Biomedical Relation Extraction Using Transfer Learning Methods

Kumulative Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der Mathematisch-Naturwissenschaftlichen Fakultät der

RHEINISCHEN FRIEDRICH-WILHELMS-UNIVERSITÄT

BONN

vorgelegt von

SUMIT MADAN

Bonn, März 2025

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

In Kooperation mit
Fraunhofer Institut für Algorithmen und Wissenschaftliches
Rechnen SCAI

Gutachter/Betreuer: Prof. Dr. rer. nat. Martin Hofmann-Apitius
Gutachter: Prof. Dr. rer. nat. Reinhard Klein

Tag der Promotion: 22. Januar 2026
Erscheinungsjahr: 2026

Abstract

The increasing volume of data in the biomedical field presents significant challenges related to information extraction and knowledge discovery. However, this large volume of data also offers substantial opportunities to enhance our understanding of disease mechanisms, identifying therapeutic targets, and advance precision medicine. To fully leverage these opportunities, advanced computational methodologies from the machine learning field are indispensable, allowing researchers to uncover valuable insights that would otherwise remain hidden.

This thesis explores the development and application of transfer learning methods — especially transformer-based models — for identifying and extracting relations from biomedical datasets. We contribute by providing a comprehensive review of the applications of transformer models across various biomedical subfields. Furthermore, we develop transformer-based methodologies in three experimental studies. Firstly, we implement a text mining workflow for extracting psychiatric attributes and psychopathological symptoms from German psychiatric reports, enabling secondary use of patient data in research. Secondly, we propose a Siamese architecture to predict virus-host protein-protein interactions using deep protein sequence embeddings, facilitating the prioritization of these interactions for drug discovery. Finally, we present an end-to-end text mining workflow designed to identify miRNA-disease associations from recent scientific literature, allowing to investigate the roles of miRNA in disease mechanisms.

In conclusion, the scientific advancements presented in this work demonstrate the potential of transformer-based methodologies for the data-driven extraction of valuable biological and medical relations, contributing to the advancement of knowledge in biomedicine.

Acknowledgments

I want to express my sincere thanks to my PhD supervisor and mentor, Prof. Dr. Martin Hofmann-Apitius, for believing in my abilities and giving me the opportunity and autonomy to pursue research in your team. Even though I tested your patience at times, your unwavering support and understanding never faltered. I'm truly grateful for the responsibilities you've entrusted me with and the doors you've opened for me.

I thank Prof. Dr. Reinhard Klein for accepting the role of my second reviewer and for being a member of my defense committee.

My sincere gratitude to Prof. Dr. Holger Fröhlich for your invaluable guidance and support during the pursuit of my scientific goals. I am truly grateful for your mentorship.

I am profoundly thankful to Prof. Dr. Juliane Fluck, whose inspiration ignited my passion for research and led me to pursue my PhD.

I would like to thank all my Fraunhofer colleagues and external collaborators with whom I've had the privilege to work with over the years. A special thanks to all the students I've had the privilege to supervise and learn together with you.

I would like to acknowledge my cousin, Shriya Bhatija, for giving me valuable feedback on my thesis. Your suggestions have helped to enhance the quality of my work.

Lastly, I am greatly indebted to my family. Mom and Dad, your endless love has been my greatest motivation. I am truly grateful for everything you have done for me. My sincere thanks to my sister, Himanshi Braun, for always standing by my side. Sharing life updates with you is just pure joy. I want to thank my wife, Annekathrin, as without your unconditional support this project of my life would have not been possible. A special shout-out to my kids, Milan and Rhio, as you provide me a wonderful escape by having playful and funny moments almost everyday.

Declaration

I hereby certify that this material is my own work, that I used only those sources and resources referred to in the thesis, and that I have identified citations as such.

Sumit Madan

Contents

List of Publications

Thesis Publications

1. **S. Madan**, M. Lentzen, J. Brandt, D. Rueckert, M. Hofmann-Apitius, and H. Fröhlich, "Transformer models in biomedicine," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 214, 2024. DOI: 10.1186/s12911-024-02600-5
2. **S. Madan**, F. Julius Zimmer, H. Balabin, S. Schaaf, H. Fröhlich, J. Fluck, I. Neuner, K. Mathiak, M. Hofmann-Apitius, and P. Sarkheil, "Deep Learning-based Detection of Psychiatric Attributes from German Mental Health Records," *International Journal of Medical Informatics*, vol. 161, p. 104 724, 2022. DOI: 10.1016/j.ijmedinf.2022.104724
3. **S. Madan**, V. Demina, M. Stapf, O. Ernst, and H. Fröhlich, "Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings," *Patterns*, vol. 3, no. 9, p. 100 551, 2022. DOI: 10.1016/j.patter.2022.100551
4. **S. Madan**, L. Kühnel, H. Fröhlich, M. Hofmann-Apitius, and J. Fluck, "Dataset of miRNA–disease relations extracted from textual data using transformer-based neural networks," *Database*, vol. 2024, baae066, 2024. DOI: 10.1093/database/baae066

Other Journal Publications

Graph Machine Learning

5. S. Krix, L. N. DeLong, **S. Madan**, D. Domingo-Fernández, A. Ahmad, S. Gul, A. Zaliani, and H. Fröhlich, "MultiGML: Multimodal graph machine learning for prediction of adverse drug events," *Heliyon*, vol. 9, no. 9, e19441, 2023. DOI: 10.1016/j.heliyon.2023.e19441

Structured EHR Data Analysis with Transformers

6. M. Lentzen, T. Linden, S. Veeranki, **S. Madan**, D. Kramer, W. Leodolter, and H. Fröhlich, "A Transformer-Based Model Trained on Large Scale Claims Data for Prediction of Severe COVID-19 Disease Progression," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 9, pp. 4548–4558, 2023. DOI: 10.1109/JBHI.2023.3288768

Semantic Search Engine for small RNA Expression

7. R.-U. Rahman, A.-M. Liebhoff, V. Bansal, M. Fiosins, A. Rajput, A. Sattar, D. S. Magruder, **S. Madan**, T. Sun, A. Gautam, S. Heins, T. Liwinski, J. Bethune, C. Trenkwalder, J. Fluck, B. Mollenhauer, and S. Bonn, "SEA-web: The small RNA Expression Atlas web application," *Nucleic Acids Research*, p. 16, 2019. DOI: 10.1093/nar/gkz869

Biomedical Ontologies

8. A. Sargsyan, P. Wegner, S. Gebel, A. Kaladharan, P. Sethumadhavan, V. Lage-Rupprecht, J. Darms, B. Schultz, J. Klein, M. Jacobs, **S. Madan**, M. Hofmann-Apitius, and A. T. Kodamullil, "The Epilepsy Ontology: A community-based ontology tailored for semantic interoperability and text mining," *Bioinformatics Advances*, vol. 3, no. 1, vbad033, 2023. DOI: 10.1093/bioadv/vbad033
9. A. Sargsyan, S. Baksi, J. Darms, **S. Madan**, S. Gebel, O. Keminer, G. M. Jose, H. Balabin, L. N. DeLong, and M. Kohler, "The COVID-19 Ontology," *Bioinformatics*, vol. 36, no. 24, pp. 5703–5705, 2020. DOI: 10.1093/bioinformatics/btaa1057

Survey

10. J. Botz, D. Wang, N. Lambert, N. Wagner, M. Génin, E. Thommes, **S. Madan**, L. Coudeville, and H. Fröhlich, "Modeling approaches for early warning and monitoring of pandemic situations as well as decision support," *Frontiers in Public Health*, vol. 10, 2022. DOI: 10.3389/fpubh.2022.994949

Biomedical Text Mining

11. P. Wegner, H. Fröhlich, and **S. Madan**, "Evaluating Knowledge Fusion Models on Detecting Adverse Drug Events in Text," *PLOS Digital Health*, 2025. DOI: 10.1371/journal.pdig.0000468
12. N. S. Babaiha, H. Elsayed, B. Zhang, A. Kaladharan, P. Sethumadhavan, B. Schultz, J. Klein, B. Freudensprung, V. Lage-Rupprecht, A. T. Kodamullil, M. Jacobs, S. Geissler, **S. Madan**, and M. Hofmann-Apitius, "A natural language processing system for the efficient updating of highly curated pathophysiology mechanism knowledge graphs," *Artificial Intelligence in the Life Sciences*, vol. 4, p. 100 078, 2023. DOI: 10.1016/j.ailsci.2023.100078
13. M. Lentzen, **S. Madan**, V. Lage-Rupprecht, L. Kühnel, J. Fluck, M. Jacobs, M. Mittermaier, M. Witzenrath, P. Brunecker, M. Hofmann-Apitius, J. Weber, and H. Fröhlich, "Critical assessment of transformer-based AI models for German clinical notes," *JAMIA Open*, vol. 5, no. 4, ooac087, 2022. DOI: 10.1093/jamiaopen/ooac087
14. L. Langnickel, K. Krockauer, M. Uebachs, S. Schaaf, **S. Madan**, T. Klockgether, and J. Fluck, "Information extraction from german clinical care documents in context of alzheimer's disease," *Applied Sciences*, vol. 11, no. 22, 2021. DOI: 10.3390/app112210717
15. R. Karki, **S. Madan**, Y. Gadiya, D. Domingo-Fernández, A. T. Kodamullil, and M. Hofmann-Apitius, "Data-Driven Modeling of Knowledge Assemblies in Understanding Comorbidity Between Type 2 Diabetes Mellitus and Alzheimer's Disease," *Journal of Alzheimer's Disease*, vol. Preprint, pp. 1–9, Preprint 2020. DOI: 10.3233/JAD-200752
16. R. Li, A. Zupanic, M. Talikka, V. Belcastro, **S. Madan**, J. Dörpinghaus, C. vom Berg, J. Szostak, F. Martin, M. C. Peitsch, and J. Hoeng, "Systems Toxicology Approach for Testing Chemical Cardiotoxicity in Larval Zebrafish," *Chemical Research in Toxicology*, 2020. DOI: 10.1021/acs.chemrestox.0c00095
17. **S. Madan**, J. Szostak, R. Komandur Elayavilli, R. T.-H. Tsai, M. Ali, L. Qian, M. Rastegar-Mojarad, J. Hoeng, and J. Fluck, "The extraction of complex relationships and their conversion to biological expression language (BEL) overview of the BioCreative VI (2017) BEL track," *Database*, vol. 2019, 2019. DOI: 10.1093/database/baz084
18. R. Li, C. vom Berg, M. Talikka, **S. Madan**, J. Dörpinghaus, J. Fluck, J. Szostak, F. Martin, M. C. Peitsch, J. Hoeng, and A. Zupanic, "Systems toxicology approach for the assessment of zebrafish cardiotoxicity," *Toxicology Letters*, vol. 295, S102, 2018. DOI: 10.1016/j.toxlet.2018.

06.611

19. A. T. Kodamullil, A. Iyappan, R. Karki, **S. Madan**, E. Younesi, and M. Hofmann-Apitius, "Of Mice and Men: Comparative Analysis of Neuro-Inflammatory Mechanisms in Human and Mouse Using Cause-and-Effect Models," *Journal of Alzheimer's Disease*, vol. 59, no. 3, pp. 1045–1055, 2017. DOI: 10.3233/JAD-170255. pmid: 28731442
20. J. Fluck, **S. Madan**, S. Ansari, A. T. Kodamullil, R. Karki, M. Rastegar-Mojarad, N. L. Catlett, W. Hayes, J. Szostak, J. Hoeng, and M. Peitsch, "Training and evaluation corpora for the extraction of causal relationships encoded in biological expression language (BEL)," *Database : the journal of biological databases and curation*, vol. 2016, 2016. DOI: 10.1093/database/baw113. pmid: 27554092
21. F. Rinaldi, T. R. Ellendorff, **S. Madan**, S. Clematide, A. van der Lek, T. Mevissen, and J. Fluck, "BioCreative V track 4: A shared task for the extraction of causal network information using the Biological Expression Language," *Database : the journal of biological databases and curation*, vol. 2016, 2016. DOI: 10.1093/database/baw067. pmid: 27402677
22. Q. Wang, S. S. Abdul, L. Almeida, S. Ananiadou, Y. I. Balderas-Martínez, R. Batista-Navarro, D. Campos, L. Chilton, H.-J. Chou, G. Contreras, L. Cooper, H.-J. Dai, B. Ferrell, J. Fluck, S. Gama-Castro, N. George, G. Gkoutos, A. K. Irin, L. J. Jensen, S. Jimenez, T. R. Jue, I. Keseler, **S. Madan**, S. Matos, P. McQuilton, M. Milacic, M. Mort, J. Natarajan, E. Pafilis, E. Pereira, S. Rao, F. Rinaldi, K. Rothfels, D. Salgado, R. M. Silva, O. Singh, R. Stefancsik, C.-H. Su, S. Subramani, H. D. Tadepally, L. Tsaprouni, N. Vasilevsky, X. Wang, A. Chatr-Aryamontri, S. J. F. Laulederkind, S. Matis-Mitchell, J. McEntyre, S. Orchard, S. Pundir, R. Rodriguez-Esteban, K. Van Auken, Z. Lu, M. Schaeffer, C. H. Wu, L. Hirschman, and C. N. Arighi, "Overview of the interactive task in BioCreative V," *Database*, vol. 2016, baw119, 2016. DOI: 10.1093/database/baw119
23. **S. Madan**, S. Hodapp, P. Senger, S. Ansari, J. Szostak, J. Hoeng, M. Peitsch, and J. Fluck, "The BEL information extraction workflow (BELIEF): Evaluation in the BioCreative V BEL and IAT track," *Database*, vol. 2016, baw136, 2016. DOI: 10.1093/database/baw136. pmid: 27694210
24. J. Szostak, S. Ansari, **S. Madan**, J. Fluck, M. Talikka, A. Iskandar, H. De Leon, M. Hofmann-Apitius, M. C. Peitsch, and J. Hoeng, "Construction of biological networks from unstructured information based on a semi-automated curation workflow," *Database*, vol. 2015, 2015. DOI: 10.1093/database/bav057. pmid: 26200752

Other Conference Publications

Semantic Search Engine

25. L. Langnickel, R. Baum, J. Darms, **S. Madan**, and J. Fluck, "COVID-19 preVIEW: Semantic Search to Explore COVID-19 Research Preprints," *Studies in Health Technology and Informatics*, vol. 281, pp. 78–82, 2021. DOI: 10.3233/SHTI210124. pmid: 34042709
26. J. Dörpinghaus, J. Klein, J. Darms, **S. Madan**, and M. Jacobs, "SCAIView – A Semantic Search Engine for Biomedical Research Utilizing a Microservice Architecture," in *Proceedings of the Posters and Demos Track of the 14th International Conference on Semantic Systems - SEMANTiCS2018*, 2018

Semantic Architecture / Ontologies

27. A. Y. Lin, S. Gebel, Q. L. Li, **S. Madan**, J. Darms, E. Bolton, B. Smithe, and M. Hofmann-Apitius, "CTO: A Community-Based Clinical Trial Ontology and its Applications in PubChemRDF and SCAIView," in *Proceedings of the 11th International Conference on Biomedical Ontologies (ICBO)*, 2020
28. S. Mora, **S. Madan**, S. Gebel, and e. al, "Proposal of an Architecture for Terminology Management in a Research Project," in *Digital Personalized Health and Medicine*, 2020, pp. 1371–1372
29. L. Langnickel, R. Baum, G. Wollnik-Korn, B. Fischer-Wagener, **S. Madan**, and J. Fluck, "The future of German MeSH: A new semi-automatic translation process and new services for search and annotation," presented at the GMDS Conference 2020, 2020
30. **S. Madan**, M. Fiosins, S. Bonn, and J. Fluck, "A semantic data integration methodology for translational neurodegenerative disease research," in *Proceedings of the 11th International Conference Semantic Web Applications and Tools for Life Sciences*, ser. CEUR Workshop Proceedings, vol. 2275, Antwerp, Belgium: CEUR, 2018

Biomedical Text Mining

31. M. Ali, **S. Madan**, A. Fischer, H. Petzka, and J. Fluck, "Automatic Extraction of BEL-Statements based on Neural Networks Automatic Extraction of BEL-Statements based on Neural Networks," in *Proceedings*

of BioCreative VI Challenge and Workshop, 2017, pp. 2013–2015

32. **S. Madan**, J. Szostak, J. Dörpinghaus, J. Hoeng, and J. Fluck, "Overview of BEL Track: Extraction of Complex Relationships and their Conversion to BEL," in *Proceedings of BioCreative VI Challenge and Workshop*, 2017
33. J. Szostak, **S. Madan**, W. Hayes, J. Doerpinghaus, J. Fluck, M. Talikka, M. C. Peitsch, and J. Hoeng, "Recent improvements of the BEL Information Extraction workFlow (BELIEF) for biomedical text mining and curation," in *10th International Biocuration Conference 2017*, vol. 6, 10th International Biocuration Conference 2017, 2017. DOI: 10.7490/F1000RESEARCH.1113812.1
34. J. Szostak, S. Ansari, M. Talikka, J. Fluck, **S. Madan**, F. Martin, M. C. Peitsch, and J. Hoeng, "BELIEF: A semi-automated curation tool to build mechanistic causal biological knowledgebase from unstructured scientific information," in *Systems Toxicology 2016 Conference - Real World Applications and Opportunities*, Les Diablerets, Switzerland, 2016
35. J. Szostak, O. Iro, L. S. Giuseppe, M. Talikka, S. Ansari, J. Fluck, **S. Madan**, F. Martin, M. C. Peitsch, and J. Hoeng, "A computational network model describing xenobiotic metabolism response in the liver built using the semi-automated curation workflow BELIEF.," 2016
36. **S. Madan**, S. Hodapp, and J. Fluck, "BELIEF Dashboard – a Web-based Curation Interface to Support Generation of BEL Networks," in *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, Sevilla, Spain, 2015, pp. 409–417
37. J. Fluck, **S. Madan**, T. Effendorf, H.-T. Mevissen, S. Clematide, A. van der Lek, and F. Rinaldi, "Track 4 Overview: Extraction of Causal Network Information in Biological Expression Language (BEL)," in *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, vol. 1, Sevilla, Spain, 2015, pp. 333–346
38. J. Fluck, **S. Madan**, S. Ansari, J. Szostak, J. Hoeng, M. Zimmermann, M. Hofmann-Apitius, and M. C. Peitsch, "BELIEF - A semiautomatic workflow for BEL network creation," in *Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM)*, 2014, pp. 109–113. DOI: 10.5167/uzh-98982
39. J. Fluck, A. Klenner, **S. Madan**, S. Ansari, T. Bobic, J. Hoeng, M. Hofmann-Apitius, and M. C. Peitsch, "BEL networks derived from qualitative translations of BioNLP Shared Task annotations," in *Workshop on Biomedical Natural Language Processing, BioNLP 2013*, Sofia, Bulgaria: Association for Computational Linguistics (ACL), 2013, pp. 80–88
40. R. Klinger, P. Senger, **S. Madan**, and M. Jacovi, "Online Communities

Support Policy-Making: The Need for Data Analysis," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7444 LNCS, 2012, pp. 132–143. DOI: 10.1007/978-3-642-33250-0_12

List of Acronyms

AI	artificial intelligence	3
AMDP	Association for Methodology and Documentation in Psychiatry	46
ALBERT	A Lite BERT	32
AUC	area under receiver operator characteristic curve	50
AUPR	area under precision-recall curve	50
BELIEF	Biological Expression Language Information Extraction Workflow	14
BERT	bidirectional encoder representations from transformers	26
bioNLP	biomedical natural language processing	3
ChEBI	Chemical Entities of Biological Interest	10
CNN	convolutional neural network	42
CRediT	Contributor Roles Taxonomy	43
COVID-19	Coronavirus disease 2019	11
DL	deep learning	3
DNA	deoxyribonucleic acid	21
EBI	European Bioinformatics Institute	9
EHR	electronic health record	3
ELECTRA	Efficiently Learning an Encoder that Classifies Token Replacements Accurately	32
ESM	Evolutionary Scale Modeling	33
FAIR	findability, accessibility, interoperability, and reusability	9
GAI	generative artificial intelligence	36
GML	graph machine learning	11
GO	Gene Ontology	10
GPT	generative pre-trained transformer	26

ICD	International Classification of Diseases	10
IMEx	International Molecular Exchange Consortium	14
INDRA	Integrated Network and Dynamical Reasoning Assembler	14
JCV	John Cunningham polyomavirus	6
Llama	Large Language Model Meta AI	36
LLM	large language model	6
LBD	literature-based discovery	19
ML	machine learning	3
MLM	masked language modeling	31
MeSH	Medical Subject Headings	17
miRNA	micro ribonucleic acid	6
mRNA	messenger ribonucleic acid	12
MSE	mental state examination	5
NCBI	National Center for Biotechnology Information	9
NER	named entity recognition	18
NEL	named entity linking	18
NLM	National Library of Medicine	17
NLP	natural language processing	17
NSP	next sentence prediction	31
PPI	protein-protein interaction	6
PMC	PubMed Central	54
RoBERTa	Robustly Optimized BERT Approach	32
RL	reinforcement learning	36
RE	relation extraction	18
RNA	ribonucleic acid	21
RO	Relation Ontology	18
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2	6
STEP	Siamese Tailored deep sequence Embedding of Proteins	49
TL	transfer learning	3
UniProtKB	UniProt Knowledgebase	22
XAI	explainable artificial intelligence	4

List of Figures

List of Tables

Part I.

Fundamentals

1. Overview of the Thesis

The fascinating discipline of biomedicine is dedicated to the advancement of human medicine by studying biological and chemical processes and has experienced a substantial surge in the volume of data, often referred to as big data. This growth can be attributed to various factors. These include advancements in information technologies for data collection, storage, and analysis, high-throughput sequencing and imaging technologies, and widespread adoption of electronic health records (EHRs) [41–43].

The big data phenomenon in biomedicine poses both challenges and opportunities. On the one hand, the sheer volume of data can be overwhelming to manage, store, and analyze. The complexity of integrating and harmonizing data from different sources presents challenges in data quality, data curation, data interoperability, and privacy protection [44, 45]. Additionally, the need for sophisticated computational infrastructure and expertise in data analytics adds to the challenge [45]. On the other hand, large-scale and diverse biomedical data inherently presents more opportunities. It enables researchers to gain deeper insights into disease pathophysiology by examining biological processes across biological scales, from the molecular to organism level. Furthermore, it facilitates the discovery of biomarkers, the development of predictive models, and the advancement of the precision medicine paradigm [45, 46].

The fusion of big data alongside the strides in artificial intelligence (AI) — especially machine learning (ML) and deep learning (DL) — has immensely impacted the field of biomedicine [47]. Recently, transfer learning (TL) has emerged as a noteworthy paradigm in biomedicine, allowing to leverage large-scale general datasets to build pre-trained models that can then be adapted to various downstream tasks with scarce datasets [48]. This approach has been particularly beneficial in biomedical natural language processing (bioNLP), computer vision, EHR data analysis, and protein sequence analysis. For instance, such models are capable of accurately modeling disease trajectories for patients in a precision medicine context [6, 46], predicting protein structures from protein sequences [49, 50], and assessing medical images for various conditions [51, 52].

Leveraging AI and ML techniques to analyze large datasets enables the

discovery of hidden patterns and transforms them into actionable knowledge. To effectively harness such derived knowledge from the vast amount of fragmented biomedical data, it is necessary to encode it into a reusable and computable form. One common way is to transform and represent the knowledge as triples, each consisting of a subject, a predicate, and an object. These triples can capture complex and multimodal biomedical relations and entities, such as drug-disease relationships, protein-protein interactions, drug side effects, or disease phenotypes [53]. Structuring biomedical knowledge as interconnected relations facilitates data integration, enables semantic search capabilities, supports reasoning over complex data, and facilitates the development of AI models through domain-specific knowledge fusion.

This thesis aims to investigate transformer-based methods for identifying and extracting information from the data necessary to build biomedical relations. The transformer-based methods represent a class of deep learning models that utilize the transformer architecture introduced by Vaswani *et al.* [54] and leverage transfer learning. Specifically, we will develop methodologies for extracting and structuring valuable insights from complex medical and biological data. These methodologies will be thoroughly evaluated to ensure their robustness and effectiveness for biomedical applications. With the application of these methodologies, we will generate new biomedical relations, which will contribute to the advancement of knowledge in the biomedical field. Additionally, we will explore potential improvements and refinements based on empirical findings from our experiments.

1.1. Objectives and Contributions

In this section, we present the key objectives that contribute to achieving the overall aim of the thesis. Additionally, Figure 1 visually illustrates the contributions of this work by depicting the relevant biomedical data modalities. Some of the modalities will be used to develop transformer-based methodologies, leading to four main contributions.

Objective 1 Generate an overview of the applications and challenges of transformer-based models in the field of biomedicine.

Contribution 1 In Chapter 6, we provide a comprehensive overview on how transformer-based models are being applied in the biomedical field. We describe that transformers have been proposed for the analysis of biomedical text, structured EHR data, biomedical graphs, biomedical imaging, and biological sequences. Additionally, the study briefly introduces explainable artificial intelligence (XAI) approaches suitable to offer explanations for predictions

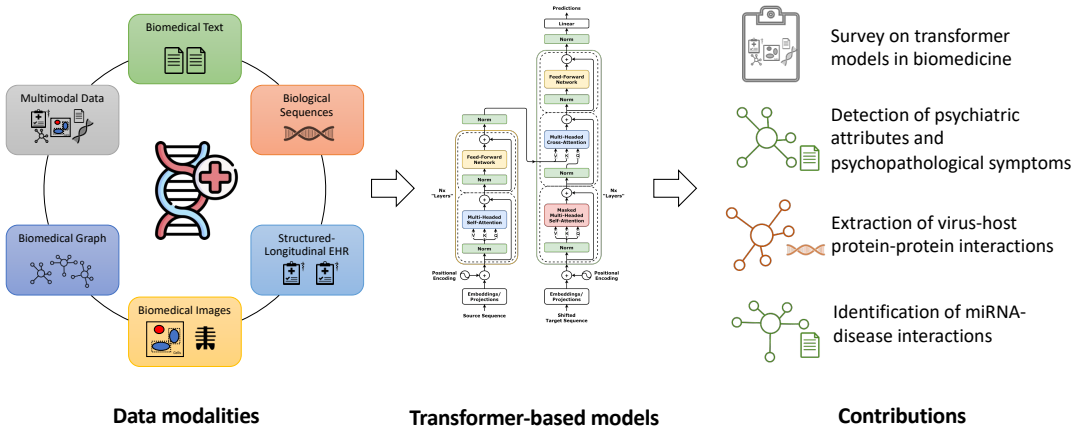


Figure 1.: Key contributions of this work: extraction of biomedical relational information using transformer-based models on different biomedical data modalities.

Source: Original transformer architecture image by Daniel Voigt Godoy (`https://github.com/dvgodoy/dl-visuals`) is licensed under CC BY 4.0. Image uses icons made by Freepik from www.flaticon.com.

made by these models. Finally, we discuss the strengths, challenges, and future directions for improving transformer-based studies in the biomedical field.

Objective 2 Explore the potential of German clinical psychiatric notes to predict psychiatric attributes and psychopathological symptoms using a transformer-based language model.

Contribution 2 Medical data residing in hospital silos contain information that has an value for research. A large portion of this data exists in the form of unstructured textual notes. Sophisticated and evaluated text mining methods are required to structure the relevant information from these notes. In the field of psychiatry, the mining of mental state examinations (MSEs) present an unique opportunity to gain large-scale insights on the psychiatric attributes and psychopathological symptoms of patients. To the best of our knowledge, no analysis has yet been conducted on German MSEs. Our study, summarized in Chapter 7, contributes to this research gap by creating a new labeled dataset of MSE reports. We utilize this dataset to create and evaluate a text mining workflow based on a German transformer-based language model to structure psychiatric traits and determine their pathological associations that can be classified as symptoms. Moreover, we employ the model to predict psychopathological symptoms from unlabeled MSEs and report the most frequently occurring symptoms. This work represents the initial step toward structuring routinely-collected psychiatric textual notes, enabling its use for secondary research analysis.

Objective 3 Investigate the efficacy of deep protein sequence embeddings to predict virus-host protein-protein interactions (PPIs) in the context of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and John Cunningham polyomavirus (JCV).

Contribution 3 The knowledge on virus-host PPIs is essential for understanding the mechanisms of viral infection and host's response, as well as developing new therapeutics. As experimental methods are costly and laborious, computational methods have become a popular alternative to predict and rank PPIs. The transformer-based models, pre-trained on billions of protein sequences [55, 56], offer a way to apply transfer learning for the detection of virus-host PPIs. In Chapter 8, we present a study that investigates transformer-based Siamese models to accurately identify relevant PPIs based solely on the protein sequence representations. The best-performing model was then employed to identify interactions between human receptors and selected proteins of two specific viruses, namely SARS-CoV-2 and JCV. With this work, we underscore the potential of deep sequence embedding techniques by reaching state-of-the-art performances on various PPI datasets. The improvement of computational PPI detection approaches will facilitate the prioritization of the PPIs for further investigation in this area.

Objective 4 Evaluate and utilize biomedical large language models (LLMs) to develop a text mining workflow for inferring associations between micro ribonucleic acids (miRNAs) and diseases from current scientific literature.

Contribution 4 MiRNAs are short sequences of nucleotides that play an important role in cellular processes and have been associated with a range of human diseases [57]. Current biomedical databases on miRNA-disease associations are often outdated and struggle to keep pace with rapidly emerging information from new scientific research. In Chapter 9, we aim to retrieve miRNA-disease associations from scientific texts by training and evaluating new transformer-based models and embedding them into an end-to-end text mining workflow. To train the models, we created a new dataset with miRNA-disease associations by utilizing the distant supervision technique. Subsequently, we applied the workflow to identify new associations from PubMed abstracts published from 2020 to 2023. Finally, we compared the predicted associations related to three diseases (Alzheimer's disease, Parkinson's disease, and epilepsy) with those in an existing biomedical database. This workflow facilitates timely updates to biomedical databases that will help researchers to advance the understanding of miRNA roles in further downstream analysis.

1.2. Study Design

The experimental studies of each contribution (excluding literature review study) adhere to a common workflow that is depicted in Figure 2.

Firstly, during the data collection and preprocessing processes, we either collect or create new labeled datasets from various sources. When no labeled data is available, we curate manually to assign labels to instances. In some cases, we enhance or filter the datasets by using secondary information (such as ontologies) from external databases. The datasets are then splitted into training, validation, and test sets. We further set aside an inferencing dataset without labels to assess the applicability of the models.

Secondly, we fine-tune one or more transformer-based models on the training data and validate them using the validation datasets. We utilize Bayesian optimization to fine-tune the hyperparameters (such as learning rate, batch size, and number of layers) of transformer-based models, as these parameters influence their training performance. Finding appropriate values of hyperparameters is crucial to learn effectively from the training data and address potential training issues (such as overfitting).

After the training and optimization processes, in the third step, we select the best model for a final evaluation. We use the held-out, independent test sets to measure the performance of models and assess their generalization ability. Depending on the availability of appropriate external test sets, we also utilize them to further ensure the reliability and robustness of the models.

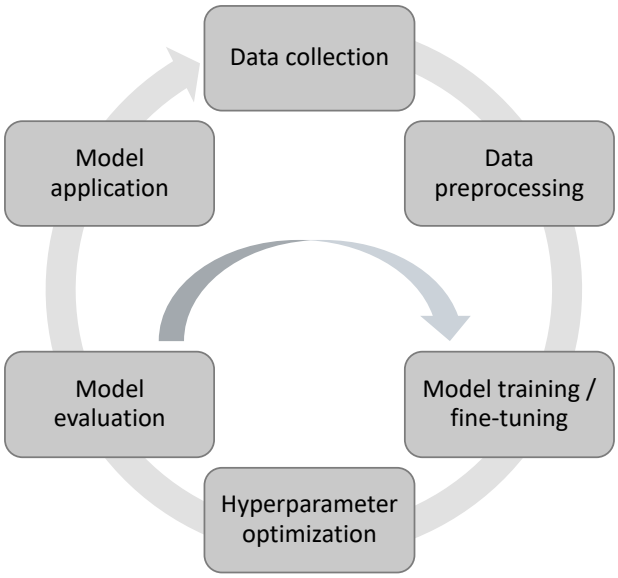


Figure 2.: Common machine learning workflow implemented by our experimental studies.

Finally, if inferencing datasets are available, we apply the best-performing model to infer new biomedical relations from these datasets.

1.3. Structure of the Thesis

This thesis is structured into three main parts. With the current Part 1, titled "Fundamentals", we lay the groundwork for the thesis by providing essential background information. Continuing with the Chapter 2, we describe what biomedical relations are, the types of relations obtained from various data modalities, and how these are represented. Thereafter, we define two specific fields relevant to our experimental studies: biomedical text mining and protein sequence analysis, along with their main challenges in Chapter 3 and Chapter 4, respectively. Subsequently, in Chapter 5, we offer a concise overview of transfer learning as well as provide an introduction to the transformer architecture including associated pre-training and fine-tuning strategy. We further elaborate on the biomedical specific pre-trained transformer models, culminating in a discussion of the prospects and limitations of recent LLMs in the biomedical field.

In Part 2, titled "Main Contributions", we summarize the four research studies, encompassing Chapter 6, Chapter 7, Chapter 8, and Chapter 9. In each of these chapters, we highlight the contributions made to the specific fields and briefly elaborate on the relevant findings of the studies.

In Part 3, titled "Recapitulation", we provide a conclusion in Chapter 10 that summarizes the key findings. In Chapter 11, we broadly consider potential directions for further exploration of open challenges.

2. Biomedical Relations

The term biomedical relation refers to the connections, links, associations, or interactions between different biomedical entities or concepts, such as genes, proteins, drugs, diseases, biological functions, and phenotypic traits. Such biomedical relations can take various forms, which include physical interactions between molecules, regulatory interactions between genes, functional relationships between proteins, and therapeutic interactions between drugs and targets. Relations are observed at multiple levels (see Table 1), ranging from basic events that describe biological roles associated with an object to more specific interactions where both the subject and object are associated with a specific (possibly causal) link. Generally, all interactions are influenced by surrounding conditions or temporal constraints. Finally, interactions may also be accompanied by additional metadata, such as supporting evidence, their provenance, a confidence score indicating the reliability of the interaction, or its detection methodology.

Moreover, understanding these biomedical relations is essential in the context of big data and AI applications in biomedicine. The ability to systematically represent and analyze these relations enhances our capacity to extract meaningful insights from vast datasets, ultimately driving advancements in biomedicine.

2.1. Representation of Biomedical Relations

The discipline of knowledge representation revolves around developing strategies for encoding information — such as relations — to make it usable and interpretable by computer systems [58, 59]. In the scientific community, numerous well-established biomedical databases serve as key resources for collecting and accessing relations. Some major database providers, such as the European Bioinformatics Institute (EBI) and the U.S. National Center for Biotechnology Information (NCBI), invest huge resources in expert curation and harmonization to ensure the quality of information provided. These databases facilitate the FAIRification process, which aims to enhance the findability, accessibility, interoperability, and reusability (FAIR) of information essential for knowledge sharing in scientific research. The Nucleic Acids

Table 1.: Overview on levels of biomedical interactions with increasing complexity.

Levels of interaction	Name	Examples (**subject**, **relationship**, object)
Level 1	Event	expression of gene A, phosphorylation of protein B, complex of protein C1 and C2
Level 2	Simple association	gene A is positively correlated with disease A, symptom B is associated with disease B1 and disease B2
Level 3	Specific (or causal) interaction	protein A catalyzes protein kinase activity, virus B is causally linked with disease B
Level 4	Context-specific relations	interaction A is found in tissue A, interaction B occurs in organ B during infection with virus B, interaction C occurred in patient C after discharge

Research database issue publishes yearly statistics on the number of new, updated, and discontinued databases, including those that offer information on relations [60, 61]. For instance, TarBase v9.0 [62], BioLiP 2.0 [63], and DGIdb 5.0 [64] provided updates on the experimentally-validated miRNA-gene, the ligand-protein, and the drug-gene interactions, respectively. One major challenge when accessing particular information is that it is scattered across several databases, each having their own unique scope. Meta databases aim to address this issue by aggregating data from a variety of sources. For instance, the aforementioned DGIdb database collects drug-gene interactions from databases such as ChEMBL [65], DrugBank [66], Drug Target Commons [67], and PharmGKB [68]. Finding a specific database can also be challenging, therefore, a curated catalog of biomedical databases is provided by Database Commons [69].

Aside from databases, there are other resources suitable for the representation of relational knowledge. Biomedical ontologies are not only limited to defining domain-specific concepts and their hierarchical relationships; they may encompass additional axioms that describe interrelations between these concepts [70]. The Chemical Entities of Biological Interest (ChEBI) [71] and the Gene Ontology (GO) [72, 73] are two such biomedical ontologies that include functional knowledge on chemical compounds and gene products, respectively. In addition, a widely recognized resource is the International

Classification of Diseases (ICD), which functions as a classification system and specifies health-related concepts and their relationships [74].

A multitude of relations can also be represented as knowledge graphs (also known as networks or pathways), where nodes represent biomedical concepts and edges represent the relationships between them [53]. These biomedical knowledge graphs can cover a variety of scopes and even include multimodal knowledge. For instance, PrimeKG [53] includes highly-curated knowledge designed for precision medicine analyses, BioKG [75] embeds knowledge from public databases and text-mined information from PubMed abstracts to enable data-driven biomedical research, while Coronavirus disease 2019 (COVID-19) Knowledge Graph integrates text-mined information and limits the scope to COVID-19 pathophysiology [76]. In addition, pathway databases such as Pathway Commons [77], Wikipathways [78], and Reactome [79] provide a structured framework to scope relational knowledge. Quality control of knowledge graphs, characterized in six dimensions including accuracy, completeness, and timeliness, is however a huge unsolved challenge [80].

Thus far, we have primarily addressed symbolic knowledge representation techniques. However, ML methodologies based on statistical foundations constitute another domain that is capable of representing data and relational knowledge [58, 81, 82]. Particularly DL — a subfield of representation learning that relies on neural networks — is highly effective at automatically learning patterns as well as abstracting features and identifying their relationships from large raw datasets [82]. Moreover, it is important to note that a trained ML model itself contains relational information, which helps it making predictions based on learned patterns. For instance, graph machine learning (GML) that utilizes graph neural networks and prior knowledge in established graphs can discover new links between concepts [83]. Furthermore, LLMs that represent another class of deep neural networks have also demonstrated to effectively predict relations from various data modalities [55, 84, 85] (see also Chapter 5 and Chapter 6).

Both symbolic and non-symbolic knowledge representation techniques are subject to certain strengths and weaknesses. Biological databases, ontologies, knowledge graphs, and pathways are human-friendly as they enable interpretability and facilitate logical reasoning. However, managing large knowledge bases and keeping them updated is challenging. Deep learning models, on the other hand, are often seen as black boxes with low explainability. Nevertheless, they are uniquely capable of learning complex patterns from data, handling large datasets, and dealing with noisy and unstructured data. Researchers have suggested hybrid approaches to overcome the challenges of both techniques. Symbolic representations can support deep learning models

and enhance their interpretability and explainability [86]. For such intents, the application of XAI methods has gained momentum in biomedicine to specifically address the transparency issues of AI. The XAI solutions are aimed at providing accurate and humanly-accessible explanations to the predictions made by AI systems [87].

2.2. Relations Across Biomedical Data Modalities

Extensive experimentation leads to the generation of a wide range of data modalities, encompassing diverse types of information across biological scales. Figure 3 visualizes common data modalities that are relevant to biomedicine. In the following, we briefly cover each modality and explore the relations researchers extracted from these.

Biological sequences comprise a type of modality, which originates from fields such as genomics, proteomics, and transcriptomics. They essentially provide unique cellular information from sequences of genes, proteins, RNA, and DNA for an organism [88]. Such data can be used to reveal insights on associations between miRNA and messenger ribonucleic acid (mRNA) for cancer [89], gather information on disease biomarkers [88], or determine causal molecular interactions [88]. The IntAct database, for instance, provides curated molecular interactions that were obtained from experimentally-derived interaction data reported in scientific literature [90]. In this work, we specifically focus on the protein sequence analysis, which is outlined in the Chapter 4.

Structured-longitudinal EHR represents the real-world patient data from clinical settings that can contain clinical variables such as patients' demographics, family histories, diagnoses, symptoms, and prescribed therapies and medications that are often embedded in a longitudinal way. Large EHR datasets, such as QResearch [91] and CRPD [92], have been utilized in recent years to perform predictive clinical modeling by using machine learning [93]. These models have become the basis for analyzing and detecting correlations between clinical features [42, 93]. Common disease-symptom relations, adverse drug events, and comorbidities have previously been studied [93]. Although large real-world EHR datasets are becoming more accessible for research, including their access in federated mode [94], computational analysis of these datasets remains challenging due to incompleteness, noise, bias, and longitudinal nature of the data. Furthermore, real-world EHR data frequently fails to offer a comprehensive patient profile, as it lacks narrative notes, imaging, and in-depth phenotyping data, often due to unavailability or privacy concerns. Therefore, such data is often inadequate for addressing

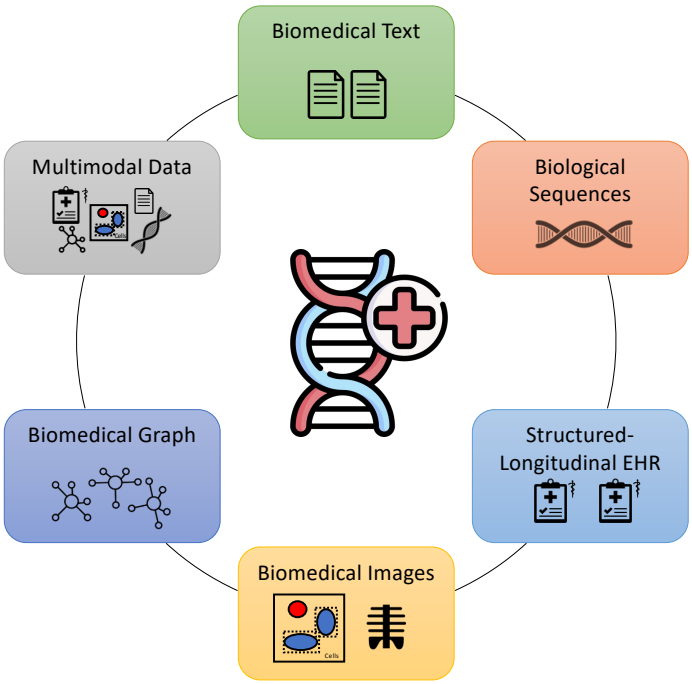


Figure 3.: Various data modalities in biology and medicine.
Credit statement: Image uses icons made by Freepik from www.flaticon.com.

certain research inquiries (e.g., for identification of disease risk factors that are casually linked) that aim to understand mechanisms of complex diseases, which are typically the focus of observational studies [95].

High-throughput imaging technologies give rise to another relevant modality, namely biomedical images such as X-ray, computer tomography, and magnetic resonance imaging. These images and the analysis derived from them serve as a critical tool for biomedical research and clinical routine care. Physicians can localize abnormalities, monitor disease progression, or plan appropriate therapies. Different types of retinal diseases can be classified through image analysis with deep learning as semantic segmentation can assign tissue or organ labels to different regions in images [96, 97]. Solutions for detection of lesions and abnormalities can, for instance, help to categorize tumors in various stages or evaluate pathological features [98]. Moreover, live cell imaging provides a way to study living cells over a certain time period using optical microscopy for an analysis of drug responses.

Biomedical graphs are another data modality as briefly discussed in the previous section. Such graphs can embed prior knowledge and facts on biological systems, where nodes correspond to biomedical concepts or entities and edges represent the various interactions occurring between them.

These graphs can comprehend a wide range of scopes and also reflect the multi-scale nature of biomedicine. For instance, a indication-specific mechanistic knowledge graph [99], curated human signaling pathways [100], or a broadly-defined graph for precision medicine [53] are examples of graphs covering vast knowledge over various human diseases, approved drugs, and biological processes. Researchers leverage these graphs to infer new insights by analyzing the underlying network, its topology, and associated attributes. This approach has been successfully applied to predict a range of biomedical relations, including interactions between drugs [101], interactions between proteins [102], identification of potential therapeutic targets [103], and the repurposing of approved drugs for new diseases [104].

Biomedical text serves as a primary medium for humans to disseminate knowledge and communicate findings in the field of biomedicine. This textual content is classified as unstructured data, can use highly-specialized vocabulary, and is written in a variety of natural languages. It encompasses scientific literature published in journals and conferences, clinical notes within EHR systems, clinical guidelines, specific case reports, social media text, and other related biomedical texts. Given the wide scope of biomedical research and real-world data, it contains diverse types of relational information. Many efforts have been focused on leveraging this data to extract genes/proteins events [105], molecular relations [106], drug-target relations [107], adverse drug events [108], radiology-related information [109], temporal relationships [110], and more. Our previous work, such as Biological Expression Language Information Extraction Workflow (BELIEF) [17, 23, 24, 35], and tools such as Integrated Network and Dynamical Reasoning Assembler (INDRA) [111, 112], have focused on detecting and capturing relations, enriched with context annotations, across biological scales from scientific literature to enable the assembly of biological graphs. In the past, initiatives like International Molecular Exchange Consortium (IMEx) have coordinated expert curation to update databases with new interaction data, notably for Alzheimer's disease research [113]. For a comprehensive overview of the biomedical text mining field, we refer to Chapter 3.

All the aforementioned data modalities can be integrated with each other, producing multimodal data. Observational clinical cohort studies often collect multimodal data on patients covering longitudinal EHR, imaging, clinical notes, cognitive functioning tests, and deep phenotyping data. Significant amount of work has been conducted to integrate two or more modalities into a single AI model in order to address various types of biomedical research questions [114]. Sakhovskiy and Tutubalina [115] built a multimodal model to classify adverse drug reactions using text and drug embeddings. In the clinical context, Liu *et al.* [116] have explored the integration of unstructured

medical text with structured EHR data for medication recommendation and diagnosis coding. A combination of EHR and imaging have been shown to improve the differentiation between common causes of acute respiratory failure [117]. Despite these efforts, the field of multimodal data analysis is still subject to several challenges. These include learning a good representation of each modality to handle the high heterogeneity of data as well as how to effectively fusing information between different modalities [114, 118].

Beyond the discussed data modalities, there are other modalities in biomedicine that can also provide relational information, such as microbiome data, metabolites data, wearable and ambient sensor data, and environmental data. For a comprehensive overview on the presented and additional biomedical data modalities, please refer to [43, 114, 119].

In the next chapters, we focus on two major application fields of data mining, namely biomedical text mining and protein sequence analysis.

3. Biomedical Text Mining

Biomedical text mining, also known as bioNLP, is a field that focus on instructing machines to understand biomedical natural language [120]. Particularly, it involves developing algorithms and tools to extract meaningful information from biomedical scientific text, clinical notes, and other related texts. It streamlines the process of knowledge extraction and curation, allowing for the rapid expansion and enrichment of biomedical databases and resources with the up-to-date information from a wide range of sources [120]. By facilitating the discovery and structuring of new knowledge, bioNLP plays a crucial role in accelerating biomedical research.

3.1. Common Tasks in Biomedical Text Mining

This section introduces a collection of tasks that are relevant to the extraction of relations within the scope of biomedical text mining.

Document classification. Document classification refers to the task in which a pre-defined set of topics are assigned to a given document [121]. In ML context, the assignment of topics to documents can be treated as a multi-label classification problem, where each topic is treated as a binary classification task [121]. The most prominent example of this task is the indexing of MEDLINE with the Medical Subject Headings (MeSH) vocabulary. MEDLINE is a journal citation database of the National Library of Medicine (NLM) that can be browsed through the search engine PubMed [122, 123]. Indexing it with MeSH vocabulary enables advanced semantic search capability in PubMed. In the MEDLINE 2022 Initiative, NLM transitioned to a fully-automated indexing of MEDLINE using the tool Medical Text Indexer that uses natural language processing (NLP) techniques and extracts MeSH terms solely based on title and abstract of an article [124]. However, NLM still uses human review and curation of random samples of the automatic indexing to ensure the MeSH indexing quality [125].

Named entity recognition (NER) and named entity linking (NEL). NER is the process of identifying and categorizing specific biomedical named entities, such as genes, proteins, diseases, drugs/medications, medical diagnoses, treatment procedures, and many more, from text [121]. Essentially, we want to find the boundaries of such named entities in text. NEL extends NER by harmonizing and linking these named entities to controlled biomedical terminologies or ontologies [121]. Both tasks are often modeled as sequence labeling tasks. Some examples of semantic search engines that allow to browse NER and NEL results are SPIKE [126], EuropePMC [127], and SCAIView [26]. These results can also be accessed through dedicated APIs, which enables efficient downstream analysis. NLM is also assessing the incorporation of genes and chemicals detection tools for PubMed [124]. We refer to Chapter 6 that provides specifics on latest biomedical NER methodologies.

Relation extraction (RE). Once NER has discovered the relevant entities in text, RE can be applied in order to identify and classify relationships between these biomedical concepts [120]. Common RE tasks in biomedicine include the extraction of PPIs, drug-disease associations, genotype-phenotype relations, and chemical-protein interactions [121]. Many such interactions between entity classes lead to the definition of biological pathways [121]. The types of these relationships can be simple associations (such as co-occurrence, tri-occurrence, correlated with, related to, causally related to, or interacts with) or complex pre-specified types including participation relation (such as part of, has participant, or agent in), spatial relation (such as contained in, contains, located in, or location of), or temporal relation types (such as derives from, precedes, or transformation of) [128] (see also Table 1). The Relation Ontology (RO) contains a collection of relationship types for standardization across ontologies and knowledge graphs [128, 129].

