



netidee

PROJEKTE

OpenBioLink

Endbericht | Call 15 | Projekt ID 5171

Lizenz CC-BY

Inhalt

1	Einleitung.....	3
2	Projektbeschreibung.....	3
3	Verlauf der Arbeitspakete.....	4
3.1	Arbeitspaket 1 - <i>Detailplanung und Formales am Projektstart</i>	4
3.2	Arbeitspaket 2 - <i>Benchmark Erweiterung</i>	4
3.3	Arbeitspaket 3 - <i>OpenBioLink Explorer tool</i>	6
3.4	Arbeitspaket 4 - <i>Durchführung und Dokumentation der "OpenBioLink 2021 Challenge"</i>	9
3.5	Arbeitspaket 5 - <i>Dokumentation und Formales am Projektende</i>	11
4	Umsetzung Förderauflagen.....	11
5	Liste Projektergebnisse.....	11
6	Verwertung der Projektergebnisse in der Praxis.....	13
7	Öffentlichkeitsarbeit/ Vernetzung.....	13
8	Eigene Projektwebsite.....	14
9	Geplante Aktivitäten nach netidee-Projektende.....	14
10	Anregungen für Weiterentwicklungen durch Dritte.....	14

1 Einleitung

Unser Projekt hilft biomedizinischen ForscherInnen, mittels AI und Webtechnologie schneller Zusammenhänge im Netzwerk biomedizischer Forschungsergebnisse zu finden, und dadurch vielversprechende Hypothesen zu formulieren und zu testen (z.B. 'Medikament X hilft bei Krankheit Y'). Andererseits hilft unser Projekt AI EntwicklerInnen, ihre Modelle zu testen und zu verbessern.

Im Projekt entwickelten wir auch generell domänenübergreifend anwendbare Algorithmen und Tools für Explainable AI und Link Prediction in Wissensgraphen.

2 Projektbeschreibung

Durchbrüche im Bereich der AI (Deep Learning, Vector Space Embeddings) haben zur Entwicklung einer Vielzahl neuer Algorithmen für Link Prediction geführt — der AI-gestützten Vorhersage von neuen Verbindungen in großen Netzwerken.

Dies bietet großes Potential für die biomedizinische Forschung -- denn ein wichtiges Ziel hierbei ist es, aus bestehenden Zusammenhängen (z.B. Gen-Protein-Prozess-Krankheit-Therapie) neue Zusammenhänge zu schließen.

Unser Projekt macht dieses Potential nutzbar. Wir entwickelten die OpenBioLink Software Suite:

- Den OpenBioLink Datensatz, ein großes, aus verschiedenen Web-Ressourcen aggregiertes Netzwerk biomedizinischen Wissens das, in den Linked Data Formaten zugänglich gemacht wird. Die Benchmark-Datenset zum Trainieren und Testen von Link Prediction Modellen wurde erweitert und ein standardisiertes Verfahren zur Evaluierung neuer AI Modelle geschaffen.
- Versatil einsetzbare Tools für die Web-basierte Interpretation von Vorhersagen von Links in großen Wissensnetzwerken wurden geschaffen, welche den internationalen State-of-the-Art in diesem Bereich vorantrieben. Daraus gingen auch zwei wissenschaftliche Publikationen hervor.

Einerseits hilft unser Projekt biomedizinischen ForscherInnen, mittels AI und Webtechnologie schneller Zusammenhänge im Netzwerk biomedizischer Forschungsergebnisse zu finden, und dadurch vielversprechende Hypothesen zu formulieren und zu testen (z.B. 'Medikament X hilft bei Krankheit Y').

Andererseits hilft unser Projekt AI EntwicklerInnen, ihre Modelle zu testen und zu verbessern.

3 Verlauf der Arbeitspakete

Aufgrund von COVID kam es zu einer Verzögerung der Stellenbesetzung, wesentliche Arbeitspakete wurden daher erst im Mai 2021 gestartet und verdichtet. Der spätere Start kann erfolgreich kompensiert werden (mehr Arbeitsstunden pro Monat, mehr Involvement von Senior Researchers/Developers).

