Supplement

BioKEEN: A library for learning and evaluating biological knowledge graph embeddings

Outline

- **Supplementary Table S1:** Knowledge Graph Embedding Models implemented in BioKEEN
- **Supplementary Table S2**: Databases included in BioKEEN via Bio2BEL
- **Supplementary Table S3**: Hyper-Parameter optimization results on ComPath
- **Supplementary Figure S1:** PyKEEN Architecture
- **Software Installation and Documentation**
- **DevOps:** Following Python Community Standards
- **Case Scenarios**

Reference	Name	Description
Bordes, *et al.* [1]	TransE	Considers a relation as a translation from the head to the tail entity.
Wang, *et al*. [2]	TransH	Extends TransE by applying the translation from head to tail entity in a relational-specific hyperplane.
Lin, *et al* [3]	TransR	Extends TransE and TransH by considering different vector spaces for entities and relations.
Ji, *et al.* [4]	TransD	Extends TransR to use fewer parameters.
Dettmers, *et al.* [5]	ConvE	Uses a convolutional neural network (CNN) for applying link prediction. An input instance is represented by the subject and predicate embeddings, these are rescaled to represent an "image", and on the rescaled input the CNN is applied. The CNN will predict the most probable object.
Bordes, *et al.* [6]	SE	Projects the head and tail entity into different matrices for each relation.
Bordes, *et al.* [7]	UM	Simplifies TransE by ignoring relation embeddings.
Nickel, *et al.* [8]	RESCAL	Represents relations as matrices and models interactions between latent features.
Dong, *et al.* [9]	ERMLP	Neural network based approach in which subject, predicate and object are concatenated and fed to the neural network.
Yang, *et al.* [10]	DistMult	Simplifies RESCAL by restricting matrices representing relations as diagonal matrices.

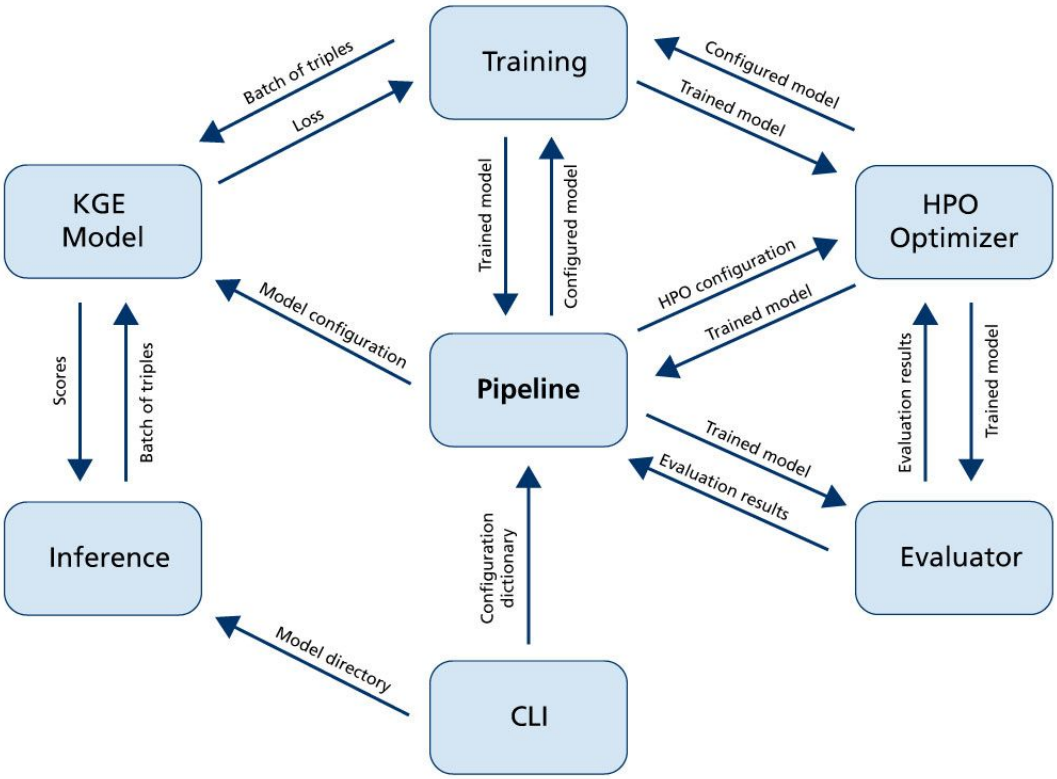
Supplementary Table S1. Knowledge graph embedding models implemented in BioKEEN.

Reference	Database	Type	Bio2BEL Zenodo
[11, 12]	ADEPTUS	disease-differential expressed genes	[13]
[14]	ComPath	pathway-pathway	[15]
[16]	DrugBank	drug-target	[17]
[18]	ExPASy	protein-enzyme class	[19]
[20]	HIPPIE	protein-protein interaction	[21]
[12, 22]	HSDN	disease-symptoms	[23]
[24]	KEGG	protein-pathway	[25]
[26]	mirTarBase	miRNA-target	[27]
[28]	MSigDB	protein-pathway	[29]
[30]	Reactome	protein-pathway	[31]
[32]	SIDER	drug-side effect	[33]
[34]	InterPro	protein-domains and protein-family	[35]
[36]	WikiPathways	protein-pathway	[37]

Supplementary Table S2. Biological databases included in BioKEEN via Bio2BEL until the date.