Many communities such as BioCreative, BioNLP Shared Task, and BioASQ have organized challenges with regards to RE. Their initiatives have played a significant role in developing new benchmarking datasets and advancing the bioNLP field. For instance, our BioCreative challenges [17, 21, 32, 37] focused on developing text mining systems for extracting complex molecular interactions between biological and chemical entities. Similarly, other challenges facilitated the extraction of PPIs [130], drug-drug interactions [131], and temporal relations in clinical records [132]. The thirteenth BioASQ challenge, organized in 2025, encourages the bioNLP research community to focus on the task of gut-brain interplay information extraction [133, 134].

We refer to Chapter 6 for an overview on the latest biomedical RE methodologies.

Literature-based discovery (LBD). The task of LBD, also known as hypothesis generation, covers the discovery of novel biomedical knowledge from the literature using computational approaches. The main objective is to leverage explicitly stated relationships in text in order to infer new relationships that are not explicitly mentioned. Using the Swanson's ABC co-occurrence model, LBD systems derive the discovery of 'A implies C' by leveraging explicit relations within the text such as 'A implies B' and 'B implies C' [135]. For example, evidence might suggest that physical activity (A) improves cardiovascular health (B). Another evidence suggests that improvement of cardiovascular health (B) could lead to improvement of cognitive function (C). Researchers could use this information to infer that physical activity (A) might have a positive effect also on cognitive function (C), leading them to investigate this relationship further by conducting new experiments. Although LBD has led to some discoveries in the past, there are critics who argue that the field lacks a formal definition, comprehensive benchmark datasets to evaluate LBD methods, and has not made significant progress in recent years, despite the advancement of ML techniques [136].

3.2. Challenges of Biomedical Text Mining

Although numerous database providers, such as DisGeNET [137], UniProt [138], PubMed [123], DrugBank [66], incorporate certain biomedical text mining tools to enrich their databases, the integrated information is often viewed as having low confidence [139]. This is mainly due to the inherent challenges in text mining and the potential for noise, errors, and biases in the extracted information. To ensure the reliability and accuracy of the curated information, database providers apply additional measures and quality control processes. Specifically, they offer the extracted information as supporting material to their curators, enabling them to thoroughly assess and validate it while simultaneously improving curation efficiency. In the following, we discuss some of the main challenges in bioNLP.

Information quality. To ensure the quality of the extracted information, the systems essentially need to clearly identify correct biomedical concepts and their interactions. They do not only need to effectively handle errors, noise, ambiguity, and conflicting information in a given text, but also infer whether the findings contradict any established knowledge. Moreover, the identification of speculation, negation, and additional evidence related to a specific interaction will further contribute to strengthening the reliability of information [139]. It is important to note that while most bioNLP research has

focused on abstracts of scientific articles, the majority of the crucial information, including study details and knowledge claims, can only be obtained by having access to the full text (in a more easily parsable format). Therefore, expanding text mining efforts beyond abstracts is essential to ensure comprehensive access to pertinent biomedical knowledge and improve the overall trustworthiness of the information.

Contextual understanding. To identify the correct context of the extracted information requires an understanding of specialized biomedical terminology. For instance, a PPI might have been identified in a particular tissue or organism. Without such contextual information that defines the boundaries within which the interaction operates, usage of such interaction in downstream experiments may yield inaccurate results. Accurately interpreting the meaning of biomedical concepts and identifying the context presents a formidable challenge, especially considering the evolving nature of scientific knowledge [139].

Data Heterogeneity. Biomedicine covers a wide range of subfields that produce a diverse nature of data modalities, as discussed in the previous chapters. Hence, biomedical texts contain complex and heterogeneous information equipped with specialized domain terminologies. Moreover, such information originates from various sources such as clinics, research institutes, experimental laboratories, academic journals, and medical conferences that focus on diverse topics, and could contain multi, inter, and transdisciplinary knowledge. This heterogeneity makes it challenging to harmonize and extract relevant information. The models need to be equipped with a general understanding of multiple languages and, at the same time, they need to be aware of domain-specific terminologies, ontologies, and the nature of data.

We refer to Chapter 6 that provides an overview on latest methodologies to analyze biomedical texts.

4. Protein Sequence Analysis

Protein sequences encompass the sequential arrangement of amino acids for building proteins. While biological sequences like deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) also exist, our primary emphasis in this work lays on analyzing protein sequences. The analysis of protein sequences — and proteins in general — has facilitated biological discoveries by unveiling detailed knowledge about their structure, function, and evolutionary relationships [140].

Protein sequences resemble natural language as both are represented by a string of characters. This shared characteristic, along with the recent advancements in NLP, has motivated researchers to leverage NLP techniques for the analysis of sequences [141]. However, there remain significant differences, since human languages contain separable structures like words, sentences, and paragraphs, whereas we have less knowledge on functional units in protein sequences [140]. Nevertheless, similar to natural language, protein sequences can also be transformed into vector representations. These vector representations of protein sequences can then be utilized to tackle a wide range of tasks and challenges [55].

4.1. Common Protein Sequence-Based Prediction Tasks

This section introduces a collection of tasks that are within the scope of protein sequence analysis.

Protein function prediction. The objective of this task is to predict one or more functions of a protein by solely analyzing its amino acid sequence. For the prediction task, GO terms are commonly used as labels [142]. The GO terms describe the known or predicted functions of proteins based on experimental evidence or computational predictions, respectively. Moreover, the prediction of protein function is valuable for understanding the involvement of proteins in disease pathobiology, determining the functions of metagenomes, and identifying potential drug targets [142]. Numerous databases serve as

valuable repositories for protein function information. Some examples include UniProt Knowledgebase (UniProtKB) [138] that organizes sequence and functional information for millions of proteins, Braunschweig Enzyme Database (BRENDA) [143] that collects information on enzymes, and MoonProt database [144] that gathers information on moonlighting proteins that have multiple functions [145].

Interaction prediction. This task refers to the prediction of protein-protein or protein-RNA interactions based on protein sequences. Protein-protein interaction prediction involves identifying pairs or groups of proteins that are likely to interact with each other in a biological system [146]. Protein-RNA interaction prediction essentially identifies RNA-binding sites on a protein's surface using its sequence [147]. It focuses on determining the specific amino acids within the sequence that are capable of interacting with RNA molecules [147]. Protein interactions play a crucial role in many cellular processes and understanding these interactions can provide insights into the functional relationships within a cell or organism [148]. They are instrumental in determining the mechanisms underlying both normal and disease states [148]. Databases such as BioGRID [149], IntAct [90], and STRING [150] offer a compilation of experimentally-validated and predicted protein-protein and protein-RNA interactions.

Site prediction. In the context of protein sequences, site prediction refers to the task of identifying specific functional sites or regions — essentially a span of amino acids — within the sequences. These sites may encompass binding sites for PPIs [151] or protein-ligand interactions [152], catalytic sites, post-translational modification sites [153], or other regions of interest that play a functional role in the biological activity of the protein. To predict the location of the functional sites, computational methods typically integrate local , contextual, and global features extracted from protein sequences [151]. The PiSite database provides information on actual binding sites of individual proteins at the residue level [154], however, their predictions are based on 3D structure rather than the sequence.

Protein structure prediction. Understanding the secondary and tertiary structure of a protein is essential for making inferences about its function as a whole, however, experimentally determining protein structures is challenging [49, 155]. The protein structure prediction task aims to determine the 3D structure of a protein based on its amino acid sequence using computational methods. The development of the deep learning model AlphaFold propelled

the task of protein structure prediction to new heights [49, 50, 156]. The open access database, namely AlphaFold Protein Structure Database, provides over 200 million AlphaFold structure predictions for proteins of various organisms [157].

4.2. Challenges of Protein Sequence Analysis

Despite significant progress in the field of protein sequence analysis, there are still numerous challenges.

Sequence variability. Sequence variability refers to the possibility of proteins with similar sequences exhibiting different functions. Additionally, individual proteins can exhibit multiple functions [145]. This makes the prediction of protein function solely based on sequence data a challenging task. While structural information can aid in identifying functions, it is often unavailable for many proteins. However, recent advancements made by models such as AlphaFold [49, 50, 156], which can predict protein structures, offer the potential to enhance protein function prediction by leveraging sequence and structural features simultaneously.

Protein sequence length. Protein sequences can vary greatly in length, consisting of hundreds or even thousands of amino acids. For instance, according to the Swiss-Prot database [138], Titin [158] is the largest known protein with a length of approx. 34,350 amino acid residues [159], whereas T cell receptor delta diversity 1 [160] is considered to be the smallest polypeptide (small protein) with a length of two amino acids in humans. Recent works on protein sequence representation learning ignored small protein sequences that are smaller than 20 amino acids [55]. Handling and analyzing such data can be complex and computationally intensive as huge memory requirements are needed for larger proteins. However, some deep learning models utilize fragmentation strategies (such as splitting sequences into smaller pieces) to reduce the computational effort [55].

Integration with other modalities. To attain a holistic understanding of biological processes, it is often necessary to use protein sequence information together with other data modalities. As mentioned earlier, incorporating protein structural information could enhance the accuracy of protein function prediction [145, 161]. Additionally, utilizing RNA sequence and structural information can be valuable in predicting protein-RNA interactions [147].

Embedding homology information in deep learning models has shown to improve the performance for secondary structure and contact prediction tasks [161, 162]. Hence, researchers [147, 161] suggest focusing on building multimodal machine learning models to learn the best features from each data modality that are suited to solve the tasks at hand.

We refer to Chapter 6 that provides an overview on latest methodologies to analyze protein sequences.

5. Transfer Learning

Transfer learning is a machine learning technique that enables models to leverage knowledge gained from one source domain and applies it to another related target domain [163]. There are two main subcategories, namely homogeneous and heterogeneous transfer learning. Homogeneous transfer learning refers to the case where the source and target domains have similar feature and label space (e.g., performing a document classification task by transferring knowledge from domain-specific textual corpus). Conversely, heterogeneous transfer learning aims to transfer knowledge in the setting where source to target domain have distinct feature and label spaces (e.g., generating textual image descriptions by using an image-based object detection model). For a comprehensive overview of current transfer learning approaches, we refer to Zhuang *et al.* [163] and Wang and Chen [164]. In this work, we focus on parameter-based homogeneous transfer learning, specifically model pre-training and fine-tuning. Here, the general idea is that the weights learned by utilizing previous knowledge from the source domain can be transferred to the target domain, thereby reducing the amount of labeled data and training time required for tackling the target domain task [164].

To enable transfer learning, the models must acquire general features, patterns, and representations from data, essentially capturing the underlying structure of the data. Although traditional machine learning methods (such as support vector machines, [165] and random forests, [166]) were simply trained on the available dataset and directly utilized, it has been shown that a two-step process of pre-training and fine-tuning can further improve model performance, particularly in deep learning [164]. First, during pre-training, the model is trained on a large dataset in a self-supervised manner to learn a general understanding of the data (see Figure 4). In self-supervision, the model learns from the data by generating and utilizing labels derived from the data itself (e.g., masking certain words in text and training the model to predict them). Then, during fine-tuning, the pre-trained model is fitted to a specific learning task by an additional supervised training on a smaller labeled dataset (see Figure 4). The two-step process allows the models to leverage the general knowledge and adapt it to perform well on a specific downstream task [164]. Pre-training typically requires significant computational resources depending on the size of the model architecture, scale of the pre-training

datasets, and the number of epochs (e.g., training rounds). Whereas, fine-tuning is considered computationally efficient as it requires less data and often very few epochs. The computational effort can be further reduced by freezing one or more layers of the pre-trained model. However, freezing layers could lead to a decrease in performance [167, 168], and therefore should be considered carefully.

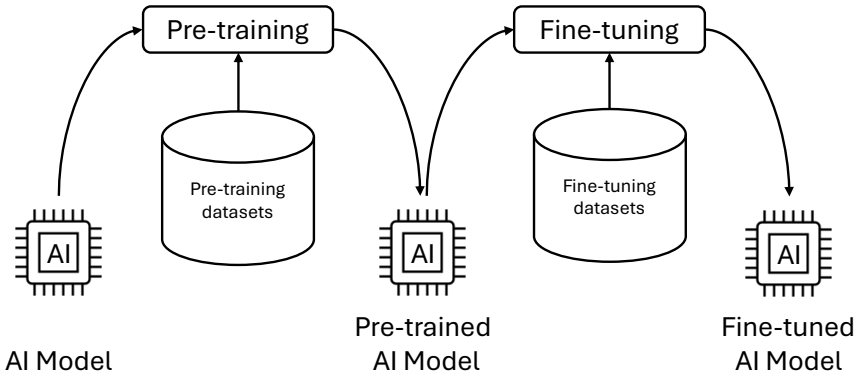


Figure 4.: Pre-training and fine-tuning paradigm

Transfer learning has been successfully applied in various domains, including computer vision, NLP, and biomedical research, where labeled data may be limited or costly to obtain. Current state-of-the-art models are built on the transformer architecture. Some prominent examples of models that utilize the pre-training and fine-tuning mechanism are bidirectional encoder representations from transformers (BERT) for language processing [169], vision transformers for image processing [170], and generative pre-trained transformer (GPT) for language generation [171]. In biomedicine, numerous studies have proposed new variants of these models and applied them to tasks such as bioNLP, image analysis, or drug discovery [84, 172].

5.1. Transformer Architecture

Before diving into the introduction of the current state-of-the-art models, we discuss the transformer architecture that represents the backbone of these models. The original transformer [54] was proposed as a neural network following an encoder-decoder architecture that utilizes the attention mechanism. Figure 5 illustrates the original transformer architecture. The objective of the encoder is to capture the contextual information in the data by finding the relationships and dependencies between the tokens of the input sequences.

The decoder is used to generate the output sequence for a given task. In other words, a given input sequence $(x_1 \dots x_n) \in \mathbb{R}^{n \times d_{\text{model}}}$ is mapped to a sequence of continuous representations $(z_1 \dots z_n) \in \mathbb{R}^{n \times d_{\text{model}}}$, which are used by the decoder to generate an output sequence $(y_1 \dots y_n) \in \mathbb{R}^{n \times d_{\text{model}}}$ one element at a time, where n are the number of tokens and d_{model} is their dimension [54].

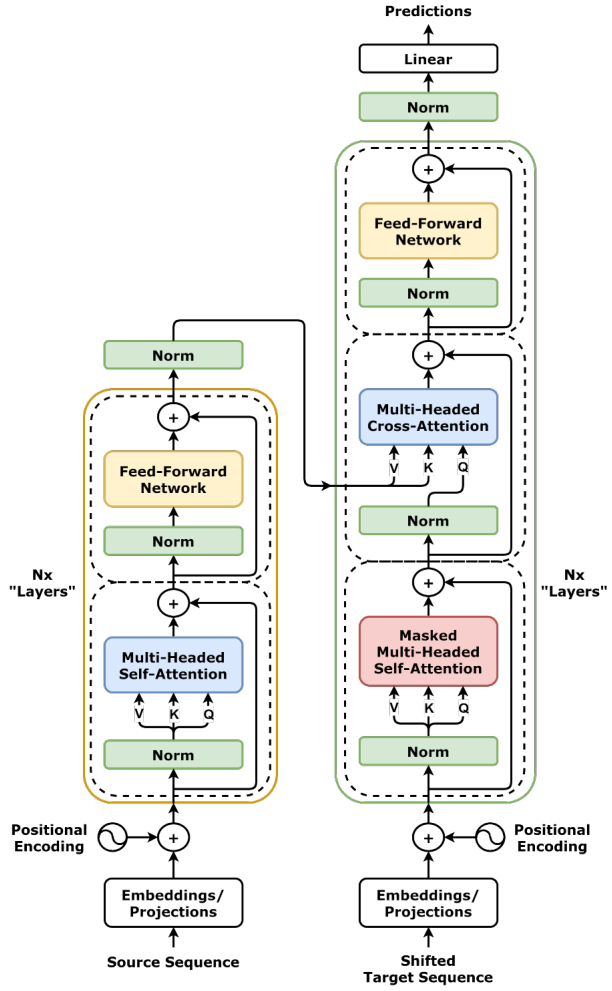


Figure 5.: The architecture of the original transformer model. The left yellow-colored box represents the encoder and the right green-colored part illustrates the decoder.

Source: This image by Daniel Voigt Godoy (`https://github.com/dvgodoy/dl-visuals`) is licensed under CC BY 4.0.

5.1.1. Encoder and Decoder

Initially, the input sequence is transformed into high-dimensional vector representations, also called word embeddings. The transformation process assigns

a unique identifier to each token in a vocabulary. An embedding matrix is used to look up the identifier and retrieve the corresponding word embedding vector. As the transformer treats the input sequence as an unordered bag of words, a positional embedding is added to the word embedding to retain the positional information of the input. The positional encoding $p \in \mathbb{R}^{n \times d_{\text{model}}}$ is computed by

$$p_{k,2i} = \sin\left(\frac{k}{10000^{2i/d_{\text{model}}}}\right) \text{ and} \tag{5.1}$$

$$p_{k,2i+1} = \cos\left(\frac{k}{10000^{2i/d_{\text{model}}}}\right) \tag{5.2}$$

using the position k of the token in the sequence, where n being the number of tokens in the input sequence if considering a single sequence, d_{model} being the dimensionality of the embedding, and i representing the index of the dimension in the embedding.

The encoder consists of the N stacks, which consist of a multi-head attention layer and a feed-forward layer (Figure 5). Both sublayers have residual connections around them and are followed by normalization layers:

$$u_i = \text{LayerNorm}(x_i + \text{Sublayer}(x_i)). \tag{5.3}$$

We refer to Ba *et al.* [173] for the definition of LayerNorm function, representing the normalization layer.

The decoder also consists of the N stacks. As depicted in Figure 5, the decoder includes an additional sublayer, known as masked multi-head attention layer, positioned before the other two sublayers that mirror those of the encoder (see next section for details). The output of the encoder is used in the multi-head attention layer as keys and queries, whereas the output of the masked multi-head attention is used as values. Similar to the encoder, all three sublayers have residual connections around them and are followed by normalization layers.

5.1.2. Attention Mechanism

The multi-head attention layer implements self-attention that allows the model to attend each token to all other tokens in the input sequence. It uses the scaled dot-product attention layer (see Figure 6) that takes the query Q, the key K, and the value V matrices as input, defined as:

$$Q = W_i^Q \cdot x, K = W_i^K \cdot x, V = W_i^V \cdot x, \tag{5.4}$$

where $x \in \mathbb{R}^{n \times d_{\text{model}}}$ and the weight matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, and $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ with d_k, d_q (where $d_k = d_q$), and d_v being the dimensions. A dot-product is computed for the query with all keys, which is scaled by $1/\sqrt{d_k}$. A softmax function [174] is applied on the scaled dot-product to obtain the weights on the values. The final matrix can be considered as a single-head attention:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V. \tag{5.5}$$

The multi-head attention allows to jointly use different representation subspaces of the given queries, keys, and values. Intuitively, the model can combine knowledge from different behaviors of the same attention mechanism, such as capturing dependencies of various ranges (short or long) within a sequence. Specifically, given a query, key, and value matrix, let h be the number of attention heads. h subparts of this matrix are calculated, which are passed through the scaled dot-product attention independently to retrieve the h heads:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \tag{5.6}$$

In the last step, the h heads are concatenated and linearly projected to obtain the final attention:

$$\text{MultiHead}(Q, K, V) = \text{concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{5.7}$$

where $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

The special case of masked multi-head attention layer adds an additional masking operation. While performing a self attention the model attends each position in the input sequence to all other positions, the masking operation ensures that the attention calculation does not consider future positions. The decoder should only have access to previous positions. This is required to ensure the autoregressive nature of the decoder during training or inference.

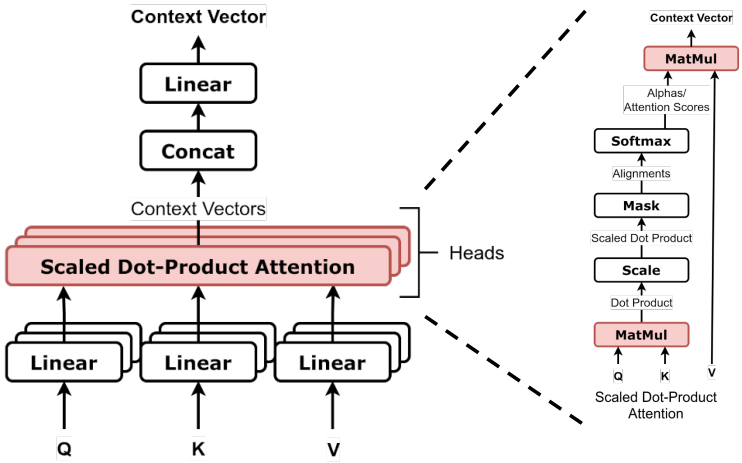


Figure 6.: The left part represents the multi-head attention. The right part visualizes the layers in scaled dot-product attention.

Source: Original images by Daniel Voigt Godoy (`https://github.com/dvgodoy/dl-visuals`) licensed under CC BY 4.0.

5.2. Pre-trained Transformer-Based Models

The transformer has enabled the development of new state-of-the-art models for various NLP tasks ranging from text classification to information extraction and text generation [169, 175]. The studies on the development of new transformer-based models share a common pattern. They utilize at least some parts of the transformer architecture (e.g., encoder, decoder, and attention mechanism), perform pre-training on huge datasets, and fine-tuning on smaller datasets for specific downstream tasks. Furthermore, the fine-tuned models are empirically evaluated on the test sets of the respective downstream tasks.

5.2.1. Bidirectional Encoder Representations from Transformers (BERT)

BERT was among the first models to apply the transformer architecture in combination with pre-training and fine-tuning to solve NLP tasks. It demonstrated effectively that large-scale pre-training on huge amounts of text can lead to significant improvements in various downstream tasks. The model proposed by Devlin *et al.* [169] employs just the encoder of the original transformer (Figure 5) by using a stack of several encoder blocks on top of each other. The encoder is specifically designed to produce contextual word embed-

dings from the given text. It captures the meaning of a word by considering its surrounding context, allowing it to produce representations that are sensitive to the specific context in which the word appears. This is achieved by using bidirectional training where every token can attend to both left and right tokens of a given textual sequence. In this way, the context from both directions can be incorporated by the model, which has been shown to improve language understanding [169]. The model also includes an additional classification head. This head is specific to the downstream tasks and takes the representation of the input sequence produced by the encoder. Typically, the classification head consists of one or more linear layers followed by an activation function (e.g., sigmoid or softmax [174]).

We now discuss the BERT architecture in more detail. BERT takes a sequence of words as input. This sequence undergoes a preprocessing step called wordpiece tokenization, which splits the sequence into individual tokens. These could be either words or smaller segments of words, which are a string of characters commonly found in other words. For example, the word "methylation" is tokenized into "meth", "##yla", and "##tion", where the "##" symbol indicates the current token's connection to the preceding token. The wordpiece tokenization allows BERT to handle words that are not part of the vocabulary, obtained from the pretraining dataset, and thereby capture more fine-grained information [169]. The tokenized input is then converted into numeric vector representations, similar to the original transformer model. The BERT output can be either a sequence embedding or token embeddings, depending on the specific prediction task at hand. The token embeddings represent every token in the sequence, while the sequence embedding is the pooled output of all tokens and, hence, represents the whole input sequence. For instance, the sequence embedding can be employed for sentence classification, while token embeddings are more suitable for tasks such as named entity recognition.

BERT utilizes two self-supervised prediction tasks, namely masked language modeling (MLM) and next sentence prediction (NSP), as pre-training objectives to train the model [169]. During training with MLM, a subset of typically 15% of the input tokens is masked at random and the task of the model is to predict only these masked tokens based on their surrounding context. MLM task helps the model to learn deep contextual relationships between words. The NSP task is especially designed to train the model to understand relationships between two consecutive sentences. By predicting the next sentence, the model gains insights into sentence coherence. In addition to the pre-training objectives, BERT pre-training made use of the two BookCorpus [176] and English Wikipedia corpora that consisted of 800M and 2,500M words, respectively. After the generation of the pre-trained BERT

model, the authors fine-tuned it by using diverse NLP datasets to evaluate its performance. BERT showed state-of-the-art performances on various NLP tasks such as general language understanding, question answering, NER, and grounded commonsense inference.

BERT Variants Many variants of BERT have been published in recent years. For instance, Robustly Optimized BERT Approach (RoBERTa) [177] removed the NSP task and instead focused exclusively on MLM as the pre-training objective. Furthermore, it had employed additional training data and optimized the pre-training phase by training longer and with different hyperparameters. The A Lite BERT (ALBERT) [178] model focused on techniques to reduce model parameters, which lowered the memory consumption and increased the pre-training speed. Additionally, ALBERT replaced the NSP objective with sentence-order prediction, which predicts the correct order of two sentences. Both techniques were found to be helpful to increase the efficacy and performance of the ALBERT model on downstream tasks. For further information and in-depth details on additional variants, including DistilBERT [179], Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) [180], and Sentence-BERT [181], please refer to [182].

5.2.2. Generative Pre-trained Transformer (GPT)

GPT or GPT-1 model is another variant of the original transformer model [183]. GPT only uses the decoder block of the transformer, but without the multi-head attention layer as there is no encoder involved. Hereby, multiple decoder blocks are stacked in a multi-layer architecture. The masked multi-head attention layer in the decoder uses the so-called left-to-right mechanism for self-attention, where every token can only attend to previous tokens. In addition to the decoder block, GPT employs the input embedding layer with positional encodings identical to the original transformer model. Furthermore, GPT is considered as a generative model that can produce human-like text by considering the context of the input textual sequence.

Radford *et al.* [183] used unsupervised pre-training by using the large BookCorpus dataset [176] to learn the underlying structure of the data. This process essentially adjusts the weights of the model based on the language modeling pre-training task. Given a sequence of words, the language modeling task predicts the probability distribution of the next word. In this way the model learns to understand the relationship between words and generate semantically meaningful text. After the pre-training, they followed with a supervised fine-tuning process to tackle the downstream target tasks. They

used datasets of various tasks such as natural language inference, question answering, sentence similarity, and text classification, where they achieved state-of-the-art results on various datasets.

The research community has built various models based on GPT-1. GPT-2 [184] and GPT-3 [175] can be considered as the direct successors of GPT-1. Both models use more decoder blocks in their architecture, which increases the size of learnable model parameters. In comparison to GPT-1 that has 117 million parameters, GPT-2 and GPT-3 contain 1.5 billion and 175 billion learnable parameters, respectively. They also utilize additional pre-training data. By doing so they improved the model performances on general language modeling. GPT-2 was trained on around 40GB, whereas GPT-3 used 570GB of text data. Further variants of GPT are GPT-J [185], DistilGPT2 [179, 186], GPT-3.5 [187], InstructGPT [188], and ChatGPT [189], of which only some are available as open source models. For more details on variants of GPT, please refer to Kalyan [190].

5.3. Biomedical Pre-trained Transformer-Based Models

The significant achievements and widespread adoption of pre-trained transformer-based models such as BERT, GPT, and their variants inspired the research community to experiment and adapt these models to the bioNLP domain. A timeline for the development of these models that are relevant to the biomedical domain is illustrated in Table 2. Particularly, the period after the publication of BERT and GPT models marks the introduction of groundbreaking transformer variants, which have fundamentally transformed the way biomedical data is analyzed.

Models including BioBERT [191], PubMedBERT [192], BioELECTRA [193], SciBERT [194], and ClinicalBERT [195] have been designed to accommodate the needs and requirements of bioNLP domain. Typically, these models perform an additional pre-training on a huge biomedical textual corpus while utilizing the already pre-trained weights of the general language understanding model BERT as their initialization.

Although the aforementioned models are designed for textual data, there are other data modalities in biomedicine, which are also of sequential nature. Studies have shown that transformer-based models also perform well on different sequential data modalities such as biomedical images, structured EHR data, and protein or DNA sequences. For instance, MedBERT [85], BEHRT [196], and our work on ExMedBERT [6] have shown significant improvements on various tasks of structured EHR data. Protein or DNA sequences have been used to pre-train models including ProtTrans [55], Evolutionary Scale

Modeling (ESM)-1b [56], and DNABERT [197]. ESM-2 [198] and the successor AlphaFold2 [50] have made significant achievements in predicting the 3D structures of proteins by using protein sequences as input.

For more details on pre-trained transformer models that are being applied in the field of biomedicine, we refer to Wang *et al.* [199] and to our publication [1] that builds the foundation for Chapter 6.

Table 2.: Timeline for the development of transformer-based models that are relevant to the biomedical domain for various data modalities. The modalities B. Text, B. Seq., and B. Graph stand for biomedical text, biological sequence, and biomedical graph, respectively. Models marked with an asterisk (∗) represent our own work.

Date	Modality	Model name
Dec. 2017	Text	Original transformer [54]
Dec. 2018	Text	GPT-1 [183]
Jun. 2019	Text	BERT [169]
Feb. 2020	B. Text	BioBERT [191]
Apr. 2020	B. Text	ClinicalBERT [195]
Apr. 2020	EHR	BEHRT [196]
Nov. 2020	B. Text	BioMegatron [200]
Dec. 2020	Text	GPT-3 [175]
Jan. 2021	Graph	Graph Transformer [201]
Apr. 2021	B. Seq.	ESM-1b [56]
Jun. 2021	Images	Vision Transformer [170]
Jul. 2021	B. Seq.	ProtTrans (ProtBERT) [55]
Aug. 2021	B. Seq.	AlphaFold 2 [50]
Jan. 2022	B. Text	PubMedBert [202]
Jul. 2022	B. Seq.	ProtGPT2 [203]
Sep. 2022	B. Seq.	∗STEP [3]
Sep. 2022	B. Text	BioGPT [204]
Nov. 2022	Text	ChatGPT [205]
Dec. 2022	B. Text	∗Bio-GottBERT [13]
Feb. 2023	EHR	Hi-BEHRT [206]
Jul. 2023	Text	LLaMA 2 [207, 208]
Sep. 2023	EHR	∗ExMed-BERT [6]
Mar. 2024	Text	GPT-4 [189]
Jun. 2024	B. Seq.	AlphaFold 3 [156]
Jan. 2025	Tabular data	TabPFN [209]

5.4. Prospects and Limitations of Recent Large Language Models

Although the contributions of our work has mostly considered the encoder-only BERT models for biomedical tasks, there are noteworthy developments in the generative artificial intelligence (GAI) field that also spark new hopes for biomedicine. We shortly want to introduce and discuss these solutions and put them in context to this work.

Various solutions that provide an interface to LLMs possessing generative capabilities (such as for text and images) were published in recent years. They have shown impressive capabilities in understanding and generating the natural language. ChatGPT (GPT-3.5 and GPT-4) from OpenAI, Gemini [210] from Google, and Large Language Model Meta AI (Llama) Chat [211] from Meta are some examples of GAI solutions, also referred to as chat bots. There are certain principles that these solutions and their integrated LLMs have in common. A noteworthy aspect is that these solutions required significant software engineering effort to scale their access to millions of users. Furthermore, the integrated LLMs have basically used a modified transformer architecture, pre-trained with self-supervised learning on large-scale multi-lingual datasets, fine-tuned using reinforcement learning with human feedback (or so-called instructions), and utilized huge computational resources. During reinforcement learning (RL), which represents a subfield of machine learning, an agent (e.g., a model) is instructed to take actions that maximize a reward signal [212]. Hereby, the human feedback provides examples of correct responses and ranking of different responses. By fine-tuning the models with human feedback in a RL context, these LLMs have been improved considerably [188]. By doing so, the models also support zero/few-short learning (also called in-context learning) using prompts, which allow the users to include one or more examples in their inquiry. By providing this additional context, the models can produce accurate predictions for tasks without being explicitly trained on.

Most of these solutions are free to use but remain closed source, only some (e.g., Llama) are openly available. Reproducing these models in academic settings is difficult due to the lack of availability of computational resources and large-scale datasets. Although this hinders an in-depth analysis of models and extensive experimentation, some researchers are actively assessing the potential of these models to address and overcome existing challenges of information extraction and representation in biomedicine. For instance, Babaiha *et al.* [213] evaluated the performance of ChatGPT for extraction of cause-and-effect triples and benchmarked the output against human-curated graphs.

They discovered various limitations and shortcomings such as missing information and inaccurate harmonization. Similar discovery has been made by others [214, 215], who benchmarked LLMs for various biomedical-related tasks including NER, RE, document classification, and question answering in a zero or few-shot scenario discovering varying performances on almost all tasks. These studies emphasized task-specific fine-tuning of LLMs for optimal performance.

Wang *et al.* [216] fine-tuned Llama 2 model to link rare disease concepts to ontologies. In another study, a fine-tuned Llama model was created by using clinical notes to improve the prediction of diagnosis-related patient groups [217]. Zhou *et al.* [218] focused on improving the clinical RE by adapting the prompts for the Llama model with a new instruction dataset. This dataset contains additional explanations for the clinical concepts and relations. Another noteworthy study utilizes LLM as agents to generate explanations for clinical decision reasoning [219]. The authors argue that their proposed multi-agent framework using generator, verifier, and reasoner agents increases the confidence in clinical decisions by creating sound and reasonable arguments. A systematic analysis of open and closed-source chat bots in comparison with web search was performed in the context of medical question answering that simulated the need of individuals seeking health advice [220]. The analysis revealed the growing potential of LLMs in providing information on diagnoses and examination as well as certain limitations on providing treatment recommendations. The authors emphasized a need for a robust AI in healthcare [220].

Although these LLMs are trained on huge datasets, the amount of biomedical-related data in these datasets is limited. This is due to the unavailability of biomedical datasets (full-text, clinical notes etc.) as they are mostly hidden in closed silos. Hence, these models are not primarily designed for biomedical use cases. Considering these gaps, Wornow *et al.* [221] defines an evaluation framework based on six categories that can be used to measure performance of clinical language and foundation models. The latter are capable of pursuing diverse tasks for different modalities without being explicitly trained for them [221, 222]. The categories include the traditional validation based on metrics, but also measuring simplified model deployment or multi-modal evaluation. Moving beyond the existing narrow versions of AI, Moor *et al.* [222] envision a generalist medical AI that will be able to perform diverse tasks with little labeled data. This type of AI should be built by leveraging various types of biomedical data modalities. However, this requires significant effort on collecting and aligning data modalities and solving many technical challenges (e.g., suitable model architecture) to support multiple modalities.

Part II.

Main Contributions

6. Transformer Models in Biomedicine

This chapter presents an overview of the following survey study [1], which is included in Appendix A.

> **S. Madan**, M. Lentzen, J. Brandt, D. Rueckert, M. Hofmann-Apitius, and H. Fröhlich, "Transformer models in biomedicine," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 214, 2024. DOI: 10.1186/s12911-024-02600-5

Summary

The transformer model [54] originated in the field of NLP and is now a highly popularized deep neural network that has signficantly transformed the landscape of modern machine learning. Its architecture forms the backbone of many prominent LLMs such as BERT [169], GPT [171], and Llama [207, 208]. Such LLMs utilize vast amounts of data to understand the hidden patterns within language itself. They can perform various NLP tasks such as translation, text generation, and question answering. As discussed in the Chapter 5, the self-attention mechanism as well as encoder-decoder structure allows them to learn complex sequence representations and handle long-range dependencies efficiently. These models leverage the transfer learning approach using a two-step process: pre-training on large amounts of unlabeled data to learn general representations, followed by fine-tuning on labeled data for specific tasks. Transformers have advanced various fields in biomedical research, where they can be employed for analyzing diverse data modalities including scientific literature, clinical notes, imaging, and genetic information.

Our study [1] aims to provide a review on the application of transformer-based models covering six biomedical modalitites, namely 1) biomedical text, 2) biological sequences, 3) structured EHR data, 4) biomedical images, 5) biomedical graphs, and 6) multimodal data (see also Figure 3). For each biomedical field, we provide an overview of proposed transformer-based models, along with data sources in use, the underlying model architecture (such as BERT or vision transformer), and on which biomedical tasks these models focused on. In the following, we discuss the application of transformer models covering these modalities.

Biomedical text. In the field of bioNLP, most studies have been conducted with transformer models and we discussed the development of such domain-specific models and their adoption on common bioNLP tasks such as document classification, NER, and RE (see also Section 5.3). For instance, such models were utilized to extract drug-drug interactions from text [223] and lung cancer staging information were extracted from computed tomography reports [224].

Biological sequences. Recently, researchers have used billions of biological sequences (such as DNA, RNA, and protein sequences) across all organisms to build pre-trained transformer models for biological sequences [55, 56, 197]. These models have produced state-of-the-art results on 3D protein structure, protein function, and protein-DNA interaction prediction [55, 56, 225]. AlphaFold 2 [50] represents a prominent model that has demonstrated the ability to predict protein structures with high accuracy. However, these sequence models face similar limitations to natural language models, as they require huge amounts of compute resources, require large training datasets, and can struggle to capture long-range interactions.

Structured EHR data. Another biomedical modality where transformer models have successfully been employed is the structured EHR data, which consists of a multivariate discrete, irregular time series data. As EHR data is inherently different from textual, sequence, or image data, studies have focused on identifying a suitable data representation [196]. They consider the sequences of diagnosis, procedure, and medication codes as a form of language, where the codes recorded during a specific visit act as tokens. In this field, the models have primarily focused on tasks such as hospitalization and mortality prediction, ventilation risk prediction, ICD coding prediction, and generation of synthetic EHR data [85, 196, 226].

Biomedical images. So called vision transformers have shown success in the analysis of biomedical images, where traditionally convolutional neural networks (CNNs) have dominated. By leveraging the self-attention mechanism, they can capture long-range dependencies and global information in order to improve the performance of image segmentation, classification, and anomaly detection. To process the images with vision transformers, they are divided into sequence of patches, which are then flattened into fixed-length vectors, quite similar to tokens. These models struggle to effectively generalize with limited data, however, their performance scales well with growing datasets. For instance, vision transformers were used to analyze lung X-rays to detect COVID-19 disease [227], to detect gastric cancer from histopathological imaging data [228], and to classify brain tumor from magnetic resonance imaging data [229].

Biomedical graphs. Variants of transformers have also been adapted to handle graph-structured data by encoding nodes, edges, and the graph topology. In some cases, they have shown improvements over traditional graph neural networks, for instance on identification of disease targets [230] or on detecting adverse effects of a certain drug [231].

Multimodal data. Modeling complex processes of biology and medicine necessitates the integration and learning across multiple modalities. Transformer-based models have been adapted to handle these multimodal inputs. The applications range from emotion recognition to clinical diagnosis using vision-and-language models [232, 233]. However, the development of universal architectures that can effectively handle diverse modalities and tasks remains a significant challenge.

We concluded the study by discussing the challenges and limitations of transformers and the future research prospects of transformer-based models in the biomedical field. We identified four possible avenues for future research, namely integration of knowledge, integration of multimodal data, generative modeling, and improvements in XAI, for which we expect new developments in the biomedical field.

Authors' contributions

To describe the contribution of each author in our publication [1], we follow the standardized roles defined in Contributor Roles Taxonomy (CRediT) [234, 235].

Sumit Madan: Conceptualization, Methodology, Investigation, Visualization, Writing - Original Draft, and Writing - Review & Editing; **Manuel Lentzen**: Investigation, Writing - Original Draft, and Writing - Review & Editing; **Johannes Brandt**: Writing - Review & Editing; **Daniel Rueckert**: Writing - Review & Editing; **Martin Hofmann-Apitius**: Conceptualization, Supervision, and Writing - Review & Editing; **Holger Fröhlich**: Conceptualization, Methodology, Supervision, Writing - Original Draft, and Writing - Review & Editing.

7. Deep Learning-Based Detection of Psychiatric Attributes from German Mental Health Records

In this chapter, we summarize our publication [2] presented fully in Appendix B.

> **S. Madan**, F. Julius Zimmer, H. Balabin, S. Schaaf, H. Fröhlich, J. Fluck, I. Neuner, K. Mathiak, M. Hofmann-Apitius, and P. Sarkheil, "Deep Learning-based Detection of Psychiatric Attributes from German Mental Health Records," *International Journal of Medical Informatics*, vol. 161, p. 104 724, 2022. DOI: 10.1016/j.ijmedinf.2022.104724

Summary

In the field of psychiatry, information on patient evaluation is predominantly captured and stored in form of text-based EHRs [236]. These documents, containing real-world mental health information, represent an important resource in understanding psychopathological factors of diseases, diagnosing conditions, and aiding further predictive analyses. Techniques of text mining are well-suited to process and analyze such text-based records, enabling the extraction of relevant psychiatric information.

In our work [2], we focused on German MSE reports that are part of the patients' discharge summaries. The MSE serves as a standardized assessment tool performed by psychiatrists for evaluating and describing the mental state and behaviors of patients' [236]. To identify the psychiatric attributes and to link them with pathological conditions (commonly referred to as psychopathological symptoms), our workflow performs three main text mining tasks: NER, RE, and NEL. During the NER task, three types of named entities are extracted: 1) psychiatric attribute, 2) normal assessment, and 3) pathological assessment. Subsequently, in the RE task, the pathological assessment entities are linked to their respective psychiatric attributes. Finally, in the NEL task, these pathological-associated attributes are mapped to the

Association for Methodology and Documentation in Psychiatry (AMDP) psychopathological symptom terms. The AMDP system represents a standard to collect psychopathological findings, physical symptoms, and anamnesis data in mentally ill patients, consisting of a terminology of 140 features of psychopathological symptoms grouped in various classes [237, 238].

Our study data consists of a sample of 660 patient MSEs, originating between 2014 and 2019, from the electronic archives of the Department of Psychiatry at University Clinic Aachen. This data is characterized by medical experts into the categories of ICD-10 [239] mental disorders (such as F20-29 schizophrenia/schizotypal/delusional disorders, F30-39 mood disorders, and more). We anonymized the study data and divided into two datasets: a labeled dataset used for model training and evaluation, and an unlabeled inference dataset intended for system application. The labeled dataset was manually annotated by a medical expert for psychiatric attributes and symptoms, and further verified by a board-certified psychiatrist. The labeled dataset consisted of 150 MSEs, which include an equal number of male and female patients (n=75 each). Furthermore, we randomly splitted this dataset for training (n=100) and for independent testing (n=50). The additional unlabeled inference dataset contained 510 patient MSEs. The gold standard annotations consisted of psychiatric attributes and assessment entities (e.g., normal and pathological). Furthermore, the pathological assessment entities were linked to their related psychiatric attributes, which were mapped to the AMDP terminology.

Our newly-constructed training set was used to fine-tune a pre-trained LLM to detect the named entities. Specifically, we leveraged GermanBERT model [240], which is a pre-trained LLM trained on the German language domain data. For the classification head, we designed a combination of a feed-forward and softmax [174] layers. We used a five-fold cross-validation to evaluate the model and to optimize the hyperparameters of the model. To assess the performance of the fine-tuning process, we measured precision, recall, and F_1-scores for each entity class individually. We tested the best-performing model on an independent test set of 50 MSEs, reaching an F_1-score of 86%-88% for the identification of attribute and the assessment entities. In order to identify the psychopathological symptoms, we designed a rule-based system that linked the psychiatric attributes to pathological assessment entities and map them to the AMDP terminology, reaching an F_1-score of 91%. Furthermore, we applied our full workflow on remaining 510 MSEs to predict patients' psychiatric attributes and symptoms, revealing *Dysphorisch* (eng. dysphoric), *Affektarm* (eng. emotionless), *Konzentrationsstörungen* (eng. concentration disorders), *Antriebsarm* (eng. less energized), and *Aufmerksamkeitsstörungen* (attention disorders) as the top five occurring symptoms.

In conclusion, this study successfully demonstrates the application of a text mining approach to extract relevant psychiatric information from MSE reports using a pre-trained LLM. The promising performance of this approach demonstrates its potential for the analysis of routine psychiatric patient data in German clinical settings, facilitating its secondary usage for further research.

Authors' contributions

To describe the contribution of each author in our publication [2], we follow the standardized roles defined in CRediT [234, 235].

Sumit Madan: Conceptualization, Supervision, Methodology, Software, Visualization, Validation, Project administration, Writing – original draft, Writing – review & editing. **Fabian Julius Zimmer**: Resources, Data curation, Validation. **Helena Balabin**: Software, Writing – original draft. **Sebastian Schaaf**: Writing – review & editing. **Holger Fröhlich**: Writing – review & editing. **Juliane Fluck**: Conceptualization, Writing – review & editing. **Irene Neuner**: Writing – review & editing. **Klaus Mathiak**: Conceptualization, Writing – review & editing. **Martin Hofmann-Apitius**: Funding acquisition, Supervision, Project administration, Writing – review & editing. **Pegah Sarkheil**: Funding acquisition, Project administration, Conceptualization, Methodology, Data curation, Validation, Writing – original draft, Writing – review & editing.

8. Accurate Prediction of Virus-host Protein-Protein Interactions via a Siamese Neural Network Using Deep Protein Sequence Embeddings

In this chapter, we summarize our publication [3] presented fully in Appendix C.

> **S. Madan**, V. Demina, M. Stapf, O. Ernst, and H. Fröhlich, "Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings," *Patterns*, vol. 3, no. 9, p. 100 551, 2022. DOI: 10.1016/j.patter.2022.100551

Summary

Viral infections can cause significant damage to tissues in the human body. For example, viruses such as John Cunningham polyomavirus (JCV) can harm brain cells, leading to neurological disorders [241]. Similarly, some viruses such as SARS-CoV-2 can affect lung cells, causing respiratory issues [242, 243]. Novel therapeutic interventions are required to combat these effects and prevent further tissue damage. Moreover, researchers are designing virus-like particles that open novel opportunities to deliver therapeutics to targeted cell types and tissues [244]. To develop effective treatments, it is crucial to understand the interactions between virus and host proteins. PPI databases such as VirHostNet [245] have limited coverage of virus-host interactions, highlighting the need for advanced ML methods to enhance such databases.

In our work [3], we developed the deep learning architecture Siamese Tailored deep sequence Embedding of Proteins (STEP), designed for predicting virus-host PPIs using only protein sequences as input. The STEP architecture learns to perform binary classification to identify whether an interaction ex-

ists between two proteins. STEP is characterized by three main aspects, 1) using pre-trained embeddings of protein sequences 2) employing a Siamese architecture consisting of two identical subnetworks, and 3) utilizing a bottleneck classification head for fine-tuning. More specifically, we obtained the pre-trained embeddings of protein sequences by using ProtBERT, which is a pre-trained transformer-based model that was trained on more than two billion protein amino acid sequences [55].

We first constructed a brain-specific dataset of 54,555 unique PPIs (involving 20,396 unique proteins) from PPT-OhmNet database [246]. Furthermore, a second PPIs dataset containing 334 interactions (with 338 unique proteins) between SARS-CoV-2 and human proteins was obtained from VirHostNet [245]. We enhanced the PPI datasets with sequences of proteins obtained from UniProt [138]. For both datasets, we employed a positive-unlabeled learning scheme to handle the lack of negative PPI samples. Random sampling was used to create pseudo-negatives, which were added to the constructed datasets.

The brain-specific PPI dataset was divided into single splits, consisting of training (60%), validation (20%), and test (20%) datasets. We used the training dataset to fine-tune and the validation dataset to optimize the hyperparameters of the so-called STEP-brain model. On the unseen test dataset, the best STEP-brain model achieved an area under receiver operator characteristic curve (AUC) and area under precision-recall curve (AUPR) of 88.78% and 88.32%, respectively. On the SARS-CoV-2 PPI dataset, we performed a nested cross-validation procedure using five outer and inner folds for validating the STEP-virus-host model's performance under various hyperparameters. The final generalization performance was 83.42% ($\pm$3.91%) AUC and 84.02% ($\pm$4.58%) AUPR. Finally, for both cases, we extended the test sets with a 1:10 positive to pseudo-negative samples ratio, showcasing stable performances of the model.

The STEP architecture was compared with various previously published methods on three different PPI detection datasets published by Tsukiyama *et al.* [247], Guo *et al.* [248], and Sun *et al.* [249]. We employed the original data splits and adopted the same evaluation strategy (e.g., 5-fold or 10-fold cross validation) as established by the initial authors. The results showed that STEP performed at least at par with state-of-the-art methods. Furthermore, the STEP architecture was also evaluated on two additional datasets published by Chen *et al.* [250] for the tasks — PPI type prediction and PPI binding affinity estimation, where it also achieved state-of-the-art performances.

Utilizing the proposed STEP-brain model, we predicted interactions between human brain receptors and the major capsid protein VP1 of the JCV. The top-scored interactions showed a strong enrichment of different neuro-

transmitters, including serotonin receptors, which is in line with the current literature. We also used the STEP-virus-host model to predict interactions between human receptors and spike glycoprotein of three different SARS-CoV-2 variants (such as Omicron), revealing a highly-probable interaction with sigma intracellular receptor 2, aligned with evidences from existing literature. Furthermore, we applied techniques of XAI (such as Integrated Gradients [251]) to identify the relevant locations in protein sequences that might have contributed to the prediction of their respective PPIs. These predicted interactions and the identified protein subsequences certainly require experimental validation in subsequent studies.

In conclusion, this work highlights the potential of protein sequence embeddings to build modern deep learning methods that can predict virus-host PPIs. This method could enhance drug development approaches that utilize virus-like particles for targeted delivery and support by predicting virus-host PPIs a better understanding of the underlying biological mechanisms.

Authors' contributions

To describe the contribution of each author in our publication [3], we follow the standardized roles defined in CRediT [234, 235].

Sumit Madan: Methodology, Data curation, Formal analysis, Visualization, Investigation, Validation, Writing - original draft, Writing - review & editing. **Victoria Demina**: Project administration, Writing - review & editing. **Marcus Stapf**: Project administration, Writing - review & editing. **Oliver Ernst**: Conceptualization, Project administration, Writing - review & editing. **Holger Fröhlich**: Conceptualization, Methodology, Supervision, Project administration, Writing - original draft, Writing - review & editing.

