HAL open science

Representation and comparison of chemotherapy protocols with ChemoKG and graph embeddings

Jong Ho Jhee, Alice Rogier, Dune Giraud, Emma Pinet, Brigitte Sabatier, Bastien Rance, Adrien Coulet

To cite this version:

Jong Ho Jhee, Alice Rogier, Dune Giraud, Emma Pinet, Brigitte Sabatier, et al.. Representation and comparison of chemotherapy protocols with ChemoKG and graph embeddings. SWAT4HCLS 2024 - 15th International Semantic Web Applications and Tools for Health Care and Life Sciences Conferenc, Feb 2024, Leiden (NL), Netherlands. hal-04455155

HAL Id: hal-04455155

https://hal.science/hal-04455155v1

Submitted on 13 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Representation and comparison of chemotherapy protocols with ChemoKG and graph embeddings

Jong Ho Jhee[1,2], Alice Rogier[1,2,3], Dune Giraud[4], Emma Pinet[4], Brigitte Sabatier[1,2,4], Bastien Rance[1,2,3] and Adrien Coulet[1,2]

[1] *Inria Paris, F-75015 Paris, France*
[2] *Inserm, Centre de Recherche des Cordeliers, Université Paris Cité, Sorbonne Université, F-75006 Paris, France*
[3] *Department of Biomedical Informatics, Hôpital Européen Georges Pompidou, AP-HP, Paris, France*
[4] *Department of Pharmacy, Hôpital Européen Georges Pompidou, AP-HP, Paris, France*

Abstract

Background: Chemotherapy, a central cancer treatment, employs antineoplastic drugs to hinder cancer cell replication by disrupting DNA synthesis or mitosis. Chemotherapies follow complex protocols composed of cycles of treatment where antineoplastic and adjuvant drugs prescribed at different doses and times. Various protocols exist, with either small or large and numerous variations to others, making it hard to compare chemotherapies to each other, comparing their differential outcomes, and in the end choosing the most adapted one for a particular patient. ***Method:*** We propose ChemoKG, a knowledge graph for chemotherapy protocols that encompasses first administration programs such as drugs, dosages, treatment durations, and second drug properties and classes imported from ChEBI, DrugBank and the ATC classification. Three resources on drugs provide complementary hierarchies and chemical properties that help to better identify similar chemotherapy protocols. To this aim, we tested on ChemoKG a novel graph embedding method employing graph neural networks (GNNs) to compare nodes in the graph that represent protocols. Unlike previous approaches that focus on triple-based embeddings, the proposed method captures subgraph structures inherited from the aggregation scheme in GNNs. ***Results:*** The resulting knowledge graph encompasses 329,164 triples with 99,901 entities and 75 predicates including 1,358 chemotherapy protocols and 226 anti-cancer drugs. We performed a cluster analysis of protocol embeddings learned on ChemoKG, to propose groups of similar protocols. This will contribute in facilitating the comparison of chemotherapy themselves, and by extension to their potential effectiveness. Additionally, it should aid in analyzing gaps between commonly accepted protocols and their real-world implementation.

Keywords

Chemotherapy protocol, chemotherapy regimen, knowledge graph, graph embedding, clustering

1. Introduction

Chemotherapy is a cancer treatment which employs antineoplastic drugs to hinder cancer cell replication for instance by disrupting DNA synthesis or mitosis. Chemotherapy remains a cornerstone in cancer treatment, administering cytotoxic drugs to limit tumor growth. This involves a nuanced balancing between reducing tumor size and minimizing side effects. Combining various drugs in a timely manner, adapted to patient profile and response is a common strategy to achieve this tradeoff [1]. Indeed, each chemotherapy treatment follows a complex, but precisely defined protocol (also named regimen) composed of repeated cycles where a set of antineoplastic and adjuvant drugs are prescribed for administration with various dose, mode (continuous *vs.* bolus infusion) and timing. This cyclic approach is not arbitrary, it aligns with the life cycle of cancer cells and ensures optimal drug efficacy [2]. Many different protocols have been described with either small or large variations, to adapt to individual factors such as age, health

Proceedings Acronym: Proceedings Name, Month XX–XX, YYYY, City, Country
jong-ho.jhee@inria.fr (J. H. Jhee); alice.rogier-ext@aphp.fr (A. Rogier); dune.giraud@aphp.fr (D. Giraud); Emma.PINET@aphp.fr (E. Pinnet); brigitte.sabatier@aphp.fr (B. Sabatier); bastien.rance@aphp.fr (B. Rance); adrien.coulet@inria.fr (A. Coulet)

0000-0001-8887-8149 (J. H. Jhee); 0000-0002-5499-3197 (A. Rogier); 0000-0002-1466-062X (A. Coulet)
© 2023 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

conditions, and genetic profiles [3]. As a result, a large number of protocols co-exist in clinical information systems and expert databases, but they are suffering from unequal evaluation and consequently make more complicated for the clinician the choice of a protocol versus another.