Model	EED	RED	LR	Loss Function	Margin	Normalization	Scoring Function	Batch Size	Epochs	WSC	FCT	Seed	Mean Rank	Hits@10 (%)
TransE	150	-	0.01	MRL	12.0	L2	L1	32	2500	-	Yes	2	**131.44**	**63.20**
	50	-	0.01	MRL	5.0	L2	L1	32	2200	-	Yes	2	130.99	62.08
	50	-	0.01	MRL	1.0	L2	L1	32	1000	-	Yes	2	226.03	19.10
TransH	50	-	0.01	MRL	4.0	-	L2	400	1500	0.03	Yes	2	481.08	25.00
TransR	30	50	0.01	MRL	0.5	-	L1	32	1500	-	Yes	2	200.13	41.01
DistMult	150	-	0.01	MRL	15.0	-	-	32	2000	-	Yes	2	230.33	38.48
UM	200	-	0.01	MRL	15.0	L2	L2	64	1000	-	Yes	2	224.83	43.26

Supplementary Table S3. Hyper-Parameter optimization results of TransE, TransH, TransR, DistMult, and UM evaluations on ComPath. We used a random 90:10 split for splitting the dataset into a training and test set. Acronyms: EED (Number of Entity Embedding Dimensions), FCT (Filtering Corrupted Triples), LR (Learning Rate), and MRL (Margin Ranking Loss), RED (Number of Relation Embedding Dimensions), WSC (Weight value for Soft Constraints). Different parameter settings are presented for TransE to illustrate the sensitivity of choosing appropriate hyper-parameter values.



Supplementary Figure S1. A diagram of the modular software architecture of PyKEEN showing the flow of information between its components. The *pykeen.Pipeline* Python module controls the workflow by interacting with the different modules. Depending on the setting in which the software is called, different modules are active.

Software Installation and Documentation

All packages described in the manuscript are available under the MIT license through GitHub (https://github.com/) and PyPI (https://pypi.org), the main packaging system for Python. All relevant information for installation is bundled in the package, so it can be easily and quickly installed independently of the operating system, running any modern version of the Python programming language. The documentation for all packages is built using the Python documenting tool *Sphinx* and is accessible at Read the Docs (https://pykeen.readthedocs.io/ and https://biokeen.readthedocs.io/). Releases to PyKEEN and BioKEEN are also tracked by Zenodo (https://zenodo.org) under [36] and [37], respectively.

DevOps

BioKEEN makes full use of the standard scientific Python stack (NumPy, SciPy, Scikit-Learn, Pandas, Jupyter) for standard mathematical operations, I/O, standard data processing, and for presentation. It uses PyTorch, one of the standards for machine learning applications for implementing the knowledge graph embedding models.

In line with community standards, BioKEEN uses flake8 to enforce code quality, setuptools to build distributions, pyroma to enforce package metadata standards, sphinx to build documentation, Read the Docs to host documentation, py.test as a testing harness, and Travis-CI as a continuous integration server to run each of these with each commit (https://travis-ci.com/SmartDataAnalytics/BioKEEN).

Because it relies on the pre-built wheels for PyTorch and the other standard Python libraries, it can be included as a requirement for other projects directly in their requirements.txt file, or included in the setup.py for other Python packages in the install_requires setting without the need for complicated build steps or any user configuration.

Case Scenarios

The technical documentation for the package is supplemented with several Jupyter notebooks outlining applications to real-world biological data sets (https://github.com/SmartDataAnalytics/BioKEEN/tree/master/notebooks) as well as a video tutorial to illustrate the command line interface process (https://vimeo.com/314252656).