9. Dataset of miRNA–Disease Relations Extracted from Textual Data Using Transformer-Based Neural Networks

In this chapter, we summarize our publication [4] presented fully in Appendix D.

> **S. Madan**, L. Kühnel, H. Fröhlich, M. Hofmann-Apitius, and J. Fluck, "Dataset of miRNA–disease relations extracted from textual data using transformer-based neural networks," *Database*, vol. 2024, baae066, 2024. DOI: 10.1093/database/baae066

Summary

MiRNAs are short sequences of nucleotides that play important roles in post-transcriptional processes and regulate major cellular functions [57]. Their deregulation impacts cellular processes and has been linked with several human diseases such as respiratory diseases [252, 253], cancer [57], and Alzheimer's disease [254]. This makes miRNAs potential candidates for targeted therapy, even more so in the light of the advancements in miRNA delivery technologies [57]. Understanding the role of specific miRNAs in biological processes and diseases requires extensive experimental research. This information is often published in scientific literature and can be extracted using text mining techniques.

In our study [4], we proposed a deep learning-based text mining workflow to extract miRNA-disease associations from scientific literature. The worfklow uses PubMed abstracts as input and conducts a sentence-level extraction of miRNA-diseases associations. To detect such associations, the workflow focuses on three main tasks: NER for detection of miRNA and disease entities in text, NEL to link both entity types to specific database identifiers, and RE to detect associations between miRNA and diseases.

To train and evaluate the models for NER, we utilized two openly-available

training datasets for disease mention annotations and two further corpora for miRNA mentions. For the RE model, we employed distant supervision to create a new training dataset by extracting and labeling miRNA-disease associations from two databases (Human microRNA Disease Database 3 [255, 256] and miR2Disease [257]), which reduced the manual curation effort significantly. This corpus was further expanded with additional miRNA-disease relations obtained from the corpus published by Bagewadi *et al.* [258]. In total, our own RE corpus has 1,928 positive and 1,322 negative miRNA-disease associations

For modelling, we conducted experiments by fine-tuning two different LLMs, such as BioBERT [191] and BioMegatron [200]. We further implemented two different learning strategies: single-task and multi-task learning. In single-task mode, the model architecture features a single classification head to fine-tune it with a unique dataset. Conversely, the multi-task learning mode leveraged additional, related datasets to fine-tune the model with multiple classification heads using a shared representation. We used precision, recall, and F_1-measure to determine the performance of NER models. While in the case of RE models, we utilized AUC and AUPR for the binary classification of positive and negative associations. Furthermore, in all cases we applied five-fold cross-validation to choose the best models.

BioMegatron-based model [200] outperformed BioBERT [191] on all NER datasets, possibly due to its larger parameter size. The experiments with the multi-task learning mode using five related datasets were not significantly better than the single-task mode. The generalization performance for NER was evaluated on a held-out test set, with an F_1-score ranging between 86,60% and 89.71% for disease mention and an F_1-score ranging between 94.71% and 97.10% for miRNA mention detection. The best-performing RE model is a fine-tuned version of the BioMegatron model trained using multi-task learning. It achieved an AUC of 98.02% and an AUPR of 98.66% on the independent test set.

To disambiguate the various mentions of a single concept, NEL is typically performed. For instance, mentions such as AD, Alzheimer's disease, Alzheimer's dementia can be disambiguated to the MeSH identifier D000544. For this purpose, we developed a rule-based system to link miRNA entities to their respective miRBase [259] identifiers, while the publicly-available software NormCo [260] was utilized to map disease entities to MeSH identifiers.

After the training and validation process, we predicted and extracted miRNA-disease associations from around 6.1 million PubMed [123] abstracts and 1.98 million PubMed Central (PMC) [261] full-text documents published between 2020 and 2023. The extraction resulted in over 370,000 (unique: 52,000) associations, with a score above 90%. In a subsequent analysis, we

focused on three diseases: epilepsy, Alzheimer's disease, and Parkinson's disease, detecting thousands of unique miRNA–disease associations for each disease. A comparison with the DisGeNET database [137] revealed that our workflow identified a significant number of new associations that are absent in the current version of DisGeNET. We conducted a final assessment by manually verifying the validity of randomly selected top-scored associations, which underscored the high performance of the automated extraction workflow.

In conclusion, our study presented a methodology for identifying miRNA-disease associations from biomedical scientific literature using pre-trained domain-specific LLMs. The introduced text mining workflow could help to extend and update existing databases with the latest findings. Our newly-published dataset consisting of extracted miRNA-disease associations could help researchers to investigate roles of specific miRNAs in diseases. To maintain the effectiveness of the extraction workflow, it will be crucial to continuously adapt and improve such automated extraction processes with advancements in bioNLP field.

Authors' contributions

To describe the contribution of each author in our publication [4], we follow the standardized roles defined in CRediT [234, 235].

Sumit Madan: Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Investigation, Validation, Supervision, Writing - original draft, Writing - review & editing. **Lisa Kühnel**: Data curation, Formal analysis, Visualization, Investigation, Validation, Writing - original draft, Writing - review & editing. **Holger Fröhlich**: Writing - review & editing. **Martin Hofmann-Apitius**: Writing - review & editing. **Juliane Fluck**: Conceptualization, Supervision, Writing - review & editing.

Part III.

Recapitulation

10. Conclusion

In this thesis, we demonstrated the feasibility of transformer-based models for the data-driven extraction of biomedical relations. Overall, the contributions of this thesis are fourfold, 1) we offered a literature survey of existing transformer models that are specifically designed for the biomedical field, including details of their downstream tasks, 2) we mined German MSEs reports with the GermanBERT LLM to gain insights on the psychiatric attributes and psychopathological symptoms of patients, 3) we predicted virus-host PPIs using deep protein sequence embeddings from ProtBERT in a Siamese model, and 4) we extracted a dataset of miRNA-disease associations from scientific literature using a text mining workflow based on BioMegatron LLM.

For the field of psychiatry, we proposed a workflow to extract psychiatric attributes and psychopathological symptoms from MSEs (see Chapter 7). As a matter of fact, there are additional notes and datasets within psychiatric EHRs that can be leveraged to extract prescribed medications, retrieve family history of psychiatric disorders, identify severity of symptoms, detect different therapies, and predict diagnoses based on symptoms [262, 263]. Structuring and harmonizing this information across hospitals could provide comprehensive patient profiles for cohort selection and secondary use [262]. However, implementing AI-based research workflows in clinical settings is difficult due its high-risk nature. It would require a dedicated effort to streamline hospital's information technology infrastructure to enable their implementation and deployment (e.g., by introducing trusted research environments [264]).

Despite the advancements in the field of LLMs, their application in German clinical settings remains challenging. This is due to the scarcity of annotated, anonymized clinical textual data, which limits the fine-tuning and evaluation of LLMs on such data. Some efforts have concentrated on generating new data, however, large-scale datasets covering different branches of medicine remain unaccessible [265]. Moreover, generative LLMs present many potential risks (such as confabulations, bias, and privacy issues), which needs to be addressed by the medical informatics community [266].

Transfer learning approaches are well-established also in protein sequence analysis. In our work, in Chapter 8, we predicted tissue-specific virus-host PPIs using a ProtBERT-based Siamese model, which facilitated the prioriti-

zation of virus-host PPIs for further investigation (such as therapeutic development). Advancing from a simple binary prediction of PPI to identify interaction types and estimate the binding affinity of the interactions would enable a granular understanding of the PPIs. Moreover, the XAI methods have the potential of explaining how the PPIs are grounded by offering deeper insights on relevant binding regions. However, validating XAI predictions is hard due to the missing ground truth. Further commitments are necessary to validate these findings. Moreover, a full parameter fine-tuning with ProtBERT alongside with hyperparameter optimization for the PPI prediction model required significant computational resources. In practical settings, however, parameter-efficient delta-tuning techniques, which includes freezing parts of the pre-trained model, could be considered to train more efficiently [267]. Nonetheless, the incorporation of protein structure embeddings derived from AlphaFold [49, 50, 156] has the potential to further improve the PPI prediction performance significantly.

In Chapter 9, we inferred a dataset of miRNA-disease associations from PubMed using LLMs, which enables a further investigation into the roles of miRNAs in various diseases. While the dataset includes new associations not contained in current DisGeNET [137], contextualizing them through the identification of mRNA target, tissue types, cell types, organisms, and used detection methods would benefit further investigation. Given that such contextual information is commonly included in the body of scientific articles, a comprehensive analysis of full-text is necessary. Despite advancements, fully automating the retrieval of such complex relations from text remains and unsolved challenge [213]. Besides the growing trend of applying zero or few-shot learning with generative AI models [205, 213], there is another approach to apply multi-task learning by training joint models that integrate NER, NEL, and RE tasks to minimize the error propagation seen in sequential multi-step workflows [268]. However, these persistent challenges of biomedical text mining (such as context identification, full-text analysis, and development of general purpose models) will require significant effort in creating an extensive and cohesive labeled dataset derived from full-text documents. Active learning strategies can be employed to efficiently scan the literature and generate a robust dataset [269].

11. Future Outlook

There are promising future research avenues within the development of transformer-based models on various biomedical data modalities for extracting relations.

Particularly, the detection of biomedical relational knowledge across biomedical fields can be significantly enhanced by integrating multiple modalities. The transformer-based models also have the capacity to advance the landscape of multimodal data analysis [270, 271] (see also Chapter 6). To learn across modalities, some have proposed general-purpose transformer-based architectures such as PercieverIO [272, 273]. However, realizing this potential requires a collaborative effort by researchers and industries to collect linked and cohesive multimodal datasets [114].

Given the significant advancements in NLP, it is worthwhile to explore LBD, which aims to generate new hypotheses leveraging knowledge from literature. The integration of generative LLMs in the LBD process has the potential to transform the generation of credible hypotheses [274–276]. Assuming that the challenges (such as confabulations) of generative models are effectively addressed, they can not only contextualize the existing relations but also incorporate new paths in these hypotheses.

Routinely-collected, structured, and narrative EHR data can reveal important clinical relations. However, most of this data is inaccessible due to privacy reasons. Federated learning enables the analysis of such data without transferring it to a potential unsafe location [277]. Although some have demonstrated the utilization of pre-trained transformer-based models to learn from federated data [278], there is a lack of experimental research in the context of EHR [277]. Nevertheless, federated learning with transformer-based models has the potential to advance the analysis of retrospective EHR data and can deliver new insights into disease risk factors, procedure outcomes, and medication efficacies while adhering to privacy laws [279, 280].

These efforts lead to models that can potentially be served as deployable solutions in biomedical applications and healthcare systems. However, these models are not perfect and their success in making predictions and providing explanations varies widely. Therefore, establishing these models is still challenging [47, 281, 282]. Nevertheless, there are many examples of AI-based

solutions approved by the US Foods and Drugs Administration (FDA) for the healthcare system [283]. In Mayo Clinic's 2030 vision, the authors anticipate leveraging AI solutions across various healthcare branches to "reduce cognitive burden for clinicians" [284]. This indicates that AI is already revolutionizing the biomedical and healthcare sectors and further breakthroughs are on the horizon.

Bibliography

[1] S. Madan, M. Lentzen, J. Brandt, D. Rueckert, M. Hofmann-Apitius, and H. Fröhlich, "Transformer models in biomedicine," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 214, 2024. DOI: 10.1186/s12911-024-02600-5.

[2] S. Madan *et al.*, "Deep Learning-based Detection of Psychiatric Attributes from German Mental Health Records," *International Journal of Medical Informatics*, vol. 161, p. 104 724, 2022. DOI: 10.1016/j.ijmedinf.2022.104724.

[3] S. Madan, V. Demina, M. Stapf, O. Ernst, and H. Fröhlich, "Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings," *Patterns*, vol. 3, no. 9, p. 100 551, 2022. DOI: 10.1016/j.patter.2022.100551.

[4] S. Madan, L. Kühnel, H. Fröhlich, M. Hofmann-Apitius, and J. Fluck, "Dataset of miRNA–disease relations extracted from textual data using transformer-based neural networks," *Database*, vol. 2024, baae066, 2024. DOI: 10.1093/database/baae066.

[5] S. Krix *et al.*, "MultiGML: Multimodal graph machine learning for prediction of adverse drug events," *Heliyon*, vol. 9, no. 9, e19441, 2023. DOI: 10.1016/j.heliyon.2023.e19441.

[6] M. Lentzen *et al.*, "A Transformer-Based Model Trained on Large Scale Claims Data for Prediction of Severe COVID-19 Disease Progression," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 9, pp. 4548–4558, 2023. DOI: 10.1109/JBHI.2023.3288768.

[7] R.-U. Rahman *et al.*, "SEAweb: The small RNA Expression Atlas web application," *Nucleic Acids Research*, p. 16, 2019. DOI: 10.1093/nar/gkz869.

[8] A. Sargsyan *et al.*, "The Epilepsy Ontology: A community-based ontology tailored for semantic interoperability and text mining," *Bioinformatics Advances*, vol. 3, no. 1, vbad033, 2023. DOI: 10.1093/bioadv/vbad033.

[9] A. Sargsyan *et al.*, "The COVID-19 Ontology," *Bioinformatics*, vol. 36, no. 24, pp. 5703–5705, 2020. DOI: 10.1093/bioinformatics/btaa1057.

[10] J. Botz *et al.*, "Modeling approaches for early warning and monitoring of pandemic situations as well as decision support," *Frontiers in Public Health*, vol. 10, 2022. DOI: 10.3389/fpubh.2022.994949.

[11] P. Wegner, H. Fröhlich, and S. Madan, "Evaluating Knowledge Fusion Models on Detecting Adverse Drug Events in Text," *PLOS Digital Health*, 2025. DOI: 10.1371/journal.pdig.0000468.

[12] N. S. Babaiha *et al.*, "A natural language processing system for the efficient updating of highly curated pathophysiology mechanism knowledge graphs," *Artificial Intelligence in the Life Sciences*, vol. 4, p. 100 078, 2023. DOI: 10.1016/j.ailsci.2023.100078.

[13] M. Lentzen *et al.*, "Critical assessment of transformer-based AI models for German clinical notes," *JAMIA Open*, vol. 5, no. 4, ooac087, 2022. DOI: 10.1093/jamiaopen/ooac087.

[14] L. Langnickel *et al.*, "Information extraction from german clinical care documents in context of alzheimer's disease," *Applied Sciences*, vol. 11, no. 22, 2021. DOI: 10.3390/app112210717.

[15] R. Karki, S. Madan, Y. Gadiya, D. Domingo-Fernández, A. T. Kodamullil, and M. Hofmann-Apitius, "Data-Driven Modeling of Knowledge Assemblies in Understanding Comorbidity Between Type 2 Diabetes Mellitus and Alzheimer's Disease," *Journal of Alzheimer's Disease*, vol. Preprint, pp. 1–9, Preprint 2020. DOI: 10.3233/JAD-200752.

[16] R. Li *et al.*, "Systems Toxicology Approach for Testing Chemical Cardiotoxicity in Larval Zebrafish," *Chemical Research in Toxicology*, 2020. DOI: 10.1021/acs.chemrestox.0c00095.

[17] S. Madan *et al.*, "The extraction of complex relationships and their conversion to biological expression language (BEL) overview of the BioCreative VI (2017) BEL track," *Database*, vol. 2019, 2019. DOI: 10.1093/database/baz084.

[18] R. Li *et al.*, "Systems toxicology approach for the assessment of zebrafish cardiotoxicity," *Toxicology Letters*, vol. 295, S102, 2018. DOI: 10.1016/j.toxlet.2018.06.611.

[19] A. T. Kodamullil, A. Iyappan, R. Karki, S. Madan, E. Younesi, and M. Hofmann-Apitius, "Of Mice and Men: Comparative Analysis of Neuro-Inflammatory Mechanisms in Human and Mouse Using Cause-and-Effect Models," *Journal of Alzheimer's Disease*, vol. 59, no. 3, pp. 1045–1055, 2017. DOI: 10.3233/JAD-170255. pmid: 28731442.

[20] J. Fluck *et al.*, "Training and evaluation corpora for the extraction of causal relationships encoded in biological expression language (BEL)," *Database : the journal of biological databases and curation*, vol. 2016, 2016. DOI: 10.1093/database/baw113. pmid: 27554092.

[21] F. Rinaldi *et al.*, "BioCreative V track 4: A shared task for the extraction of causal network information using the Biological Expression Language," *Database : the journal of biological databases and curation*, vol. 2016, 2016. DOI: 10.1093/database/baw067. pmid: 27402677.

[22] Q. Wang *et al.*, "Overview of the interactive task in BioCreative V," *Database*, vol. 2016, baw119, 2016. DOI: 10.1093/database/baw119.

[23] S. Madan *et al.*, "The BEL information extraction workflow (BELIEF): Evaluation in the BioCreative V BEL and IAT track," *Database*, vol. 2016, baw136, 2016. DOI: 10.1093/database/baw136. pmid: 27694210.

[24] J. Szostak *et al.*, "Construction of biological networks from unstructured information based on a semi-automated curation workflow," *Database*, vol. 2015, 2015. DOI: 10.1093/database/bav057. pmid: 26200752.

[25] L. Langnickel, R. Baum, J. Darms, S. Madan, and J. Fluck, "COVID-19 preVIEW: Semantic Search to Explore COVID-19 Research Preprints," *Studies in Health Technology and Informatics*, vol. 281, pp. 78–82, 2021. DOI: 10.3233/SHTI210124. pmid: 34042709.

[26] J. Dörpinghaus, J. Klein, J. Darms, S. Madan, and M. Jacobs, "SCAIView – A Semantic Search Engine for Biomedical Research Utilizing a Microservice Architecture," in *Proceedings of the Posters and Demos Track of the 14th International Conference on Semantic Systems - SEMANTiCS2018*, 2018.

[27] A. Y. Lin *et al.*, "CTO: A Community-Based Clinical Trial Ontology and its Applications in PubChemRDF and SCAIView," in *Proceedings of the 11th International Conference on Biomedical Ontologies (ICBO)*, 2020.

[28] S. Mora, S. Madan, S. Gebel, and e. al, "Proposal of an Architecture for Terminology Management in a Research Project," in *Digital Personalized Health and Medicine*, 2020, pp. 1371–1372.

[29] L. Langnickel, R. Baum, G. Wollnik-Korn, B. Fischer-Wagener, S. Madan, and J. Fluck, "The future of German MeSH: A new semi-automatic translation process and new services for search and annotation," presented at the GMDS Conference 2020, 2020.

[30] S. Madan, M. Fiosins, S. Bonn, and J. Fluck, "A semantic data integration methodology for translational neurodegenerative disease research," in *Proceedings of the 11th International Conference Semantic Web Applications and Tools for Life Sciences*, ser. CEUR Workshop Proceedings, vol. 2275, Antwerp, Belgium: CEUR, 2018.

[31] M. Ali, S. Madan, A. Fischer, H. Petzka, and J. Fluck, "Automatic Extraction of BEL-Statements based on Neural Networks Automatic Extraction of BEL-Statements based on Neural Networks," in *Proceedings of BioCreative VI Challenge and Workshop*, 2017, pp. 2013–2015.

[32] S. Madan, J. Szostak, J. Dörpinghaus, J. Hoeng, and J. Fluck, "Overview of BEL Track: Extraction of Complex Relationships and their Conversion to BEL," in *Proceedings of BioCreative VI Challenge and Workshop*, 2017.

[33] J. Szostak *et al.*, "Recent improvements of the BEL Information Extraction workFlow (BELIEF) for biomedical text mining and curation," in *10th International Biocuration Conference 2017*, vol. 6, 10th International Biocuration Conference 2017, 2017. DOI: 10.7490/F1000RESEARCH.1113812.1.

[34] J. Szostak *et al.*, "BELIEF: A semi-automated curation tool to build mechanistic causal biological knowledgebase from unstructured scientific information," in *Systems Toxicology 2016 Conference - Real World Applications and Opportunities*, Les Diablerets, Switzerland, 2016.

[35] J. Szostak *et al.*, "A computational network model describing xenobiotic metabolism response in the liver built using the semi-automated curation workflow BELIEF.," 2016.

[36] S. Madan, S. Hodapp, and J. Fluck, "BELIEF Dashboard – a Web-based Curation Interface to Support Generation of BEL Networks," in *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, Sevilla, Spain, 2015, pp. 409–417.

[37] J. Fluck *et al.*, "Track 4 Overview: Extraction of Causal Network Information in Biological Expression Language (BEL)," in *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, vol. 1, Sevilla, Spain, 2015, pp. 333–346.

[38] J. Fluck *et al.*, "BELIEF - A semiautomatic workflow for BEL network creation," in *Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM)*, 2014, pp. 109–113. DOI: 10.5167/uzh-98982.

[39] J. Fluck *et al.*, "BEL networks derived from qualitative translations of BioNLP Shared Task annotations," in *Workshop on Biomedical Natural Language Processing, BioNLP 2013*, Sofia, Bulgaria: Association for Computational Linguistics (ACL), 2013, pp. 80–88.

[40] R. Klinger, P. Senger, S. Madan, and M. Jacovi, "Online Communities Support Policy-Making: The Need for Data Analysis," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7444 LNCS, 2012, pp. 132–143. DOI: 10.1007/978-3-642-33250-0_12.

[41] F. F. Costa, "Big data in biomedicine," *Drug Discovery Today*, vol. 19, no. 4, pp. 433–440, 2014. DOI: 10.1016/j.drudis.2013.10.012.

[42] M. K. Ross, W. Wei, and L. Ohno-Machado, ""Big Data" and the Electronic Health Record," *Yearbook of Medical Informatics*, vol. 23, no. 01, pp. 97–104, 2014. DOI: 10.15265/IY-2014-0003.

[43] C. J. Cremin, S. Dash, and X. Huang, "Big data: Historic advances and emerging trends in biomedical research," *Current Research in Biotechnology*, vol. 4, pp. 138–151, 2022. DOI: 10.1016/j.crbiot.2022.02.004.

[44] D. Howe *et al.*, "Big data: The future of biocuration," *Nature*, vol. 455, no. 7209, pp. 47–50, 2008. DOI: 10.1038/455047a. pmid: 18769432.

[45] T. Hulsen *et al.*, "From Big Data to Precision Medicine," *Frontiers in Medicine*, vol. 6, p. 34, 2019. DOI: 10.3389/fmed.2019.00034. pmid: 30881956.

[46] H. Fröhlich *et al.*, "From hype to reality: Data science enabling personalized medicine," *BMC Medicine*, vol. 16, no. 1, p. 150, 2018. DOI: 10.1186/s12916-018-1122-7.

[47] M. Wainberg, D. Merico, A. Delong, and B. J. Frey, "Deep learning in biomedicine," *Nature Biotechnology*, vol. 36, no. 9, pp. 829–838, 2018. DOI: 10.1038/nbt.4233.

[48] C. V. Theodoris *et al.*, "Transfer learning enables predictions in network biology," *Nature*, vol. 618, no. 7965, pp. 616–624, 7965 2023. DOI: 10.1038/s41586-023-06139-9.

[49] A. W. Senior *et al.*, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, 7792 2020. DOI: 10.1038/s41586-019-1923-7.

[50] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 7873 2021. DOI: 10.1038/s41586-021-03819-2.

[51] M. Groh *et al.*, "Deep learning-aided decision support for diagnosis of skin disease across skin tones," *Nature Medicine*, vol. 30, no. 2, pp. 573–583, 2024. DOI: 10.1038/s41591-023-02728-3.

[52] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," *npj Digital Medicine*, vol. 5, no. 1, pp. 1–8, 2022. DOI: 10.1038/s41746-022-00592-y.

[53] P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Scientific Data*, vol. 10, no. 1, p. 67, 1 2023. DOI: 10.1038/s41597-023-01960-3.

[54] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010. DOI: 10.5555/3295222.3295349.

[55] A. Elnaggar *et al.*, "ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2021. DOI: 10.1109/tpami.2021.3095381. pmid: 34232869.

[56] A. Rives *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, 2021. DOI: 10.1073/pnas.2016239118.

[57] R. Rupaimoole and F. J. Slack, "MicroRNA therapeutics: Towards a new era for the management of cancer and other diseases," *Nature Reviews Drug Discovery*, vol. 16, no. 3, pp. 203–222, 3 2017. DOI: 10.1038/nrd.2016.246.

[58] D. Riaño, M. Peleg, and A. ten Teije, "Ten years of knowledge representation for health care (2009–2018): Topics, trends, and challenges," *Artificial Intelligence in Medicine*, vol. 100, p. 101 713, 2019. DOI: 10.1016/j.artmed.2019.101713.

[59] R. Brachman and H. Levesque, *Knowledge Representation and Reasoning*, 1 edition. Amsterdam ; Boston: Morgan Kaufmann, 2004, 381 pp.

[60] D. J. Rigden and X. M. Fernández, "The 2023 Nucleic Acids Research Database Issue and the online molecular biology database collection," *Nucleic Acids Research*, vol. 51, no. D1, pp. D1–D8, 2023. DOI: 10.1093/nar/gkac1186.

[61] D. J. Rigden and X. M. Fernández, "The 2024 Nucleic Acids Research database issue and the online molecular biology database collection," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1–D9, 2024. DOI: 10.1093/nar/gkad1173. pmid: 38035367.

[62] G. Skoufos *et al.*, "TarBase-v9.0 extends experimentally supported miRNA–gene interactions to cell-types and virally encoded miRNAs," *Nucleic Acids Research*, vol. 52, no. D1, pp. D304–D310, 2024. DOI: 10.1093/nar/gkad1071.

[63] C. Zhang, X. Zhang, P. L. Freddolino, and Y. Zhang, "BioLiP2: An updated structure database for biologically relevant ligand–protein interactions," *Nucleic Acids Research*, vol. 52, no. D1, pp. D404–D412, 2024. DOI: 10.1093/nar/gkad630.

[64] M. Cannon *et al.*, "DGIdb 5.0: Rebuilding the drug–gene interaction database for precision medicine and drug discovery platforms," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1227–D1235, 2024. DOI: 10.1093/nar/gkad1040.

[65] B. Zdrazil *et al.*, "The ChEMBL Database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1180–D1192, 2024. DOI: 10.1093/nar/gkad1004.

[66] C. Knox *et al.*, "DrugBank 6.0: The DrugBank Knowledgebase for 2024," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1265–D1275, 2024. DOI: 10.1093/nar/gkad976.

[67] J. Tang *et al.*, "Drug Target Commons: A Community Effort to Build a Consensus Knowledge Base for Drug-Target Interactions," *Cell Chemical Biology*, vol. 25, no. 2, 224–229.e2, 2018. DOI: 10.1016/j.chembiol.2017.11.009. pmid: 29276046.

[68] M. Whirl-Carrillo *et al.*, "An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine," *Clinical Pharmacology and Therapeutics*, vol. 110, no. 3, pp. 563–572, 2021. DOI: 10.1002/cpt.2350. pmid: 34216021.

[69] L. Ma *et al.*, "Database Commons: A Catalog of Worldwide Biological Databases," *Genomics, Proteomics & Bioinformatics*, vol. 21, no. 5, pp. 1054–1058, 2023. DOI: 10.1016/j.gpb.2022.12.004.

[70] R. Stevens, C. A. Goble, and S. Bechhofer, "Ontology-based knowledge representation for bioinformatics," *Briefings in Bioinformatics*, vol. 1, no. 4, pp. 398–414, 2000. DOI: 10.1093/bib/1.4.398.

[71] J. Hastings *et al.*, "The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013," *Nucleic Acids Research*, vol. 41, no. D1, pp. D456–63, 2013. DOI: 10.1093/nar/gks1146. pmid: 23180789.

[72] Gene Ontology Consortium, "The Gene Ontology (GO) database and informatics resource," *Nucleic acids research*, vol. 32, pp. D258–D261, suppl_1 2004.

[73] J. A. Blake *et al.*, "Gene ontology consortium: Going forward," *Nucleic Acids Research*, vol. 43, no. D1, pp. D1049–D1056, 2015. DOI: 10.1093/nar/gku1179. pmid: 25428369.

[74] K. Mabon, O. Steinum, and C. G. Chute, "Postcoordination of codes in ICD-11," *BMC Medical Informatics and Decision Making*, vol. 21, no. 6, p. 379, 2022. DOI: 10.1186/s12911-022-01876-9.

[75] Y. Zhang *et al.* "BioKG: A comprehensive, high-quality biomedical knowledge graph for AI-powered, data-driven biomedical research." (2023), [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.10.13.562216v2, pre-published.

[76] D. Domingo-Fernández *et al.*, "COVID-19 Knowledge Graph: A computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology," *Bioinformatics (Oxford, England)*, vol. 37, no. 9, pp. 1332–1334, 2021. DOI: 10.1093/bioinformatics/btaa834. pmid: 32976572.

[77] I. Rodchenkov *et al.*, "Pathway Commons 2019 Update: Integration, analysis and exploration of pathway data," *Nucleic Acids Research*, vol. 48, no. D1, pp. D489–D497, 2020. DOI: 10.1093/nar/gkz946.

[78] A. Agrawal *et al.*, "WikiPathways 2024: Next generation pathway database," *Nucleic Acids Research*, vol. 52, no. D1, pp. D679–D689, 2024. DOI: 10.1093/nar/gkad960.

[79] M. Milacic *et al.*, "The Reactome Pathway Knowledgebase 2024," *Nucleic Acids Research*, vol. 52, no. D1, pp. D672–D678, 2024. DOI: 10.1093/nar/gkad1025.

[80] X. Wang *et al.*, "Knowledge graph quality control: A survey," *Fundamental Research*, vol. 1, no. 5, pp. 607–626, 2021. DOI: 10.1016/j.fmre.2021.09.003.

[81] R. Hoehndorf and N. Queralt-Rosinach, "Data Science and symbolic AI: Synergies, challenges and opportunities," *Data Science*, vol. 1, no. 1–2, pp. 27–38, 2017. DOI: 10.3233/DS-170004.

[82] Y. Bengio, A. Courville, and P. Vincent. "Representation Learning: A Review and New Perspectives." arXiv: 1206.5538 [cs]. (2014), [Online]. Available: http://arxiv.org/abs/1206.5538, pre-published.

[83] X.-M. Zhang, L. Liang, L. Liu, and M.-J. Tang, "Graph Neural Networks and Their Current Applications in Bioinformatics," *Frontiers in Genetics*, vol. 12, p. 690 049, 2021. DOI: 10.3389/fgene.2021.690049. pmid: 34394185.

[84] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940, 8 2023. DOI: 10.1038/s41591-023-02448-8.

[85] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–13, 2021.

[86] D. A. Lee, "Knowledge Representation and Reasoning in AI: Analyzing Different Approaches to Knowledge Representation and Reasoning in Artificial Intelligence Systems," *Journal of Artificial Intelligence Research*, vol. 4, no. 1, pp. 14–29, 1 2024.

[87] W. Saeed and C. Omlin. "Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities." arXiv: 2111.06420 [cs]. (2021), [Online]. Available: http://arxiv.org/abs/2111.06420, pre-published.

[88] P. S. Reel, S. Reel, E. Pearson, E. Trucco, and E. Jefferson, "Using machine learning approaches for multi-omics data analysis: A review," *Biotechnology Advances*, vol. 49, p. 107 739, 2021. DOI: 10.1016/j.biotechadv.2021.107739.

[89] S. Vineetha, C. C. S. Bhat, and S. M. Idicula, "MicroRNA–mRNA interaction network using TSK-type recurrent neural fuzzy network," *Gene*, vol. 515, no. 2, pp. 385–390, 2013. DOI: 10.1016/j.gene.2012.12.063.

[90] S. Orchard *et al.*, "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic acids research*, vol. 42, no. D1, pp. D358–D363, 2014.

[91] J. Hippisley-Cox, D. Stables, and M. Pringle, "QRESEARCH: A new general practice database for research," *Informatics in primary care*, vol. 12, no. 1, pp. 49–50, 2004.

[92] E. Herrett *et al.*, "Data resource profile: Clinical practice research datalink (CPRD)," *International journal of epidemiology*, vol. 44, no. 3, pp. 827–836, 2015.

[93] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012. DOI: 10.1038/nrg3208.

[94] M. J. Sheller *et al.*, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, no. 1, p. 12 598, 2020. DOI: 10.1038/s41598-020-69250-1.

[95] German National Cohort (GNC) Consortium, "The German National Cohort: Aims, study design and organization," *European Journal of Epidemiology*, vol. 29, no. 5, pp. 371–382, 2014. DOI: 10.1007/s10654-014-9890-7.

[96] R. Arunkumar and P. Karthigaikumar, "Multi-retinal disease classification by reduced deep learning features," *Neural Computing and Applications*, vol. 28, no. 2, pp. 329–334, 2017. DOI: 10.1007/s00521-015-2059-9.

[97] T. Schlegl, S. M. Waldstein, W.-D. Vogl, U. Schmidt-Erfurth, and G. Langs, "Predicting Semantic Descriptions from Medical Images with Convolutional Neural Networks," in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, and M. J. Cardoso, Eds., vol. 9123, Cham: Springer International Publishing, 2015, pp. 437–448. DOI: 10.1007/978-3-319-19992-4_34.

[98] Y. Jiang, M. Yang, S. Wang, X. Li, and Y. Sun, "Emerging role of deep learning-based artificial intelligence in tumor pathology," *Cancer Communications*, vol. 40, no. 4, pp. 154–166, 2020. DOI: 10.1002/cac2.12012.

[99] D. Domingo-Fernández *et al.*, "Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): A web server for mechanism enrichment," *Bioinformatics*, vol. 33, no. 22, pp. 3679–3681, 2017. DOI: 10.1093/bioinformatics/btx399. pmid: 28651363.

[100] K. Kandasamy *et al.*, "NetPath: A public resource of curated signal transduction pathways," *Genome Biology*, vol. 11, no. 1, R3, 2010. DOI: 10.1186/gb-2010-11-1-r3. pmid: 20067622.

[101] M. H. Al-Rabeah and A. Lakizadeh, "Prediction of drug-drug interaction events using graph neural networks based feature extraction," *Scientific Reports*, vol. 12, no. 1, p. 15 590, 2022. DOI: 10.1038/s41598-022-19999-4.

[102] K. Jha, S. Saha, and H. Singh, "Prediction of protein–protein interaction using graph neural networks," *Scientific Reports*, vol. 12, no. 1, p. 8360, 1 2022. DOI: 10.1038/s41598-022-12201-9.

[103] Ö. Muslu, C. T. Hoyt, M. Lacerda, M. Hofmann-Apitius, and H. Fröhlich, "GuiltyTargets: Prioritization of Novel Therapeutic Targets With Network Representation Learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 1, pp. 491–500, 2022. DOI: 10.1109/TCBB.2020.3003830.

[104] K. Hsieh *et al.*, "Drug repurposing for COVID-19 using graph neural network and harmonizing multiple evidence," *Scientific Reports*, vol. 11, no. 1, p. 23 179, 1 2021. DOI: 10.1038/s41598-021-02353-5.

[105] L. Zhu and H. Zheng, "Biomedical event extraction with a novel combination strategy based on hybrid deep neural networks," *BMC Bioinformatics*, vol. 21, no. 1, p. 47, 2020. DOI: 10.1186/s12859-020-3376-2.

[106] Z. Li, Y. Lian, X. Ma, X. Zhang, and C. Li, "Bio-semantic relation extraction with attention-based external knowledge reinforcement," *BMC Bioinformatics*, vol. 21, no. 1, p. 213, 2020. DOI: 10.1186/s12859-020-3540-8.

[107] L. Hong *et al.*, "A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 347–355, 6 2020. DOI: 10.1038/s42256-020-0189-y.

[108] B. Dandala, M. Diwakar, and D. Murthy, "IBM Research System at TAC 2017: Adverse Drug Reactions Extraction from Drug Labels," *Text Analysis Conference (TAC2017)*, 2017.

[109] M. Jantscher, F. Gunzer, R. Kern, E. Hassler, S. Tschauner, and G. Reishofer, "Information extraction from German radiological reports for general clinical text and language understanding," *Scientific Reports*, vol. 13, no. 1, p. 2353, 2023. DOI: 10.1038/s41598-023-29323-3.

[110] S. Sohn *et al.*, "Comprehensive temporal information detection from clinical text: Medical events, time, and TLINK identification," *Journal of the American Medical Informatics Association : JAMIA*, vol. 20, no. 5, pp. 836–842, 2013. DOI: 10.1136/amiajnl-2013-001622. pmid: 23558168.

[111] B. M. Gyori, J. A. Bachman, K. Subramanian, J. L. Muhlich, L. Galescu, and P. K. Sorger, "From word models to executable models of signaling networks using automated assembly," *Molecular Systems Biology*, vol. 13, no. 11, p. 954, 2017. DOI: 10.15252/msb.20177651. pmid: 29175850.

[112] J. A. Bachman, B. M. Gyori, and P. K. Sorger, "Automated assembly of molecular mechanisms at scale from text mining and curated databases," *Molecular Systems Biology*, vol. 19, no. 5, e11325, 2023. DOI: 10.15252/msb.202211325.

[113] L. Breuza *et al.*, "A Coordinated Approach by Public Domain Bioinformatics Resources to Aid the Fight Against Alzheimer's Disease Through Expert Curation of Key Protein Targets," *Journal of Alzheimer's Disease*, vol. 77, no. 1, pp. 257–273, 2020. DOI: 10.3233/JAD-200206. pmid: 32716361.

[114] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical AI," *Nature Medicine*, vol. 28, no. 9, pp. 1773–1784, 9 2022. DOI: 10.1038/s41591-022-01981-2.

[115] A. Sakhovskiy and E. Tutubalina, "Multimodal model with text and drug embeddings for adverse drug reaction classification," *Journal of Biomedical Informatics*, vol. 135, p. 104 182, 2022. DOI: 10.1016/j.jbi.2022.104182.

[116] S. Liu *et al.*, "Multimodal data matters: Language model pre-training over structured and unstructured electronic health records," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 504–514, 2022. DOI: 10.1109/jbhi.2022.3217810.

[117] S. Jabbour, D. Fouhey, E. Kazerooni, J. Wiens, and M. W. Sjoding, "Combining chest X-rays and electronic health record (EHR) data using machine learning to diagnose acute respiratory failure," *Journal of the American Medical Informatics Association*, vol. 29, no. 6, pp. 1060–1068, 2022. DOI: 10.1093/jamia/ocac030.

[118] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[119] S. Rajendran, W. Pan, M. R. Sabuncu, Y. Chen, J. Zhou, and F. Wang, "Learning across diverse biomedical data modalities and cohorts: Challenges and opportunities for innovation," *Patterns*, 2024.

[120] J. Fluck and M. Hofmann-Apitius, "Text mining for systems biology," *Drug Discovery Today*, vol. 19, no. 2, pp. 140–144, 2014. DOI: 10.1016/j.drudis.2013.09.012. pmid: 24070668.

[121] S. Zhao, C. Su, Z. Lu, and F. Wang, "Recent advances in biomedical literature mining," *Briefings in Bioinformatics*, vol. 22, no. 3, bbaa057, 2021. DOI: 10.1093/bib/bbaa057.

[122] National Library of Medicine, *Medline overview*, https://www.nlm.nih.gov/medline/medline_overview.html, Last accessed on 2025-01-12.

[123] National Library of Medicine, *Pubmed overview*, https://pubmed.ncbi.nlm.nih.gov/about/, Last accessed on 2025-01-12.

[124] National Library of Medicine, "MEDLINE 2022 Initiative: Transition to Automated Indexing," *NLM Tech Bull.*, no. 443, 2021.

[125] A. Krithara, J. G. Mork, A. Nentidis, and G. Paliouras, "The road from manual to automatic semantic indexing of biomedical literature: A 10 years journey," *Frontiers in Research Metrics and Analytics*, vol. 8, p. 1 250 930, 2023. DOI: 10.3389/frma.2023.1250930. pmid: 37841902.

[126] H. T. Tabib *et al.*, "Interactive Extractive Search over Biomedical Corpora," in *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, Eds., Online: Association for Computational Linguistics, 2020, pp. 28–37. DOI: 10.18653/v1/2020.bionlp-1.3.

[127] S. Rosonovski *et al.*, "Europe PMC in 2023," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1668–D1676, 2024. DOI: 10.1093/nar/gkad1085.

[128] G. D. Guardia, R. Z. Vêncio, and C. R. de Farias, "A UML profile for the OBO relation ontology," *BMC Genomics*, vol. 13, no. 5, S3, 2012. DOI: 10.1186/1471-2164-13-S5-S3.

[129] C. Mungall *et al.*, *Oborel/obo-relations: 2023-08-18 Release*, version v2023-08-18, Zenodo, 2023. DOI: 10.5281/zenodo.8263469.

[130] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia, "Overview of the protein-protein interaction annotation extraction task of BioCreative II," *Genome Biology*, vol. 9, S4, SUPPL. 2 2008. DOI: 10.1186/gb-2008-9-s2-s4. pmid: 18834495.

[131] I. Segura-Bedmar *et al.*, "SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, vol. 2, 2013, pp. 341–350. DOI: 10.1.1.310.783.

[132] W. Sun, A. Rumshisky, and O. Uzuner, "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge," *Journal of the American Medical Informatics Association: JAMIA*, vol. 20, no. 5, pp. 806–813, 2013. DOI: 10.1136/amiajnl-2013-001628. pmid: 23564629.

[133] U. o. P. Department of Information Engineering, *Gutbrainie clef 2025*, https://hereditary.dei.unipd.it/challenges/gutbrainie/2025, Last accessed on 2025-02-04.

[134] T. B. Challenge, *Bioasq - task gutbrainie: Gut-brain interplay information extraction*, https://participants-area.bioasq.org/general_information/GutBrainIE/, Last accessed on 2025-02-04.

[135] D. R. Swanson, "Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge," *Perspectives in Biology and Medicine*, vol. 30, no. 1, pp. 7–18, 1986. DOI: 10.1353/pbm.1986.0087.

[136] E. Moreau, "Literature-based discovery: Addressing the issue of the subpar evaluation methodology," *Bioinformatics*, vol. 39, no. 2, btad090, 2023. DOI: 10.1093/bioinformatics/btad090.

[137] J. Piñero *et al.*, "The DisGeNET knowledge platform for disease genomics: 2019 update," *Nucleic acids research*, vol. 48, no. D1, pp. D845–D855, 2020.

[138] The UniProt Consortium, "UniProt: The Universal Protein Knowledgebase in 2023," *Nucleic Acids Research*, vol. 51, no. D1, pp. D523–D531, 2023. DOI: 10.1093/nar/gkac1052.

[139] H. Kilicoglu, "Biomedical text mining for research rigor and integrity: Tasks, challenges, directions," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1400–1414, 2018.

[140] D. Ofer, N. Brandes, and M. Linial, "The language of proteins: NLP, machine learning & protein sequences," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 1750–1758, 2021. DOI: 10.1016/j.csbj.2021.03.022.

[141] H. Iuchi *et al.*, "Representation learning applications in biological sequence analysis," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 3198–3208, 2021. DOI: 10.1016/j.csbj.2021.05.039.

[142] M. Kulmanov and R. Hoehndorf, "DeepGOPlus: Improved protein function prediction from sequence," *Bioinformatics*, vol. 36, no. 2, pp. 422–429, 2020. DOI: 10.1093/bioinformatics/btz595.

[143] A. Chang *et al.*, "BRENDA, the ELIXIR core data resource in 2021: New developments and updates," *Nucleic Acids Research*, vol. 49, no. D1, pp. D498–D508, 2021. DOI: 10.1093/nar/gkaa1025.

[144] C. Chen *et al.*, "MoonProt 3.0: An update of the moonlighting proteins database," *Nucleic Acids Research*, vol. 49, no. D1, pp. D368–D372, 2021. DOI: 10.1093/nar/gkaa1101.

[145] C. J. Jeffery, "Current successes and remaining challenges in protein function prediction," *Frontiers in Bioinformatics*, vol. 3, p. 1 222 182, 2023. DOI: 10.3389/fbinf.2023.1222182. pmid: 37576715.

[146] F. Soleymani, E. Paquet, H. Viktor, W. Michalowski, and D. Spinello, "Protein–protein interaction prediction with deep learning: A comprehensive review," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 5316–5341, 2022. DOI: 10.1016/j.csbj.2022.08.070. pmid: 36212542.

[147] J. Wei, S. Chen, L. Zong, X. Gao, and Y. Li, "Protein–RNA interaction prediction with deep learning: Structure matters," *Briefings in Bioinformatics*, vol. 23, no. 1, bbab540, 2022. DOI: 10.1093/bib/bbab540.

[148] M. W. Gonzalez and M. G. Kann, "Chapter 4: Protein Interactions and Disease," *PLoS Computational Biology*, vol. 8, no. 12, e1002819, 2012. DOI: 10.1371/journal.pcbi.1002819. pmid: 23300410.

[149] R. Oughtred *et al.*, "The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions," *Protein Science*, vol. 30, no. 1, pp. 187–200, 2021.

[150] D. Szklarczyk *et al.*, "STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research*, vol. 47, no. D1, pp. D607–D613, 2019. DOI: 10.1093/nar/gky1131. pmid: 30476243.

[151] M. Zeng, F. Zhang, F.-X. Wu, Y. Li, J. Wang, and M. Li, "Protein–protein interaction site prediction through combining local and global features with deep neural networks," *Bioinformatics*, vol. 36, no. 4, pp. 1114–1120, 2020. DOI: 10.1093/bioinformatics/btz699.

[152] R. Nikam, K. Yugandhar, and M. M. Gromiha, "DeepBSRPred: Deep learning-based binding site residue prediction for proteins," *Amino Acids*, vol. 55, no. 10, pp. 1305–1316, 2023. DOI: 10.1007/s00726-022-03228-3.

[153] D. Wang *et al.*, "MusiteDeep: A deep-learning based webserver for protein post-translational modification site prediction and visualization," *Nucleic Acids Research*, vol. 48, no. W1, W140–W146, 2020. DOI: 10.1093/nar/gkaa275.

[154] M. Higurashi, T. Ishida, and K. Kinoshita, "PiSite: A database of protein interaction sites using multiple binding states in the PDB," *Nucleic Acids Research*, vol. 37, pp. D360–D364, Database issue 2009. DOI: 10.1093/nar/gkn659. pmid: 18836195.

[155] M. Torrisi, G. Pollastri, and Q. Le, "Deep learning methods in protein structure prediction," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1301–1310, 2020. DOI: 10.1016/j.csbj.2019.12.011.

[156] J. Abramson *et al.*, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, vol. 630, no. 8016, pp. 493–500, 2024. DOI: 10.1038/s41586-024-07487-w.

[157] M. Varadi *et al.*, "AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models," *Nucleic Acids Research*, vol. 50, no. D1, pp. D439–D444, 2022. DOI: 10.1093/nar/gkab1061. pmid: 34791371.

[158] The UniProt Consortium, *Titin (q8wz42 - titin_human)*, https://www.uniprot.org/uniprotkb/Q8WZ42, Last accessed on 2025-01-12.

[159] S. Labeit and B. Kolmerer, "Titins: Giant Proteins in Charge of Muscle Ultrastructure and Elasticity," *Science*, vol. 270, no. 5234, pp. 293–296, 1995. DOI: 10.1126/science.270.5234.293.

[160] The UniProt Consortium, *T cell receptor delta diversity 1 (p0dpr3 - trdd1_human)*, https://www.uniprot.org/uniprotkb/P0DPR3, Last accessed on 2025-01-12.

[161] S. Unsal, H. Atas, M. Albayrak, K. Turhan, A. C. Acar, and T. Doğan, "Learning functional properties of proteins with language models," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 227–245, 3 2022. DOI: 10.1038/s42256-022-00457-9.

[162] R. M. Rao *et al.*, "MSA Transformer," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, 2021, pp. 8844–8856.

[163] F. Zhuang *et al.*, "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021. DOI: 10.1109/JPROC.2020.3004555.

[164] J. Wang and Y. Chen, *Introduction to Transfer Learning: Algorithms and Practice* (Machine Learning: Foundations, Methodologies, and Applications). Singapore: Springer Nature, 2023. DOI: 10.1007/978-981-19-7584-4.

[165] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. DOI: 10.1007/BF00994018.

[166] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, IEEE, 1995, pp. 278–282.

[167] D. Vucetic, M. Tayaranian, M. Ziaeefard, J. J. Clark, B. H. Meyer, and W. J. Gross, "Efficient fine-tuning of BERT models on the edge," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 2022, pp. 1838–1842.

[168] J. Lee, R. Tang, and J. Lin, "What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning," 2019. arXiv: 1911.03090 [cs].

[169] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[170] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, Online, 2021.

[171] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[172] J. Clusmann *et al.*, "The future landscape of large language models in medicine," *Communications Medicine*, vol. 3, no. 1, pp. 1–8, 1 2023. DOI: 10.1038/s43856-023-00370-1.

[173] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," in *NIPS 2016 Deep Learning Symposium*, 2016. arXiv: 1607.06450.

[174] J. S. Bridle, "Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition," in *Neurocomputing*, F. F. Soulié and J. Hérault, Eds., Berlin, Heidelberg: Springer, 1990, pp. 227–236. DOI: 10.1007/978-3-642-76153-9_28.

[175] T. B. Brown *et al.*, "Language models are few-shot learners," 2020. arXiv: 2005.14165.

[176] Y. Zhu *et al.*, "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 19–27. DOI: 10.1109/ICCV.2015.11.

[177] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," 2019. arXiv: 1907.11692.

[178] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," 2020. arXiv: 1909.11942 [cs].

[179] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter." arXiv: 1910.01108 [cs]. (2020), [Online]. Available: http://arxiv.org/abs/1910.01108, pre-published.

[180] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," 2020. arXiv: 2003.10555.

[181] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[182] J. K. Tripathy *et al.*, "Comprehensive analysis of embeddings and pre-training in NLP," *Computer Science Review*, vol. 42, p. 100 433, 2021. DOI: 10.1016/j.cosrev.2021.100433.