The subtlety of variations in term for instance of timing (e.g., time lapse between two administrations), mode of administration (e.g., bolus *vs.* continuous) motivates the need for a low grain knowledge representation of protocols, especially for future studies aiming at evaluating the comparative effectiveness of treatment strategies [4]. In addition, such representation would enable the definition of distances between protocols, in regard to their multidimensional definition, composed of the drugs they leverage, their dose, timing, mode, classes, etc.

We introduce ChemoKG, a knowledge graph that represents chemotherapy protocols, their various dimensions, and links their constitutive drugs to various properties from ChEBI, DrugBank and the anatomical therapeutic chemical (ATC) classification. As a first illustration of the interest of ChemoKG, we propose here a clustering that group similar protocols in regard to their description in ChemoKG. The clustering relies on a novel embedding framework named RAGE, which leverages graph neural networks (GNNs) to learn a representation of protocols that considers the properties of the drug administrations present in the neighborhood of protocols in ChemoKG. Unlike other approaches that focus on triple-based embeddings, RAGE captures subgraph structures inherited from the aggregation scheme in GNNs. The clustering analysis provides a classification of protocols that we compare with two reference classifications: the first is based on cancer locations associated with protocols, the second on pharmacological groups of drugs. We evaluated RAGE embedding approach on a link prediction task; and RAGE outperforms or shows competitive performance against the selected baselines. The cluster analysis we performed with classical algorithms shows that RAGE allows for a reasonably good grouping of protocols by cancer locations, despite the fact that the graph does not contain this information. For ATC as a reference, RAGE showed the best result, in comparison to a classical method (not based on machine learning) named cumulative dose intensity (CDI).

To our knowledge, this is the first attempt to classify chemotherapy protocols on the basis of several of their features. We believe that our effort to compare chemotherapies will find applications first in the management of protocols in hospitals that historically recorded every small variation in protocols, resulting in large collections in need of structuration and cleaning; second in the definition of standard protocols as institutions have adopted different, but sometimes similar ones; and in the identification of concurrent protocols. Indirectly, we hope that comparing chemotherapy protocols will help in their relative evaluation and in the guidance for the choice for one among a set of similar ones.

The remainder of the paper is organized as follows. First, previous works on chemotherapy representation, graph embeddings and clustering from graph embeddings are presented in Section 2. Section 3 introduces ChemoKG. Section 4 describes the methodology of both the proposed graph embedding framework and its use for a clustering task. Section 5 presents our experimental results and is followed by elements of discussion and a conclusion in Section 6.

2. Related works

Chemotherapy databases The growing variety of chemotherapy protocols has led to a recent interest first in naming protocols in non-ambiguous ways [5], and second in proposing repositories of the various protocols. The larger available one is HemOnc, which includes >4,000 regimens [3]. It is a collaborative database that includes regimens description and general information about them. It started in 2011 through a collaboration of oncologists from several US University hospitals with an initial focus on the field of hematology cancers. HemOnc proposes a data schema to represent and share protocols in Owl. However, this schema does not include detailed properties of administrations and drugs. Another initiative developed in the UK is SACT [6] that contains both adult and pediatric oncology protocols. SACT has the particularity to store data not only about protocols, but also about patients, their diagnoses and outcomes. For this

reason, SACT is not shared in open access. Worth to note, a seminal work is DIOS [7], that consisted of 260 protocols at the time of publication (2013) and is not accessible anymore.

Knowledge graph embeddings The aim of knowledge graph embeddings is to project entities and relations into some continuous vector space while preserving the relation between entities [8]. Those entity and relation embeddings can further be used in downstream tasks, such as link prediction [9], triple classification [10] and entity clustering [11]. TransE [12] is a representative approach based on a translational distance model. Given a triple (s, r, o), the relation is interpreted as a translation vector r so that the embedded entities s and o can be connected by r with low error having $s + r \approx o$. TransE has the advantage of being simple, but has difficulty in learning 1-to-many and many-to-many relations [13]. DistMult [14] exploits a similarity-based scoring function to match the latent semantics of entities. It represents pairwise relations between entities in the vector space along the same dimension of relations. However, since relations between entities are over-simplified, the model consider all relations symmetric. ComplEx [15] is an extension of DistMult that uses complex-valued embeddings so as to better model asymmetric relations. MuRE [16] employs hyperbolic embedding instead of Euclidean analogues to represent hierarchical structures. RDF2Vec [17] uses random walks on the RDF graph to create sequences of entities, which are then used as input for the model. However, it is focused on embedding entities without considering the semantics of relations. CompGCN [35] represents relationships between entities using Graph Convolutional Networks (GCNs), which focuses on local neighborhood entities. It utilizes relation-type specific parameters to learn embeddings. More embedding techniques can be found in the following surveys [18, 19].

