertaining to subgraph EA exhibited minimal variation in comparison to the EA utilizing all KGs. The CRMs exhibited a 1.61% decline, while the entities pertaining to relations demonstrated an average reduction of 2.71%. In contrast, other entities acting as object nodes had larger match decreases.

2) ATTRIBUTES LEVEL

The results shown in Fig. 8A and Table 11, are comparable to those obtained for the AvsA alignments without RefSeq (Table 12).

Ten methods yielded hits@1 values exceeding 80% (BootEA RotatE>BootEA>GCN Align>RotatE>AlignE>SEA>RSN4EA>AttrE>AliNet>IMUSE), and seven of which exceeded 90%. These methods are analogous to those obtained in the AvsA alignments, exhibiting slightly diminished hits@1 outcomes and a lower CV between samples. Only IMUSE exhibited a more pronounced decline.

The clustering of methods based on both the hits@1 values and the time values yielded comparable results, as did the clustering based on either of these two variables alone. As IMUSE demonstrated a decline of over 9% in the

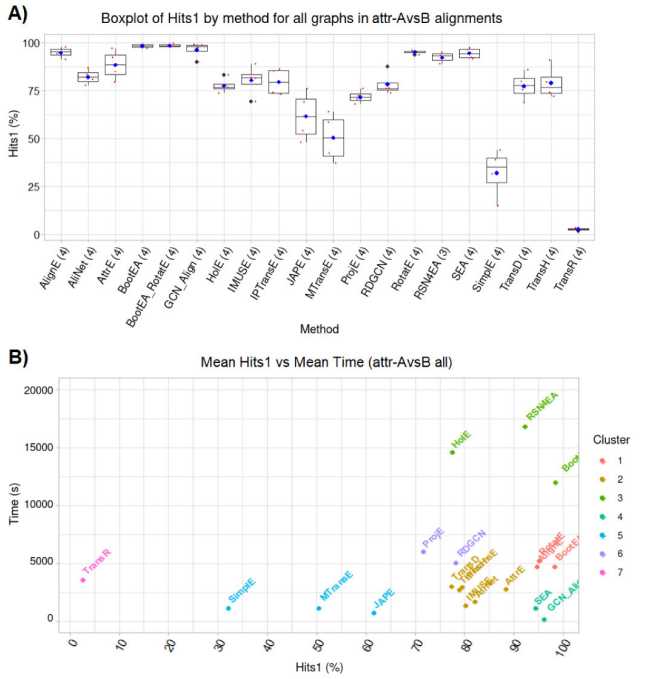

FIGURE 8. Hits@1 results corresponding to EA of different KGs at attributes level (attr-AvsB). The RefSeq dataset was omitted due to the absence of values for several methods tested. A) The boxplot depicts the distribution of hits@1 values by method. B) Plot of hits@1 versus time, with clustered methods also by hits@1 and time values.

hits@1 values, this approach yielded a cluster with inferior results in the method clustering. Therefore, the methods that demonstrated the greatest efficacy in the hits@1-time relation were the cluster of GCN Align and SEA, as well as the BootEA and AttrE methods, which are the optimal ones within their respective clusters (Fig. 8B). Additionally, BootEA-RotatE exhibits high hits@1 values despite its prolonged execution times.

The EA of subgraphs also resulted in patterns similar to those discussed above.

In the entity count utilizing the entire data set, there were also no significant differences with respect to AvsA alignments excluding RefSeq. On the one hand, the classes pertaining to CRMs were aligned in 87.10% of cases, with a 5.68% CV (86.68% with a 4.76% CV in AvsA without RefSeq). The instances of these classes were aligned in 81.61% of cases, with a 5.48% CV, a slight decrease from the previous value of 82.15% and 4.88% CV.

Conversely, classes corresponding to relations between CRMs and other entities were aligned on average 80.66%, with 7.02% CV, while instances were aligned 77.05% with 4.20% CV. Previously, the respective values were 81.43% with 3.96% CV and 77.50% with 5.28% CV.

Nevertheless, the difference in EA by subdomains was greater in attr-AvsB. The application of classes pertaining to CRM sequences resulted in a 20.87% reduction in the percentage of matches, whereas instances exhibited a 13.11% decline. With regard to the entities over relations, the reduction was 12.28% and 4.52%, respectively.

TABLE 11. Hits@1 values (%) per method in EA of different KGs at attributes level, and using all data domains (attr-AvsB all). RefSeq was excluded due to its incompatibility with several methods.

Method	DisEnh	ENdb	EnDis	VISTA	Mean H@1	Mean CV
AlignE	91.13	96.23	94.21	97.39	94.74	0.65
AliNet	86.65	83.54	80.56	77.52	82.06	0.87
AttrE	79.38	92.13	84.76	96.96	88.31	4.22
BootEA	96.97	97.84	99.05	98.96	98.20	0.64
BootEA_R	98.55	97.78	99.44	97.80	98.39	0.28
GCN Align	89.65	98.45	97.38	99.05	96.13	0.47
HolE	73.45	76.83	76.67	83.16	77.53	5.36
IMUSE	69.14	88.86	81.75	81.30	80.26	0.68
IPTransE	73.82	84.98	73.02	86.08	79.47	3.28
JAPE	47.75	68.90	76.03	53.97	61.66	3.21
MTransE	37.12	58.31	63.89	42.16	50.37	5.94
ProjE	72.27	70.57	67.70	75.93	71.62	2.08
RDGCN	73.50	87.42	76.11	76.07	78.27	1.70
RotatE	95.52	93.51	95.24	95.99	95.06	0.82
RSN4EA	88.74	95.00	93.16	NA	92.30	0.45
SEA	92.31	91.46	96.27	97.37	94.35	0.89
SimplE	14.63	43.74	31.19	38.67	32.06	14.31
TransD	75.38	79.77	68.49	85.70	77.33	5.27
TransH	73.95	79.21	71.83	90.75	78.93	1.38
TransR	2.73	3.22	2.70	1.68	2.58	38.16
Mean	71.63	79.39	76.47	77.71	76.48	4.53

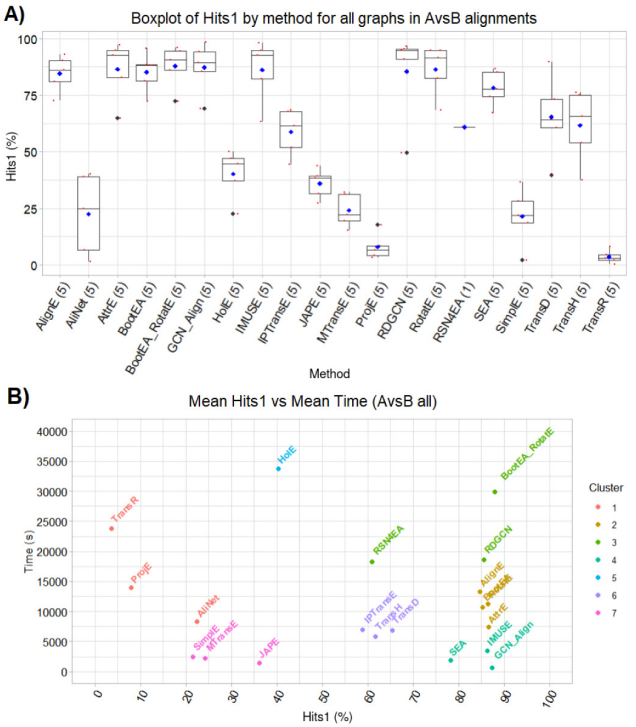

FIGURE 9. Hits@1 results corresponding to EA of different KGs at attributes and relations level, and using the modified methodology and all data (mod-AvsB all). The RefSeq dataset was omitted due to the absence of values for several tested methods. A) The boxplot depicts the distribution of hits@1 values by method. B) Plot of hits@1 versus time, with clustered methods also by hits@1 and time values.

