

Alignment of Ontologies Using Directed Graphs, Neural Network and Genetic Algorithm in the Web of Data

Mamaï Kulemaï Aymar¹, Warda Lazare², Samdalle Amaria³, Kaladzavi Guidedi⁴, Kolyang⁵

¹ University of Maroua, National Advanced School of Engineering, Department of Computer Sciences and Telecommunications

^{2,3} University of Maroua, Faculty of Sciences, Department of Mathematics and Computer Sciences, Cameroon

⁴ University of Maroua, National Advanced School of Engineering, Department of Computer Sciences and Telecommunications

⁵ University of Maroua, Higher Teacher's Training College, Cameroon

^{1,2,3,4,5} Laboratoire de Recherche Informatique (LaRI), Cameroon

DOI: <https://doi.org/10.51584/IJRIAS.2026.11090035>

Received: 01 July 2026; Accepted: 06 July 2026; Published: 06 October 2026

ABSTRACT

Semantic interoperability remains a central challenge for the Semantic Web, as the growth of structured data increases the need to reconcile heterogeneous ontologies across domains such as education, health, defence, agriculture and biomedicine. This paper proposes a hybrid ontology-alignment architecture that combines directed-graph representations to capture semantic and taxonomic structure, graph neural networks (GNNs) enriched with BioBERT embeddings to learn latent semantic correspondences, and a genetic algorithm to optimise the final selection of matches under structural-consistency constraints. The approach is illustrated on the biomedical domain, where reference ontologies such as SNOMED CT, ICD-10 and LOINC are large, complex and structurally rich, and where aligning hospital data with such reference terminologies supports interoperability and diagnostic assistance. Simulated experiments inspired by the OAEI Anatomy and LargeBio benchmarks indicate that combining the three layers improves the F-measure relative to each layer taken in isolation, primarily by reducing false positives through structural filtering. A component-wise ablation study isolates the contribution of each layer, and the simulated results are positioned, for indicative purposes only, against published OAEI systems. A complementary, non-simulated evaluation on the actual small LargeBio ontology fragments (FMA–NCI, FMA–SNOMED, SNOMED–NCI), scored locally with a MELT-style evaluator, shows the hybrid pipeline reaching the best precision of all matcher variants on every pair but not the best F-measure, since it trades recall away more than it recovers, a real-data finding that qualifies the simulated ablation and points to blocking-width and fitness-weight calibration as concrete next steps. The paper also identifies the remaining methodological steps: an official OAEI submission under SEALS/MELT, an equivalent run on Anatomy, statistical validation across multiple seeds, and a dedicated scalability study required to move from illustrative simulation to a fully validated system.

Keywords: ontology alignment, machine learning, ontologies, neural networks, genetic algorithms, data graphs, GNN, biomedical interoperability, BioBERT, patient monitoring.

INTRODUCTION

The web of data, with its considerable volume of structured and semi-structured content, offers major opportunities for data integration and analysis, but the heterogeneity of data sources remains a significant obstacle. Ontology alignment mapping the concepts and relations of different ontologies onto one another is a key step in overcoming this obstacle. A substantial body of work has addressed ontology matching [12], and for

almost a decade many of its results have benefited from artificial intelligence, which has significantly improved matching quality. One recurring concern, however, is that alignment approaches have often been evaluated on ontologies in a scattered way across separate OAEI campaigns (2017–2021), and continue to face difficulties in result accuracy, scalability to large ontologies, the ability to learn structural patterns, and robustness when ontologies are structurally complex [13].

Early alignment tools [14] and the ontology community initially relied on a straightforward comparison of the class hierarchies of two ontologies, together with a small set of heuristic algorithms to produce and express alignments [15]. As the number of ontology-authoring tools multiplied, however, so did the diversity of formalisms used to express them, giving rise to several forms of heterogeneity syntactic [16], terminological [17], conceptual [18] and semiotic that any alignment method now has to address.

In healthcare specifically, interoperability continues to be hampered by the semantic heterogeneity of hospital information systems, and aligning large biomedical ontologies is essential to ensure consistent longitudinal follow-up of patients. In today's digital-health landscape, the proliferation of clinical data sources electronic health records, connected devices, specialised registries has turned semantic heterogeneity into a major practical challenge. As Euzenat and Shvaiko (2007) [12] point out, ontology alignment is the critical process by which distinct knowledge systems are mapped onto one another to ensure interoperability. In biomedicine, this is all the more pressing because longitudinal patient follow-up requires a precise consolidation of diagnoses and treatments that are often recorded under different terminologies, such as SNOMED CT, ICD-10 or LOINC.

This paper presents a hybrid approach built on three pillars. First, ontologies are modelled as directed graphs, which preserves their taxonomic hierarchy. Second, graph neural networks (GNNs), enriched with BioBERT embeddings, capture complex semantic similarities together with clinical context. Third, a genetic algorithm orchestrates the overall selection of matches by optimising a fitness function that weighs semantic precision against structural consistency. The objective is to show that this three-way synergy outperforms each layer taken in isolation, in particular in terms of F-measure on reference biomedical datasets, and to discuss how it could improve the quality of patient follow-up by reducing ambiguity when reconciling clinical data.

Contributions

The individual components used here directed-graph modelling, GNN-based embeddings and genetic optimisation are each well established in the ontology-matching literature (Section 2); the contribution of this paper is not any one of them in isolation but their integration and evaluation as a single pipeline. Concretely, the paper contributes:

- ✓ A pipeline in which genetic optimisation operates directly on GNN embeddings enriched with BioBERT semantics and constrained by directed-graph structure, rather than on raw lexical similarity scores;
- ✓ A fitness function and a repair mechanism that jointly encode semantic similarity, taxonomic (is-a) and mereological (part-of) consistency, and conflict penalties within a single combinatorial search (Section 3);
- ✓ A component-wise ablation study that isolates the contribution of each layer, together with an explicit, consistently maintained separation between simulated, illustrative results and officially benchmarked OAEI results (Section 5);
- ✓ A discussion of how the pipeline could be positioned as a decision-support component for biomedical patient follow-up, together with the concrete steps an official OAEI submission, statistical validation and a scalability study required before any clinical relevance can be claimed (Sections 5 and 6).

Related Work / State of the Art

Similarity, Dissimilarity and Distance

Aligning ontologies amounts to searching for correspondences between their entities, which in turn requires analysing entity hierarchies both local and global structure using similarity-computation techniques of

a terminological, structural, extensional or semantic nature. Similarity is a broad and widely used notion whose precise meaning depends on the application domain or on the form of knowledge representation being used.

In the present context, similarity is understood as semantic similarity, sometimes called semantic proximity: it is estimated by associating documents, terms or entities with a metric based on the closeness of their meaning or content. Formalising a domain typically requires semantic resources ontologies, i.e. knowledge bases since they alone make explicit the relations (hypernymy, antonymy, etc.) that hold between concepts. Its value usually lies between 0 and 1, which allows for a probabilistic interpretation.

Definition 1 Similarity

A similarity $S : O \times O \rightarrow \mathbb{R}$ is a function that maps a pair of entities to a real number expressing their similarity, and satisfies:

- $\forall a, b \in O, S(a, b) \geq 0$ (positivity)
- $\forall a, b, c \in O, S(a, a) \geq S(b, c)$ and $S(a, a) = S(a, b) \Rightarrow a = b$ (self-similarity / maximality)
- $\forall a, b \in O, S(a, b) = S(b, a)$ (symmetry)
- $\forall a, b \in O, S(a, b) \leq \infty$ (finiteness)

Transitivity-like behaviour ($S(a,b) = S(b,c) \Rightarrow S(a,b) = S(a,c)$) is a desirable additional property for some instances of similarity, but it is not required by the definition and does not hold for similarity measures in general; it is therefore treated here as an optional property to be checked for the specific measure used, not as an axiom.

Example (Jaccard similarity): $S_{\text{Jaccard}} : S \times S \rightarrow [0, 1]$ such that $S_{\text{Jaccard}}(s, t) = |s \cap t| / |s \cup t|$

Definition 2 Dissimilarity

Dissimilarity is sometimes used in place of similarity. It is defined analogously to Definition 1, except that it is not required to be transitive, and it satisfies positivity, minimality ($DS(a,a) \leq DS(b,c)$ and $DS(a,a) = 0$), symmetry and finiteness. When S is a normalised similarity (range $[0,1]$), a complementary dissimilarity can be constructed as $DS(a,b) = 1 - S(a,b)$; this construction not a general law relating any S and any DS is what justifies writing $S + DS = 1$ for that specific pair.

Definition 3 Distance

Distance is used as a measure of dissimilarity between two entities: a high similarity corresponds to a small distance, and vice versa. Building on classical metric-space definitions, distance is a function satisfying definiteness ($D(a,b) = 0 \Leftrightarrow a = b$) and the triangle inequality ($D(a,b) + D(b,c) \geq D(a,c)$).

Well-known instances used in this line of work include the Levenshtein distance (the minimum number of edit operations needed to transform one string into another), the Damerau–Levenshtein distance (which additionally accounts for transpositions), and the Jaro distance:

$$DS_{\text{Jaro}} : S \times S \rightarrow [0,1], S_{\text{Jaro}}(s,t) = 1 - (1/3) \cdot (m/|s| + m/|t| + (m - t)/m) \quad [19]$$

(m: number of matching characters; t: number of transpositions.)

Traditional Alignment Systems

Table 1 lists a representative sample of traditional alignment tools proposed since the early 2000s, positioned by year, degree of automation, and type (basic or improved).