Netidee wurde in allen Projektseiten bzw. Publikationen als Fördergeber gelistet.

3.1 **Arbeitspaket 1 - *Detailplanung und Formales am Projektstart***

Der Projektstart wurde erfolgreich abgewickelt (Vertragsprüfung, Unterzeichnung des Vertrages, Projektplanung basierend auf Excel-Vorlage, Start der Projektwebsite, Förderabruf für Förderrate 1).

3.2 **Arbeitspaket 2 - *Benchmark Erweiterung***

Eine umfassende Analyse von weiteren Datenquellen wurde erstellt und Anknüpfungspunkte bzw. Erweiterungen des derzeitigen Wissensgraphen wurden identifiziert. Dieses Mapping hilft Anwendern, unseren Wissensgraphen mit anderen Wissensquellen zu verknüpfen und zu erweitern.

Der Code für die Durchführung des Benchmarks (generieren von Vorhersagen, generieren von Scores) wurde neu implementiert und erlaubt nun eine bessere Vergleichbarkeit verschiedener Methode. Dies ist von grundlegender Bedeutung auch für die faire Durchführung der „Challenge“ (Arbeitspaket 4).

Dokumentation wurde erweitert.



OpenBioLink

pypi v0.1.4
docs passing
license MIT
Uptime passing

OpenBioLink is a resource and evaluation framework for evaluating link prediction models on heterogeneous biomedical graph data. It contains benchmark datasets as well as tools for creating custom benchmarks and evaluating models.

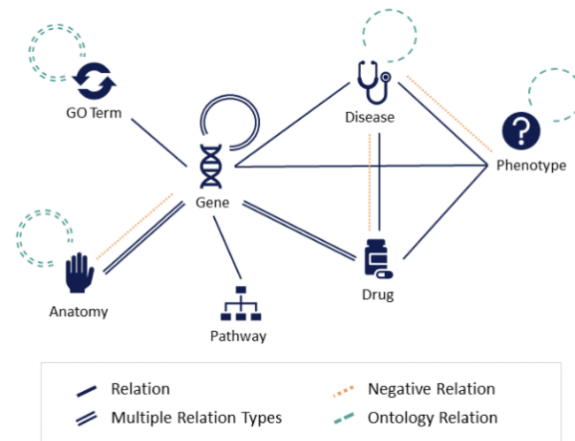


Fig. 1. An overview of the OpenBioLink benchmark graph.

[Documentation](#)

[Paper preprint on arXiv](#) • [Peer reviewed paper in the journal Bioinformatics \(for citations\)](#) • [Supplementary data](#)

Ausschnitt der OpenBioLink dataset / benchmark Github Seite

<https://github.com/OpenBioLink/OpenBioLink>

3.3 Arbeitspaket 3 - OpenBioLink Explorer tool

Das OpenBioLink Explorer Softwarepaket und User Interface wurden entwickelt und fertig gestellt. Dieses Interface erlaubt die Explorierung von Zusammenhängen in großen Wissensbasen, sowie eine Darstellung der Vorhersagen, die durch unseren Explainable Link Prediction Algorithmus generiert werden.

Der Explainable Link Prediction Algorithmus (SAFRAN) wurde im Projekt weiter entwickelt und liefert State-of-the-Art Ergebnisse. Eine wissenschaftliche Publikation zu dem Algorithmus wurde fertiggestellt und publiziert.

SAFRAN: An interpretable, rule-based link prediction method outperforming embedding models

Simon Ott

SIMON.OTT@MEDUNIWIEN.AC.AT

Institute of Artificial Intelligence and Decision Support, Medical University of Vienna, Austria

Christian Meilicke

CHRISTIAN@INFORMATIK.UNI-MANNHEIM.DE

Data and Web Science Research Group, University Mannheim, Germany

Matthias Samwald

MATTHIAS.SAMWALD@MEDUNIWIEN.AC.AT

Institute of Artificial Intelligence and Decision Support, Medical University of Vienna, Austria