3) EDGES AND ATTRIBUTES LEVEL

In this case, eight methods yielded hits@1 results exceeding 80% (BootEA RotatE>GCN Align>AttrE>RotatE>IMUSE >RDGCN>BootEA>AlignE). However, no method demonstrated values exceeding 90% (Fig. 9A and Table 13).

TABLE 12. Hits@1 values (%) comparison between different EA types using all KGs. The RefSeq dataset was omitted due to the absence of values for several tested methods. Standard deviation (SD) values were also provided, as well as the difference of hits1 compared to the AvsA alignments was also included.

method	AvsA hits1	AvsA SD	attr-AvsB hits1	attr-AvsB SD	attr-AvsB diff	rel-AvsB hits1	rel-AvsB SD	rel-AvsB diff	AvsB hits1	AvsB SD	AvsB diff
AlignE	94.76	0.44	94.74	0.61	-0.02	93.50	0.80	-1.26	84.64	1.76	-10.12
AliNet	82.63	0.67	82.06	0.72	-0.57	40.20	1.91	-42.43	22.48	2.34	-60.15
AttrE	89.03	4.95	88.31	3.49	-0.72	91.50	0.09	2.47	86.53	2.08	-2.50
BootEA	98.46	0.31	98.20	0.63	-0.26	92.98	1.22	-5.48	85.26	1.24	-13.20
BootEA_R	98.52	0.32	98.39	0.27	-0.13	93.77	2.20	-4.75	87.94	0.87	-10.58
GCN Align	96.64	0.43	96.13	0.46	-0.51	95.58	0.26	-1.06	87.28	0.47	-9.36
HolE	75.80	0.89	77.53	4.14	1.73	45.93	8.01	-29.87	40.32	9.44	-35.48
IMUSE	89.66	0.99	80.26	0.49	-9.40	91.12	1.53	1.46	86.26	1.91	-3.40
IPTransE	78.85	1.91	79.47	2.58	0.62	67.83	9.58	-11.02	58.87	11.68	-19.98
JAPE	61.88	2.05	61.66	2.01	-0.22	46.41	1.89	-15.47	36.10	2.46	-25.78
MTransE	51.50	1.00	50.37	2.94	-1.13	33.84	4.01	-17.66	24.14	2.81	-27.36
ProjE	70.87	0.86	71.62	1.44	0.75	15.67	0.59	-55.20	8.00	1.24	-62.87
RDGCN	78.30	0.36	78.27	1.33	-0.03	93.84	0.61	15.54	85.49	1.19	7.19
RSN4EA	89.36	2.82	92.30	0.42	2.94	93.69	1.23	4.33	60.87	1.16	-28.49
RotatE	94.77	0.94	95.06	0.78	0.29	71.30	2.02	-23.47	86.41	4.20	-8.36
SEA	94.25	0.22	94.35	0.84	0.10	86.26	3.37	-7.99	78.29	4.39	-15.96
SimplE	32.90	2.31	32.06	4.08	-0.84	30.04	5.91	-2.86	21.51	5.76	-11.39
TransD	78.49	2.08	77.33	4.30	-1.16	65.49	14.92	-13.00	65.45	5.21	-13.04
TransH	80.28	2.46	78.93	1.18	-1.35	73.43	5.43	-6.85	61.70	7.48	-18.58
TransR	2.56	0.95	2.58	1.04	0.02	8.46	1.48	5.90	3.58	0.80	1.02
mean	76.98	1.35	76.48	1.69	-0.50	66.54	3.35	-10.44	58.56	3.43	-18.42

TABLE 13. Hits@1 values (%) per method in EA of different KGs at both levels (attributes and relations), using all data domains and the modified methodology (mod-AvsB all). RefSeq was excluded due to its incompatibility with several methods.

Method	ENdb DisEnh	ENdb EnDis	EnDis DisEnh	VISTA DisEnh	VISTA EnDis	Mean Hits1	Mean CV
AlignE	93.09	81.08	72.70	90.21	86.11	84.64	2.03
AliNet	38.95	40.27	1.48	6.70	25.00	22.48	9.37
AttrE	97.24	82.97	64.84	94.85	92.78	86.53	2.78
BootEA	95.58	81.35	72.35	88.66	88.33	85.26	1.46
BootEA_RotatE	96.13	86.22	72.20	90.72	94.44	87.94	0.97
GCN_Align	98.34	85.41	68.89	94.33	89.44	87.28	0.57
HolE	37.29	47.03	22.43	44.85	50.00	40.32	21.73
IMUSE	98.07	82.16	63.47	94.85	92.78	86.27	2.37
IPTransE	67.68	51.89	44.55	68.56	61.67	58.87	17.96
JAPE	38.40	43.78	27.41	31.44	39.44	36.10	6.55
MTransE	32.04	31.35	19.63	15.46	22.22	24.14	10.07
ProjE	3.32	6.49	17.72	4.12	8.33	8.00	28.00
RDGCN	96.69	90.81	49.57	94.85	95.56	85.49	1.52
RotatE	94.75	82.43	68.34	94.85	91.67	86.41	1.55
RSN4EA	NA	60.87	NA	NA	NA	60.87	6.06
SEA	85.36	74.32	67.37	86.60	77.78	78.29	5.45
SimplE	21.82	18.65	2.14	36.60	28.33	21.51	30.72
TransD	73.20	64.32	39.49	89.69	60.56	65.45	8.19
TransH	65.75	54.05	37.42	76.29	75.00	61.70	11.68
TransR	8.01	2.97	0.39	2.06	4.44	3.58	23.34
Mean	65.35	58.42	42.76	63.46	55.80	58.56	9.62

In contrast, in the modified rel-AvsB alignments (different relations and same attributes), nine methods yielded results higher than 80% (eight of which were greater than 90%). This discrepancy is illustrated in Table 12, which summarizes the hits@1 results of the different methods in the scenarios that were studied.

When considering these methods as reference, it becomes evident that the observed difference is greater than the intrinsic variability observed between replicates. Furthermore, these methods with notable results were also present in the methods obtained in rel-AvsB, but with a different ranking.

This similarity was also evident in the clustering of methods based on their values of hits@1, time, and both. In the clustering with both variables, GCN Align, SEA, and IMUSE emerged as the most prominent cluster. Additionally, AttrE and BootEA-RotatE were identified as the most effective methods within their respective clusters.

Regarding the results of entity count, the classes belonging to CRM sequences were aligned, on average, in 44.69 % of the cases, with a high degree of variability between replicates (24.79 % CV). In the case of classes pertaining to relations between CRM sequences and other entities, the percentage of aligned entities was 56.88%, with a similarly high degree

of variability (22.83% CV). These results differ significantly from those obtained in the EA of AvsB type (without RefSeq), where 91.50% of CRM sequences were aligned (4.66% CV), and 89.24% of relation classes (4.45% CV). This represents a difference of 46.81% and 32.36%, respectively.

In contrast, the EA by subgraphs revealed a 4.00% reduction in the classes pertaining to CRM sequences, and a 2.60% reduction in the classes related to relations. The entities that acted as object nodes also showed worse results than in the EA with all KGs.

