Predicting protein and pathway associations for understudied dark kinases using pattern-constrained knowledge graph embedding

Mariah V. Salcedo[1,*], Nathan Gravel[2,*], Abbas Keshavarzi[3], Liang-Chin Huang[2], Krzysztof J. Kochut[3] and Natarajan Kannan[1,2]

[1] Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, United States of America
[2] Institute of Bioinformatics, University of Georgia, Athens, GA, United States of America
[3] School of Computing, University of Georgia, Athens, GA, United States of America
[*] These authors contributed equally to this work.

Corresponding authors
Krzysztof J. Kochut,
kkochut@uga.edu
Natarajan Kannan, nkannan@uga.edu

ABSTRACT

The 534 protein kinases encoded in the human genome constitute a large druggable class of proteins that include both well-studied and understudied "dark" members. Accurate prediction of dark kinase functions is a major bioinformatics challenge. Here, we employ a graph mining approach that uses the evolutionary and functional context encoded in knowledge graphs (KGs) to predict protein and pathway associations for understudied kinases. We propose a new scalable graph embedding approach, RegPattern2Vec, which employs regular pattern constrained random walks to sample diverse aspects of node context within a KG flexibly. RegPattern2Vec learns functional representations of kinases, interacting partners, post-translational modifications, pathways, cellular localization, and chemical interactions from a kinase-centric KG that integrates and conceptualizes data from curated heterogeneous data resources. By contextualizing information relevant to prediction, RegPattern2Vec improves accuracy and efficiency in comparison to other random walk-based graph embedding approaches. We show that the predictions produced by our model overlap with pathway enrichment data produced using experimentally validated Protein-Protein Interaction (PPI) data from both publicly available databases and experimental datasets not used in training. Our model also has the advantage of using the collected random walks as biological context to interpret the predicted protein-pathway associations. We provide high-confidence pathway predictions for 34 dark kinases and present three case studies in which analysis of meta-paths associated with the prediction enables biological interpretation. Overall, RegPattern2Vec efficiently samples multiple node types for link prediction on biological knowledge graphs and the predicted associations between understudied kinases, pseudokinases, and known pathways serve as a conceptual starting point for hypothesis generation and testing.

INTRODUCTION

Protein kinases play a fundamental role in cell signaling by phosphorylating peptides and proteins on serine, threonine, and tyrosine residues. The 534 protein kinases encoded in the human genome regulate protein activity in diverse pathways including those involved in DNA repair, cell cycle control, and metabolism (*Hunter, 2009*; *Johnson et al., 2023*; *Lahiry et al., 2010*; *Ochoa et al., 2020*). Because dysregulation of protein kinase signaling is causally associated with the pathogenesis of many diseases (*Ardito et al., 2017*; *Oprea et al., 2018*), much effort has gone into the characterization of their physiological functions (*Ayala-Aguilera et al., 2022*; *Xie et al., 2021*) and the development of selective protein kinase inhibitors (*Ardito et al., 2017*; *Attwood et al., 2021*; *Xie et al., 2021*). Despite these efforts, nearly 164 out of the 534 known kinases remain relatively understudied and are referred to as "dark" kinases by the NIH Illuminating the Druggable Genome consortium (IDG) (*Berginski et al., 2021*; *Gillespie et al., 2022*; *Kelleher et al., 2023*). Characterizing the functions of these dark kinases is crucial because they work in conjunction with other well-studied kinases in signaling pathways and are also frequently mutated, or abnormally expressed, in human diseases such as cancers (*Berginski et al., 2021*; *Brognard & Hunter, 2011*; *Collins et al., 2018*; *Ravanmehr et al., 2021*; *Soleymani et al., 2022*). While considerable progress has been made in illuminating the functions of several dark kinases, placing these kinases in a pathway or a cell signaling network context remains a major bioinformatics challenge.

The compendium of genomic, proteomic and interactome data now available through the efforts of the IDG consortium and numerous investigator-initiated efforts allows for the possibility of inferring the functions and pathways for dark kinases through integrative mining of the known patterns and relationships in existing data. In particular, the development of network-based approaches, such as knowledge graphs (KGs), that relate and link diverse types of protein kinase data in the forms of networks (composed of nodes and edges) enables the prediction of dark kinase functions or pathways through network context (*Soleymani et al., 2022*). Indeed, KG mining approaches and machine learning (ML) methods applied to KGs have been successfully employed in the identification of kinase substrates (*Gavali et al., 2022*; *Nováček et al., 2020*), prioritization of understudied kinases (*Huang et al., 2018*), and identification of disease associations (*Bachman, Gyori & Sorger, 2022*; *Gyori et al., 2017*; *Moret et al., 2021*; *Soleymani et al., 2022*).

The working premise of these network-based approaches is that proteins (kinases in this case) that share similar network neighbors (such as interacting proteins, downstream substrates, cellular localization, or molecular functions) with well-studied kinases are likely to be involved in common pathways. Several predictive ML methods utilize this premise, however many of these methods are designed for use with homogenous networks (*Grover &*

Leskovec, 2016; *Perozzi, Al-Rfou & Skiena, 2014*). Homogenous networks have just a single node and edge type. Whereas heterogenous networks, like KGs, are defined by multiple node and relationship types such as *Drug–targets–Protein* and *Protein–isAssociatedWith– DiseaseState,* where *Drug* and *Protein* are node types and *targets* and *isAssociatedWith* are edge types (or relationships). Because heterogenous networks offer a broader range of knowledge representation over homogenous networks, they are suitable for biological representations that require data integration and conceptualization from multiple data sources. However, models that can fully leverage heterogenous graphs often suffer from scalability issues, limiting their predictive power and making the construction of the KG difficult (*Alshahrani, Thafar & Essack, 2021*; *Bonner et al., 2022*; *Gavali et al., 2022*). Due to these limitations, KGs have not been previously employed to predict pathway associations for dark kinases because they are sparsely connected in KGs (due to low information content), making it difficult to predict functional associations through local context alone.

Embedding-based approaches have been proposed to overcome the limitations of sparse data within KGs, as they have the capability to represent the graph in dense and low dimensional feature space (*Dai et al., 2020*). However, many seminal methods developed for graph embedding are not optimized for heterogenous graphs as is the case with many message-passing GNN-based models (*Hamilton, Ying & Leskovec, 2017*; *Pei et al., 2020*; *Tang, Li & Yu, 2019*; *Velickovic et al., 2017*; *Zhang et al., 2018*; *Zhang et al., 2019*). Other methods, such as proximity-preserving methods, have had success in large-scale knowledge graph embedding, however using these methods to fully capture complex relations is still an active area of research as many of these models lack full expressivity (accounting for both local and global dependencies) (*Bordes et al., 2013*; *Gao et al., 2020*; *Ge et al., 2022*; *Huang et al., 2021*; *Sadeghian et al., 2021*; *Sun et al., 2019*; *Wang et al., 2021*; *Yang & Liu, 2021*; *Zhou, Yi & Jia, 2021*). However, a recent advancement, the spatio-translational model BoxE, is fully expressive, but due to the model's use of the "rule injection" method, yields high false positivity rates with sparse graphs (*Abboud et al., 2020*).

Relation-preserving methods attempt to capture complex relationships by matching semantic similarities of entities and relations and learn "context" from nodes within the KG. Many base-line relation-preserving methods are still widely used for link prediction tasks with biological data (*Chen et al., 2020*; Gan et al. 2023; *Ha & Park, 2022*; *Li et al., 2020*; *Long & Luo, 2020*; *Peng, Guan & Shang, 2019*; *Samizadeh & Minaei-Bidgoli, 2020*; *Wang et al., 2021*; *Wong et al., 2020*). This deep learning approach is based on a family of models from Natural Language Processing (NLP) called word2vec and can be combined with constrained random walks to sample the graph and allow for use with heterogenous networks (*Dong, Chawla & Swami, 2017*). By sampling subsections of the KG, random walk-based methods can scale with larger KGs that would otherwise be too computationally expensive to use in full. While many techniques are used for data reduction in KGs (*Wang et al., 2017*), constrained random walks have distinct advantages. Using a constrained "random walk" method on the KG, the graph is stochastically "sampled" such that multiple data types present within the network can be appropriately leveraged for predictions. The sampled section of the graph can then be converted into an embedding of the KG such that the biological context relevant for the prediction is accurately captured enabling nontrivial

associations between dark kinase and associated pathways (*Dong, Chawla & Swami, 2017*; *Nickel et al., 2015*).

One of the most well-known relation-preserving methods that utilizes random walks is metpath2vec. metapath2vec preserves the relationships in the graph during the sampling process by utilizing schemas (ordered sets of specific node or edge types). By specifying what node/edge types should be sampled, integration of different data types can be fully leveraged for predictions, thus accurately capturing the structure of heterogenous graphs. However, creating such meta-paths for complex heterogeneous KGs, often with no well-defined schema, is challenging and time-consuming with the performance of the model being highly dependent on the chosen meta-path (series of nodes).

Several previous attempts to automatically generate meta-paths have been reported in the literature (*Meng et al., 2015*; *Wan et al., 2020*; *Yang et al., 2018*) but they all rely on fixed-length meta-paths, which prevents accurate representation of latent and hierarchical structure encoded in large KGs. Other works have expanded upon the random walk framework proposed in metapath2vec, including W-MetaPath2Vec (*Pham & Do, 2019*), HIN2Vec (*Fu, Lee & Lei, 2017*), and RW-k (*Anil et al., 2019*). Additional works have combined the use of meta-path-based sampling with graph neural networks (GNNs) to make the aggregated message passing method more appropriate with heterogenous networks, as is the case with MAGNN (*Fu et al., 2020*). Finally, attempts to exploit semantics explicit in structural relations found within the graph have been leveraged for graph sampling, as is the case with Relation Structure-Aware Heterogenous Information Network Embedding (RHINE) (*Shi et al., 2020*), which defines relations that link nodes according to their similar properties or ability to bridge compatible nodes. We explore several of the methods listed above (MAGNN, RHINE, BoxE) for use with our graph (See Results 'Benchmarking RegPattern2Vec's predictive performance with metapath2vec and other graph embedding approaches'). However due to the focus of our question relating to the functionality of dark kinases there is inherent data imbalance within the graph making these methods not appropriate for use.

To address this, we employ RegPattern2Vec (*Keshavarzi, Kannan & Kochut, 2021*), a novel approach for learning on semantic data to predict new pathway associations for dark kinases. RegPattern2Vec directly addresses the limitations of previous models and provides more accurate predictions by precisely capturing relevant node/edge context in areas of graph sparsity using regular patterns and hyperparameters (See Methods & Materials 'Regular pattern selection and its usage in random walks'–Biased random walk constrained by a regular pattern). Unlike metapath2vec, RegPattern2Vec can make node-association predictions without specifying the entire schema. This allows us to meaningfully sample the KG without establishing multiple meta-paths (*Keshavarzi, Kannan & Kochut, 2021*). Here we show the application of RegPattern2Vec for link prediction on a large heterogenous KG.

Overall, we successfully employed RegPattern2Vec to capture the node/edge context more accurately within our KG. After sampling, the ML technique (link prediction) is used to predict new associations between dark kinases and characterized pathways. We place 34 dark kinases for which we have consistent, high-confidence predictions produced through

replicate runs (with differing hyperparameters) in a pathway context and based on an analysis of the meta-paths navigated for three selected dark kinases, we provide biological interpretations of the predicted pathway associations. The RegPattern2Vec pipeline is available at the GitHub repository (https://github.com/gravelCompBio/RegPattern2Vec).

MATERIALS & METHODS

Knowledge graph architecture

KGs are very similar to Heterogeneous Information Networks. An Information Network is a directed graph $G = (V, E)$, composed of vertices (also called nodes) and edges, with an associated node type mapping function $\phi : V \to A$ and an edge type mapping function $\psi : E \to R$ (*Shi et al., 2016*). Each node $v \in V$ belongs to one particular node type in the node type set $A : \phi(v) \in A$, and each edge $e \in E$ belongs to a particular edge type in the edge type set $R : \psi(e) \in R$. An Information Network is called a Heterogeneous Information Network if the sets A and R both contain more than one element, that is, there are multiple labels (types) for graph nodes and multiple labels (types) for edges. If sets A and R are singletons, the Information Network is called a Homogeneous Information Network (all nodes in the network are of the same type, and all edges are of the same type). While Heterogeneous Information Networks require that if two edges belong to the same edge type, the two edges share the same starting node type and ending node type, KGs do not. That is, the same edge (relation) type can be applied to different starting node types and different ending node types. This is also the case in the Resource Description Framework (RDF) (*Cyganiak et al., 2014*; *W3C, 2014*), a notation often used to represent KGs. RDF is based on the notion of triples of the form *Subject-predicate-Object* (*Cyganiak et al., 2014*) representing edges connecting nodes in the graph.