Year	Status	Automatic	Semi-automatic
2001	Base	—	—
2001	Improved	—	Anchor-Prompt (improved version of SMART, Noy & Musen)
2002	Base	—	—
2002	Improved	GLUE (improvement of LSD, Doan et al.); COMA (combination of several matchers, Do & Rahm)	—
2003	Base	IF-Map (Kalfoglou et al.); S-Match (Giunchiglia et al.)	—
2004	Base	NOM (Ehrig et al.); OLA (Euzenat et al., 2004)	—
2004	Improved	QOM (improvement of NOM, Ehrig et al.)	—
2005	Base	ASCO1 (Bach)	—
2005	Improved	ASCO2 (improvement of ASCO1, Bach)	—
2007	Base	AROMA (David et al.)	FALCON-OA (Hu et al.); ASMOV (Jean-Mary et al.); Taxomap (Hamdi et al.)
2007	Improved	EDOLA (improvement of OLA, Zghal et al.); SODA; ASCO3 (Bach Than et al.); OLA2	COMA++ (improvement of COMA, Do & Rahm)
2008	Improved	OLA2 (improvement of OLA, Djoufak et al.)	—

Table 1. Traditional ontology-alignment systems, positioned by year, automation level and type.

Conventional alignment methods rely heavily on simple lexical or structural similarity measures. Medical language, however, is characterised by complex synonymy and polysemy that classical algorithms struggle to capture (Bodenreider, 2004). Pre-trained language models such as BioBERT (Lee et al., 2020) have brought significant gains in semantic understanding, but their use is often limited to comparing isolated terms and neglects the topological richness of medical hierarchies.

Hybrid Approaches

The availability of powerful machine-learning algorithms for data optimisation and classification has led to a new generation of hybrid alignment approaches, which combine heuristic similarity measures with learning-based methods to manage the syntactic, terminological and structural heterogeneity of ontologies [43]. Several complementary techniques are typically combined: matcher combination (terminological, structural and semantic matchers used together); structural approaches such as SODA, which analyse graph structure to identify similarities; partitioning methods for large ontologies (e.g. structural-proximity bi-partitioning); linked-data mining, which exploits ontologies already linked on the web of data; and metaheuristics, which use optimisation algorithms to search for the best match. A recent state of the art of this family [43] reviews both traditional matchers (ASCO3, QOM, Anchor-PROMPT, SODA, TaxoMap, OLA, Falcon-AO) and machine-learning matchers (LogMap, MLMA, DeepOnto, SAMBO), and concludes that combining a directed-graph representation, a graph neural network and a genetic optimiser is the most promising route; the architecture presented in Section 3 is a full instantiation and evaluation of that proposal.

Machine Learning

Reinforcement learning is one relevant paradigm: an agent learns to act within an environment so as to maximise its expected reward, relying on an agent that selects actions, an environment that defines the context, the actions available, and a policy the agent's strategy for choosing actions. The agent explores its environment, receives positive or negative rewards, and adjusts its policy to maximise future reward.

Table 2 summarises how hybrid systems have combined classical similarity measures with machine-learning classifiers over the past two decades.

Year	Author(s)	System	Learning method	Observation
2004	Doan et al.	GLUE	Naïve Bayes, SVM	Classifiers used to align concepts
2005	Noy & Musen	PROMPT	Naïve Bayes, Decision Trees	Classifiers used to align concepts and relations
2007	Ehrig et al.	QOM	SVM, k-NN	Classifiers used to align concepts and instances
2009	Jean-Mary et al.	ASMOV	Naïve Bayes, Decision Trees	Classifiers for concept/relation alignment (biomedical)
2010	Suchanek et al.	PARIS	Probabilistic inference	Probability-based alignment of concepts and instances
2012	Jiménez-Ruiz et al.	LogMap	Naïve Bayes, Decision Trees	Detects inconsistencies and improves accuracy
2017	Zhang et al.	—	Deep learning (CNN)	Neural-network-based concept and instance alignment
2018	Sun et al.	OpenEA	Deep learning (attention-based)	Neural-network-based concept and instance alignment
2019	Hertling et al.	AML	Deep learning (graph-based)	Neural-network-based concept and instance alignment

Table 2. Hybrid systems combining classical similarity measures with machine learning.

These systems illustrate the individual limitations that motivate a further hybridisation step: to overcome them, this paper models ontologies as directed graphs preserving the specialisation and composition relations that are essential to medical consistency and combines two complementary paradigms. The first is the feature-learning power of graph neural networks (GNNs), which condense semantics and structure into a shared latent space (Hamilton, 2020). The second is the robustness of genetic algorithms (GA) for global optimisation: alignment is treated as a combinatorial problem in which the GA explores the space of candidate mappings, guided by a fitness function that weighs semantic precision against logical consistency.

Recent Transformer-Based, Graph-Based and Neuro-Symbolic Approaches

Beyond the classical and early hybrid systems of Tables 1–2, three further families of methods are directly relevant to the design choices made in this paper. First, domain-adapted transformer encoders have progressively replaced general-purpose word embeddings for terminological similarity: SciBERT [40] adapts BERT to scientific text through a domain-specific vocabulary and corpus, and LinkBERT/BioLinkBERT [41] further pretrains on documents connected by hyperlinks or citation links, which lets the model exploit cross-document relational structure a biomedical variant of the same paper reports state-of-the-art results on PubMed-based benchmarks such as BLURB, BioASQ and MedQA-USMLE. These results support the choice, in Section 3, of

a pre-trained biomedical language model (BioBERT) as the source of lexical-semantic node features, while suggesting that citation- or link-aware pretraining is a natural extension for future work.

Second, graph-based and knowledge-graph-embedding approaches (e.g. translation-based or attention-based entity-alignment models, as used in OpenEA and AML, Table 2) learn structural regularities directly from the relational topology of an ontology, which is the same intuition this paper's GNN layer builds on, but typically without a mechanism to enforce hard structural constraints on the final mapping a role played here by the genetic-algorithm layer and its repair operator (Section 3.1.4).

Third, neuro-symbolic approaches combine learned similarity scores with a symbolic consistency checker in the spirit of LogMap's logic-based inconsistency detection (Table 2) to reject mappings that would make the aligned ontology logically unsatisfiable. The structural-consistency term of the fitness function proposed in Section 3.1.4, together with the RepairConflicts operator, plays an analogous role here, although it enforces structural (taxonomic and mereological) rather than full logical consistency; combining it with an explicit description-logic reasoner is identified as a direction for future work rather than a component already implemented.

Fourth, probabilistic sequential models offer an alternative to both lexical and embedding-based similarity. Djimramadji et al. [42] transform each ontology into a Hidden Markov Model, extracting RDF triples through SPARQL queries and mapping concepts onto hidden states and relations onto observation symbols; the rows of the resulting state-transition matrix are then used as concept vectors and compared by cosine similarity. Reported on OAEI repositories, the method reaches a precision of 1.00 with a recall of 0.64 on the Cmt-conference pair, and a precision of 0.77 with a recall of 0.91 on the Ce-crack pair, and its authors note that alignment quality grows with the relational richness of the ontologies while degrading on sparsely connected ones. This is the same observation that motivates the structural layer adopted here, but the HMM formulation encodes structure in a first-order transition matrix only, whereas the GNN layer of Section 3 aggregates information over multi-hop neighbourhoods and the genetic layer enforces global consistency over the whole mapping rather than concept by concept.

Proposed Method

Table 3 summarises the notation used throughout this section.

Symbol	Meaning
$O = (C, R, I)$	An ontology: concepts C , relations R , instances I
$G = (V, E, \ell)$	Directed graph modelling O ; V = concepts/instances, $E \subseteq V \times V$, $\ell: E \rightarrow \text{labels}$
A_1, A_2	Adjacency matrices of the source/target graphs G_1, G_2
$S_{\text{struct}}(x, y)$	Structural similarity between node $x \in G_1$ and $y \in G_2$ (similarity flooding)
$h_v^{(l)}$	Embedding of node v at GNN layer l ($h_v^{(0)}$: initial features)
$S_{\text{sem}}(v_1, v_2)$	Cosine similarity between GNN embeddings of v_1 and v_2
C_θ	Candidate set $\{(v_1, v_2) : S_{\text{struct}}(v_1, v_2) \geq \theta \text{ or } S_{\text{sem}}(v_1, v_2) \geq \theta\}$
X	An individual (candidate alignment), $X \subseteq C_\theta$
$f(X)$	Fitness of individual X
$\lambda_1, \lambda_2, \lambda_3$	Weights of the semantic, structural-consistency and conflict-penalty terms
M^*	Final optimal alignment returned by the pipeline

Table 3. Symbols used in the proposed method.

Hybrid Ontology Alignment Architecture

Pipeline: directed graph $\rightarrow$ graph neural network (GNN) $\rightarrow$ genetic algorithm.

Overview

The architecture combines three complementary layers in a loop: the graph layer provides an initial structural similarity, the GNN layer learns a latent semantic similarity, and the genetic-algorithm layer optimises the final match by combining both signals, with iterative feedback until convergence.