IV. DISCUSSION

EA methods can contribute to find matches between entities in disparate KGs, but given the growing number of EA methods with different approaches and their different performances in different domains, we need to have a benchmark to compare the different methods. The work in this paper addresses this issue by comparing different methods in multiple scenarios on an equitable baseline.

A total of 2,800 alignments have been executed for studying the performance of different methods on biological KGs pertaining to gene regulation. In particular, we have focused on KGs related to enhancer sequences and their relationships with other biological entities, including target genes and phenotypes.

Identifying optimal methods for integrating disparate gene regulatory resources would facilitate the attainment of data interoperability in a domain of clinical and research significance. This is particularly pertinent given the demonstrated clinical relevance of enhancer sequences (and, indeed, CRMs more broadly) in the context of disease pathogenesis [69], [74]. In turn, the interoperability between KGs in a federated environment could generate an extensive network suitable to query and analyze heterogeneous data, formulate hypotheses, and discover new knowledge in relevant fields [2], [13], [57], [58]. Two examples are the development of new drugs [88], and the simulation of digital twins [89], [90] in the context of personalized medicine. However, it is important to note that alignments made using EA methods are predictions that require validation. Consequently, while these tools facilitate the work, their practical application will be further enhanced by the assistance of a human expert who can distinguish successful cases from false positives.

The workflow used [82] enabled a fair comparison of the EA methods included in the study. This is a fundamental task to provide a benchmark on the performance of EA methods in a novel domain. Furthermore, the use of this workflow in different contexts also contributes to its validation.

In addition, our work has identified difficulties in performing EA when the KGs to be aligned have few entities in common, that can limit the training of different methods. This issue required to extend the workflow with a methodological variant, which allowed us to address these situations and provide positive results. However, its suitability in other scenarios needs to be further studied.

The performance of the 20 methods under study was reported in terms of hits@1 values and execution times, as well as by type of entity or biological concept aligned. This provided information about which biological entities exhibited superior performance and which were suitable for improvement.

Other metrics such as hits@5, hits@10, MR, and MRR were analyzed, but they did not contribute to providing better results than those already obtained with hits@1 and execution times. For this reason, these results are reported in the supplementary material available in our repository.

Although EA methods should adapt to the structure of current KGs, in this paper we have also explored the performance of methods with a focus on the evaluation of existing schemas. Several methods with different approaches have been used, and their performance provides information on which entities have a more standardized semantic modeling, considering that these methods are designed with features of common KGs. Thus, the results of the alignments provide clues about which features of the used schemas can be improved, e.g. by adding more relations or attributes. In conclusion, this improvement can facilitate the establishment of equivalence relations between entities from disparate KGs, thereby enhancing the interoperability of the underlying resources. A task that is also necessary if we want to exploit the potential of artificial intelligence, since the different semantic representation prevents intelligent computer agents from discovering information about the same entities that are not uniformly represented in the KGs.

Next, we discuss the performance of EA methods and the main limitations of our work. Table 14 summarizes the best performing methods in the different alignment scenarios analyzed.

A. PERFORMANCE OF THE ALIGNMENT METHODS

The hit and execution time results exhibited variability contingent on the methodology employed, the datasets utilized as input, and the strategy or type of alignment undertaken. The most favorable outcomes were observed in experiments conducted with GCN Align, SEA, IMUSE, and AttrE. The GCN Align method yielded the most favorable outcomes, while the AttrE approach demonstrated particular efficacy in KGs where entities exhibited disparate relations and attributes, which is the most probable scenario.

The hit results for BootEA, AlignE, and RotatE were also satisfactory. However, their execution times were longer, and generally required more memory. In particular, these methods were not always able to perform alignments involving the RefSeq dataset, as they required more than the 250 GB RAM limit. For example, in AvsA alignments. RefSeq was the largest dataset analyzed in this study. Consequently, these methods may not be suitable for large datasets and services without high throughput.

RDGCN and AliNet were also methods that required high memory requirements. While RDGCN required more than 250 GB for AvsA alignments with RefSeq, AliNet

TABLE 14. Summary of best performing methods in the different alignment scenarios analyzed. Only methods that showed hits@1 greater than 80% across all analyzed KGs (all KGs and subgraphs) were included. For the best performance category, we also considered the coincidence of the methods in the best-performing cluster across the different analyzed KGs.

Scenario	Hits@1 >80 %	Hits@1 >80 % without incidences	Best hits@1-time performance
AvsA	AlingE, BootEA, BootEA-RotatE, GCN Align, IMUSE, SEA	BootEA-RotatE, GCN Align, IMUSE, SEA	GCN Align, IMUSE, SEA
rel-AvsB	AlingE, AttrE, BootEA, BootEA-RotatE, GCN Align, IMUSE, SEA	AttrE, BootEA-RotatE, GCN Align, IMUSE, SEA	GCN Align, IMUSE, SEA
attr-AvsB	AlignE, BootEA, BootEA-RotatE, GCN Align, SEA	BootEA-RotatE, GCN Align, SEA	GCN Align, SEA
AvsB	GCN Align	GCN Align	GCN Align, IMUSE, SEA
Common methods	GCN Align	GCN Align	GCN Align, SEA

could complete the alignment with requirements close to 100 GB. The other analyzed methods had low memory requirements, close to or less than 10 GB. On the other hand, RSN4EA was sensitive to execution times, and the JAPE and ProjE methods to small subgraphs. Consequently, the aforementioned methods demonstrated lower execution capacity and scalability problems with respect to the existing diversity of KGs.

Additionally, BootEA-RotatE is worthy of mention due to its ability to achieve highly accurate results in diverse EA, while maintaining robustness. Nevertheless, it is also one of the most time-consuming methods, and consequently, it may also have scalability problems with large KGs.

According to the results obtained, summarized in Table 12, the best hits@1 were obtained for the alignment of identical graphs (AvsA), which was an expected finding. These were followed by alignments of different graphs at the attribute level (attr-AvsB), at the relation level (rel-AvsB) and, lastly, at both levels (AvsB). This pattern was observed for most methods, although some exceptions or discrepancies were found, which can be justified by the intrinsic variability between replicates (observable in the standard deviation values). At this point, the stochasticity in the distribution of entities between training, testing and validation sets is an aspect that can contribute to this variability.

It should be noted that two methods presented better results in AvsB alignments than in AvsA beyond intrinsic variability. Namely, TransR and RDGCN, for which those results are not completely explained by the standard deviation values. TransR exhibited low hit values and high variability in all alignments performed, so this difference outside the standard deviation ranges could be motivated by the small number of replicates used. This is because the methods that produced the worst hits results also exhibited greater intrinsic variability. Increasing the number of replicates could also help to explain the differences in RDGCN, but the behavior of both methods could require further study.

A variety of methods have been demonstrated to yield high results in self-alignment, with hits@1>80%: BootEA, RotatE, AlignE, BootEA-RotatE, GCN Align, SEA, IMUSE, AttrE, and RSN4EA. This is also indicative of KGs that containing entities with suitable attributes and relations, enabling the identification of similarities and differences between entities.

The same methods (except for RDGCN instead of RSN4EA) yielded hits@1>80% for the EA of KGs with different relations. In the case of EA of KGs with different attributes, the outcomes were also analogous to those of self-alignment, with the exception of the IMUSE method, whose results exhibited a more pronounced influence, but higher than 80% in hits@1. Poorer performance was more significant when the differences were at the relation level, but the best performing methods still maintained proximal results. In this case, RotatE and SEA were some of the best performing methods that were most affected.