[183] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, 2018.

[184] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[185] B. Wang and A. Komatsuzaki, *GPT-J-6B: A 6 billion parameter autoregressive language model*, 2021.

[186] V. Sanh. "DistilGPT2," GitHub. (2020), [Online]. Available: https://github.com/huggingface/transformers/tree/main/examples/research_projects/distillation.

[187] OpenAI. "OpenAI GPT-3 API [text-davinci-003]," GPT-3.5 Turbo. (2023), [Online]. Available: https://platform.openai.com/docs/models/gpt-3-5-turbo.

[188] L. Ouyang *et al.* "Training language models to follow instructions with human feedback." arXiv: 2203.02155 [cs]. (2022), [Online]. Available: http://arxiv.org/abs/2203.02155, pre-published.

[189] OpenAI *et al.* "GPT-4 Technical Report." arXiv: 2303.08774 [cs]. (2024), [Online]. Available: http://arxiv.org/abs/2303.08774, pre-published.

[190] K. S. Kalyan, "A survey of GPT-3 family large language models including ChatGPT and GPT-4," *Natural Language Processing Journal*, vol. 6, p. 100 048, 2024. DOI: 10.1016/j.nlp.2023.100048.

[191] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020. DOI: 10.1093/bioinformatics/btz682.

[192] Y. Gu *et al.*, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," 2020.

[193] K. raj Kanakarajan, B. Kundumani, and M. Sankarasubbu, "BioELECTRA:Pretrained Biomedical text Encoder using Discriminators," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, Online: Association for Computational Linguistics, 2021, pp. 143–154. DOI: 10.18653/v1/2021.bionlp-1.16.

[194] I. Beltagy, A. Cohan, and K. Lo, "SciBERT: Pretrained Contextualized Embeddings for Scientific Text," 2019. arXiv: 1903.10676 [cs].

[195] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," presented at the CHIL 2020: ACM Conference on Health, Inference, and Learning, Toronto, ON: ACM, New York, NY, USA, 2020. arXiv: 1904.05342 [cs].

[196] Y. Li *et al.*, "BEHRT: Transformer for electronic health records," *Scientific Reports*, vol. 10, no. 1, p. 7155, 2020. DOI: 10.1038/s41598-020-62922-y.

[197] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," *Bioinformatics (Oxford, England)*, vol. 37, no. 15, J. Kelso, Ed., pp. 2112–2120, 2021. DOI: 10.1093/bioinformatics/btab083.

[198] Z. Lin *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023. DOI: 10.1126/science.ade2574.

[199] B. Wang *et al.*, "Pre-trained Language Models in Biomedical Domain: A Systematic Survey," *ACM Computing Surveys*, vol. 56, no. 3, 55:1–55:52, 2023. DOI: 10.1145/3611651.

[200] H.-C. Shin *et al.*, "BioMegatron: Larger Biomedical Domain Language Model," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, 2020, pp. 4700–4706. DOI: 10.18653/v1/2020.emnlp-main.379.

[201] V. P. Dwivedi and X. Bresson, "A Generalization of Transformer Networks to Graphs," presented at the AAAI Workshop on Deep Learning on Graphs: Methods and Applications, 2021. DOI: 10.48550/arXiv.2012.09699. arXiv: 2012.09699 [cs].

[202] Y. Gu *et al.*, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, 2022. DOI: 10.1145/3458754.

[203] N. Ferruz, S. Schmidt, and B. Höcker, "ProtGPT2 is a deep unsupervised language model for protein design," *Nature Communications*, vol. 13, no. 1, p. 4348, 1 2022. DOI: 10.1038/s41467-022-32007-7.

[204] R. Luo *et al.*, "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, bbac409, 2022. DOI: 10.1093/bib/bbac409.

[205] OpenAI, *ChatGPT*, OpenAI, 2022.

[206] Y. Li *et al.*, "Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 1106–1117, 2023. DOI: 10.1109/JBHI.2022.3224727.

[207] H. Touvron *et al.* "LLaMA: Open and Efficient Foundation Language Models." arXiv: 2302.13971 [cs]. (2023), [Online]. Available: http://arxiv.org/abs/2302.13971, pre-published.

[208] H. Touvron *et al.* "Llama 2: Open Foundation and Fine-Tuned Chat Models." arXiv: 2307.09288 [cs]. (2023), [Online]. Available: http://arxiv.org/abs/2307.09288, pre-published.

[209] N. Hollmann *et al.*, "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, no. 8045, pp. 319–326, 2025. DOI: 10.1038/s41586-024-08328-6.

[210] Google, *Gemini*, Google, 2023.

[211] Meta, *Llama Chat*, Meta, 2023.

[212] A. Plaat, *Deep Reinforcement Learning*. Singapore: Springer Nature, 2022. DOI: 10.1007/978-981-19-0638-1.

[213] N. S. Babaiha, S. G. Rao, J. Klein, B. Schultz, M. Jacobs, and M. Hofmann-Apitius, "Rationalism in the face of GPT hypes: Benchmarking the output of large language models against human expert-curated biomedical knowledge graphs," *Artificial Intelligence in the Life Sciences*, p. 100 095, 2024.

[214] Q. Chen *et al.* "A Comprehensive Benchmark Study on Biomedical Text Generation and Mining with ChatGPT." (2023), [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.04.19.537463v1, pre-published.

[215] I. Jahan, M. T. R. Laskar, C. Peng, and J. X. Huang, "A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks," *Computers in Biology and Medicine*, vol. 171, p. 108 189, 2024. DOI: 10.1016/j.compbiomed.2024.108189.

[216] A. Wang, C. Liu, J. Yang, and C. Weng. "Fine-tuning Large Language Models for Rare Disease Concept Normalization." (2023), [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.12.28.573586v1, pre-published.

[217] H. Wang, C. Gao, C. Dantona, B. Hull, and J. Sun, "DRG-LLaMA : Tuning LLaMA model to predict diagnosis-related group for hospitalized patients," *npj Digital Medicine*, vol. 7, no. 1, pp. 1–9, 2024. DOI: 10.1038/s41746-023-00989-3.

[218] H. Zhou, M. Li, Y. Xiao, H. Yang, and R. Zhang, "LLM Instruction-Example Adaptive Prompting (LEAP) Framework for Clinical Relation Extraction," *medRxiv*, p. 2023.12.15.23300059, 2023. DOI: 10.1101/2023.12.15.23300059.

[219] S. Hong, L. Xiao, X. Zhang, and J. Chen. "ArgMed-Agents: Explainable Clinical Decision Reasoning with Large Language Models via Argumentation Schemes." arXiv: 2403.06294 [cs]. (2024), [Online]. Available: http://arxiv.org/abs/2403.06294, pre-published.

[220] S. Sandmann, S. Riepenhausen, L. Plagwitz, and J. Varghese, "Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks," *Nature Communications*, vol. 15, no. 1, p. 2050, 2024. DOI: 10.1038/s41467-024-46411-8.

[221] M. Wornow *et al.*, "The shaky foundations of large language models and foundation models for electronic health records," *npj Digital Medicine*, vol. 6, no. 1, pp. 1–10, 2023. DOI: 10.1038/s41746-023-00879-8.

[222] M. Moor *et al.*, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 7956 2023. DOI: 10.1038/s41586-023-05881-4.

[223] Y. Zhu, L. Li, H. Lu, A. Zhou, and X. Qin, "Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions," *Journal of Biomedical Informatics*, vol. 106, p. 103 451, 2020. DOI: 10.1016/j.jbi.2020.103451.

[224] D. Hu, H. Zhang, S. Li, Y. Wang, N. Wu, and X. Lu, "Automatic Extraction of Lung Cancer Staging Information From Computed Tomography Reports: Deep Learning Approach," *JMIR medical informatics*, vol. 9, no. 7, e27955, 2021. DOI: 10.2196/27955. pmid: 34287213.

[225] Q. Yuan, S. Chen, J. Rao, S. Zheng, H. Zhao, and Y. Yang, "AlphaFold2-aware protein–DNA binding site prediction using graph transformer," *Briefings in Bioinformatics*, vol. 23, no. 2, bbab564, 2022. DOI: 10.1093/bib/bbab564.

[226] C. Pang *et al.* "CEHR-GPT: Generating Electronic Health Records with Chronological Patient Timelines." arXiv: 2402.04400 [cs]. (2024), [Online]. Available: http://arxiv.org/abs/2402.04400, pre-published.

[227] K. S. Krishnan and K. S. Krishnan, "Vision Transformer based COVID-19 Detection using Chest X-rays," in *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, 2021, pp. 644–648. DOI: 10.1109/ISPCC53510.2021.9609375.

[228] H. Chen *et al.*, "GasHis-Transformer: A multi-scale visual transformer approach for gastric histopathological image detection," *Pattern Recognition*, vol. 130, p. 108 827, 2022. DOI: 10.1016/j.patcog.2022.108827.

[229] S. Tummala, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, "Classification of Brain Tumor from Magnetic Resonance Imaging using Vision Transformers Ensembling," *Current Oncology*, vol. 29, no. 10, pp. 7498–7511, 10 2022. DOI: 10.3390/curroncol29100590.

[230] H. Wang, F. Guo, M. Du, G. Wang, and C. Cao, "A novel method for drug-target interaction prediction based on graph transformers model," *BMC Bioinformatics*, vol. 23, no. 1, p. 459, 2022. DOI: 10.1186/s12859-022-04812-w.

[231] E.-d. El-allaly, M. Sarrouti, N. En-Nahnahi, and S. Ouatik El Alaoui, "An attentive joint model with transformer-based weighted graph convolutional network for extracting adverse drug event relation," *Journal of Biomedical Informatics*, vol. 125, p. 103 968, 2022. DOI: 10.1016/j.jbi.2021.103968.

[232] Q. Shi, J. Fan, Z. Wang, and Z. Zhang, "Multimodal channel-wise attention transformer inspired by multisensory integration mechanisms of the brain," *Pattern Recognition*, vol. 130, p. 108 837, 2022. DOI: 10.1016/j.patcog.2022.108837.

[233] M. Monajatipoor, M. Rouhsedaghat, L. H. Li, C.-C. J. Kuo, A. Chien, and K.-W. Chang, "BERTHop: An Effective Vision-and-Language Model for Chest X-ray Disease Diagnosis," *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 13435, pp. 725–734, 2022. DOI: 10.1007/978-3-031-16443-9_69. pmid: 37093922.

[234] L. Allen, J. Scott, A. Brand, M. Hlava, and M. Altman, "Publishing: Credit where credit is due," *Nature*, vol. 508, no. 7496, pp. 312–313, 2014. DOI: 10.1038/508312a.

[235] A. Brand, L. Allen, M. Altman, M. Hlava, and J. Scott, "Beyond authorship: Attribution, contribution, collaboration, and credit," *Learned Publishing*, vol. 28, no. 2, pp. 151–155, 2015. DOI: 10.1087/20150211.

[236] J. J. Silverman *et al.*, *Practice Guidelines for the Psychiatric Evaluation of Adults*, Third Edition. American Psychiatric Association, 2016.

[237] R.-D. Stieglitz, A. Haug, E. Fähndrich, M. Rösler, and W. Trabert, "Comprehensive Psychopathological Assessment Based on the Association for Methodology and Documentation in Psychiatry (AMDP) System: Development, Methodological Foundation, Application in Clinical Routine, and Research," *Frontiers in Psychiatry*, vol. 8, 2017. DOI: 10.3389/fpsyt.2017.00045. pmid: 28439242.

[238] *Das AMDP-System – Manual Zur Dokumentation Psychiatrischer Befunde*, 10th ed. Arbeitsgemeinschaft für Methodik und Dokumentation, 2018.

[239] W. H. Organization, *ICD-10 : international statistical classification of diseases and related health problems : tenth revision*. World Health Organization, 2004.

[240] Branden Chan, Timo Möller, Malte Pietsch, Tanay Soni, and Chin Man Yeung. "Open Sourcing German BERT," deepset. (2019), [Online]. Available: https://deepset.ai/german-bert.

[241] W. Querbes, A. Benmerah, D. Tosoni, P. P. Di Fiore, and W. J. Atwood, "A JC virus-induced signal is required for infection of glial cells by a clathrin- and eps15-dependent pathway," *Journal of Virology*, vol. 78, no. 1, pp. 250–256, 2004. DOI: 10.1128/jvi.78.1.250-256.2004. pmid: 14671106.

[242] P. A. Swanson and D. B. McGavern, "Viral diseases of the central nervous system," *Current Opinion in Virology*, Viral Pathogenesis • Preventive and Therapeutic Vaccines, vol. 11, pp. 44–54, 2015. DOI: 10.1016/j.coviro.2014.12.009.

[243] H. R. van Doorn and H. Yu, "33 - Viral Respiratory Infections," in *Hunter's Tropical Medicine and Emerging Infectious Diseases (Tenth Edition)*, E. T. Ryan, D. R. Hill, T. Solomon, N. E. Aronson, and T. P. Endy, Eds., London: Elsevier, 2020, pp. 284–288. DOI: 10.1016/B978-0-323-55512-8.00033-8.

[244] D. Ye *et al.*, "Trafficking of JC virus-like particles across the blood–brain barrier," *Nanoscale Advances*, 2021.

[245] T. Guirimand, S. Delmotte, and V. Navratil, "VirHostNet 2.0: Surfing on the web of virus/host molecular interactions data," *Nucleic Acids Research*, vol. 43, no. D1, pp. D583–D587, 2015. DOI: 10.1093/nar/gku1121.

[246] M. Zitnik, R. Sosič, S. Maheshwari, and J. Leskovec. "BioSNAP Datasets: Stanford Biomedical Network Dataset Collection." (2018), [Online]. Available: http://snap.stanford.edu/biodata.

[247] S. Tsukiyama, M. M. Hasan, S. Fujii, and H. Kurata, "LSTM-PHV: Prediction of human-virus protein–protein interactions by LSTM with word2vec," *Briefings in Bioinformatics*, bbab228 2021. DOI: 10.1093/bib/bbab228.

[248] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucleic Acids Research*, vol. 36, no. 9, pp. 3025–3030, 2008. DOI: 10.1093/nar/gkn159.

[249] T. Sun, B. Zhou, L. Lai, and J. Pei, "Sequence-based prediction of protein protein interaction using a deep-learning algorithm," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–8, 2017.

[250] M. Chen *et al.*, "Multifaceted protein–protein interaction prediction based on Siamese residual RCNN," *Bioinformatics*, vol. 35, no. 14, pp. i305–i314, 2019. DOI: 10.1093/bioinformatics/btz328.

[251] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," 2017. arXiv: 1703.01365 [cs].

[252] J. Takamizawa *et al.*, "Reduced Expression of the let-7 MicroRNAs in Human Lung Cancers in Association with Shortened Postoperative Survival," *Cancer Research*, vol. 64, no. 11, pp. 3753–3756, 2004. DOI: 10.1158/0008-5472.CAN-04-0637.

[253] H. Rupani, T. Sanchez-Elsner, and P. Howarth, "MicroRNAs and respiratory diseases," *European Respiratory Journal*, vol. 41, no. 3, pp. 695–705, 2013. DOI: 10.1183/09031936.00212011. pmid: 22790917.

[254] S. Kumar *et al.*, "Synaptosome microRNAs regulate synapse functions in Alzheimer's disease," *NPJ Genomic Medicine*, vol. 7, p. 47, 2022. DOI: 10.1038/s41525-022-00319-8. pmid: 35941185.

[255] Y. Li *et al.*, "HMDD v2.0: A database for experimentally supported human microRNA and disease associations," *Nucleic Acids Research*, vol. 42, no. D1, pp. 1–5, 2014. DOI: 10.1093/nar/gkt1023. pmid: 24194601.

[256] Z. Huang *et al.*, "HMDD v3.0: A database for experimentally supported human microRNA-disease associations," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1013–D1017, 2019. DOI: 10.1093/nar/gky1010. pmid: 30364956.

[257] Q. Jiang *et al.*, "miR2Disease: A manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, SUPPL. 1 2009. DOI: 10.1093/nar/gkn714. pmid: 18927107.

[258] S. Bagewadi, T. Bobić, M. Hofmann-Apitius, J. Fluck, and R. Klinger, "Detecting miRNA Mentions and Relations in Biomedical Literature," *F1000Research*, 2015. DOI: 10.12688/f1000research.4591.3. pmid: 26535109.

[259] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "miRBase: From microRNA sequences to function," *Nucleic Acids Research*, vol. 47, no. D1, pp. D155–D162, 2019. DOI: 10.1093/nar/gky1141.

[260] D. Wright, Y. Katsis, R. Mehta, and C.-N. Hsu, "NormCo: Deep Disease Normalization for Biomedical Knowledge Base Construction," 2018.

[261] National Library of Medicine, *About pubmed central (pmc)*, https://pmc.ncbi.nlm.nih.gov/about/intro/, Last accessed on 2025-01-12.

[262] H. Dalianis, *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer Nature, 2018.

[263] A. Le Glaz *et al.*, "Machine Learning and Natural Language Processing in Mental Health: Systematic Review," *Journal of Medical Internet Research*, vol. 23, no. 5, e15708, 2021. DOI: 10.2196/15708. pmid: 33944788.

[264] E. Mansouri-Benssassi *et al.*, "Disclosure control of machine learning models from trusted research environments (TRE): New challenges and opportunities," *Heliyon*, vol. 9, no. 4, 2023. DOI: 10.1016/j.heliyon.2023.e15143. pmid: 37123891.

[265] U. Hahn. "Clinical Document Corpora and Assorted Domain Proxies: A Survey of Diversity in Corpus Design, with Focus on German Text Data." arXiv: 2412.00230 [cs]. (2024), [Online]. Available: http://arxiv.org/abs/2412.00230, pre-published.

[266] S. Tian *et al.*, "Opportunities and challenges for ChatGPT and large language models in biomedicine and health," *Briefings in Bioinformatics*, vol. 25, no. 1, bbad493, 2024. DOI: 10.1093/bib/bbad493.

[267] N. Ding *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220–235, 2023. DOI: 10.1038/s42256-023-00626-4.

[268] T. Lai, H. Ji, C. Zhai, and Q. H. Tran. "Joint Biomedical Entity and Relation Extraction with Knowledge-Enhanced Collective Inference." arXiv: 2105.13456 [cs]. (2021), [Online]. Available: http://arxiv.org/abs/2105.13456, pre-published.

[269] C. Schröder and A. Niekler, "A Survey of Active Learning for Text Classification using Deep Neural Networks," 2020. arXiv: 2008.07267 [cs].

[270] G. Liu *et al.*, "Medical-vlbert: Medical visual language bert for covid-19 ct report generation with alternate learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 9, pp. 3786–3797, 2021.

[271] S. Koorathota, Z. Khan, P. Lapborisuth, and P. Sajda, "Multimodal Neurophysiological Transformer for Emotion Recognition," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022, pp. 3563–3567. DOI: 10.1109/EMBC48229.2022.9871421.

[272] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *International Conference on Machine Learning*, PMLR, 2021, pp. 4651–4664.

[273] A. Jaegle *et al.*, "Perceiver IO: A General Architecture for Structured Inputs & Outputs," presented at the International Conference on Learning Representations, 2022.

[274] Z. Yang, X. Du, J. Li, J. Zheng, S. Poria, and E. Cambria, "Large language models for automated open-domain scientific hypotheses discovery," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 13 545–13 565. DOI: 10.18653/v1/2024.findings-acl.804.

[275] I. Taleb, A. N. Navaz, and M. A. Serhani, "Leveraging Large Language Models for Enhancing Literature-Based Discovery," *Big Data and Cognitive Computing*, vol. 8, no. 11, p. 146, 11 2024. DOI: 10.3390/bdcc8110146.

[276] W. Wang *et al.* "SciPIP: An LLM-based Scientific Paper Idea Proposer." arXiv: 2410.23166 [cs]. (2024), [Online]. Available: http://arxiv.org/abs/2410.23166, pre-published.

[277] F. Zhang *et al.* "Recent Methodological Advances in Federated Learning for Healthcare." arXiv: 2310.02874 [cs]. (2023), [Online]. Available: http://arxiv.org/abs/2310.02874, pre-published.

[278] H. Li *et al.*, "FedTP: Federated Learning by Transformer Personalization," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023. DOI: 10.1109/TNNLS.2023.3269062.

[279] T. K. Dang, X. Lan, J. Weng, and M. Feng, "Federated Learning for Electronic Health Records," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 5, 72:1–72:17, 2022. DOI: 10.1145/3514500.

[280] L. Peng *et al.*, "An in-depth evaluation of federated learning on biomedical natural language processing for information extraction," *npj Digital Medicine*, vol. 7, no. 1, pp. 1–9, 2024. DOI: 10.1038/s41746-024-01126-4.

[281] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, p. 195, 2019. DOI: 10.1186/s12916-019-1426-2.

[282] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nature Medicine*, vol. 28, no. 1, pp. 31–38, 1 2022. DOI: 10.1038/s41591-021-01614-0.

[283] S. Benjamens, P. Dhunnoo, and B. Meskó, "The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–8, 2020. DOI: 10.1038/s41746-020-00324-0.

[284] N. A. Caine *et al.*, "A 2030 Vision for the Mayo Clinic Department of Medicine," *Mayo Clinic Proceedings*, vol. 97, no. 7, pp. 1232–1236, 2022. DOI: 10.1016/j.mayocp.2022.02.010. pmid: 35787852.

Appendices

A. Transformer Models in Biomedicine

Reprinted with permission from:

S. Madan, M. Lentzen, J. Brandt, D. Rueckert, M. Hofmann-Apitius, and H. Fröhlich, "Transformer models in biomedicine," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 214, 2024. DOI: 10.1186/s12911-024-02600-5

Madan *et al.*
BMC Medical Informatics and Decision Making (2024) 24:214
https://doi.org/10.1186/s12911-024-02600-5

BMC Medical Informatics and Decision Making

REVIEW

Open Access

Transformer models in biomedicine

Sumit Madan[1,2*], Manuel Lentzen[1,3], Johannes Brandt[4], Daniel Rueckert[4,5,6], Martin Hofmann-Apitius[1,3] and Holger Fröhlich[1,3*]

Abstract

Deep neural networks (DNN) have fundamentally revolutionized the artificial intelligence (AI) field. The transformer model is a type of DNN that was originally used for the natural language processing tasks and has since gained more and more attention for processing various kinds of sequential data, including biological sequences and structured electronic health records. Along with this development, transformer-based models such as BioBERT, MedBERT, and MassGenie have been trained and deployed by researchers to answer various scientific questions originating in the biomedical domain. In this paper, we review the development and application of transformer models for analyzing various biomedical-related datasets such as biomedical textual data, protein sequences, medical structured-longitudinal data, and biomedical images as well as graphs. Also, we look at explainable AI strategies that help to comprehend the predictions of transformer-based models. Finally, we discuss the limitations and challenges of current models, and point out emerging novel research directions.

Keywords Transformer, Biomedicine, Life Science, Deep learning, Neural networks, Machine learning

Introduction

The transformer [1] is a well-known deep neural network (DNN) model, which has revolutionized the artificial intelligence (AI) field. The architecture of the transformer builds the backbone of large language models (LLM), enabling them to harness the power of vast amounts of data to gain a more profound understanding of the underlying information. The architecture was initially developed for comprehending natural language, achieving this by analyzing every input sentence and capturing the context of each word through focusing on other words. Generic LLMs have brought significant advancements to various natural language processing (NLP) tasks ranging from machine translation over text generation to question answering. Most common examples of generic LLMs include Generative Pre-trained Transformer (GPT) [2], Bidirectional Encoder Representations from Transformers (BERT) [3], Large Language Model Meta AI (LLamA) [4, 5], and BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) [6].

The success of transformer-based models can be attributed to the self-attention mechanism, integrated encoder-decoder architecture, and scalable as well as modular structure. These characteristics allow it to learn effective representations of the underlying data, encode long-range dependencies, and process huge amounts of data in an efficient way. The basic building block of the transformer is the self-attention mechanism [1, 7]. This mechanism allows the model to learn complex sequence representations by incorporating or attending to the

*Correspondence:
Sumit Madan
sumit.madan@scai.fraunhofer.de
Holger Fröhlich
holger.froehlich@scai.fraunhofer.de
[1] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, Sankt Augustin 53757, Germany
[2] Institute of Computer Science, University of Bonn, Bonn 53115, Germany
[3] Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, Bonn 53115, Germany
[4] School of Medicine, Klinikum Rechts der Isar, Technical University Munich, Munich, Germany
[5] School of Computation, Information and Technology, Technical University Munich, Munich, Germany
[6] Department of Computing, Imperial College London, London, UK

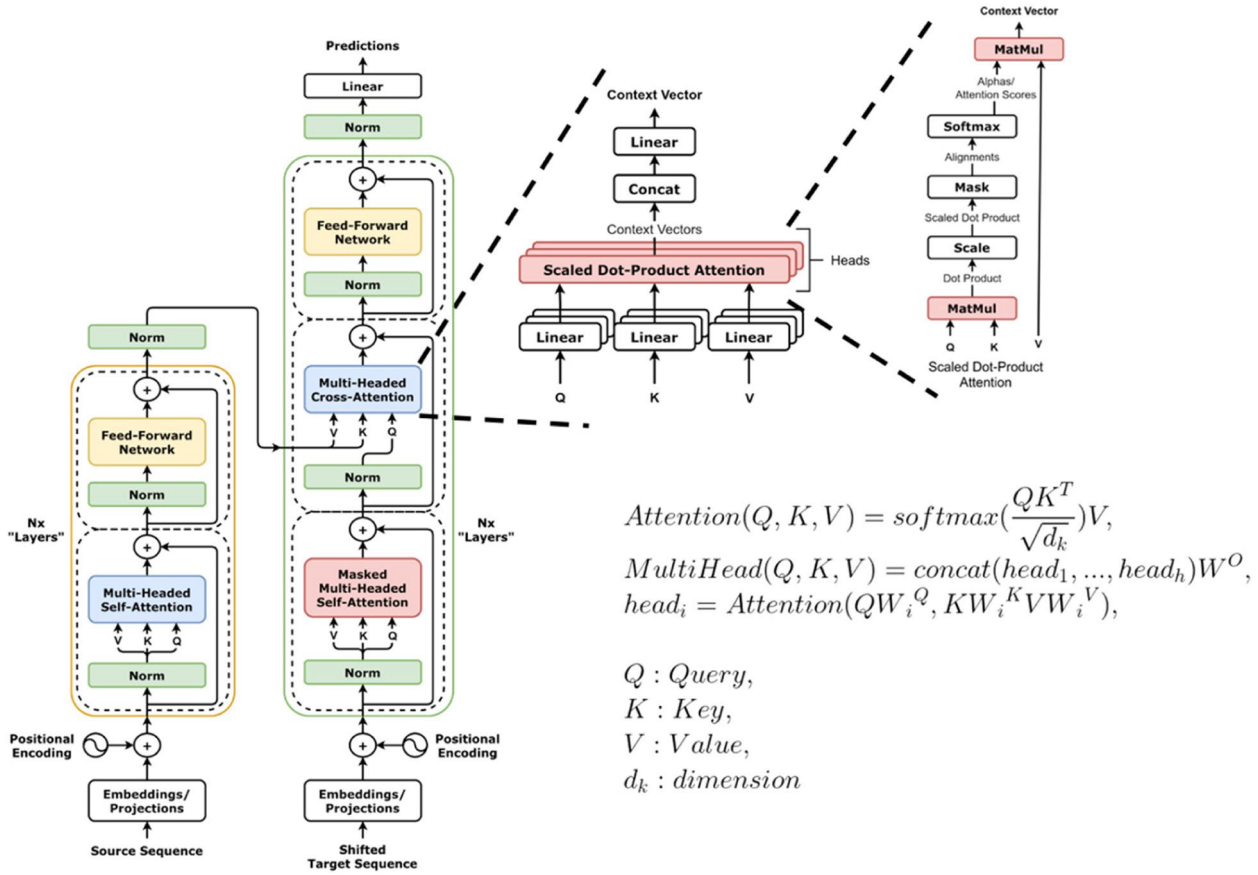


Fig. 1 The transformer architecture with its self-attention mechanism. Original transformer images by https://github.com/dvgodoy/dl-visuals / CC BY 4.0.)

information throughout the other parts of the same sequence. Equally important is the encoder-decoder structure of the transformer while both comprise multiple layers and variants of the self-attention mechanism (Fig. 1). This type of architecture facilitates sequence-to-sequence learning; therefore, transformers were originally used to solve the machine translation problem (e.g., translation from English to German). The encoder-only architecture (for instance, utilized in BERT) can be used for classification and understanding tasks [3], whereas decoders-only (such as GPT, LLamA, and BLOOM) are used for generative tasks [8, 9]. Furthermore, the modular and scalable architecture of the transformer allows the stacking of encoder and decoder blocks on each other, which substantially increases the capacity of the model. By processing huge amounts of data with larger models, the performance of transformers has been significantly increased on various tasks [8].

Transformer-based models such as BERT or GPT apply a two step process in their approach to understand the data provided to them and handling various downstream tasks. In a pre-training phase, they leverage the abundant unlabeled data to learn a general representation through an embedding model of the underlying objects in a self-supervised manner. Unlabeled data, characterized by the absence of labels or tags, is widely available. For instance, the web contains a vast amount of textual content in the form of web pages, blogs, and forums that are not categorized or labeled. In contrast, labeled datasets contain data that have been annotated with specific labels or categories, such as the label "gene" in case of biomedical texts. Due to the manual annotation process, obtaining labeled data is often more challenging and time-consuming compared to unlabeled data. In the fine-tuning phase, the pre-trained general representation model is used to train a supervised use case-specific task model using the limited labeled data. Over time they have been applied successfully beyond language to process other modalities and brought significant advancements to speech processing, computer vision (CV), and many more areas.

Transformers are now in the spotlight of many areas of biomedical-related AI research. They have been proven instrumental in addressing diverse biomedical-related questions, facilitating the analysis of data

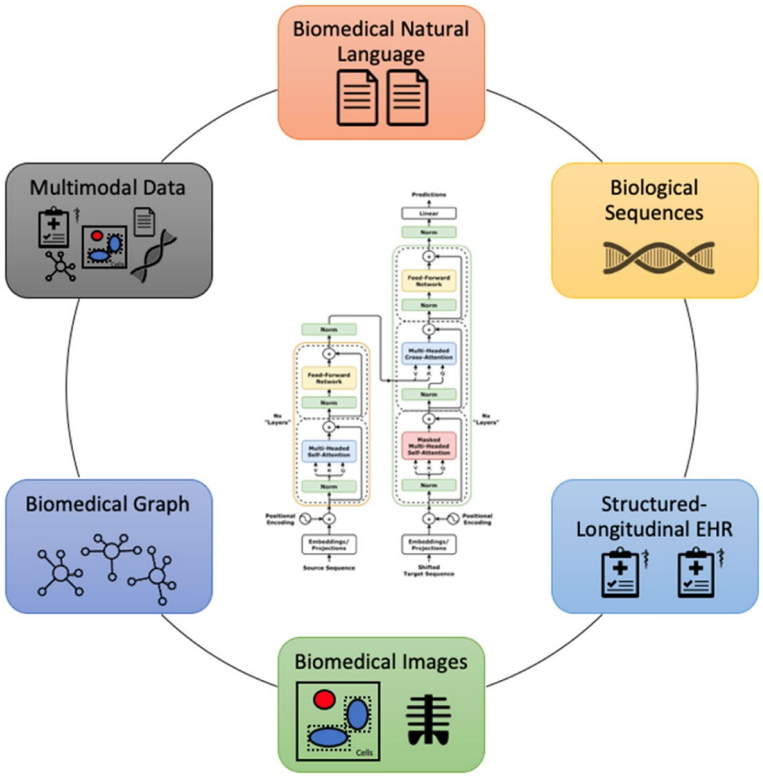


Fig. 2 Application fields of transformers in biomedicine. Transformer image by https://github.com/dvgodoy/dl-visuals / CC BY 4.0

modalities ranging from biomedical literature to complex imaging and genetic information. The pace of progress has reached a limit that is difficult to grasp and, therefore, requires a thorough survey of the field. To our knowledge, such a thorough review is missing so far. Our paper thus tries to fill a gap. In the following, we highlight and discuss transformer-based models in five application fields (Fig. 2): 1) biomedical natural language processing (including biomedical literature, clinical notes, and social media text), 2) biological sequences (including protein sequences), 3) structured-longitudinal electronic health records (EHR), 4) biomedical images, and 5) biomedical graphs. We also introduce some studies that have pursued learning on multiple modalities jointly. Finally, we discuss methods to make transformer-based predictions, and we conclude by providing a prospect for future research.

Table 1 provides a glossary of concepts of AI that are discussed in this work. The mathematical details on transformers will not be elaborated in this work, however, we refer the readers to [10, 11] for more details.

Biomedical natural language processing

Domain-specific transformers

Transformer-based models have made major strides in the biomedical NLP field, largely through adapting general language models for the biomedical domain by pre-training on huge publicly available biomedical corpora including documents from databases such as PubMed, PubMed Central (PMC), and Medical Information Mart for Intensive Care-III (MIMIC-III) [12, 13]. The majority of the studies introducing domain-specific language models often follow a familiar pattern, focusing on a specific transformer-based model architecture, initializing it with random weights or the weights of a general language model, pre-training the initialized model with domain-specific corpora as well as multiple objective tasks, and finally evaluating the models with different sizes on various biomedical downstream tasks.

For instance, BioBERT, which is initialized with the weights of the general English language model Bidirectional Encoder Representations from Transformers (BERT) [3], is a domain-specific model further pre-trained on PubMed abstracts and PMC full-text documents [14]. BioBERT was fine-tuned for various

Table 1 Glossary of AI and machine learning terminology sorted alphabetically

Concept	Definition
Classification head	It consists of one or multiple layers attached to the head of the model that outputs predictions, typically accepting embeddings and performing classification, prediction or other tasks.
Contrastive learning	A technique that can be used to improve performance of machine learning models by learning representations that bring similar samples closer to each other and move dissimilar ones further apart in the embedding space.
Decoder	Similar to the encoder, the decoder in the transformer architecture also captures the relevant context of the input data and generates an output sequence by translating the high-dimensional embedded information in a step-by-step process. It can be used for many generative tasks performed by models such as GPT, BioGPT, and CEHR-GPT.
Distant supervision	A technique in machine learning that utilizes indirect labels to generate or augment training datasets for model training.
Domain-specific transformer-based models	These specialized models have been pre-trained or fine-tuned on data from a specific domain such as biomedicine.
Embeddings	They are numerical representations of data needed to process them with machine learning algorithms. Embeddings can be generated for various kinds of objects such as words, proteins, diagnosis codes, and medications.
Encoder	As a part of the transformer architecture captures the relevant context of the input data and generates a high-dimensional embedding that can be used in many downstream tasks (such as text classification, graph node classification, and image segmentation). Transformer-based models such as BERT are based solely on the encoder.
Explainable AI (XAI)	XAI refers to the field of designing algorithms that can comprehend predictions of machine learning models, providing more insights into their decision-making process.
Generative modeling	Refers to the process of generating new samples from a certain distribution that is learned from the underlying training data. The pre-trained transformer decoder-based models can generate new text, protein sequences, structured EHRs, and images.
Graph neural networks	These are neural networks specifically designed to learn from homogeneous and heterogeneous knowledge graphs and can perform tasks such as node classification, link prediction, and graph or sub-graph classification. They learn from the underlying structure and interconnections in graphs.
Fine-tuning phase	In this phase, the pre-trained model can subsequently be tuned by an additional supervised training for a specific task using labeled data.
Large language models (LLM)	LLMs are large neural networks based on transformer architecture pre-trained on vast amounts of textual data. They are capable of understanding and generating natural language.
Machine reading comprehension task	This task aims to train algorithms to comprehend and extract relevant information from textual data. The algorithm takes in a document and a query as input. The goal is to derive the correct answer based on the provided text. One common application is span extraction, where the output consists of the relevant span of text from the document that answers the query. During query definition additional knowledge can also be utilized to improve the performance.
Machine translation	This is a task of automatically translating from one human language to another using advanced machine learning algorithms.
Pre-training phase	In this phase, the machine learning models leverage the data to learn a general representation of the underlying objects (such as text or images) in a self-supervised manner.
Self-attention mechanism	It is a technique utilized in deep neural networks to focus on all or different parts of the input while processing this input. For instance, to learn the context of each word in a sentence, the model attends to every other word. This technique has been proven as beneficial for natural language processing and other sequential data.
Self-supervised learning	It is a machine learning approach in which the intrinsic data properties are utilized to create pseudo-labels for the data itself. Subsequently, the models are trained using this self-labeled data to understand and learn the underlying patterns and relationships.
Sequence labeling task	The objective of this task is to assign labels to units of a sequence. For instance, words or tokens in a sentence can be labeled as biological concepts. Similarly, in case of protein sequences, amino acids can be assigned labels for secondary structure elements (such as alpha-helix, beta-strand, or coil).
Sequence-to-sequence learning	The models trained with this learning technique are designed to map input sequences of one domain to output sequences of another domain. Summarization or translation of text or predicting the secondary structure from protein sequence are typical examples of sequence-to-sequence learning tasks.
Relational graph attention networks	These are a type of graph neural network that, as the name suggests, apply the self-attention mechanism to graph-relational data modeling the different relationships embedded in the graph.
Representation learning	It enables the data such as text, images, protein sequences for processing with mathematical operations by representing them as compact and dense vectors. These vectors are also called embeddings that carry relevant features of input data.

Table 1 (continued)

Concept	Definition
Transfer learning	With this technique the model designed for one task is reused or fine-tuned to perform a different, however related task, leveraging its pre-trained knowledge to improve learning efficiency and potentially achieve better performance with less data.
Transformer	It is a deep neural network architecture that utilizes self-attention to process sequential data (such as text, protein sequences and images) in a modular and scalable encoder-decoder architecture.
Transformer-based models	Deep learning models that employ the self-attention mechanism to handle data. Their structure utilizes encoder, decoder, or both parts of the transformer architecture.
Vision transformer (ViT)	Transformer architecture specifically designed for computer vision tasks.

Table 2 Overview of pre-trained biomedical language models

Study	Data sources	Model architecture	Biomedical tasks
BioBERT [3]	PubMed, PMC	BERT	Biomedical named entity recognition (NER), relation extraction (RE), and question answering (QA)
PubMedBERT [21]	PubMed	BERT	Biomedical NER, RE, QA, evidence-based medicine information extraction, document classification, and sentence similarity
BioMegatron [22]	PubMed	Megatron	Biomedical NER, RE, and QA
BioELECTRA [23]	PubMed	ELECTRA	Biomedical NER, RE, QA, evidence-based medicine information extraction, document classification, medical natural language inference, and sentence similarity
BioALBERT [24]	PubMed, PMC	ALBERT	Biomedical NER, RE, QA, evidence-based medicine information extraction, document classification, medical natural language inference, and sentence similarity
BioMed-RoBERTa [25]	PubMed, ChemProt	RoBERTa	Chemical relation classification
BioGPT [26]	PubMed	GPT-2	Biomedical RE, QA, document classification, and text generation
ClinicalBERT [27, 28]	MIMIC-III, i2b2 datasets	BERT	Identification of clinical entities, natural language inferencing
ClinicalXLNet [29]	MIMIC-III	XLNet	Identifying patient reports with prolonged mechanical ventilation and 90-day mortality
RoBERTa-MIMIC, ALBERT-MIMIC [30]	MIMIC-III, i2b2 datasets	RoBERTa, ALBERT	Identification of clinical entities
Clinical-Longformer, Clinical-BigBird [31]	MIMIC-III	Longformer, BigBird	Document classification, question answering, named entity recognition, natural language inference
GatorTron [32]	University of Florida Health, MIMIC-III, PubMed, Wikipedia	BERT, BioMegatron	Clinical concept extraction, medical relation extraction, semantic textual similarity, medical natural language inference, medical QA
Bioreddit-BERT [33]	Reddit health-related articles	BERT	Biomedical named entity recognition, adverse reaction mention detection
Bio-GottBERT [18]	German medical text	BERT	Identification of procedures, diagnoses, and medications
CamemBERT-bio [19]	French clinical documents	RoBERTa	Detection of clinical entities
KM-BERT [20]	Korean medical literature	BERT	Identification of diseases and treatment entities

downstream biomedical NLP tasks and achieved new state-of-the-art performances for named entity recognition (NER), question answering (QA), and relation extraction (RE). More studies have introduced various variants of biomedical pre-trained models using different transformer-based architectures such as ELECTRA [15], RoBERTa [16] and GPTs (GPT1, GPT2, GPT3) [2, 9, 17]. Furthermore, BERT variants have been pre-trained on

different types of biomedical corpora, see Table 2 for an overview.

Finally, different efforts have been made to develop language-specific transformer variants for biomedical texts in different regions of the world (Table 2). Some examples of these variants are Bio-GottBERT [18], CamemBERT-bio [19], KM-BERT [20] dedicated to languages like German, French, and Korean, respectively. Noteworthy, the main limitation of these efforts is often the limited availability of language-specific data.

Applications to document and topic classification

Document and topic classification are typical NLP downstream application tasks to which pre-trained transformer models have been applied in biomedicine: During the Coronavirus disease 2019 (COVID-19) pandemic, a new search engine LitCovid [34] was introduced by the United States National Library of Medicine (NLM), which provides an overview of the latest COVID-19 literature and allows users to filter the literature based on different categories such as case reports, mechanism, prevention, or diagnosis. The classification of the literature was done manually by the creators. However, in a later stage, various experiments with transformer-based models like BioBERT, PubMedBERT, and others showed high performance with an F_1-score of approx. 94% to automatically assign categories to new literature [35]. CO-Search is another example of a COVID-19 search engine that used a Siamese-BERT-based document retrieval engine with a strong evaluation performance [36]. Nentidis et al. [37] report results of a semantic indexing challenge in which the best participating system utilized BERT and BERTMeSH [38] models.

Applications to Named Entity Recognition (NER) and linking (NEL)

After identifying relevant documents for a certain topic, one is often interested in finding hidden but valuable biomedical concepts inside them. NER and named entity linking (NEL) tasks are specifically designed to extract these relevant concepts and link them to biological databases. Such concepts appear in various areas of biomedicine, ranging from molecular biology (genes, proteins, microRNAs, biological functions, and cellular components) to the clinical domain (medication/drug, adverse drug reactions, diagnoses, and diseases). For instance, the sentence "Apolipoprotein E: Structural Insights and Links to Alzheimer Disease Pathogenesis" (PMID:33,176,118) contains the mention of the protein *Apolipoprotein E*, the disease *Alzheimer disease*, and the biological process *Pathogenesis* that can be linked to Uniprot term *APOE_HUMAN (ID: P02649)*, disease ontology term *Alzheimer's disease (DOID:10,652)*, and National Cancer Institute Thesaurus (NCIT) term *Pathogenesis (NCIT: C18264)*, respectively. In the case of NER, the majority of studies have considered this task as a sequence labeling task (Table 1), in which they used BERT-based models to predict labels for each token in a sequence. Rather than a sequence labeling task, NER has also been formulated as a machine reading comprehension task (Table 1), which allows easy integration of prior knowledge into models [39].

Most authors fine-tune domain-specific transformer models, such as BioBERT, to detect one specific entity, for example drugs or genes [14]. However, multi-task learning strategies have also been proposed to detect chemical or disease mentions with one single model [40]. Some work has also been performed to capture complex cases of entities (such as discontinuous or overlapping entities) by Khandelwal et al. [41], where they combined BERT and GloVe embeddings with a new label-tagging schema to train an NER model in a distant supervision setting showing a significant performance boost in detection of disorder entities obtained from clinical free-text notes. Zaratiana et al. [42] have studied an integration of a BERT-based model with graph neural networks to create a span representation that can reduce the number of overlapping spans of disease mentions. They reported an F1 performance of 87.43%, however, the best F1-score reported on the used dataset is at 90.48%[1]. An overview of different studies employing transformers for NER and NEL is shown in Table 3.

Applications to relation extraction

Relation extraction, often performed after NER, is one of the main tasks in information extraction, which creates semantic links between two or more entities appearing in the text. These links, among others, can be loose (associates, interacts, correlates, etc.), quite specific (increase/decrease, binds, has participants, etc.), or even causal (directly increases, directly decreases, determined by) as defined by the relation ontology [49]. For instance, the sentence "STK38 is associated with PPARgamma" (PMID:34,670,478) contains a simple association relation between two proteins, whereas "Mitotic exit kinase Dbf2 directly phosphorylates chitin synthase Chs2" (PMID:27,086,703) describes a causal relation. The extracted relations from unstructured text are mostly used to construct biomedical knowledge graphs and expand existing ones with new knowledge [50, 51].

Transformer-based models have achieved remarkable success in extracting relations from textual content. Most

[1] https://paperswithcode.com/sota/named-entity-recognition-ner-on-ncbi-disease.

Table 3 Overview of biomedical named entity recognition studies employing transformer models

Study	Data sources	Model architecture	Biomedical tasks
[43]	Emergency department notes from Stanford Health Care	BioBERT-based model	Extraction of COVID-19 symptoms and risk factors
[44]	Mental state examinations from University Hospital Aachen	GermanBERT	Extraction of psychiatric symptoms
PLM-ICD [45]	MIMIC-II, MIMIC-III	BioBERT, ClinicalBERT, PubMed-BERT, RoBERTa-PM	Prediction of clinical coding of medical records
[46]	i2b2 corpora, Physionet corpus, Dernoncourt-Lee corpus	BERT, SciBERT, BioBERT	Deidentification of clinical records
[47]	EHRs, Stockholm EPR corpora,	Swedish BERT	Deidentification of clinical records
BERN2 [48]	PubMed	BioBERT	Recognition and normalization of genes/proteins, disease, drugs, species, and mutations

Table 4 Overview of biomedical relation extraction applications

Study	Data sources	Model architecture	Biomedical tasks
[64]	PubMed	BioPREP based on BioBERT	Detection of Xanthium compound-diabetes associations.
BioPrep [65]	SemMedDB	BioBERT, SciBERT	General biomedical predicate classification
[66]	Wikipedia and Mayo Clinic articles from DISNET	BioBERT	Creation of disease network
[67]	DrugTargetCommons, ChEMBL, DisGeNet, PubMed	SciBERT, BioBERT, BioMed-RoBERTa, BlueBERT	Prediction of drug-target interactions
KSM [68]	PubMed, BioCreative VI Track 4 PPI extraction task dataset	Multiple transformers with knowledge selector	Identification of protein-protein interactions
Patent_BERT [69]	CLEF 2020 - ChEMU Task data, Chemical Patents	BioBERT	Extraction of chemical reactions
[70]	Drug-adverse event corpus, PubMed	RoBERTa	Identification of drug-adverse effect relations
[71]	PubMed	BioBERT	Identification of plant-phenotype relations

of the studies have typically fine-tuned BERT-like models on subject-predicate-object relations of one dataset in a supervised manner. For instance, Zhu et al. [52] utilized BioBERT to extract drug-drug interactions from text with an overall F1 performance of 80.9% beating previous deep learning approaches. Other approaches for relation extraction involve multi-task learning, where multiple datasets are used for fine-tuning with the intuition that a model will learn a general representation of encoded relations that are of different types. To extract associations between drug-drug, chemical-protein, and medical diagnosis-treatment concepts, Moscato et al. [53] proposed a transformer-based architecture with multiple classification heads each designed to learn features for a specific type of relation. With their multi-task model, they could improve the performance by approx. 1.5% for chemical-protein and medical diagnosis-treatment associations. However, the model showed a decline of performance by 0.6% for drug-drug interactions in comparison to the single-task model. This showed that the effectiveness of multi-task learning can vary across different datasets. Solutions have also been proposed to simultaneously link entities and extract relations either by integrating multiple models in a pipeline manner [54, 55] or train a joint model responsible for extracting entities and relations at once [56–58]. Some have also experimented with datasets that were created either using distant supervision [59] or even without any supervision [60]. An overview about different studies employing transformers for relation extraction is shown in Table 4.

To get an even broader view of transformer-based models used in biomedical text mining - especially on tasks this work has not focused on - we refer to various surveys published by many researchers around the world [30, 61–63].

In summary, transformer-based models are well set in the biomedical NLP field. One main challenge is however the lack of clinical datasets due to privacy reasons, which hinders the development and evaluation of models specific to clinical settings. Another challenge is the limited diversity of datasets used in studies evaluating pre-trained models as they often focus on single entity types like disease and chemical mentions. More efforts to utilize and generate NLP datasets that cover a wider range

Table 5 Overview of biological sequencing analysis studies

Study	Data sources	Model architecture	Biomedical tasks
ProtTrans [75]	UniRef, UniParc, and Big Fantastic Database	BERT, T5, Transformer-XL, Albert, Electra, XLNet	Prediction of secondary structure and per-protein location and membrane prediction
ESM-1b Transformer [76]	UniParc	Transformer	Remote homology detection, prediction of secondary structure and tertiary contacts, prediction of mutational effects
ProteinBERT [77]	UniRef, Gene Ontology	BERT extended	Prediction of secondary structure, remote homology, fluorescence, and protein stability
MSA Transformer [96]	Multiple sequence alignments (MSA) based on UniRef	Modified transformer	Contact prediction, secondary structure prediction
ProtGPT2 [82]	UniRef	GPT2	Sequence generation, homology detection, disorder prediction,
SignalP 6.0 [97]	UniProt, PROSITE, TOPDB	ProtBERT + CRF	Detection of signal peptide types
Tranception [98]	UniProt, ProteinGym	Autoregressive transformer	Protein fitness prediction
[99]	UniProt, EVmutation	ESM-1b transformer, variational autoencoder and more	Protein fitness prediction
TMBed [100]	OPM, SignalP 6.0	ProtT5 + CNN	Prediction of transmembrane classes for each residue
ReLSO [101]	GIFFORD, GB1, GFP, TAPE	Transformer-based encoder	Designing new protein sequences
[102]	BIOSNAP, DAVIS, and BindingDB	ProtBERT + ChemBERTa	Prediction of drug-target interactions
STEP [103]	BIOSNAP [104]	Siamese ProtBERT	Prediction of protein-protein interactions

of biomedical entities and relations are required. Furthermore, the processing and analysis of longer biomedical texts still poses a challenge, which require sophisticated models. Newer models including LLaMa, BLOOM, and GPT4 implement techniques to cope with these challenges by enabling in-context learning and allowing to process longer texts. However, since these models are not specifically designed for the biomedical domain, thorough evaluation efforts are necessary to identify their advantages and limitations.