For example, in Fig. 1A, for a KG representing information about protein kinases, a node (entity) representing a dark protein kinase *CDK* 13 (cyclin dependent kinase 13) can be connected by an edge (relationship) labeled *hasPathway* to a node representing a pathway *NeutrophilDegranulation*, which represents the knowledge that *CDK* 13 participates in the pathway *NeutrophilDegranulation*. In such a KG, *CDK* 13 may be connected to other nodes using different labels (*hasPathway*), such as *CDK13–has*Molelcular*Function—CyclinBinding*, *CDK13–hasBiologicalProcess—GranulocyteActivation* and *CDK13–hasCellularComponent—GolgiApparatus*. Other protein kinases can additionally be included in the neighborhood context through shared nodes, such as the shared molecular function of *CyclinBinding* between dark kinase *CDK13* and light kinase *CDK6* (cyclin-dependent kinase 6). Here, edges have multiple labels, and destination nodes are of different types, which indicates that it is a heterogeneous KG.

Given a Heterogeneous Information Network $G = (V, E)$ (as defined above), the network's schema is a directed graph, $S = (A, R)$, based on G's node type mapping $\phi : V \to A$ and its edge (relation) type mapping $\psi : E \to R.S$ is a directed graph defined over node types A, with edges as relations from R. Similarly, the schema of a KG represented in RDF is represented in RDFS (RDF Schema (*W3C, 2014*)). For example, given the *CDK13* examples above, the schema would contain the edges *Protein —participatesIn—Pathway*, *Protein–hasBiologicalProcess—BiologicalProcess,* and *Protein—isLocatedIn—CellularLocation*. The
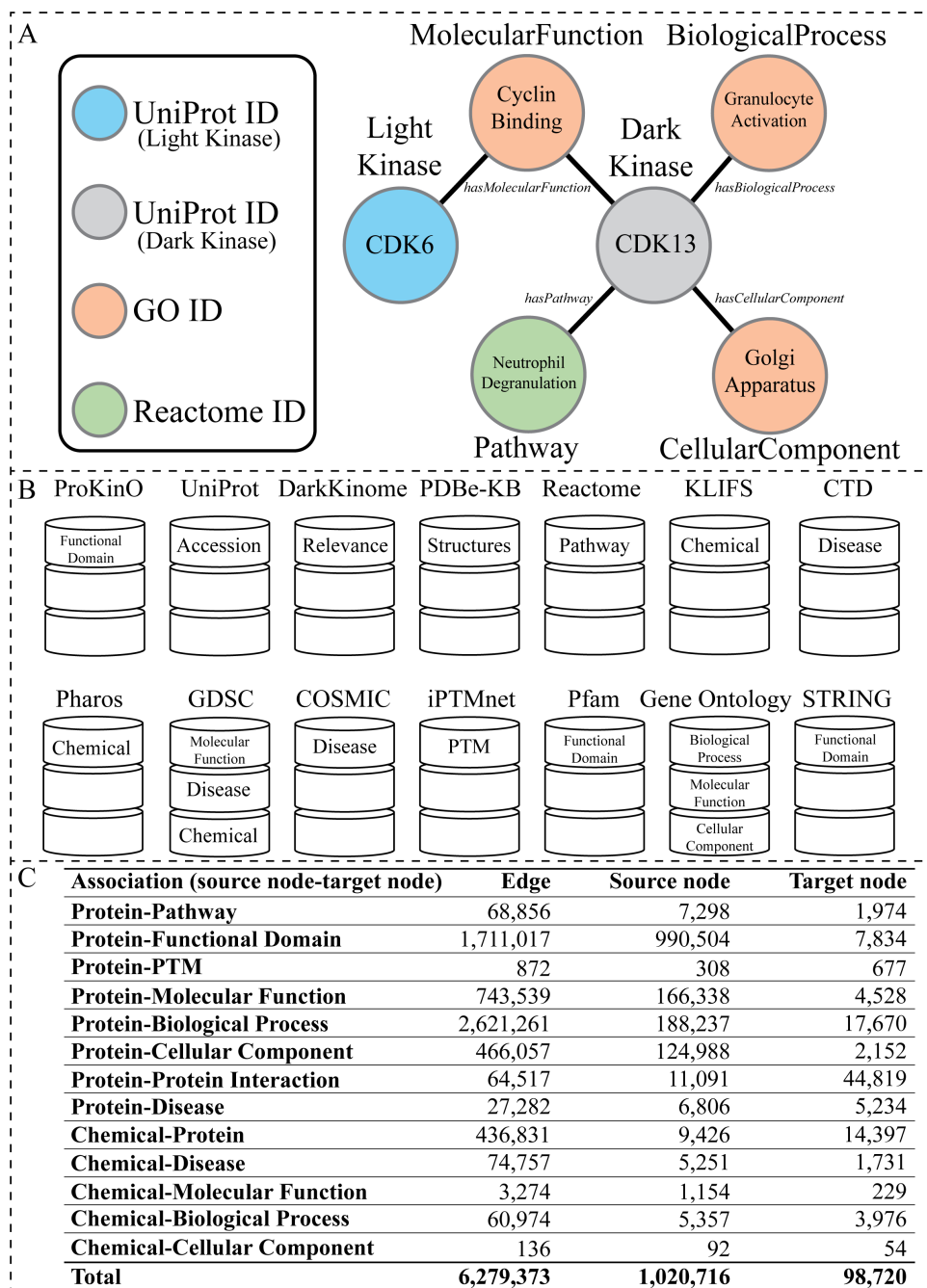
Figure 1 **Overview of the KG.** (A) Visual representation of the node types present within the KG generated from the listed databases above. The circles represent nodes (entities), and the lines represent edges (relationships). The circles are colored according to node type. The labels inside the circle are specific examples of data found in our KG related to the dark kinase CDK13. The labels outside of the circles correspond to the node type in bold text and the edge type in italics. (B) A graphical representation of all databases utilized in the construction of the KG with the node type label listed inside the representation of each corresponding database. (C) Abundance of associations with corresponding edge and node counts that exist in the KG.

schema, sometimes referred to as a meta-knowledge graph, or meta-graph, specifies constraints on using the edge labels to certain types of starting and ending nodes (subjects and objects in RDF). Also, given a KG schema, we can create a KG conforming to the schema containing many individuals (nodes).

Data sources and curation of datasets for knowledge graph generation

To construct our human kinase-centric KG, we gathered data from multiple publicly available curated resources (Fig. 1B). Nodes and edges within the graph were further integrated under more general node and relationship labels (also referred to as node types and edge types respectively) such as *Disease* (Figs. 1B and 2B). We further filtered the information included within our KG to avoid redundancy (described below). Fig. 1C shows basic statistics of different edges and their source and target nodes. The final KG contained 1,064,097 nodes of 11 types and 6, 279, 374 associations of 13 types.

The node types *Protein* and *FunctionalDomain* were populated with information on human kinase classification, functional domains, structure, and Protein–Protein Interaction (PPI). Information on these data types was extracted from the following databases: ProKinO (*McSkimming et al., 2015*), Dark Kinase Knowledgebase (darkkinome.org) (*Berginski et al., 2021*), UniProt (*UniProt, 2021*), PDBe-KB (*Velankar et al., 2010*), Pfam v33.1 (*Mistry et al., 2021*), and STRING v11 (*Szklarczyk et al., 2019*). Of note, included within our graph are numerous ortholog (non-human) *Protein* nodes collected from PPI databases. *Protein* Post-Translational Modifications (PTMs) associations were also included within our KG under node type *PTM* and were retrieved from iPTMnet v5 (*Ross et al., 2017*). The *Chemical* and *Disease* node types were populated from the Comparative Toxicogenomics Database (*Davis et al., 2021*) (protein-chemical, protein-disease, chemical-disease, and chemical-GO term associations) as well as cancer mutation data from the Catalogue of Somatic Mutations in Cancer (COSMIC) (*Tate et al., 2019*). The Genomics of Drug Sensitivity in Cancer (GDSC) (*Yang et al., 2013*) database was used for drug activity data. The *Chemical* node type additionally includes information on kinase-associated and ligand interaction motifs and ligand activity retrieved from Kinase-Ligand Interaction Fingerprints and Structures (KLIFS) (*Kanev et al., 2021*) and Pharos (*Sheils et al., 2021*), respectively. Gene ontology terms and pathway information regarding human kinases were retrieved from the Gene Ontology v2019_11 (*Gene Ontology, 2021*) (*GOTerm* node type) and Reactome v76 (*Jassal et al., 2020*) (*Pathway* node type), respectively.

Further filtering of data sources to reduce redundancy

The retrieved datasets were further curated prior to KG generation and population. All protein-pathway associations obtained from Reactome were split into manually curated associations (evidence = TAS) and predicted associations (evidence = IEA), only manually curated associations were included in our KG. Additionally, due to the hierarchical nature of the data taken from Reactome (i.e., Pathways are often organized into parent–child relationships), we eliminated high level pathways as they were deemed too general for
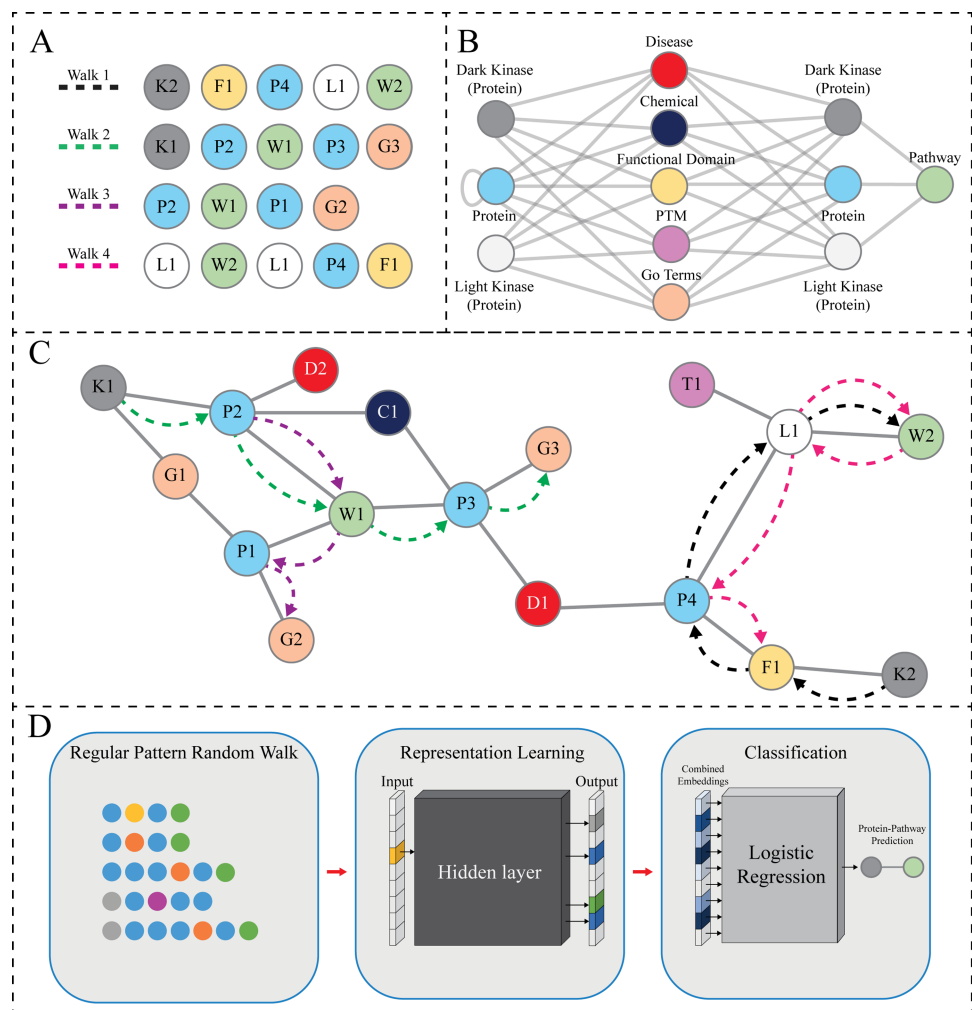
Figure 2 RegPattern2Vec Overview. (A) The hypothetical sequence of nodes produced by random walks guided by the regular pattern, shown in (B) (see legend for panel B for more details) and with the limitation of collecting five nodes (hyperparameter Length of Walk). This can result in several walks shown in (A). Walk1 shows a full walk that ends on a pathway node, Walk2 shows a full walk that ends on a node other than pathway, Walk3 shows a walk that ends early, and finally Walk4 shows a reverse walk. (B) The regular pattern designed for random walk sampling (schema). This pattern is defined as *Protein* [^*Pathway*] + *Protein Pathway*. Each random walk instance is constrained based on this pattern. All meta-paths must follow the schema in which they start with either a DarkKinase, LightKinase or Protein node and must collect a pathway node during the random walk. Once a pathway node is collected, if there are no other appropriate nodes according to the limitations put into place by both the schema and hyperparameters then reverse walks are allowed in which the pathway node will be collected again, and the walk resumes in a random direction until hyperparameters are satisfied. This feature is added to capture local relevant nodes in areas of data sparsity. (C) A hypothetical subgraph of the KG. The dashed lines show the node sequences highlighted in (A). Nodes are labeled and colored according to node types as shown in panel B with numbering corresponding to the order in which it is sampled (IE: P3 is the third Protein node sampled). This figure serves as a graphical depiction of how various data types exist within a neighborhood for sampling. (D) The meta-paths produced through our modified random walk approach are turned into embeddings for representation learning. Once the node context has been vectorized we then use logistic regression to perform link prediction as a binary classification task.

our experiments. For example, we filtered out the pathways beneath "Disease" (R-HSA-1643685) but not beneath "Infectious disease" (R-HSA-5663205). This was done not only to reduce redundancy in the data but also to ensure the vector embeddings created from our KG for link prediction were not overfitted to highly connected nodes (often referred to as *hubs*). Similar filtering was done with Protein-GO term associations retrieved from Gene Ontology-GO terms with more than 5000 associations were removed. Additionally, Protein-GO terms were assigned separate node and edge types in the KG according to the *MolecularFunction*, *CellularComponent*, and *BiologicalProcess* terms.

