Knowledge Enhanced Representation Learning for Drug Discovery

Thanh Lam Hoang[1], Marco Luca Sbodio[1], Marcos Martinez Galindo[1], Mykhaylo Zayats[1]
Raul Fernandez-Diaz[1,3], Victor Valls[1], Gabriele Picco[1], Cesar Berrospi[2], Vanessa Lopez[1]
[1] IBM Research, Dublin research lab, Dublin, Ireland
[2] IBM Research, Zurich research lab, Zurich, Switzerland
[3] University College Dublin, Ireland
Emails: {t.l.hoang, marco.sbodio, vanlopez}@ie.ibm.com, ceb@zurich.ibm.com
{Marcos.Martinez.Galindo, raulfd, mykhaylo.zayats1, victor.valls}@ibm.com

Abstract

Recent research on predicting the binding affinity between drug molecules and proteins use representations learned, through unsupervised learning techniques, from large databases of molecule SMILES and protein sequences. While these representations have significantly enhanced the predictions, they are usually based on a limited set of modalities, and they do not exploit available knowledge about existing relations among molecules and proteins. Our study reveals that enhanced representations, derived from multimodal knowledge graphs describing relations among molecules and proteins, lead to state-of-the-art results in well-established benchmarks (first place in the leaderboard for Therapeutics Data Commons benchmark "Drug-Target Interaction Domain Generalization Benchmark", with an improvement of 8 points with respect to previous best result). Moreover, our results significantly surpass those achieved in standard benchmarks by using conventional pre-trained representations that rely only on sequence or SMILES data. We release our multimodal knowledge graphs, integrating data from seven public data sources, and which contain over 30 million triples. Pretrained models from our proposed graphs and benchmark task source code are also released.

Introduction

Developing a concise representation of proteins and small molecules is a crucial task in AI-based drug discovery. Recent studies (Rao et al. 2019; Rives et al. 2021) have focused on utilizing large databases of protein sequences or molecules for self-supervised representation learning. These representations are then fine-tuned using limited labeled data for tasks like predicting the binding affinity between drugs and targets. In the field of protein representation learning, (Zhang et al. 2022) and (Zhou et al. 2023) have demonstrated that enhancing protein representations with additional information from knowledge graphs (KGs), such as comprehensive textual data from the gene ontology (GO) (Ashburner et al. 2000), can enhance the performance of pre-trained representations on various protein properties and protein-protein interaction tasks. Despite promising initial findings, important research directions remain unexplored.

Firstly, it is important to investigate if joint robust representations of proteins, molecules and other associated entities (such as diseases) may improve downstream prediction tasks. A joint representation of each of these entities should take into account not only the intrinsic properties of the entity itself, but also its relations with other entities. In contrast to previous studies (Zhang et al. 2022; Zhou et al. 2023), while their focus was narrow, concentrating on protein sequences, GO terms, and text descriptions, our work encompasses a wider array of modalities—text, numbers, protein sequences, SMILES (string representation of molecular graphs), categories, and entities like drugs, diseases, and pathways.

Secondly, previous studies (Zhang et al. 2022; Zhou et al. 2023) assumed that the graphs or datasets used for training were carefully combined into a single source. However, in our research, we deal with different KGs built using several datasets from various sources. It's worth highlighting that merging these datasets into a single source is complex due to the difficulty of aligning their structures automatically. For example, in the STITCH knowledge graph (Szklarczyk et al. 2016), the connection labeled "interaction with", denoting the relationship between chemicals and proteins, is established by analyzing their co-occurrence within a Pubmed abstract. This form of association might bear resemblance to the "target of" relation found in the Uniprot knowledge graph, however confirming their equivalence presents a challenging task because co-orcurrence in a Pubmed abstract does not mean a protein is a target of a drug. Moreover, creating a comprehensive graph by combining multiple ones demands significant computational resources. To address these issues, we propose an approach based on ensemble methods, which can effectively capture information from separate and unaligned KGs.

Finally, while there are datasets for studying the problem of drugs-proteins interactions, there are no publicly available multimodal knowledge graphs for this purpose. We aim to provide the research community with such graphs, by selectively integrating knowledge from existing datasets.

The contributions of this work are summarised as follows:

- We release to the research community a set of multimodal knowledge graphs, and a set of models pretrained on those graphs using graph neural networks to derive protein and drug representations[1]. We hope that this will

[1] https://github.com/IBM/otter-knowledge

foster further research in the area.

- We provide experimental results showing that a multimodal knowledge enhanced representation surpasses the state of the art for predicting drugs-proteins interactions, even for challenging discovery scenario, where the majority of the entities in the test data (e.g., molecules) are unseen in the training data and might only have one available modality (e.g, its SMILES). Previous work focused on enhancing Protein Language Models (PLMs) with textual descriptions, and did not exploit KGs that capture heterogenous relations between entities and multimodal entity attributes (including molecules SMILES, proteins sequences, textual, numerical and categorical attributes).
- We tackle the problem of learning from partially connected multimodal knowledge graphs via ensemble methods, yielding promising outcomes. We also study how our approach remains robust across diverse pretraining objectives, such as regression and link prediction.