References

1. Bordes, A., *et al.* (2013). Translating embeddings for modeling multi-relational data. *NIPS.*.
2. Wang, Z., *et al*. (2014). Knowledge Graph Embedding by Translating on Hyperplanes. *AAAI. Vol. 14.*
3. Lin, Y., *et al.* (2015). Learning entity and relation embeddings for knowledge graph completion. *AAAI. Vol. 15.*
4. Ji, G., *et al. (2015)*. Knowledge graph embedding via dynamic mapping matrix. *ACL.*
5. Dettmers, T., *et al.* (2017) Convolutional 2d knowledge graph embeddings. *arXiv preprint arXiv:1707.01476.*
6. Bordes, A., *et al.* (2011). Learning Structured Embeddings of Knowledge Bases. *AAAI. Vol. 6. No. 1.*
7. Bordes, A., *et al.* (2014). A semantic matching energy function for learning with multi-relational data. *Machine Learning* 94.2 : 233-259.
8. Nickel, M., *et al.* (2011) A Three-Way Model for Collective Learning on Multi-Relational Data. *ICML. Vol. 11.*
9. Dong, X., *et al.* (2014) Knowledge vault: A web-scale approach to probabilistic knowledge fusion. *ACM.*
10. Yang, B. *et al.* Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575.*
11. Amar, D., *et al.* (2015). Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets. *Nucleic Acids Research*, *43*(16), 7779–7789.
12. Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., … Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *ELife*, *6*.
13. Charles Tapley Hoyt. (2018, November 20). bio2bel/adeptus v0.0.1 (Version v0.0.1). Zenodo. http://doi.org/10.5281/zenodo.1492143
14. Domingo-Fernández, *et al.* (2018). ComPath: An ecosystem for exploring, analyzing, and curating pathway databases. *npj Syst Biol Appl.* 5(1):3.
15. Daniel Domingo-Fernández, & Charles Tapley Hoyt. (2018, November 20). ComPath/resources v0.0.6 (Version v0.0.6). Zenodo. http://doi.org/10.5281/zenodo.1492201
16. Wishart, D. S., *et al.* (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, *46*(D1), D1074–D1082.
17. Charles Tapley Hoyt. (2018, May 8). bio2bel/drugbank v0.0.1 (Version v0.0.1). Zenodo. http://doi.org/10.5281/zenodo.1243727
18. Artimo, P., *et al.* (2012) ExPASy: SIB bioinformatics resource portal. *Nucleic acids research* 40.W1: W597-W603.
19. Aram Grigoryan, Charles Tapley Hoyt, & Christian Ebeling. (2018, November 15). bio2bel/expasy v0.2.0 (Version v0.2.0). Zenodo. http://doi.org/10.5281/zenodo.1489291
20. Alanis-Lobato, G., *et al.* (2017). HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, *45*(D1), D408–D414.
21. Charles Tapley Hoyt. (2018, September 26). bio2bel/hippie v0.0.1 (Version v0.0.1). Zenodo. http://doi.org/10.5281/zenodo.1435930
22. Zhou, X., *et al.* (2014). Human symptoms–disease network. *Nature communications*, *5*, 4212.
23. Charles Tapley Hoyt. (2018, November 20). bio2bel/hsdn v0.0.1 (Version v0.0.1). Zenodo. http://doi.org/10.5281/zenodo.1492163

24. Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, *28*(1), 27-30.
25. Daniel Domingo-Fernández, & Charles Tapley Hoyt. (2018, November 13). bio2bel/kegg v0.2.1 (Version v0.2.1). Zenodo. http://doi.org/10.5281/zenodo.1486130
26. Chou, C. H., *et al.* (2016). miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Research*, *44*(D1), D239–D247.
27. Charles Tapley Hoyt, & Colin Birkenbihl, (2018, May 8). bio2bel/mirtarbase v0.1.2 (Version v0.1.2). Zenodo. http://doi.org/10.5281/zenodo.1243731
28. Liberzon, A., *et al.* (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), 1739-1740.
29. Daniel Domingo-Fernández, & Charles Tapley Hoyt. (2018, October 13). bio2bel/msig: KEEN compatible (Version 0.1.1). Zenodo. http://doi.org/10.5281/zenodo.1461390
30. Joshi-Tope, G., *et al.* (2005). Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, *33*(suppl_1), D428-D432.
31. Daniel Domingo-Fernández, & Charles Tapley Hoyt. (2018, October 13). bio2bel/reactome: KEEN compatible (Version 0.1.4). Zenodo. http://doi.org/10.5281/zenodo.1461389
32. Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2015). The SIDER database of drugs and side effects. *Nucleic acids research*, *44*(D1), D1075-D1079.
33. Charles Tapley Hoyt. (2018, December 3). bio2bel/sider v0.0.1 (Version v0.0.1). Zenodo. http://doi.org/10.5281/zenodo.1882884
34. Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., … Mitchell, A. L. (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research*, *45*(D1), D190–D199. https://doi.org/10.1093/nar/gkw1107
35. Charles Tapley Hoyt, & Aram Grigoryan. (2018, May 8). bio2bel/interpro v0.1.1 (Version v0.1.1). Zenodo. http://doi.org/10.5281/zenodo.1243733
36. Pico, A. R., *et al.* (2008). WikiPathways: pathway editing for the people. *PLoS biology*, *6*(7), e184.
37. Daniel Domingo-Fernández, & Charles Tapley Hoyt. (2018, November 13). bio2bel/wikipathways v0.2.0 (Version v0.2.0). Zenodo. http://doi.org/10.5281/zenodo.1486073
38. Mehdi Ali, Charles Tapley Hoyt, Daniel Domingo-Fernández, & Gezim Sejdiu. (2018, November 19). SmartDataAnalytics/PyKEEN v0.0.12 (Version v0.0.12). Zenodo. http://doi.org/10.5281/zenodo.1491380
39. Charles Tapley Hoyt, Mehdi Ali, & Daniel Domingo-Fernández. (2018, November 8). SmartDataAnalytics/BioKEEN v0.0.3 (Version v0.0.3). Zenodo. http://doi.org/10.5281/zenodo.1480774