Additional filtering to increase confidence in the data within our KG was performed using various strategies specific to the data source. Only PPIs (STRING) directly involving a kinase and with an experimental data score of more than 700 were included in our KG, and only PTMs (iPTMnet) with a confidence score of more than 1.0 were included in the graph. Additionally, Pfam entries with the protein kinase domains ("Pkinase" and "Pkinase_Tyr") were removed.

Creation of curated kinase knowledge graph for kinase-pathway link prediction

The KG used in our link prediction experiments described here contains several node types such as *Protein*, GO Terms (*BiologicalProcess*, *MolecularFunction*, and *CellularComponent*), *Disease*, and other types. The nodes collected during a hypothetical random walk are shown in Fig. 2A. The hypothetical random walk is constrained by a schema organization that is shown in Fig. 2B the same schema is applied throughout the rest of the paper which indicates [*Protein*, any node not *Protein* and not *Pathway*, *Protein*, *Pathway*] (*Jassal et al., 2020*). In addition, the *Protein* type includes the understudied kinases (referred to as *DarkKinases*), well-studied kinases (referred to as *LightKinases*), and other proteins. This distinction is essential, as it gives us the ability to make predictions for a subset of proteins. The edge types (edges in the schema) are not labeled here, as we do not consider them in our method described here. Instead, we only rely on the types of source and target nodes. The loop edge returning to the *Protein* type indicates the *Protein–Protein* Interaction relation, which is included in our knowledge graph. Finally, Fig. 2C depicts the hypothetical subgraph in which the random walks are performed relating back to the nodes displayed in Fig. 2A.

Link prediction workflow

In the experiments and results presented in this paper, we used RegPattern2Vec. It is used as the first step in our link prediction process to produce vector representations for the nodes in the graph and formulate the link prediction as a classification problem. A ML model was trained using the combined vectors of existing pairs of nodes connected by an edge (link) of interest. The outline of the method is illustrated in Fig. 2D. We discuss each step of our link prediction method in the subsequent sections.

Regular pattern selection and its usage in random walks

Given a Heterogenous Information Network (as defined above), we have used regular patterns of the form $H[^\wedge T]+ H\ T$ for link prediction tasks, *i.e.,* edges (links) between

nodes of type *H* and *T*. In the pattern, *H* is a source node type and *T* is a target node type of the link of interest, respectively. Additionally, [^T] denotes any node type different from *T* (in regular expressions terminology it is known as the complement operator), and [^T]+ denotes a repetition (sequence) of one or more node types that are different from *T*. The overall pattern requires that any random walk must match the specified sequence of node types, in that a walk must begin with a node of type *H* then it must be followed by one or more nodes of type different than *T*, then a node of type *H*, which must then be immediately followed by a node of type *T*.

Simply put, the regular pattern is defined as a schema subgraph that should include only the relevant node types for a specific problem. An example regular pattern subgraph is shown in Fig. 2C. Unlike in the metapath2vec method (*Dong, Chawla & Swami, 2017*), the schema graph pattern represents multiple paths (walks) that can be used to predict missing links. Here, we aim to predict protein-pathway links and the random walks of interest are constrained by a regular pattern connecting node types and edge types most relevant to the prediction links. Here, when generating protein-pathway predictions, we have defined the regular pattern as *Protein [^Pathway] + Protein Pathway,* and each random walk instance must match it. Each walk starts from a Protein node. Selecting the next nodes on a walk is based on the existing graph nodes and matching the neighbor's type in the regular pattern. When a random walk reaches a Pathway node, it follows the pattern in reverse order until a certain number of steps (nodes) in the walk are reached, or when it reaches a termination node, *i.e.,* when there is no neighbor to match the pattern. We designed this specific aspect of the model due to the challenge of data sparsity for dark kinases. By allowing the reverse pattern to be followed we allow for additional relevant neighborhood nodes to be collected- when walking backwards once the pathway node is reached, the walk travels in a random direction until the hyperparameters are satisfied, capturing potentially relevant neighbors for our prediction task.

Biased random walk constrained by a regular pattern

As this work considered undirected graphs, enumerating a given graph's path is computationally infeasible. The solution is to sample a subset of paths from all possible paths from a random distribution. The random walk constrained by a regular pattern is selected to generate walks of arbitrary length controlled by the "*walk length*" hyperparameter. Having the undirected heterogeneous network and a selected regular pattern, the random walk can be started from each instance of a starting node type in the pattern. As we want all the nodes to appear in our walks, iterating over them would be desirable. It is obvious that if we repeat the walk from each node (not all KG nodes may begin a walk, as constrained by the regular pattern), we will discover more walks as the node might link to multiple nodes of the same type. We call this hyperparameter the "*number of walks*" (referred to as NW in Results 'Benchmarking RegPattern2Vec's predictive performance with metapath2vec and other Graph Embedding Approaches'–Investigating the impact of hyperparameters on regpattern2vec's predictive variability among replicate runs'). We will discuss how to choose the hyperparameter and the analysis of their impact in the Results 'Benchmarking RegPattern2Vec's predictive performance with metapath2vec and other Graph Embedding
Approaches'–Investigating the impact of hyperparameters on regpattern2vec's predictive variability among replicate runs'. The next step for each node is to select a node from the adjacent nodes based on the established regular pattern. This might result in multiple choices, and this is where randomization comes to play. RegPattern2Vec can utilize random distribution created by a user-defined function to generate the same probability for all the nodes or use an arbitrary distribution. To implement the random walk collection process, a regular pattern is converted to a Deterministic Finite Automaton (DFA), denoted by M. Each possible walk step is mapped to the DFA. The DFA M is used to check if transitions are allowed (an edge between two nodes); hence, a disallowed change gets a zero probability and is not used in the random walks.

On the other hand, in scale-free networks where the degree distribution follows the power law, some nodes may have a high degree of incoming/outgoing edges (known as *hubs*). Because such high degree nodes can dominate random walks and, consequently, representation learning, one popular way to address this issue is to bias the walks by the inverse of degrees of nodes (*Grover & Leskovec, 2016*), where the probability of choosing a node v^{i+1} from v^i is calculated by normalizing the inverse of degrees of all neighbors of v^i. Although this approach lowers the probability of selecting high-degree nodes, it biases the random walk toward low-degree nodes, thereby capturing pertinent information related to low-degree "dark kinase" nodes. We used the formula below for node selection:

$$P\left(v^{i+1}|v^i,M\right) = \begin{cases} g\left(r^i\right)\dfrac{\frac{1}{|N_{v^{i+1}}|}}{\sum_{v\in N_{v^i}}\frac{1}{|N_v|}} & \left(v^i,r^i,v^{i+1}\right)\text{ is an edge in the graph and the} \\ & \text{transition from v to v is allowed in M} \\ 0 & \left(v^i,r^i,v^{i+1}\right)\text{ is an edge, but the transition is undefined} \\ 0 & \left(v^i,r^i,v^{i+1}\right)\text{ does not exist in the graph} \end{cases}$$

In the formula, v^i denotes the current node and the candidate node for next step is $v^{i+1}\in N_{v^i}$. $|N_v|$ denotes the degree of node v, and N_{v^i} is the set of all the neighbors of node v^i. $g\left(r^i\right)$ is the proportion of the edges of type r ^{i} among all edges of node v^i. Therefore, we randomly choose one edge (relation) type and then use the probability distribution by the inverse of node degrees to select the next node in the walk, but only among the nodes connected by the edge type chosen.

Vectorization of nodes and deep learning of semantic relationships between proteins and pathways

RegPattern2Vec uses a modified skip-gram model presented in (*Dong, Chawla & Swami, 2017*) to generate vector representations for the nodes of the KG. The random walks generate sequences of nodes, which resemble natural language sentences. The ML model simultaneously captures the local structure of the graph and the types of the nodes and encodes them as vector representations.

Link prediction as a classification problem

For each pair of protein-pathway associations, we combined their vector embeddings utilizing a widely used Hadamard product (*Grover & Leskovec, 2016*; *Lin et al., 2015*;

Minervini et al., 2015). The resulting vector is used as features to train a Logistic Regression model. Training the model additionally requires generating negative examples for protein-pathway association pairs which are difficult as resources with true biological negatives, such as Negatome (*Blohm et al., 2014*), do not exist for our *Pathway* data type. To generate negative examples, we must first work under the closed-world assumption. This assumes that all information needed is provided, so for a statement to be true it must be explicitly stated to be so. Thus, a lack of a statement denotes that it is false (this contrasts with the open-world assumption in which statements can be true without them being explicitly known to be so). For our relation of interest (predicted link), we randomly select a head and tail node not connected by an edge in the graph. The embeddings of such nodes are then combined to produce negative examples. This process is repeated until the number of negative examples matches the positive examples.

Our logistic regression method utilizes the one-versus-rest strategy, in which a multi-class classification is split into one binary classification problem per class, meaning a classifier will be trained for each task, and therefore all the other data points in the split data are used as negative examples with the methods discussed above (under the closed-world assumption). Thus, the number of negative examples is dependent on the number of positive examples for each classification task. Although our dataset is large, this strategy is employed due to the inherent data imbalance between well-studied and dark kinases.

RESULTS

In this work, we propose a new guided random walk approach for KGs, known as RegPattern2Vec, used to predict pathway associations for dark kinases. To generate protein-pathway predictions, we first constructed a protein kinase KG by integrating curated data from various resources populating the graph with nodes on molecular function, disease association, protein–protein associations and more (see Materials & Methods 'Data sources and curation of datasets for knowledge graph generation'). The resulting KG consisted of 1,064,097 nodes (entities) and 6, 279, 373 edges (relationships) (Table S1, Fig. 1C). RegPattern2Vec was then used to sample the large KG into sequences of nodes. Regular patterns from RegPattern2Vec guided sampling. Using such patterns, we generate several walk paths (or a list of sequential nodes in the network) for representation learning. This allows the sampling to capture different sub-structures in the graph (covering different meta-paths) without explicitly designing walks for them.

After sampling, the model learned vector embeddings for each node in the collected paths, utilizing a Natural Language Processing (NLP) model similar to metapath2vec (*Dong, Chawla & Swami, 2017*). Link predictions utilizing the embeddings were then formulated as binary classification tasks and predictions were made for protein-pathway associations using a logistic regression model. For more details, refer to the Materials & Methods 'Link prediction as a classification problem'.

Capturing semantic relationships between proteins and pathways in deep learning

As previously mentioned in more detail, (see Materials & Methods 'Vectorization of nodes and deep learning of semantic relationships between proteins and pathways') RegPattern2Vec uses a modified skip-gram model presented in (*Dong, Chawla & Swami, 2017*) to generate vector representations for the nodes of the KG. Figure 3 shows the learned vector representation of nodes in the vector space using principal component analysis (PCA), a standard dimensionality reduction technique. For the protein-pathway predictions, we just consider the nodes to be of three types: "*Protein*", "*Pathway*", and "*Others*" when learning representation for the nodes. The separation of nodes in PCA shows our vector representation captures the node types and their network context. Several key differences exist between protein nodes and other nodes in our graph that may explain why they cluster separately from all other nodes. Protein nodes (including those annotated as dark/light kinases) are the default starting node in the schema, and many edge types describe proteins and connect them to other nodes in our graph (for example, a node describing a functional domain will be connected to the protein node it is describing and we additionally illustrate protein–protein interactions in our graph). This may allow protein nodes to occupy a unique space within the embedding reflected in the PCA.