Clustering with graph embeddings Embeddings provide a representation of objects in the form of numeric vectors that is convenient for computing distances between them and consequently driving clustering analyses. To cite only few works from the biomedical domain, Monnin *et al*. [20] clustered embeddings of pharmacogenomic relationships, *Mohamed et al.* [21] clustered polypharmacy side-effects and Fernández-Torras *et al.* [22] clustered drugs, diseases and genes to predict drug responses.

3. ChemoKG

ChemoKG is an original knowledge graph in RDF (Resource Description Framework) of chemotherapy protocols. It encompasses 1,358 protocols by instantiating the ontology ChemoOnto [23, 24], which provides the necessary classes and relations to represent the various dimensions of protocols. As illustrated by Figure 1, this includes the administration program of a chemotherapy composed of its drugs, their dosages, the duration of their administration (bolus vs. continuous infusion), drug properties (such as their half-life) imported from ChEBI [25], DrugBank and their classification in the ATC classification. The protocols themselves have been extracted from a local database of the Pharmacy Service of the European Georges Pompidou Hospital of the AP-HP, Paris. The resulting knowledge graph encompasses 329,164 triples with 99,901 entities and 75 relations including 1,358 chemotherapy protocols and 226 anti-cancer drugs. ChemoKG includes both protocols used in standard treatments and in clinical studies. Because protocols evaluated by clinical studies are confidential, we defined a subset of our knowledge graph named ChemoKG-open that excludes them and that is consequently sharable. ChemoKG-open includes 513 protocols and is available on Zenodo at https://zenodo.org/records/10263831.[1] The statistics of the main classes of ChemoKG and ChemoKG-open are presented in Table 1.

[1] A SPARQL endpoint will be made available upon publication at https://chemokg.inria.fr.

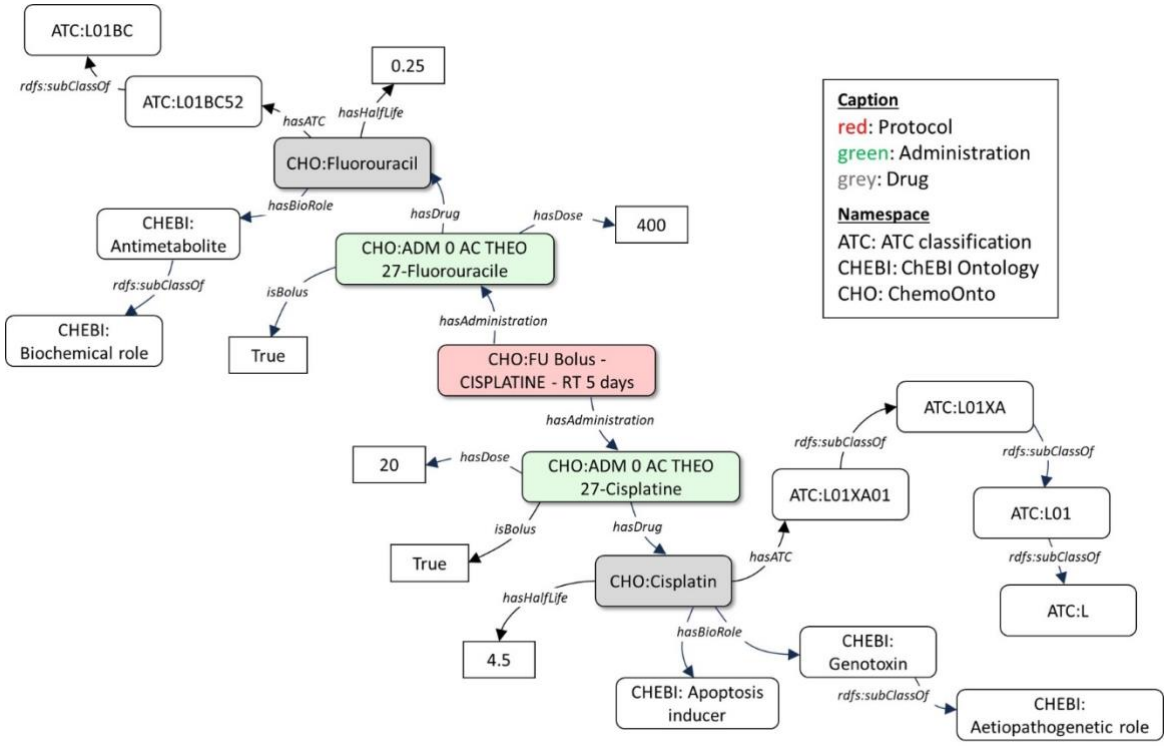

Figure 1: Sample from ChemoKG. The central entity is a protocol (in red), with two drug administrations (in green), each associated with its drug (in grey). Arrows describe a relation type from one entity to another. 'Dose', 'half life', 'bolus' are literals.