Multimodal Knowledge Representation Learning Framework

The diagram in Figure 1 illustrates the overall process of our system named *Otter-Knowledge*. This process involves constructing multimodal knowledge graphs from diverse sources, generating initial embeddings for each modality using pretrained models available in the model zoo, and subsequently improving the embeddings by incorporating information from the knowledge graphs through the utilization of graph neural networks (GNN). We discuss each component of the given system in the following subsections.

Multimodal Knowledge Graph Construction

A Multimodal Knowledge Graph (MKG) is a directed labeled graph where labels for nodes and edges have well-defined meanings, and each node has a modality, a particular mode that qualifies its type (text, image, protein, molecule, etc.). We consider two node subsets: *entity nodes* (or entities), which corresponds to concepts in the knowledge graph (for example protein, or molecule), and *attribute nodes* (or attributes), which represent qualifying attributes of an entity (for example the mass of a molecule, or the description of a protein). We refer to an edge that connects an entity to an attribute as *data property*, and an edge that connects two entities as *object property*. Each node in the graph has a unique identifier, and a unique modality (specified as a string).

We developed a framework for automating the construction of a multimodal knowledge graph by extracting and fusing data from a variety of sources, including text delimited files, JSON, and proprietary data sources (Staar, Dolfi, and Auer 2020). The framework takes as input a schema file (specified in JSON), which declaratively describes how to build the desired graph from a set of data sources.

The framework that builds the MKG ensures that each triple is unique, and it automatically merges entities having the same unique identifier, but whose data is extracted from different data sources. It is also possible to use the special relation `sameAs` [2] to indicate that two entities having different unique identifiers are to be considered as the same entity. The `sameAs` relation is useful when creating a MKG from multiple partially overlapping data sources; when the graph is built. Additionally, it is possible to build an MKG incrementally, by merging two or more graphs built using different schemas; The merge operation automatically combines entities with matching unique identifiers or distinct attributes (e.g., proteins with the same sequence).

The framework builds the graph in memory, but can also provide support for building the graph using a database on disk; the graph triples can also be serialised using GML [3] or any RDF [4] serialization formats. Finally, the framework offers scalable parallel and GPU-based computation of initial node embeddings, tailored to each node's modality.

Computing Initial Embeddings

As explained before, the MKG contains nodes representing entities and nodes representing attributes of those entities. In the MKG, each node has a modality assigned, e.g: entity nodes can have a modality *Protein* or *Drug*, nodes containing text could have a modality *text* or protein sequence. We assign a model for each one of the modalities in our graphs, as specified by the user in the *schema*. The models, referred to as *handlers*, are capable of preprocessing the values in the nodes and computing their initial embeddings. Our framework allows to easily retrieve all the nodes in the graph with the same modality to efficiently compute the initial embeddings with the assigned *handler*, facilitating parallelization, GPU utilization, batching, and avoiding the need to load different models in memory simultaneously.

Handlers are exclusively assigned to nodes representing attributes. For each modality, only one *handler* is being used, although it is possible to change the *handler* assigned. For instance, for SMILES it is possible to use *morgan-fingerprint* or *MolFormer*. The comparison between different models is not within the scope of this work. These are the *handlers* that we have used for computing the initial embeddings of the graph:

- *morgan-fingerprint* We use the Morgan fingerprint (Rogers and Hahn 2010) in RDKit[5] for processing the SMILES. We employ a shape of 2048 (nBits) and a radius of 2. If RDKit fails to generate the fingerprint, we supply an embedding of the same shape filled with zeros.
- *MolFormer* (Ross et al. 2022) is a large-scale chemical language model designed with the intention of learning a model trained on small molecules which are represented as SMILES strings. *MolFormer* leverges Masked Language Modeling and employs a linear attention Transformer combined with rotary embeddings.
- *protein-sequence-mean* For the protein sequences, we use the mean token embeddings extracted from the 33rd

[2] We borrow the semantic of `owl:sameAs` - see https://www.w3.org/TR/owl-ref/\#sameAs-def