Benchmarking RegPattern2Vec's predictive performance with metapath2vec and other graph embedding approaches

To evaluate the accuracy of our method and note any improvements from metapath2vec, a test set was first generated by excluding 50% of the known protein-pathway associations from the training set. The training examples were generated using the process detailed in the Materials & Methods (Link prediction workflow–Link prediction as a classification problem'). To compare our method to metapath2vec, we produced AUC ROC curves for both models. Multiple curves were generated for the metapath2vec model as multiple meta-paths were used resulting in variability in the model's performance (Fig. 4). Metapath 1 ('*Protein*', '*FunctionalDomain*', '*Protein*', '*Pathway*', '*Protein*') showed the best performance, based on 10-fold cross-validation, with an f1-score of 0.87 and AUC ROC of 0.94 for protein-pathway prediction. In contrast, RegPattern2Vec achieved an f1-score of 0.90 and AUC ROC of 0.96 on the same training and testing datasets (Fig. 4). We also compared RegPattern2Vec to other recently proposed embedding approaches for heterogenous graphs, namely BoxE (*Abboud et al., 2020*), MAGNN (*Fu et al., 2020*), and RHINE (*Shi et al., 2020*). These models were chosen mostly due to their ability to scale with larger knowledge graphs. Surprisingly, these models displayed poor performance in the task of predicting dark kinase pathway associations (Fig. S1), presumably because of the sparsity of knowledge related to dark kinases in the KG. RegPattern2Vec overcomes some of the challenges imposed by data sparsity by resampling a node as a starting point, allowing for constraint backward walking for a greater sampling of neighbor nodes, and use of inverse node degree as a consideration to select the next node in the walk.
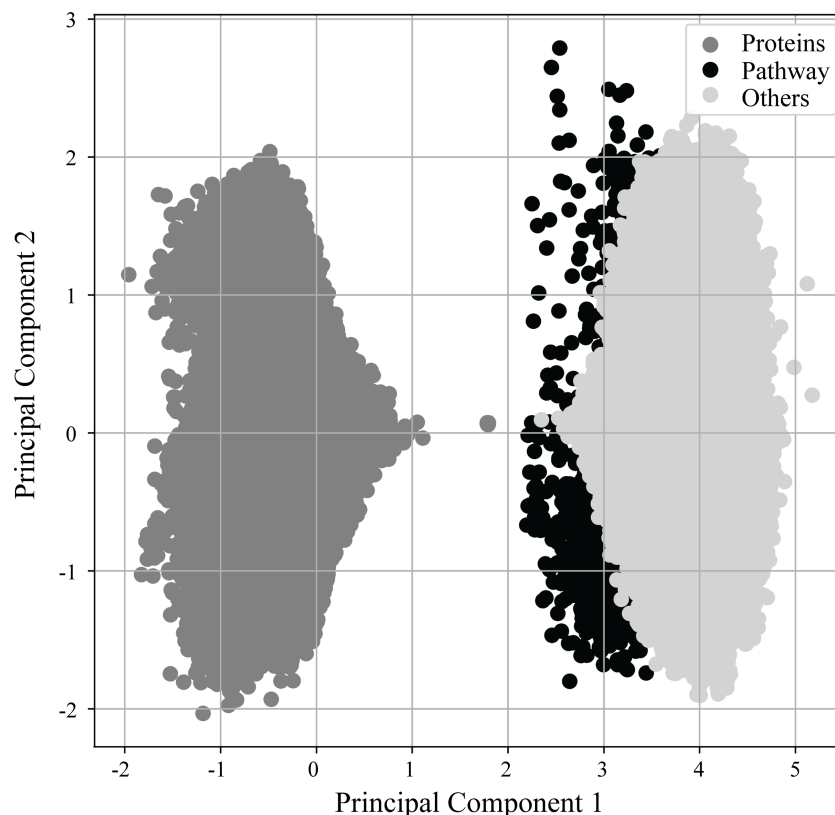
Figure 3 **PCA of vector embeddings.** Principle component analysis of the vector embedding produced for all nodes generated by ML model. Given that the defined schema of RegPattern2Vec is defined as *Protein* [^*Pathway*] + *Protein Pathway* when clustering the vector embeddings, we considered all nodes to be of type: *Protein*, anything that is not a *Pathway* or *Protein* (*Others*), and *Pathways*.

Investigating the impact of hyperparameters on regpattern2vec's predictive variability among replicate runs

Although RegPattern2Vec is better suited for the task of learning on sparse biological KGs, due to the nature of the random walks, there is an intrinsic level of variability with the predictions made by our model. Although the exact same sequence of nodes will not be sampled each time, there should be a larger context within the embedding space that identifies patterns useful for predictions for the dark kinases sampled. Therefore, overlaps in the *Pathway-Protein* predictions made by our model using different hyperparameters are expected. To demonstrate our model's robustness, we measured the predictive consistency while modifying hyperparameters across three replicates (runs that all have the same hyperparameters) and compared these results. The percentage of protein-pathway predictions that overlap (within all replicates) was used as a metric for our model's robustness. This analysis also revealed the variability in the overlap of the protein-pathway predictions on a per-kinase basis (for both light and dark kinases). This variability may be due to imbalances in the data available for each kinase. Notably, greater variability was seen amongst dark kinases, with the highest percentage overlap being 87.5% and the lowest
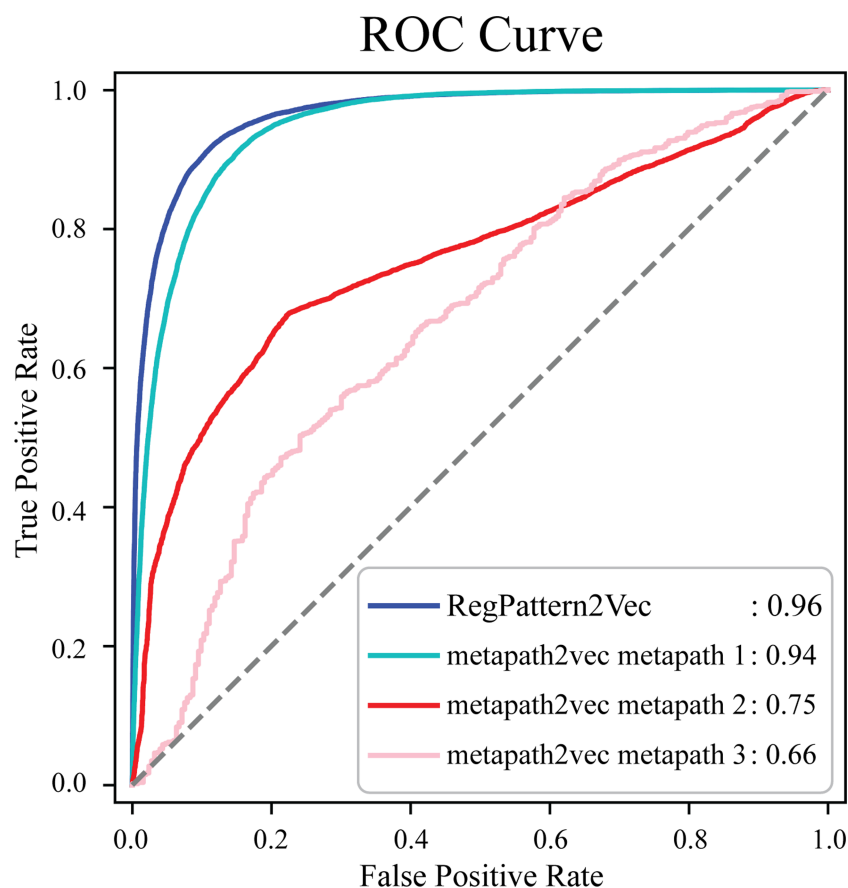
Figure 4 **Comparison of AUC ROC curves generated with RegPattern2Vec and metapath2vec.**
The comparison of AUC ROC curves of RegPattern2Vec and differing schemas utilized with the
metapath2vec model. AUC ROC curves were generated by excluding 50% of the known associations
from the training set, link prediction was then carried out using a logistic regression algorithm type as the
binary classification model. metapath 1 consists of the schema ['*Protein*', '*FunctionalDomain*', '*Protein*',
'*Pathway*', '*Protein*']. metapath 2 consists of the schema ['*Protein*', '*DarkKinase*', '*Protein*', '*Pathway*',
'*Protein*']. metapath 3 consists of the schema ['*LightKinase*', '*Protein*', '*DarkKinase*', '*Protein*', '*Pathway*',
'*LightKinase*']. metapath 1 was the best performing schema tested, metapath 2 was still among our top
performing schemas tested, and metapath 3 was among the worst performing schemas tested.

percentage overlap being 0% (for the hyperparameter 40 NW with 95% confidence cutoff).
A supplementary table (Table S2) provides further information indicating the overlap seen
for each kinase. Additionally, we provided a supplementary figure (Fig. S2), displaying the
confidence (0.6 to 1.0) in overlap for all hyperparameters tested.

One source of variability within our model is directly controlled by the hyperparameter
"a *number of walks* (NW)", which specifies the number of times a node should be sampled
as the starting node during the random walk process (refer to Materials & Methods
'Vectorization of nodes and deep learning of semantic relationships between proteins
and pathways'). Repeating the walk from a specified node allows for a more complete
characterization of neighbor nodes, often allowing for more node/edge types to be captured

and considered in downstream analysis which is especially important when working with sparse data.

To find the optimal NW, the same starting node was used for generating paths in the random walk anywhere from 10 to 80 times. This process was repeated, resulting in a total of three replicates for each variation of the NW hyperparameter. Overlap between the protein-pathway predictions obtained through replicate hyperparameter conditions was then compared. From this analysis, we sought to identify the smallest value for the hyperparameter NW that can accurately capture local structure and node context within the graph to make consistent predictions (as determined by replicate overlap). This analysis revealed that changing the hyperparameter NW did not drastically change the number of overlapping protein-pathway predictions between replicate datasets (Fig. S3). Of note, the value of 40 for the hyperparameter NW (abbreviated 40 NW) generated the most consistent predictions among the NW parameters tested, with a mean of ~22% overlapping pathways across three 0.7 cutoff replicates when averaging overlap over all kinases even when utilizing a range of confidence cutoffs from 0.7−0.95 (Fig. S3). As a result of this benchmark, we determined that 40 NW is the optimal hyperparameter for predicting protein-pathway relationships in the current KG, as it performs slightly better (as measured by replicate overlap) than the more computationally intensive hyperparameter values tested.

Hyperparameter optimization results in consistent replicate pathway predictions for 34 dark kinases across different replicate runs

After finding the optimal value for the hyperparameter NW (Fig. S3), we decided to investigate the overlap of all protein-pathway predictions produced by our previous analysis in which we test the hyperparameter NW (from 10–80 with a total of three replicate runs for each condition). The overlapping predictions present amongst highly variable walks suggest that these predictions are made based on nodes more often sampled within our network. After filtering for predictions that have over 0.95 confidence, only 95 dark kinase protein-pathway predictions emerged that were present in all replicates of all variations of NW tested (a total of 24 datasets) (Fig. 5A), resulting in representative protein-pathway predictions for 34 of the unique dark kinases in our dataset (Table S3). The top ten kinases (ranked according to the number of predictions that overlap) are shown in Fig. 5B. Of note, many of the protein-pathway predictions generated by our model were highly consistent (either in pathways with similar function or protein family involvement) on a per kinases basis (Table S3). For example, for the dark kinase VRK serine/threonine kinase 3 (VRK3), seven out of ten predictions were related to Toll Like Receptor (TLR) cascades, and five out of five protein-pathway predictions for the dark kinase protein kinase, membrane-associated tyrosine/threonine 1 (PKMYT1) were related to cell cycle control (Table S3). We further focused on a subset of dark kinases for path analysis described below.