Therefore, the top-ranked methods exhibit reduced variation when some of the relations or attributes are altered. Moreover, in the EA of different KGs at both levels (edges and attributes), the methods that demonstrated the greatest efficacy also coincided. However, the hits@1 values were reduced more significantly due to the larger differences between entities. Consequently, the methods that showed better hits results were fundamentally common among strategies, although with a different ranking distribution depending on the approach. Furthermore, the differences at the relations level had a greater impact.

The EA of subgraphs did not yield enhanced results when compared to the EA of the entire KGs. This can be attributed to the fact that EA methods demonstrate enhanced performance when more information is available.

The different EA methods employ distinct approaches to determine which entities may correspond to the same real-world entity. To accomplish this, the methods must extract the main features of each entity. These features can include the graph structure, associated text, and ontological categories. Partitioning KGs by biological domains generates nodes that act as objects, but have reduced features because their detailed modeling has been disaggregated into an independent KG. Consequently, aligning those entities is more difficult for EA methods, thus decreasing the overall hits performance. On the other hand, since the KG structure is affected, the matches of the subject entities can also be affected.

However, it is important to note that some methods produced similar or slightly better results for some subgraphs than for all KGs. This can be explained by the intrinsic vari-

ability of the executions, but also by the superior performance of certain methods in the EA of KGs with entities that have a smaller number of relationships and attributes. Conversely, since not all subgraphs exhibited the same behavior, the exclusion of poorly enriched entities that hinder the hit computation may also contribute to these results.

The best performing methods in subgraphs by hits@1 and time values were a subset of those obtained in EA with all KGs. The most notable methods were GCN Align, IMUSE, SEA, BootEA, AttrE, and BootEA-RotatE, the first three being the most consistent across different domains or subgraphs. AttrE showed greater variability between subgraphs, and BootEA had memory issues with some RefSeq subgraphs. Finally, although BootEA-RotatE was one of the methods with the best hit performance, it required a longer runtime. Therefore, although subgraph partitioning may improve the scalability of some methods by reducing the required memory, other graph partitioning approaches and improvements to the methods themselves should be explored.

On the other hand, the use of different KG combinations was an unexplored aspect, an interesting issue to optimize the subgraph partitioning strategy. Since complete KGs correspond to the merging of the different subgraphs or subdomains derived from a database, this is an aspect that could be explored in future work reusing the workflow.

The experiments also demonstrated intrinsic variability between replicates, as evidenced primarily by the CV values. In the EA of same and different KGs, the CV of hits@1 values was low in most methods, particularly in those with the most favorable outcomes. This indicates a high degree of consistency in the hits@1 results between replicates ($<$10% CV).

This variability is higher in the EA of subgraphs. A high negative correlation between hits@1 values and their CVs was also observed in the majority of alignments. Therefore, high hits@1 values are accompanied by low variability between replicates. Consequently, a higher variability was observed in AvsB alignments, followed by rel-AvsB, attr-AvsB and AvsA.

However, this correlation was not obtained for time requirements and their CVs, as these CVs of time are higher than those obtained for hits@1. This discrepancy could be justified due to the fact that each replication does not involve the same number of iterations in each pairwise alignment.

B. PERFORMANCE BY BIOLOGICAL ENTITY TYPE

Analyzing performance by type of biological entity is useful because it allows the performance of the methods to be broken down by biological entity type. This is relevant because the final success rate of each method is determined by the combined matches of the different entities.

The results reported in this article have focused on the main entities of each KG, as these are domain-specific KGs. However, the supplementary material details the results for each specific entity. In this discussion, we address all entities based on their topological design in the model, that is, based on a common design. This approach enables us to evaluate the performance of each entity, and the possible reasons for misalignments. Each method presents different hit values, since they use different strategies to determine which entities may correspond to the same real entity.

In line with the results, the methods that rely heavily on structural similarity propagation tend to underperform when subgraphs are heterogeneous, while methods that incorporate semantic or textual representations generally show improved robustness when entity labels and biological annotations are informative.

The core entities of the model yielded the best alignment results and the lowest CV values. These core entities are associated with CRM sequences and entities about relations between these sequences and other biological entities, like target genes, transcription factors, or diseases. These entities possess a greater number of attributes and relations, which represents a significant feature. Nevertheless, the performance of this category of entities varies according to the type of alignment. When the entities differ in the number of both attributes and relations, the percentage of aligned entities is significantly diminished.

The diversity of attributes and relations should be accompanied by the standardization of the nomenclature of the sequences. The findings indicate that the presence of shared attributes can mitigate the discrepancies associated with the varying relations between entities.

Entities such as chromosome types possess attributes but exhibit limited diversity in their relations due to their role as object nodes of the central entities. Furthermore, the relations demonstrate minimal diversity, as only a single entity type, namely, CRM sequences, is linked to the chromosome nodes. These type of entities demonstrate proficiency in AvsA alignments but exhibit diminished hits capacity when confronted with rel-AvsB alignments, wherein the relationships undergo a transformation. The incorporation of additional relations involving nodes, both as subject and object nodes, has the potential to enhance the performance of entities exhibiting this pattern in a schema.

Entities such as genes, proteins, phenotypes, and taxa engage in relations but lack attributes. This phenomenon may occur as a result of the reuse of entities from other KGs or schemas. Importing the information present in those KGs or schemas would enhance the overall outcomes.

A metric has been provided which assigns a score to each dataset in accordance with the ranking by hits@1 score. VISTA was the highest scoring dataset, followed by EnDisease, ENdb, DiseaseEnhancer and RefSeq. While EA methods are most effective when entities possess a multitude of attributes and relations, VISTA is not the dataset with the greatest number of relations to other entities. The entity count indicates that the attributes and relations presented in these CRMs are sufficient to obtain optimal hits results. Conversely, the lower hits values observed in the entities corresponding to proteins and phenotypes result in lower scores for EnDisease, ENdb, and DiseaseEnhancer. Consequently, while the existence of relations between entities facilitates the

identification of features and differences between entities, the final hits@1 score may be affected by incomplete entities if the entities in question are not sufficiently enriched.

In light of these considerations, we propose that the evaluation of KGs based on their successful alignment with entities represents a promising avenue for a systematic approach to evaluation. Such evaluation has typically been conducted through case studies due to the absence of established systematic evaluation standards [32]. Nevertheless, the efficacy of EA techniques can be assessed according to entity type, enabling the identification of those entities exhibiting superior or inferior performance and, consequently, those entities that can be more efficiently modeled.

C. LIMITATIONS AND FURTHER WORK

The large number of EA performed is the result of two factors: the use of two replicates for each alignment and the application of 20 EA methods, five reference datasets, and different alignment strategies.

In this study, we report the differences obtained between different types of alignments. However, a larger number of replicates would allow for a contrast of means, thereby providing results with greater statistical significance. A larger sample size, i.e, a higher number of databases used, also would allow for more robust statistical analysis.

Furthermore, the databases used as a sample correspond to repositories manually curated by experts. Therefore, they are databases of higher authority and their data are typically reused in other repositories. Other databases contain sequences collected from high-throughput experiments and contain a larger volume of data to be validated.

Besides, in this work the EA were executed using the default OpenEA configuration, i.e. we did not go into individual parameter optimization because this was not a proposed objective. In addition, we used the default configuration to ensure a fair comparison and to provide comparable results with other state-of-the-art studies and possible future works. However, in the pipeline used, we generated a validation set that can be exploited in future work to optimize the methods and achieve better performance.