Abstract

Neural embedding-based machine learning models have shown promise for predicting novel links in knowledge graphs. Unfortunately, their practical utility is diminished by their lack of interpretability. Recently, the fully interpretable, rule-based algorithm AnyBURL yielded highly competitive results on many general-purpose link prediction benchmarks. However, current approaches for aggregating predictions made by multiple rules are affected by redundancies. We improve upon AnyBURL by introducing the SAFRAN rule application framework, which uses a novel aggregation approach called Non-redundant Noisy-OR that detects and clusters redundant rules prior to aggregation. SAFRAN yields new state-of-the-art results for fully interpretable link prediction on the established general-purpose benchmarks FB15K-237, WN18RR and YAGO3-10. Furthermore, it exceeds the results of multiple established embedding-based algorithms on FB15K-237 and WN18RR and narrows the gap between rule-based and embedding-based algorithms on YAGO3-10.

Abstract des Papers zum SAFRAN Link Prediction Algorithmus

https://openreview.net/pdf?id=jCt9S_3w_S9



 CMake Build Matrix passing  docs passing

SAFRAN (Scalable and fast non-redundant rule application) is a framework for fast inference of groundings and aggregation of predictions of logical rules in the context of knowledge graph completion/link prediction. It uses rules learned by [AnyBURL](#) (Anytime Bottom Up Rule Learning), a highly-efficient approach for learning logical rules from knowledge graphs.

[Paper preprint on arXiv](#) • [AKBC 2021 conference paper \(for citations\)](#)

Documentation

Can be found [here](#).

Citation

```
@inproceedings{
  ott2021safran,
  title={{SAFRAN}: An interpretable, rule-based link prediction method outperforming embedding models},
  author={Simon Ott and Christian Meilicke and Matthias Samwald},
  booktitle={3rd Conference on Automated Knowledge Base Construction},
  year={2021},
  url={https://openreview.net/forum?id=jCt9S_3w_S9},
  doi={}
}
```

This project received funding from [netidee](#).

Ausschnitt der SAFRAN Github Seite
<https://github.com/OpenBioLink/SAFRAN>

Eine weitere Publikation zum LinkExplorer User Interface und Framework wurde ebenfalls fertiggestellt und im peer-reviewten Top-Journal Bioinformatics publiziert.

Data and text mining

LinkExplorer: Predicting, explaining and exploring links in large biomedical knowledge graphs

Simon Ott¹, Adriano Barbosa-Silva¹ and Matthias Samwald^{1,*}

¹Institute of Artificial Intelligence, Medical University of Vienna, Vienna, 1090, Austria

*To whom correspondence should be addressed.

Abstract


Summary: Machine learning algorithms for link prediction can be valuable tools for hypothesis generation. However, many current algorithms are black boxes or lack good user interfaces that could facilitate insight into why predictions are made. We present LinkExplorer, a software suite for predicting, explaining and exploring links in large biomedical knowledge graphs. LinkExplorer integrates our novel, rule-based link prediction engine SAFRAN, which was recently shown to outcompete other explainable algorithms and established black box algorithms. Here, we demonstrate highly competitive evaluation results of our algorithm on multiple large biomedical knowledge graphs, and release a web interface that allows for interactive and intuitive exploration of predicted links and their explanations.

Availability and Implementation: A publicly hosted instance, source code and further documentation can be found at <https://github.com/OpenBioLink/Explorer>.

Contact: matthias.samwald -at- meduniwien.ac.at

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Abstract des LinkExplorer Papers <https://www.biorxiv.org/content/10.1101/2022.01.09.475537v2>



LinkExplorer

The LinkExplorer is a web-based tool for exploring nodes and relations of link prediction benchmark datasets and explanations of predictions done with the rule-based approach SAFRAN. A running instance of this tool can be found at:

<http://explore.ai-strategies.org>

Included are three biomedical knowledge bases:

- OpenBioLink
- Hetionet
- Pheknowlator

and two general-domain benchmarks:

- YAGO3-10
- WN18RR

[Paper preprint on bioRxiv](#) • [Peer reviewed paper in the journal Bioinformatics \(for citations\)](#) • [Supplementary data](#) • [Citation \(bibTex\)](#)