[3] https://en.wikipedia.org/wiki/Graph_Modelling_Language

[4] https://www.w3.org/RDF/

[5] https://www.rdkit.org/docs/index.html

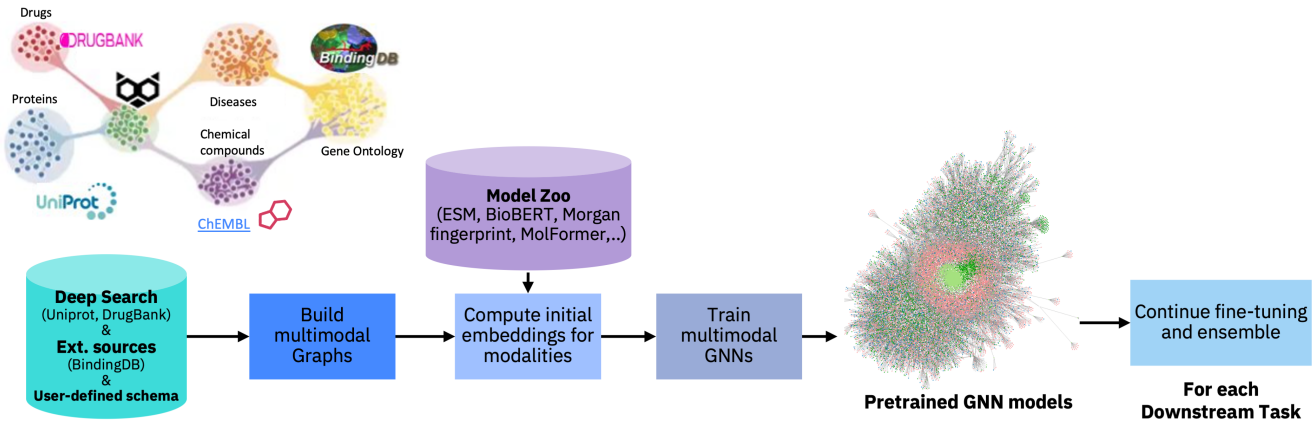

Figure 1: *Otter-Knowledge* workflow.

layer of the *esm1b_t33_650M_UR50S* model (Rives et al. 2021). All sequences are truncted to 1022 amino acids.

- *text* For the textual values, we use the 'sentence-transformers/paraphrase-albert-small-v2' (Reimers and Gurevych 2019) [6].
- *number* In the case of numbers, we do not use any model to get initial embeddings. We convert the numerical value to a torch tensor and use it as embedding.
- *categorical and entities* nodes depicting entity modalities or categories that do not possess handlers, resulting in the absence of an initial embedding. Initial embeddings start as zeros and improve through GNN training, based on the neighborhood connections to its data properties and related entities (e.g, for an entity Protein that could be its sequence, family, function, interactions with other proteins, etc., provided such a modality or relations exist).

Using distinct models lets us leverage existing competitive pretrained language models for meaningful modality representations (e.g., gauging embedding similarity). This aids the GNN in exploiting cross-modal links via KG geometry, enhancing generalization, negating the necessity for a larger dataset for unseen entities in a single model setup.

Pretraining with Inductive R-GNN

GNN architecture To improve representations of MKG entities we assimilate initial embeddings with the information about the MKG structure, i.e., connection patterns between graph nodes. To this end we train GNN (Zhou et al. 2020) that propagates initial embeddings of attribute nodes through a set of layers that transform input embedding according to the entity data and object properties. Our GNN is inductive since it may compute entities embeddings only from their available neighbours and the corresponding initial embeddings. This enables computing embeddings for unseen nodes during training.

The architecture of GNN adheres to a standard design and consists of two main blocks: encoder and decoder. For the encoder, we first define a projection layer, which consists of a set of linear transformations for each attribute node modality and projects those into a common dimensionality. Then, we apply several multi-relational graph convolutional layers (R-GCN) (Schlichtkrull et al. 2018), which distinguish between different types of edges connecting source and target nodes by having a set of trainable parameters for each edge type. For the decoder, we consider link prediction task between entities and regression tasks for numerical data properties. We also considered skip-connections (He et al. 2016) in GNN layers to provide shortcut paths for the gradients to flow directly through the deep network.

Learning objective For link prediction, we evaluate three common scoring functions: DistMult, TransE, and a Binary Classifier. We compare their scores against actual labels using negative log likelihood loss. Additionally, for numerical attributes, we include a regression objective to minimize root mean square error (MSE) of predicted numerical properties. The learning process combines the link prediction and regression objectives into a single function. GNN hyperparameters values are detailed in the Appendix.

Negative sampling To train the GNN for link prediction we need to provide both positive and negative examples. While positive examples come from the MKG itself, negative triples are generated synthetically. To achieve this, for each relation type we extract a set of admissible source and target nodes, subsequently, we randomly sample sources and targets from the corresponding admissible sets. We use an equal ratio between positive and negative links.

Scaling the GNN training Due to the integration of data from various sources, the size of the integrated data can become significantly large. To address this, we employ a graph auto-scaling approach (GAS) described in (Fey et al. 2021). This method divides the graph into smaller partitions using Metis[7]. GAS performs training on each partition separately, so it is able scale to arbitrarily large graphs. It avoids information loss due to the connection between partitions via

[6]https://huggingface.co/sentence-transformers/paraphrase-albert-small-v2

[7]https://github.com/KarypisLab/METIS

booking-keep historical embedding of one-hop nodes which have a direct connection toward to nodes inside the partition. We also considered an alternative to GAS, where instead of keeping historical embedding of one-hop nodes, the partition is enhanced with data properties of the one-hop nodes.

Model versus representation fine-tuning we experimented both cases when only representation of drugs/proteins given by the models are transferred to the downstream tasks versus fine-tuning the projection layers of the GNN for the drug/proteins.

Ensemble learning The pretrained representation is sensitive to various factors, such as the chosen objectives for the GNN and the specific graphs used for training. Additionally, combining all the datasets into a single large graph requires substantial computational resources and poses challenges in aligning different databases. In this study, we propose a simple approach that involves combining pretrained embeddings obtained from different settings and KGs. This allows us to train multiple GNN models simultaneously without the need to merge all KGs into a single location, resulting in savings in computational resources and effort required for data and schema alignment. We form a linear ensemble by assigning equal weights to downstream models, each transferred/fine-tuned with distinct GNN models.