Biological sequences

Biological sequences, such as deoxyribonucleic acid (DNA), ribonucleic acid (RNA), or protein sequences, are relatively similar to natural languages. In the same way that characters in a natural language construct meaningful words, phrases, or sentences to convey some meaning, the building blocks of sequences arranged in different combinations form structures or support specific biological functions. It is not a surprise that the recent success of transformer-based models in NLP tasks also motivated the development of dedicated models to represent and analyze biological sequences. This trend is further supported by the availability of large databases (UniProt; [72], ENSEMBL [73], GenBank [74] containing vast amounts of biological sequences that can be used to perform pre-training of transformer-based models based on amino-acid as well as DNA sequences (Table 5).

Trained protein embeddings are, among other things, used to evaluate whether the prediction of per-residue secondary structure and subcellular localization shows a similar accuracy as methods that use evolutionary information [75], the recovery of proteins along the species or gene axes is possible, the biochemical properties (such as hydrophobic or aromatic nature) of amino acids can be recovered [76], or the retrieved embeddings generalize over different protein sequence lengths [77]. For instance, Elnaggar et al. [75] used the pre-trained ProtTrans models to predict the secondary structure labels (such as alpha-helix, beta-strand, or coil) for each amino acid, reaching state-of-the-art performance on multiple datasets. Although the majority of studies that develop these pre-trained models using protein sequences employ them for various downstream classification tasks, such models can also generate de novo protein sequences with the same fundamental characteristics as the natural ones [78–82]. ProtGPT2, a protein autoregressive pre-trained language model trained on 50 million sequences, is such a model that can predict subsequent amino acid sequences given a certain context (such as a number of amino acids as input) [82]. The generated protein sequences have shown properties of globular proteins

and preserved functional hotspots [82]. However, major limitations still exist as there is no way of anticipating the discovery of functional traits underlying new protein sequences, which would require costly high-throughput experimental approaches.

Recent studies have more systematically explored amino acid sequence representations learned by pre-trained transformers [83–85]. For example, the analysis by Detlefsen et al. [83] shows that pre-trained transformer models have difficulties separating details of a single protein family. In consequence, the authors propose fine-tuning (evo-tuning) on the respective protein family to increase their capacity to show clear phylogenetic separation. They also show that enforcing specific biological properties on representations is not a straightforward task and that it is currently steered by model architecture, specific preprocessing (such as using multiple sequence alignment) of underlying data, objective functions for the pre-training, and placing prior distributions on parts of the model to better mimic certain biological traits.

Fine-tuned transformer models for amino acid sequences have been used for various downstream tasks such as protein function classification, protein fitness prediction, and detection of protein interactions with chemical substances (Table 5). Furthermore, AlphaFold [86, 87] has achieved considerable improvements in the protein 3D structure prediction by using protein sequences as input. AlphaFold2 [87], builds upon two core deep learning-based modules, namely Evoformer and Structure modules, that has significantly improved the performance on the Critical Assessment of Protein Structure Prediction (CASP) 14 dataset by setting a new state-of-the-art [86, 87]. The transformer-based Evoformer module uses representation of multiple sequence alignment (MSA) and pairwise representation of protein sequence as input. The MSA, which is precomputed by conducting a search through sequence databases to find sequences that resemble the input protein sequence, informs the model about evolutionary conservation and variation. Whereas, the pairwise representation captures the interactions between pairs of amino acid residues, which is crucial for understanding the spatial geometry of protein. The Structure module uses these representations to construct an atomistic model of the protein's structure. It employs an additional attention mechanism and optimization procedure to ensure that the predicted 3D structure is physically plausible and adheres to known biophysical constraints. AlphaFold2 combines both modules to refine the representations and 3D structure prediction in an iterative process to produce the final structure. Like AlphaFold, the transformer-based models RoseTTAFold [88] and ESMFold [89] were independently developed to also predict accurate 3D-protein structures by learning patterns appearing in protein sequences.

Some recent studies have proposed to learn and capture global representations of DNA sequences [90, 91]. Ji et al. [91] pre-trained a DNABERT model, which is based on the BERT model with masked language modeling (MLM) objective and used tokenized k-mer (with best k at 6) sequences as input instead of regarding each nucleotide as a single token. Due to the specific tokenization, the vocabulary size of DNABERT was set to 4^k+5 (using permutations of 4 nucleotides with additional 5 special tokens such as for separator and padding). The pre-trained DNABERT model was analyzed using various fine-tuning tasks showing particularly that it can effectively identify proximal and core promoter regions, transcription factor binding sites, and functional genetic variants. Furthermore, DNABERT can also be used for interpretability by using learned attention weighting that characterizes the contextual relationships within a sequence to visualize its important regions and motifs.

Another relevant biological problem of how non-coding DNA regions influence gene expression in cells has been analyzed by Avsec et al. [92], who propose a transformer-based architecture called Enformer that enables the integration of long-range interactions in the genome producing significant improvements in predicting tissue and cell-type-specific gene expression. Similar to Kelley et al. [93], to read long sequences with a size of around 197.000 base pairs, the Enformer uses a number of convolutional layers that perform convolution on input sequences to reduce the spatial dimensionality. After the convolutional layers, instead of dilated convolution as used by Kelley et al. [93], the Enformer implements transformer layers that use attention mechanisms to represent the long-range interactions. The Enformer has shown significant performance gains in gene expression prediction; however, it has not yet reached the accuracies of experimental approaches. Furthermore, Enformer has also shown improvements in variant effect prediction that was performed on expression quantitative trait loci (eQTL) data [92].

In summary, studies on pre-trained transformer-based models for biological sequences have highlighted their capabilities to produce state-of-the-art results on 3D structures, functions, and interactions prediction. These sequence models however have similar limitations to NLP models. They require huge amounts of training data, which can represent a bottleneck for certain sequences (such as small RNAs). Additionally, these models often struggle to capture long-range interactions due to fixed-length context windows, which can be crucial in biological sequences. In the case of protein structure prediction, AlphaFold and others are highly accurate in predicting

single protein chains; they however lack the ability to generate precise multi-chain protein complexes. Newer studies such as AlphaFold-Multimer [94] and ESMPair [95] have extended the previous models to also predict accurate protein complex structures. While transformer-based models show the ability to generalize on biological sequences data, further research is required to identify additional methods (for e.g. layers or architecture) to overcome the aforementioned limitations.

Structured-longitudinal electronic health records

Electronic health records (EHRs) are now routinely and in vast quantities collected by many healthcare systems. Typically, they contain unstructured information like clinical notes but also structured data, including time-stamped diagnosis and medication codes as well as time-stamped codes for medical procedures. The latter provide excellent opportunities for the efficient development of machine learning models for better personalized healthcare. However, it is difficult to utilize such data due to high dimensionality, heterogeneity, temporal dependency, sparsity, and irregularity [105]. More specifically, structured EHRs can be regarded as an instance of multivariate discrete, irregular time series data.

Several studies have recently proposed transformer-based models for the analysis of structured EHR data. The intuition behind these approaches is that sequences of diagnosis, procedure and medication codes might be interpreted as a kind of language, in which codes recorded at one particular visit might be viewed as tokens. Accordingly, transformers have been pre-trained on large amounts of patient data to generate numeric representations of a patient's medical history, which are then used for downstream tasks like medication recommendation or mortality prediction. For example, Shang et al. [106] developed the graph-augmented transformer model G-BERT. It uses the hierarchical information from the International Statistical Classification of Diseases and Related Health Problems (ICD) and Anatomical Therapeutic Chemical (ATC) ontologies to train a graph neural network, which encodes in a first step diagnosis and medication codes in a lower dimensional space. In a second step, corresponding concept embeddings are used as a modified position encoding in a BERT-like transformer architecture. The authors pre-trained their model on 20,000 patients from the MIMIC-III dataset, then applied it to a medication recommendation task and found it slightly superior to baseline techniques (1.06% gain in AUPR to the second-best approach).

Later, Li et al. [107] developed BERT for EHR (BEHRT), which uses an altered embedding layer to process a sequence of diagnosis codes. Unlike G-BERT, the model provides a patient representation for the entire medical history rather than each visit. When applied to a diagnosis code prediction task, BEHRT surpassed baseline methods (1.2–1.5% higher area under the receiver operating characteristic curve (AUROC) and 8.0-10.8% increased area under the precision-recall curve (AUPR) for the disease prediction task). Since BEHRT – like many other transformer-based models – is restricted to a maximum sequence length of 512 codes, Li et al. [108] devised a hierarchical BEHRT (HI-BEHRT) variant in a subsequent study. This method applies BEHRT to parts of the medical history using a sliding window separately before aggregating the information by forwarding the individual representations to a final transformer. In addition to the hierarchical modification, the authors included information on medications, procedures, and laboratory tests. In disease prediction tasks, it was discovered that HI-BEHRT outperforms BEHRT by 1–5% and 3–6%, respectively, in terms of AUROC and AUPR. Another variant is the Med-BERT model [109]. Compared to G-BERT and BEHRT, it employs a more extensive vocabulary of diagnosis codes. Furthermore, it introduces a new training objective called prediction of prolonged length of stay (LOS). During pre-training, the model predicts whether patients had hospital visits of seven or more days (LOS > 7 days) for their entire EHR sequences. After pre-training on data from 28 million patients, the model was applied to a disease prediction task. On three datasets originating from two clinical databases, the AUROC performance was increased by 1.21–6.14% compared to the baseline approaches. Very recent work further extended Med-BERT by adding demographic information, medications as well as quantitative lab measurements [110].

Other studies addressed the potential shortcomings of the approaches mentioned above. For instance, Pang et al. [111] proposed CEHR-BERT that, unlike Med-BERT and BEHRT, employs a different method to embed the time-series data before passing it to the transformer layers. It uses embeddings initialized with time2vec model [112] to encode the relative time between visits and the patient's age. The age, time, and concept embeddings are concatenated and passed through a fully connected layer to generate the BERT architecture's temporal concept embeddings. In addition, it incorporates a new pre-training task called visit type prediction (VTP) alongside MLM. This task requires the model to determine if the visit was inpatient, outpatient, emergency, or masked. Compared to baseline approaches, including the retrained versions of BEHRT and Med-BERT, CEHR-BERT increased AUPRs and AUROCs by 0.6–4.2% and 0.4–2.51%, respectively. The aspect of appropriate time encoding was also covered in several other studies [113–117].

In contrast, Agarwal et al. [118] based their Transmed approach on the notion of a hierarchical transformer for EHRs. On the one hand, a static context encoder was employed to handle information such as a patient's age, sex, race, and prior conditions such as diabetes or smoking. On the other hand, temporal context encoders were used to process the information at individual visits. The aggregated representations from the static and temporal encoders are then used to predict a patient's risk of hospital stay or mechanical ventilation following a COVID-19 diagnosis. Across all four tasks, Transmed outperformed a newly pre-trained version of BEHRT (11–20% higher AUROC) and was mostly on par or better than a baseline gated recurrent unit (GRU) model.

There have also been attempts to combine structured and unstructured EHR data into joint patient representations. For instance, the Bidirectional Representation Learning model with Transformer architecture on Multimodal EHR (BRLTM) [119] utilizes diagnosis, drug, and procedure codes as well as information derived from unstructured clinical notes via latent Dirichlet allocation (LDA). When the authors compared BRLTM to other models, including a retrained version of BEHRT, they discovered that it was superior at accurately predicting diseases over multiple time frames. Liu et al. [120] followed a different approach with their Med-PLM model. Instead of deriving features from clinical notes, they use ClinicalBERT for processing clinical notes and a G-BERT-like model for processing structured EHR data before combining both using a cross-attentional module. The authors found that the final model outperformed unimodal counterparts (e.g., ClinicalBERT or G-BERT) in all tasks, highlighting the potential of merging both data modalities. Similarly, Darabi et al. [113] used both data modalities for their TAPER model and reached comparable results.

Another recent development in the context of EHRs is the synthetic generation of EHRs with transformers. Cheng et al. [121] recently proposed CEHR-GPT, a model that builds upon their previous work on CEHR-BERT to generate synthetic EHRs using GPT. Unlike CEHR-BERT, CEHR-GPT includes additional information on demographics, patient history, and temporal dependencies. Each visit is represented by a visit type token (VTT), and time is encoded using artificial time tokens (ATT) and long-term (LT) tokens. In their experiments, the authors compared three different patient representations of GPT. They found that CEHR-GPT was the most suitable variant for generating realistic synthetic EHR data while preserving patient privacy and temporal dependencies. However, they reported that the prevalence of concepts in the generated data was skewed compared to the original data and that the representation of time intervals is currently limited, suggesting that further improvements could be made in training the model and the representation of EHR data.

A broad overview of different studies employing transformers for structured-longitudinal EHR analysis is shown in Table 6.

In summary, transformer-based models are promising for working with structured EHR data. However, applying these models to EHR data also presents several challenges. Firstly, EHR data is highly heterogeneous and diverse, making it relatively unclear how to best represent it compared to text, sequence, and image data. Many studies focus on finding a suitable data representation. In addition to this challenge, comparing these models and their results is difficult. Since most pre-trained models and datasets are publicly unavailable due to privacy concerns, a direct comparison of the models is often impossible. Although studies often use other models as baselines and perform pre-training on available data to compare model architectures, a direct comparison of initially pre-trained models is not feasible, as is common in the NLP field. Furthermore, Kumar et al. [122] point out that simple linear models could not only be data and computationally efficient but could achieve comparable performance to transformer-based models. For instance, they propose an attention-free architecture called SANSformer that outperformed BEHRT and BRLTM models. Despite these challenges, transformer-based models remain a promising tool for analyzing EHR data. Further research is important to understand their full potential as well as limitations and how they can improve patient outcomes and provide better decision support.

Biomedical images

Due to the self-attention mechanism employed in transformer-based models, they have shown superior ability to model long-term dependencies in data, however, mostly in cases where the data is of sequential nature. Recently, transformers have also been adapted successfully to a wide variety of image analysis cases. For the purpose of image analysis, the image is first split into a sequence of patches (regions), which are then flattened to fixed vector length - quite similar to tokens. The flattened image patches are then linearly projected and combined with their positional embeddings that provide spatial information on each patch. The sequence of transformed patches can then be fed to a transformer. This approach is referred to as a vision transformer (ViT) in the literature [123–126]. Dosovitskiy et al. [124] formulated image classification as a sequence prediction task, which he addressed via a ViT. They examined two approaches for aggregating spatial information from images: the use of a CLS token and global pooling [124]. The CLS token in

Table 6 Summary of models performing structured-longitudinal EHR analysis

Study	Data sources	Model architecture	Biomedical tasks
BEHRT [107]	CPRD	Transformer-based encoder	Disease prediction
Hi-BEHRT [108]	CRPD	Hierarchical BEHRT	Disease prediction
G-BERT [106]	MIMIC-III	Graph neural network and BERT	Drug recommendation
BRLTM [119]	UCLA EHR data	Transformer-based encoder	Disease prediction
Med-BERT [109]	Cerner Health Facts , Truven Health MarketScan	Transformer-based encoder	Disease prediction
ExMed-BERT [110]	IBM Explorys Therapeutic Dataset	Extended Med-BERT	Disease prediction
CEHR-BERT [111]	CUIMC-NYP OMOP	Transformer-based encoder with additional FFN for temporal embedding	Various predictive tasks (disease, readmission, death, hospitalization)
Med-PLM [120]	MIMIC-III	G-BERT / Med-BERT + ClinicalBERT + Cross-modal module	ICD coding, readmission, drug recommendation
TransMED [118]	STARR OMOP	Hierarchical use of BERT	Hospital stay, ventilation risk
T^3Net [114]	KPMAS	Transformer-based encoder	Hospitalization and mortality prediction
TAPER [113]	MIMIC-III	Transformer-based encoder, BERT	Readmission and mortality prediction
CEHR-GPT [121]	EHR data from the Columbia University Irving Medical Center-New York Presbyterian Hospital	Transformer-based decoder, GPT	Generation of synthetic EHR data

ViTs aggregates global information through self-attention, dynamically adjusting to capture complex image relationships. Global pooling, including methods like global average and max pooling, simplifies feature aggregation by applying straightforward mathematical operations across all image patches. While the CLS token's aggregation is learnable and adapts to task specifics, global pooling offers a more generalized and computationally efficient summary [124].

ViTs have been applied to medical images derived from imaging techniques such as X-ray, computer tomography (CT), MRI, ultrasonography, optical coherence tomography (OCT), and high-content cell imaging screens. For instance, ViTs were used to analyze lung X-rays to detect COVID-19 disease [127–129], breast sonography images to classify breast cancer [130, 131], or femur X-rays to check for fractures [132]. Chen et al. [133] have proposed a ViT to detect gastric cancer from histopathological imaging data. Furthermore, CT images were used by Wu et al. [134] to build a medical application for classification of emphysema that can be further divided into three different subtypes, whereas Wang et al. [135] screened rare medical OCT imaging dataset for lesions associated with genitourinary syndrome of menopause. Nonetheless, MRI datasets have also been classified using ViTs for brain tumors [136] or for intraductal papillary mucosal neoplasms in the pancreas by [137]. Upon closer examination of the work of Tanzi et al. [132], you can observe potential benefits of ViT architectures compared to conventional approaches. Based on their results, it seems worth exploring the superiority of embedding space representations generated by ViTs, which can boost performance for medical classification tasks. Examining attention layers, that are commonly part of ViT architectures, makes these models inherently explainable, an attribute highly regarded by clinicians for model evaluation. Lastly, their retrospective analysis of integration into clinical practice, allows for the conclusion that a ViT-based computer aided diagnosis (CAD) system can contribute to improving clinical workflows and decision making for young residents and experienced doctors alike.

Another relevant task in the biomedical computer vision field is to detect segments of object instances such as lesions in functional magnetic resonance images, tumors in histopathological images, brain tissues in magnetic resonance images, retinal vessels in fundus imagery, or single-cell information from microscopy imagery [138, 139]. Transformer-based models are being heavily used for segmentation as they often improve accuracy compared to the traditional convolutional neural network-based (CNN) methods. Although most studies use

Table 7 Summary of transformer models for biomedical image analysis

Study	Data sources	Model architecture	Biomedical tasks
[127–129]	Lung X-rays	Vision transformer	Detection of COVID-19
[130, 131]	Breast sonography images	Vision transformer	Classification of breast cancer
[132]	Femur X-rays	Vision transformer	Detection of fractures
GasHis-Transformer [133]	Histopathological imaging	Vision transformer	Detection of gastric cancer
[134]	CT imaging	Vision transformer	Classification of emphysema and their subtypes
ViT-P [135]	OCT imaging	Vision transformer	Detection of lesions associated with genitourinary syndrome of menopause
[136]	MRI dataset	Vision transformer	Detection of brain tumors
[137]	MRI dataset	Vision transformer	Classification of intraductal papillary mucosal neoplasms in the pancreas
Cell-DETR [139]	Live-cell microscopy dataset	Based on transformer encoder-decoder	Segmentation of yeast cells in microstructures
[143]	Stained breast cancer cell images	Vision Transformer	Mechanism of action prediction of cells treated with compounds
[144]	High-content imaging screen	Vision Transformer	Classification of cell phenotypes

hybrid transformer architectures, some have also built pure transformer-based models. For instance, Gao et al. [140] have proposed a hybrid transformer-based architecture UTNet by integrating a complexity-reduced self-attention into a CNN for segmentation. In comparison Huang et al. [141] introduced the pure transformer-based method MISSFormer, optimized especially for medical image segmentation tasks. Most studies have focused on the medical field, but some have also applied transformer-based methods for segmenting cells in images that originated in in-vitro experiments. Prangemeier et al. [139] have proposed a cell detection transformer for direct end-to-end instance segmentation, reaching a similar accuracy as the CNN-based methods while showing the simplicity and improved runtime of the proposed model.

In the drug discovery field, it is common nowadays to perform an automated high-content screening of cells treated with specific chemical substances. These screening experiments might identify substances that have desirable effects on the phenotypes of cells. High-content images of cells are also used for image-based profiling, where the profiles are derived by extracting relevant features from screened images [142]. Such phenotypic profiles can be used in downstream applications such as identifying a disease-associated phenotype, identifying lead compounds, bioactivity and toxicity assessment, and detecting a compound's mechanism of actions [142], where recently transformer-based models are being applied to [143, 144]. For instance, Cross-Zamirski et al. [143] proposed a ViT-based model that uses weak labels to learn phenotypic representations from a publicly available dataset containing high-content images of cells and evaluate the model on two mechanism-of-action classification tasks. Furthermore, the authors show that the representations are biologically meaningful by analyzing the attention maps. Table 7 provides a broader overview of recent applications of ViTs.

Even though ViTs have proven to be powerful architectures for a variety of problems in biomedical imaging, they can not be recommended unlimitedly in favor of more established computer vision models, e.g. convolutional neural networks (CNNs) [145, 146]. It is important to understand how both architectures "perceive" images, in order to understand its particularities, advantages and disadvantages. The architecture of convolutional networks is inspired by the visual cortex of the brain [147]. They use receptive fields to learn kernels enabling them to recognize features crucial to their task. A subsequent pooling operation relatively increases the receptive fields of the kernels. This process is repeated iteratively, so the kernels can interpret more distant areas of the image [148]. This, by design, creates inductive biases, like translation equivariance and locality [124], important properties for image classification.

In contrast, as described earlier, a ViT treats an image as a sequence of patches, and through self-attention, every patch of the sequence attends to every other patch, so it needs to learn all spatial relations from data training [124]. Essentially, this causes ViT to struggle in effectively generalizing with limited data [124]. However, their performance scales well with growing datasets, outperforming CNNs as the number of training samples increases. Unfortunately, especially in the biomedical domain, large publicly available datasets are scarce.

Nonetheless recent work by He et al. [149] has shown that pre-training techniques, such as training a masked autoencoder (MAE) for patch embeddings, can reduce the number of training samples, training times, and boost performance in natural images. Zhou et al. [150] later showed that this can be applied in the medical domain as well. Varma et al. [151] tackled the issue of ViTs relying on predefined image sizes, necessitating pre-processing steps that can degrade image information. Through their flexible positional embedding and alternate batching strategies, they can reduce image manipulation while maintaining fine-grained image features.

Driven by its popularity and constant developments through ongoing research, one can assume that ViT architectures will increase in value and impact for biomedical imaging tasks in the near future.

Biomedical graphs

Besides textual content, biological sequences, imaging data, and structured EHR data, graphs are frequently used in biomedicine to describe relations between concepts. Graphs can cover various aspects of life sciences, hence, they can connect different types of nodes and edges with each other. Graph representation learning with machine learning methods enables the usage of graphs for various biomedically relevant downstream tasks such as protein-protein interaction prediction, prediction of adverse drug reactions, cell-type-association prediction, disease-subgraph classification drug-interaction prediction, patient-treatment prediction [152]. These tasks can be modeled as graph or sub-graph classification, node classification, or link prediction, which are often performed by encoding the information included in graphs, such as the graph structure, local graph neighborhoods, and the distinguishing features of nodes and edges [152].

Graph Transformer [153], Graph Transformer Networks [154], GTransformer [155], Structured Transformer [156], GraphFormers [157], and Relphormer [158] are some adaptations of transformer-based models suitable for graph representation learning. Transformers for graphs are conceptually similar to relational graph attention networks (RGATs) [159]. They regard each node of a graph as an entity in a pseudo-sequence. However, unlike transformers for sequences, the attention is restricted to neighboring nodes, hence taking into account the graph topology.

Graph-based transformers have, for instance, been applied in the drug discovery field where the focus lies on the identification of targets [160, 161], prediction of response [162], prediction of ATC code [163], or adverse reactions for a certain drug [164]. Additional work has also been performed to predict the properties of molecules involving toxicity, carcinogenicity, or blood-brain barrier penetration [165–167]. Recently, also textual and image analysis tasks (such as relation extraction or deformable image registration) have been successfully pursued using graph transformers [168, 169]. A further example is the prediction of interactions between transcription factors and DNA, which can be formulated as a link prediction task in a bipartite graph [170]. Other authors have delved into engineering new proteins by generative graph representations of 3D protein structures [156, 171]. Also, the prediction of protein-protein interactions has been performed using graph neural networks while using protein 3D structure graphs and learned sequence embeddings of ProtBERT [172].

Although the analysis or usage of graphs to support biomedical tasks with graph transformers is yet focused only on some niche areas, researchers are already prospecting new fields, such as the analysis of single-cell multi-omics in immuno-oncology to characterize cellular heterogeneity [173], where this technology could also be helpful. Nonetheless, further experiments are required to assess whether the transfer learning strategies, along with graph transformers, will ultimately prevail over the general graph neural networks like RGATs.

Transformers for multimodal data

The majority of existing research studies have addressed biomedical tasks using just one single data modality, however, modeling complex processes of biology and medicine inherently requires the integration of and learning on multiple modalities, such as genetic, proteomics, pharmacogenomic, imaging, and text [174]. Recently, transformer-based models have been adapted to process multiple data modalities simultaneously. Koorathota et al. [175] introduced a multimodal neurophysiological transformer for recognizing emotions using multiple modalities (such as time series and extracted features) obtained through electroencephalography, galvanic skin response, and photoplethysmogram techniques. Inspired by the multisensory integration mechanism of the brain, Shi et al. [176] proposed an adapted transformer-based model to integrate visual and auditory modalities to improve emotion and bird species recognition using video-audio clips.

Furthermore, vision-and-language models are a recent development, which take textual content and images as input and jointly learn to capture the relationships between both modalities. These models have also been adapted to the clinical domain, for instance, for chest X-ray disease diagnosis [177] or to automatically generate reports for abnormal COVID-19 chest CT scans [178]. Similarly, paired images and textual reports of chest and musculoskeletal X-rays were used with contrastive

learning to build new pre-trained models that improved upon medical image classification and retrieval on various datasets [179]. Others have explored an integration of molecular structures using simplified molecular-input line-entry system (SMILES) signatures in biomedical text to build a transformer-based multimodal system that can predict molecular properties, classify chemical reactions, and improve NER as well as relation extraction [180]. Finally, Lentzen et al. [110] proposed a multimodal transformer architecture to combine structured EHRs with quantitative clinical measures. Their idea was a concatenation of the latent representation learned by the transformer encoder with a feature vector representing quantitative data. The concatenated representations are then passed forward through the classification head during the fine-tuning phase.

Development of transformer-based models capable of learning from multimodal data is a non-trivial challenge. These models are highly specific to the particular modalities (for e.g. text, image, or structured EHR) and tasks at hand. There is a pressing need for further investigation into how transformer-based architectures can evolve into universal architectures that are agnostic to various biomedical modalities and the underlying tasks.

Making transformers explainable

Specifically in biomedicine, it is essential to study which features a model used to make predictions in order to identify potential flaws and build trust in the results. Since the first appearance of transformer-based models, several studies have proposed different approaches for post-hoc model explanation based on techniques developed in the booming field of Explainable AI (XAI) [181].

Most approaches focus on the implicitly learned attention weights of transformer-based models. For instance, Vig [182] developed the BertViz tool for displaying the attention weights for analysis and debugging purposes. Later, Ji et al. [91] used similar visualizations for their pre-trained DNABERT model. Following the evaluation, the authors studied the attention landscape and found, for instance, that the model prioritizes intronic sequence sections when predicting splice sites. Similarly, Avsec et al. [92] investigated their model, which was built to predict gene expression and chromatin states using the average attention weights. Their analysis revealed that the developed model attends to parts of the sequence located up to 100 kb from the gene site. A slightly different strategy was followed by Koorathota et al. [175], who proposed a multimodal neurophysiological transformer for predicting valence and arousal as a response to music. They created a metric known as the sum of absolute activation differences to interpret the interactions between the different modalities. Unlike the majority of attention-weight analyses, this analysis is neither affected by individual samples nor the selection of attention layers or heads. The study revealed that, for instance, electroencephalography and photoplethysmogram signals significantly affect the model's prediction.

Other studies investigate the application of general-purpose XAI methodologies. For instance, Kokalj et al. [183] introduced TransShap, an adaptation of Shapley Additive Explanations (SHAP) [184] that may be utilized to evaluate and understand the functioning of text classifiers. Lastly, Madan et al. [103] applied the integrated gradients method [185] instead of focusing on the attention weights of the model. The authors utilized this method to explain their predictions on virus-host protein-protein interactions while discovering sections of sequences that contribute to the model's predictions. Advances in the field of XAI methods, in general, have opened up new opportunities to interpret the models while gaining new insights on predictions, although significant limitations still exist due to the lack of validation datasets, hence, careful investigation of the reliability of these XAI strategies is highly necessary [186]. Furthermore, a general caveat is the possible misinterpretation of XAI approaches as providing a causal understanding of the prediction problem.

Discussion

Strengths of transformers

Transformer-based models have pushed the boundaries for processing and analyzing various data modalities such as text, EHRs, biological sequences, images, and graphs across a wide variety of biomedical tasks, as demonstrated by the examples shown in the previous sections. Since transformers originated in the NLP field, the biomedical NLP has seen a certain momentum with these models earlier than other disciplines, resulting in a greater number of transformer-related research studies within the NLP field. At the moment, transformers have mainly been applied to discrete data, but also first adaptations to continuous time series data have been proposed [187].

The success of the transformer can be mainly explained by two factors:

a) the attention mechanism, which allows for capturing long-range dependencies in the input.
b) the self-supervised learning paradigm that supports pre-training from huge amounts of unlabeled data and subsequent fine-tuning / transfer-learning of a domain-specific task.

Specifically, the second aspect allows for effective utilization of background information, which explains

the often-observed superior prediction performance compared to more conventional machine learning approaches.

Challenges when using transformers

The pre-training of transformers using the self-supervised learning paradigm depends on huge training datasets. Accordingly, the training of transformers is computationally intensive. It should be noted that transformers have millions of parameters (one of the largest model PaLM published by Chowdhery et al. [188] has 540 billion parameters), and the underlying attention mechanism has a quadratic time complexity with regard to the input sequence length. To overcome these challenges, new solutions have been proposed such as optimizing the transformer model [189–193] or applying knowledge distillation technique [194]. For instance, Kitaev et al. [191] proposed the Reformer model that improves the efficiency of the transformer by reducing the complexity of the dot-product attention mechanism and by optimizing the storage of activations in the model.

Future direction 1: knowledge integration

Another line of research focuses on the utilization of background knowledge during the training procedure. For example, in the NLP field K-BERT is an extension of BERT, in which the input token stream is expanded by background information extracted from a knowledge graph [195]. ERNIE uses two encoders, a T-encoder for the original tokens and a K-encoder for entities in a knowledge graph, and both representations are fused [196, 197]. While the authors of these papers report enhanced prediction performances of NLP-related tasks outside the biomedical domain, there is the question of how according methods might be impacted by incompleteness and errors in the knowledge graph, which could be a major concern in the biomedical field. Furthermore, not all knowledge can be effectively represented as a graph. Depending on the respective application, other knowledge representations, such as logical rules and mathematical equations, could be worthwhile to consider in future research as well.

Future direction 2: multimodal data integration

Integrating multimodal data is key for many systems and precision medicine tasks. Heterogeneous information across different data modalities, such as genetics, epigenetics, proteomics, metabolomics, imaging, text, and clinical observations, must be aligned and fused to perform multimodal learning with transformer-based models. Although first publications are now focusing on multimodal transformers (see section above), this line of research is still at the beginning. For example, one general challenge in the area of multimodal data integration are varying dimensions and numerical ranges of input modalities [198]. Recent studies have begun to explore general-purpose architectures that can handle different modalities of varying dimensionalities [199–201], but we expect more work to come along those lines.

Future direction 3: generative modeling

More recently, generative transformer models have shown impressive advancements in the NLP field. One of the most prominent examples, which is however not particularly devoted to biomedicine, is ChatGPT [202]. ChatGPT has shown remarkable performances on generating near-human level textual content and leading dialogues with humans. Generative transformer models such as ChatGPT or its freely available variants (e.g., GPT4All; [203]) could in the future support many tasks in medical routine, such as generating synthetic clinical notes [32], writing discharge letters, or coding and billing diagnosis and medications. Furthermore, these models could also support the field of biomedical research. Researchers have already started experimenting with generative transformers to generate synthetic protein sequences [82, 204]. However, a huge challenge of applying such models in biomedicine is to verify the trustworthiness of the generated content. For instance, engineered protein sequences need to be experimentally tested. Automatically generated discharge letters have to be validated manually.

Future direction 4: better explainable models

By being able to explain and understand the predictions through XAI techniques, trust and confidence can be built in biomedical AI models, which is even more relevant for decision-making processes in the clinical domain. Several general-purpose XAI techniques have been adapted for transformers recently [205–207]. Some have shown that trust in models can also be increased by producing counterfactual explanations that show under which hypothetical changes to the input a different output will be generated, a method often used by humans to understand unfamiliar processes [208, 209]. However, the XAI field as such is still in its infancy. For example, there is no generally accepted definition of "explainability", and there is a lack of gold standards against which new methods could be compared. Accordingly, existing attempts to make transformers explainable have to be seen relative to the advances of the XAI field as a whole. While first approaches in the XAI field mainly focused on images, the development of general-purpose model explanation techniques, such as SHAP is still relatively recent. We can thus expect that with the increasing advances of the XAI

field also better explanation techniques for transformers will become available.

Conclusion

Transformers, originally created in the NLP field, are still a relatively new deep learning approach. Recent years have witnessed a dramatically increased use for various data types with transformers, which are of relevance in biomedicine, including structured EHRs, graphs, images, and biological sequences. The main strengths of transformers are the in-built attention mechanism and the possibility for self-supervised pre-training, which, however, requires huge datasets. Accordingly, transformers have currently found little use in domains where such datasets are not available, e.g., signals coming from wearable devices, or clinical studies and registries. Also, despite research on modeling time-series data with transformers [187], dedicated studies in biomedicine for this type of data are yet to emerge. Currently emerging directions of research include better strategies for knowledge integration, multimodal data fusion, and the adaptation of novel XAI techniques. We expect that efforts to integrate data across the entire healthcare system, such as those in United Kingdom (UK) like Health Data Research UK (https://www.hdruk.ac.uk/), UK Biobank (https://www.ukbiobank.ac.uk/) and Genomics England (https://www.genomicsengland.co.uk/), will enable an even more wide-spread use of transformers in the future.

Abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CAD	Computer aided diagnosis
COVID-19	Coronavirus Disease 2019
CV	Computer Vision
CT	Computer Tomography
EHR	Electronic Health Records
GPT	Generative pre-trained transformers
MRI	Magnetic Resonance Imaging
NER	Named Entity Recognition
NEL	Named Entity Linking
NLP	Natural Language Processing
NLM	United States National Library of Medicine
PMC	PubMed Central
ViT	Vision Transformers
ICD	International Statistical Classification of Diseases and Related Health Problems
ATC	Anatomical Therapeutic Chemical
VTP	Visit Type Prediction
GRU	Gated Recurrent Unit
LDA	Latent Dirichlet Allocation
SHAP	Shapley Additive Explanations
XAI	Explainable Artificial Intelligence

Acknowledgements

Not applicable.

Authors' contributions

Sumit Madan: Conceptualization, Methodology, Investigation, Visualization, Writing - Original Draft, and Writing - Review & Editing; Manuel Lentzen: Investigation, Writing - Original Draft, and Writing - Review & Editing; Johannes Brandt: Writing - Review & Editing; Daniel Rueckert: Writing - Review & Editing; Martin Hofmann-Apitius: Conceptualization, Supervision, and Writing - Review & Editing; Holger Fröhlich: Conceptualization, Methodology, Supervision, Writing - Original Draft, and Writing - Review & Editing.

Funding

Open Access funding enabled and organized by Projekt DEAL. Research reported in this publication was supported by Integration of Heterogeneous Data and Evidence towards Regulatory and HTA Acceptance (IDERHA), an Innovative Health Initiative (IHI) Joint Undertaking (JU) under grant agreement No 101112135. The JU receives support from the European Union's Horizon Europe research and innovation programme, and life science industries represented by COCIR, EFPIA / Vaccines Europe, EuropaBio and MedTech Europe. Views and opinions expressed in this paper are those of the author(s) only and do not necessarily reflect those of the aforementioned parties. Neither of the aforementioned parties can be held responsible for them.

Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 18 June 2023 Accepted: 8 July 2024
Published online: 29 July 2024

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems. Red Hook, NY, USA: Curran Associates Inc; 2017. p. 6000–10.
2. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.
3. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019. p. 4171–86.
4. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. 2023.
5. Touvron H, Martin L, Stone K, et al. Llama 2. Open foundation and fine-tuned chat models. 2023.
6. Workshop B, Scao TL, Fan A et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. 2023. https://doi.org/10.48550/arXiv.2211.05100.
7. Bahdanau D, Cho KH, Bengio Y. Neural machine translation by jointly learning to align and translate. San Diego: 3rd International Conference on Learning Representations, ICLR 2015; 2015.
8. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A. Language models are few-shot learners. arXiv. 2020;2005:14165.
9. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog. 2019;1:9.
10. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, Massachusetts: The MIT Press; 2016.
11. Lin T, Wang Y, Liu X, Qiu X. A survey of transformers. 2021. https://doi.org/10.48550/arXiv.2106.04554.

12. Johnson A, Pollard T, Mark R. MIMIC-III clinical database. 2015. https://doi.org/10.13026/C2XW26.
13. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.
14. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36:1234–40.
15. Clark K, Luong M-T, Le QV, Manning CD. Electra: pre-training text encoders as discriminators rather than generators. arXiv. 2020;2003:10555.
16. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: a robustly optimized bert pretraining approach. arXiv. 2019;1907:11692.
17. OpenAI, Achiam J, Adler S et al. GPT–4 Technical Report. 2024. https://doi.org/10.48550/arXiv.2303.08774.
18. Lentzen M, Madan S, Lage-Rupprecht V, et al. Critical assessment of transformer-based AI models for German clinical notes. JAMIA Open. 2022;5:ooac087.
19. Copara Zea JL, Knafou JDM, Naderi N, Moro C, Ruch P, Teodoro D. Contextualized French language models for biomedical named entity recognition. Actes de la 6e conférence conjointe Journées d'Études sur la parole (JEP, 33e édition), Traitement Automatique Des Langues Naturelles (TALN, 27e édition), Rencontre Des Étudiants chercheurs en Informatique pour le Traitement Automatique Des Langues (RÉCITAL, 22e édition). Nancy, France: ATALA et AFCP: Atelier DÉfi Fouille de Textes; 2020. p. 36–48.
20. Kim Y, Kim J-H, Lee JM, Jang MJ, Yum YJ, Kim S, Shin U, Kim Y-M, Joo HJ, Song S. A pre-trained BERT for Korean medical natural language processing. Sci Rep. 2022;12:13847.
21. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-specific language model pretraining for biomedical natural language processing. 2020.
22. Shin HC, Zhang Y, Bakhturina E, Puri R, Patwary M, Shoeybi M, Mani R. BioMegatron: larger biomedical domain language model. In: Proceedings of the 2020 conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 4700–6. https://doi.org/10.18653/v1/2020.emnlp-main.379.
23. Kanakarajan Kraj, Kundumani B, Sankarasubbu M. BioELECTRA: pre-trained biomedical text encoder using discriminators. In: proceedings of the 20th workshop on biomedical language processing. Online: Association for Computational Linguistics; 2021. p. 143–54.
24. Naseem U, Dunn AG, Khushi M, Kim J. Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT. BMC Bioinformatics. 2022;23:144.
25. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. 2020. https://doi.org/10.48550/arXiv.2004.10964.
26. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, Liu T-Y. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022;23:bbac409.
27. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd clinical natural language processing workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019. p. 72–8.
28. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. 2019;1904:05342 [cs].
29. Huang K, Singh A, Chen S, Moseley E, Deng C-Y, George N, Lindvall C. Clinical XLNet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. In: Proceedings of the 3rd clinical natural language processing workshop. 2020. p. 94–100.
30. Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. J Am Med Inform Assoc. 2020;27:1935–42.
31. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. 2022. https://doi.org/10.48550/arXiv.2201.11838.
32. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. NPJ Digit Med. 2022;5:194.
33. Basaldella M, Liu F, Shareghi E, Collier N. COMETA: a corpus for medical entity linking in the social media. arXiv. 2020;2010:03295 [cs].
34. Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID–19 literature. Nucleic Acids Res. 2021;49:D1534–40.
35. Chen Q, Allot A, Leaman R, et al. Multi-label classification for biomedical literature: an overview of the BioCreative VII LitCovid Track for COVID–19 literature topic annotations. Database. 2022;2022:baac069.
36. Esteva A, Kale A, Paulus R, Hashimoto K, Yin W, Radev D, Socher R. COVID–19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. Npj Digit Med. 2021;4:1–9.
37. Nentidis A, Krithara A, Bougiatiotis K, Paliouras G. Overview of BioASQ 8a and 8b: results of the Eighth Edition of the BioASQ tasks a and b. In: Cappellato L, Eickhoff C, Ferro N, Névéol A, eds. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. Thessaloniki, Greece: CEUR; 2020. Available from: https://ceur-ws.org/Vol-2696/#paper_164.
38. You R, Liu Y, Mamitsuka H, Zhu S. BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text. Bioinformatics. 2021;37:684–92.
39. Sun C, Yang Z, Wang L, Zhang Y, Lin H, Wang J. Biomedical named entity recognition using BERT in the machine reading comprehension framework. J Biomed Inform. 2021;118:103799.
40. Peng Y, Chen Q, Lu Z. An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing. Online: Association for Computational Linguistics; 2020. p. 205–14. Available from: https://aclanthology.org/2020.bionlp-1.22.
41. Khandelwal A, Kar A, Chikka VR, Karlapalem K. Biomedical NER using novel schema and distant supervision. In: Proceedings of the 21st workshop on biomedical language processing. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 155–60.
42. Zaratiana U, Tomeh N, Holat P, Charnois T. GNNer: reducing overlapping in span-based NER using graph neural networks. In: Proceedings of the 60th annual meeting of the Association for Computational Linguistics: student research workshop. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 97–103.
43. Fries JA, Steinberg E, Khattar S, Fleming SL, Posada J, Callahan A, Shah NH. Ontology-driven weak supervision for clinical entity classification in electronic health records. Nat Commun. 2021;12:2017.
44. Madan S, Julius Zimmer F, Balabin H, Schaaf S, Fröhlich H, Fluck J, Neuner I, Mathiak K, Hofmann-Apitius M, Sarkheil P. Deep learning-based Detection of Psychiatric Attributes from German Mental Health Records. International Journal of Medical Informatics 104724; 2022.
45. Huang C-W, Tsai S-C, Chen Y-N. PLM-ICD: automatic ICD coding with pretrained language models. In: Proceedings of the 4th clinical natural language processing workshop. 2022. p. 10–20.
46. Johnson AE, Bulgarelli L, Pollard TJ. Deidentification of free-text medical records using pre-trained bidirectional transformers. In: Proceedings of the ACM conference on health, inference, and learning. 2020. p. 214–21.
47. Vakili T, Lamproudis A, Henriksson A, Dalianis H. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In: Proceedings of the thirteenth language resources and evaluation conference. 2022. p. 4245–52.
48. Sung M, Jeong M, Choi Y, Kim D, Lee J, Kang J. BERN2: an advanced neural biomedical named entity recognition and normalization tool. Bioinformatics. 2022;38:4837–9.
49. Mungall C, Matentzoglu N, Balhoff J et al. Oborel/obo-relations: release 2022-10-26. 2022. https://doi.org/10.5281/zenodo.7254604.
50. Karki R, Madan S, Gadiya Y, Domingo-Fernández D, Kodamullil AT, Hofmann-Apitius M. Data-driven modeling of knowledge assemblies in understanding comorbidity between type 2 diabetes mellitus and alzheimer's disease. J Alzheimers Dis. 2020;78:1–9.
51. Kodamullil AT, Iyappan A, Karki R, Madan S, Younesi E, Hofmann-Apitius M. Of mice and men: comparative analysis of neuro-inflammatory mechanisms in human and mouse using cause-and-effect models. J Alzheimers Dis. 2017;59:1045–55.
52. Zhu Y, Li L, Lu H, Zhou A, Qin X. Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions. J Biomed Inform. 2020;106:103451.

53. Li D, Xiong Y, Hu B, Tang B, Peng W, Chen Q. Drug knowledge discovery via multi-task learning and pre-trained models. BMC Med Inf Decis Mak. 2021;21:251.
54. Hu D, Zhang H, Li S, Wang Y, Wu N, Lu X. Automatic extraction of lung cancer staging information from computed tomography reports: deep learning approach. JMIR Med Inf. 2021;9:e27955.
55. Zhang X, Zhang Y, Zhang Q, Ren Y, Qiu T, Ma J, Sun Q. Extracting comprehensive clinical information for breast cancer using deep learning methods. Int J Med Inf. 2019;132:103985.
56. Bansal T, Verga P, Choudhary N, McCallum A. Simultaneously linking entities and extracting relations from biomedical text without mention-level supervision. arXiv. 2019;1912:01070 [cs].
57. Chen M, Lan G, Du F, Lobanov V. oint Learning with Pre-trained Transformer on Named Entity Recognition and Relation Extraction Tasks for Clinical Analytics. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: Association for Computational Linguistics; 2020. p. 234–42.
58. Verga P, Strubell E, McCallum A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In: Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics; 2018. p. 872–84.
59. Iinuma N, Miwa M, Sasaki Y. Improving supervised drug-protein relation extraction with distantly supervised models. In: Proceedings of the 21st workshop on biomedical language processing. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 161–70.
60. Papanikolaou Y, Roberts I, Pierleoni A. Deep bidirectional transformers for relation extraction without supervision. In: Proceedings of the 2nd workshop on deep learning approaches for low-resource NLP (DeepLo 2019). Hong Kong, China: Association for Computational Linguistics; 2019. p. 67–75.
61. Hall K, Chang V, Jayne C. A review on natural language processing models for COVID–19 research. Healthc Analytics. 2022;2:100078.
62. Kalyan KS, Rajasekharan A, Sangeetha S. AMMU: a survey of transformer-based biomedical pretrained language models. J Biomed Inform. 2022;126:103982.
63. Wang B, Xie Q, Pei J, Tiwari P, Li Z, fu J. Pre-trained language models in biomedical domain: a systematic survey. 2021. https://doi.org/10.48550/arXiv.2110.05006.
64. Syafiandini AF, Song G, Ahn Y, Kim H, Song M. An automatic hypothesis generation for plausible linkage between xanthium and diabetes. Sci Rep. 2022;12:17547.
65. Hong G, Kim Y, Choi Y, Song M. BioPREP: deep learning-based predicate classification with SemMedDB. J Biomed Inform. 2021;122:103888.
66. García del Valle EP, Lagunes García G, Prieto Santamaría L, Zanin M, Menasalvas Ruiz E, Rodríguez-González A. Leveraging network analysis to evaluate biomedical named entity recognition tools. Sci Rep. 2021;11:13537.
67. Aldahdooh J, Vähä-Koskela M, Tang J, Tanoli Z. Using BERT to identify drug-target interactions from whole PubMed. BMC Bioinformatics. 2022;23:245.
68. Zhou H, Li X, Yao W, Liu Z, Ning S, Lang C, Du L. Improving neural protein-protein interaction extraction with knowledge selection. Comput Biol Chem. 2019;83:107146.
69. Wang J, Ren Y, Zhang Z, Xu H, Zhang Y. From tokenization to self-supervision: building a high-performance information extraction system for chemical reactions in patents. Front Res Metr Anal. 2021;6:691105.
70. Jain H, Raj N, Mishra S. A Sui Generis QA Approach using RoBERTa for adverse drug event identification. BMC Bioinformatics. 2021;22:330.
71. Cho H, Kim B, Choi W, Lee D, Lee H. Plant phenotype relationship corpus for biomedical relationships between plants and phenotypes. Sci Data. 2022;9:235.
72. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49:D480–9.
73. Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. Nucleic Acids Res. 2022;50:D988–95.
74. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res. 2020;48:D84–6.
75. Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: towards cracking the Language of Lifes Code through Self-supervised Deep Learning and High Performance Computing. IEEE Trans Pattern Anal Mach Intell. 2021. https://doi.org/10.1109/tpami.2021.3095381.
76. Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. PNAS. 2021. https://doi.org/10.1073/pnas.2016239118.
77. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 2022;38:2102–10.
78. Madani A, McCann B, Naik N, Keskar NS, Anand N, Eguchi RR, Huang P-S, Socher R. ProGen: Language Modeling for Protein Generation. 2020. https://doi.org/10.48550/arXiv.2004.03497.
79. Madani A, Krause B, Greene ER et al. (2021) Deep neural language modeling enables functional protein generation across families. 2021.07.18.452833.
80. Hesslow D, Zanichelli N, Notin P, Poli I, Marks D. RITA: a study on scaling up generative protein sequence models. arXiv. 2022;2205:05789.
81. Nijkamp E, Ruffolo J, Weinstein EN, Naik N, Madani A. ProGen2: Exploring the Boundaries of Protein Language Models. 2022. https://doi.org/10.48550/arXiv.2206.13517.
82. Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun. 2022;13:4348.
83. Detlefsen NS, Hauberg S, Boomsma W. Learning meaningful representations of protein sequences. Nat Commun. 2022;13:1914.
84. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen P, Canny J, Abbeel P, Song Y. Evaluating protein transfer learning with TAPE. Adv Neural Inf Process Syst. 2019;32.
85. Unsal S, Atas H, Albayrak M, Turhan K, Acar AC, Doğan T. Learning functional properties of proteins with language models. Nat Mach Intell. 2022;4:227–45.
86. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577:706–10.
87. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
88. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373:871–6.
89. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379:1123–30.
90. Clauwaert J, Waegeman W. Novel transformer networks for Improved sequence labeling in genomics. IEEE/ACM Trans Comput Biol Bioinf. 2020;19:97–106.
91. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics. 2021;37:2112–20.
92. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18:1196–203.
93. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. 2018;28:739–50.
94. Evans R, O'Neill M, Pritzel A et al. Protein complex prediction with AlphaFold-Multimer. 2022;2021.10.04.463034.
95. Chen B, Xie Z, Qiu J, Ye Z, Xu J, Tang J. Improved the Protein Complex Prediction with Protein Language Models. 2022;2022.09.15.508065.
96. Rao RM, Liu J, Verkuil R, Meier J, Canny J, Abbeel P, Sercu T, Rives A. MSA Transformer. In: Proceedings of the 38th International Conference on Machine Learning. Online: PMLR; 2021. p. 8844–56. Available from: https://proceedings.mlr.press/v139/rao21a.html.
97. Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol. 2022;40:1023–5.
98. Notin P, Dias M, Frazer J, Hurtado JM, Gomez AN, Marks D. Tranception: Protein Fitness Prediction with Autoregressive Transformers and Inference-time Retrieval. In: Proceedings of the 39th International Conference on Machine Learning. Online: PMLR; 2022. p. 16990–7017
99. Hsu C, Nisonoff H, Fannjiang C, Listgarten J. Learning protein fitness models from evolutionary and assay-labeled data. Nat Biotechnol. 2022;40:1114–22.