Path analysis provides context for prediction associations between understudied PRKACB and DAG/IP3 signaling

The dark protein kinase cAMP-activated catalytic subunit beta (PRKACB) was amongst the kinases identified with the highest overlap of protein-pathway predictions (Fig. 5B).
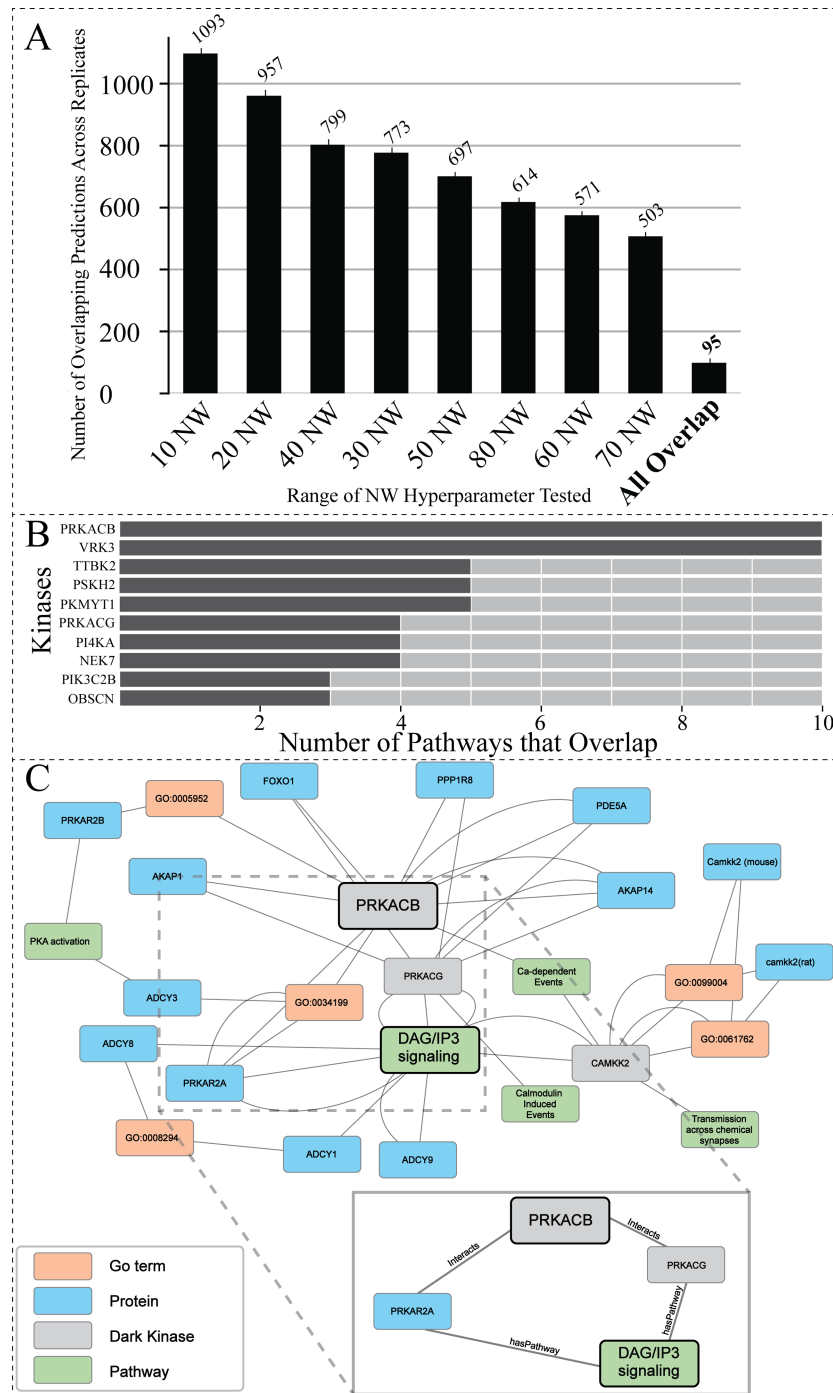
Figure 5 **Investigating the overlap among all replicates and NW hyperparameter values.** (A) Total number of overlapping *DarkKinase-Pathway* predictions for all variations of the NW hyperparameter tested (10–80) with replicates (each hyperparameter tested three times) for a total of 24 datasets compared. Replicates were averaged and then graphed. Finally, overlap

Figure 5 (…continued)
between all 24 datasets was compared in the "All Overlap" bar (B) Graph depicting the number of overlapping pathways for the top ten (out of 34) kinases ranked by number (abundance) of overlapping predictions (C) We highlight a portion of the subgraph for the nodes collected through our modified random walk process for the protein-pathway prediction generated through link prediction between *PRKACB* and *DAG/IP3 signaling*. The nodes shown are common between five paths in the KG, obtained through the variation in NW datasets described above. PRKACB was chosen for this analysis based on panel B. The key indicates the color-coded node types, and the magnified panel highlights the highly connected hubs and their edge types.

PRKACB, the beta-catalytic subunit of Protein Kinase A (PKA), appears to be involved in calcium regulation and modification, a known role for PKA (*Barodia, Sophronea & Luthra, 2022*; *Kania et al., 2017*; *Ould Amer & Hebert-Chatelain, 2018*; *Palencia-Campos et al., 2020*). Eight of the ten pathways identified were directly linked to calcium regulation (Table S3). To contextualize the predictions made by our model for PRKACB, we created a subgraph displaying common nodes traversed during the random walk process for the node representing the Reactome protein-pathway prediction R-HSA-1489509 (*DAG/IP3Signaling*). A portion of the subgraph highlighting the common nodes sampled between different paths are shown in Fig. 5C. The nodes found in all sampled paths are more likely to represent nodes that are important for providing the context that results in the shared protein-pathway prediction. The entirety of the subgraph (Data S1) displays all nodes with the addition of unconnected nodes that only appear in one path generated during the random walk process.

The highly sampled nodes of the subgraph shown in Fig. 5C, reveal that *PRKACG* (protein kinase cAMP-activated catalytic subunit gamma), *PRKAR2A* (protein kinase cAMP-dependent type II regulatory subunit alpha) (*Manchev et al., 2014*; *Wei et al., 2021*) and (calcium/calmodulin dependent protein kinase kinase 2) *CAMKK2* (*Najar et al., 2021*) are nodes often sampled for the *DAG/IP3Signaling* prediction. The nodes being sampled multiple times demonstrate our model's ability to leverage PPIs to infer pathway involvement. Additional nodes of interest that are often sampled for this prediction include PPIs between PRKACB (protein of interest) and both A-kinase anchor protein 28 (AKA28), and cGMP-specific 3′, 5′-cyclic phosphodiesterase (PDE5A). Both proteins have biologically relevant functions supporting the protein-pathway prediction for PRKACB's involvement with DAG/IP3 signaling. AKA28 has been shown to bind to type II regulatory subunits of PKA (related to our dark kinase) and anchors/targets PKA to discrete locations within the cell (*Kennedy & Scott, 2015*; *Kultgen et al., 2002*; *Omar & Scott, 2020*; *Sarma et al., 2010*). PDE5A is a phosphodiesterase that regulates intracellular levels of cAMP and AMP (*Peng et al., 2020*). These observations support the predicted link/association between understudied PRKCB kinase and DAG/IP3 signaling pathway.

Predicted association of understudied CDK19 in TGF-beta receptor signaling: interpretability with path analysis and validation with PPI datasets

To further explore the biological relevance of our predictions for dark kinases, we compared our protein-pathway predictions to Reactome datasets generated using Protein Interaction

and Proximity (PPI) data for dark kinases not included in our training data. PPI data was obtained from the Dark Kinase Knowledgebase (*Berginski et al., 2021*). Of note, many kinases in this dataset did not have sufficient protein-interactors to perform Reactome pathway enrichment analysis. For the 50 dark kinases with an adequate number of identified interactors, the overlap between the protein-pathway predictions generated by our link prediction and the Reactome enrichment generated by PPI data were compared. Figure 6A shows the top ten dark kinases ranked according to protein-pathway prediction overlap abundance and Table S4 provides the pathway information for all dark kinases with overlap seen. Among the dark kinases with the highest overlap from this comparison was cyclin-dependent kinase 19 (CDK19), a dark kinase belonging to the cyclin-dependent kinase superfamily (*Malumbres, 2014*).

After performing similar subgraph analysis to that previously described for PRKACB (Results,, 'Hyperparameter optimization results in consistent replicate pathway predictions for 34 dark kinases across different replicate runs'), we were able to isolate nodes commonly traversed for the protein node of interest (*CDK19*) and the pathway node of interest *TGF-betaReceptorSignaling* (R-HSA-170834) (Fig. 6B). Again, only a subset of the paths is shown for ease of viewing. The entirety of the subgraph (Data S2) displays all nodes with the addition of unconnected nodes that only appear in one path generated during the random walk process. For the dark protein kinase CDK19 with the link prediction of TGF-beta receptor signaling, two "key" node hubs were identified including *CCNC* (cyclin-c) and *CDK8* (cyclin dependent kinase 8). Recent total mRNA sequencing has demonstrated that loss of CCNC could activate the transforming growth factor (TGF)-beta signaling pathway (*Tang et al., 2021*). CDK8 has also been implicated in TGF-beta signaling through previous studies showing they drive Smad transcriptional action and turnover in TGF-beta pathways (*Alarcón et al., 2009*). Through interactions with several members of the mediator of RNA polymerase II transcription (MED) protein family, CDK19 can also indirectly exploit the relationship between CDK8 and TGF-beta receptor signaling. It appears that CDK19, and CDK8 have been shown to directly interact with not just each other but MED12, as well as CCNC, forming a complex collectively termed the "Mediator Complex" (*Fant & Taatjes, 2019*). This complex interacts with DNA-bound transcription factors and RNA polymerase II (Pol II) to activate and repress gene expression, with mediator subunit MED12 promoting TGF-beta signaling through both canonical regulation of transcription and non-genomic activity (*Weber & Garabedian, 2018*). Taken together this suggests that while direct connections contribute to the predictive power of the model the surrounding node context can also reinforce these connections and can contribute to predictions made.

Predicted association between understudied PSKH2 and cilium assembly

Protein-pathway predictions were additionally produced by our model for an understudied protein kinase PSKH2 (*Byrne et al., 2022*) which is a paralog of the Golgi-associated protein serine kinase H1 (PSKH1) (*Brede et al., 2000*). Currently, the Dark Kinase Knowledgebase (*Berginski et al., 2021*) lists only ten known protein interactors for PSKH2 and no pathway annotation for this kinase exists in the literature. We utilized this and another recently
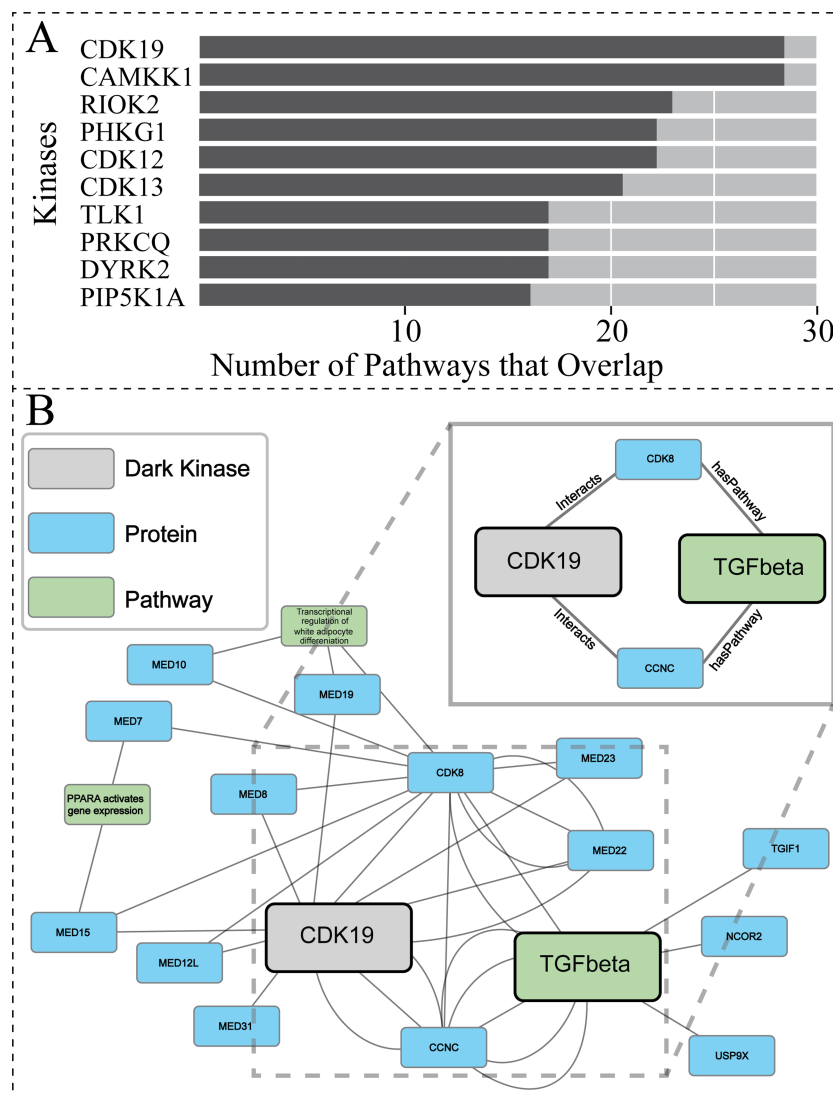
Figure 6 Investigating protein-pathway predictions and dark kinase knowledgebase experimentally validated predictions. Protein Interaction and Proximity (PPI) data for dark kinases obtained through the Dark Kinase Knowledgebase were used as input for Reactome enrichment. These predictions were then compared to those produced by our model using data produced during the hyperparameters exercise (24 datasets). (A) Graph depicting the number of overlapping pathways between the two datasets for the top ten kinases ranked by number (abundance) of overlapping predictions. (B) We highlight a portion of the subgraph for the nodes collected through our modified random walk process for the protein-pathway prediction generated through link prediction between *CDK19* and *TGF-beta*. The nodes shown are common between five paths in the KG, obtained through the 40 NW dataset. CDK19 was chosen for this analysis based on (A). The key indicates the color-coded node types, and the magnified panel highlights the highly connected hubs and their edge types.

published dataset (*Byrne et al., 2022*) not included in training, for the Reactome enrichment analysis and independent validation of RegPattern2Vec predictions on PSKH2.