Pretraining Dataset Collection

Table 1 summarises the KGs used for pretraining the models. In this subsection, we discuss data preprocessing methods. We open-sourced these preprocessed KGs, for details readers can refer to the github repository.

Uniprot (Consortium 2022) comprises 573,227 SwissProt proteins (from curated UniProt subset). The UBC KG combines all UniProt proteins with diverse attributes, including sequence (567,483 entries), full name, organism, protein family, function, catalytic activity, pathways and length. The KG also features 38,665 *target_of* edges linking UniProt IDs to ChEMBL and Drugbank IDs, along with 196,133 interactants connecting UniProt protein IDs.

BindingDB (Liu et al. 2007) consists of 2,656,221 data points involving 1.2 million compounds and 9,000 targets. In additional to the BingdingDB KG with affinity score, we create triples for every drug-protein pair. This results in a final set of 2,232,392 triples included in the UBC KG.

ChEMBL (Gaulton et al. 2011) consists of drug-like bioactive molecules, in the UBC KG we included 10,261 ChEMBL molecule IDs with their corresponding SMILES downloaded from OpenTargets (Ochoa et al. 2022), from which 7,610 have a *sameAs* link to Drugbank ID molecules.

Drugbank (Law et al. 2013) consists of detailed chemical data on 9,749 drugs (such as SMILES, description, indication, mechanism of action, organism, average mass, toxicity and other calculated and experimental properties), drug pathways and 1,301,422 drug interactions. Due to licenses restrictions in UBC we only release drugbank IDs linked from other sources, such as Uniprot and ChemBL.

DUDe (Mysinger et al. 2012) comprises a collection of 22,886 active compounds and their corresponding affinities towards 102 targets. For our study, we utilized a preprocessed version of the DUDe (Sledzieski et al. 2022), which includes 1,452,568 instances of drug-target interactions. To prevent any data leakage, we eliminated the negative interactions and the overlapping triples with the TDC DTI dataset. This yielded 40,216 pairs of drug-target interactions.

PrimeKG (the Precision Medicine Knowledge Graph) (Chandak, Huang, and Zitnik 2023) integrates 20 biomedical resources, it describes 17,080 diseases with 4 million relationships. PrimeKG includes nodes describing Gene/Proteins (29,786) and Drugs (7,957 nodes). Our MKG holds 13 types of modalities and 12,757,300 edges. These edges consist of 154,130 data properties and 12,603,170 object properties. Among them, 642,150 edges depict protein interactions, 25,653 edges represent drug-protein interactions, and 2,672,628 edges capture drug interactions.

STITCH (Search Tool for Interacting Chemicals) (Szklarczyk et al. 2016) is a database of known and predicted interactions between chemicals represented by SMILES strings and proteins whose sequences are taken from STRING database (Szklarczyk et al. 2023). For the MKG curation we filtered only the interaction with highest confidence, i.e., the one which is higher 0.9. This resulted into 10,717,791 triples for 17,572 different chemicals and 1,886,496 different proteins. Furthermore, the graph is split into 5 subgraphs of roughly similar size so the GNN is trained sequentially on each of them, building on previous subgraph's model.

Experiments

Here we summarise the main results from our experiments. Further details are available in out github repo[8].

Downstream benchmarks To evaluate the performance of the proposed approach on drug-target binding affinity prediction task we use three datasets: DTI DG, DAVIS and KIBA, which are available from the TDC (Huang et al. 2022) benchmark. The DTI DG dataset features a leaderboard with the state-of-the-art metrics reported for different methods. The dataset's temporal split, based on patent application dates, making this dataset suitable for evaluating method generalization. In contrast, the DAVIS and KIBA datasets employ random splits, including two additional splits based on target or drug. These latter splits assess learning methods with new drugs/proteins. To avoid data leakage we removed overlapping triples between the test sets of downstream data and the pretrained KGs.

Downstream models TDC framework adapts the DeepDTA approach (Öztürk, Özgür, and Ozkirimli 2018) for drug protein binding affinity prediction model referred to as the downstream model. The provided approach is suboptimal when both drug and protein embeddings are provided as input. We compared the original downstream model with a simplified variant that uses only concatenation and transformation blocks. Using the original model with ESM protein

[8]https://github.com/IBM/otter-knowledge

Datasets	# triples	Entities	Modalities	Data license	Released
UBC	6,207,654	Proteins/Drugs	sequences, SMILES, text, number	Open	Yes
PrimeKG	12,757,257	Proteins/Drugs/Diseases	sequences, SMILES, text	Open	Yes
DUDe	40,216	Proteins/Drugs	sequences, SMILES	Open	Yes
STITCH	12,621,873	Proteins/Chemicals	sequences, SMILES	Open	Yes
BindingDB	7,855,545	Proteins/Chemicals	sequences, SMILES	Open	Yes

Table 1: Summary of the KGs for pretraining models with the number of triples used for GNN training.