100. Bernhofer M, Rost B. TMbed: transmembrane proteins predicted through language model embeddings. BMC Bioinformatics. 2022;23:326.
101. Castro E, Godavarthi A, Rubinfien J, Givechian K, Bhaskar D, Krishnaswamy S. (2022) Transformer-based protein generation with regularized latent space optimization. Nat Mach Intell 1–12.
102. Kang H, Goo S, Lee H, Chae J, Yun H, Jung S. Fine-tuning of BERT Model to accurately predict drug–target interactions. Pharmaceutics. 2022;14:1710.
103. Madan S, Demina V, Stapf M, Ernst O, Fröhlich H. Accurate prediction of virus-host protein-protein interactions via a siamese neural network using deep protein sequence embeddings. Patterns. 2022;3:100551.
104. Zitnik M, Sosič R, Maheshwari S, Leskovec J. BioSNAP Datasets: Stanford Biomedical Network Dataset Collection. 2018. http://snap.stanford.edu/biodata.
105. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19:1236–46.
106. Shang J, Ma T, Xiao C, Sun J. Pre-training of graph augmented transformers for medication recommendation. In: 28th International Joint Conference on Artificial Intelligence, IJCAI 2019. Macao: International Joint Conferences on Artificial Intelligence (IJCAI); 2019. p. 5953–9.
107. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G. BEHRT: Transformer for electronic health records. Sci Rep. 2020;10:7155.
108. Li Y, Mamouei M, Salimi-Khorshidi G, Rao S, Hassaine A, Canoy D, Lukasiewicz T, Rahimi K. Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records. IEEE J Biomed Health Inform. 2023;27:1106–17.
109. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. 2021;4:1–13.
110. Lentzen M, Linden T, Veeranki S, Madan S, Kramer D, Leodolter W, Fröhlich H. A transformer-based model trained on large scale Claims Data for prediction of severe COVID–19 disease progression. IEEE J Biomedical Health Inf. 2023;27:4548–58.
111. Pang C, Jiang X, Kalluri KS, Spotnitz M, Chen R, Perotte A, Natarajan K. (2021) CEHR-BERT: incorporating temporal information from structured EHR data to improve prediction tasks. Mach Learn Health 239–60.
112. Kazemi SM, Goel R, Eghbali S, Ramanan J, Sahota J, Thakur S, Wu S, Smyth C, Poupart P, Brubaker M. (2019) Time2Vec: learning a vector representation of time. https://doi.org/10.48550/ARXIV.1907.05321.
113. Darabi S, Kachuee M, Fazeli S, Sarrafzadeh M. TAPER: time-aware patient EHR representation. IEEE J Biomedical Health Inf. 2020;24:3268–75.
114. Finch A, Crowell A, Chang Y-C, Parameshwarappa P, Martinez J, Horberg M. A comparison of attentional neural network architectures for modeling with electronic medical records. JAMIA Open. 2021;4:ooab064.
115. Luo J, Ye M, Xiao C, Ma F. HiTANet: Hierarchical time-aware attention networks for risk prediction on electronic health records. HiTANet. 2020. https://doi.org/10.1145/3394486.3403107.
116. Peng X, Long G, Shen T, Wang S, Jiang J. (2021) Sequential diagnosis prediction with transformer and ontological representation. https://doi.org/10.48550/ARXIV.2109.03069.
117. Ren H, Wang J, Zhao WX, Wu N. RAPT: pre-training of time-aware transformer for learning robust healthcare representation. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. New York, NY, USA: Association for Computing Machinery; 2021. p. 3511–3503.
118. Agarwal K, Choudhury S, Tipirneni S, et al. Preparing for the next pandemic via transfer learning from existing diseases with hierarchical multi-modal BERT: a study on COVID–19 outcome prediction. Sci Rep. 2022;12:10748.
119. Meng Y, Speier W, Ong MK, Arnold CW. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. IEEE J Biomedical Health Inf. 2021;25:3121–9.
120. Liu S, Wang X, Hou Y, Li G, Wang H, Xu H, Xiang Y, Tang B. (2022) Multimodal data matters: Language model pre-training over structured and unstructured electronic health records. IEEE J Biomedical Health Inf 1–12.
121. Pang C, Jiang X, Pavinkurve NP, Kalluri KS, Minto EL, Patterson J, Zhang L, Hripcsak G, Elhadad N, Natarajan K. CEHR-GPT: Generating Electronic Health Records with Chronological Patient Timelines. 2024. https://doi.org/10.48550/arXiv.2402.04400.
122. Kumar Y, Ilin A, Salo H, Kulathinal S, Leinonen MK, Marttinen P. (2024) Self-Supervised Forecasting in Electronic Health Records with attention-free models. IEEE Trans Artif Intell 1–17.
123. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Vedaldi A, Bischof H, Brox T, Frahm JM, eds. Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol. 12346. Cham: Springer; 2020. https://doi.org/10.1007/978-3-030-58452-8_1.
124. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: ICLR 2021 The Ninth International Conference on Learning Representations. Online: International Conference on Learning Representations (ICLR). 2021.
125. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021. p. 10012–22.
126. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. Online: PMLR; 2021. p. 10347–57.
127. Krishnan KS, Krishnan KS. Vision transformer based COVID–19 detection using chest X-rays. In: 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC). 2021. p. 644–8.
128. Park S, Kim G, Oh Y, Seo JB, Lee SM, Kim JH, Moon S, Lim J-K, Ye JC. Multi-task vision transformer using low-level chest X-ray feature corpus for COVID–19 diagnosis and severity quantification. Med Image Anal. 2022;75:102299.
129. Shome D, Kar T, Mohanty SN, Tiwari P, Muhammad K, AlTameem A, Zhang Y, Saudagar AKJ. Covid-transformer: interpretable covid–19 detection using vision transformer for healthcare. Int J Environ Res Public Health. 2021;18:11086.
130. Gheflati B, Rivaz H. Vision transformers for classification of breast ultrasound images. In: 2022 44th annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2022. p. 480–3.
131. Wang W, Jiang R, Cui N, Li Q, Yuan F, Xiao Z. Semi-supervised vision transformer with adaptive token sampling for breast cancer classification. Front Pharmacol. 2022;13:929755.
132. Tanzi L, Audisio A, Cirrincione G, Aprato A, Vezzetti E. Vision transformer for femur fracture classification. Injury. 2022;53:2625–34.
133. Chen H, Li C, Wang G, et al. GasHis-Transformer: a multi-scale visual transformer approach for gastric histopathological image detection. Pattern Recogn. 2022;130:108827.
134. Wu Y, Qi S, Sun Y, Xia S, Yao Y, Qian W. A vision transformer for emphysema classification using CT images. Phys Med Biol. 2021;66:245016.
135. Wang H, Ji Y, Song K, Sun M, Lv P, Zhang T. ViT-P: classification of genitourinary syndrome of menopause from OCT images based on vision transformer models. IEEE Trans Instrum Meas. 2021;70:1–14.
136. Tummala S, Kadry S, Bukhari SAC, Rauf HT. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. Curr Oncol. 2022;29:7498–511.
137. Salanitri FP, Bellitto G, Palazzo S, et al. Neural transformers for Intraductal Papillary Mucosal Neoplasms (IPMN) classification in MRI images. In: 2022 44th annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2022. p. 475–9.
138. He K, Gan C, Li Z, Rekik I, Yin Z, Ji W, Gao Y, Wang Q, Zhang J, Shen D. Transformers in medical image analysis: a review. 2022. https://doi.org/10.48550/arXiv.2202.12165.
139. Prangemeiz T, Reich C, Koeppl H. Attention-based transformers for instance segmentation of cells in microstructures. In: 2020 IEEE international conference on Bioinformatics and Biomedicine (BIBM). 2020. p. 700–7.
140. Gao Y, Zhou M, Metaxas DN. UTNet: a hybrid transformer architecture for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Online: Springer; 2021. p. 61–71.

141. Huang X, Deng Z, Li D, Yuan X. MISSFormer: an effective medical image segmentation transformer. 2021. https://doi.org/10.48550/arXiv.2109.07162.
142. Chandrasekaran SN, Ceulemans H, Boyd JD, Carpenter AE. Image-based profiling for drug discovery: due for a machine-learning upgrade? Nat Rev Drug Discov. 2021;20:145–59.
143. Cross-Zamirski JO, Williams G, Mouchet E, Schönlieb C-B, Turkki R, Wang Y. (2022) Self-supervised learning of phenotypic representations from cell images with weak labels. https://doi.org/10.48550/arXiv.2209.07819.
144. Wieser M, Siegismund D, Heyse S, Steigele S. Vision transformers show improved robustness in high-content image analysis. In: 2022 9th Swiss conference on Data Science (SDS). 2022. p. 72–71.
145. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770–8.
146. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60:84–90.
147. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol. 1962;160:106.
148. Li Z, Liu F, Yang W, Peng S, Zhou J. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans Neural Networks Learn Syst. 2022;33:6999–7019.
149. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. p. 16000–9.
150. Zhou L, Liu H, Bae J, He J, Samaras D, Prasanna P. Self pre-training with masked autoencoders for medical image classification and segmentation. In: 2023 IEEE 20th international symposium on biomedical imaging (ISBI). IEEE. 2023. p. 1–6.
151. Varma A, Shit S, Prabhakar C, Scholz D, Li HB, Menze B, Rueckert D, Wiestler B. VariViT: A vision transformer for variable image sizes. In: Medical imaging with deep learning. Paris, France. 2024.
152. Li MM, Huang K, Zitnik M. Graph representation learning in biomedicine and healthcare. Nat Biomed Eng. 2022;1–17.
153. Dwivedi VP, Bresson X. A Generalization of Transformer Networks to Graphs. 2021. https://doi.org/10.48550/arXiv.2012.09699.
154. Yun S, Jeong M, Yoo S, Lee S, Yi SS, Kim R, Kang J, Kim HJ. Graph Transformer networks: learning meta-path graphs to improve GNNs. Neural Netw. 2022;153:104–19.
155. Rong Y, Bian Y, Xu T, Xie W, Wei Y, Huang W, Huang J. Self-supervised graph transformer on large-scale molecular data. Adv Neural Inf Process Syst. 2020;33:12559–71.
156. Ingraham J, Garg VK, Barzilay R, Jaakkola T. Generative models for graph-based protein design. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc; 2019. p 15820–31.
157. Yang J, Liu Z, Xiao S, Li C, Lian D, Agrawal S, Singh A, Sun G, Xie X. GraphFormers: GNN-nested Transformers for Representation Learning on Textual Graph. arXiv. 2021;2105.02605. https://doi.org/10.48550/arXiv.2105.02605.
158. Bi Z, Cheng S, Chen J, Liang X, Xiong F, Zhang N. Relphormer: Relational Graph Transformer for Knowledge Graph representations. Neurocomputing. 2024;566:127044.
159. Busbridge D, Sherburn D, Cavallo P, Hammerla NY. Relational graph attention networks. arXiv. 2019;1904:05811 [cs, stat].
160. Wang H, Guo F, Du M, Wang G, Cao C. A novel method for drug-target interaction prediction based on graph transformers model. BMC Bioinformatics. 2022;23:459.
161. Zhang P, Wei Z, Che C, Jin B. DeepMGT-DTI: Transformer network incorporating multilayer graph information for drug–target interaction prediction. Comput Biol Med. 2022;142:105214.
162. Chu T, Nguyen TT, Hai BD, Nguyen QH, Nguyen T. Graph transformer for drug response prediction. IEEE/ACM Trans Comput Biol Bioinform. 2022. https://doi.org/10.1109/TCBB.2022.3206888.
163. Yan C, Suo Z, Wang J, Zhang G, Luo H. DACPGTN: drug ATC code prediction method based on graph transformer network for drug discovery. Front Pharmacol. 2022;13:907676.
164. El-allaly E, Sarrouti M, En-Nahnahi N, Ouatik El Alaoui S. An attentive joint model with transformer-based weighted graph convolutional network for extracting adverse drug event relation. J Biomed Inform. 2022;125:103968.
165. Chen D, Gao K, Nguyen DD, Chen X, Jiang Y, Wei G-W, Pan F. Algebraic graph-assisted bidirectional transformers for molecular property prediction. Nat Commun. 2021;12:3521.
166. Fradkin P, Young A, Atanackovic L, Frey B, Lee LJ, Wang B. A graph neural network approach for molecule carcinogenicity prediction. Bioinformatics. 2022;38:i84–91.
167. Zhang T, Guo X, Chen H, Fan S, Li Q, Chen S, Guo X, Zheng H. (2022) TG-GNN: transformer based geometric enhancement graph neural network for molecular property prediction. https://doi.org/10.21203/rs.3.rs-1795724/v1.
168. Lai P-T, Lu Z. (2021) BERT-GT: cross-sentence n-ary relation extraction with BERT and graph transformer. Bioinformatics btaa1087.
169. Yang T, Bai X, Cui X, Gong Y, Li L. GraformerDIR: graph convolution transformer for deformable image registration. Comput Biol Med. 2022;147:105799.
170. Yuan Q, Chen S, Rao J, Zheng S, Zhao H, Yang Y. AlphaFold2-aware protein–DNA binding site prediction using graph transformer. Brief Bioinform. 2022;23:bbab564.
171. Dong S, Wang S. Assembled graph neural network using graph transformer with edges for protein model quality assessment. J Mol Graph Model. 2022;110:108053.
172. Jha K, Saha S, Singh H. Prediction of protein–protein interaction using graph neural networks. Sci Rep. 2022;12:8360.
173. Ma A, Xin G, Ma Q. The use of single-cell multi-omics in immuno-oncology. Nat Commun. 2022;13:2728.
174. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med. 2022;28:1773–84.
175. Koorathota S, Khan Z, Lapborisuth P, Sajda P. Multimodal neurophysiological transformer for emotion recognition. In: 2022 44th annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2022. p. 3563–7.
176. Shi Q, Fan J, Wang Z, Zhang Z. Multimodal channel-wise attention transformer inspired by multisensory integration mechanisms of the brain. Pattern Recogn. 2022;130:108837.
177. Monajatipoor M, Rouhsedaghat M, Li LH, Chien A, Kuo CCJ, Scalzo F, Chang KW. BERTHop: An Effective Vision-and-Language Model for Chest X-ray Disease Diagnosis. 2021. https://doi.org/10.48550/arXiv.2108.04938.
178. Liu G, Liao Y, Wang F, Zhang B, Zhang L, Liang X, Wan X, Li S, Li Z, Zhang S. Medical-vlbert: medical visual language bert for covid–19 ct report generation with alternate learning. IEEE Trans Neural Networks Learn Syst. 2021;32:3786–97.
179. Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive learning of medical visual representations from paired images and text. In: Proceedings of machine learning for health care 2022. 2022.
180. Zeng Z, Yao Y, Liu Z, Sun M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. Nat Commun. 2022;13:862.
181. Speith T. A review of taxonomies of explainable artificial intelligence (XAI) methods. In: 2022 ACM conference on fairness, accountability, and transparency. New York, NY, USA: Association for Computing Machinery; 2022. p. 2239–50.
182. Vig J. BertViz: a tool for visualizing multihead self-attention in the BERT model. ICLR Workshop: Debugging Machine Learning Models. New Orleans: ICLR; 2019.
183. Kokalj E, Škrlj B, Lavrač N, Pollak S, Robnik-Šikonja M. BERT meets shapley: extending SHAP explanations to transformer-based classifiers. In: Proceedings of the EACL hackashop on news media content analysis and automated report generation. 2021. p. 16–21.
184. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc.; 2017. p. 4768–77
185. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. arXiv. 2017;1703:01365 [cs].
186. Saporta A, Gui X, Agrawal A et al. Benchmarking saliency methods for chest X-ray interpretation. 2022;2021.02.28.21252634.

187. Lim B, Arik SO, Loeff N, Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. 2020. https://doi.org/10.48550/arXiv.1912.09363.
188. Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling language modeling with pathways. arXiv. 2022;2204:02311 [cs].
189. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. arXiv. 2020;2004:05150.
190. Choromanski KM, Likhosherstov V, Dohan D, Song X, Gane A, Sarlos T, Hawkins P, Davis JQ, Mohiuddin A, Kaiser L. Rethinking attention with performers. International Conference on Learning Representations. Online: ICLR. 2021.
191. Kitaev N, Kaiser Ł, Levskaya A. Reformer: the efficient transformer. ArXiv. 2020;2001:04451 [cs, stat].
192. Tay Y, Dehghani M, Bahri D, Metzler D. Efficient transformers: a survey. ACM Comput Surv. 2022;55:1–109.
193. Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontanon S, Pham P, Ravula A, Wang Q, Yang L. Big bird: transformers for longer sequences. Adv Neural Inf Process Syst. 2020;33:17283–97.
194. Gou J, Yu B, Maybank SJ, Tao D. Knowledge distillation: a survey. Int J Comput Vis. 2021;129:1789–819.
195. Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, Deng H, Wang P. K-BERT: enabling language representation with knowledge graph. ArXiv. 2019;1909:07606 [cs].
196. Sun Y, Wang S, Li YK, Feng S, Tian H, Wu H, Wang H. ERNIE 2.0: a continual pre-training framework for language understanding. In: AAAI. 2020. p. 8968–75.
197. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: enhanced language representation with informative entities. arXiv. 2019;1905:07129.
198. Ahmad A, Fröhlich H. Integrating heterogeneous omics data via statistical inference and learning techniques. Genomics and computational biology. 2016. https://doi.org/10.18547/gcb.2016.vol2.iss1.e32.
199. Baevski A, Hsu W-N, Xu Q, Babu A, Gu J, Auli M. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. 2022. https://doi.org/10.48550/arXiv.2202.03555.
200. Jaegle A, Borgeaud S, Alayrac J-B, et al. Perceiver IO. A aeneral architecture for structured Inputs & outputs. 2022.
201. Jaegle A, Gimeno F, Brock A, Vinyals O, Zisserman A, Carreira J. Perceiver: general perception with iterative attention. In: International conference on machine learning. Online: PMLR. 2021. p. 4651–64.
202. OpenAI. ChatGPT (Mar 14 version) Large language model. 2023. https://chat.openai.com/chat.
203. Anand Y, Nussbaum Z, Duderstadt B, Schmidt B, Treat A. GPT4All: an ecosystem of open-source assistants that run on local hardware. 2023.
204. Verkuil R, Kabeli O, Du Y, Wicky BIM, Milles LF, Dauparas J, Baker D, Ovchinnikov S, Sercu T, Rives A. Language models generalize beyond natural proteins. 2022;2022.12.21.521521.
205. Ali A, Schnake T, Eberle O, Montavon G, Müller K-R, Wolf L. XAI for transformers: better explanations through conservative propagation. 2022. https://doi.org/10.48550/arXiv.2202.07304.
206. Deb M, Deiseroth B, Weinbach S, Schramowski P, Kersting K. AtMan: understanding transformer predictions through memory efficient attention manipulation. 2023. https://doi.org/10.48550/arXiv.2301.08110.
207. Gavito AT, Klabjan D, Utke J. Multi-layer attention-based explainability via transformers for tabular data. 2023. https://doi.org/10.48550/arXiv.2302.14278.
208. Del Ser J, Barredo-Arrieta A, Díaz-Rodríguez N, Herrera F, Saranti A, Holzinger A. On generating trustworthy counterfactual explanations. Inf Sci. 2024;655:119898.
209. Metsch JM, Saranti A, Angerschmid A, Pfeifer B, Klemt V, Holzinger A, Hauschild A-C. CLARUS: an interactive explainable AI platform for manual counterfactuals in graph neural networks. J Biomed Inform. 2024;150:104600.

Publisher's Note

B. Deep Learning-Based Detection of Psychiatric Attributes from German Mental Health Records

Reprinted with permission from:

S. Madan, F. Julius Zimmer, H. Balabin, S. Schaaf, H. Fröhlich, J. Fluck, I. Neuner, K. Mathiak, M. Hofmann-Apitius, and P. Sarkheil, "Deep Learning-based Detection of Psychiatric Attributes from German Mental Health Records," *International Journal of Medical Informatics*, vol. 161, p. 104 724, 2022. DOI: 10.1016/j.ijmedinf.2022.104724

International Journal of Medical Informatics 161 (2022) 104724

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Deep Learning-based detection of psychiatric attributes from German mental health records

Sumit Madan [a,b,*], Fabian Julius Zimmer [c], Helena Balabin [a], Sebastian Schaaf [d,1], Holger Fröhlich [a,e], Juliane Fluck [f,g,1], Irene Neuner [c], Klaus Mathiak [c], Martin Hofmann-Apitius [a,e], Pegah Sarkheil [c,h,*]

[a] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Schloss Birlinghoven, 53757 Sankt Augustin, Germany
[b] Institute of Computer Science, University of Bonn, 53113 Bonn, Germany
[c] Department of Psychiatry, Psychotherapy and Psychosomatics, Faculty of Medicine, RWTH Aachen, Pauwelsstraße 30, 52074 Aachen, Germany
[d] HPC and Scientific Computing, German Center for Neurodegenerative Diseases (DZNE), 53127 Bonn, Germany
[e] Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53113 Bonn, Germany
[f] Knowledge Management, ZB MED – Information Centre for Life Sciences, 50931 Cologne, Germany
[g] The Agricultural Faculty, University of Bonn, 53115 Bonn, Germany
[h] Department of Psychiatry and Psychotherapy, University of Münster, 48149 Münster, Germany

ARTICLE INFO

Keywords:
Electrical Health Records
Mental State Examination
Clinical Text Mining
Deep Learning, AMDP

ABSTRACT

Background: Health care records provide large amounts of data with real-world and longitudinal aspects, which is advantageous for predictive analyses and improvements in personalized medicine. Text-based records are a main source of information in mental health. Therefore, application of text mining to the electronic health records – especially mental state examination – is a key approach for detection of psychiatric disease phenotypes that relate to treatment outcomes.

Methods: We focused on the mental state examination (MSE) in the patients' discharge summaries as the key part of the psychiatric records. We prepared a sample of 150 text documents that we manually annotated for psychiatric attributes and symptoms. These documents were further divided into training and test sets. We designed and implemented a system to detect the psychiatric attributes automatically and linked the pathologically assessed attributes to AMDP terminology. This workflow uses a pre-trained neural network model, which is fine-tuned on the training set, and validated on the independent test set. Furthermore, a traditional NLP and rule-based component linked the recognized mentions to AMDP terminology. In a further step, we applied the system on a larger clinical dataset of 510 patients to extract their symptoms.

Results: The system identified the psychiatric attributes as well as their assessment (normal and pathological) and linked these entities to the AMDP terminology with an F_1-score of 86% and 91% on an independent test set, respectively.

Conclusion: The development of the current text mining system and the results highlight the feasibility of text mining methods applied to MSE in electronic mental health care reports. Our findings pave the way for the secondary use of routine data in the field of mental health, facilitating further clinical data analyses.

1. Introduction

Beside the common advantages of clinical routine data like availability and cost-effectiveness, the use of routine data in mental health research has additional values regarding the longitudinal information. Because mental health conditions are highly dependent on dynamic brain-related processes like developmental, adaptive, and degenerative changes, follow-ups and reevaluations over an extended period can provide essential information in understanding the psychopathological aspects of the diseases. Retrospective studies based on large electronic

* Corresponding authors at: Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Schloss Birlinghoven, 53757 Sankt Augustin, Germany (S. Madan). Department of Psychiatry and Psychotherapy, University of Münster, 48149 Münster, Germany (P. Sarkheil).
E-mail addresses: sumit.madan@scai.fraunhofer.de, sumit.madan@gmx.de (S. Madan).
[1] These authors worked at 1 during conduction of the study.

https://doi.org/10.1016/j.ijmedinf.2022.104724
Received 24 November 2021; Received in revised form 7 February 2022; Accepted 18 February 2022
Available online 22 February 2022

collections of mental health records can be an attractive source of longitudinal information. To enable the use of unstructured clinical routine data in mental health research, considerable work needs to be done to extract structured information, which involves coding of free-text reports through text mining techniques.

Commonly, a psychiatric clinical examination consists of an interview regarding the past and present symptoms and observing the current pathological signs. Mental status exam (MSE) [1] is a standardized form of examination with methods for observing and describing the mental state and behaviors of each patient, based on both objective observations of the clinician and subjective descriptions given by the patient herself. To capture the information related to MSE, a medical expert examines the patient for any possible signs or symptoms of a psychiatric condition and provides documentation as a part of the patient's medical record. The MSE has a great potential to be a main information source to extract phenotype information from the electronic health records [2]. Text documentation of MSE follows a specified form and semantics [3] that intends to facilitate rapid medical communication and training (Fig. 1). To identify the entities that contain the main relevant and classifiable information, advanced text mining approaches are needed as MSE documentation often contains incomplete sentences, abbreviations, acronyms, negations. Also, the documentation is very individual and often contains vague or uncertain expressions.

Mapping the MSE documentation to a reference system is an important step for generation of the comparable results. For relating the mentioned symptoms to the standard psychopathological concepts, we relied on the assessment based on the Association for Methodology and Documentation in Psychiatry (AMDP) system [4], which has been developed for the standardized assessment of mental state. The AMDP system, which has been internationally recognized and translated into many languages (English, French, German, Italian, Portuguese, etc.), represents a terminology of psychopathological symptoms and their rating. It contains a short definition for each symptom, notes on the severity (mild, medium, severe), and a list of distinct examples. The symptoms listed by the AMDP system (in total 140 features) are numbered and grouped together in the AMDP manual. Altogether, AMDP can serve as a terminology for the MSE.

Researchers have proposed rule and machine learning-based text mining methods to extract various kind of information from electronic health records (EHR). Barak-Corren et al. [5] extracted demographic characteristics, diagnostic codes, laboratory results and prescribed medications from English EHRs to predict suicidal behavior. Hazewinkel et al. [6] have analyzed textual data included in notes and reports of Dutch EHRs of patients that were admitted in a psychiatric hospital in The Netherlands. Clinical notes in English language of psychiatry wards from Mayo Clinic were utilized by Sohn et al. [7] to detect drug side effects using a rule-based approach. Named entity recognition (NER) has been particularly popular in mental health care in identification of the relevant concepts from the clinical records [8,9]. It has been already applied to identify predictors of suicide from EHR [8] and social media [9].

Recently, transfer learning-based methods such as Bidirectional Encoder Representations from Transformers (BERT) have gained a lot of attention [10]. Briefly, transfer learning rests on the idea that pre-trained word embeddings [11] learned from large amounts of training data (e.g., Wikipedia articles) using deep learning models already contain a significant amount of information that is relevant for more specific downstream tasks, including NER in clinical documents. Lee et al. [12] published BioBERT which is additionally trained on PubMed and PubMed Central articles, achieving state-of-the-art results in several biomedical natural language processing (NLP) tasks. Similarly, for the clinical domain, Alsentzer et al. [13] have created Clinical BERT embeddings by performing a pre-training on 2 million freely available clinical notes [14] using the BERT architecture [10]. To our knowledge no transformers-based language model has yet been applied to German clinical data. An extended description of the related work is included in Supplementary.

In this work, we introduce a text mining approach to extract key clinical information from MSE documents. As a first step, we extracted MSE containing documents from the clinical information system and prepared a manually annotated dataset. The text mining system was implemented as a two-step procedure for 1) deep learning-based recognition of relevant entities, such as psychological assessments (NER), and 2) mapping to the standard AMDP terminology (entity linking). For NER, we fine-tuned a freely available deep learning-based general language understanding model, so called GermanBERT [15], that is pre-trained on German textual content. Additionally, we evaluated our text mining system thoroughly based on an independent test set and demonstrate the promising prediction performance. Finally, we applied our workflow to identify psychopathological symptoms from an enhanced set of psychiatric patient data. We also self-assess the quality of our medical AI work that employs medical data using the IJMEDI checklist [16] (included in Supplementary).

2. Materials and methods

2.1. Study data

2.1.1. Sample selection

We selected a set of MSE texts from discharge summaries, which are issued when or after the patient leaves the care of the hospital as the primary communication mechanism between hospitals and other healthcare providers. More than 30.000 German documents of this kind are available in the electronic archives of the Department of Psychiatry and Psychotherapy, University Hospital Aachen (UKAachen). For the current study, MSE sections of 660 patients were isolated from the discharge summaries of (pseudo-)randomly selected patients from various ranges of mental disorders, who received treatment in the inpatient services of UKAachen between 2014 and 2019. Patients' identification information (such as names, gender) were removed from all study data. The study data consists of two different datasets – an annotated dataset used for system training and evaluation, and an additional unlabeled dataset for later system application. Note that the additional dataset has no label annotations and thus cannot be used for model evaluation purposes. The dataset for training consisted of 100 documents, which were randomly chosen out of the 150 annotated documents. The remaining 50 annotated documents were used as independent test data.

Table 1 shows the demographics of the patient collective (n = 150) of the dataset for system training and evaluation in which 75 (50%) were

"The patient was conscious. He was oriented to time, place and person. In social contact he was friendly and cooperative. Speech monotonous, no disturbance of attention and memory, concentration disturbed. He exhibited loosening of associations and flight of ideas, no compulsions, no anxiety. Delusions were not observed. He reported auditory hallucinations. Affect labile, mood depressed, partly restless, no agitation, no aggressive behavior. He denied suicidal ideas."

Fig. 1. An exemplary MSE report.

Table 1
Demographics of the patient collective of two different datasets for system training and evaluation, and for system application.

	Dataset for system training and evaluation	Dataset for system application
Total	n = 150 (training = 100 and test = 50)	n = 510
Gender	75 female, 75 male	223 female, 287 male
Age at date of discharge (years)	44.52 (mean) 17.56 (standard deviation)	48.66 (mean) 12.08 (standard deviation)
Retrospective time span	2014 – 2018	2017 – 2019

Table 2
Total number of manual annotations for each class appearing in the annotated set of 150 documents.

Class	Total annotations
Attribute	3,423
NormalAssessment	1,734
PathologicalAssessment	1,302
AMDP concept	1,276

female and 75 (50%) were male. The sample is further characterized by the categories of mental disorders encoded with ICD-10 [17] diagnoses, since various disorders are expected to be differential in MSE outcomes (Supplementary Table S1). Whereas the additional dataset of anonymized, unannotated psychiatric discharge summaries from years 2017 to 2019 was used to predict the patients' symptoms by applying the developed system. In total, we extracted 510 MSEs (170 MSEs for each of the three categories; see Supplementary Table S1) from discharge summaries.

2.2. Data annotation

To build the gold standard, MSEs have been tagged by a medical expert with the following label types: 1) Attribute: assessed components, 2) NormalAssessment (related to a component), and 3) PathologicalAssessment (related to a component) and further verified by a board-certified psychiatrist. Fig. 2 shows exemplary excerpts of two MSEs. Furthermore, the phrases that have been labeled with the type *PathologicalAssessment* in combination with their related attributes were mapped to the AMDP terminology (Fig. 3). The mapping was performed only for this label type, because the AMDP terminology only covers the pathological mental states. The annotation guidelines are included in Supplementary.

2.3. Ethics and data protection

The data used for this retrospective research is considered as "real-world data" collected during the primary mental health care in hospital. The use of patients' data for the current research was approved by "the Medical Ethics Committee" at the RWTH Aachen Faculty of Medicine (EK 349/20).

3. Methodology

In order to automatically detect entities in MSEs, we employed the widely used approach of fine-tuning a pre-trained language model on a given task-specific dataset. One of the most well-known deep learning models used for this type of approach is Bidirectional Encoder Representations from Transformers (BERT) [10]. BERT fully relies on using attention functions [18] to learn token embeddings. The original BERT model is limited to the English language. However, recently, a version for the German language (GermanBERT [15]) has been made publicly available. Following the example of the BERT transfer learning approach, GermanBERT [15] forms the equivalent adaptation of the language model to the German language domain. Using the same hyperparameters as the original $BERT_{BASE}$ [10] model, the GermanBERT [15] model was trained from scratch on the German Wikipedia dump, the OpenLegalData [19] dump and German news articles. In this work, we further fine-tuned GermanBERT on our study data to recognize the defined entity classes. Once the model is fine-tuned on the MSEs, it can be used to predict new labels on other unseen examinations. Lastly, the identified entities are mapped / normalized to the best matching AMDP concepts. A summary of the overall workflow is shown in Fig. 4.

3.1. Preprocessing datasets for Fine-Tuning process

For NER, the offsets of the annotated entities, which represent the actual position in text, were first converted into an inside-outside-beginning (IOB) format [20] (see Fig. 5). Overall, there were a total of nine labels present in the token classification setting. Seven of them represented the IOB scheme, namely one outside class, together with the beginning and inside labels for each of the three original annotation classes. Additionally, there were two more labels for padding and tagging sub words of a labelled entity. Afterwards, the documents were segmented into single sentences. Only sentences containing entities were considered for fine-tuning GermanBERT.

3.2. Fine-tuning GermanBERT on study data for entity recognition

Based on the pre-trained GermanBERT model, a tokenizer and a language model were initialized. To adapt the general-purpose language model to the given entity recognition task, a token classification head consisting of a feed-forward and a softmax layer was added on top of the output of the GermanBERT model. More precisely, this final output layer represented each possible token label (resulting in $dim_{out} = 9$) and, additionally, was fully connected to the previous layer (of dimension 768).

After tokenization and preprocessing, the training data was used within a 5-fold cross-validation to fit and optimize the parameters of the model. The respective hyperparameters are listed in Supplementary Table S3. For assessing the performance of the fine-tuned models, we used entity-level precision, recall, and F_1-scores, both separately for each class, as well as the micro and macro averages of all classes.

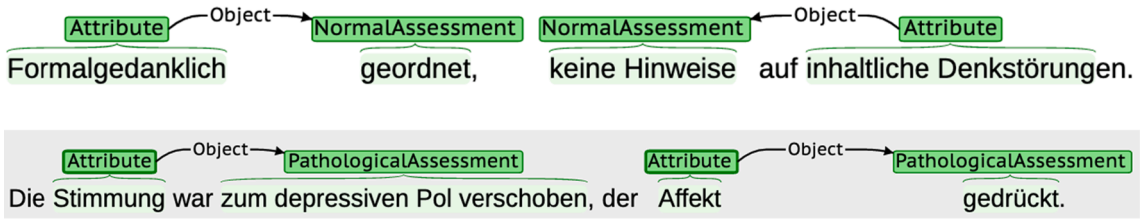


Fig. 2. Exemplary excerpts from the MSE document. The gold standard annotation of the attributes is represented by the green boxes. Each attribute of the type NormalAssessment or PathologicalAssessment is related to an annotation of the type Attribute. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

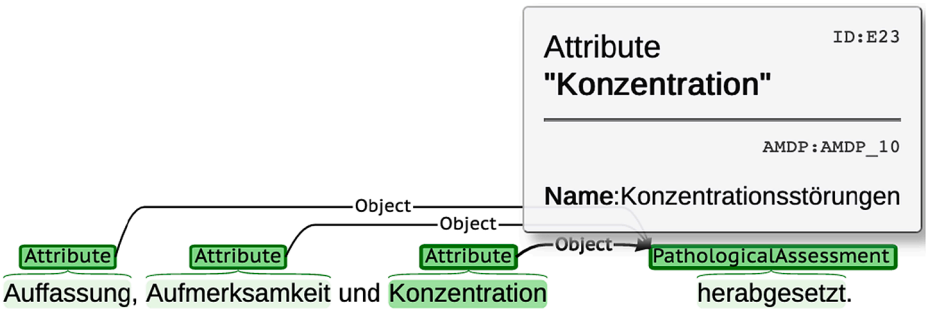


Fig. 3. An annotated sentence that relates the Attribute "Konzentration" (English: concentration) and its pathological assessment "herabgesetzt" (English: reduced) to the AMDP concept "Konzentrationsstörungen" (ID: AMDP:10) (English: concentration disturbance).

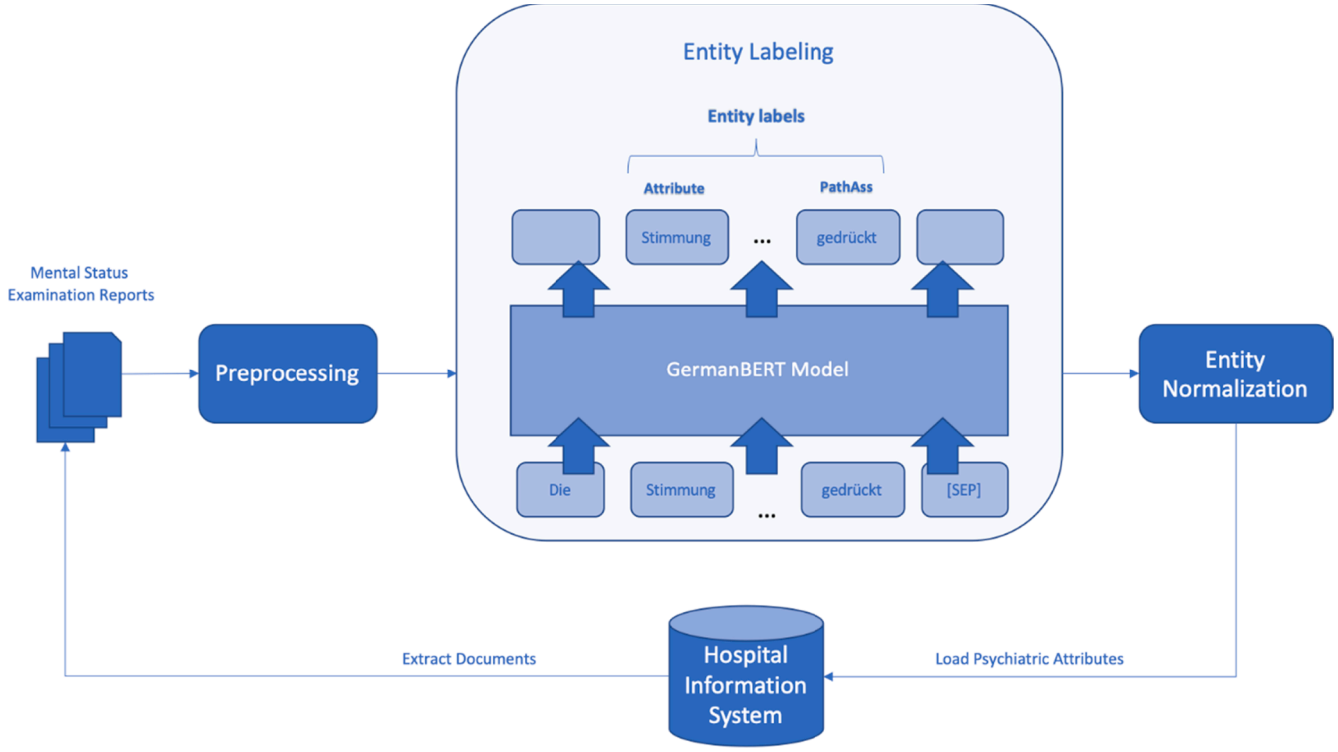


Fig. 4. Outline of the general workflow used for the analysis of MSE reports, consisting of the fine-tuning procedure of the pre-trained GermanBERT model, as well as additional pre- and post-processing steps. Sentences are used as input, as described by the (simplified) input cells in the *entity labeling* box. The entity labels are normalized in the entity normalization procedure to the AMDP terminology. The results are then further loaded in the database of the HIS.

$$\text{Precision} = \frac{\sum true\ positive}{\sum true\ positive + false\ positive}$$

$$recall = \frac{\sum true\ positive}{\sum true\ positive + false\ negative}$$

$$F_1 - score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

3.3. Linking to AMDP terminology

Our initial empirical experiments showed that due to the small size of AMDP annotations, training of a robust machine learning (ML) model to map pathologically assessed attributes to AMDP terminology was not yet feasible. Therefore, we implemented a traditional dictionary-lookup algorithm to link the mentions to AMDP. The dictionary is derived in a multi-step process. First, we gathered all linked mentions of pathologically assessed attributes from the training data. In a second step, we concatenated the mentions of both classes to generate synonyms for the AMDP concepts appearing in training data. Next, we derived further synonyms by also considering the labels of all available AMDP concepts in the terminology themselves. Finally, we merged the synonyms and harmonized (such as lower casing, removal of special characters) them in a post-processing step.

This dictionary was used as the main resource for the matching algorithm that was used to link mentions of attribute and pathological assessment classes to AMDP terminology. After processing the documents with the ML-based named entity recognition, the detected entities were used as an input to the matching algorithm. As a first step, the sentences containing attribute and pathological assessment entities were filtered for further processing. A rule was defined to link the mentions of attributes to pathological assessments. The entities were linked to build pairwise combinations. The mentions of pairwise combinations are concatenated, which are further labeled as queries. Furthermore, a stemming approach, using the German Snowball stemmer, was applied on the synonyms of the dictionary and on the queries. Now, with the exact string match the queries were compared with the synonyms. If no

Die	O
Patienten	O
ist	O
wach	B-ATTR
,	O
bewusstseinsklar	B-ATTR
und	O
zu	B-ATTR
allen	I-ATTR
Qualitäten	I-ATTR
orientiert	B-NORM
.	O

Fig. 5. Exemplary sentence taken from the preprocessed MSE reports. B-ATTR, I-ATTR and B-NORM denote the beginning of an attribute, inside of an attribute, and beginning of a normal assessment entity, respectively. Any word that is not belonging to one of the three classes is labelled with O (meaning outside).

result was found, a fuzzy string match was performed. In our experiments, on the validation set we found a threshold of 91 as the best value for the fuzzy string-matching algorithm (see Section Implementation). This value was further used as a threshold for the final system.

3.4. Implementation

The system has been implemented with Java and Python. We use a modified version of BratReader from the DKPro Core [21] library to read the annotated documents and create a JSON document for further processing. We employed the spaCy library [22] for converting text into IOB format. The deep learning-based model for entity recognition was based on the Framework for Adapting Representation Models (FARM) [23] that internally uses the PyTorch implementation of the pre-trained model in the Transformers [24] library. The machine learning lifecycle was managed with the Mlflow [25] library that allows for logging all training and evaluation experiments as well as metrics and models. For entity normalization, German Snowball stemmer [26] integrated in pystemmer [27] was used to perform stemming of German tokens and fuzzywuzzy [28] was used to execute fuzzy string matching.

3.5. Data availability

The data that support the findings of this study are available on request from the corresponding authors, but restrictions apply to the availability of these data. The data are not publicly available due to data protection and privacy reasons as they represent sensitive patient data.

4. Results

4.1. Datasets

Data was obtained from the health care records of the Dept. of Psychiatry and Psychotherapy, RWTH Aachen University Hospital for training and evaluation of the models. A total of 150 documents were annotated manually with several classes. Table 2 contains the total number of annotations for each class. All the annotations were performed on the sentence level. In the annotated dataset 1,089 Attribute annotations can be found. These attributes have been further linked to mentions that are annotated with two classes NormalAssessment (569 entries) and PathologicalAssessment (386 entries). The pathological assessed attributes are normalized with the AMDP terminology. In total 773 AMDP concepts have been linked to 386 pathological assessed attributes.

Table 3 lists 10 top lower-cased annotations for each of the important classes that appear in the annotated dataset. The most common attributes are *wach*, *stimmung, konzentration, and antrieb*. The most common pathological AMDP concepts that appear in the dataset are *Dysphorisch* (eng. Dysphoric), *Affektarm (eng. Emotionless), Konzentrationsstörungen (eng. Concentration disorders), Antriebsarm (eng. Less energized)*.

4.2. Entity recognition

To detect the mentions of each class, we used the pre-trained GermanBERT model and fine-tuned it on the MSE documents. For this purpose, we initially split the entire dataset into a training dataset (100 documents) and a test dataset (50 documents) at random. To identify the best possible model variant based on the training data we employed 5-fold cross-validation. Based on the performance assessed through cross-validation (detailed results are included under Section Cross-Validation Results in Supplementary), we use the optimized hyperparameters to train the final model on the whole training dataset. The generalization performance of the final model was assessed on the held-out test set of 50 documents. Table 4 presents the classification scores for each class on the test set averaged over five runs (the results of the training performance are included in Supplementary Table S2). We reached a precision of 89.0%, a recall of 87.4%, and an F_1-score of

Table 4
Precision, recall and F_1-score on the test dataset (50 documents) of the NER on MSEs task, averaged over five runs. Support column informs about the total number of instances of each class in the test dataset.

	Precision (%)	Recall (%)	F_1-score (%)	Support
Attribute	89.0 ± 0.0	87.4 ± 1.2	88.6 ± 0.8	795
Pathological Assessment	87.4 ± 2.0	85.0 ± 0.0	86.2 ± 1.0	314
Normal Assessment	86.6 ± 2.0	85.6 ± 1.2	86.2 ± 1.6	282
macro average	88.0 ± 0.0	87.0 ± 0.0	87.0 ± 0.0	1,391
micro average	88.0 ± 0.0	87.0 ± 0.0	87.0 ± 0.0	1,391

Table 3
The most frequent annotations for Attributes, NormalAssessments, PathologicalAssessments, and AMDP concepts appearing in the annotated dataset. Some of the entries (such as “wach”, eng. Alert, or “ängste”, eng. Fears) count as Attributes and normal/pathological assessments at the same time and might appear in different classes.

Attributes	(n)	NormalAssessments	(n)	PathologicalAssessments	(n)	AMDP Concepts	(n)
wach	47	kein	126	reduziert	60	Dysphorisch (AMDP:67)	67
stimmung	47	orientiert	50	gedrückt	21	Affektarm (AMDP:61)	54
konzentration	46	wach	46	ängste	17	Konzentrationsstörungen (AMDP:10)	52
antrieb	46	klar	37	verlangsamt	14	Antriebsarm (AMDP:80)	47
kontakt	41	geordnet	33	distanziert	14	Aufmerksamkeitsstörungen (AMDP:152)	44
aufmerksamkeit	40	kein anhalt	26	herabgesetzt	11	Auffassungsstörungen (AMDP:9)	31
ich-störung	40	freundlich	24	angespannt	11	Motorisch unruhig (AMDP:83)	24
suizidalität	39	gepflegt	21	depressiv	11	Affektstarr (AMDP:79)	22
affekt	36	keine hinweise	20	vermindert	10	Ängste (AMDP:153)	21
bewusstsein	36	ruhig	16	beeinträchtigt	8	Affektlabil (AMDP:77)	20

88.6% for the detection of Attribute class. The PathologicalAssessment and NormalAssessment classes are both detected with an F_1-score of 86.2%. For all three classes, the scores are significantly better than achieved through cross validating the models. In summary, the validation on an unseen test set reveals that the model is quite generalizable and robust in terms of entity detection.

4.3. Extraction of AMDP concepts

One of the goals of the current work is to delineate the pathological attributes as AMDP concepts. For this purpose, the extracted patient attributes are first linked with their assessment. If pathological, they have been mapped to the corresponding concepts from the AMDP terminology. Table 5 shows the results of the extraction of AMDP concepts by using the algorithm mentioned in Section *Normalization to AMDP* on the test set. We reached a precision of 90.0%, a recall of 92.0%, and an F_1-score of 91.0%.

4.4. System application on additional patient records

Next, we applied the developed system to the additional dataset of 510 unannotated MSEs. The system could detect 7,047 Attribute (unique: 183), 3,073 NormalAssessment (unique:67), and 2,254 PathologicalAssessment entities (unique: 157). Furthermore, the mapping to the AMDP terminology retrieved a total of 2,197 AMDP concepts (unique: 44). Table 6 provides an overview of the top 10 annotations with their associated total amounts of the available classes. Most importantly, 7 out of top 10 AMDP concepts in the annotated training dataset are identical to the results of this dataset (Table 6).

5. Discussion

Most of the clinical routine data in the mental health discipline is recorded as text documents. Therefore, text mining techniques are becoming crucial for extraction of relevant information. In this work, we present the first text mining approach to mental health data analytics in the German speaking region. We focused on MSE as the main part of the clinical evaluation of psychiatric patients and a standard for communicating the evaluation results. A pre-trained deep neural network (GermanBERT [15]) have been fine-tuned to identify relevant attributes and psychological assessments in German clinical routine data. In a further step, a method to relate the extracted information to AMDP, a standard terminology for psychopathology, have been implemented. We validated the results of the approach on an independent test set to demonstrate the robustness of the method. Finally, we applied the workflow on a larger clinical dataset, which returns a set of symptom variables for further clinical and research data analyses.