Comparing the overlap in protein-pathway predictions (40 NW dataset) with the Reactome enrichment, 18 pathways were shown to overlap (Table S5). The previously
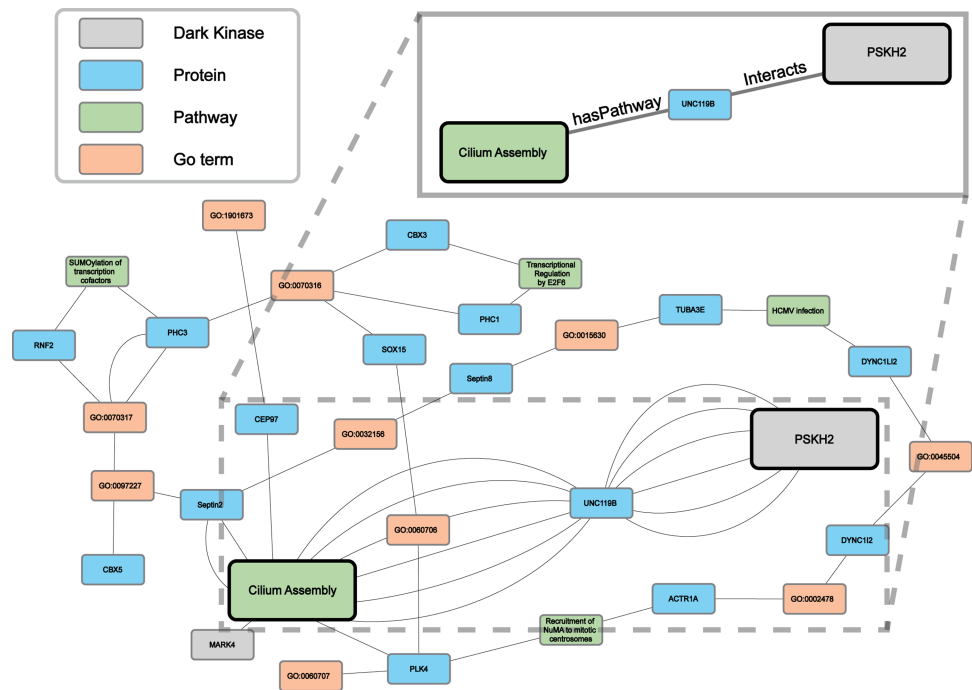
Figure 7 **A subgraph of the network context of the KG surrounding PSKH2.** As demonstrated in the previous Fig. 5B our model can generate consistent predictions for dark kinase PSKH2. We compared the predictions produced by our model (40 NW) and Reactome enrichment using unpublished mass-spectrometry data produced by our collaborators. We highlight a portion of the subgraph for the nodes collected through our modified random walk process for the protein-pathway prediction generated through link prediction between *PSKH2* and *CiliumAssembly*. The nodes shown are common between five paths in the KG, obtained through the 40 NW dataset. The key indicates the color-coded node types, and the magnified panel highlights the highly connected hubs and their edge types.

characterized subgraph visualization (Results, 'Hyperparameter optimization results in consistent replicate pathway predictions for 34 dark kinases across different replicate runs') was then performed, with five random paths chosen for visualization of the nodes commonly traversed for the protein node of interest (*PSKH2*) and the prediction node of interest (*CiliumAssembly* R-HSA- 5617833) (Fig. 7 and Data S3). Through this analysis one major "key" node hub was identified, *UNC119B*. unc-119 lipid binding chaperone B (UNC119B) is a protein that plays a key role in the localization of ciliary proteins by binding to N-myristoylated proteins and masking the hydrophobic lipid from hydrophilic cytosol, therefore facilitating trafficking to the primary cilium or the immunological synapse (*Stroukov et al., 2019*; *Yelland et al., 2021*). Many of the additional commonly traversed nodes belong to larger families of proteins known to play important roles in cilium assembly such as tubulins and dyneins. Thus, further characterization of these proteins can shed light on the role of PSKH2 in cilium assembly.

Overall, through our analysis of the paths utilized by the model as context for the link prediction process, we demonstrated that our modified random walk approach, RegPattern2Vec, can successfully sample multiple types of nodes and edges. These findings
support our notion that our model preserves both local and latent structure through appropriate neighbor sampling and therefore provides the node context needed to make predictions. The predictions made by our model regarding dark kinases allow us to explore dark kinases of interest in a pathway context, giving us an idea of how these kinases may function and relate to other proteins.

DISCUSSION

In this work, we present a network approach to predict pathway associations for understudied dark kinases by leveraging the network context of well-studied kinases. We predict pathway associations for dark kinases and propose a new graph embedding approach, RegPattern2Vec, which has significant advantages over the previously proposed metapath2vec approach for KG embedding. First, unlike metapath2vec, RegPattern2Vec does not require pre-defined knowledge of KG schema to define the paths. Second, because RegPattern2Vec does not use fixed length meta-paths, it captures the structure of the graph and the characteristics of both local and latent representations more effectively, as indicated by improved performance measures (Fig. 4). Third, our hyperparameter optimization strategies, combined with analysis of meta-paths, enable biological interpretation of predicted links.

While the KG embedding and ML approaches described here have distinct advantages over metapath2vec for link prediction, our model can be further improved through the generation of curated training and testing datasets and additional data that capture pathway context. Currently, to generate negative datasets, any missing link between nodes is used as an indication that a relationship does not exist. In short, the negative sample is simply non-connected nodes. In certain knowledge domains turned into KGs this is appropriate, however just because a link does not exist within our graph does not mean it does not occur in a cellular context (*i.e.,* two proteins interacting or a protein having a certain association). To address this limitation, we are currently working on a KG that includes subcellular localization information of proteins, and tissue/cell specificity to create a more informed negative set that better reflects the natural boundaries placed by biology rather than the boundaries of our current knowledge. Additional future directions include fully incorporating edge information into the embeddings used for our model's predictions. The use of edge labels could allow for the use of additional Natural Language Processing (NLP) based models such as Transformers (*Bi et al., 2022*; *Koncel-Kedziorski et al., 2019*; *Xie et al., 2022*; *Yao, Mao & Luo, 2019*).

Overall, we have demonstrated that our model can make valid biological predictions even with sparse input through the use of node context extracted from the KG. While this current work was focused on placing dark kinases in a pathway context, within the KG are nodes and edge types relevant for PPI and substrate predictions as well (Fig. 1C, Table S1). Therefore, our KG can be further utilized for other predictive tasks and the functionality of the KG can be further expanded by incorporating other forms of data. For example, including information such as cell-type specific expression of kinases could potentially improve the overall model performance while enabling predictions for cell-type

specific dark kinase functions and interactions. However, the addition of new data types can significantly alter the topology of the KG. As data is added, new node/edge types and schemas will need to be incorporated into the graph and hyperparameters will again need to be optimized. The protocol described in Results ('Benchmarking RegPattern2Vec's predictive performance with metapath2vec and other graph embedding approaches') for testing hyperparameters serves as a framework for continued usage of the KG, and exploration of variability within the random walk process. The code for RegPattern2Vec has been published on GitHub and these methodologies can be applied to additional KGs for other understudied proteins and protein families.

CONCLUSIONS

We sought to characterize understudied dark kinases utilizing a network-based approach that combines multi-domain knowledge on both well-studied kinases and dark kinases within a single heterogeneous KG. Our method focuses on improvements within the random walk component of the graph embedding process. To accomplish this, we utilized RegPattern2Vec, allowing for the selection of the relevant part(s) of the KG using labeled semantics (regular patterns) to not only learn the local similarity between nodes and edges but also more complex and non-trivial associations which lead to novel protein-pathway predictions for dark kinases. Our method performs better in terms of scalability for large graphs compared to other network-based methods by sampling the graph more accurately while preserving node/edge context and structure.

ACKNOWLEDGEMENTS

We acknowledge members of the Kannan lab for their valuable comments and suggestions. We additionally acknowledge the various contributions to the databases we utilized made through the efforts of the IDG consortium and numerous investigator-initiated efforts.

Funding

This work was supported by the National Institutes of Health (U01CA239106). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Institutes of Health: U01CA239106.

Competing Interests

Natarajan Kannan is an Academic Editor for PeerJ.

Author Contributions

- Mariah V. Salcedo conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Nathan Gravel conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Abbas Keshavarzi conceived and designed the experiments, performed the experiments, prepared figures and/or tables, constructed Knowledge Graph, and approved the final draft.
- Liang-Chin Huang analyzed the data, prepared figures and/or tables, constructed Knowledge Graph, and approved the final draft.
- Krzysztof J. Kochut conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Natarajan Kannan conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The Knowledge Graph, Example Embedding, UniProt IDs to Gene Names File, Reactome IDs to UniProt ID Mapping File, and Folder Hierarchy Need for Example Code are available at Zenodo: Nathan Gravel. (2023). Supplemental Data 4 (v1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7541827.

The data related to the code are available at Github and Zenodo:

Available at https://github.com/gravelCompBio/RegPattern2Vec

gravelCompBio. (2023). gravelCompBio/RegPattern2Vec: v1.0 (v1.0-alpha.0). Zenodo. Available at https://doi.org/10.5281/zenodo.7541894.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.15815#supplemental-information.

REFERENCES

Abboud R, Ceylan I, Lukasiewicz T, Salvatori T. 2020. Boxe: a box embedding model for knowledge base completion. *Advances in Neural Information Processing Systems* **33**:9649–9661.

Alarcón C, Zaromytidou AI, Xi Q, Gao S, Yu J, Fujisawa S, Barlas A, Miller AN, Manova-Todorova K, Macias MJ, Sapkota G, Pan D, Massagué J. 2009. Nuclear CDKs drive Smad transcriptional activation and turnover in BMP and TGF-beta pathways. *Cell* **139**:757–769 DOI 10.1016/j.cell.2009.09.035.

Alshahrani M, Thafar MA, Essack M. 2021. Application and evaluation of knowledge graph embeddings in biomedical data. *PeerJ Computer Science* **7**:e341 DOI 10.7717/peerj-cs.341.

Anil A, Singhal S, Jain P, Singh SR, Ladhar A, Singh S, Chugh U. 2019. Network sampling using k-hop random walks for heterogeneous network embedding. In: *Proceedings of the ACM India joint international conference on data science and management of data*. New York: ACM, 354–357.

Ardito F, Giuliani M, Perrone D, Troiano G, Muzio LLo. 2017. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). *International Journal of Molecular Medicine* **40**:271–280 DOI 10.3892/ijmm.2017.3036.

Attwood MM, Fabbro D, Sokolov AV, Knapp S, Schiöth HB. 2021. Trends in kinase drug discovery: targets, indications and inhibitor design. *Nature Reviews Drug Discovery* **20**:839–861 DOI 10.1038/s41573-021-00252-y.

Ayala-Aguilera CC, Valero T, Lorente-Macías Á, Baillache DJ, Croke S, Unciti-Broceta A. 2022. Small molecule kinase inhibitor drugs (1995-2021): medical indication, pharmacology, and synthesis. *Journal of Medicinal Chemistry* **65**:1047–1131 DOI 10.1021/acs.jmedchem.1c00963.