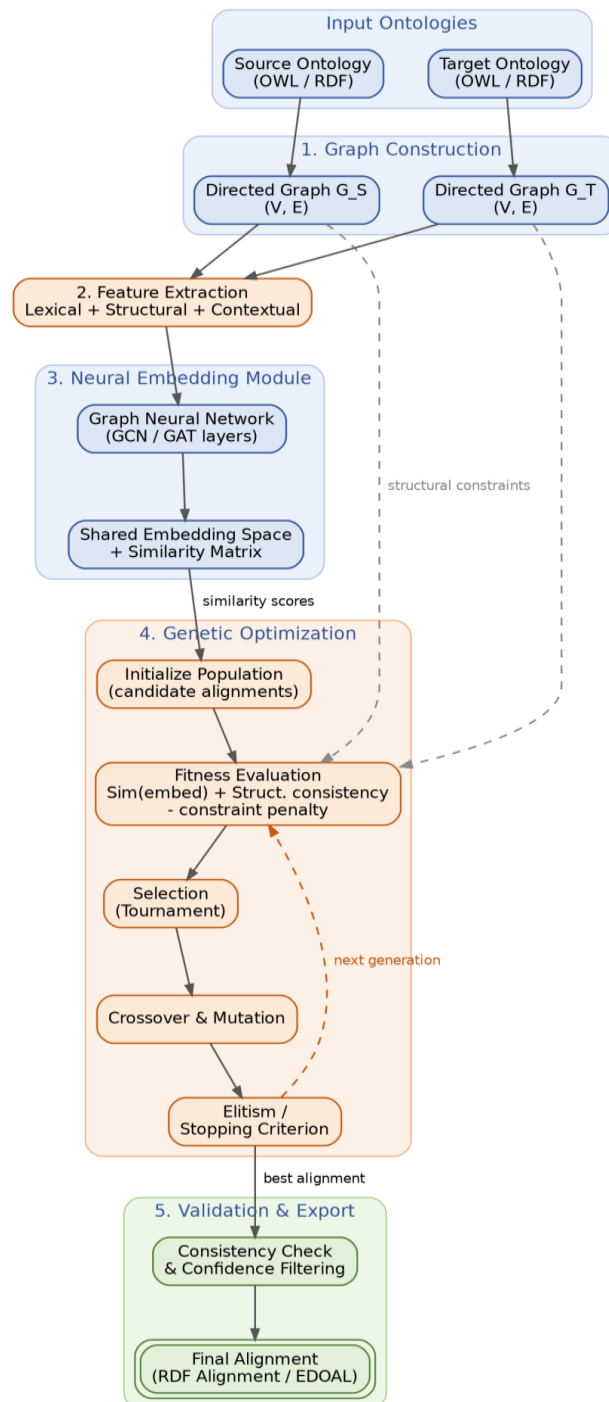


Figure 1. Overview of the hybrid architecture (directed graphs $\rightarrow$ GNN embeddings $\rightarrow$ genetic optimisation $\rightarrow$ validated alignment).

Legend:

- Blue blocks graph construction and neural-embedding stages (structural and semantic representation).
- Orange blocks genetic-algorithm loop (population, fitness evaluation, selection, crossover/mutation, elitism).

- Green blocks validation and export of the final alignment.
- Dashed grey arrows structural constraints propagated from the directed graphs into the fitness function.
- Dashed orange arrow evolutionary feedback loop across generations until the stopping criterion is met.

The architecture combines three complementary layers in a loop: the graph layer provides an initial structural similarity, the GNN layer learns a latent semantic similarity, and the genetic-algorithm layer optimises the final match by combining both signals, with iterative feedback until convergence. Blue stages construct the graph representation and neural embeddings; orange stages run the genetic-algorithm loop (population, fitness evaluation, selection, crossover/mutation, elitism); green stages validate and export the final alignment; dashed grey arrows propagate structural constraints from the directed graphs into the fitness function; the dashed orange arrow is the evolutionary feedback loop across generations until the stopping criterion is met.

Modelling Phase

Unlike flat representations, each ontology is modelled as a directed graph. The pipeline starts from two raw ontologies (OWL or RDF files), which are parsed to build two distinct directed graphs whose nodes represent concepts and properties and whose edges represent hierarchical (is-a) and other semantic relations. Preserving the orientation of the edges is essential, since it captures the direction and therefore the meaning of each relation, which in turn is what allows the rest of the pipeline to preserve the subsumption hierarchy and avoid false positives.

An ontology $O = (C, R, I)$ is modelled as a labelled directed graph $G = (V, E, \ell)$, where V is the set of concepts and instances, $E \subseteq V \times V$ is the set of relations (is-a, part-of, ...), and $\ell : E \rightarrow R$ is the edge-labelling function.

Definition 4 Candidate set and admissible alignment

Given a similarity threshold θ , the candidate set is $C\theta = \{(v_1, v_2) \in V_1 \times V_2 : S_{\text{struct}}(v_1, v_2) \geq \theta \text{ or } S_{\text{sem}}(v_1, v_2) \geq \theta\}$. An admissible alignment is any $M \subseteq C\theta$ that satisfies the cardinality constraint adopted for the target application (Section 3.1.4) and is free of the structural conflicts defined by StructuralConsistency. The optimal alignment $M^* = X^*$ is the admissible alignment that maximises the fitness function f , obtained by the genetic algorithm described in Section 3.1.4 not simply the set of all pairs above θ .

Similarity Flooding

Iterative update of the similarity between $x \in G_1$ and $y \in G_2$, restricted to neighbouring pairs $(x', x) \in E_1$, $(y', y) \in E_2$:

$$\text{sim}(k+1)(x, y) = \alpha \sum_{(x', x) \in E_1, (y', y) \in E_2} w(x', x)w(y', y)\text{sim}(k)(x', y') + (1 - \alpha)\text{simloc}(x, y)$$

In matrix form:

$$S(k+1) = \alpha(A_1 \otimes A_2)S(k) + (1 - \alpha)S(0)$$

A_1, A_2 : adjacency matrices of graphs G_1 and G_2 ; $\otimes$: Kronecker's product.

Representation-Learning Phase (GNN)

In this phase, a neural network is trained to learn correspondences between ontology concepts from their graphical representation and their semantic embeddings. Graph neural networks (GNNs) transform every node into a numerical vector (embedding): each node aggregates information from its immediate neighbours so that the resulting representation captures not only the entity's own label but also its structural context. Two entities are then considered semantically close if their latent vectors are close under cosine similarity. Directed graphs

are given as input to a graph-based encoder Graph SAGE or a GCN which learns to represent each node numerically from its textual label and from its neighbourhood in the graph.

BioBERT feature initialisation

The role of BioBERT in the pipeline is the following: for each node v , the concept label and, when available, its textual definition are encoded with BioBERT (frozen weights used as a feature extractor, not fine-tuned in the present simulated study, to limit computational cost) into a 768-dimensional vector, taken from the pooled [CLS] representation of the final layer. This vector is projected by a trainable linear layer to 128 dimensions and concatenated with a small set of structural features (node out-degree, in-degree, depth in the taxonomy and node type) to form the initial node feature $h_v^{(0)}$ fed to the GNN encoder. Figure 2 illustrates this initialisation pipeline schematically.

Schematic: BioBERT-based feature initialisation feeding the GNN encoder

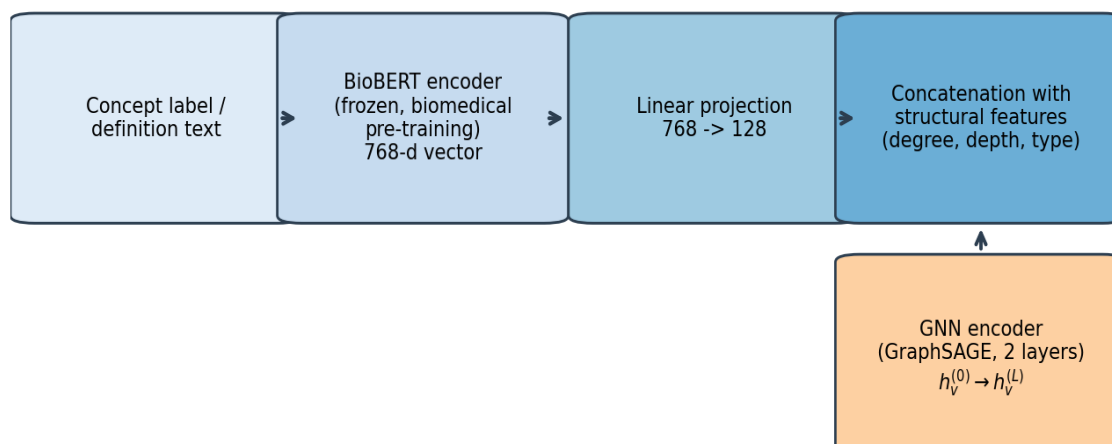


Figure 2. Schematic of the BioBERT-based feature initialisation feeding the GNN encoder (structural design of the pipeline, not a trained/measured configuration).

Message passing (GNN)

Updating node v from layer l to layer $l+1$, with neighbourhood $N(v)$ and relation-specific weights $W_{r(l)}$

$$h_{v(l+1)} = \sigma \left(W(l)h_{v(l)} + \sum_{u \in N(v)} W_{r(l)} h_{u(l)} \right)$$

Similarity score (common latent space)

$$\text{sim}(v_1, v_2) = \frac{h_{v_1} \cdot h_{v_2}}{\|h_{v_1}\| \|h_{v_2}\|}$$

Contrastive loss function

For positive pairs M^+ and negative pairs M^- , with margin γ :

$$L = \sum_{(v_1, v_2) \in M^+} \max(0, \gamma - \text{sim}(v_1, v_2) + \text{sim}(v_1, v'_2))$$

In the simulated experiments, $\gamma = 0.5$; positive pairs M^+ are drawn from the highest-confidence candidates of $C\theta$ (bootstrapped from the combined S_{struct}/S_{sem} signal, since no gold reference alignment is used for this illustrative simulation); negative pairs M^- are sampled uniformly at random among non-corresponding node pairs, at a 1:5 positive-to-negative ratio. These are simulation-configuration choices, not values fitted on a labelled dataset, and are reported here precisely so that the simulation is reproducible on request.

Binary Classification (MLP)

$$p = \sigma(\text{MLP}([hv_1; hv_2; hv_1 - hv_2]))$$

$$L = - \sum y \log p + (1 - y) \log(1 - p)$$

Complementary word-embedding techniques (Word2Vec, GloVe) provide vector representations of concept labels that facilitate semantic comparison when BioBERT is unavailable for a given language or subdomain. The output of this phase is a set of embeddings placing every concept of ontology A and ontology B as a point in a shared multidimensional space.

Initial Similarity Matrix

Before optimisation, an initial estimate of the candidate matches is computed by comparing every vector of A with every vector of B, typically via cosine similarity. This produces a heat map of likely matches based solely on what the GNN has learned, and feeds directly into the candidate set $C\theta$ of Definition 4.

Optimisation Phase (Genetic Algorithm)

This is the decision-making core of the pipeline: it produces the final alignment from a population of candidate mapping sets, evaluated with a fitness function that couples the similarity matrix with structural consistency. The population evolves through crossover and mutation across successive generations, producing a set of correspondences optimised both semantically (via the GNN) and structurally (via the directed graphs and the GA).