Furthermore, it would also be of particular interest to explore and/or develop complementary metrics that address misalignment issues in detail. The biomedical terminology is complex and has hierarchical structures, so it would be interesting to consider some misaligned terms.

It should be noted that the execution time experiments were not conducted under completely isolated conditions. Therefore, although this variability can be attributed to the discrepancy in the number of iterations in each alignment, it is plausible that other processes running on the server may have been influenced by the competition for available resources.

The data were modeled using a common schema [13], since no alternative semantic schema could be identified that would represent the data from the selected sources in RDF. Consequently, we recreated different real-life scenarios by aligning datasets with disparate relations and generating distinct attributes for the modeled sequences. However, this study did not include a comparison of KGs with a higher level of semantic complexity. It seems likely that the growing interest in biological KGs will result in the development of alternative models that will facilitate further exploration of this topic.

Ultimately, this study has been conducted within the domain of gene regulation, where the results illustrate that these methods can contribute to the EA and thereby to the interoperability between KGs. However, it would be beneficial to replicate this study in other biological and biomedical domains to ascertain how the performance of the methods may vary across different domains. Multilingual systems, such as clinical systems, would also be an interesting application area, because the variability of the languages used is greater than in scientific databases. In this context, the methodology and workflow provided in this work can be applied to the study of other domains of interest to identify efficient methods that can contribute to the field.

Moreover, since EA is a field in constant evolution and new methods are proposed continuously, the inclusion of other EA methods at the corresponding stage also facilitate the exploration of methods not addressed in this work and future methods for adapting the results to advances in the field. Similarly, the evaluation and comparison with methods using other approaches also constitute a future work of interest. At this point, the best performing methods from OAEI initiatives, such as LogMap [91] or Matcha [92].

V. CONCLUSION

The GCN Align method yielded the most favorable outcomes in terms of hit values and time requirements among the methods tested. Other methods yielded noteworthy results, including BootEA, BootEA-RotatE, AttrE, SEA, IMUSE, RotatE, and AlignE. However, not all of these methods demonstrated the same degree of robustness as GCN Align, and they exhibited disparate hits@1-time ratios. The alignment by subgraphs yielded inferior results compared to the use of the entire graph. The method allows for the degree of alignment of the entities in the schema to be determined and for the schema's suitability for the datasets to be evaluated.

REFERENCES

[1] N. Fanourakis, V. Efthymiou, D. Kotzinos, and V. Christophides, "Knowledge graph embedding methods for entity alignment: Experimental review," *Data Mining Knowl. Discovery*, vol. 37, no. 5, pp. 2070–2137, Sep. 2023, doi: 10.1007/s10618-023-00941-9.

[2] A. F. Al Musawi, S. Roy, and P. Ghosh, "A review of link prediction applications in network biology," *IEEE Access*, vol. 13, pp. 54997–55016, 2025, doi: 10.1109/ACCESS.2025.3553732.

[3] B.-W. Zhao, X.-R. Su, Y. Yang, D.-X. Li, G.-D. Li, P.-W. Hu, Z.-H. You, X. Luo, and L. Hu, "Regulation-aware graph learning for drug repositioning over heterogeneous biological network," *Inf. Sci.*, vol. 686, Jan. 2025, Art. no. 121360, doi: 10.1016/j.ins.2024.121360.

[4] K. G. Cortes, S. Sundar, S. Gehrke, K. Manpearl, J. Lin, D. R. Korn, H. Caufield, K. Schaper, J. Reese, K. Koirala, L. E. Hunter, E. K. Carter, M. DeLuca, A. Krishnan, C. Mungall, and M. Haendel, "Improving biomedical knowledge graph quality: A community approach," 2025, *arXiv:2508.21774*.

[5] J. K. Chaudhari, S. Pant, R. Jha, R. K. Pathak, and D. B. Singh, "Biological big-data sources, problems of storage, computational issues, and applications: A comprehensive review," *Knowl. Inf. Syst.*, vol. 66, no. 6, pp. 3159–3209, Jun. 2024, doi: 10.1007/s10115-023-02049-4.

[6] S. Pal, S. Mondal, G. Das, S. Khatua, and Z. Ghosh, "Big data in biology: The hope and present-day challenges in it," *Gene Rep.*, vol. 21, Dec. 2020, Art. no. 100869, doi: 10.1016/j.genrep.2020.100869.

[7] M. Danielewski, M. Szalata, J. K. Nowak, J. Walkowiak, R. Słomski, and K. Wielgus, "History of biological databases, their importance, and existence in modern scientific and policy context," *Genes*, vol. 16, no. 1, p. 100, Jan. 2025, doi: 10.3390/genes16010100.

[8] B.-H. Yoon, S.-K. Kim, and S.-Y. Kim, "Use of graph database for the integration of heterogeneous biological data," *Genomics Informat.*, vol. 15, no. 1, pp. 19–27, 2017, doi: 10.5808/gi.2017.15.1.19.

[9] K. Fecho et al., "Announcing the biomedical data translator: Initial public release," *Clin. Transl. Sci.*, vol. 18, no. 7, p. 70284, Jul. 2025, doi: 10.1111/cts.70284.

[10] B. E. Santangelo, M. Apgar, A. S. B. Colorado, C. G. Martin, J. Sterrett, E. Wall, M. P. Joachimiak, L. E. Hunter, and C. A. Lozupone, "Integrating biological knowledge for mechanistic inference in the host-associated microbiome," *Frontiers Microbiology*, vol. 15, Apr. 2024, Art. no. 1351678, doi: 10.3389/fmicb.2024.1351678.

[11] A. Riquelme-García, J. Mulero-Hernández, and J. T. Fernández-Breis, "Annotation of biological samples data to standard ontologies with support from large language models," *Comput. Struct. Biotechnol. J.*, vol. 27, pp. 2155–2167, May 2025, doi: 10.1016/j.csbj.2025.05.

[12] H. Chen, T. Yu, and J. Y. Chen, "Semantic Web meets integrative biology: A survey," *Briefings Bioinf.*, vol. 14, no. 1, pp. 109–125, Jan. 2013, doi: 10.1093/bib/bbs014.

[13] J. Mulero-Hernández, V. Mironov, J. A. Miñarro-Giménez, M. Kuiper, and J. T. Fernández-Breis, "Integration of chromosome locations and functional aspects of enhancers and topologically associating domains in knowledge graphs enables versatile queries about gene regulation," *Nucleic Acids Res.*, vol. 52, no. 15, p. e69, Aug. 2024, doi: 10.1093/nar/gkae566.

[14] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: Tool for the unification of biology," *Nature Genet.*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.

[15] K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner, "The sequence ontology: A tool for the unification of genome annotations," *Genome Biol.*, vol. 6, no. 5, pp. 1–12, Apr. 2005, doi: 10.1186/gb-2005-6-5-r44.

[16] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration," *Nature Biotechnol.*, vol. 25, no. 11, pp. 1251–1255, Nov. 2007, doi: 10.1038/nbt1346.

[17] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C.-N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann, "Knowledge graphs," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–37, May 2022, doi: 10.1145/3447772.

[18] M. Salvadores, P. R. Alexander, M. A. Musen, and N. F. Noy, "BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF," *Semantic Web*, vol. 4, no. 3, pp. 277–284, 2013.