Ausschnitt der LinkExplorer Github Seite
<https://github.com/OpenBioLink/explorer>

LinkExplorer Overview Entities Feedback Dataset: obl Explanation: max Load other

cytochrome P450 family 2 subfamily C member 18 can be overexpressed in liver
Confidence: 0.58576
because

Rule	Confidence	Correctly predicted	Predicted
aldehyde oxidase 1 is in an interaction with cytochrome P450 family 2 subfamily C member 18	0.58571	41	65
NADPH is catalyzed by cytochrome P450 family 2 subfamily C member 18	0.44218	65	142
cytochrome P450 family 2 subfamily C member 18 associated with Triphosphopyridine nucleotide	0.42056	90	209
cytochrome P450 family 2 subfamily C member 18 associated with 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone	0.37838	14	32
cytochrome P450 family 2 subfamily C member 18 can be underexpressed in frontal cortex	0.29116	214	730

Screenshot der Prediction-View des LinkExplorer Interfaces (aus <https://openbiolink.github.io/Explorer/tutorial/explain.html>)

3.4 Arbeitspaket 4 - Durchführung und Dokumentation der "OpenBioLink 2021 Challenge"



Das Regelwerk, Prozedere und Tools für die Durchführung der Challenge wurden fertiggestellt. Eine Website für die Challenge wurde erstellt (<https://openbiolink.github.io/>). Eine Bewerbung der Challenge in sozialen und professionellen Netzwerken wurde durchgeführt.




Challenge

To foster progress in link prediction on large-scale, heterogeneous biomedical data, we invite you to participate in the OpenBioLink Challenge. Link Prediction on knowledge graphs is a versatile paradigm for generating new insights about relationships between entities. It is especially important in fields such as biomedical research, where it can help with hypothesis generation, prioritizing drug targets or therapeutic substances for experimental screening etc.

Timeline

 Extended deadline to 10.12.2021 

Task	Deadline	
Team registration	-	Link
Submission deadline	10.12.2021 - 23:59 GMT-0	Link

There is no deadline for registering your team however, you'll need to register to OpenBioLink2021 before being able to make a submission.

Prizes

- Best predictive performance prize (500 €)
- Best explainability prize (500 €)

The **best predictive performance prize** will be awarded to the model with the best result in terms of predictive accuracy (measured with the Hits@10 metric). It can only be given to submissions that improve on current baseline results.

The **best explainability prize** will be awarded to innovative models that provide good explanations for their predictions. Explainability is an important and still largely unaddressed issue in biomedical link prediction, and we want to foster the creation of explainable algorithms. The OpenBioLink challenge team will evaluate the quality of explanations and the innovativeness of the approach. Predictive performance is not a major criterion for winning the explainability prize, but it must exceed a Hits@10 value of 0.440.

OpenBioLink 2021 Challenge Website

<https://openbiolink.github.io/>

Wider Erwarten schaffte aber leider keines der partizipierenden Teams, die von uns zuvor erreichten 'Baseline' Ergebnisse zu übertreffen und somit den Gewinn-Kriterien zu entsprechen. Es konnte somit leider kein Preis vergeben werden und keine Publikation zu den Gewinner-Systemen gemacht werden.

Trotz des Mangels an den Erfolgskriterien entsprechenden Submissions haben die in AP4 getätigten Outreach-Aktivitäten die Sichtbarkeit des OpenBioLink Benchmarks erhöht. Das für die Challenge geschaffene Evaluierungstoolset kann auch weiterhin verwendet werden und es ist anzunehmen, dass in Zukunft bessere AI Systeme unsere eigenen Baseline-Resultate übertreffen werden können.

3.5 Arbeitspaket 5 - Dokumentation und Formales am Projektende

Dokumentation und formales am Projektende wurden gemäß der Vorgaben abgewickelt. Gemäß der Vorgaben wurden Dokumentations-PDFs generiert; wir möchten aber darauf hinweisen, dass die Dokumentation über die diversen Projektwebseiten leichter zugänglich ist.