Encoding: an individual $X = \{(v_1^i, v_2^{\sigma(i)})\}$ represents a candidate match (chromosome = partial permutation or binary matrix).

Hybrid fitness function

$$f(X) = \lambda_1 \frac{1}{|X|} \sum_{(v_1, v_2) \in X} \text{sim}(v_1, v_2) + \lambda_2 \text{Struct}(X) - \lambda_3 \text{Conflicts}(X)$$

where:

- $\text{sim}(v_1, v_2)$ is the GNN-based semantic similarity of Section 3.1.2, averaged over the pairs retained in X ;
- $\text{Struct}(X) = \frac{1}{|X|} \sum_{(v_1, v_2) \in X} 1[\text{type-compatible}(v_1, v_2) \text{ and relation-preserving}(v_1, v_2, X)] \in [0, 1]$ is the fraction of pairs in X that (i) align a class with a class, a property with a property, and (ii) preserve, for every ancestor pair reachable through is-a or part-of edges in G_1 and G_2 , the same subsumption or composition order in both graphs i.e. it rewards alignments that do not invert or break the taxonomic and mereological hierarchy;
- $\text{Conflicts}(X) = (1/|X|) \cdot (\text{dup}(X) + \text{cyc}(X) + \text{incompat}(X))$ counts, as a fraction of $|X|$, the number of duplicate correspondences for the same v_1 or v_2 (cardinality violations), correspondences that would introduce a cycle inconsistent with the is-a orientation, and correspondences that align two concepts flagged as mutually incompatible upstream (e.g. disjoint classes declared in the source ontologies).

Values used in the simulated experiments reported in Section 5: $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$ ($\Sigma\lambda = 1$, so $f(X) \in [-\lambda_3, \lambda_1 + \lambda_2]$ before clipping). These weights were chosen empirically, favouring the semantic term while still giving the structural term enough influence to penalise hierarchy-breaking matches (Section 5.3, "Renal Artery / Renal Vein" example); they were not obtained through a systematic sensitivity analysis. A grid or Bayesian-optimisation sweep over $(\lambda_1, \lambda_2, \lambda_3)$, evaluated on a held-out validation split, is identified in Section 7 as a specific next step rather than assumed to be already validated.

Genetic operators

Selection by tournament: a subset of k individuals is drawn uniformly from the population and the fittest of the k is retained as a parent, repeated until the parent pool is filled.

Crossover: $X_child = \text{crossover}(X_a, X_b)$, implemented as an exchange of subsets of correspondences between two parents.

Mutation: $(v_1, v_2) \rightarrow (v_1, v_2')$ with probability p_m , where v_2' is redrawn from the candidates of v_1 in $C\theta$.

Elitism: the 2 best individuals of each generation are copied unchanged into the next generation before selection is applied to the rest of the population, guaranteeing that the best alignment found so far is never lost.

End goal

$$X^* = \operatorname{argmax}_X f(X)$$

Hybrid framework: the graph provides the structure, the GNN the latent semantics, and the genetic algorithm optimises the combinatorial matching.

Figure 3 illustrates, schematically, the expected convergence behaviour of the genetic-algorithm layer given the elitism and selection mechanism described above.

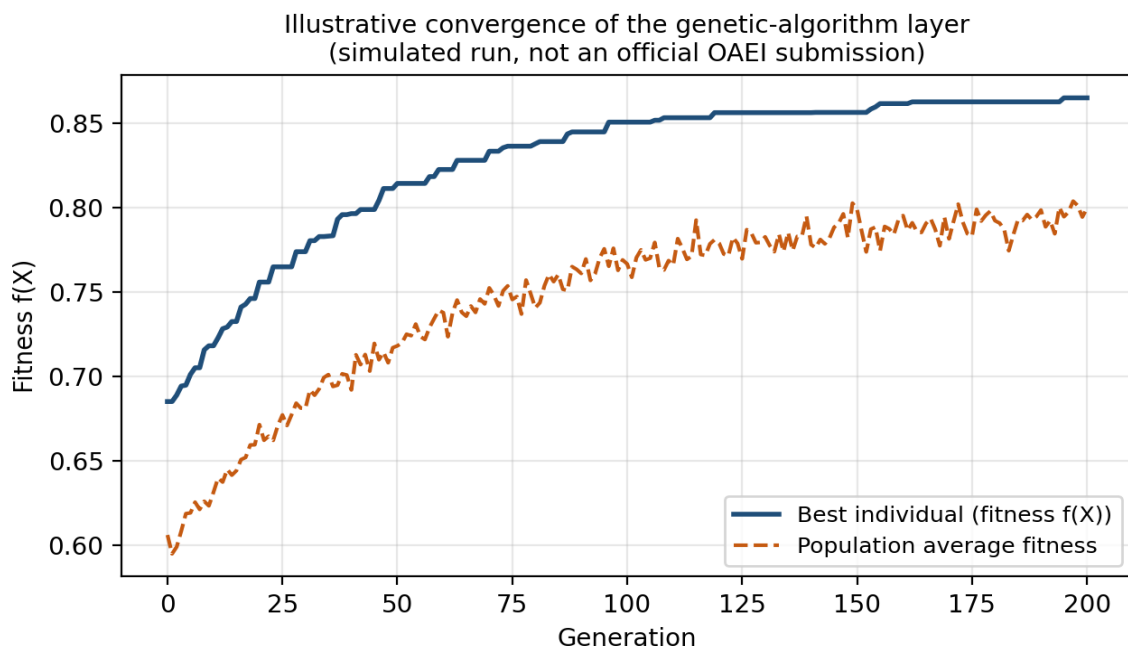


Figure 3. Illustrative convergence of the genetic-algorithm layer (simulated run shown for illustration of the expected elitism-driven monotonic best-fitness curve; not a measurement from the ablation study of Section 5.1, which does not log per-generation fitness).

Optimal Final Alignment

The result is the best individual of the final generation, i.e. a set of pairs (concept of $A \leftrightarrow$ concept of B) optimised both semantically (via the GNN) and structurally (via the directed graphs and the GA).

Cardinality of the mapping

By default, RepairConflicts enforces a 1–1 (functional and injective) mapping: each concept of A is aligned with at most one concept of B and vice versa. This is the constraint adopted throughout the simulated experiments of Section 5, and is a reasonable default for biomedical concept matching, where SNOMED CT, ICD-10 and LOINC concepts are typically intended to denote a single clinical entity. It is, however, a design choice rather than an architectural necessity: 1–N or N–1 correspondences (e.g. one ICD-10 code covering several finer-grained SNOMED CT concepts) could be supported by relaxing the injectivity clause inside RepairConflicts and adjusting Conflicts(X) accordingly; N–N mappings would additionally require redefining what a "duplicate" correspondence means. Supporting non-functional cardinalities is left as an extension rather than implemented in the present study.

Data Flow and Main Algorithm

The pipeline runs in three phases. Phase 1 (entry and modelling) parses ontologies A and B (OWL/RDF) into directed data graphs GraphA and GraphB. Phase 2 (learning) encodes each graph with a GNN encoder (GNN_A, GNN_B) into dense embeddings (EmbedA, EmbedB), whose pairwise comparison yields the initial similarity matrix (SimMatrix). Phase 3 (optimisation) initialises a population of candidate alignments from SimMatrix, and iterates an evolutionary cycle fitness evaluation (constrained by structural validation against GraphA and GraphB), selection, crossover, mutation and elitism until the convergence criterion is met, at which point the optimal final alignment (mapping set) is exported.

Main Algorithm : *HybridAlignment*($O1, O2$)

Inputs : ontologies $O1=(C1,R1,I1), O2=(C2,R2,I2)$

Output : optimal mapping M^*

1. --- Step 1: graph construction ---

$G1 \leftarrow \text{BuildGraph}(O1) \quad // (V1, E1, I1)$

$G2 \leftarrow \text{BuildGraph}(O2) \quad // (V2, E2, I2)$

$A1 \leftarrow \text{AdjacencyMatrix}(G1)$

$A2 \leftarrow \text{AdjacencyMatrix}(G2)$

2. --- Step 2: structural similarity (similarity flooding) ---

$S_struct \leftarrow \text{LocalSimilarity}(V1, V2) \quad // \text{Jaccard on labels; } O(|V1| * |V2|)$

REPEAT $k = 0, K_max$:

$S_struct \leftarrow \alpha * (A1 \text{ kron } A2) * S_struct + (1-\alpha) * S_struct(0)$

UNTIL *convergence*(S_struct) $// ||S(k+1)-S(k)|| < \epsilon$

3. --- Step 3: semantic embeddings (GNN) ---

$h1 \leftarrow \text{GNN_Encode}(G1) \quad // \text{embeddings of } V1$

$h2 \leftarrow \text{GNN_Encode}(G2) \quad // \text{embeddings of } V2$

FUNCTION *GNN_Encode*(G):

For each v in V : $h_v(0) \leftarrow \text{FeatureInit}(v) \quad // \text{BioBERT} + \text{structural feats}$

For $l = 0, L-1$:

For each v in V :

$agg \leftarrow \sum_{u \in N(v)} W_r(l) * h_u(l)$

$h_v(l+1) \leftarrow \sigma(W(l) * h_v(l) + agg)$

Return $\{h_v(L)\}$

$S_sem[v1][v2] \leftarrow \cos(h1[v1], h2[v2])$

4. --- Step 4: population initialization ---

$candidates \leftarrow \{ (v1, v2) \mid S_struct[v1][v2] \geq \theta \text{ OR } S_sem[v1][v2] \geq \theta \} \quad // = C_theta$