[19] F. Abad-Navarro, C. Martínez-Costa, and J. T. Fernández-Breis, "Ontology coverage analysis through language models," *Eng. Appl. Artif. Intell.*, vol. 162, Dec. 2025, Art. no. 112671, doi: 10.1016/j.engappai.2025.112671.

[20] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne, "Knowledge graphs: Opportunities and challenges," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13071–13102, Nov. 2023, doi: 10.1007/s10462-023-10465-9.

[21] J. Chen, H. Dong, J. Hastings, E. Jiménez-Ruiz, V. López, P. Monnin, C. Pesquita, P. Škoda, and V. Tamma, "Knowledge graphs for the life sciences: Recent developments, challenges and opportunities," *Trans. Graph Data Knowl.*, pp. 5–33, 2023, doi: 10.4230/tgdk.1.1.5.

[22] A. Santos, A. R. Colaço, A. B. Nielsen, L. Niu, M. Strauss, P. E. Geyer, F. Coscia, N. J. W. Albrechtsen, F. Mundt, L. J. Jensen, and M. Mann, "A knowledge graph to interpret clinical proteomics data," *Nature Biotechnol.*, vol. 40, no. 5, pp. 692–702, May 2022, doi: 10.1038/s41587-021-01145-6.

[23] P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Sci. Data*, vol. 10, no. 1, p. 67, Feb. 2023, doi: 10.1038/s41597-023-01960-3.

[24] B. J. Stear, T. Mohseni Ahooyi, J. A. Simmons, C. Kollar, L. Hartman, K. Beigel, A. Lahiri, S. Vasisht, T. J. Callahan, C. M. Nemarich, J. C. Silverstein, and D. M. Taylor, "Petagraph: A large-scale unifying knowledge graph framework for integrating biomolecular and biomedical data," *Sci. Data*, vol. 11, no. 1, p. 1338, Dec. 2024, doi: 10.1038/s41597-024-04070-w.

[25] C. Königs, M. Friedrichs, and T. Dietrich, "The heterogeneous pharmacological medical biochemical network PharMeBINet," *Sci. Data*, vol. 9, no. 1, p. 393, Jul. 2022, doi: 10.1038/s41597-022-01510-3.

[26] A. Bueckle, B. W. Herr, J. Hardi, E. M. Quardokus, M. A. Musen, and K. Börner, "Construction, deployment, and usage of the human reference atlas knowledge graph," *Sci. Data*, vol. 12, no. 1, p. 1100, Jul. 2025, doi: 10.1038/s41597-025-05183-6.

[27] J. H. Caufield et al., "KG-Hub—Building and exchanging biological knowledge graphs," *Bioinformatics*, vol. 39, no. 7, p. 418, Jul. 2023, doi: 10.1093/bioinformatics/btad418.

[28] E. Cavalleri, A. Cabri, M. Soto-Gomez, S. Bonfitto, P. Perlasca, J. Gliozzo, T. J. Callahan, J. Reese, P. N. Robinson, E. Casiraghi, G. Valentini, and M. Mesiti, "An ontology-based knowledge graph for representing interactions involving RNA molecules," *Sci. Data*, vol. 11, no. 1, p. 906, Aug. 2024, doi: 10.1038/s41597-024-03673-7.

[29] A. Altenhoff et al., "The SIB Swiss institute of bioinformatics semantic Web of data," *Nucleic Acids Res.*, vol. 52, no. D1, pp. D44–D51, Jan. 2024, doi: 10.1093/nar/gkad902.

[30] D. R. Unni et al., "Biolink model: A universal schema for knowledge graphs in clinical, biomedical, and translational science," *Clin. Transl. Sci.*, vol. 15, no. 8, pp. 1848–1855, Aug. 2022, doi: 10.1111/cts.13302.

[31] S. Lobentanzer et al., "Democratizing knowledge representation with BioCypher," *Nature Biotechnol.*, vol. 41, no. 8, pp. 1056–1059, Aug. 2023, doi: 10.1038/s41587-023-01848-y.

[32] T. J. Callahan et al., "An open source knowledge graph ecosystem for the life sciences," *Sci. Data*, vol. 11, no. 1, p. 363, Apr. 2024, doi: 10.1038/s41597-024-03171-w.

[33] Z. Sun, C. Wang, W. Hu, M. Chen, J. Dai, W. Zhang, and Y. Qu, "Knowledge graph alignment network with gated multi-hop neighborhood aggregation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 222–229, doi: 10.1609/aaai.v34i01.5354.

[34] B. D. Trisedya, J. Qi, and R. Zhang, "Entity alignment between knowledge graphs using attribute embeddings," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 297–304, doi: 10.1609/aaai.v33i01.3301297.

[35] Z. Sun, W. Hu, Q. Zhang, and Y. Qu, "Bootstrapping entity alignment with knowledge graph embedding," *IJCAI*, pp. 4396–4402, 2018, doi: 10.24963/ijcai.2018/611.

[36] Z. Wang, Q. Lv, X. Lan, and Y. Zhang, "Cross-lingual knowledge graph alignment via graph convolutional networks," *Assoc. Comput. Linguistics*, vol. 18, pp. 349–357, Jan. 2018, doi: 10.18653/v1/d18-1032.

[37] M. Nickel, L. Rosasco, and T. Poggio, "Holographic embeddings of knowledge graphs," in *Proc. AAAI Conf. Artif. Intell.*, 2016, vol. 30, no. 1, pp. 1955–1961.

[38] F. He, Z. Li, Y. Qiang, A. Liu, G. Liu, P. Zhao, L. Zhao, M. Zhang, and Z. Chen, "Unsupervised entity alignment using attribute triples and relation triples," in *Database Systems for Advanced Applications* (Lecture Notes in Computer Science), vol. 11446, G. Li, J. Yang, J. Gama, J. Natwichai, and Y. Tong, Eds., Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-18576-3_22.

[39] H. Zhu, R. Xie, Z. Liu, and M. Sun, "Iterative entity alignment via joint knowledge embeddings," *IJCAI*, vol. 17, pp. 4258–4264, 2017, doi: 10.24963/ijcai.2017/595.

[40] Z. Sun, W. Hu, and C. Li, "Cross-lingual entity alignment via joint attribute-preserving embedding," in *The Semantic Web–ISWC 2017* (Lecture Notes in Computer Science), vol. 10587, C. d'Amatoet et al., Eds., Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-68288-4_37.

[41] M. Chen, Y. Tian, M. Yang, and C. Zaniolo, "Multilingual knowledge graph embeddings for cross-lingual knowledge alignment," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1511–1517.

[42] B. Shi and T. Weninger, "ProjE: Embedding projection for knowledge graph completion," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 1236–1242, doi: 10.1609/aaai.v31i1.10677.

[43] Y. Wu, X. Liu, Y. Feng, Z. Wang, R. Yan, and D. Zhao, "Relation-aware entity alignment for heterogeneous knowledge graphs," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5278–5284, doi: 10.24963/ijcai.2019/733.

[44] L. Guo, Z. Sun, and W. Hu, "Learning to exploit long-term relational dependencies in knowledge graphs," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2505–2514. [Online]. Available: https://proceedings.mlr.press/v97/guo19c/guo19c.pdf

[45] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," in *Proc. Int. Conf. Learn. Represent.*, 2019. [Online]. Available: https://openreview.net/forum?id=HkgEQnRqYQ

[46] S. Pei, L. Yu, R. Hoehndorf, and X. Zhang, "Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference," in *Proc. World Wide Web Conf.*, May 2019, pp. 3130–3136, doi: 10.1145/3308558.3313646.