Table 1
Statistics of entities in ChemoKG and ChemoKG-open

Class	# entities (ChemoKG)	# entities (ChemoKG-open)	Source
Protocol	1,358	513	ChemoOnto
Anti-cancer drug	226	82	ChemoOnto
Bolus (True/False)	2	2	ChemoOnto
Dose	119	71	ChemoOnto
ATC	6,691	6,691	ATC
Biological role	13,286	13,286	ChEBI
Half-life	81	81	DrugBank

4. Method

In order to compare chemotherapy protocols, we propose first to learn embedding for protocols represented in ChemoKG, second to cluster similar protocols and third to compare the resulting clustering to two reference classifications: (*i*) protocols classified by cancer location, and (*ii*) by pharmacological and therapeutic subgroups.

Protocol embeddings We propose an original approach named RAGE, standing for Relation-Aware knowledge Graph Embedding, to compute node embeddings. Inspired from relation learning in [26, 27], RAGE builds on the GNN model to aggregate information from each entity's neighborhood, plus relations to characterize the type of links that connect the entity to its neighbors. An overview and naming of main variables involved in RAGE is depicted in Figure 2.

Let $\mathcal{G}$ be a knowledge graph such as $\mathcal{G} = \{(s, r, v) | s, v \in \mathcal{V}, r \in \mathcal{R}\}$, where $\mathcal{V}$ is a set of nodes, here named entities, $\mathcal{R}$ a set of labeled and oriented edges named relations or predicates and the triple (s, r, v) denotes that the entity s is related to the entity v through the relation r. For

example, the triple $(cisplatin, hasBioRole, genotoxin)$ indicates that "$cisplatin$ has a $biological$ $role$ $genotoxin$". The l-th layer embedding of a given entity s is formulated as:

$$e_s^l = \frac{1}{|\mathcal{N}_s|} \sum_{(r,v)\in\mathcal{N}_s} e_r^{l-1} \circ e_v^{l-1}, \tag{1}$$

where $\mathcal{N}_s = \{(r,v)|(s,r,v) \in \mathcal{G}\}$ is the neighborhood of the entity s, $l = \{1 \dots L\}$ is the number of layers and $\circ$ is the element-wise product. For L layers, the final embedding of an entity s is defined as the sum of each layer embedding:

$$e_s^* = e_s^0 + \cdots + e_s^L. \tag{2}$$

Accordingly, the final embedding is the sum of embeddings from each layer including the input e_d^0. In this way, we gather all the information of the target entity s and its "L-hop" neighbors. Because our task is to perform a clustering of protocols, it is important to consider drug property in the representation of protocols. For instance, 'ATC classes' or 'biological roles' can be captured within 3-hop neighbors (Figure 1).

Given a set of protocols $\mathcal{P} \subset \mathcal{V}$, the evaluation of how the relation between the protocol and a drug administration is likely is defined as follows:

$$\hat{y}_{pa} = {e_p^*}^T e_a^*. \tag{3}$$

If the final embeddings of protocol p and administration a derived from (2) are close (connected) to each other the evaluation value $\hat{y}_{pa}$ is high. i.e., if the drug administration a is part of a protocol p, $\hat{y}$ should be high and inversely low if the drug administration is not. We define the objective function using the Bayesian personalized ranking loss [30]:

$$\mathcal{L} = \sum_{(p,a,a')\in\mathcal{S}} -\ln\sigma\left(\hat{y}_{pa} - \hat{y}_{pa'}\right), \tag{4}$$

where $\mathcal{S} = \{(p,a,a')|(p,a) \in \mathcal{S}^+, (p,a') \in \mathcal{S}^-\}$ and σ is the sigmoid function. $\mathcal{S}^+$ is the set of protocol and administered drug pair and $\mathcal{S}^-$ is the set of protocol and non-administered drug pair. The loss is minimized when the likely score of the administered drug increases and the likely score of the non-administered drug decreases.