Bachman JA, Gyori BM, Sorger PK. 2022. Automated assembly of molecular mechanisms at scale from text mining and curated databases. *bioRxiv*.

Barodia SK, Sophronea T, Luthra PM. 2022. A(2A) R mediated modulation in IP(3) levels altering the [Ca(2+)](i) through cAMP-dependent PKA signalling pathway. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1866**:130242 DOI 10.1016/j.bbagen.2022.130242.

Berginski ME, Moret N, Liu C, Goldfarb D, Sorger PK, Gomez SM. 2021. The Dark Kinase Knowledgebase: an online compendium of knowledge and experimental results of understudied kinases. *Nucleic Acids Research* **49**:D529–D535 DOI 10.1093/nar/gkaa853.

Bi Z, Cheng S, Zhang N, Liang X, Xiong F, Chen H. 2022. Relphormer: relational graph transformer for knowledge graph representation. ArXiv preprint. arXiv:220510852.

Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, Ruepp A, Frishman D. 2014. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Research* **42**:D396–D400 DOI 10.1093/nar/gkt1079.

Bonner S, Kirik U, Engkvist O, Tang J, Barrett IP. 2022. Implications of topological imbalance for representation learning on biomedical knowledge graphs. *Briefings in Bioinformatics* **23**:bbac279 DOI 10.1093/bib/bbac279.

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**:1–9.

Brede G, Solheim J, Tröen G, Prydz H. 2000. Characterization of PSKH1, a novel human protein serine kinase with centrosomal, golgi, and nuclear localization. *Genomics* **70**:82–92 DOI 10.1006/geno.2000.6365.

Brognard J, Hunter T. 2011. Protein kinase signaling networks in cancer. *Current Opinion in Genetics and Development* **21**:4–11 DOI 10.1016/j.gde.2010.10.012.

Byrne D, Shrestha S, Daly LA, Marensi V, Ramakrishnan K, Eyers CE, Kannan N, Eyers PA. 2022. Evolutionary and cellular analysis of the 'dark' pseudokinase PSKH2. *Biochemical Journal* **480**:141–160 DOI 10.1042/bcj20220474.

Chen ZH, You Z-H, Guo Z-H, Yi H-C, Luo G-X, Wang Y-B. 2020. Prediction of drug–target interactions from multi-molecular network based on deep walk embedding model. *Frontiers in Bioengineering and Biotechnology* **8**:338.

Collins KAL, Stuhlmiller TJ, Zawistowski JS, East MP, Pham TT, Hall CR, Goulet DR, Bevill SM, Angus SP, Velarde SH, Sciaky N, Oprea TI, Graves LM, Johnson GL, Gomez SM. 2018. Proteomic analysis defines kinase taxonomies specific for subtypes of breast cancer. *Oncotarget* **9**:15480–15497 DOI 10.18632/oncotarget.24337.

Cyganiak R, Wood D, Lanthaler M, Klyne G, Carroll JJ, McBride B. 2014. RDF 1.1 concepts and abstract syntax. *W3C Recommendation* **25**:1–22.

Dai Y, Wang S, Xiong NN, Guo W. 2020. A survey on knowledge graph embedding: approaches, applications and benchmarks. *Electronics* **9**:750 DOI 10.3390/electronics9050750.

Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, Mattingly CJ. 2021. Comparative toxicogenomics database (CTD): update 2021. *Nucleic Acids Research* **49**:D1138–D1143 DOI 10.1093/nar/gkaa891.

Dong Y, Chawla NV, Swami A. 2017. metapath2vec: scalable representation learning for heterogeneous networks. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. New York: ACM, 135–144.

Fant CB, Taatjes DJ. 2019. Regulatory functions of the Mediator kinases CDK8 and CDK19. *Transcription* **10**:76–90 DOI 10.1080/21541264.2018.1556915.

Fu T-Y, Lee W-C, Lei Z. 2017. Hin2vec: explore meta-paths in heterogeneous information networks for representation learning. In: *Proceedings of the 2017 ACM on conference on information and knowledge management*, 1797–1806.

Fu X, Zhang J, Meng Z, King I. 2020. Magnn: metapath aggregated graph neural network for heterogeneous graph embedding. In: *Proceedings of the web conference 2020*, 2331–2341.

Gao C, Sun C, Shan L, Lin L, Wang M. 2020. Rotate3d: representing relations as rotations in three-dimensional space for knowledge graph embedding. In: *Proceedings of the 29th ACM international conference on information & knowledge management*. New York: ACM, 385–394.

Gavali S, Ross K, Chen C, Cowart J, Wu CH. 2022. A knowledge graph representation learning approach to predict novel kinase-substrate interactions. *Molecular Omics Journal* **18**:853–864 DOI 10.1039/d1mo00521a.

Ge X, Wang Y-C, Wang B, Kuo C-CJ. 2022. Compounde: knowledge graph embedding with translation, rotation and scaling compound operations. ArXiv preprint. arXiv:2207.05324.

Gene Ontology C. 2021. The gene ontology resource: enriching a gold mine. *Nucleic Acids Research* **49**:D325–D334 DOI 10.1093/nar/gkaa1113.

Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, Griss J, Sevilla C, Matthews L, Gong C, Deng C, Varusai T, Ragueneau E, Haider Y, May

B, Shamovsky V, Weiser J, Brunson T, Sanati N, Beckman L, Shao X, Fabregat A, Sidiropoulos K, Murillo J, Viteri G, Cook J, Shorser S, Bader G, Demir E, Sander C, Haw R, Wu G, Stein L, Hermjakob H, D'Eustachio P. 2022. The reactome pathway knowledgebase 2022. *Nucleic Acids Research* **50**:D687–D692 DOI 10.1093/nar/gkab1028.

Grover A, Leskovec J. 2016. node2vec: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* New York: ACM, 855–864.

Gyori BM, Bachman JA, Subramanian K, Muhlich JL, Galescu L, Sorger PK. 2017. From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology* **13**:954 DOI 10.15252/msb.20177651.

Ha J, Park S. 2022. NCMD: Node2vec-based neural collaborative filtering for predicting miRNA-disease association. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatic.*

Hamilton W, Ying Z, Leskovec J. 2017. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems* **30**:1–11.

Huang LC, Ross KE, Baffi TR, Drabkin H, Kochut KJ, Ruan Z, D'Eustachio P, McSkimming D, Arighi C, Chen C, Natale DA, Smith C, Gaudet P, Newton AC, Wu C, Kannan N. 2018. Integrative annotation and knowledge discovery of kinase post-translational modifications and cancer-associated mutations through federated protein ontologies and resources. *Scientific Reports* **8**:6518 DOI 10.1038/s41598-018-24457-1.

Huang X, Tang J, Tan Z, Zeng W, Wang J, Zhao X. 2021. Knowledge graph embedding by relational and entity rotation. *Knowledge-Based Systems* **229**:107310 DOI 10.1016/j.knosys.2021.107310.

Hunter T. 2009. Tyrosine phosphorylation: thirty years and counting. *Current Opinion in Cell Biology* **21**:140–146 DOI 10.1016/j.ceb.2009.01.028.

Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. 2020. The reactome pathway knowledgebase. *Nucleic Acids Research* **48**:D498–D503 DOI 10.1093/nar/gkz1031.

Johnson JL, Yaron TM, Huntsman EM, Kerelsky A, Song J, Regev A, Lin TY, Liberatore K, Cizin DM, Cohen BM, Vasan N, Ma Y, Krismer K, Robles JT, van de Kooij B, van Vlimmeren AE, Andree-Busch N, Kaufer NF, Dorovkov MV, Ryazanov AG, Takagi Y, Kastenhuber ER, Goncalves MD, Hopkins BD, Elemento O, Taatjes DJ, Maucuer A, Yamashita A, Degterev A, Uduman M, Lu J, Landry SD, Zhang B, Cossentino I, Linding R, Blenis J, Hornbeck PV, Turk BE, Yaffe MB, Cantley LC. 2023. An atlas of substrate specificities for the human serine/threonine kinome. *Nature* DOI 10.1038/s41586-022-05575-3.

Kanev GK, de Graaf C, Westerman BA, de Esch IJP, Kooistra AJ. 2021. KLIFS: an overhaul after the first 5 years of supporting kinase research. *Nucleic Acids Research* **49**:D562–D569 DOI 10.1093/nar/gkaa895.

Kania E, Roest G, Vervliet T, Parys JB, Bultynck G. 2017. IP(3) receptor-mediated calcium signaling and its role in autophagy in cancer. *Frontiers in Oncology* **7**:140 DOI 10.3389/fonc.2017.00140.

Kelleher KJ, Sheils TK, Mathias SL, Yang JJ, Metzger VT, Siramshetty VB, Nguyen DT, Jensen LJ, Vidovic D, Schurer SC, Holmes J, Sharma KR, Pillai A, Bologa CG, Edwards JS, Mathe EA, Oprea TI. 2023. Pharos 2023: an integrated resource for the understudied human proteome. *Nucleic Acids Research* **51**:D1405–D1416 DOI 10.1093/nar/gkac1033.

Kennedy EJ, Scott JD. 2015. Selective disruption of the AKAP signaling complexes. *Methods in Molecular Biology* **1294**:137–150 DOI 10.1007/978-1-4939-2537-7_11.

Keshavarzi A, Kannan N, Kochut K. 2021. RegPattern2Vec: link prediction in knowledge graphs. In: *2021 IEEE international IOT, electronics and mechatronics conference (IEMTRONICS)*. Piscataway: IEEE, 1–7.

Koncel-Kedziorski R, Bekal D, Luan Y, Lapata M, Hajishirzi H. 2019. Text generation from knowledge graphs with graph transformers. ArXiv preprint. arXiv:190402342.

Kultgen PL, Byrd SK, Ostrowski LE, Milgram SL. 2002. Characterization of an A-kinase anchoring protein in human ciliary axonemes. *Molecular Biology of the Cell* **13**:4156–4166 DOI 10.1091/mbc.e02-07-0391.

Lahiry P, Torkamani A, Schork NJ, Hegele RA. 2010. Kinase mutations in human disease: interpreting genotype-phenotype relationships. *Nature Reviews Genetics* **11**:60–74 DOI 10.1038/nrg2707.

Li G, Luo J, Wang D, Liang C, Xiao Q, Ding P, Chen H. 2020. Potential circRNA-disease association prediction using DeepWalk and network consistency projection. *Journal of biomedical informatics* **112**:103624.

Lin Y, Liu Z, Sun M, Liu Y, Zhu X. 2015. Learning entity and relation embeddings for knowledge graph completion. In: *Twenty-ninth AAAI conference on artificial intelligence*. Washington, D.C.: AAAI.

Long Y, Luo J. 2020. Association mining to identify microbe drug interactions based on heterogeneous network embedding representation. *IEEE Journal of Biomedical and Health Informatics* **25**:266–275.

Malumbres M. 2014. Cyclin-dependent kinases. *Genome Biology* **15**:1–10 DOI 10.1186/gb-2014-15-1-r1.

Manchev VT, Hilpert M, Berrou E, Elaib Z, Aouba A, Boukour S, Souquere S, Pierron G, Rameau P, Andrews R, Lanza F, Bobe R, Vainchenker W, Rosa JP, Bryckaert M, Debili N, Favier R, Raslova H. 2014. A new form of macrothrombocytopenia induced by a germ-line mutation in the PRKACG gene. *Blood* **124**:2554–2563 DOI 10.1182/blood-2014-01-551820.

McSkimming DI, Dastgheib S, Talevich E, Narayanan A, Katiyar S, Taylor SS, Kochut K, Kannan N. 2015. ProKinO: a unified resource for mining the cancer kinome. *Human Mutation* **36**:175–186 DOI 10.1002/humu.22726.

Meng C, Cheng R, Maniu S, Senellart P, Zhang W. 2015. Discovering meta-paths in large heterogeneous information networks. In: *Proceedings of the 24th international conference on World Wide Web*, 754–764.

Minervini P, Fanizzi N, d'Amato C, Esposito F. 2015. Scalable learning of entity and predicate embeddings for knowledge graph completion. In: *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*. Piscataway: IEEE, 162–167.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: the protein families database in 2021. *Nucleic Acids Research* **49**:D412–D419 DOI 10.1093/nar/gkaa913.

Moret N, Liu C, Gyori BM, Bachman JA, Steppi A, Hug C, Taujale R, Huang L-C, Berginski ME, Gomez SM. 2021. A resource for exploring the under-studied human kinome for research and therapeutic opportunities. *BioRxiv*. DOI 10.1101/2020.04.02.022277.