Based on the expert annotation of our dataset, consisting of 150 MSE, 90% of the MSE pathological entities could be referred to the AMDP symptom list, which confirms the efficacy of this system in normalizing the unstructured MSE in routine data. The fine-tuned model that detects various entities such as attributes and their assessment as normal or pathological reached an F_1-score of around 86% on test dataset for all entity classes, which is a quite reasonable performance. Furthermore, the mapping / normalization of the pathological symptoms to AMDP terminology achieved a high F_1-score of 91%. These promising results encourage future efforts towards automatically structuring the clinical notes from the EHRs. Our results revealed that the AMDP concept of dysphoria has been most frequently identified in the MSE reports, suggesting evaluation of dysphoria to be the focus of clinicians. Other frequent AMDP concepts include emotionality, concentration, and drive. Altogether, it can be inferred that the clinicians use the MSE as a tool to observe and assess the patient's current mental state with a focus on affective evaluations. Text mining of MSE reports in EHR might primarily inform about the affective signs and symptoms. Several psychiatric attributes are reported rarely. As for now, we have only analyzed MSE of 660 patients through our information extraction pipeline, which may not comprehensively cover more diverse psychiatric attributes. Therefore, we plan to extend the work with a broader analysis of additional MSE reports that could reveal interesting associations.

Table 5
Precision, recall, and F_1-score on the test dataset of extraction of AMDP concepts. Support column shows the number of total AMDP concepts in the dataset.

	Precision (%)	Recall (%)	F_1-score (%)	Support
AMDP concept	90.0	92.0	91.0	341

The AMDP terminology has been developed to introduce a systematic to the terminology of psychopathology. The purpose for its development was a comparable and reliable documentation of evaluation results in clinical practice and research [4]. The AMDP system offers 100 psychopathological (and 40 somatic) definable symptoms, sorted in main categories as individual entries. We suggest the AMDP system can be used as a normative reference for the identified entities in text mining of MSE documents. That mentioned, the standard clinical terminology SNOMED-CT is quite popular for documenting patient clinical information in many countries. Recently, Germany has become a new member of the SNOMED International consortium and will start to apply this terminology in clinical research and practice. We suggest that in the near future a mapping of AMDP to SNOMED-CT will be required for a consistent harmonization of mental health care data to assure the convergence of clinical interpretations and machine-readable codes. This mapping might be also indicated as SNOMED-CT is likely to code specific items, while AMDP is a comprehensive system that includes the normal findings and unmentioned attributes.

The current workflow is based on costly manual extraction of the MSE section from the discharge summaries for training data creation. To speed up and improve the workflow further approaches are needed to include automatic segmentation of the MSE section in the clinical documents. A further point for future development is identifying the severity of disease symptoms and predicting ICD-10 codes directly from the collection of symptoms. Such approaches offer a great potential for a more cost-effective coding for secondary use of clinical data, for example in scientific research or in the context of insurance claims. Extending the text mining techniques to other clinical text sections like medical history and nursing reports is a useful further step for capturing the data needed for a broad biopsychosocial phenotyping. To test and improve the generalizability of the workflow, further future research is planned by applying the workflow to documents from multiple clinical centers.

6. Conclusion

In this study, we constructed a text analysis system composed of a neural network model and a traditional NLP and rule-based analysis methods extracting mental state attributes from the psychiatric discharge summaries. Routine clinical data was used for training, test, and validation. The proposed approach achieved promising results. Automatized transforming unstructured texts into a structured format, the standard AMDP terminology, enables identification of meaningful patterns and new insights from the clinical routine data in the psychiatric discipline.

CRediT authorship contribution statement

Sumit Madan: Conceptualization, Supervision, Methodology, Software, Visualization, Validation, Writing – original draft. **Fabian Julius Zimmer:** Resources, Data curation, Validation. **Helena Balabin:** Software, Writing – original draft. **Sebastian Schaaf:** Writing – review & editing. **Holger Fröhlich:** Writing – review & editing. **Juliane Fluck:** Conceptualization, Writing – review & editing. **Irene Neuner:** Writing – review & editing. **Klaus Mathiak:** Conceptualization, Writing – review & editing. **Martin Hofmann-Apitius:** Funding acquisition, Supervision,

Table 6
Top 10 annotations of Attribute, NormalAssessment, PathologicalAssessment, and AMDP concept with their associated total amounts appearing in the additional dataset of 510 patients.

Attributes	(n)	NormalAssessments	(n)	PathologicalAssessments	(n)	AMDP Concepts	(n)
antrieb	436	kein hinweis	448	reduziert	481	Dysphorisch (AMDP:67)	334
wach	392	orientiert	402	gedrückt	169	Antriebsarm (AMDP:80)	263
konzentration	384	geordnet	235	unruhig	126	Konzentrationsstörungen (AMDP:10)	212
suizidalität	374	distanziert	227	verlangsamt	110	Affektarm (AMDP:61)	185
stimmung	364	freundlich	226	depressiv	109	Aufmerksamkeitsstörungen (AMDP:152)	167
aufmerksamkeit	359	kein anhalt	224	herabgesetzt	95	Auffassungsstörungen (AMDP:9)	149
auffassung	326	ruhig	184	vermindert	63	Affektstarr (AMDP:79)	145
zwänge	324	kein	136	angespannt	60	Motorisch unruhig (AMDP:83)	107
kontakt	319	regelrecht	130	labil	45	Psychomotorisch verlangsamt (AMDP:156)	96
fremdgefährdung	310	zugewandt	118	gesteigert	42	Verlangsamt (AMDP:16)	72

Project administration, Writing – review & editing. **Pegah Sarkheil:** Funding acquisition, Project administration, Conceptualization, Methodology, Data curation, Validation, Writing – original draft.

Acknowledgements

S.M., F.J.Z, M.HA., and P.S's work was supported and funded from the EU's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2 by the HBP Medical Informatics Platform). S.M., H.B., S.S., and J.F.'s work was performed in context of the project IDSN that is supported by the German Federal Ministry of Education and Research (BMBF) as part of the program "i:Dsem – Integrative Data Semantics in the Systems Medicine", project number 031L0029 [A-C]. We would like to acknowledge the organization AMDP e.V. (Die Arbeitsgemeinschaft für Methodik und Dokumentation in der Psychiatrie) for the permission to use AMDP terminology for this research.

Summary Points.

"What was already known on the topic":

- Application of routine data in mental health investigations has been very limited to date.
- The major obstacle has been the unstructured information, mainly as text-based documents.
- Hazewinkel et al. [6] and Sohn et al. [7] have analyzed Dutch and English EHRs from psychiatry to obtain frequently used concepts and drug side effects.
- To our knowledge, extraction of psychiatric attributes and symptoms has not been scientifically explored yet.

"What this study added to our knowledge":

- We propose a text mining system for extraction of the relevant information from the mental health examination records, which we evaluated on an independent dataset.
- A deep-learning approach has been applied to identify the relevant attributes in the mental state examination records.
- We created a system to link the attributes to the AMDP standard terminology to semantically enhance the data and make it interoperable.
- We achieved encouraging results and demonstrated the feasibility of using text mining methods to extract relevant information from patient data, which can be used in future for mental health research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2022.104724.

References

[1] J.J. Silverman, M. Galanter, M. Jackson-Triche, D.G. Jacobs, J.W. Lomax, M. B. Riba, L.D. Tong, K.E. Watkins, L.J. Fochtmann, R.S. Rhoads, Practice Guidelines for the Psychiatric Evaluation of Adults, Third Edition, American Psychiatric Association (2016), https://doi.org/10.1176/appi.books.9780890426760.
[2] D.C. Martin, The Mental Status Examination, in: H.K. Walker, W.D. Hall, J.W. Hurst (Eds.), Clinical Methods: The History, Physical, and Laboratory Examinations, 3rd ed., Butterworths, Boston, 1990. http://www.ncbi.nlm.nih.gov/books/NBK320/ (accessed February 9, 2021).
[3] M. Soltan, J. Girguis, How to approach the mental state examination, BMJ. 357 (2017), j1821, https://doi.org/10.1136/sbmj.j1821.
[4] R.-D. Stieglitz, A. Haug, E. Fähndrich, M. Rösler, W. Trabert, Comprehensive Psychopathological Assessment Based on the Association for Methodology and Documentation in Psychiatry (AMDP) System: Development, Methodological Foundation, Application in Clinical Routine, and Research, Front Psychiatry. 8 (2017), https://doi.org/10.3389/fpsyt.2017.00045.
[5] Y. Barak-Corren, V.M. Castro, S. Javitt, A.G. Hoffnagle, Y. Dai, R.H. Perlis, M. K. Nock, J.W. Smoller, B.Y. Reis, Predicting Suicidal Behavior From Longitudinal Electronic Health Records, AJP. 174 (2) (2017) 154–162, https://doi.org/10.1176/appi.ajp.2016.16010077.
[6] M.C. Hazewinkel, R.F.P. de Winter, R.W. van Est, D. van Hyfte, D. Wijnschenk, N. Miedema, E. Hoencamp, Text Analysis of Electronic Medical Records to Predict Seclusion in Psychiatric Wards: Proof of Concept, Front. Psychiatry. 10 (2019), https://doi.org/10.3389/fpsyt.2019.00188.
[7] S. Sohn, J.-P.-A. Kocher, C.G. Chute, G.K. Savova, Drug side effect extraction from clinical narratives of psychiatry and psychology patients, J Am Med Inform Assoc. 18 (2011) i144–i149, https://doi.org/10.1136/amiajnl-2011-000351.
[8] M. Senior, M. Burghart, R. Yu, A. Kormilitzin, Q. Liu, N. Vaci, A. Nevado-Holgado, S. Pandit, J. Zlodre, S. Fazel, Identifying Predictors of Suicide in Severe Mental Illness: A Feasibility Study of a Clinical Prediction Rule (Oxford Mental Illness and Suicide Tool or OxMIS), Front Psychiatry. 11 (2020) 268, https://doi.org/10.3389/fpsyt.2020.00268.
[9] J. Du, Y. Zhang, J. Luo, Y. Jia, Q. Wei, C. Tao, H. Xu, Extracting psychiatric stressors for suicide from social media using deep learning, BMC Med Inform Decis Mak. 18 (2018) 43, https://doi.org/10.1186/s12911-018-0632-8.
[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ArXiv:1810.04805 [Cs]. (2018). http://arxiv.org/abs/1810.04805 (accessed February 11, 2019).
[11] A. Dudchenko, G. Kopanitsa, Comparison of Word Embeddings for Extraction from Medical Records, Int J Environ Res Public Health. 16 (2019) E4360, https://doi.org/10.3390/ijerph16224360.
[12] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, ArXiv: 1901.08746 [Cs]. (2019). http://arxiv.org/abs/1901.08746 (accessed February 6, 2019).
[13] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly Available Clinical BERT Embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019: pp. 72–78. https://doi.org/10.18653/v1/W19-1909.
[14] A.E.W. Johnson, T.J. Pollard, L. Shen, L.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data. 3 (2016), 160035, https://doi.org/10.1038/sdata.2016.35.
[15] Branden Chan, Timo Möller, Malte Pietsch, Tanay Soni, Chin Man Yeung, Open Sourcing German BERT, Deepset. (2019). https://deepset.ai/german-bert (accessed September 14, 2019).
[16] F. Cabitza, A. Campagner, The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies, International Journal of Medical Informatics. 153 (2021) 104510, https://doi.org/10.1016/j.ijmedinf.2021.104510.
[17] W.H. Organization, ICD-10: international statistical classification of diseases and related health problems : tenth revision, World Health Organization, 2004 https://apps.who.int/iris/handle/10665/42980 (accessed February 9, 2021.
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, in, Advances in Neural Information

Processing Systems (2017) 5998–6008, https://doi.org/10.1017/S0140525X16001837.
[19] Freier Zugang zu juristischen Daten - Open Legal Data, (n.d.). https://de.openlegaldata.io/ (accessed June 21, 2020).
[20] L.A. Ramshaw, M.P. Marcus, Text chunking using transformation-based learning, in: Natural Language Processing Using Very Large Corpora, Springer, 1999: pp. 157–176.
[21] R.E. De Castilho, I. Gurevych, A broad-coverage collection of portable NLP components for building shareable analysis pipelines, in, in: Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT, 2014, pp. 1–11.
[22] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python, Zenodo (2020), https://doi.org/10.5281/zenodo.1212303.
[23] deepset.ai, deepset-ai/FARM, deepset, 2021. https://github.com/deepset-ai/FARM (accessed August 20, 2021).
[24] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, Transformers: State-of-the-art natural language processing, in: in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.
[25] MLflow - A platform for the machine learning lifecycle, MLflow. (n.d.). https://mlflow.org/ (accessed August 20, 2021).
[26] German stemming algorithm, (n.d.). http://snowball.tartarus.org/algorithms/german/stemmer.html (accessed August 20, 2021).
[27] PyStemmer, Snowball Stemming language and algorithms, 2021. https://github.com/snowballstem/pystemmer (accessed August 20, 2021).
[28] seatgeek/fuzzywuzzy, SeatGeek, 2021. https://github.com/seatgeek/fuzzywuzzy (accessed August 20, 2021).

C. Accurate Prediction of Virus-Host Protein-Protein Interactions via a Siamese Neural Network Using Deep Protein Sequence Embeddings

Reprinted with permission from:

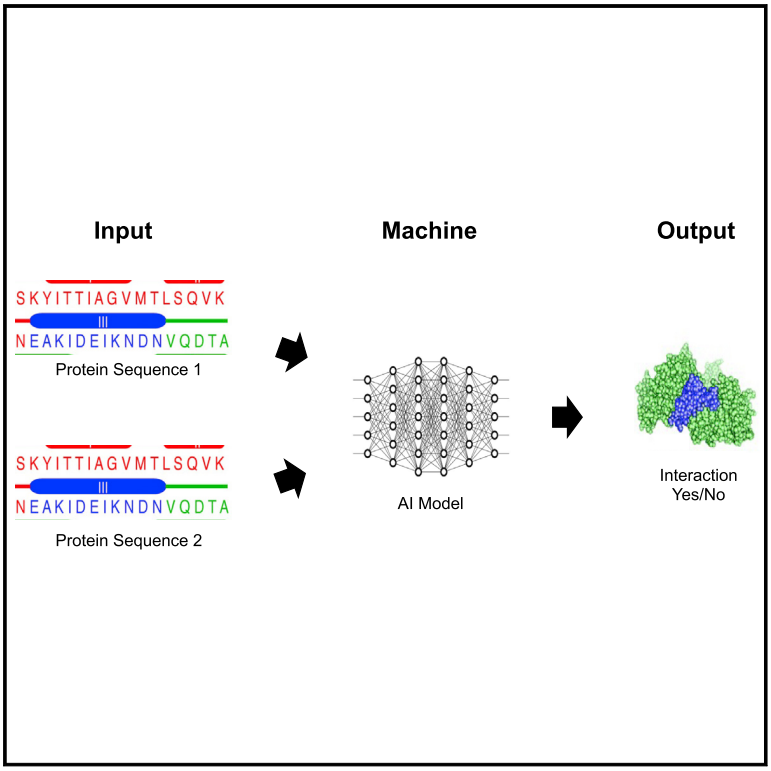
S. Madan, V. Demina, M. Stapf, O. Ernst, and H. Fröhlich, "Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings," *Patterns*, vol. 3, no. 9, p. 100 551, 2022. DOI: 10.1016/j.patter.2022.100551

Article

Patterns

Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings

Graphical abstract



Highlights

- Deep learning approach (STEP) predicts virus protein to human host protein interactions
- It is based on recent deep protein sequence embeddings and Siamese neural network
- Prediction of PPIs of the JCV VP1 protein and of the SARS-CoV-2 spike protein
- Identify parts of sequences that most likely contribute to the PPI using explainable AI

Authors

Sumit Madan, Victoria Demina, Marcus Stapf, Oliver Ernst, Holger Fröhlich

Correspondence

sumit.madan@scai.fraunhofer.de (S.M.), holger.froehlich@scai.fraunhofer.de (H.F.)

In brief

Protein-protein interaction (PPI) databases that include already-known PPIs represent an important resource in bioinformatics. A major challenge is to extend our knowledge of PPIs, which are highly relevant for the development of novel virus-like particles that can deliver therapeutics to targeted cells and tissues. Here, we use these PPI databases and the protein sequence information to train deep Siamese neural network architecture while using transfer learning and apply them to predict new virus-host PPIs with high accuracy.

Madan et al., 2022, Patterns 3, 100551
September 9, 2022
https://doi.org/10.1016/j.patter.2022.100551

CellPress

Please cite this article in press as: Madan et al., Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings, Patterns (2022), https://doi.org/10.1016/j.patter.2022.100551

Patterns

CellPress
OPEN ACCESS

Article

Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings

Sumit Madan,[1,2,*] Victoria Demina,[3] Marcus Stapf,[3] Oliver Ernst,[3] and Holger Fröhlich[1,4,5,*]
[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53757 Sankt Augustin, Germany
[2]Institute of Computer Science, University of Bonn, 53115 Bonn, Germany
[3]NEUWAY Pharma GmbH, In den Dauen 6A, 53117 Bonn, Germany
[4]Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53113 Bonn, Germany
[5]Lead contact
*Correspondence: sumit.madan@scai.fraunhofer.de (S.M.), holger.froehlich@scai.fraunhofer.de (H.F.)
https://doi.org/10.1016/j.patter.2022.100551

THE BIGGER PICTURE The development of novel cell and tissue-specific therapies requires a profound knowledge about protein-protein interactions (PPIs). Identifying these PPIs with experimental approaches such as biochemical assays or yeast two-hybrid screens is cumbersome, costly, and at the same time difficult to scale. Computational approaches can help to prioritize huge amounts of possible PPIs by learning from biological sequences plus already known PPIs. In this work, we developed an approach that is based on recent deep protein sequence embedding techniques, which we integrate into a Siamese neural network architecture. We use this approach to train models by using protein sequence information and known PPIs. We apply the models to two use cases to predict virus protein to human host interactions. Altogether our work highlights the potential of deep sequence embedding techniques as well as explainable artificial intelligence methods for the analysis of biological sequence data.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Prediction and understanding of virus-host protein-protein interactions (PPIs) have relevance for the development of novel therapeutic interventions. In addition, virus-like particles open novel opportunities to deliver therapeutics to targeted cell types and tissues. Given our incomplete knowledge of PPIs on the one hand and the cost and time associated with experimental procedures on the other, we here propose a deep learning approach to predict virus-host PPIs. Our method (Siamese Tailored deep sequence Embedding of Proteins [STEP]) is based on recent deep protein sequence embedding techniques, which we integrate into a Siamese neural network. After showing the state-of-the-art performance of STEP on external datasets, we apply it to two use cases, severe acute respiratory syndrome coronavirus 2 and John Cunningham polyomavirus, to predict virus-host PPIs. Altogether our work highlights the potential of deep sequence embedding techniques originating from the field of NLP as well as explainable artificial intelligence methods for the analysis of biological sequences.

INTRODUCTION

Viral infections can cause severe tissue-specific damage to human health. In case of the infection of brain cells, severe neurological disorders can be the consequence.[1] Accordingly, prediction and understanding of tissue-specific virus-host interactions is important for designing targeted therapeutic intervention strategies. At the same time virus-like particles (VLPs), such as John Cunningham VLPs, open novel opportunities to deliver therapeutic compounds to targeted brain cells and tissues, because

Please cite this article in press as: Madan et al., Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings, Patterns (2022), https://doi.org/10.1016/j.patter.2022.100551

these proteins have the ability to cross the blood-brain barrier.[2] Hence, it is also relevant from a therapeutic perspective to know the binding of VLPs to potential drug receptors in the brain.

The knowledge about virus-host interactions covered in databases like VirHostNet[3] is limited. While various experimental approaches exist to measure PPIs, including yeast two-hybrid screens, biochemical assays, and chromatography,[4] these methods are often time consuming, laborious, costly, and difficult to scale to large numbers of possible PPIs. Thus, computational methods have been proposed that use various types of protein information to predict PPIs. Older approaches focused on predicting PPIs either using structure and/or genomic context of proteins.[5] Other approaches[6,7] suggested classical machine learning algorithms (such as support vector machines) in combination with manually engineered features derived from protein sequences to predict PPIs.

In recent years, deep learning-based approaches[8–11] have become popular and have increasingly superseded traditional machine learning approaches for the prediction of PPIs. Often these approaches use known PPIs from established PPI databases (e.g., BioGrid, IntAct, STRING, human protein references database, VirHostNet)[3,12–15] to generate datasets to train deep neural network architectures. Some of these methods use recent network representation learning techniques to complete a known virus-host PPI graph.[16] Other authors focused on protein sequences to predict PPIs. For example, Sun et al.[8] and Wang et al.[9] proposed using a stacked autoencoder. Chen et al.[17] developed a deep learning framework using a Siamese neural architecture to predict binary and multi-class PPIs. Tsukiyama et al.[10] recently proposed a long short-term memory (LSTM)-based model on top of a classical word2vec embedding of sequences to predict human-virus PPIs by using protein sequences. Using the same embedding technique, Liu-Wei et al.[18] developed an approach that predicts host-virus PPIs for multiple viruses considering their taxonomic relationships.

In the last few years, transfer learning-based approaches from the natural language processing (NLP) area have massively impacted the field of protein bioinformatics.[19–21] These methods are trained on a huge amount of protein sequences to learn informative features of protein sequences. For instance, Elnaggar et al.[19] used 2.1 billion protein sequences for the pre-training of ProtTrans, a collection of transformer models originally stemming from the NLP field. Such methods allow the transformation of a protein sequence into a vector representation, which can subsequently be used efficiently for various downstream tasks, e.g., protein family classification.[22] There are several advantages of using the available pre-trained transformer models, such as avoiding the error-prone design of hand-crafted features to encode protein sequences and, correspondingly, a much more efficient development of new AI models with a potentially higher prediction performance.

In this article, we introduce a novel deep learning architecture combining the recently published ProtBERT[19] deep sequence embedding approach with a Siamese neural network to predict PPIs by using the primary sequences of protein pairs. While recent publications generally follow a similar strategy, they have used more traditional sequence embedding methods.[10] To our knowledge, our work thus constitutes the first attempt to evaluate the use of the most recent, pre-trained transformer models to obtain a deep learning-based biological sequence embedding for PPI prediction. After evaluating the promising prediction performance of our method (Siamese Tailored deep sequence Embedding of Proteins [STEP]), we use it for two cases: (i) predicting interactions of the John Cunningham polyomavirus (JCV) major capsid protein VP1 (UniProt:P03089) with human receptors in the brain, and (ii) predicting interactions of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike glycoprotein (UniProt:P0DTC2) with human receptors. Predicted interactions in both cases demonstrate a clear interpretation in the light of existing literature knowledge, hence supporting the biological relevance of predictions made by our method.

In this study, we make four contributions to the state-of-the-art. First, we construct a novel deep learning architecture STEP for virus-host PPI prediction that requires only the protein sequences as the input and discards the need of handcrafted or other types of features. Second, we demonstrate that using transformer-based models for PPI prediction achieves at least state-of-the-art performance for PPI prediction. In computer vision and NLP, such transformer-based models have shown that they are well suited for learning contextual relationships hidden in sequential data. However, these have not yet been applied to the field of PPI prediction. Hence, we use and build on the huge effort of Elnaggar et al.,[19] who published a pre-trained ProtBERT model that was trained on more than 2 billion amino acid sequences. In addition, we demonstrate that using transfer learning in STEP achieves state-of-the-art performance, for which we evaluated STEP on multiple publicly available virus-host and host-host PPI datasets. Third, we predict interactions for two viruses that are known to cause serious diseases and provide an interpretation on those predictions demonstrating the support through existing literature knowledge. Last, we show how experimental explainable AI (XAI) techniques could be used to identify regions in protein amino acid sequences that attribute to the prediction of PPI.

RESULTS

Comparative evaluation of STEP with state-of-the-art work

We performed a head-to-head comparison of our STEP architecture (Figure 2) on three different datasets published by Tsukiyama et al.,[10] Guo et al.,[23] and Sun et al.[8] Tsukiyama et al.[10] recently published the LSTM-PHV Siamese model, which uses a more traditional word2vec sequence embedding. The dataset published by the authors consists of host-virus PPIs that were retrieved through the Host-Pathogen Interaction Database[24] 3.0. In total, the dataset consists of 22,383 PPIs with 5,882 human and 996 virus proteins. Additionally, it includes artificially sampled negative instances with the positive to negative ratio of 1:10. The authors themselves compared LSTM-PHV on their dataset against a random Forest approach by Yang et al.[25] Guo et al.[23] published a yeast PPI dataset and used support vector machines to build a PPI detection model. Sun et al.[8] created a dataset using human protein references database, which contains human-human PPIs. Tsukiyama et al.[10] and Guo et al.[23] performed a five-fold cross-validation (CV) experiment, whereas Sun et al.[8] used a 10-fold CV setting. We evaluated our STEP

Please cite this article in press as: Madan et al., Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings, Patterns (2022), https://doi.org/10.1016/j.patter.2022.100551

CellPress
OPEN ACCESS

Table 1. Overview of the results of comparative evaluation of STEP on LSTM-PHV,[10] yeast,[23] and human PPI[8] datasets

	AUC	AUPR	F_1	MCC
Comparative analysis on host-virus PPI dataset from Tsukiyama et al.[10] via 5-fold CV				
Tsukiyama et al.[10]	97.58% (±0.13%)	93.86% (±0.35%)	91,00% (±0.53%)	90.30% (±0.53%)
STEP (ours)	98.72% (±0.16%)*	95.71% (±0.51%)*	91.53% (± 0.65%)*	90.82% (±0.72%)*
Comparative analysis on single independent host-virus PPI test dataset from Tsukiyama et al.[10]				
Yang et al.[25]	96.30%	81.00%	72.40%	69.70%
Tsukiyama et al.[10]	97.30%	93.80%	91.10%*	90.40%*
STEP (ours)	98.50%*	94.50%*	89.69%	88.76%
Comparative analysis on Yeast PPI dataset from Guo et al.[23] via 5-fold CV				
Guo et al.[23]	NA	NA	87.34% (±1.33)	75.09% (±2.51%)
Chen et al.[17]	NA	NA	97.09% (±0.23%)	94.17% (±0.48%)
STEP (ours)	99.61% (±0.10%)	99.58% (±0.17%)	97.37% (±0.27%)*	94.77% (±0.54%)*
Comparative analysis on Human PPI dataset from Sun et al.[8] via 10-fold CV				
Sun et al.[8]	NA	NA	97.15%	NA
STEP (ours)	99.74% (±0.03%)	99.66% (±0.04%)	98.84% (±0.09%)*	97.67% (±0.18%)

NA, not available in original publication.
For LSTM-PHV and Yeast PPI datasets, we applied a 5-fold CV similar to the authors of the given studies. For the Human PPI dataset of Sun et al.,[8] we applied a 10-fold CV for training the STEP models. The highest values are marked with asterisks. More details of each experiment can be found in Tables S1–S3.

architecture using the exact same datasets with the exact same data splits as the authors of the compared methods. STEP was initialized with the hyperparameters shown in Table S1. Table 1 shows the results of all experiments, demonstrating at least state-of-the-art performance of our method. Additionally, we can conclude that our approach compared on exactly the same data published by Tsukiyama et al.[10] performs similar to their LSTM-PHV method and better than the approach by Yang et al[25]

Finally, we also evaluated our STEP architecture on two additional tasks, namely, PPI type prediction and a PPI binding affinity estimation using the data and the CV setup provided by Chen et al.[17] For both tasks, we reached at least state-of-the-art performances with our approach (see Note S1.1. and Table S4).

Prediction of JCV major capsid protein VP1 interactions

We split the brain tissue-specific interactome dataset including all positive and pseudo-negative interactions into training (60%), validation (20%), and test (20%) datasets. The validation set was used for tuning hyperparameters of the model only (see Table S5). After tuning on the validation set, we used our best model to make predictions on the hold-out test set. Figure 1 illustrates the area under receiver operator characteristic curve (AUC) and precision-recall curve (AUPR). The model achieved an AUC and AUPR of 88.78% and 88.32% on the unseen test set, respectively. Also, on an extended test set with a ratio 1:10 of positive to pseudo-negative samples the results are quite stable (see Table S6).

We used this STEP-brain model to predict interactions of the JCV major capsid protein VP1 with all human receptors. Table 2 shows the top 10 predicted interactions that are ranked by the score retrieved by the logistic output function of the model. File S3 contains all the predicted interactions. According to the method of integrated gradients, large parts of the VP1 sequence contribute to our model's prediction of the PPI with the top ranked receptor KIAA1549 (Figure S4). More specifically, signal peptide *N*-regions in KIAA1549 negatively contribute to the predicted class, whereas the beginning of the non-cytoplasmic domain region is contributing positively.

Altogether, we observed a strong enrichment of VP1 interactions predicted with olfactory, serotonin, amine, taste, and acetylcholine receptors (Figure S2). Notably, neurotransmitter (and specifically serotonin) receptors have previously been suggested to be the entry of the virus into myelin-producing glial brain cells,[26] causing progressive multifocal leukoencephalopathy as a fast progressing and life-threatening neurodegenerative disorder.[27] Furthermore, we found an enrichment of tyrosine kinase activity (Figure S3), which is in line with the fact that tyrosine kinase inhibitors have been suggested as therapy against JCV.[28,29]

We further performed an enrichment analysis with InterPro[30] protein domains for the predicted interactions between JCV major capsid protein VP1 and human receptors (Figure S5, Table S7). In line with the gene ontology (GO) enrichment analysis, the two top-ranked protein domains Inter-Pro:IPR006029 and Inter-Pro:IPR006202 are neurotransmitter-gated ion channel transmembrane domains that open transiently upon binding of specific ligands, which then allow transmission of signals at chemical synapses.[31,32] Furthermore, the receptor-type tyrosine-protein phosphatase/carbonic anhydrase domain is enriched, which is in line with the enrichment of tyrosine kinase activity found via GO analysis. The enriched domains Inter-Pro:IPR013106 (immunoglobulin V-set domain) and Inter-Pro:IPR007110 (immunoglobulin-like domain) are both immunoglobulin-like domains that are involved in cell-cell recognition, cell surface receptors, and immune system response,[33] which play a role in the recognition of a virus protein.

Please cite this article in press as: Madan et al., Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings, Patterns (2022), https://doi.org/10.1016/j.patter.2022.100551

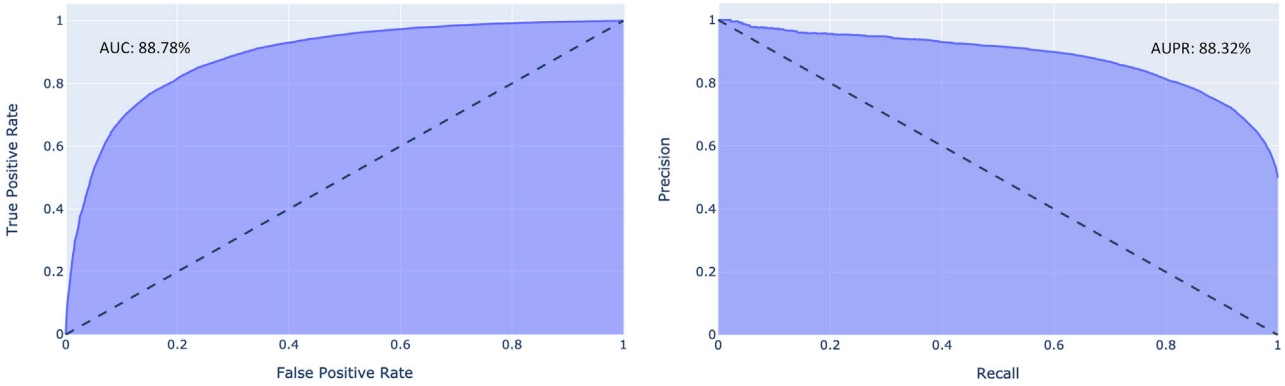


Figure 1. Receiver operator characteristic (ROC) curve (left) and AUPR (right) obtained by applying the STEP-brain model on unseen test data

Prediction of SARS-CoV-2 spike glycoprotein interactions

We performed a nested CV procedure on the given SARS-CoV-2 interactions dataset. We used five outer and five inner loops to validate the generalization performance and while performing the hyperparameter optimization in the inner loop. In each outer run, we created a stratified split of the interactome into train (4/5) and test (1/5) datasets. In the nested run, we further split the outer train dataset into train (1/5) and validation (1/5) datasets, which were used to optimize the hyperparameters of the model using the respective training data. The performance of the classifiers was evaluated with AUC and was averaged over all nested runs. The best identified hyperparameters (see Table S8) were used to train the models in the outer loop. We retrieved a final generalization performance of 83.42% (±3.91%) AUC and 84.02% (±4.58%) AUPR that was calculated by averaging the prediction results of the outer loop (see Table 3). On an extended test set with a ratio 1:10 of positive to pseudo-negative samples, the results are stable for the AUC; however, the AUPR decreases significantly (Tables S9 and S10).

We used the STEP-virus-host model obtained from the best outer fold to predict interactions of the SARS-CoV-2 spike protein (alpha, delta, and omicron variants) with all human receptors that were not already contained in VirHostNet (see Tables S11–S13). File S4 contains all the predicted interactions for the omicron variant. Interestingly, for all virus variants the sigma intracellular receptor 2 (GeneCards:TMEM97; UniProt:Q5BJF2) was the only one predicted with an outstanding high probability (of >70% in all cases) (Tables S11–S13). The sigma 1 and 2 receptors are thought to play a role in regulating cell survival, morphology, and differentiation.[34,35] In addition, the sigma receptors have been proposed to be involved in the neuronal transmission of SARS-CoV-2.[36] They have been suggested as targets for therapeutic intervention.[37–39] Our results suggest that the antiviral effect observed in cell lines treated with sigma receptor binding ligands might be due to a modulated binding of the spike protein, thus inhibiting virus entry into cells. In this context, an analysis via the integrated gradients method shows that only parts of the sigma 2 receptor and the SARS-CoV-2 spike protein contribute to our model's prediction of the PPI (Figure S6). More specifically, the non-cytoplasmic domain and EXPERA domains demonstrate positive integrated gradient scores, i.e., the existence of these domains influences our model to make the according prediction.

DISCUSSION

Huge advancements have been made recently by applying deep learning algorithms from NLP to protein bioinformatics. Protein language models such as ProtTrans and ProtBERT,[19] which are trained on billions of protein sequences, learn informative features through the transformation of sequences to vector representations. These models previously showed their predictive power in various tasks such as prediction of secondary structure or classification of membrane proteins.[19]

In our work, we used ProtBERT within a specifically designed Siamese neural network architecture to predict PPIs by only using the primary sequences of protein pairs. We trained our models following a positive unlabeled (PU) learning scheme and performed an extensive evaluation and hyperparameter optimization of our models, demonstrating high prediction performances for virus protein to human receptor interactions of JCV and SARS-CoV-2. An additional head-to-head comparison with the recently published method by Tsukiyama et al.[10] using a more traditional word2vec sequence embedding combined with an LSTM unit revealed state-of-the-art prediction performance of our STEP approach.

Interactions predicted by our proposed model between JCV major capsid protein VP1 and receptors in brain cells showed a strong enrichment of different neurotransmitters, including serotonin receptors, which is in line with the current literature. For the SARS-Cov-2 spike protein, our model interestingly predicted for all virus variants an interaction with the sigma intracellular receptor 2, which might explain the cytopathic effects of sigma receptor binding ligands reported in the literature.[38–40] In both cases, recent techniques coming from the field of XAI allowed us to interpret model predictions and identify those parts of protein sequences that, according to our model, mostly influence the prediction of respective PPIs. Of course, a validation of these predictions would require experimental procedures that are beyond the scope of this article.

Altogether, our work demonstrates the potential of modern deep learning-based biological sequence embeddings and modern XAI techniques for bioinformatics. While in this article

Please cite this article in press as: Madan et al., Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings, Patterns (2022), https://doi.org/10.1016/j.patter.2022.100551

Table 2. Top 10 predicted interactions of the JCV major capsid protein VP1 and human receptors ranked by the probability obtained by our model

Rank	Receptor protein ID	Receptor protein name	Score (in %)	Associated GO molecular function
1	Q9HCM3	UPF0606 protein KIAA1549	99.31	–
2	O94991	SLIT and NTRK-like protein 5	99.09	protein binding
3	Q7Z443	polycystic kidney disease protein 1-like 3	98.68	calcium channel activity, sour taste receptor activity
4	O60840	voltage-dependent L-type calcium channel subunit alpha-1F	98.63	high voltage-gated calcium channel activity, metal ion binding
5	P13611	versican core protein	98.51	calcium ion binding, hyaluronic acid binding, glycosaminoglycan binding, extracellular matrix structural constituent conferring compression resistance
6	P23471	receptor-type tyrosine-protein phosphatase zeta	98.33	protein tyrosine phosphatase activity, integrin binding, protein binding, phosphatase activity, hydrolase activity, phosphoprotein phosphatase activity, transmembrane receptor protein tyrosine phosphatase activity
7	Q8N2Q7	neuroligin-1	98.33	neurexin family protein binding, signaling receptor activity, identical protein binding, cell adhesion molecule binding, scaffold protein binding, PDZ domain binding, amyloid-beta binding
8	Q9BZV3	interphotoreceptor matrix proteoglycan 2	98.23	heparin binding, hyaluronic acid binding, extracellular matrix structural constituent
9	P41968	melanocortin receptor 3	98.19	peptide hormone binding, G protein-coupled receptor activity, melanocyte-stimulating hormone receptor activity, neuropeptide binding, melanocortin receptor activity
10	P23470	receptor-type tyrosine-protein phosphatase gamma	98.14	protein tyrosine phosphatase activity, identical protein binding, phosphatase activity, transmembrane receptor protein tyrosine phosphatase activity, hydrolase activity, phosphoprotein phosphatase activity

we focused on JCV and SARS-CoV-2, our proposed model could in future work be easily trained to predict interactions of other viruses as well and, thus, contribute to the emerging set of computational methods that might help to respond to future epidemic and pandemic situations more effectively. In addition, there is the potential to use our method in the context of modern drug development approaches, which use virus-like particles to deliver compounds to specific tissues and receptors.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for code and data should be directed to and will be fulfilled by the lead contact, Holger Fröhlich (holger.froehlich@scai.fraunhofer.de).

Materials availability

This study did not generate any physical materials.

Data and code availability

The data and source code are available at https://github.com/SCAI-BIO/STEP.

Construction of datasets

Primary data sources

The following primary resources were used to create training and test datasets in this work:

1. UniProt protein sequence dataset[41] containing human protein sequences.
2. UniProt mapping dataset[41] containing mappings to other databases.
3. VirHostNet dataset[3] including virus-host interactions of SARS-CoV-2 spike glycoprotein.
4. PPT-Ohmnet dataset[42] (https://snap.stanford.edu/biodata/datasets/10013/10013-PPT-Ohmnet.html, accessed November 18, 2021) containing brain tissue-specific protein-protein-interactions.
5. The GO[43] receptor protein dataset containing annotation of proteins as receptors and parts of protein complexes.
6. Sequences of JCV major capsid protein VP1 (https://www.uniprot.org/uniprot/P03089, accessed on 18 November 2021) and SARS-CoV-2 spike glycoprotein (https://www.uniprot.org/uniprot/P0DTC2, accessed November 18, 2021).
7. Pathogen-host PPI training and test set provided by Tsukiyama et al.[10] (http://kurata35.bio.kyutech.ac.jp/LSTM-PHV/download_page, accessed November 18, 2021) (used for comparative analysis).
8. Yeast PPI dataset from Guo et al.[23] (used for comparative analysis).
9. Human PPI dataset from Sun et al.[8] (used for comparative analysis).
10. PPI type prediction dataset SHS27k from Chen et al.[17] (used for comparative analysis).
11. PPI binding affinity estimation dataset from Chen et al.[17] (used for comparative analysis).

Construction of brain-specific protein-protein interactome dataset

We chose the PPT-Ohmnet database[42] that includes tissue-specific human PPIs collected from various sources. PPT-Ohmnet only takes physical PPIs into account that are supported by experimental evidence (https://snap.stanford.edu/biodata/datasets/10013/10013-PPT-Ohmnet.html). More specifically, interactions contained in PPT-Ohmnet were collected from various curated databases such as TRANSFAC, IntAct, and MINT.[44] The tissue information for an interaction was inferred through the low-throughput tissue-specific gene expression data.[45] The protein-protein interactome can be considered as a graph, in which the proteins represent nodes and the interactions between them are considered as edges. Furthermore, every edge contains

Please cite this article in press as: Madan et al., Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings, Patterns (2022), https://doi.org/10.1016/j.patter.2022.100551

Table 3. Results of the outer loop folds retrieved during the nested CV of STEP-virus-host model by using the test set with a ratio of 1:1 positive to pseudo-negative instances

Outer fold	AUC	AUPR
1	88.17%	89.93%
2	86.83%	88.62%
3	77.03%	77.73%
4	82.52%	81.67%
5	82.56%	82.15%
Mean	83.42% (± 3.91%)	84.02% (±4.58%)

the information about the tissue type. In total, there are 144 tissue types with 4,510 proteins (nodes) and about 3,666,563 non-unique edges (interactions) in the whole PPT-Ohmnet graph. More details about the creation and content of the PPT-Ohmnet database can be found in Menche et al.[44] and Greene et al.[45]

We extracted all tissue types and manually filtered the ones specific for the brain. In total, 36 brain-specific tissue types could be found from a total of 144 in the PPT-Ohmnet database (Figure S1). Using the information about brain tissue specific co-expression of proteins, we filtered the PPT-Ohmnet interactome. The final brain tissue-specific interactome contains 3,548 proteins (nodes) and 977,990 non-unique edges (interactions). Furthermore, the interactome contains 56,021 unique edges, from which 1,466 PPIs that interact with themselves were excluded. In total, 54,555 PPIs were used for further analysis. Figure S1 shows the distribution of proteins and their interactions for each brain-specific tissue type. File S1 contains the brain-specific tissue types.

We further enriched each interaction with information about the experimental detection methods that were used. This information is not included in PPT-Ohmnet; hence, we used BioGRID and IntAct as the two largest PPI databases to extract the experimental procedures, such as "pull down," "two hybrid," by which the interactions were originally discovered. The list of experimental procedures was further manually curated to filter out detection methods considered as unreliable. Only PPIs detected by methods considered as reliable were used for further processing.

To train deep learning models, we retrieved the sequences of all proteins in our PPIs from the UniProt database. We downloaded the human proteins dataset from the manually curated part of UniProt—the so-called SwissProt.[41] Next, we extracted for all proteins their sequences and metadata such as name, ID, and label. In total, sequences for 20,396 human proteins could be found. Finally, we filtered the PPIs and human receptor proteins for which we found the sequences.

Construction of SARS-CoV-2 protein-protein interactome dataset

As a second dataset, we used the VirHostNet[3] database to collect all PPIs between SARS-CoV-2 and human proteins. We extracted for all human and SARS-CoV-2 proteins their sequences and metadata such as name, ID, and label from SwissProt. Our VirHostNet interactome contained 334 PPIs involving 338 proteins between SARS-CoV-2 and *Homo sapiens*.

Collection of human receptor proteins

To extract human receptor proteins, we first performed a search in GO for the term "receptor." The GO branch annotation "cellular components" was used to filter only for proteins. The GO annotation "organism" was used to filter for human proteins. In total, 2,075 results were found, in which 2,059 human receptor proteins and 16 human protein complexes were included. For further analyses, we only focused on human receptor proteins, for which we retrieved associated protein sequences from SwissProt. In total, sequences for 2,027 human receptor proteins could be found. File S2 includes the list of identified human receptor proteins.

Preparation for PU learning

The goal of PPI detection is to learn a model that is able to detect whether there exists an interaction between two proteins. This task is often considered as a binary classification problem that can be solved by training a classifier to distinguish between positive and negative instances. However, the available PPI databases just contain positive, true interactions. Interactions not listed in a PPI database might still exist, but are possibly unknown today. PU learning is a scheme where a machine learning algorithm only has access to positive and unlabeled instances.[46,47] In PU learning all non-existent or unknown PPIs can be considered as "unlabeled" or as "pseudo-negatives"; however, they might also contain an unknown fraction of positive instances. Therefore, PU learning amounts to constructing a binary classifier that ranks instances with respect to the positive class conditional probability.

A popular strategy of PU learning is to first focus on the selection of reliable negative instances. In a second step, a conventional binary classifier is trained on positive and selected negative instances.[46] There are two types of strategies to sample pseudo-negative instances: random sampling or similarity-based sampling. With the random sampling strategy, the negative instances are created by randomly exchanging one of the partners in an interaction protein pair. While the similarity-based sampling considers the sequence similarity (or dissimilarity) of proteins. An example of this strategy is the dissimilarity-random-sampling method,[48] also used by Tsukiyama et al.,[10] which follows the hypothesis that, if two viral proteins have similar sequences, a human protein that interacts with one of them cannot be paired with the other as a negative example. A sampling of highly dissimilar negative samples might result in overly optimistic classification performances.[10] Therefore, in our work, we applied the random sampling approach to create negative instances. A major challenge in this context is the high-class imbalance between positive and unlabeled training instances in our data. Hence, we decided to randomly subsample an equal number of pseudo-negatives.

Architecture and transfer learning of STEP

We used a deep Siamese neural network architecture while using transfer learning to learn relevant, latent features of PPI pairs based on protein sequences.

ProtBERT: Pre-trained embeddings of protein sequences

ProtBERT[19] is a pre-trained model trained on approximately 2 billion protein sequences using a masked language modeling objective.[49] It is based on the BERT model[49] that was developed for the natural language domain. Hereby, ProtBERT considers protein sequences as sentences and the so-called building blocks of proteins—amino acids—as vocabulary. The ProtBERT model, specifically the BFD variant[19] used in this work, consists of 30 layers with 16 attention heads and 1,024 hidden layers. It was trained by using the Lamb[50] optimizer for around 23.5 days on 128 compute nodes each containing 1,024 tensor processing units. During training, the language model learns to extract the biophysical characteristics of proteins from billions of protein sequences.

Siamese neural network architecture

Given a pair of proteins, we first obtained their sequences. These sequences were then fed into a Siamese model architecture (Figure 2), in which the pre-trained ProtBERT model was used to obtain embeddings of both protein sequences. There are various ways to infer the relation between sequence embeddings. Some researchers focus on concatenation and others focus on element-wise multiplication (also known as Hadamard product) of both sequence embeddings. In this work, we implemented an integration layer that uses the Hadamard product to combine the sequence embeddings, as it is often found to be the most effective way to model symmetric characteristics of proteins.[17]

Classification head for PU learning

On top of the integration layer, we added a classification head represented by multiple hidden layers (Figure 2). We designed the classification head as a bottleneck-shaped architecture with a combination of dropout and linear layers, which ended in an output layer using a logistic function and thus allowed to rank protein pairs as either more likely to interact (positive) or not (negative). Notably, a network with bottleneck structure introduces a gradual decrease of the number of neurons per layer that allows the network to focus on relevant information and discards redundant or irrelevant information.

Evaluation criteria

We evaluated our models using an independent test dataset. This consisted of a defined fraction of known PPIs taken at random and excluded from training plus a specified fraction of pseudo-negatives that were not part of the training set. The performance was measured using the AUC and the AUPR.