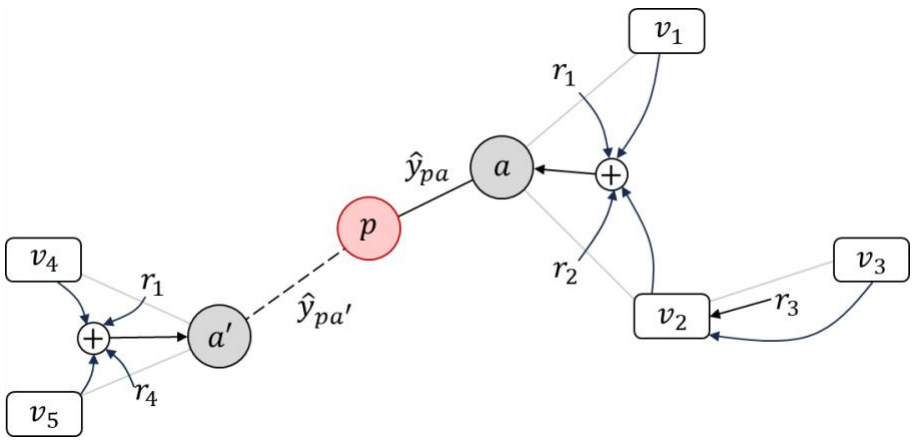

Figure 2: Overview of the relation-aware aggregation scheme in RAGE. p is the node we learn final embedding for, here a protocol. a and a' are nodes related and not related to p, respectively; here drug administrations. v_i are other nodes of the graph. r_i are predicates relating nodes. Plus signs represent the aggregative sum of embeddings. $\hat{y}_{pa}$ is an evaluation of the probability for nodes p and a to be linked.

Evaluation of embeddings on a link prediction task The task of link prediction aims at predicting potentially missing links between entities within a knowledge graph, on the basis of what is already stated in the graph. To compare RAGE with other graph embedding approaches, we evaluate their different capabilities in predicting "5as Administration" predicate between

protocols and drug administrations in ChemoKG. We particularly consider TransE [12], DistMult [14], MuRE [16], ComplEx [15] and CompGCN [35]. TransE and MuRE are translational distance models that aim at finding a vector representation of entities with relation to the translation of the entities based on distance measures. DistMult and ComplEx are semantic matching models that use similarity-based scoring functions. CompGCN is a GCN-based model which considers aggregating the neighborhood information for the entity relation embeddings. The models consider the relations between entities, but require adaptation to capture similarity between strings and numerical values that compose literals or distant relations in a graph. To compare the performances of these different models, we use the Mean Reciprocal Rank (MRR) and Hits@N (H@N). MRR evaluates models that return a ranked list of answers to queries by weighting results proportionally of their place in the ranking. H@N is the count of how many positive triples are ranked in the top-N positions against a set of negative triples.

Cluster computations and evaluation We cluster a subset of nodes $\mathcal{P} \subset \mathcal{V}$ of a knowledge graph on the basis of the Euclidean distance between their embeddings computed with a selection of two embedding approaches (RAGE included). In this exploratory study we compare performances of several classical clustering algorithms, namely *k*-means, Single and OPTICS [28, 29]. Both *k*-means and Single take as an input parameter the number of desired clusters. Single differs from *k*-means in that it is a hierarchical clustering algorithm that successively merges clusters whose distance between their closest observations is minimal. OPTICS is fundamentally different as it finds zones of high density and expands clusters from them. It takes as a main input parameter the minimal size of a cluster.

We evaluate our clusters in comparison to two referential classification of protocols. The first reference classification groups protocols by their primary indication *i.e.,* the cancer localization they primarily target according to our pharmacology experts. The second classification is based on the sets of 3rd level ATC classes of drugs involved in protocols. These classifications are available at https://chemokg.inria.fr, for protocols of ChemoKG-open.

Clustering analyses are evaluated with Adjusted Rand Index (ARI), Normalized Mutual Information (NMI) and Fowlkes-Mallows Index (FMI). ARI measures the overlapping between two clustering (or between one and a reference classification in our case). ARI equals 0 for a random labeling and 1 for an exactly similar labeling and is adjusted to limit the effect of chance. NMI measures the mutual information between two clustering, normalized by the entropy of each clustering. NMI is equal to 1 for an exactly similar labeling. FMI is the geometric mean of precision and recall. FMI ranges from 0 to 1 and a high value indicates a high similarity between the clustering and the reference classification. In addition, we compare our method based on graph embeddings to a state-of-the-art mean to compare chemotherapies, named the Cumulative Dose Intensity (CDI) [31]. The CDI is defined as a vector of normalized cumulative doses of administered drugs. It is used to compare the course of chemotherapies, as well as protocols.