[47] S. M. Kazemi and D. Poole, "SimplE embedding for link prediction in knowledge graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4289–4300.

[48] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, "Knowledge graph embedding via dynamic mapping matrix," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 687–696, doi: 10.3115/v1/p15-1067.

[49] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proc. AAAI Conf. Artif. Intell.*, 2014, vol. 28, no. 1, pp. 1112–1119, doi: 10.1609/aaai.v28i1.8870.

[50] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proc. AAAI Conf. Artif. Intell.*, 2015, vol. 29, no. 1, pp. 2181–2187, doi: 10.1609/aaai.v29i1.9491.

[51] M. U. Akhtar, J. Liu, Z. Xie, X. Cui, X. Liu, and B. Huang, "Multilingual entity alignment by abductive knowledge reasoning on multiple knowledge graphs," *Eng. Appl. Artif. Intell.*, vol. 139, Jan. 2025, Art. no. 109660, doi: 10.1016/j.engappai.2024.109660.

[52] Z. Li, N. Ding, C. Liang, S. Cao, M. Zhai, R. Huang, Z. Zhang, and B. Hu, "A semi-supervised framework fusing multiple information for knowledge graph entity alignment," *Expert Syst. Appl.*, vol. 259, Jan. 2025, Art. no. 125282, doi: 10.1016/j.eswa.2024.125282.

[53] H. Li, C. Zhang, X. Chen, C. Wang, and Y. Yue, "A multi-modal entity alignment method based on image enhancement and credibility iteration," *Complex Intell. Syst.*, vol. 11, no. 11, pp. 1–19, Nov. 2025, doi: 10.1007/s40747-025-02077-3.

[54] X. Chen, T. Lu, and Z. Wang, "LLM-align: Utilizing large language models for entity alignment in knowledge graphs," 2024, *arXiv:2412.04690*.

[55] X. Jiang, Y. Shen, Z. Shi, C. Xu, W. Li, Z. Li, J. Guo, H. Shen, and Y. Wang, "Unlocking the power of large language models for entity alignment," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics*, 2024, pp. 7566–7583, doi: 10.18653/v1/2024.acl-long.408.

[56] X. Jiang, Y. Shen, Z. Shi, C. Xu, W. Li, H. Zihe, J. Guo, and Y. Wang, "MM-ChatAlign: A novel multimodal reasoning framework based on large language models for entity alignment," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2024, pp. 2637–2654, doi: 10.18653/v1/2024.findings-emnlp.148.

[57] V. Sriram, A. M. Conard, I. Rosenberg, D. Kim, T. S. Saponas, and A. K. Hall, "Addressing biomedical data challenges and opportunities to inform a large-scale data lifecycle for enhanced data sharing, interoperability, analysis, and collaboration across stakeholders," *Sci. Rep.*, vol. 15, no. 1, p. 6291, Feb. 2025, doi: 10.1038/s41598-025-90453-x.

[58] C. A. A. Chahal, F. Alahdab, B. Asatryan, D. Addison, N. Aung, M. K. Chung, S. Denaxas, J. Dunn, J. L. Hall, N. Pamir, D. J. Slotwiner, J. D. Vargas, and A. A. Armoundas, "Data interoperability and harmonization in cardiovascular genomic and precision medicine," *Circulation, Genomic Precis. Med.*, vol. 18, no. 3, Jun. 2025, Art. no. 004624, doi: 10.1161/circgen.124.004624.

[59] B. Cheng, J. Zhu, and P. De Meo, "Dual context representation learning framework for entity alignment," *Big Data Mining Anal.*, vol. 8, no. 2, pp. 346–363, Apr. 2025, doi: 10.26599/bdma.2024.9020082.

[60] R. Zhang, B. D. Trisedya, M. Li, Y. Jiang, and J. Qi, "A benchmark and comprehensive survey on knowledge graph entity alignment via representation learning," *VLDB J.*, vol. 31, no. 5, pp. 1143–1168, Sep. 2022, doi: 10.1007/s00778-022-00747-z.

[61] M. Ali, M. Berrendorf, C. T. Hoyt, L. Vermue, M. Galkin, S. Sharifzadeh, A. Fischer, V. Tresp, and J. Lehmann, "Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8825–8845, Dec. 2022, doi: 10.1109/TPAMI.2021.3124805.

[62] S. Garg and D. Roy, "A birds eye view on knowledge graph embeddings, software libraries, applications and challenges," 2022, *arXiv:2205.09088*.

[63] S. Hertling and H. Paulheim, "The knowledge graph track at oaei: Gold standards, baselines, and the golden hammer bias," in *Proc. Eur. Semantic Web Conf.*, Jun. 2020, pp. 343–359, doi: 10.1007/978-3-030-49461-2_20.

[64] J. Mulero Hernández and J. T. Fernández-Breis, "Analysis of the landscape of human enhancer sequences in biological databases," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 2728–2744, May 2022, doi: 10.1016/j.csbj.2022.05.045.

[65] L. A. Pennacchio, W. Bickmore, A. Dean, M. A. Nobrega, and G. Bejerano, "Enhancers: Five essential questions," *Nature Rev. Genet.*, vol. 14, no. 4, pp. 288–295, Apr. 2013, doi: 10.1038/nrg3458.

[66] A. Panigrahi and B. W. O'Malley, "Mechanisms of enhancer action: The known and the unknown," *Genome Biol.*, vol. 22, no. 1, p. 108, Dec. 2021, doi: 10.1186/s13059-021-02322-1.

[67] Y. He, W. Long, and Q. Liu, "Targeting super-enhancers as a therapeutic strategy for cancer treatment," *Frontiers Pharmacol.*, vol. 10, p. 361, Apr. 2019, doi: 10.3389/fphar.2019.00361.

[68] Y. Jia, W.-J. Chng, and J. Zhou, "Super-enhancers: Critical roles and therapeutic targets in hematologic malignancies," *J. Hematology Oncol.*, vol. 12, no. 1, pp. 1–17, Dec. 2019, doi: 10.1186/s13045-019-0757-y.

[69] A. Claringbould and J. B. Zaugg, "Enhancers in disease: Molecular basis and emerging treatment strategies," *Trends Mol. Med.*, vol. 27, no. 11, pp. 1060–1073, Nov. 2021, doi: 10.1016/j.molmed.2021.07.012.

[70] Q. Liu, L. Guo, Z. Lou, X. Xiang, and J. Shao, "Super-enhancers and novel therapeutic targets in colorectal cancer," *Cell Death Disease*, vol. 13, no. 3, p. 228, Mar. 2022, doi: 10.1038/s41419-022-04673-4.

[71] M. Bacabac and W. Xu, "Oncogenic super-enhancers in cancer: Mechanisms and therapeutic targets," *Cancer Metastasis Rev.*, vol. 42, no. 2, pp. 471–480, Jun. 2023, doi: 10.1007/s10555-023-10103-4.

[72] X.-P. Li, J. Qu, X.-Q. Teng, H.-H. Zhuang, Y.-H. Dai, Z. Yang, and Q. Qu, "The emerging role of super-enhancers as therapeutic targets in the digestive system tumors," *Int. J. Biol. Sci.*, vol. 19, no. 4, pp. 1036–1048, 2023, doi: 10.7150/ijbs.78535.

[73] Z. Yao, P. Song, and W. Jiao, "Pathogenic role of super-enhancers as potential therapeutic targets in lung cancer," *Frontiers Pharmacol.*, vol. 15, Apr. 2024, Art. no. 1383580, doi: 10.3389/fphar.2024.1383580.