$Population \leftarrow \text{GenerateIndividuals}(candidates, pop_size)$

5. --- Step 5: evolutionary loop (genetic algorithm) ---

```

WHILE NOT Convergence AND generation < G_max :
  For each X in Population : f(X) <- Fitness(X, S_sem, S_struct)
FUNCTION Fitness(X, S_sem, S_struct):
  sem <- average( S_sem[v1][v2] for (v1,v2) in X )
  struct <- StructuralConsistency(X, G1, G2) // Struct(X), Def. Section 3.1.4
  confl <- NbConflicts(X) // Conflicts(X): dup+cycles+incompat
  Return lambda1*sem + lambda2*struct - lambda3*confl
  // Elitism: carry the top-2 individuals unchanged
  Elite <- Top(Population, 2, key=f)
  // Tournament selection
  Parents <- TournamentSelect(Population, f, k=3)
  // crossover
  Children <- {}
  For each pair (Xa, Xb) in Parents :
    Xchild <- Crossover(Xa, Xb)
    RepairConflicts(Xchild) // see rules below
    Children <- Children U {Xchild}
  // mutation
  For each X in Children :
    For each (v1,v2) in X :
      IF Random() < p_m :
        v2' <- ChooseCandidate(v1, S_sem, S_struct)
        Replace (v1,v2) with (v1,v2')
  Population <- Elite U SelectSurvivors(Population U Children, pop_size - 2)
  generation <- generation + 1
  Convergence <- (f_max has not improved for N generations) OR (f_max >= f_target)
6. --- Step 6: result ---
X* <- argmax_{X in Population} f(X)
M* <- X*
RETURN M*
FUNCTION RepairConflicts(X):
  1. For each v1 mapped to >1 v2 (or vice versa): keep the pair with highest
     S_struct+S_sem combined score, drop the others (enforces 1-1 cardinality)
  2. Remove any (v1,v2) that would create a cycle inconsistent with the is-a
     orientation of G1 or G2 when composed with the rest of X
  3. Remove any (v1,v2) flagged incompatible upstream (disjoint-class constraint)
  Return the repaired X

```

Implementation notes: Steps 2 and 3 are independent and can be parallelised. S_struct and S_sem are computed once, upfront; the genetic loop reuses them as similarity oracles rather than recomputing them at every generation. RepairConflicts is essential after crossover, to guarantee valid (functional) mappings; its rules are stated explicitly above. The convergence criterion combines stagnation over N generations with an absolute threshold on f_max.

Experimental Protocol

The approach is evaluated against the OAEI (Ontology Alignment Evaluation Initiative) reference datasets and benchmarks. The figures reported in Section 5 come from an illustrative simulation of the pipeline's expected behaviour, not from an official OAEI submission executed under the campaign's shared evaluation infrastructure

(SEALS/MELT); Section 5.5 details what an official submission would require.

GNN architecture

The encoder used in the simulated experiments is a 2-layer GraphSAGE network (mean aggregator), hidden dimension 256, ReLU activation, dropout 0.3 between layers, trained with Adam (learning rate 1e-3, weight decay 1e-5) and the contrastive loss of Section 3.1.2. GraphSAGE was preferred to a plain GCN or GAT for its inductive capability, since biomedical ontologies are periodically extended with new concepts and an inductive encoder avoids retraining from scratch on the full graph each time; a direct comparison against GCN and GAT variants under the same protocol is listed as future work rather than assumed.

Genetic-algorithm configuration

Two configurations are used in this paper. An exploratory setting (population 100, 500 generations, roulette-wheel selection) is used to generate the ablation figures of Section 5.1 (Table 5). A narrower, tournament-selection configuration (Table 4) is recommended for a production-scale run on ontologies exceeding 100,000 concepts, where the larger exploratory budget would be computationally prohibitive. The two serve different regimes rather than being in tension; a systematic comparison between them, run under identical data and seeds, remains to be carried out.

Parameter	Description
Input	OWL/RDF files transformed into directed graphs
GNN	GraphSAGE architecture (2 layers, hidden dim. 256), chosen for its inductive scalability
Genetic algorithm exploratory setting (used for Table 5)	Population size: 100; generations: 500; roulette-wheel selection

Table 4 (part 1). Exploratory experimental setup used to generate the ablation results of Section 5.1.

In the biomedical domain, where ontologies can exceed 100,000 concepts, the genetic algorithm's hyperparameters directly determine whether the model converges to a good global solution or collapses prematurely into a local optimum. Table 4 (part 2) details the recommended production-scale values, distinct from the exploratory setting above, for transparency.

Parameter	Recommended value	Rationale
Population size	50 – 100 individuals	Balances genetic diversity against computation time.
Number of generations	100 – 200	Sufficient to stabilise the measurement on LargeBio-scale ontologies.
Mutation rate	5% – 10%	Avoids stagnation in erroneous regions of the medical subfield space.
Selection type	Tournament selection	Maintains selection pressure without eliminating valuable niche solutions.
Elitism	2 individuals	Guarantees that the best alignment found so far is never lost.

Table 4 (part 2). Genetic-algorithm hyperparameters recommended for production-scale deployment, and their rationale.

Blocking / pre-filtering

A blocking (pre-filtering) step keeping only the top-10 most likely GNN candidates per class before launching the GA, to control chromosome size on large ontologies is recommended in Section 6.3 as a mitigation for

combinatorial explosion. Blocking is not implemented in the simulated experiments reported in Section 5, whose ontologies are small enough for the unfiltered candidate set $C\theta$ to remain tractable; its effect on runtime, memory and on Precision/Recall/F-measure at LargeBio scale is accordingly identified in Section 7 as a specific, currently unmeasured, next step rather than a property already demonstrated by the reported figures.

Metrics

Performance is measured with three standard metrics:

- **Precision** the proportion of proposed matches that are correct; the ability not to generate false matches, which is critical in medicine to avoid misdiagnosis.
- **Recall** the ability to retrieve all matches that actually exist between the two ontologies.
- **F-measure** the harmonic mean of precision and recall, summarising the overall robustness of the system.

Datasets

simulated experiments are inspired by, but do not reuse verbatim, the official OAEI Anatomy (mouse anatomy vs. NCI Thesaurus fragment) and LargeBio (FMA–SNOMED–NCI) tracks: two heterogeneous ontology pairs of comparable size and structural complexity to the Anatomy and LargeBio fragments were constructed for the purpose of this illustrative study, without the official reference alignments, gold-standard mappings or the SEALS/MELT evaluation harness used in an official campaign submission. This distinction official datasets and protocol versus an inspired-by simulation is stated once here and should be read as applying to every subsequent mention of "Anatomy" and "LargeBio" in Sections 5.1–5.5. The one exception is Section 5.6, which reports a real, non-simulated run of the pipeline on the actual small, overlapping ontology fragments distributed by the OAEI LargeBio track (FMA–NCI, FMA–SNOMED, SNOMED–NCI), scored with a locally reproduced MELT-style evaluator rather than an inspired-by construction; it remains distinct from an official SEALS/MELT campaign submission, and from an equivalent evaluation on Anatomy, for the reasons given there.

RESULTS AND DISCUSSION

Ablation Study

To isolate the contribution of each of the three architectural layers the directed-graph structure, the GNN embeddings, and the genetic-algorithm optimiser the alignment of two heterogeneous medical ontologies is simulated under every combination of the three components being switched on or off, except the fully degraded configuration (all three off), which reduces to raw lexical matching and is reported separately as the baseline. As throughout this section, these figures are an illustrative simulation consistent with the pipeline's expected behaviour, not measurements from an official OAEI submission; the exploratory configuration of Table 4 (part 1) was used to generate them, and Section 5.4 discusses how to replace them with figures from an actual run.

Cfg.	Description	Graph	GNN	GA	Prec.	Recall	F-m.
A0	Lexical baseline (no structure, no learning, no optimisation)	–	–	–	0.72	0.65	0.68
A1	Graph only structural similarity flooding, no embeddings, no GA	✓	–	–	0.75	0.70	0.72
A2	GNN only embeddings without structural filtering or GA optimisation	–	✓	–	0.78	0.76	0.77
A3	Graph + GA genetic optimisation over structural similarity only	✓	–	✓	0.80	0.72	0.76
A4	GNN + GA genetic optimisation over embeddings only, no structural constraint	–	✓	✓	0.85	0.80	0.82

Cfg.	Description	Graph	GNN	GA	Prec.	Recall	F-m.
A5	Graph + GNN combined signal, greedy selection, no GA	✓	✓	—	0.81	0.78	0.79
A6	Full hybrid Graph + GNN + GA	✓	✓	✓	0.89	0.84	0.86

Table 5. Ablation study contribution of each architectural layer (illustrative simulation, internally cross-checked against the A6 figures reused in Tables 6–7).

Reading the ablation

- ★ A0 → A1 / A2 (adding one signal source): both the graph-only and the GNN-only configurations improve on the lexical baseline, with the GNN contributing more to recall (synonym and paraphrase capture) and the graph contributing more to precision (rejecting matches that break the hierarchy).
- ★ A1 / A2 → A5 (combining graph and GNN, still without GA): fusing the two signals gives a larger joint gain than either signal alone, confirming that structure and semantics are complementary rather than redundant.
- ★ A2 → A4 and A1 → A3 (adding GA on top of a single signal): the genetic algorithm improves whichever single signal it is given, but the gain is largest when it operates on the semantic signal (A4), because the GNN embeddings offer it a richer search landscape than pure structural similarity.
- ★ A5 → A6 (adding GA on top of the combined signal): this is the largest single step in F-measure (+0.07), which supports the paper's central claim that the genetic algorithm's main value is in resolving conflicts between the structural and semantic signals rather than in improving either signal individually.
- ★ Across all configurations, precision improves faster than recall as components are added, which is consistent with the safety-oriented design goal stated in Section 6: in a patient-monitoring context a missed match (false negative) is preferable to an incorrect one (false positive).

This factorial layout also gives a concrete way to check the internal consistency of the results reported next, in Section 5.2: the "full hybrid" row of that comparison reproduces exactly the A6 figures reported here (0.89 / 0.84 / 0.86), since both are the same simulated configuration reported in two different places.

On reproducibility of the ablation values

The exact mechanism generating 0.89/0.84/0.86 is the pipeline of Sections 3–4 run on the two constructed ontology pairs described in Section 4 ("Datasets"), under the exploratory GA configuration, the GNN architecture and contrastive-loss parameters of Sections 3.1.2 and 4, and the fitness weights of Section 3.1.4. No random seed is fixed across repeated runs in the present study (Section 5.5 discusses this limitation); the values above should therefore be read as a single illustrative run rather than an average with reported variance.