5. Experimental results

The experiment was conducted within two steps. First, to perceive the effectiveness of the chemotherapy protocol representation we compared RAGE with other graph embedding methods on a link prediction task on ChemoKG. Second, to seek the features and patterns within the group of chemotherapy protocols clustering was performed with the embeddings of protocols obtained from the first step. This work is implemented with PyTorch [32], PyKeen [33] for graph embeddings and scikit-learn [34] for clustering.

5.1. Graph embeddings on ChemoKG

We compared the performance of RAGE and state-of-the-art approaches on a link prediction task. 10-fold cross validation was conducted on the triples in ChemoKG. The initialization of parameters was done using Xavier uniform initialization [36]. For RAGE, pre-trained vectors of

entities were initialized using TransE. Next, the learning of entity embeddings is continued using the RAGE model with three layers ($L = 3$). The number of layers used in CompGCN was also 3. All the baseline models were optimized using Adam [37], the learning rate was 0.01 with exponential decay and the output dimension of entities was 100.

The performance of RAGE and baseline methods on the link prediction are reported in Table 2. Overall, we observed that RAGE outperformed four baselines and was competitive to MuRE on all metrics. MuRE showed the best performance for all metrics in average. We observed that translational distance models (*i.e.*, MuRE) performed well in regard to semantic matching models (*i.e.*, DistMult, ComplEx) and CompGCN. The performance of RAGE is relatively lower than MuRE probably because it learns relations between protocols and administrations only rather than between all the entities and relations. This observation seems to also impact clustering results reported in Section 5.2.

Table 2
Performance of various considered models for the task of link prediction on ChemoKG.

Model	MRR	H@1	H@3	H@10
TransE [12]	0.1445±0.1340	0.0866±0.0610	0.2263±0.1152	0.3247±0.1759
DistMult [14]	0.2187±0.2182	0.1874±0.1874	0.2430±0.1333	0.2662±0.1239
ComplEx [15]	0.1064±0.0876	0.0730±0.0163	0.1013±0.0632	0.1402±0.0966
MuRE [16]	**0.3771±0.1632**	**0.3402±0.1307**	**0.4688±0.1221**	**0.5345±0.1424**
CompGCN [35]	0.3018±0.1485	0.1546±0.1542	0.4083±0.1918	0.4958±0.1625
RAGE [proposed method]	0.3538±0.1899	0.3307±0.1618	0.4538±0.2065	0.5116±0.1876

5.2. Chemotherapy protocol clustering

Clustering was performed on the embeddings of protocols obtained using RAGE and MuRE. Protocols assigned to the same cluster are expected to be similar in regard to their definition in ChemoKG. Results of our comparative study of three clustering algorithms and their ability to reflect our two reference classifications are shown in Table 3 and 4. For cancer locations, RAGE and the combination of CDI and RAGE showed better performance than CDI alone and MuRE. For ATC, RAGE still showed better performance than CDI and MuRE. We deduce that drug properties such as biological roles and half-lifes were beneficial for grouping protocols into cancer locations, which are absent from the graph. The ATC level information present in ChemoKG should be considered by RAGE what should explain the good grouping of protocols according to ATC classes.

Table 3
Performance of protocol clustering in comparison with a reference classification based on their primary indication (i.e., cancer locations). Parameter K is the number of clusters and S the minimum size of clusters.

	CDI			MuRE			RAGE			CDI+RAGE		
	ARI	NMI	FMI	ARI	NMI	FMI	ARI	NMI	FMI	ARI	NMI	FMI
K-means (K=32)	0.1045	0.4127	0.1553	0.0581	0.3484	0.1074	**0.4266**	**0.7150**	**0.4783**	0.3508	0.7135	0.4059
Single (K=32)	0.0429	0.3538	0.1924	0.0021	0.1730	0.2384	0.6153	0.7706	0.6433	**0.7628**	**0.8106**	**0.7862**
OPTICS (S=20)	0.0183	0.2198	0.1467	0.0046	0.1723	0.1138	**0.4946**	**0.7314**	**0.5706**	0.4402	0.6896	0.5444

Table 4
Performance of protocol clustering in comparison with a reference classification based on the pharmaceutical class of their drugs. Parameter K is the number of clusters and S the minimum size of clusters.