	Downstream	**DTI DG**	**DAVIS**			**KIBA**		
Upstream	Splits	Temporal	Random	Target	Drug	Random	Target	Drug
	Leaderboard	0.538	NA	NA	NA	NA	NA	NA
	Baseline	0.569	0.805	0.554	**0.264**	0.852	0.630	0.576
	Otter DistMult	0.578	0.808	0.572	0.152	0.859	0.627	0.593
UBC	Otter TransE	0.577	0.807	0.571	0.130	0.858	0.644	0.583
	Otter Classifier	0.580	0.810	0.573	0.104	0.861	0.631	0.616
	Otter DistMult	0.577	0.805	0.573	0.132	0.857	0.650	0.607
DUDe	Otter TransE	0.576	0.807	0.570	0.170	0.858	0.653	0.604
	Otter Classifier	0.579	0.808	0.574	0.167	0.860	0.641	0.630
	Otter DistMult	0.575	0.806	0.571	0.162	0.858	0.611	0.617
PrimeKG	Otter TransE	0.573	0.807	0.568	0.186	0.858	0.642	0.607
	Otter Classifier	0.576	0.813	0.576	0.133	0.861	0.630	0.635
	Otter DistMult	0.575	0.808	0.573	0.138	0.859	0.615	0.603
STITCH	Otter TransE	0.578	0.814	0.572	0.119	0.859	0.636	0.635
	Otter Classifier	0.576	0.804	0.571	0.156	0.856	0.627	0.585
	Otter Ensemble	**0.588**	**0.839**	**0.578**	0.168	**0.886**	**0.678**	**0.638**

Table 2: Results of knowledge enhanced representation transfer on three standard drug-target binding affinity prediction benchmarks datasets with different splits. Evaluation utilized Pearson correlation, and results, including pretraining on distinct upstream datasets and an ensemble of models, were reported.

embeddings and Morgan fingerprints for molecule SMILES inputs yielded inferior results (0.539, the baseline in Table 2) compared to the simplified downstream model (0.569) on the DTI DG dataset. This led us to devise a new model architecture for evaluating pretrained embeddings. We augmented the existing network, which combines ESM and Morgan fingerprints, with a parallel network that combines and transforms GNN embeddings. The final binding affinity prediction is obtained by summing the outputs from both networks. In the model fine-tuning experiment, we replicated the weights of the GNN's projection layers and continued fine-tuning them to predict the binding affinity target.

Knowledge enhanced representation versus vanilla representation In Table 2, we present the outcomes of the *Otter-Knowledge* models, which were pretrained on graphs generated from Uniprot (U), BindingDB (B), ChemBL (C), DUDe, PrimeKG and STITCH with three different training objectives: TransE, DistMult, and binary classifier respectively. Also, as previously described, we control the information propagated to the Drug/Protein entities, and manually handpick a subset of links from each database that are relevant to drug discovery. In all these methods, we started with the initial embeddings of sequences using ESM-1b models, while Morgan fingerprints were utilized for SMILES. We call this baseline method vanilla representation as oppose to methods utilizing knowledge-enhanced representations. The embeddings were then fine-tuned with knowledge graphs. Our results demonstrate that *Otter-Knowledge* outperforms the baseline vanilla representation that lacks the enhanced knowledge from the graphs. Notably, a significant improvement was observed when we created an ensemble of 12 models trained on UBC, DUDe, PrimeKG and STITCH. We achieved state-of-the-art results on the leaderboard[9] of the DTI DG dataset (Table 2). Nevertheless, the enhancements were not significant for the drug split in the KIBA dataset. As shown in Table 4, the KIBA dataset comprises only 68 drugs. The challenge arises due to the limited number of drugs in this split; however, fine-tuning the GNN layers, as opposed to embedding transfer, resolves this issue, as illustrated in Table 3.

Model fine-tuning Table 3 illustrates the outcomes achieved through the continued fine-tuning of the projection layers in pretrained GNNs models. These models were initially trained without the use of GAS but included skip connections. Moreover, skip connections were integrated into the downstream models which results in a stronger baseline model compared to the baseline model presented in Table 2, attributed to the incorporation of skip connections into the downstream models. Importantly, fine-tuning consistently outperforms representation transfer, and the ensem-

[9]https://tdcommons.ai/benchmark/dti_dg_group/bindingdb_patent/

Experiment	**DTI DG**	**DAVIS**			**KIBA**		
Splits	Temporal	Random	Target	Drug	Random	Target	Drug
Baseline	0.557	0.840	0.559	0.215	0.857	0.644	0.679
UBC	0.512	0.844	0.575	0.170	0.862	0.648	0.684
BindingDB	0.600	0.829	0.569	0.232	0.858	0.563	0.683
STITCH	0.546	0.838	0.587	0.217	0.865	0.637	0.642
PrimeKG	0.547	0.839	0.573	0.188	0.857	0.634	0.676
DUDe	0.534	0.762	0.572	0.223	0.806	0.594	0.522
Otter ensemble	**0.617**	**0.867**	**0.632**	**0.290**	**0.891**	**0.687**	**0.712**

Table 3: Results when the projection layers of the pretrained GNNs are fine-tuned on downstream tasks.