4 Umsetzung Förderauflagen

Aufgrund des Feedbacks in den Förderunterlagen wurde deutlich mehr Augenmerk auf die domänen-übergreifenden Entwicklungen im Projekt gelegt (d.h., auf jene Komponenten, die auch außerhalb der biomedizinischen Domäne relevant sind). Wir haben mit SAFRAN einen State-of-the-Art Algorithmus für Explainable Link Prediction geschaffen. Der LinkExplorer wurde ebenfalls für nicht-biomedizinische Use-Cases entwickelt und kann auch hier ein global wichtiges Tool darstellen.

5 Liste Projektergebnisse

Projektzwischenbericht	CC-BY-3.0 AT	https://netidee.at/openbiolink
Projektendbericht	CC-BY-3.0 AT	https://netidee.at/openbiolink

<p>Entwickler_innen-DOKUMENTATION des Projektergebnisses für andere Entwickler_innen ("Dritte"), die das Projektergebnis nach Projektende nutzen/weiterentwickeln wollen</p> <p><u>Für Entwickler_innen (Systemkonzept, ggf. Grobspezifikationen):</u></p> <p>a. WAS IST ES b. FÜR WEN IST ES /WEM HILFT ES WODURCH c. WIE FUNKTIONIERT ES (für Entwickler_innen: Übersicht und detailliertes Systemkonzept, SW-Struktur)</p>	<p>CC-BY-3.0 AT</p>	<p>https://netidee.at/openbiolink</p>
<p>Anwender_innen-DOKUMENTATION des Projektergebnisses für Anwender_innen, die das Projektergebnis nach Projektende nutzen wollen</p> <p><u>Für Anwender_innen ("Bedienungsanleitung") :</u></p> <p>a. WAS IST ES b. FÜR WEN IST ES /WEM HILFT ES WODURCH c. WIE FUNKTIONIERT ES</p>	<p>CC-BY-3.0 AT</p>	<p>https://netidee.at/openbiolink</p>
<p>Veröffentlichungsfähiger Einseiter</p> <ul style="list-style-type: none"> * Kurzfassung WAS FÜR WEN WIE * Liste Projektergebnisse - also diese Liste, ggf. kompromiert * mit Angabe Open Source Lizenz/Webadresse * wo finden Dritte die Projektergebnisse (inkl. Dokumentation Anwender_innen bzw. Entwickler_innen) * mögliche Weiterentwicklungen/ weitere Einsatz-/ Nutzungsmöglichkeiten 	<p>CC-BY-3.0 AT</p>	<p>https://netidee.at/openbiolink</p>
<p>Dokumentation Externkommunikation zur Erreichung Sichtbarkeit /Nachhaltigkeit (separates Dokument oder als Teil des Endberichtes)</p> <ul style="list-style-type: none"> * Welche Maßnahmen wurden in welchem Umfang gesetzt * Jeweils Bewertung Aufwand / Nutzen * Lessons Learned / Empfehlungen für andere Projekte 	<p>CC-BY-3.0 AT</p>	<p>https://netidee.at/openbiolink</p>
<p>Web Graph Creation Module 2.0</p> <p>Dieses Software-Modul greift auf verteilte, im Web zugängliche biomedizinische Ressourcen zu und aggregiert die gefundenen Daten. Im Zuge des webidee Projektes wird die Funktionalität des Software-Moduls erweitert, um noch reichhaltigere Daten gewinnen zu können.</p>	<p>MIT</p>	<p>https://github.com/OpenBioLink/OpenBioLink</p>