	CDI			MuRE			RAGE			CDI+RAGE		
	ARI	NMI	FMI	ARI	NMI	FMI	ARI	NMI	FMI	ARI	NMI	FMI
K-means (K=10)	0.1204	0.2156	0.2326	0.0741	0.1939	0.1891	**0.8420**	**0.9336**	**0.8651**	0.8281	0.9311	0.8542
Single (K=10)	0.0872	0.2419	0.3007	0.0059	0.0751	0.3717	**0.9509**	**0.9687**	**0.9591**	0.8071	0.9355	0.8503
OPTICS (S=20)	0.0436	0.1811	0.2466	0.0167	0.0438	0.3441	**0.7157**	**0.8662**	**0.7557**	0.5885	0.8126	0.6546

6. Discussion and conclusion

This paper is an initial attempt to evaluate the suitability of learning graph embeddings to classify and identify similar chemotherapy protocols, in a setting where many protocols are offered to clinicians, with potentially unequal levels of evaluation, consequently generating a clinical decision-making challenge.

From a knowledge representation and open resource point of view our knowledge graph is aligned with standard ontologies, but we would win in providing mapping to other initiatives, in particular HemOnc. This task is indeed not trivial as the HemOnc schema is different and less precise than ours. However, the graph embedding approach we described can spotlight highly similar protocols from HemOnc and ChemoKG, and in this sense has the potential of guiding the mapping between the two resources.

We acknowledge that the work presented here would gain from additional experiments. First, our graph of protocols encompasses some numerical values, such as drug dose and half-life. All graph embedding approaches we considered do not enable arithmetic comparisons of numerical values. This lets us think that approaches that enable such comparison, such as KEN [38] would lead to improvements. Also, we experimented recursive approaches (RAGE and CompGCN) only with $L = 3$ (*i.e.*, size of the considered neighborhood), and without including inferable links in the graph. Increasing L would enable embeddings to consider more of the graph, potentially leading to improvements. The inclusion of inferable links would add direct links between drugs and higher ATC or ChEBI ontology classes, what could help in identifying similar protocols. However, this extension is associated with a risk of flooding embeddings with general classes and is consequently to test with caution. Regarding the evaluation of the clustering, we only provided external evaluation metrics, *i.e.*, metrics that compare one clustering with a reference classification. This could be enriched with internal evaluation metrics such as the DUNN index or silhouette score that evaluate how many instances assigned to one cluster are both close to each other and distant from instances assigned to other clusters. Indeed, the two reference classifications are not a ground truth, but references one may want to compare to. In this setting, the internal quality of the clustering is a pertinent metrics, which could also enable the comparison of various clustering strategies.

Nonetheless, our experiments illustrate the level of performance of various graph embedding approaches on ChemoKG and their usability for learning meaningful clusters. We will pursue our efforts to facilitate the data and knowledge management associated with chemotherapy protocols and would like to expand our work from protocols to patient data, and study how protocols are respected, or modified to adapt to individuals and what is the impact on patient outcomes.

Contributions

JHJ populated ChemoOnto with protocols, enriched it with additional drug properties; co-designed the study; designed RAGE; implemented and ran the experiments; wrote the first version of this manuscript. AR created ChemoOnto and designed the instantiation of ChemoOnto with protocols; participated in the writing. DG, EP and BS guided the motivation of the work, provided with the two reference classifications. BR participated in the design of ChemoOnto, ChemKG and of this study, and in the writing. AC participated in the design of ChemoOnto, ChemKG, co-designed this study, and contributed to the writing.

Acknowledgements

This work is supported by Inria Paris and the CombO project. This work has benefited from a government grant managed by the Agence Nationale de la Recherche under the France 2030 program, reference ANR-22-PESN-0007, ShareFAIR and ANR-22-PESN-0008, NEUROVASC.

References

[1] DeVita Jr, V. T., & Chu, E. (2008). A history of cancer chemotherapy. Cancer research, 68(21), 8643-8653.

[2] Khongorzul, P., Ling, C. J., Khan, F. U., Ihsan, A. U., & Zhang, J. (2020). Antibody–drug conjugates: a comprehensive review. Molecular Cancer Research, 18(1), 3-19.

[3] Warner, J. L., Cowan, A. J., Hall, A. C., & Yang, P. C. (2015). HemOnc. Org: A collaborative online knowledge platform for oncology professionals. Journal of Oncology Practice, 11(3), e336-e350.

[4] Riaño, D., Peleg, M., & Ten Teije, A. (2019). Ten years of knowledge representation for health care (2009–2018): Topics, trends, and challenges. Artificial intelligence in medicine, 100, 101713.

[5] Rubinstein, S. M., Yang, P. C., Cowan, A. J., & Warner, J. L. (2020). Standardizing chemotherapy regimen nomenclature: a proposal and evaluation of the HemOnc and National Cancer Institute Thesaurus Regimen Content. JCO Clinical Cancer Informatics, 4, 60-70.

[6] Bright, C. J., Lawton, S., Benson, S., Bomb, M., Dodwell, D., Henson, K. E., ... & Smittenaar, R. (2020). Data resource profile: the systemic anti-cancer therapy (SACT) dataset. International journal of epidemiology, 49(1), 15-15l.