Datasets	**DTI DG**	**DAVIS**	**KIBA**
# triples	232460	27621	118036
# drugs	140745	68	2068
# proteins	569	379	299
Type of splits	Temporal	Random/Drug/Target	Random/Drug/Target

Table 4: TDC benchmark datasets and statistics.

ble approach yields markedly improved outcomes across all datasets and splits compared to the new baseline models.

Information flow control and Noisy link prediction In GNN pretraining, bridging the gap between initial and later task data is crucial. Initially, diverse attributes are considered, but in fine-tuning, only specific data (protein sequences and SMILES) is available. Our experiments explore controlled information flow to Drug/Protein entities to understand its impact on subsequent tasks. Additionally, we consider noisy links in upstream data and manually select relevant links for drug discovery, comparing outcomes when training with these restricted links versus all potentially noisy links.

Table 5 shows the results of *Otter-Knowledge* for UBC when (i) we do *not* control the information that is propagated to Drug/Protein entities, (ii) we do *not* cherry-pick a subset of links from each database that are relevant to the downstream task, (iii) regression for numerical data properties is added to the objective in addition to link prediction. Observe from the table that the results are similar to the results in Table 2, with minor variations across different scoring functions and datasets. Notably, Otter Classifier with noisy links (N) and no information flow control (C) achieves comparable or even better performance than when we cherry-pick links and control the flow of information (Table 2). Slight variations indicate GNN embeddings' resilience to irrelevant noisy triples for downstream tasks. Incorporating regression objectives and information flow control doesn't seem to impact generalization or significantly improve benchmark results.

Related Work

We review methods to learn an effective representation from proteins, molecules and their interactions.

Representation learning for proteins and small molecules Representation learning focuses on encoding essential information about entities, such as proteins or molecules, into low-dimensional tensors. Self-supervised algorithms using language models (LMs) have achieved remarkable success in learning protein and molecule representations by training on extensive datasets of protein sequences or linear serialization of small molecules, such as SMILES. State of the art examples of transformer-based protein language models (PLMs) are TAPE (Rao et al. 2019), ProteinLM (Xiao et al. 2021), ProteinBERT (Brandes et al. 2022), ESM (Rives et al. 2021), Prottrans (Elnaggar et al. 2020), and MSA (Rao et al. 2021). They are typically trained on masked reconstruction – they learn the likelihood that a particular amino acid appears in a sequence context. Because the probability that a residue will be conserved or not across related sequences is intrinsically tied to its biological role, existent PLMs can capture co-evolutionary and inter-residue contact information (Rao et al. 2019; Rives et al. 2021), and have shown impressive performance on various tasks, such as predicting protein structure (Lin et al. 2022) and function (Rao et al. 2019). Regarding small molecules, their molecular structure can be condensed into linear notations like SMILES or SELFIES. LMs have also been used to interpret these representations, e.g., MolFormer (Ross et al. 2022), MolBERT (Fabian et al. 2020), SmilesFormer (Owoyemi and Medzhidov 2023) or SELFormer (Yüksel et al. 2023). Both Protein and molecular representations have been fine-tuned using a contrastive learning co-embedding by Conplex (Singh et al. 2022) achieving good performance in Drug-Target Interaction (DTI) prediction, surpassing state of the art approaches in the TDC-DG leaderboard which evaluates of out-of-domain generalisation (Huang et al. 2021) and achieving high specificity while detecting false positive bindings in "decoy" datasets like DUD-E.

Knowledge enhanced pre-trained language models for proteins LMs do not consider existent extensive knowledge, in the form of manually curated functional and structural annotations, in human-curated domain datasets and

Datasets (UBC)	**DTI DG**	**DAVIS**			**KIBA**		
Splits	Temporal	Random	Target	Drug	Random	Target	Drug
Otter DistMult (C)	0.575	0.809	0.571	0.126	0.861	**0.643**	0.617
Otter TransE (C)	0.576	0.809	0.570	**0.157**	0.858	0.632	0.585
Otter Classifier (C)	0.578	0.814	**0.577**	0.097	0.861	0.633	0.631
Otter DistMult (N+C)	0.578	0.809	0.574	0.105	0.862	**0.643**	0.615
Otter TransE (N+C)	0.579	0.809	0.573	0.108	0.857	0.633	0.583
Otter Classifier (N+C)	0.580	**0.816**	**0.577**	0.147	**0.864**	0.639	**0.641**
Otter DistMult (N+C+R)	0.579	0.810	0.572	0.145	0.862	0.629	0.625
Otter TransE (N+C+R)	0.580	0.811	0.576	0.073	0.859	0.627	0.594
Otter Classifier (N+C+R)	**0.582**	0.812	0.574	0.124	0.860	0.619	0.600

Table 5: Information flow control and noisy links results for UBC for different scoring functions. The table results should be compared with the results in Table 2 (UBC). N (noisy links); C (no flow control); R (regression).