<p>OpenBioLink 2021 Benchmark Erweitertes und verbessertes Benchmark-Dataset, das mit dem im Projekt geschaffenen Web Graph Creation Module 2.0 automatisch generiert wird.</p>	<p>CC0 (d.h. besonders unrestrictive Lizenz) für selbst-generierte Daten; Aggregierte externe Daten haben diverse eigene Open Source Lizenzen</p>	<p>https://github.com/OpenBioLink/OpenBioLink</p>
<p>OpenBioLink Explorer Software, welche es erlaubt, über den Webbrowser die durch AI generierte Vorhersage neuer Verbindungen im Wissensnetzwerk zu explorieren, zu analysieren, zu vergleichen, und zu erklären. Nutzbar sowohl für wenig IT-affine biomed. ForscherInnen, als auch für EntwicklerInnen zum Debugging.</p>	<p>MIT</p>	<p>https://github.com/OpenBioLink/Explorer https://github.com/OpenBioLink/SAFRAN</p>
<p>Durchführung und Ergebnisse der "OpenBioLink 2021 Challenge". Um den Impact der netidee geschaffenen Ressourcen kosteneffizient zu vervielfachen führen wir eine offene Community-Challenge mit dem OpenBioLink 2021 Benchmark durch, bei der das AI-Modell mit den besten Vorhersagen ein Preisgeld im Wert von 1000 Euro verliehen bekommt (vgl. Kaggle).</p>	<p>CC-BY-3.0 AT</p>	<p>https://openbiolink.github.io/</p>

6 Verwertung der Projektergebnisse in der Praxis

SAFRAN und LinkExplorer werden derzeit bereits von einigen akademischen Gruppen in aktuellen Forschungsprojekten verwendet. OpenBioLink und damit verbundene Ressourcen wurden von der Pharmafirma AstraZeneca in manchen Projekten verwendet.

7 Öffentlichkeitsarbeit/ Vernetzung

Der im Projekt weiterentwickelte Algorithmus wurde bei der AKBC 2021 Konferenz (<https://www.akbc.ws/2021/virtual-details/>) einem internationalen Publikum präsentiert. Die Entwicklungen wurden bei unseren Kontakten in der Pharma-Firma AstraZeneca beworben, die an OpenBioLink Interesse gezeigt und uns für einen Vortrag eingeladen hat. Die OpenBioLink

Challenge wurde in Mailinglists und per Direktkontakt an relevante Forschungsgruppen beworben.

Projektergebnisse wurden auf Github publiziert. Social media (Twitter) wurde zur breiten Externkommunikation verwendet

Die Kombination aus Github-Repos, Preprint-Publikationen auf arXiv und Promotion auf Twitter ist in unserer Ansicht die effektivste Strategie zur Externkommunikation.

8 Eigene Projektwebsite

OpenBioLink (dataset) Github: <https://github.com/OpenBioLink/OpenBioLink>

OpenBioLink 2021 Challenge website: <https://openbiolink.github.io/>

Link Explorer (User interface + framework) Github: <https://github.com/OpenBioLink/Explorer>

SAFRAN (Link prediction algorithm that works together with Link Explorer) Github: <https://github.com/OpenBioLink/SAFRAN>

9 Geplante Aktivitäten nach netidee-Projektende

Die Software-Pakete sind dank der Entwicklung im Rahmen von netidee als Versionen 1.0 feature-complete und einsatzbereit. Die Projektergebnisse werden von uns nach Abschluss des Projektes als state-of-the-art Lösungen zur Web-basierten vorhersage in großen Wissensgraphen bereitgestellt und beworben werden. Die Software wird weiterhin frei als Open Source Produkt mit permissiven Lizenzen verfügbar gemacht werden.

Wir werden die LinkExplorer Suite in verschiedenen Anwendungsfällen (z.B. Arzneimittelentwicklung) zusammen mit externen Gruppen anwenden und evaluieren.

10 Anregungen für Weiterentwicklungen durch Dritte

OpenBioLink und die dafür entwickelten Evaluierungswerkzeuge können als Basis für die Entwicklung von biomedizinischen Link Prediction Algorithmen verwendet werden.

SAFRAN und LinkExplorer sind generelle Werkzeuge für web-basierte Link Prediction in großen Wissensnetzwerken und können daher für eine breite Palette an Anwendungsfällen verwendet werden. Wir laden externe Gruppen ein, diese Tools für ihre spezifischen Anwendungsfälle zu verwenden und, wenn notwendig, zu adaptieren.