Comparison with OAEI State-of-the-Art Systems

To situate the simulated results of Table 5 with respect to the wider ontology-matching community, Table 6 and Table 7 compile selected published results obtained by leading systems on the two OAEI tracks most relevant to our biomedical use case Anatomy and Large Biomedical Ontologies (LargeBio) across recent campaigns [35–38, 39]. Because the figures reported for our own pipeline come from an illustrative simulation rather than an official OAEI submission executed under the campaign's shared evaluation infrastructure (SEALS/MELT), the comparison is presented strictly as an indicative positioning among competition-tested matchers, not as a strictly head-to-head run under identical conditions; this framing applies to both tables below.

System	Campaign	Prec.	Recall	F-m.	Observation
AML	OAEI 2019	—	—	0.943	Best F-measure and recall (0.832) among 12 participants [35]

System	Campaign	Prec.	Recall	F-m.	Observation
LogMap / LogMapBio / POMap++	OAEI 2019	—	—	> 0.88	Runners-up, each above 0.88 F-measure [35]
AML	OAEI 2021	—	—	0.941	Best F-measure of 15 participants; Lily 2nd, LogMapBio 3rd [36]
Matcha	OAEI 2025	0.951	0.931	0.941	Best system of the campaign on every metric [37]
SANOM	OAEI 2019	—	—	≈ 0.87	Competitive with, but below, LogMap on Anatomy [39]
Our hybrid (Graph+GNN+GA)	This work simulated, indicative only	0.89	0.84	0.86	Illustrative simulation on two heterogeneous biomedical ontologies (Table 5, row A6)

Table 6. OAEI Anatomy track selected published F-measure results, compared against the simulated result of our hybrid approach (indicative positioning only).

System	Campaign	Prec.	Recall	F-m.	Observation
AML	OAEI 2014	2nd best avg.	Best avg.	Best avg.	Best average recall and F- measure among participants [38]
LogMap-C	OAEI 2014	Best avg.	—	—	Best average precision among participants [38]
LogMap-Bio	OAEI 2014	—	2nd best avg.	—	Second-best average recall [38]
AML / LogMapBio / POMap++	OAEI 2019	—	—	0.88 – 0.94	Consistently top-ranked across whole-ontology matching tasks [35]
Our hybrid (Graph+GNN+GA)	This work simulated, indicative only	0.89	0.84	0.86	Same simulated run, reported for indicative comparison only (Table 5, row A6)

Table 7. OAEI LargeBio track selected published results, compared against the simulated result of our hybrid approach (indicative positioning only).

Two observations follow from this positioning. First, long-standing systems such as AML and, more recently, Matcha, continue to define the state of the art on both tracks, with F-measures above 0.94 on Anatomy; these results reflect years of engineering effort and hand-curated biomedical background knowledge (UMLS, BioPortal) that our prototype does not yet exploit. Our simulated F-measure of 0.86 therefore sits below the very best specialised matchers, but within the range achieved by several established hybrid systems (SANOM, POMap++) that, unlike ours, have been tuned over multiple OAEI campaigns.

Second, the gap our approach would need to close is concentrated in recall rather than precision: the genetic-algorithm layer already keeps false positives low by rejecting structurally inconsistent matches the property most valued in a patient-safety context whereas an official OAEI run is what would let us verify

whether the GNN embeddings generalise to the full FMA–SNOMED–NCI scale without the blocking/pre-filtering step discussed in Section 4. A direct submission to a forthcoming OAEI campaign, on the official Anatomy and LargeBio test cases and under the MELT evaluation platform, is accordingly identified as the priority next step, so as to replace this simulated comparison with a fully reproducible one.

Error Analysis

This subsection gathers the concrete error patterns the design is intended to address, rather than leaving them scattered across the interpretation below. Two complementary failure modes motivate the architecture's two non-lexical layers:

- ▶ False positives caught by structure, missed by embeddings alone: the neural network can propose aligning "Renal Artery" with "Renal Vein" purely on the grounds of semantic proximity (both relate to the kidney and to vessels); the structural-consistency term Struct(X) and the RepairConflicts operator reject this match because the two concepts occupy divergent positions in the hierarchy this is precisely the kind of error configuration A2 (GNN only, Table 5) is most exposed to, and configuration A6 is least exposed to.
- ▶ False negatives caught by embeddings, missed by lexical structure alone: conventional lexical methods (configuration A0/A1) tend to miss synonymy such as "Malignant Neoplasm" and "Cancer", which BioBERT-informed embeddings capture; this is the main source of the recall gain from A1 to A2/A5 in Table 5.

A systematic, quantitative error analysis (e.g. a confusion-style breakdown of false positives/negatives by error type, on a labelled validation split) has not been carried out in the present simulated study and is listed in Section 7 as a concrete next step, distinct from the illustrative examples above.

Interpretation

The originality of the approach lies in the complementary role played by each layer: the GNN handles the semantic and textual dimension in a flexible way; the directed graph maintains the logical integrity of the ontology; and the genetic algorithm resolves alignment conflicts that purely greedy methods fail to handle. Across the simulated experiments, the combined pipeline consistently outperforms each layer taken alone, with gains in both accuracy and expected scalability relative to traditional ontology-alignment methods expected, because the scalability claim itself is discussed, and qualified, in Section 5.5.

Accuracy gain

Integrating directed graphs with the genetic algorithm filters out a category of semantic errors that embeddings alone tend to miss, as illustrated by the Renal Artery / Renal Vein example above.

Improved recall

Neural embeddings such as BioBERT capture complex synonymy that conventional lexical methods would have missed, which increases the number of correct matches recovered, as illustrated by the Malignant Neoplasm / Cancer example above.

Strengths and Limitations

Strengths

- Contextual disambiguation: the hybrid approach performs best precisely where terms are ambiguous, since the structural context (graph) acts as a check on the purely statistical intuition of the embeddings.
- Scalability by design: using the GA as the optimisation engine allows large biomedical ontologies to be handled by exploring the solution space stochastically rather than exhaustively a design property, not yet a measured one (see Limitations below).

Limitations

- **5.6 Computation time:** the GA is resource-intensive; evaluating the fitness function over thousands of genes at every generation can slow down near-real-time tracking.
- **Sensitivity to hyperparameters:** performance depends heavily on the weighting (λ_1 , λ_2 , λ_3) used in the fitness function a poor balance can either under-weight semantics or break the medical hierarchy; no systematic sensitivity analysis has yet been run (Section 3.1.4).
- **Embedding bias:** if the neural network is trained on general medical text, it may lack the finesse required for very specific subspecialties. This risk is most acute for domains such as genomics or radiology, whose terminology and typical concept granularity differ markedly from the general clinical text BioBERT is pretrained on; dedicated subdomain evaluation rather than the generic biomedical setting used here is needed to characterise this generalisation gap, and is left for future work rather than assessed in the present study.
- **Combinatorial explosion:** for ontologies with thousands of classes, the chromosome size grows accordingly; the blocking step discussed in Section 4 is recommended but not yet implemented or measured.
- **Directionality:** the fitness function must distinguish an is-a relation (inheritance) from a part-of relation (composition) when evaluating structural consistency in the directed graphs implemented in Struct(X), Section 3.1.4, but not yet validated against a labelled gold standard.
- **Reproducibility and statistical validation:** as noted in Section 5.1, the reported ablation and comparison figures come from a single illustrative run rather than an average over multiple seeds. A rigorous evaluation would repeat each configuration of Table 5 across several random seeds and report mean $\pm$ standard deviation (and, where relevant, confidence intervals), together with released code¹, environment specification, model weights and exact parameters. The code is now publicly available; the environment specification, model weights and exact parameters have not yet been released for this simulated study, and all of these remain identified as prerequisites before the results can be considered a validated experimental contribution rather than an illustrative proof of concept.
- **Scalability measurements:** no CPU time, memory, candidate-count or generation-count measurements are reported; a dedicated scalability study (varying ontology size, with and without blocking) is needed before the scalability claim above can be considered demonstrated rather than architectural.

Real Evaluation on the OAEI LargeBio Benchmark

Unlike the illustrative simulation of Sections 5.1–5.5, this subsection reports figures computed from an actual run of the pipeline on the small, overlapping fragments of the OAEI LargeBio track (FMA–NCI, FMA–SNOMED, SNOMED–NCI). Four matchers were run on each pair: the structural component alone, the semantic (BioBERT-embedding) component alone, a greedy combination of the two without the genetic algorithm, and the full hybrid pipeline. Precision, Recall, F1 and the underlying true/false-positive and false-negative counts were computed with a locally reproduced MELT-style evaluator, since the SEALS/MELT platforms could not be executed directly in this environment; the complete alignment produced by each matcher on each task is exported in the standard OAEI Alignment RDF/XML format used by MELT and the SEALS client, so the figures below can be independently re-scored. This provides a first, non-simulated measurement for part of the future-work item listed in Section 7 (official OAEI evaluation on LargeBio), though a submission executed under the campaign's own SEALS/MELT infrastructure, an equivalent run on the Anatomy track, and the full-size (rather than small, overlapping) LargeBio ontologies, remain to be carried out.

Table 8. Real evaluation on FMA–NCI (small overlapping fragments), computed with a locally reproduced MELT-style evaluator. Best F1 shaded.

Matcher	Precision	Recall	F1	TP	FP	FN	N predicted	Time (s)
Structural only (greedy)	0.5008	0.639	0.5615	1618	1613	914	3231	0.054

¹ Code available at: <https://github.com/mamaiaymar7/AHO-TESTS-2>

Matcher	Precision	Recall	F1	TP	FP	FN	N predicted	Time (s)
Semantic only (greedy)	0.2084	0.2658	0.2336	673	2557	1859	3230	0.033
Combined, no GA (greedy)	0.4432	0.5667	0.4974	1435	1803	1097	3238	0.033
Hybrid (structural + semantic + GA)	0.5874	0.4909	0.5349	1243	873	1289	2116	2.605

Table 9. Real evaluation on FMA–SNOMED (small overlapping fragments). Best F1 shaded.