Najar MA, Rex DAB, Modi PK, Agarwal N, Dagamajalu S, Karthikkeyan G, Vijayakumar M, Chatterjee A, Sankar U, Prasad TSK. 2021. A complete map of the calcium/calmodulin-dependent protein kinase kinase 2 (CAMKK2) signaling pathway. *Cell Communication and Signaling* **15**:283–290 DOI 10.1007/s12079-020-00592-1.

Nickel M, Murphy K, Tresp V, Gabrilovich E. 2015. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE* **104**:11–33 DOI 10.1109/JPROC.2015.2483592.

Nováček V, McGauran G, Matallanas D, Vallejo Blanco A, Conca P, Muñoz E, Costabello L, Kanakaraj K, Nawaz Z, Walsh B, Mohamed SK, Vandenbussche PY, Ryan CJ, Kolch W, Fey D. 2020. Accurate prediction of kinase-substrate networks using knowledge graphs. *PLOS Computational Biology* **16**:e1007578 DOI 10.1371/journal.pcbi.1007578.

Ochoa D, Jarnuczak AF, Vieitez C, Gehre M, Soucheray M, Mateus A, Kleefeldt AA, Hill A, Garcia-Alonso L, Stein F, Krogan NJ, Savitski MM, Swaney DL, Vizcaino JA, Noh KM, Beltrao P. 2020. The functional landscape of the human phosphoproteome. *Nature Biotechnology* **38**:365–373 DOI 10.1038/s41587-019-0344-3.

Omar MH, Scott JD. 2020. AKAP signaling Islands: venues for precision pharmacology. *Trends in Pharmacological Sciences* **41**:933–946 DOI 10.1016/j.tips.2020.09.007.

Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, Gomez SM, Guha R, Hersey A, Holmes J, Jadhav A, Jensen LJ, Johnson GL, Karlson A, Leach AR, Ma'ayan A, Malovannaya A, Mani S, Mathias SL, McManus MT, Meehan TF, Mering Cvon, Muthas D, Nguyen DT, Overington JP, Papadatos G, Qin J, Reich C, Roth BL, Schurer SC, Simeonov A, Sklar LA, Southall N, Tomita S, Tudose I, Ursu O, Vidovic D, Waller A, Westergaard D, Yang JJ, Zahoranszky-Kohalmi G. 2018. Unexplored therapeutic opportunities in the human genome. *Nature Reviews Drug Discovery* **17**:317–332 DOI 10.1038/nrd.2018.14.

Ould Amer Y, Hebert-Chatelain E. 2018. Mitochondrial cAMP-PKA signaling: what do we really know? *Biochimica et Biophysica Acta - Bioenergetics* **1859**:868–877 DOI 10.1016/j.bbabio.2018.04.005.

Palencia-Campos A, Aoto PC, Machal EMF, Rivera-Barahona A, Soto-Bielicka P, Bertinetti D, Baker B, Vu L, Piceci-Sparascio F, Torrente I, Boudin E, Peeters S,

Van Hul W, Huber C, Bonneau D, Hildebrand MS, Coleman M, Bahlo M, Bennett MF, Schneider AL, Scheffer IE, Kibæk M, Kristiansen BS, Issa MY, Mehrez MI, Ismail S, Tenorio J, Li G, Skålhegg BS, Otaify GA, Temtamy S, Aglan M, Jønch AE, De Luca A, Mortier G, Cormier-Daire V, Ziegler A, Wallis M, Lapunzina P, Herberg FW, Taylor SS, Ruiz-Perez VL. 2020. Germline and mosaic variants in PRKACA and PRKACB cause a multiple congenital malformation syndrome. *American Journal of Human Genetics* **107**:977–988 DOI 10.1016/j.ajhg.2020.09.005.

Pei H, Wei B, Chang KC-C, Lei Y, Yang B. 2020. Geom-gcn: geometric graph convolutional networks. ArXiv preprint. arXiv:2002.05287.

Peng J, Guan J, Shang X. 2019. Predicting parkinson's disease genes based on node2vec and autoencoder. *Frontiers in Genetics* **10**:226 DOI 10.3389/fgene.2019.00226.

Peng T, Qi B, He J, Ke H, Shi J. 2020. Advances in the development of phosphodiesterase-4 inhibitors. *Journal of Medicinal Chemistry* **63**:10594–10617 DOI 10.1021/acs.jmedchem.9b02170.

Perozzi B, Al-Rfou R, Skiena S. 2014. Deepwalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, New York: ACM, 701–710.

Pham P, Do P. 2019. W-MetaPath2Vec: the topic-driven meta-path-based model for large-scaled content-based heterogeneous information network representation learning. *Expert Systems with Applications* **123**:328–344 DOI 10.1016/j.eswa.2019.01.015.

Ravanmehr V, Blau H, Cappelletti L, Fontana T, Carmody L, Coleman B, George J, Reese J, Joachimiak M, Bocci G, Hansen P, Bult C, Rueter J, Casiraghi E, Valentini G, Mungall C, Oprea TI, Robinson PN. 2021. Supervised learning with word embeddings derived from PubMed captures latent knowledge about protein kinases and cancer. *NAR Genomics and Bioinformatics* **3**:lqab113 DOI 10.1093/nargab/lqab113.

Ross KE, Huang H, Ren J, Arighi CN, Li G, Tudor CO, Lv M, Lee J-Y, Chen S-C, Vijay-Shanker K. 2017. iPTMnet: integrative bioinformatics for studying PTM networks. In: Wu CH, Arighi CN, Ross KE, eds. *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics*. New York, NY: Springer, 333–353.

Sadeghian A, Armandpour M, Colas A, Wang DZ. 2021. Chronor: rotation based temporal knowledge graph embedding. In: *Proceedings of the AAAI conference on artificial intelligence*, 6471–6479.

Samizadeh M, Minaei-Bidgoli B. 2020. Drug-target interaction prediction by Metapath2vec node embedding in heterogeneous network of interactions. *International Journal on Artificial Intelligence Tools* **29**:2050001.

Sarma GN, Kinderman FS, Kim C, von Daake S, Chen L, Wang BC, Taylor SS. 2010. Structure of D-AKAP2:PKA RI complex: insights into AKAP specificity and selectivity. *Structure* **18**:155–166 DOI 10.1016/j.str.2009.12.012.

Sheils TK, Mathias SL, Kelleher KJ, Siramshetty VB, Nguyen DT, Bologa CG, Jensen LJ, Vidović D, Koleti A, Schürer SC, Waller A, Yang JJ, Holmes J, Bocci G, Southall N, Dharkar P, Mathé E, Simeonov A, Oprea TI. 2021. TCRD and Pharos 2021: mining the human proteome for disease biology. *Nucleic Acids Research* **49**:D1334–D1346 DOI 10.1093/nar/gkaa993.

Shi C, Li Y, Zhang J, Sun Y, Philip SY. 2016. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering* **29**:17–37.

Shi C, Lu Y, Hu L, Liu Z, Ma H. 2020. Rhine: relation structure-aware heterogeneous information network embedding. *IEEE Transactions on Knowledge and Data Engineering* **34**:433–447.

Soleymani S, Gravel N, Huang L-C, Yeung W, Bozorgi E, Bendzunas NG, Kochut KJ, Kannan N. 2022. Dark kinase annotation, mining and visualization using the Protein Kinase Ontology. *bioRxiv*.

Stroukov W, Rösch A, Schwan C, Jeney A, Römer W, Thuenauer R. 2019. Synchronizing protein traffic to the primary cilium. *Frontiers in Genetics* **10**:163 DOI 10.3389/fgene.2019.00163.

Sun Z, Deng Z-H, Nie J-Y, Tang J. 2019. Rotate: knowledge graph embedding by relational rotation in complex space. ArXiv preprint. arXiv:190210197.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV. 2019. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**:D607–D613 DOI 10.1093/nar/gky1131.

Tang M, Pei G, Su D, Wang C, Feng X, Srivastava M, Chen Z, Zhao Z, Chen J. 2021. Genome-wide CRISPR screens reveal cyclin C as synthetic survival target of BRCA2. *Nucleic Acids Research* **49**:7476–7491 DOI 10.1093/nar/gkab540.

Tang S, Li B, Yu H. 2019. ChebNet: efficient and stable constructions of deep neural networks with rectified power units using chebyshev approximations. ArXiv preprint. arXiv:1911.05467.

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, Forbes SA. 2019. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research* **47**:D941–D947 DOI 10.1093/nar/gky1015.

UniProt C. 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**:D480–D489 DOI 10.1093/nar/gkaa1100.

Velankar S, Alhroub Y, Alili A, Best C, Boutselakis HC, Caboche S, Conroy MJ, Dana JM, Van Ginkel G, Golovin A. 2010. PDBe: protein data bank in Europe. *Nucleic Acids Research* **39**:D402–D410.

Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. 2017. Graph attention networks. ArXiv preprint. arXiv:1710.10903.

W3C. 2014. RDF Schema 1.1. *Available at https://www.w3.org/TR/rdf-schema/* (accessed on 31 December 2020).

Wan G, Du B, Pan S, Haffari G. 2020. Reinforcement learning based meta-path discovery in large-scale heterogeneous information networks. In: *Proceedings of the AAAI conference on artificial intelligence*, 6094–6101.

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**:2724–2743 DOI 10.1109/TKDE.2017.2754499.

Wang S, Fu K, Sun X, Zhang Z, Li S, Jin L. 2021. Hierarchical-aware relation rotational knowledge graph embedding for link prediction. *Neurocomputing* **458**:259–270 DOI 10.1016/j.neucom.2021.05.093.

Weber H, Garabedian MJ. 2018. The mediator complex in genomic and non-genomic signaling in cancer. *Steroids* **133**:8–14 DOI 10.1016/j.steroids.2017.11.007.

Wei L, Zhang R, Zhang J, Li J, Kong D, Wang Q, Fang J, Wang L. 2021. PRKAR2A deficiency protects mice from experimental colitis by increasing IFN-stimulated gene expression and modulating the intestinal microbiota. *Mucosal Immunology* **14**:1282–1294 DOI 10.1038/s41385-021-00426-2.

Wong L, You Z-H, Guo Z-H, Yi H-C, Chen Z-H, Cao M-Y. 2020. MIPDH: a novel computational model for predicting microRNA–mRNA interactions by DeepWalk on a heterogeneous network. *ACS Omega* **5**:17022–17032.

Xie X, Zhang N, Li Z, Deng S, Chen H, Xiong F, Chen M, Chen H. 2022. From discrimination to generation: knowledge graph completion with generative transformer. ArXiv preprint. arXiv:220202113.

Xie Z, Yang X, Duan Y, Han J, Liao C. 2021. Small-molecule kinase inhibitors for the treatment of nononcologic diseases. *Journal of Medicinal Chemistry* **64**:1283–1345 DOI 10.1021/acs.jmedchem.0c01511.

Yang C, Liu M, He F, Zhang X, Peng J, Han J. 2018. Similarity modeling on heterogeneous networks via automatic path discovery. In: Berlingerio M, Bonchi F, Gärtner T, Hurley N, Ifrim G, eds. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science*, vol. 11052. Cham: Springer DOI 10.1007/978-3-030-10928-8_3.

Yang H, Liu J. 2021. Knowledge graph representation learning as groupoid: unifying TransE, RotatE, QuatE, ComplEx. In: *Proceedings of the 30th ACM international conference on information & knowledge management*, New York: ACM, 2311–2320.

Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, Ramaswamy S, Futreal PA, Haber DA, Stratton MR, Benes C, McDermott U, Garnett MJ. 2013. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research* **41**:D955–D961 DOI 10.1093/nar/gks1111.

Yao L, Mao C, Luo Y. 2019. KG-BERT: BERT for knowledge graph completion. ArXiv preprint. arXiv:1909.03193.

Yelland T, Garcia E, Samarakoon Y, Ismail S. 2021. The structural and biochemical characterization of UNC119B cargo binding and release mechanisms. *Biochemistry* **60**:1952–1963 DOI 10.1021/acs.biochem.1c00251.

Zhang J, Shi X, Xie J, Ma H, King I, Yeung D-Y. 2018. Gaan: gated attention networks for learning on large and spatiotemporal graphs. ArXiv preprint. arXiv:1803.07294.

Zhang J, Shi X, Zhao S, King I. 2019. Star-gcn: stacked and reconstructed graph convolutional networks for recommender systems. ArXiv preprint. arXiv:1905.13129.

Zhou X, Yi Y, Jia G. 2021. Path-RotatE: knowledge graph embedding by relational rotation of path in complex space. In: *2021 IEEE/CIC international conference on communications in China (ICCC)*. Piscataway: IEEE, 905–910.