[7] Klimes, D., Smid, R., Kubásek, M., Vyzula, R., & Dusek, L. (2013). DIOS-database of formalized chemotherapeutic regimens. In EFMI-STC (pp. 165-169).

[8] Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12), 2724-2743.

[9] Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web, 8(3), 489-508.

[10] Socher, R., Chen, D., Manning, C. D., & Ng, A. (2013). Reasoning with neural tensor networks for knowledge base completion. Advances in neural information processing systems, 26.

[11] Gad-Elrab, M. H., Stepanova, D., Tran, T. K., Adel, H., & Weikum, G. (2020, November). Excut: Explainable embedding-based clustering over knowledge graphs. In International Semantic Web Conference (pp. 218-237). Cham: Springer International Publishing.

[12] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. Advances in neural information processing systems, 26.

[13] Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014, June). Knowledge graph embedding by translating on hyperplanes. In Proceedings of the AAAI conference on artificial intelligence (Vol. 28, No. 1).

[14] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

[15] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016, June). Complex embeddings for simple link prediction. In International conference on machine learning (pp. 2071-2080). PMLR.

[16] Balazevic, I., Allen, C., & Hospedales, T. (2019). Multi-relational poincaré graph embeddings. Advances in Neural Information Processing Systems, 32.

[17] Ristoski, P., & Paulheim, H. (2016). Rdf2vec: Rdf graph embeddings for data mining. In The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part I 15 (pp. 498-514). Springer International Publishing.

[18] Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12), 2724-2743.

[19] Ji, S., Pan, S., Cambria, E., Marttinen, P., & Philip, S. Y. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. IEEE transactions on neural networks and learning systems, 33(2), 494-514.
[20] Monnin, P., Raïssi, C., Napoli, A., & Coulet, A. (2022). Discovering alignment relations with Graph Convolutional Networks: A biomedical case study. *Semantic Web*, *13*(3), 379-398.
[21] Mohamed, S. K., Nounu, A., & Nováček, V. (2021). Biological applications of knowledge graph embedding models. Briefings in bioinformatics, 22(2), 1679-1693.
[22] Fernández-Torras, A., Duran-Frigola, M., Bertoni, M., Locatelli, M., & Aloy, P. (2022). Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the Bioteque. Nature Communications, 13(1), 5304.
[23] Rogier, A., Rance, B., Coulet, A. (2023). ChemoOnto, an ontology to qualify the course of chemotherapies. Bio-ontologies COSI 2023, Poster.
[24] ROGIER, A., Rance, B., & Coulet, A. (2024, January 22). ChemoOnto, an ontology to qualify the course of chemotherapies. ISMB-ECCB 2023, Lyon. Zenodo. https://doi.org/10.5281/zenodo.10548491
[25] Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., ... & Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic acids research, 44(D1), D1214-D1219.
[26] Monnin, P., Raïssi, C., Napoli, A., & Coulet, A. (2022). Discovering alignment relations with Graph Convolutional Networks: A biomedical case study. *Semantic Web*, *13*(3), 379-398.
[27] Wang, X., Huang, T., Wang, D., Yuan, Y., Liu, Z., He, X., & Chua, T. S. (2021, April). Learning intents behind interactions with knowledge graph for recommendation. In Proceedings of the web conference 2021 (pp. 878-887).
[28] Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster analysis. John Wiley & Sons.
[29] Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. ACM Sigmod record, 28(2), 49-60.
[30] Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2012). BPR: Bayesian personalized ranking from implicit feedback. arXiv preprint arXiv:1205.2618.
[31] Longo, D. L., Duffey, P. L., DeVita Jr, V. T., Wesley, M. N., Hubbard, S. M., & Young, R. C. (1991). The calculation of actual or received dose intensity: a comparison of published methods. J Clin Oncol, 9(11), 2042-2051.
[32] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.
[33] Ali, M., Berrendorf, M., Hoyt, C. T., Vermue, L., Galkin, M., Sharifzadeh, S., Fisher, A., Tres, V. & Lehmann, J. (2021). Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12), 8825-8845.
[34] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al.. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.
[35] Vashishth, S., Sanyal, S., Nitin, V., & Talukdar, P. (2019). Composition-based multi-relational graph convolutional networks. arXiv preprint arXiv:1911.03082.
[36] Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249-256). JMLR Workshop and Conference Proceedings.
[37] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
[38] Cvetkov-Iliev, A., Allauzen, A., & Varoquaux, G. (2023). Relational data embeddings for feature enrichment with background information. *Machine Learning*, *112*(2), 687-720.