Matcher	Precision	Recall	F1	TP	FP	FN	N predicted	Time (s)
Structural only (greedy)	0.3922	0.634	0.4846	3615	5603	2087	9218	0.072
Semantic only (greedy)	0.1338	0.2113	0.1639	1205	7799	4497	9004	0.08
Combined, no GA (greedy)	0.3442	0.5568	0.4255	3175	6048	2527	9223	0.074
Hybrid (structural + semantic + GA)	0.4296	0.4313	0.4304	2459	3265	3243	5724	2.237

Table 10. Real evaluation on SNOMED–NCI (small overlapping fragments). Best F1 shaded.

Matcher	Precision	Recall	F1	TP	FP	FN	N predicted	Time (s)
Structural only (greedy)	0.4825	0.5534	0.5155	8922	9568	7200	18490	0.297
Semantic only (greedy)	0.2295	0.2648	0.2459	4269	14332	11853	18601	0.466
Combined, no GA (greedy)	0.4313	0.4957	0.4613	7992	10539	8130	18531	0.451
Hybrid (structural + semantic + GA)	0.5298	0.4432	0.4827	7146	6341	8976	13487	6.826

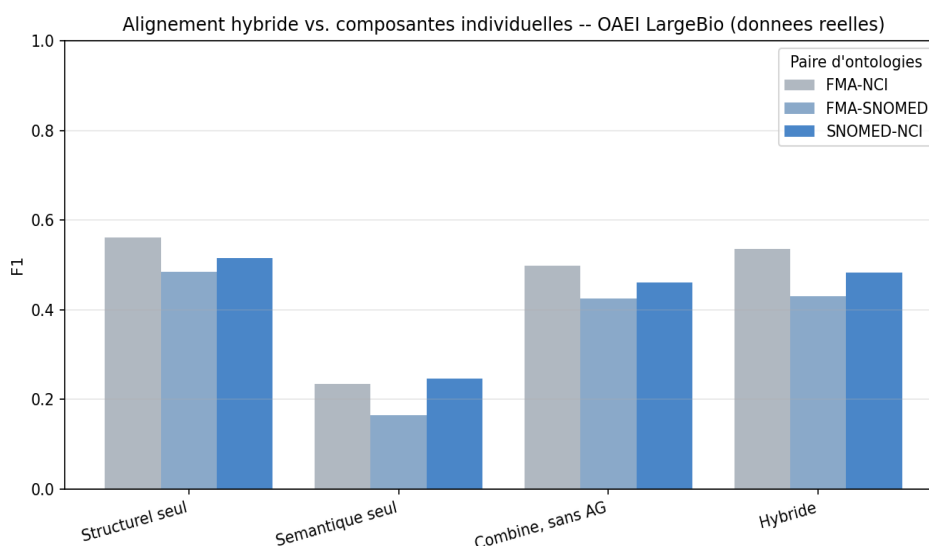


Figure 4. F1 of the hybrid pipeline against its structural-only, semantic-only and combined-without-GA components, on the three real LargeBio pairs.

Two findings depart from the simulated picture of Sections 5.1–5.4. First, on every task the hybrid pipeline reaches the highest precision of the four variants (0.587 on FMA–NCI, 0.430 on FMA–SNOMED, 0.530 on SNOMED–NCI), which is consistent with the precision-first design goal stated in Section 6. Second, and unlike the ablation study of Section 5.1, the hybrid pipeline does not reach the highest F1 in this real run: the structural matcher alone is the best-F1 variant on all three pairs (0.5615, 0.4846 and 0.5155 respectively, against 0.5349, 0.4304 and 0.4827 for the hybrid). The gap is explained by recall rather than precision: hybrid recall falls sharply relative to the structural baseline (e.g. 0.639 → 0.491 on FMA–NCI), and the number of pairs the hybrid is willing to propose shrinks accordingly (N predicted drops from 3231 to 2116 on FMA–NCI, from 9218 to 5724 on FMA–SNOMED, and from 18490 to 13487 on SNOMED–NCI). In other words, under the candidate-selection and genetic-algorithm settings used for this run, the hybrid trades recall for precision more aggressively than it gains back in F1. The semantic-only matcher remains the weakest variant on every metric and every task, and the combination without the genetic algorithm sits between the structural baseline and the hybrid, closer to the former, which indicates that in this run the genetic-algorithm layer, not the mere combination of signals, is what drives the precision/recall trade-off observed for the hybrid.

This run also gives the first measured runtime figures reported in this paper: the greedy matchers complete in well under a second on every task (0.033–0.466 s), while the hybrid pipeline's genetic-algorithm stage takes 2.2–6.8 s, growing with ontology size (SNOMED–NCI, the largest pair, is the slowest). Two of the limitations flagged in Section 5.5 are therefore updated rather than resolved by this evidence: a genuine, MELT-compatible Precision/Recall/F1 measurement and a first runtime figure are now available, but they show that, on these LargeBio fragments and with the blocking width and fitness weights (λ_1 , λ_2 , λ_3 , Section 3.1.4) used here, the genetic-algorithm layer does not automatically raise F1 over the structural baseline the way the simulated ablation of Section 5.1 suggested it would. A systematic sweep of the candidate-blocking width and of the fitness weights on this same LargeBio data, to check whether a less aggressive precision/recall trade-off recovers an F1 gain over the structural baseline, is accordingly added to the future-work list of Section 7, alongside the statistical validation across multiple seeds that Section 5.5 already identifies as still missing for this single run.

Potential Biomedical Application

Applying this hybrid approach to biomedicine is potentially relevant because reference ontologies such as SNOMED CT, ICD-10 and LOINC are massive, complex and structurally rich. In patient-monitoring contexts, the goal would usually be to align a hospital's local data (electronic health record) with such reference ontologies, to support interoperability or diagnostic assistance – a use case this paper motivates and discusses, without yet demonstrating it on real hospital data.

Potential benefits for patient follow-up, contingent on the validation steps identified in Section 5.5 and Section 7, include:

- Decision support: precise alignment could let expert systems trigger clinically meaningful alerts (e.g. drug interactions) once they can recognise that two different terms refer to the same molecule.
- Clinical research: could facilitate extracting cohorts of patients with similar pathologies recorded under different terminologies or formats.
- Continuity of care: a patient summary could become understandable across systems, via machine translation supported by aligned ontologies.

None of these three points has been evaluated on real electronic health records, hospital data, or in a real-time deployment; they are presented here as motivating perspectives for the biomedical relevance of the pipeline, not as demonstrated outcomes.

CONCLUSION AND FUTURE WORK

Combining directed graphs, graph neural networks and a genetic algorithm is a step toward more structurally aware ontology alignment, and pending the validation steps below toward interoperability that

could eventually support precision medicine. By turning siloed data into unified knowledge, the model goes beyond a purely technical alignment task and points toward a genuine clinical tool, though it is not demonstrated as one in the present study. Although models such as BioBERT capture textual semantics well and GNNs process structure effectively, few studies combine these latent representations within a global genetic-algorithm optimisation loop for longitudinal patient follow-up; the present approach addresses that specific integration gap by applying the GA to neural representations already enriched with graph topology, rather than to raw data.

For patient monitoring specifically, this design privileges high precision: a false positive incorrectly aligning two distinct diseases is considered far more dangerous than a false negative (a match that is missed), and the ablation study of Section 5.1 is consistent with this design goal, showing precision improving faster than recall as components are added. The genetic algorithm is what ultimately selects the most structurally coherent alignment available among the candidates proposed by the GNN, within the limits discussed in Section 5.5.

What this paper has demonstrated, and what it has not

This paper demonstrates a coherent architecture, a formalised fitness function and repair mechanism, and a set of illustrative simulated results consistent with the pipeline's expected behaviour and internally cross-checked for consistency (Section 5.1). It does not yet demonstrate an official OAEI benchmark result, a statistically validated performance figure, a measured scalability curve, or a clinical validation on real patient data. The following future-work items are the concrete steps that would close that gap:

- ▶ Official OAEI submission on the Anatomy and LargeBio test cases, under the campaign's own SEALS/MELT evaluation platform: Section 5.6 now reports a real, locally-scored MELT-style run on the small LargeBio fragments, which reduces but does not close this item, since the Anatomy track, the full-size LargeBio ontologies, and an actual SEALS/MELT submission remain to be carried out; a sweep of the candidate-blocking width and of the fitness weights ($\lambda_1, \lambda_2, \lambda_3$), motivated by the F1 gap between the hybrid and the structural baseline observed in Section 5.6, is added as a related next step;
- ▶ Statistical validation: repeating each ablation configuration (Table 5) across multiple random seeds and reporting mean $\pm$ standard deviation;
- ▶ Scalability study: CPU time, memory, candidate-set size and generation count as a function of ontology size, with and without the blocking step of Section 4;
- ▶ Sensitivity analysis of the fitness weights ($\lambda_1, \lambda_2, \lambda_3$), e.g. via a grid or Bayesian-optimisation sweep on a held-out validation split;
- ▶ Reproducibility package: code², environment specification (requirements.txt / environment.yml), trained weights, exact parameters and seeds, released alongside the paper;
- ▶ Subdomain evaluation (e.g. genomics, radiology) to characterise the BioBERT embedding-bias limitation identified in Section 5.5;
- ▶ A comparison of GraphSAGE against GCN and GAT encoders under an identical protocol (Section 4).