effectively leveraging all this available factual knowledge to enhance representation learning is an open challenge. Nonetheless, prior research indicates that it can improve results in downstream learning tasks. OntoProtein (Zhang et al. 2022) fine-tuned a PLM by reconstructing masked amino acids while minimizing the embedding distance between the contextual representation of proteins and associated gene ontology (GO) functional annotations (Central et al. 2023). For this purpose they built ProteinKG25, a KG consisting of 600k protein sequences and nearly five million triples. Their results show that the representations obtained where useful for classification tasks such as protein-protein interaction type, protein function, and contact prediction; but underperformed in regression tasks like homology, fluorescence, and stability. KeAP (Zhou et al. 2023) employs a token-level approach, using non-masked amino acids to iteratively query knowledge tokens from Gene Ontology (GO) for restoring masked amino acids via cross-attention. It focuses solely on mask token reconstruction, while OntoProtein combines contrastive learning and masked modeling. Trained on ProteinKG25 (Zhang et al. 2022), KeAP outperforms OntoProtein on nine downstream tasks.

Graph-based approaches for therapeutics Graphs are a natural way to represent molecular interactions, signalling pathways, and disease co-morbidities. They can also be used for representation learning as they allow for the distillation of high-dimensional information about a node's neighborhood into low-dimensional vector spaces. The training objective of these representations is that similar neighborhoods are embedded close to each other in the vector space. Optimised representations can then be used to train downstream models to predict properties of specific nodes (e.g., protein function), as well as, novel associations between nodes (e.g., drug-target interactions). An overview on graph representation learning in biomedicine can be found in (Li, Huang, and Zitnik 2022). State-of-the-art approaches have shown that incorporating multiple knowledge sources improves downstream performance. For example, DTiGEMS+ (Thafar et al. 2020) formulates the prediction of DTIs as a link prediction problem in an heterogeneous graph. TxGNN (Huang et al. 2023) predicts drug indications/contraindications for rare diseases using the PrimeKG heterogeneous and multimodal KG (Chandak, Huang, and Zitnik 2023). We're the first to demonstrate that combining knowledge graphs (KGs) for protein and SMILES sequence representation boosts drug-target interaction (DTI) tasks on discovery benchmarks (Huang et al. 2021), where most test entities are unseen in training data.

Conclusion and Future Work

In this paper, we studied representation learning for multimodal knowledge graphs fused from multiple sources for drug discovery. Our study lays the foundation for future investigations in this area with the release of multimodal KGs selectively integrating data from several datasets, and pretrained models constructed with these KGs. Our models can be utilized to acquire representations specifically tailored for drug discovery applications. Furthermore, they can serve as benchmarks for comparing and evaluating against other representation learning techniques. We have also publicly released the standard evaluation framework for assessing pretrained representations in drug-target binding affinity prediction. We extensively analyzed different representation learning methods on three established drug-target binding datasets. Our approach outperformed existing methods, achieving state-of-the-art results on the TDC DG dataset.

Nevertheless, numerous unresolved research directions remain unexplored. Firstly, the inclusion of additional modalities, such as the 3D structure of molecules or proteins, can provide valuable insights for representation learning. Secondly, a substantial challenge lies in effectively handling a vast number of datasets, where aligning them into a single multimodal KG is not a straightforward task. Developing a learning approach capable of accommodating the dynamic input schema from diverse sources is a crucial problem to address. Finally, assessing how representations generalize across tasks, and developing robust learning methods for representation generalization across multiple tasks under changing data distributions, remain as key research goals.

Acknowledgements

We would like to thank Michele Dolfi and Peter Staar for their assistance in preparing datasets for the construction of knowledge graphs.

References

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S. E.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; and Sherlock, G. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics*.

Brandes, N.; Ofer, D.; Peleg, Y.; Rappoport, N.; and Linial, M. 2022. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8): 2102–2110.

Central, G.; Aleksander, S. A.; Balhoff, J.; Carbon, S.; Cherry, J. M.; Drabkin, H. J.; Ebert, D.; Feuermann, M.; Gaudet, P.; Harris, N. L.; et al. 2023. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1).

Chandak, P.; Huang, K.; and Zitnik, M. 2023. Building a knowledge graph to enable precision medicine. *Nature Scientific Data*.

Consortium, T. U. 2022. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1): D523–D531.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; and Rost, B. 2020. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *CoRR*, abs/2007.06225.

Fabian, B.; Edlich, T.; Gaspar, H.; Segler, M.; Meyers, J.; Fiscato, M.; and Ahmed, M. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*.

Fey, M.; Lenssen, J. E.; Weichert, F.; and Leskovec, J. 2021. Gnnautoscale: Scalable and expressive graph neural networks via historical embeddings. In *International Conference on Machine Learning*, 3294–3304. PMLR.

Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; and Overington, J. P. 2011. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1): D1100–D1107.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, K.; Chandak, P.; Wang, Q.; Havaldar, S.; Vaid, A.; Leskovec, J.; Nadkarni, G.; Glicksberg, B.; Gehlenborg, N.; and Zitnik, M. 2023. Zero-shot Prediction of Therapeutic Use with Geometric Deep Learning and Clinician Centered Design. *medRxiv*.

Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C.; Xiao, C.; Sun, J.; and Zitnik, M. 2021. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; and Zitnik, M. 2022. Artificial intelligence foundation for therapeutic science. *Nature chemical biology*.

Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B.; Zhou, Y.; and Wishart, D. S. 2013. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, 42(D1): D1091–D1097.

Li, M. M.; Huang, K.; and Zitnik, M. 2022. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering*, 1–17.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*.

Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; and Gilson, M. K. 2007. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1): D198–D201.

Mysinger, M. M.; Carchia, M.; Irwin, J. J.; and Shoichet, B. K. 2012. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*, 55(14): 6582–6594.

Ochoa, D.; Hercules, A.; Carmona, M.; Suveges, D.; Baker, J.; Malangone, C.; Lopez, I.; Miranda, A.; Cruz-Castillo, C.; Fumis, L.; Bernal-Llinares, M.; Tsukanov, K.; Cornu, H.; Tsirigos, K.; Razuvayevskaya, O.; Buniello, A.; Schwartzentruber, J.; Karim, M.; Ariano, B.; Martinez Osorio, R.; Ferrer, J.; Ge, X.; Machlitt-Northen, S.; Gonzalez-Uriarte, A.; Saha, S.; Tirunagari, S.; Mehta, C.; Roldán-Romero, J.; Horswell, S.; Young, S.; Ghoussaini, M.; Hulcoop, D.; Dunham, I.; and McDonagh, E. 2022. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Research*, 51(D1): D1353–D1359.

Owoyemi, J.; and Medzhidov, N. 2023. SmilesFormer: Language Model for Molecular Design.

Öztürk, H.; Özgür, A.; and Ozkirimli, E. 2018. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17): i821–i829.

Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, P.; Canny, J.; Abbeel, P.; and Song, Y. 2019. Evaluating Protein Transfer Learning with TAPE. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Rao, R. M.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; Sercu, T.; and Rives, A. 2021. MSA Transformer. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8844–8856. PMLR.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in*

Natural Language Processing. Association for Computational Linguistics.

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C.; Ma, J.; and Fergus, R. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15).

Rogers, D.; and Hahn, M. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5): 742–754.

Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; and Das, P. 2022. Molformer: Large Scale Chemical Language Representations Capture Molecular Structure and Properties.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; van den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling Relational Data with Graph Convolutional Networks. In Gangemi, A.; Navigli, R.; Vidal, M.-E.; Hitzler, P.; Troncy, R.; Hollink, L.; Tordai, A.; and Alam, M., eds., *The Semantic Web*, 593–607. Cham: Springer International Publishing.

Singh, R.; Sledzieski, S.; Cowen, L.; and Berger, B. 2022. Learning the Drug-Target Interaction Lexicon. *bioRxiv*.

Sledzieski, S.; Singh, R.; Cowen, L.; and Berger, B. 2022. Adapting protein language models for rapid DTI prediction. *bioRxiv*, 2022–11.

Staar, P. W. J.; Dolfi, M.; and Auer, C. 2020. Corpus processing service: A Knowledge Graph platform to perform deep data exploration on corpora. *Applied AI Letters*, 1(2): e20.

Szklarczyk, D.; Kirsch, R.; Koutrouli, M.; Nastou, K.; Mehryary, F.; Hachilif, R.; Gable, A.; Fang, T.; Doncheva, N.; Pyysalo, S.; Bork, P.; Jensen, L. J.; and von Mering, C. 2023. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1): D638–D646.

Szklarczyk, D.; Santos, A.; von Mering, C.; Jensen, L. J.; Bork, P.; and Kuhn, M. 2016. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic acids research*, 44(D1): D380–D384.

Thafar, M.; Olayan, R.; Ashoor, H.; Albaradei, S.; Bajic, V.; Gao, X.; Gojobori, T.; and Essack, M. 2020. DTiGEMS+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *Journal of Cheminformatics*, 12(1).

Xiao, Y.; Qiu, J.; Li, Z.; Hsieh, C.; and Tang, J. 2021. Modeling Protein Using Large-scale Pretrain Language Model. *CoRR*, abs/2108.07435.

Yüksel, A.; Ulusoy, E.; Ünlü, A.; Deniz, G.; and Doğan, T. 2023. SELFormer: Molecular Representation Learning via SELFIES Language Models. *arXiv preprint arXiv:2304.04662*.

Zhang, N.; Bi, Z.; Liang, X.; Cheng, S.; Hong, H.; Deng, S.; Zhang, Q.; Lian, J.; and Chen, H. 2022. OntoProtein: Protein Pretraining With Gene Ontology Embedding. In *International Conference on Learning Representations*.

Zhou, H.-Y.; Fu, Y.; Zhang, Z.; Cheng, B.; and Yu, Y. 2023. Protein Representation Learning via Knowledge Enhanced Primary Structure Reasoning. In *The Eleventh International Conference on Learning Representations*.

Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; and Sun, M. 2020. Graph neural networks: A review of methods and applications. *AI Open*, 1: 57–81.