[74] S. Liu, W. Dai, B. Jin, F. Jiang, H. Huang, W. Hou, J. Lan, Y. Jin, W. Peng, and J. Pan, "Effects of super-enhancers in cancer metastasis: Mechanisms and therapeutic targets," *Mol. Cancer*, vol. 23, no. 1, p. 122, Jun. 2024, doi: 10.1186/s12943-024-02033-8.

[75] E. Smith and A. Shilatifard, "Enhancer biology and enhanceropathies," *Nature Struct. Mol. Biol.*, vol. 21, no. 3, pp. 210–219, Mar. 2014, doi: 10.1038/nsmb.2784.

[76] X. Bai, S. Shi, B. Ai, Y. Jiang, Y. Liu, X. Han, M. Xu, Q. Pan, F. Wang, Q. Wang, J. Zhang, X. Li, C. Feng, Y. Li, Y. Wang, Y. Song, K. Feng, and C. Li, "ENdb: A manually curated database of experimentally supported enhancers for human and mouse," *Nucleic Acids Res.*, pp. D51–D57, Oct. 2019, doi: 10.1093/nar/gkz973.

[77] W. Zeng, X. Min, and R. Jiang, "EnDisease: A manually curated database for enhancer-disease associations," *Database*, vol. 2019, p. 020, Jan. 2019, doi: 10.1093/database/baz020.

[78] G. Zhang, J. Shi, S. Zhu, Y. Lan, L. Xu, H. Yuan, G. Liao, X. Liu, Y. Zhang, Y. Xiao, and X. Li, "DiseaseEnhancer: A resource of human disease-associated enhancer catalog," *Nucleic Acids Res.*, vol. 46, no. 1, pp. 78–84, Jan. 2018, doi: 10.1093/nar/gkx920.

[79] A. Visel, S. Minovitsky, I. Dubchak, and L. A. Pennacchio, "VISTA enhancer browser—A database of tissue-specific human enhancers," *Nucleic Acids Res.*, vol. 35, pp. D88–D92, Jan. 2007, doi: 10.1093/nar/gkl822.

[80] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Res.*, vol. 35, pp. D61–D65, Jan. 2007, doi: 10.1093/nar/gkl842.

[81] Z. Sun, Q. Zhang, W. Hu, C. Wang, M. Chen, F. Akrami, and C. Li, "A benchmarking study of embedding-based entity alignment for knowledge graphs," *Proc. VLDB Endowment*, vol. 13, no. 12, pp. 2326–2340, Aug. 2020, doi: 10.14778/3407790.3407828.

[82] G. Almagro-Hernández, J. Mulero-Hernández, P. Deshmukh, J. A. Bernabé-Díaz, J. L. Sánchez-Fernández, P. Espinoza-Arias, J. Mueller, and J. T. Fernández-Breis, "Evaluation of alignment methods to support the assessment of similarity between e-commerce knowledge graphs," *Knowl.-Based Syst.*, vol. 315, Apr. 2025, Art. no. 113283, doi: 10.1016/j.knosys.2025.113283.
[83] J. A. Bernabé-Díaz, M. D. C. Legaz-García, J. M. García, and J. T. Fernández-Breis, "Efficient, semantics-rich transformation and integration of large datasets," *Expert Syst. Appl.*, vol. 133, pp. 198–214, Nov. 2019, doi: 10.1016/j.eswa.2019.05.010.
[84] J. Arenas-Guerrero, D. Chaves-Fraga, J. Toledo, M. S. Pérez, and O. Corcho, "Morph-KGC: Scalable knowledge graph materialization with mapping partitions," *Semantic Web*, vol. 15, no. 1, pp. 1–20, Jan. 2024, doi: 10.3233/sw-223135.
[85] E. Antezana, W. Blondé, M. Egaña, A. Rutherford, R. Stevens, B. De Baets, V. Mironov, and M. Kuiper, "BioGateway: A semantic systems biology tool for the life sciences," *BMC Bioinf.*, vol. 10, no. S10, pp. 1–15, Oct. 2009, doi: 10.1186/1471-2105-10-s10-s11.
[86] X. Zhao, W. Zeng, J. Tang, W. Wang, and F. M. Suchanek, "An experimental study of state-of-the-art entity alignment approaches," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 6, pp. 2610–2625, Jun. 2022, doi: 10.1109/TKDE.2020.3018741.
[87] K. Zeng, C. Li, L. Hou, J. Li, and L. Feng, "A comprehensive survey of entity alignment for knowledge graphs," *AI Open*, vol. 2, pp. 1–13, 2021, doi: 10.1016/j.aiopen.2021.02.002.
[88] A. J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E. L. Willighagen, C. T. Evelo, N. Blomberg, G. Ecker, C. Goble, and B. Mons, "Open PHACTS: Semantic interoperability for drug discovery," *Drug Discovery Today*, vol. 17, nos. 21–22, pp. 1188–1198, Nov. 2012, doi: 10.1016/j.drudis.2012.05.016.
[89] J. Moreira, "The role of interoperability for digital twins," in *Proc. Int. Conf. Enterprise Design*, 2024, pp. 139–157, doi: 10.1007/978-3-031-54712-6_9.
[90] I. David, G. Shao, C. Gomes, D. Tilbury, and B. Zarkout, "Interoperability of digital twins: Challenges, success factors, and future research directions," in *Proc. Int. Symp. Leveraging Appl. Formal Methods*, 2024, pp. 27–46, doi: 10.1007/978-3-031-75390-9_3.
[91] E. Jiménez-Ruiz and B. C. Grau, "LogMap: logic-based and scalable ontology matching," in *Proc. Int. Semantic Web Conf.*, 2011, pp. 273–288, doi: 10.1007/978-3-642-25073-6_18.
[92] D. Faria, M. C. Silva, P. Cotovio, L. Ferraz, L. Balbi, and C. Pesquita, "Results in OAEI 2024 for matcha," in *Proc. CEUR Workshop*, vol. 3897, 2024, pp. 110–117.

JUAN MULERO-HERNÁNDEZ received the degree in biotechnology and the master's degree in bioinformatics from the University of Murcia, Spain, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree in computer science in the field of ontologies, semantic web and knowledge-based systems. His research interests address gene regulation and semantic web solutions for biological knowledge modeling, and the integration and interoperability of heterogeneous sources. He was awarded as the Best Academic Record.

GINÉS ALMAGRO-HERNÁNDEZ received the degree in biotechnology and the Ph.D. degree in computer science from the University of Murcia (UMU), Spain, in 2013 and 2020, respectively. He is currently a Postdoctoral Researcher with the Faculty of Computer Science, UMU. He is also a member of the IMIB-Arrixaca Bio-Health Research Institute. His current research interests include the application of semantic technologies for the development of learning health systems and the development of quality assurance methods for ontologies and terminologies.

JESUALDO TOMÁS FERNÁNDEZ-BREIS (Senior Member, IEEE) received the bachelor's degree in computer engineering and the Ph.D. degree in computer science from the University of Murcia (UMU), Spain, in 1999 and 2003, respectively. He is currently a Full Professor with the Faculty of Computer Science, UMU. Since 2004, he has been leading research projects related to semantic web technologies. He is the Co-founder of the Spin-Off Longseq Applications. He is also a member of the IMIB-Arrixaca Bio-Health Research Institute. His current research interests include the application of semantic technologies for the development of learning health systems and the development of quality assurance methods for ontologies and terminologies.