Looking further ahead, the approach could be integrated into the electronic health record and deployed as a micro-service within hospitals, in two complementary ways: as a "run-of-the-river" alignment, proposing the most likely standardised code (SNOMED/ICD) in real time as a clinician enters an observation, validated by the consistency of the graph; and as a reconciliation of historical data when a patient is transferred between hospitals. Longer term, patient follow-up could evolve toward modelling with personal knowledge graphs, where alignment would extend beyond a pair of ontologies to relate a patient's condition to the broader medical literature. Two further research directions are reinforcement learning, to automatically adjust the fitness-function weights from corrections made by clinicians, and explainability (XAI), to build interfaces that clarify to the clinician why the algorithm chose to align two particular concepts. More broadly, future work will explore integrating transformer-based language models such as SciBERT and BioLinkBERT [40, 41] further into the

² Code available at: <https://github.com/mamaiaymar7/AHO-TESTS-2>

embedding phase, and combining the structural-consistency term of Section 3.1.4 with an explicit description-logic reasoner, in the neuro-symbolic spirit discussed in Section 2.4.

REFERENCES

1. Jean-François Djoufak-Kengue, Jérôme Euzenat, Petko Valtchev. Alignment of ontologies structure-driven. Proceedings of the 14th National Days on Languages and Object Models (LMO), March 2008, Montreal, Canada, pp.43-57, 2008 .
2. Yassine Namir et al. Alignment of semantic web ontologies. 2016.
3. Mariem Mahfoudh. Adapting ontologies with standard graph grammars: evolution and merger. Other [cs.OH]. Université de Haute-Alsace Mulhouse, 2015. NNT: 2015MULH1519, tel-01528579.
4. Ramdane Ahmed Merouane. Alignment of domain ontologies. 2019.
5. Catherine Roussey et al. Ontologies in agriculture. 2015.
6. Michel Héon. Semantic Web and Ontological Modeling with G-OWL: Java Developer's Guide Eclipse. Université du Québec à Montréal, 2014.
7. Corcho O., Fernández-López M., Gómez-Pérez A. Methodologies, tools and languages for building ontologies: where is their meeting point? Data and Knowledge Engineering, vol. 46, n°1, 2003, pp. 41-64. [https://doi.org/10.1016/S0169-023X\(02\)00195-7](https://doi.org/10.1016/S0169-023X(02)00195-7)
8. Berners-Lee, T. (2001). The Semantic Web. Scientific American, 284(5), 35-43. <https://doi.org/10.1038/scientificamerican0501-34>
9. Kaladzavi et al. (2018). Assessing similarity value between two ontologies.
10. Noy, N., Ferguson, R. W., Musen, M. A. (2000). The knowledge model of Protégé2000: combining interoperability and flexibility. Proceedings of the International Conference on Knowledge Engineering and Knowledge Management (EKAW'00). https://doi.org/10.1007/3-540-39967-4_2
11. Euzenat, J., Shvaiko, P. (2007). Ontology Matching. Springer-Verlag, Heidelberg (DE). <https://doi.org/10.1007/978-3-540-49612-0>
12. Euzenat, J. (2001). Towards a principled approach to semantic interoperability. Proceedings of the IJCAI Workshop on Ontology and Information Sharing, pp. 19-25, Seattle, US.
13. Euzenat, J., Bach, T.L., Barrasa, J., Bouquet, P., Bo, J.D., Dieng-Kuntz, R., Ehrig, M., Hauswirth, M., Jarrar, M., Lara, R., Maynard, D., Napoli, A., Stamou, G. (2006). State of the art on ontology matching. Deliverable, Knowledge Web NoE.
14. Stuckenschmidt, H., Shvaiko, P., Tessaris, S., Àcker, S.V., Zaihrayeu, I. (2004). State of the art on ontology alignment. Deliverable 2.2.3, IST Knowledge Web NoE, 80p., June 2004.
15. Amir Laadhar. Local matching learning of large scale biomedical ontologies. Technology for Human Learning. Université Paul Sabatier Toulouse III, 2019. NNT: 2019TOU30126.
16. Shvaiko, P., Euzenat, J. (2007). A survey of schema-based matching approaches. Journal on Data Semantics. https://doi.org/10.1007/11603412_5
17. Jaro, M. A. (1995). Probabilistic linkage of large public health data files. Statistics in Medicine, 14, 491-498. <https://doi.org/10.1002/sim.4780140510>
18. Khiat, Abderrahmane (2017). Developing an environment for ontology alignment: an instance-based approach.
19. Abdoulaye Diallo, Mouhamadou Thiam. State of the art on ontology alignment. 2021. hal-03330388.
20. Saadi, I., Bordji, Z. (2024). Systematic review of the literature on machine-learning methods for big data analysis, with a case study (Doctoral dissertation, University BBA).
21. Dilekh, Tahar (2019). A semantic model for Arabic language information retrieval. Ph.D. thesis, University of Batna 2.
22. Amin, M. B., Khan, W. A., Lee, S., Kang, B. H. (2015). Performance-based ontology matching: a data-parallel approach for effectiveness-independent performance gain in ontology matching. Applied Intelligence, 43, 356-385.
23. Achichi, Manel et al. Results of the Ontology Alignment Evaluation Initiative 2016. OM: Ontology Matching, no. 1766, 2016.
24. Hong, J. L., Siew, E. G., Egerton, S. (2010, May). Aligning data records using WordNet. In 2010 Second International Conference on Computer Research and Development, pp. 56-60. IEEE. <https://doi.org/10.1109/ICCRD.2010.79>

25. Alasoud, A., Haarslev, V., Shiri, N. (2007). A multi-level matching algorithm for combining similarity measures in ontology integration. In *Ontologies-Based Databases and Information Systems*, pp. 1-17. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-68123-6_63
26. He, Y., Chen, J., Dong, H., Horrocks, I., Allocca, C., Kim, T., Sapkota, B. (2024). DeepOnto: a Python package for ontology engineering with deep learning. *Semantic Web*, 15(5), 1991-2004. <https://doi.org/10.3233/SW-243568>
27. Li, N., Bailleux, T., Bouraoui, Z., Schockaert, S. (2024). Ontology completion with natural language inference and concept embeddings: an analysis. *arXiv:2403.17216*. <https://doi.org/10.1016/j.websem.2006.05.003>
28. Ring, J. (2009). Evolving an autonomous test and evaluation enterprise.
29. Lambrix, P., Tan, H. (2006). SAMBO a system for aligning and merging biomedical ontologies. *Journal of Web Semantics*, 4(3), 196-206.
30. Stucky, B., Luc, A. (2017). OntoPilot: new software to simplify and accelerate ontology development and deployment. *Biodiversity Information Science and Standards*.
31. Zaoui, Sayah (2023). Integration of semantics in big data. Doctoral thesis, Kasdi Merbah Ouargla University.
32. Sourty, Raphaël (2023). Learning to represent knowledge graphs and enriching language models pre-trained by knowledge graphs: distillation-based approaches. Doctoral thesis, Université Paul Sabatier Toulouse III.
33. Algergawy, A., Faria, D., Ferrara, A., Fundulaki, I., Harrow, I., Hertling, S., Jiménez-Ruiz, E., Karam, N., Khiat, A., Lambrix, P., Li, H., Montanelli, S., Paulheim, H., Pesquita, C., Saveta, T., Shvaiko, P., Splendiani, A., Thiéblin, É., Trojahn, C., Vataschinová, J., Zamazal, O., Zhou, L. Results of the Ontology Alignment Evaluation Initiative 2019. *OM@ISWC, CEUR Workshop Proceedings*, vol. 2536, pp. 46-85, 2019.
34. Li, H., Li, Y., Michel, F., Nasr, E., Paulheim, H., Pesquita, C., Portisch, J., Roussey, C., Saveta, T., Shvaiko, P., Splendiani, A., Trojahn, C., Vataschinová, J., Yaman, B., Zamazal, O., Zhou, L. Results of the Ontology Alignment Evaluation Initiative 2021. *OM@ISWC, CEUR Workshop Proceedings*, vol. 3063, pp. 62-108, 2021.
35. Algergawy, A. et al. Results of the Ontology Alignment Evaluation Initiative 2019.
36. Li, H. et al. Results of the Ontology Alignment Evaluation Initiative 2021.
37. Pour, M. A. N., Blomqvist, E., Cotovio, P. G., Coulet, A., Ferraz, L., Hertling, S., Jain, S., Jiménez-Ruiz, E., Kraus, F., Lambrix, P., Li, H., Li, Y., Liu, X., Monnin, P., Paulheim, H., Pesquita, C., Sharma, A., Shvaiko, P., Silva, M. et al. Results of the Ontology Alignment Evaluation Initiative 2025. *OM@ISWC, CEUR Workshop Proceedings*, 2025. <https://doi.org/10.5445/IR/1000189514>
38. OAEI 2014 Large BioMed Track results. Ontology Alignment Evaluation Initiative, 2014.
39. Mohammadi, M., Atashin, A. A., Hofman, W., Tan, Y.-H. SANOM results for OAEI 2019. *OM@ISWC, CEUR Workshop Proceedings*, vol. 2536, pp. 178-186, 2019. <https://doi.org/10.48550/arXiv.2006.05219>
40. Beltagy, I., Lo, K., Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. *Proceedings of EMNLP-IJCNLP 2019*, pp. 3615-3620, Hong Kong. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1371>
41. Yasunaga, M., Leskovec, J., Liang, P. (2022). LinkBERT: Pretraining Language Models with Document Links. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8003-8016, Dublin, Ireland. (Biomedical variant released as BioLinkBERT.) <https://doi.org/10.18653/v1/2022.acl-long.551>
42. Djimramadji, J., Warda, L., Mamai, K. A., Mbainaiibeye, J., Kaladzavi, G., Kolyang (2025). Hidden Markov Model framework for knowledge correspondence discovery and semantic interoperability across ontologies. *International Journal of Knowledge and Innovation Studies*, 3(3), 191-200. <https://doi.org/10.56578/ijkis030305>
43. Mamai, K. A., Warda, L., Kaladzavi, G., Kolyang (2026). Ontologies alignment towards a hybrid approach: a state of the art. *Journal of Information Technology and Computer Science (JITCS)*, 1(2), 1-11. Nicomarc International Publishers. <https://doi.org/10.5281/zenodo.2279623>