It should be re-emphasized that in our data negative samples are those protein pairs for which an interaction is unknown. Therefore, we evaluated the

Please cite this article in press as: Madan et al., Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings, Patterns (2022), https://doi.org/10.1016/j.patter.2022.100551

CellPress
OPEN ACCESS

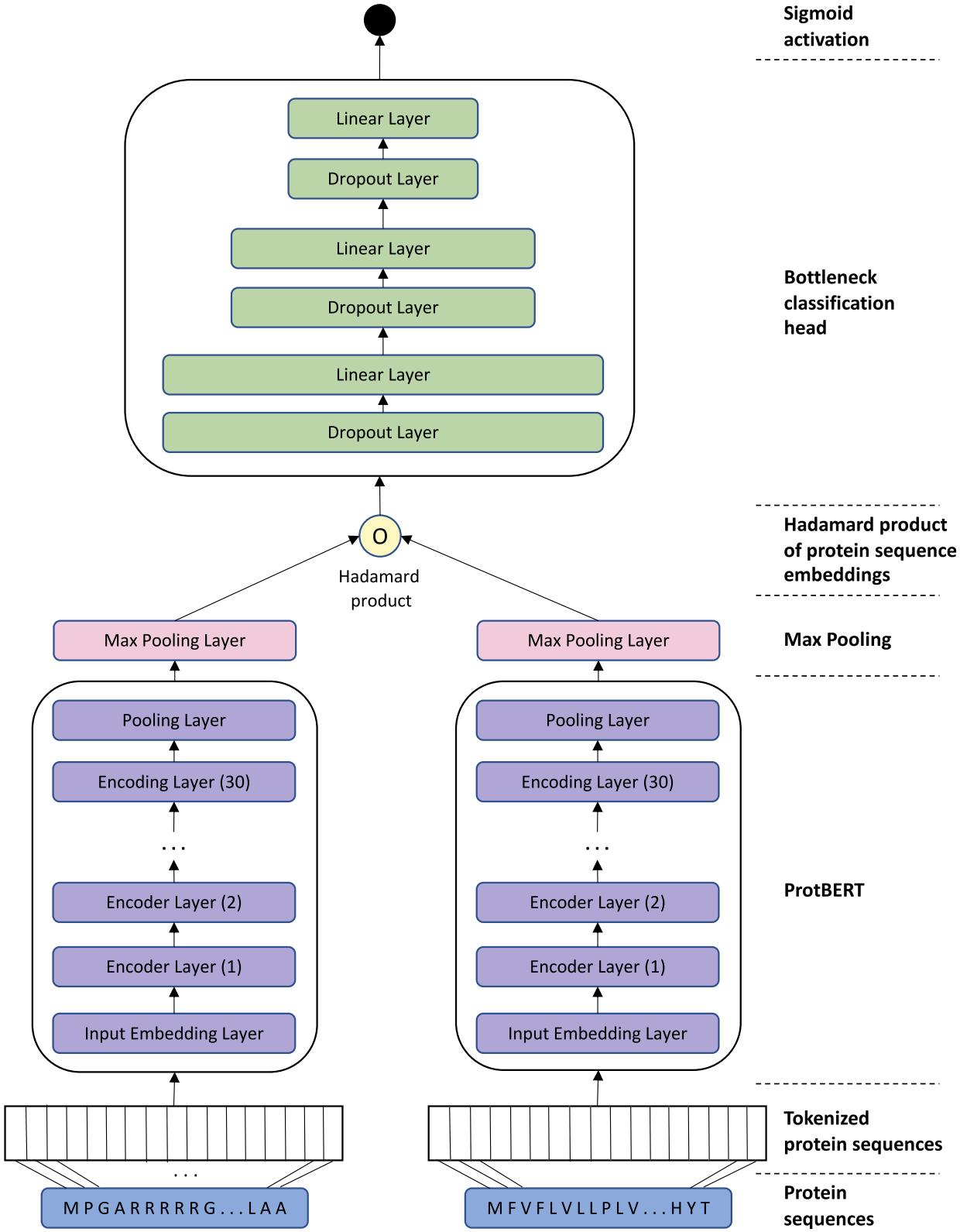


Figure 2. Architecture of our STEP model that uses the Siamese neural network while using the ProtBERT embeddings

Please cite this article in press as: Madan et al., Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings, Patterns (2022), https://doi.org/10.1016/j.patter.2022.100551

ability of our models to enrich true positives at the beginning of a predicted ranking of potential PPIs. This ability is exactly reflected by AUC and AUPR measures, which are thus frequently used in the literature about PU learning.[47] Notably, from a theoretical point of view the AUC estimated via PU learning and the one from a fully labeled dataset are provably linearly correlated.[51]

Hyperparameter optimization

To tune our system, we performed an extensive Bayesian hyperparameter optimization[52] using the training data. Owing to the huge amount of training time for a single trial, hyperparameter candidates were evaluated using a single validation set consisting of a specified fraction of known PPIs plus an equal amount of sub-sampled negatives. For each trial, intermediate and final performances were assessed using the AUC measure and captured in an SQL database for later analyses. The captured data were also used by the pruning process of Optuna to stop unpromising trials at an early stage.[53] Each optimization trial was executed on a 2× A100 NVIDIA GPUs with VMEM of 32 GB and five trials were executed parallelly by using 10× GPUs in total. The whole optimization process took 10 full days by executing 116 trials in total. The evaluated hyperparameter ranges and the best parameters are illustrated in Tables S5 and S8.

Making STEP models explainable: An analysis of integrated gradients

One of the main criticisms of modern deep learning approaches is their often-perceived black box character. To address this concern, we aimed to understand the influence of individual amino acids on model predictions. For that purpose, we used the integrated gradients method,[54] which offers an intuitive and mathematically sound approach to explain predictions made by a deep neural network. Integrated gradients require no modifications to the trained model. Given an input sample ($x \in R^n$), integrated gradients rely on a baseline/reference input sample ($x' \in R^n$), which we constructed using the concatenation of one class, multiple padding, and one separator token. For a STEP model $F : R^n \rightarrow [0, 1]$, integrated gradients are then obtained by accumulating the partial derivatives $\frac{\partial F(x)}{\partial x_i}$ with respect to input feature i while moving from the reference x' to the observed input x:[54]

$$IntegratedGrads_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

We used 1,000 steps to approximate the integrated gradients, as suggested by Sundararajan et al.[54] for highly nonlinear networks.

Gene set enrichment analysis

To better understand the biology of all ranked predictions in the individual use cases, we performed a gene set enrichment analysis to investigate an enrichment of gene sets listed in the Molecular Signatures Database[55] (MsigDB). We downloaded molecular function gene sets of the GO included as the collection C5 from MsigDB (v7.4, MsigDB/c5.go.mf.v7.4.symbols.gmt and MsigDB/c5.go.bp.v7.4.symbols.gmt). We considered a GO term to be statistically significant if, after applying the multiple hypothesis testing correction with the Benjamini-Hochberg method,[56] its adjusted p value was less than 0.01.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100551.

ACKNOWLEDGMENTS

We thank André Gemünd for his support regarding the computational infrastructure of Fraunhofer SCAI. We thank NEUWAY Pharma GmbH who provided the funding for the work presented in this study.

AUTHOR CONTRIBUTIONS

Conceptualization, O.E. and H.F.; Methodology, H.F. and S.M.; Data Curation, Formal Analysis, Visualization, Investigation, Validation, S.M.; Supervision, H.F.; Project Administration, V.D., M.S., O.E., and H.F.; Writing—Original Draft, S.M. and H.F.; Writing—Review and Editing, S.M., V.D., M.S., O.E., and H.F.

DECLARATION OF INTERESTS

V.D., M.S., and O.E. are employees of NEUWAY Pharma GmbH. The company funded the work presented in this article but had no influence on scientific results.

Received: March 11, 2022
Revised: March 28, 2022
Accepted: June 16, 2022
Published: July 28, 2022

REFERENCES

1. Swanson, P.A., and McGavern, D.B. (2015). Viral diseases of the central nervous system. Curr. Opin.Virol. *11*, 44–54. https://doi.org/10.1016/j.coviro.2014.12.009.
2. Ye, D., Zimmermann, T., Demina, V., Sotnikov, S., Ried, C.L., Rahn, H., Stapf, M., Untucht, C., Rohe, M., Terstappen, G.C., et al. (2021). Trafficking of JC virus-like particles across the blood–brain barrier. Nanoscale Adv. *3*, 2488–2500. https://doi.org/10.1039/d0na00879f.
3. Guirimand, T., Delmotte, S., and Navratil, V. (2015). VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. Nucleic Acids Res. *43*, D583–D587. https://doi.org/10.1093/nar/gku1121.
4. Lalonde, S., Ehrhardt, D.W., Loqué, D., Chen, J., Rhee, S.Y., and Frommer, W.B. (2008). Molecular and cellular approaches for the detection of protein–protein interactions: latest techniques and current limitations. Plant J. *53*, 610–635. https://doi.org/10.1111/j.1365-313x.2007.03332.x.
5. Skrabanek, L., Saini, H.K., Bader, G.D., and Enright, A.J. (2008). Computational prediction of protein–protein interactions. Mol. Biotechnol. *38*, 1–17. https://doi.org/10.1007/s12033-007-0069-2.
6. Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. (2007). Predicting protein–protein interactions based only on sequences information. Proc. Natl. Acad. Sci. USA *104*, 4337–4341. https://doi.org/10.1073/pnas.0607879104.
7. Zhou, X., Park, B., Choi, D., and Han, K. (2018). A generalized approach to predicting protein-protein interactions between virus and host. BMC Genom. *19*, 568. https://doi.org/10.1186/s12864-018-4924-2.
8. Sun, T., Zhou, B., Lai, L., and Pei, J. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinf. *18*, 277. https://doi.org/10.1186/s12859-017-1700-2.
9. Wang, Y.-B., You, Z.-H., Li, X., Jiang, T.-H., Chen, X., Zhou, X., and Wang, L. (2017). Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. Mol. Biosyst. *13*, 1336–1344. https://doi.org/10.1039/c7mb00188f.
10. Tsukiyama, S., Hasan, M.M., Fujii, S., and Kurata, H. (2021). LSTM-PHV: prediction of human-virus protein–protein interactions by LSTM with word2vec. Briefings Bioinf. *22*, bbab228. https://doi.org/10.1093/bib/bbab228.
11. Xu, W., Gao, Y., Wang, Y., and Guan, J. (2021). Protein–protein interaction prediction based on ordinal regression and recurrent convolutional neural networks. BMC Bioinf. *22*, 485. https://doi.org/10.1186/s12859-021-04369-0.
12. Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., et al. (2021). The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. *30*, 187–200. https://doi.org/10.1002/pro.3978.
13. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., Del-Toro, N., et al. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. *42*, D358–D363. https://doi.org/10.1093/nar/gkt1115.
14. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019).

STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613. https://doi.org/10.1093/nar/gky1131.

15. Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human protein reference database—2009 update. Nucleic Acids Res. *37*, D767–D772. https://doi.org/10.1093/nar/gkn892.
16. Du, H., Chen, F., Liu, H., and Hong, P. (2021). Network-based virus-host interaction prediction with application to SARS-CoV-2. Patterns *2*, 100242.
17. Chen, M., Ju, C.J.-T., Zhou, G., Chen, X., Zhang, T., Chang, K.-W., Zaniolo, C., and Wang, W. (2019). Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. Bioinformatics *35*, i305–i314. https://doi.org/10.1093/bioinformatics/btz328.
18. Liu-Wei, W., Kafkas, Ş., Chen, J., Dimonaco, N.J., Tegnér, J., and Hoehndorf, R. (2021). DeepViral: prediction of novel virus–host interactions from protein sequences and infectious disease phenotypes. Bioinformatics *37*, 2722–2729. https://doi.org/10.1093/bioinformatics/btab147.
19. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2021). ProtTrans: Towards Cracking the Language of Lifes Code through Self-Supervised Deep Learning and High Performance Computing (IEEE Trans Pattern Anal Mach Intell).
20. Min, S., Park, S., Kim, S., Choi, H.-S., and Yoon, S. (2019). Pre-training of deep bidirectional protein sequence representations with structural information. Preprint at arXiv. https://doi.org/10.48550/arXiv.1912.05625.
21. Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., and Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinf. *20*, 723. https://doi.org/10.1186/s12859-019-3220-8.
22. Nambiar, A., Heflin, M., Liu, S., Maslov, S., Hopkins, M., and Ritz, A. (2020). Transforming the language of life: transformer neural networks for protein prediction tasks. In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 1–8.
23. Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res. *36*, 3025–3030. https://doi.org/10.1093/nar/gkn159.
24. Ammari, M.G., Gresham, C.R., McCarthy, F.M., and Nanduri, B. (2016). HPIDB 2.0: a curated database for host–pathogen interactions. Database *2016*, baw103.
25. Yang, X., Yang, S., Li, Q., Wuchty, S., and Zhang, Z. (2020). Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. Comput. Struct. Biotechnol. J. *18*, 153–161. https://doi.org/10.1016/j.csbj.2019.12.005.
26. Ferenczy, M.W., Marshall, L.J., Nelson, C.D.S., Atwood, W.J., Nath, A., Khalili, K., and Major, E.O. (2012). Molecular biology, epidemiology, and pathogenesis of progressive multifocal leukoencephalopathy, the JC virus-induced demyelinating disease of the human brain. Clin. Microbiol. Rev. *25*, 471–506. https://doi.org/10.1128/cmr.05031-11.
27. Boothpur, R., and Brennan, D.C. (2010). Human polyoma viruses and disease with emphasis on clinical BK and JC. J. Clin. Virol. *47*, 306–312. https://doi.org/10.1016/j.jcv.2009.12.006.
28. Querbes, W., Benmerah, A., Tosoni, D., Di Fiore, P.P., and Atwood, W.J. (2004). A JC virus-induced signal is required for infection of glial cells by a clathrin- and eps15-dependent pathway. J. Virol. *78*, 250–256. https://doi.org/10.1128/jvi.78.1.250-256.2004.
29. Bennett, C.L., Berger, J.R., Sartor, O., Carson, K.R., Hrushesky, W.J., Georgantopoulos, P., Raisch, D.W., Norris, L.B., and Armitage, J.O. (2018). Progressive multi-focal leucoencephalopathy among ibrutinib-treated persons with chronic lymphocytic leukaemia. Br. J. Haematol. *180*, 301–304. https://doi.org/10.1111/bjh.14322.
30. Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., et al. (2021). The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. *49*, D344–D354. https://doi.org/10.1093/nar/gkaa977.
31. Kofuji, P., Wang, J.B., Moss, S.J., Huganir, R.L., and Burt, D.R. (1991). Generation of two forms of the gamma-aminobutyric acidA receptor gamma 2-subunit in mice by alternative splicing. J. Neurochem. *56*, 713–715. https://doi.org/10.1111/j.1471-4159.1991.tb08209.x.
32. Wagner, K., Edson, K., Heginbotham, L., Post, M., Huganir, R.L., and Czernik, A.J. (1991). Determination of the tyrosine phosphorylation sites of the nicotinic acetylcholine receptor. J. Biol. Chem. *266*, 23784–23789. https://doi.org/10.1016/s0021-9258(18)54351-9.
33. Teichmann, S.A., and Chothia, C. (2000). Immunoglobulin superfamily proteins in Caenorhabditis elegans 1 1Edited by G. von Heijne. J. Mol. Biol. *296*, 1367–1383. https://doi.org/10.1006/jmbi.1999.3497.
34. Huang, Y.-S., Lu, H.-L., Zhang, L.-J., and Wu, Z. (2014). Sigma-2 receptor ligands and their perspectives in cancer diagnosis and therapy: sigma-2 receptor ligands. Med. Res. Rev. *34*, 532–566. https://doi.org/10.1002/med.21297.
35. Guo, L., and Zhen, X. (2015). Sigma-2 receptor ligands: neurobiological effects. Comput. Mater. Continua *22*, 989–1003. https://doi.org/10.2174/0929867322666150114163607.
36. Yesilkaya, U.H., Balcioglu, Y.H., and Sahin, S. (2020). Reissuing the sigma receptors for SARS-CoV-2. J. Clin. Neurosci. *80*, 72–73. https://doi.org/10.1016/j.jocn.2020.08.014.
37. Abate, C., Niso, M., Abatematteo, F.S., Contino, M., Colabufo, N.A., and Berardi, F. (2020). PB28, the sigma-1 and sigma-2 receptors modulator with potent anti–SARS-CoV-2 activity: a Review about its pharmacological properties and structure affinity relationships. Front. Pharmacol. *11*, 589810. https://doi.org/10.3389/fphar.2020.589810.
38. Das, A.B., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., White, K.M., O'Meara, M.J., Rezelj, V.V., Guo, J.Z., Swaney, D.L., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug-repurposing. Nature *583*, 459. https://doi.org/10.1186/s12920-021-01079-7.
39. Ostrov, D.A., Bluhm, A.P., Li, D., Khan, J.Q., Rohamare, M., Rajamanickam, K., Bhanumathy, K., Lew, J., Falzarano, D., Vizeacoumar, F.J., et al. (2021). Highly specific sigma receptor ligands exhibit anti-viral properties in SARS-CoV-2 infected cells. Pathogens *10*, 1514. https://doi.org/10.3390/pathogens10111514.
40. Abbate, S., Avvenuti, M., and Light, J. (2014). Usability Study of a wireless monitoring system among Alzheimer's disease elderly population. Int. J. Telemed. Appl. *2014*, 617495. https://doi.org/10.1155/2014/617495.
41. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.
42. Zitnik, M., Sosič, R., Maheshwari, S., and Leskovec, J. (2018). BioSNAP Datasets (Stanford Biomedical Network Dataset Collection).
43. Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. *32*, D258–D261. https://doi.org/10.1093/nar/gkh036.
44. Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. Science *347*, 1257601. https://doi.org/10.1126/science.1257601.
45. Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. *47*, 569–576. https://doi.org/10.1038/ng.3259.
46. Bekker, J., and Davis, J. (2020). Learning from positive and unlabeled data: a survey. Mach. Learn. *109*, 719–760. https://doi.org/10.1007/s10994-020-05877-5.

47. Sansone, E., De Natale, F.G.B., and Zhou, Z.-H. (2019). Efficient training for positive unlabeled learning. IEEE Trans. Pattern Anal. Mach. Intell. *41*, 2584–2598. https://doi.org/10.1109/tpami.2018.2860995.

48. Eid, F.-E., ElHefnawi, M., and Heath, L.S. (2016). DeNovo: virus-host sequence-based protein–protein interaction prediction. Bioinformatics *32*, 1144–1150. https://doi.org/10.1093/bioinformatics/btv737.

49. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.

50. You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. (2019). Large batch optimization for deep learning: training bert in 76 minutes. Preprint at arXiv. https://doi.org/10.48550/arXiv:1904.00962.

51. Menon, A., Rooyen, B.V., Ong, C.S., and Williamson, B. (2015). Learning from corrupted binary labels via class-probability estimation. In Proceedings of the 32nd International Conference on Machine Learning (PMLR)), pp. 125–134.

52. Bergstra, J., Yamins, D., and Cox, D. (2013). Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In International Conference on Machine Learning (PMLR), pp. 115–123.

53. Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: a next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623–2631.

54. Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1703.01365.

55. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

56. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

D. Dataset of miRNA–Disease Relations Extracted from Textual Data Using Transformer-based Neural Networks

Reprinted with permission from:

S. Madan, L. Kühnel, H. Fröhlich, M. Hofmann-Apitius, and J. Fluck, "Dataset of miRNA–disease relations extracted from textual data using transformer-based neural networks," *Database*, vol. 2024, baae066, 2024. DOI: 10.1093/database/baae066

Database, 2024, baae066
DOI: https://doi.org/10.1093/database/baae066
Original article

Dataset of miRNA–disease relations extracted from textual data using transformer-based neural networks

Sumit Madan[1,*], Lisa Kühnel[2,3], Holger Fröhlich[1,4], Martin Hofmann-Apitius[1,4], Juliane Fluck[2,3,5]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53757 Sankt Augustin, Germany
[2]Knowledge Management, German National Library of Medicine (ZB MED)—Information Centre for Life Sciences, Friedrich-Hirzebruch-Allee 4, Bonn 53115, Germany
[3]Graduate School DILS, Bielefeld Institute for Bioinformatics Infrastructure (BIBI), Faculty of Technology, Bielefeld University, Postfach 10 01 31, Bielefeld, Nordrhein-Westfalen 33501, Germany
[4]Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, Friedrich-Hirzebruch-Allee 6, Bonn 53113, Germany
[5]Information management, Institute of Geodesy and Geoinformation, University of Bonn, Katzenburgweg 1a, Bonn 53115, Germany

[*]Corresponding author. Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53757 Sankt Augustin, Germany. E-mail: sumit.madan@scai.fraunhofer.de

Citation details: Madan, S., Kühnel, L., Fröhlich, H. *et al.* Dataset of miRNA–disease relations extracted from textual data using transformer-based neural networks. *Database* (2024) Vol. 2024: article ID baae066; DOI: https://doi.org/10.1093/database/baae066

Abstract

MicroRNAs (miRNAs) play important roles in post-transcriptional processes and regulate major cellular functions. The abnormal regulation of expression of miRNAs has been linked to numerous human diseases such as respiratory diseases, cancer, and neurodegenerative diseases. Latest miRNA–disease associations are predominantly found in unstructured biomedical literature. Retrieving these associations manually can be cumbersome and time-consuming due to the continuously expanding number of publications. We propose a deep learning-based text mining approach that extracts normalized miRNA–disease associations from biomedical literature. To train the deep learning models, we build a new training corpus that is extended by distant supervision utilizing multiple external databases. A quantitative evaluation shows that the workflow achieves an area under receiver operator characteristic curve of 98% on a holdout test set for the detection of miRNA–disease associations. We demonstrate the applicability of the approach by extracting new miRNA–disease associations from biomedical literature (PubMed and PubMed Central). We have shown through quantitative analysis and evaluation on three different neurodegenerative diseases that our approach can effectively extract miRNA–disease associations not yet available in public databases.

Database URL: https://zenodo.org/records/10523046

Introduction

Short RNA molecules such as microRNAs (miRNAs) that bind to target messengerRNAs (mRNAs) play important roles in post-transcriptional processes and regulate major cellular functions [1]. Deregulation of expression of miRNAs, which impacts the gene expression patterns and disrupts cellular processes, has been associated with several human diseases such as respiratory diseases [2–4], cancer [1], and Alzheimer's disease (AD) [5–7]. Targeting disease-associated mRNAs through selected miRNAs makes these molecules interesting candidates for therapy, which is even more of significance with further clinical advancements in miRNA delivering technologies [1]. However, this requires a thorough knowledge of the involvement of specific miRNAs in normal biological processes and in diseases, which is obtained through *in vivo* and *in vitro* experiments and published in research literature.

Extraction of such miRNA–disease associations from the literature can be performed through text mining techniques. In the past, Bagewadi *et al.* [8] proposed the extraction of miRNA, species, genes/proteins, and disease annotations and their relations by creating new corpora and utilizing rule-based methods (such as regular expressions) and machine learning methods (such as support vector machines). They reached an F1-score of up to 76% for miRNA relations. In addition, Li *et al.* [9] created a rule-based text mining system called miRTex that focused on extracting just miRNA–gene and gene–miRNA regulation relations from scientific literature. Their final system achieved an F1-score of 88% on a test set of 150 PubMed abstracts; however, the recall (81%) was significantly lower than precision (96%), which is a common characteristic of a rule-based system. Gupta *et al.* [10] proposed the miRiaD text mining tool, which reached an F1-score of 89.4% on a set of 200 sentences, to extract miRNA–disease

Received 21 January 2024; Revised 23 June 2024; Accepted 10 July 2024

relations from the entire Medline identifying 8301 abstracts containing such relations. The tool BeFree, proposed by Bravo *et al.* [11], that exploits morphosyntactic information of the text reached an F1-score of 85% for the extraction of gene–disease associations that also includes miRNAs. The results of BeFree are integrated in the DisGeNET database, a platform for disease genomics [12].

In the meantime, transformer-based general language models such as Bidirectional Encoder Representations from Transformes (BERT) [13] or Generative Pre-trained Transformer (GPT) [14] have revolutionized the field of natural language processing (NLP), as they can effectively represent long-term interactions in text using the built-in attention mechanism [15]. These models are pretrained on large text corpora to model the English language. Furthermore, various biomedical domain-specific models such as BioBERT [16], BioMegatron [17], and ClinicalBERT [18] have been created by additional pretraining on PubMed abstracts, PMC full-text documents, and clinical notes. These biological language models have been proposed for various biomedical NLP (bioNLP) tasks, such as named entity recognition (NER), relation extraction (RE), and document classification. In the past, the bioNLP research has mostly focused on extracting protein–protein interactions [19], drug–drug interactions [20], adverse effects detection [21], clinical entity extraction [22, 23], molecular event extraction [24], and more [25, 26].

In this paper, we introduce a deep learning-based text mining workflow that extracts miRNA–disease associations from the literature. The text mining workflow defines three different tasks: (I) detection of miRNA and disease entities (NER), (II) linking of miRNA and disease entities to specific database identifiers [entity linking (EL)], and (III) detection of their associations (RE). We also create a new training dataset containing miRNA–disease associations using distant learning from multiple databases, which is used to train the relation extraction model. After evaluating the promising prediction performance of our workflow, we use it to extract miRNA–disease associations from PubMed between 2020 and 2023. We further discuss the predicted associations in the context of three diseases of interest. For re-usage, we publish the new corpus, the predicted associations, and the source code of our workflow.

Materials and methods

First, we describe all datasets that are required for the three tasks NER, EL, and RE. Next, the training, evaluation, and application of machine learning modeling approach are described in detail

Datasets

Collection of miRNA and disease entity recognition datasets

We used the openly available National Center for Biotechnology Information (NCBI) Disease published by NCBI [27] and BioCreative V Chemical Disease Relation (BC5CDR Disease) [28] corpora that both contain disease mention annotations. These annotations also include entity links to concept identifiers from the Medical Subject Headings (MeSH) database, whereas miRNA mentions are included in miRNA [8] and miRTex [9] corpora. For all datasets, we used the so-called Beginning-Inside-Outside-standoff format [29] for labeling the datasets, where 'O' is assigned to every token that does not represent an entity, 'B' corresponds to the first token of an entity, and 'I' is assigned to following tokens of an entity.

Building a corpus of miRNA–disease relations using distant supervision

Distant or weak supervision aims to create a training dataset (or corpus) by extracting instances from a single or multiple existing knowledge bases, in order to reduce the amount of manual curation effort [30]. To create a suitable training corpus containing miRNA–disease relations, we used two different databases, namely Human microRNA Disease Database 3 [31, 32] and miR2Disease [33]. We first applied rule-based approaches to extract miRNA and disease entities from PubMed abstracts using MiRNADetector [8] and JProMiner, a re-engineered NER algorithm based on the ProMiner software developed by Hanisch *et al.* [34]. In a post-processing step, we filtered out sentences with no miRNA or disease annotations. Furthermore, sentences containing multiple miRNA or disease annotations were manually curated. We further extended our corpus with miRNA–disease relations published by [8]. An overview of all datasets used for training can be seen in Table 1, including some descriptive statistics on the number of mentions and relations for each individual dataset.

Training and application of the miRNA–disease detection pipeline

General workflow

The miRNA–disease association detection workflow consists of two pipelines, which are illustrated in Fig. 1. The training and evaluation pipeline is used to train models that are able to detect miRNA and disease entities and their underlying associations between them. The inferencing pipeline is used to apply the trained models to detect miRNA–disease associations from huge text collections.

In the training and evaluation pipeline (Fig. 1), the first step consists of reading and preprocessing the NER and RE corpora. In the next step, we split the whole corpus in various training, validation, and test sets. For NER, we performed tokenization of sentences and prepared the entities and resulting tokens for IOB-tagging. For RE, we also tokenized the sentences and masked the miRNA and disease entities with predefined tokens for further processing. As each model has its own specific wordpiece tokenization scheme, we utilized the model-specific tokenizer that converts the instances into fixed-sized vectors. In the next stage, these instances are used to fine-tune and optimize the pretrained models for both NER and RE tasks. A model evaluation and selection reveals the best models that can be used for prediction. The inferencing pipeline (Fig. 1) is designed to predict associations from text. First, documents from databases [PubMed and PubMed Central (PMC)] are prepared for inferencing. Subsequently, the models for NER and RE are applied to detect miRNA entities, disease entities, and their associations. In a normalization step, the miRNA and disease entities are normalized to the specific database concepts, namely to Mirbase and MeSH identifiers.

Fine-tuning of BERT-based models

We used the BioBERT [16] and BioMegatron [17] models for our experiments. Both are based on the BERT model published by Google [13], which is trained in a self-supervised manner on huge amounts of text that were obtained from

Table 1. Overview of training and test dataset including number of sentences, mentions, and relations in each dataset

		Training		Test	
NER class	Dataset name	Sentences (%)	Mentions (%)	Sentences (%)	Mentions (%)
Disease	NCBI Disease [27]	6224 (87)	5920 (86)	907 (13)	960 (14)
Disease	BC5CDR Disease [28]	9278 (65)	8427 (66)	4950 (35)	4424 (34)
miRNA	miRNA corpus [8]	1864 (70)	528 (58)	780 (30)	375 (42)
miRNA	miRTex corpus [9]	2063 (57)	1540 (56)	1556 (43)	1217 (44)
		Training relations		Test relations	
RE class	Dataset name	Positive (%)	Negative (%)	Positive (%)	Negative (%)
gene–disease	GAD corpus [11]	2520 (90)	2276 (90)	281 (10)	253 (10)
gene–disease	EU-ADR [53]	235 (90)	83 (90)	27 (10)	10 (10)
miRNA–disease	SCAI-MDC (ours)	1468 (76)	1032 (78)	460 (24)	290 (22)

The numbers in brackets represent the proportions in the training and test sets. The proportions of all external datasets are kept as defined in the original studies. In the case of relation datasets, the number of sentences is identical to the number of relations.

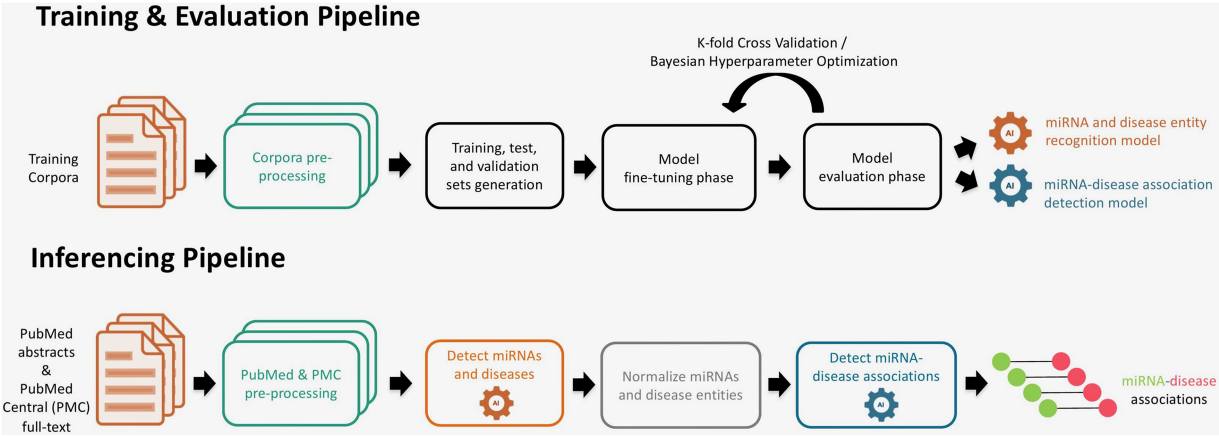


Figure 1. Training, evaluation, and inferencing pipelines for extraction of miRNA and disease entities (NER) and their associations (RE).

OpenBooks, Wikipedia, etc. BioBERT and BioMegatron used the pretrained BERT model and its wordpiece tokenizer. Both were trained further using both PubMed and PMC articles to obtain a domain-specific model for biomedicine. BERT, BioBERT, and BioMegatron are so-called general purpose language models that can be used for various text mining tasks such as NER, RE, document classification, or question answering. To use them for these tasks, they need to be further fine-tuned in a supervised manner on datasets that are specific to the underlying tasks.

For RE, we experimented with two different training modes, namely single-task mode (STM) and multi-task mode (MTM). In the STM, the models were fine-tuned on a single dataset, whereas in MTM, related datasets were used for fine-tuning the various classification heads of the BioBERT model. In MTM, we apply the paradigm of multi-task learning, where a single model is trained to accomplish multiple closely-related tasks simultaneously by using a shared representation [35]. Previous studies have shown that multi-task learning can be beneficial as it improves the generalization by focusing on the commonalities of the tasks and learning relevant features contained in training data of different tasks [35]. The architecture of the final model that is used for fine-tuning BioBERT and BioMegatron is depicted in Fig. 2. We also experimented with different variants for the classification head (such as multiple linear layers, bottleneck architecture). However, the experiments revealed that a simple linear layer works best in all cases. Therefore, our final model contains a single linear layer on top of the pre-trained BioBERT and BioMegatron models.

Linking of miRNA and disease entities

We implemented a rule-based system to link miRNA entities to miRBase identifiers. miRBase [36] is a database that includes published miRNA sequences and annotations, and furthermore, it provides a registry with unique names for miRNAs. To link the recognized disease entities to MeSH identifiers, we used the software NormCo [37].

Evaluation of NER and RE models

For NER, we used precision, recall, and F_1-measure to determine the performance of the models. For RE, which is defined as a binary classification, we used the area under receiver operator characteristic curve (AUROC) and precision–recall curve (AUPR) to evaluate performance. We also provide a confusion matrix report for tasks where it is appropriate, which includes true positive, false positive, false negative, and true negative cases.

In an initial stage, we prepared training and test splits for each dataset. It is important to note that the proportions of the splits are kept as defined in their original studies. Furthermore, we applied five-fold cross-validation to choose the best models. For each iteration, we created a stratified split of the training dataset into training (n – 1 folds) and validation (1 fold) datasets. We then trained on the training dataset and

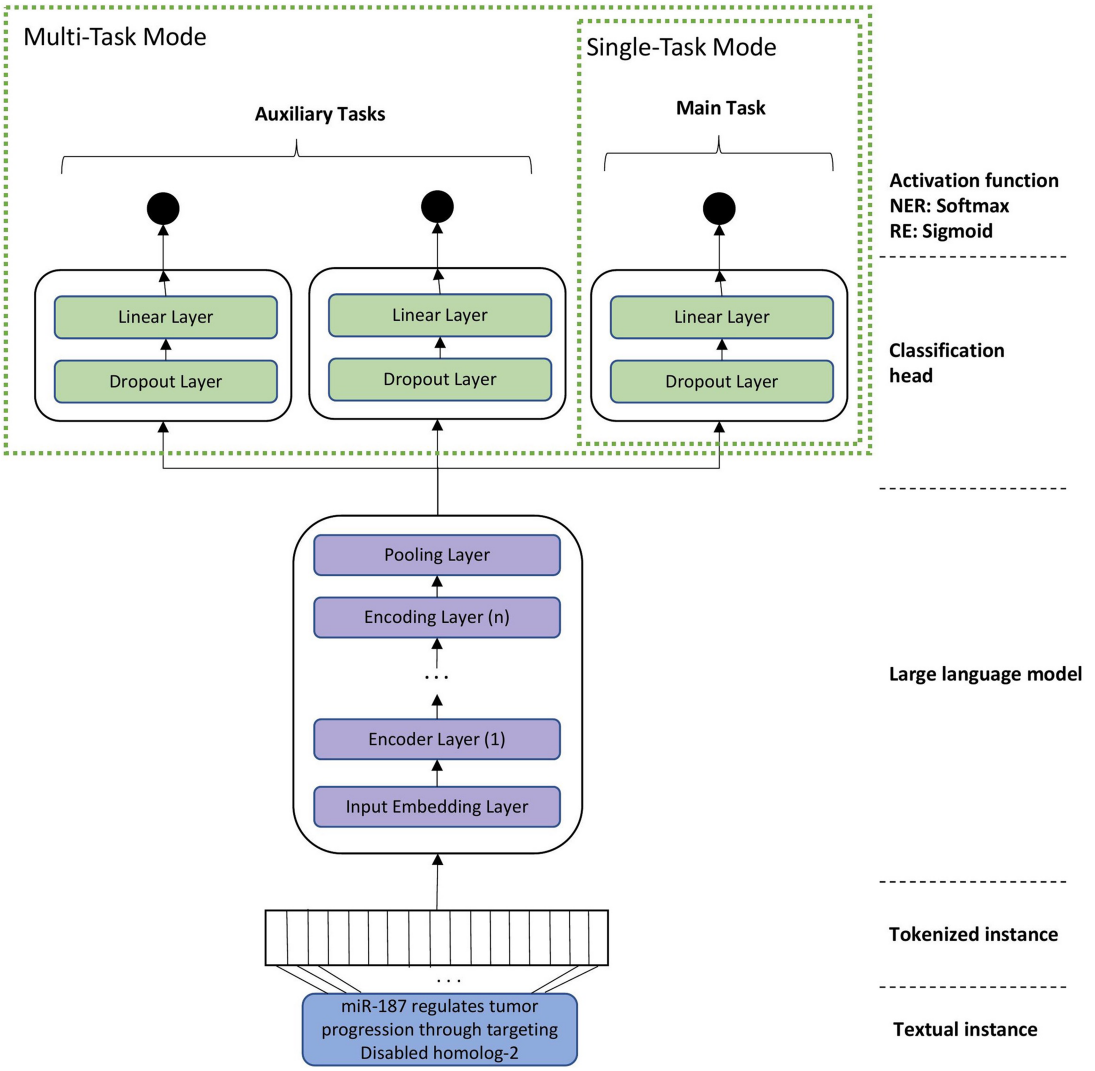


Figure 2. A general architecture of the model for task-specific fine-tuning of domain-specific language models (such as BioBERT and BioMegatron). The STM contains just one head. MTM contains additional heads for each auxiliary task or corpus.

evaluated (and optimized the hyperparameters) on the validation dataset for n iterations. The results of n evaluations are aggregated and the standard deviation is reported. The final evaluation of the best models was performed on the withheld independent test set.

Hyperparameter optimization

We performed a Bayesian hyperparameter optimization [38] using the Optuna [39] framework for all our models with the appropriate training data. We assessed the intermediate and final performances of each experimental trial using the F_1-measure (NER) and AUROC (RE). The results were captured in an SQL database for later analyses, such as identifying the best experimental trials. The captured trial data were also used by the Optuna pruner to identify and halt unpromising trials already at an early stage.

Comparison of predicted associations with DisGeNET

In a consecutive analysis, we compare our predicted associations with data from DisGeNET, where we focus on three different diseases, namely epilepsy, AD, and Parkinson's disease (PD). To compare the associations, it was necessary to retrieve MeSH and UMLS concept identifiers for the disease terms as our workflow normalizes to MeSH and DisGeNet include UMLS identifiers. To retrieve the MeSH and UMLS classes for these diseases, we first gathered all subclasses of the disease from the MONDO ontology [40] and then retrieved their MeSH and UMLS associated identifiers. Both tasks were performed using the OLS4 API (https://www.ebi.ac.uk/ols4). After gathering the associations, we filtered them using the disease identifiers.

Results

Detection of miRNA and disease entities

To detect miRNA and disease mentions, we used the pre-trained BioBERT and BioMegatron models and fine-tuned them on various datasets. To identify the best possible model variant based on the training data, we employed a 5-fold cross-validation during training. Based on the performance assessed through cross-validation, we used the

Table 2. Evaluation results of NER task models trained and tested on various datasets.

		BioBERT			BioMegatron		
Entity class	Dataset	Prec.	Recall	F_1	Prec.	Recall	F_1
Disease	NCBI Disease	84.62	90.09	87.27	**88.22**	**91.25**	**89.71**
	BC5CDR	82.07	85.39	83.70	**85.49**	**87.75**	**86.60**
	NCBI Disease + BC5CDR	–	–	–	**86.26**	**87.83**	**87.04**
miRNA	miRNA	91.32	**98.13**	94.60	**91.75**	9787	**94.71**
	miRTex	93.93	95.79	94.85	**96.59**	**97.62**	**97.10**
	miRNA + miRTex	–	–	–	**94.51**	**96.23**	**95.36**

The confusion matrix of the BioMegatron model is included in Supplementary Table S7. – indicates data are not available. Bold entries represent the best results.

optimized hyperparameters to train the final model on the whole training dataset. The generalization performance of the final models was assessed on the held-out test set. Table 2 presents the classification scores for each dataset in the specific test set.

For the NCBI dataset, we achieved the highest performance with an F1-score of 89.71%, precision of 88.22%, and recall of 91.25%. For the BC5CDR dataset, the best F1-score was 86.60% with precision of 85.49% and recall of 87.75%. We also trained a model with the combination of both datasets, where a micro F1-score of 87.04% was reached on the combined test set. Overall, BioMegatron performed better than BioBERT, which is probably due to the large parameter size of the BioMegatron model.

In the case of miRNA entity detection, the best F1-measure for the miRNA dataset was 94.70%, precision was 91.75%, and recall was 97.86%, and the best performance for the miRTex dataset was achieved with an F1-score of 97.10% and a precision of 96.59%. The training on the combined dataset reached a micro F1-score of 95.36%. Similar to the disease category, the BioMegtron model performed significantly better than BioBERT, while the BioBERT model delivered the best recall of 98.13% on the miRNA dataset. The confusion matrices of the best NER models are provided in Supplementary Table S7. The optimized hyperparameters of the best NER models are included in Supplementary Table S1–S4 and S6.

We also experimented with MTM; however, the results were not significantly better in comparison to the STM. Although the NER datasets and tasks share with each other certain similarities, the significant differences in the annotation guidelines and their varying levels of complexity of the mentions likely reduced the effectiveness of the multitask approach. The shared representations in the model might have led to negative transfer, showing a drop in the model performance. Similar observations have also been made by Crichton *et al.* [41]. Hence, for further analysis we focused only on STM.

It is important to note that the BC5CDR corpus is a subcorpus of the CTD-Pfizer corpus [42]. The creators of the corpus aimed to investigate the potential involvement of pharmaceutical drugs in cardiovascular, neurological, renal and hepatic toxicity. Therefore, the BC5CDR corpus is focused on drugs and their role in toxicity [42]. In contrast, the NCBI disease corpus is intended to represent the entire PubMed. In an analysis of both corpora performed by Kühnel and Fluck [43], they revealed that the BC5CDR corpus contains more complex contexts, including abbreviations from diseases but also mentions several gene names, such as *BRCA1*, resembling the structure of an abbreviation. This could explain why the model performances for the NCBI Disease dataset are slightly better.

Table 3. Evaluation results of RE task on test dataset for STM and MTM training modes based on BioMegatron model.

Datasets	Mode	AUROC (in %)	AUPR (in %)
miRNA–disease	STM	97.58	97.55
	MTM	**98.02**	**98.66**

Bold entries represent the best results.

Detection of miRNA–disease associations

We trained an association detection model using the BioMegatron model on our own training dataset (80% of the whole dataset). As BioMegatron, in comparison to BioBERT, delivered the best results in almost all cases, we only focused on experimenting with the BioMegatron model further. The model selection was based on five-fold cross-validation. After choosing the best hyperparameters, we evaluated the final model performance using measures, such as AUROC and AUPR on an independent test set (20% of the whole dataset). Table 3 illustrates the evaluation performances. We reached a high rate of 9758% AUROC and 9755% AUPR with the STM. Even higher scores are reached with the MTMs, amounting to 98.02% and 98.66% for AUROC and AUPR, respectively. The receiver operator characteristic and precision–recall curves of the best model are depicted in Supplementary Figure S1. The optimized hyperparameters of the best RE model are included in Supplementary Tables S5 and S6.

Prediction of miRNA–disease associations from PubMed

We applied our miRNA–disease association extraction workflow on around 6.1 million PubMed abstracts and 1.98 million PMC full-text documents published between 2020 and 2023. Overall, the workflow predicted 727 009 (unique: 75 887) normalized positive associations found in 69 816 PMC and PubMed documents. These associations include 2730 disease and 2427 miRNA concepts. Overall, 374029 positive associations (unique: 52624; found in 59187 PMC documents) of them have a high confidence score of 90% (retrieved through a sigmoid function). Figure 3 provides an overview of the total predicted miRNA–disease associations for PubMed abstracts, PMC full-text documents, and both corpora combined.

In a subsequent analysis, we filtered for associations of three different diseases, namely epilepsy, AD, and PD

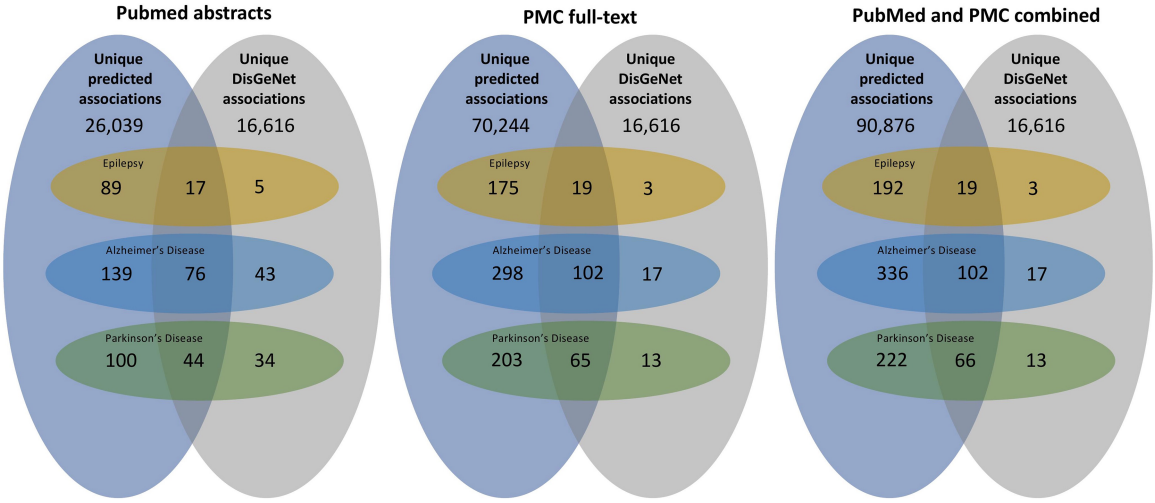


Figure 3. An overview of the total predicted unique associations between miRNA and diseases in comparison to the DisGeNET database. The three subfigures represent the results extracted between 2020 and 2023 from PubMed abstracts, PMC full-text documents, and both combined. Furthermore, it provides an overview over the miRNA–disease associations of three diseases (epilepsy, AD, and PD).

(see Fig. 3). For epilepsy, AD, and PD, the workflow detected 2226 (unique: 211), 6306 (unique: 438), and 3159 (unique: 287) miRNA–disease associations, respectively. In a first step, we compared the extracted miRNA associations with those in the existing database DisGeNET, which contains curated miRNA–disease associations extracted from different resources before 2020. In all cases, we could significantly increase the number of miRNA–disease associations and found a high number of new relations not contained in DisGeNET (see Fig. 3). Since we focus on new findings, only a small number of the relations overlap with the relations contained in DisGeNet and others are only available in DisGeNet.

We also performed an analysis of the missed DisGeNet associations for the year 2020 of the three diseases. In total, DisGeNet contains four unique miRNA–disease associations for epilepsy, eight for AD, and eight for PD from publications published in the year 2020. Only two associations (one for AD and one for PD) were missed by our workflow. In these cases, the workflow predicted wrong association labels (no association). All the other missed associations were from publications published before 2020, which we have not included in our workflow. To expand this analysis, we randomly analyzed additional unique associations that were missed by our pipeline. In some cases, the association was detected, however, with a lower score (<0.9). In other cases, the disease and miRNA normalizer were not able to properly normalize the mentions. We provide some examples of these cases in the Supplementary Section 'Examples of Workflow Issues'.

Evaluation of newly detected miRNA–disease associations

For all three diseases, AD, PD, and epilepsy, we randomly choose associations from the predicted results of the PubMed corpora that had a high score (>0.9). Examples of extracted associations with their corresponding sentences are shown in Table 4. For AD, our workflow detected three miRNA–disease associations from a study by Kumar *et al.* [5]. In this study, by analyzing postmortem brains of AD and control samples using a miRNAs microarray platform, the authors have addressed the question of whether synaptosomal miRNAs affect AD synapse activity. They found that three specific miRNAs are potentially associated with AD Braak stages. In the case of PD, our workflow detected two miRNA–disease associations from the study published by Chen *et al.* [44]. The authors investigated blood circulating miRNAs that are proposed to be promising biomarkers for neurodegenerative diseases such as PD. They analyzed the plasma of PD patients, multiple system atrophy patients, and healthy controls. Our workflow detected two associations from the study [45], where the authors studied the role of let-7b miRNAs in temporal lobe epilepsy (TLE). They found a novel noncoding RNA-mediated mechanism involving the miRNA let-7b and H19 [a long noncoding RNA (lncRNA)] in seizure-induced glial cell activation.

For a systematic analysis of the newly found associations, we analyzed the precision and recall for the PD–miRNA associations. To check the overall precision of the newly predicted associations, 30 sentences were examined. This analysis showed that only two extractions were incorrect. In the sentence 'PD was associated with postoperative expression of GFAP; ePOCD was associated with postoperative expression of microRNA-21-5p and GFAP as well as intraoperative expression of NSP' (PMID:34 300 256) [46], the abbreviation 'PD' means postoperative delirium and hence the association with the disease PD is incorrect. The second error occurred in an extraction from the text fragment '[…] Mitochondrial complex I deficiency and functional abnormalities are implicated in the development of PD. MicroRNA-29a […]' (PMID: 36 174 668) [47] that consists of more than one sentence and could not be verified as the correct source for the extracted association, although the relation was mentioned later in the abstract. In summary, this analysis shows a precision of over 93%.

In order to analyze the recall, we utilized a systematic review of PD–miRNA associations published by [48]. All referenced associations from publications in 2020 to 2023 were compared with our automatically extracted set. Out of 23 associations, a total of 15 associations were extracted from the same abstract also referenced by the review, but 8 could

Table 4. Examples of predicted miRNA–disease associations for AD, PD, and epilepsy with their corresponding sentences.

Disease	miRNA	Sentence	PMID
AD	hsa-miR-501-3p (MIRBASE:MIMAT0004774)	The **miR-501-3p**, miR-502-3p, and miR-877-5p were identified as potential synaptosomal miRNAs upregulated with disease progression based on **AD** Braak stages.	36454178
AD	hsa-miR-502-3p (MIRBASE:MIMAT0004775)	The miR-501-3p, **miR-502-3p**, and miR-877-5p were identified as potential synaptosomal miRNAs upregulated with disease progression based on **AD** Braak stages.	36454178
AD	hsa-miR-877-3p (MIRBASE:MIMAT0004949)	The miR-501-3p, miR-502-3p, and **miR-877-5p** were identified as potential synaptosomal miRNAs upregulated with disease progression based on **AD** Braak stages.	36454178
PD	hsa-miR-133b (MIRBASE:MIMAT0000770)	Elevated **miR-133b** and miR-221-3p distinguished **PD** from controls with 84.8% sensitivity and 88.9% specificity.	34315950
PD	hsa-miR-221-3p (MIRBASE:MIMAT0000278)	Elevated miR-133b and **miR-221-3p** distinguished **PD** from controls with 84.8% sensitivity and 88.9% specificity.	34315950
Epilepsy	hsa-let-7b-5p (MIRBASE:MIMAT0000063)	Overexpression of **let-7b** inhibited hippocampal glial cell activation, inflammatory response and **epileptic seizures** by targeting Stat3.	32648622
Epilepsy	hsa-let-7b-5p (MIRBASE:MIMAT0000063)	LncRNA H19 could competitively bind to **let-7b** to promote hippocampal glial cell activation and epileptic seizures by targeting Stat3 in a rat model of **TLE**.	32648622

The normalized miRNA names mentioned in the second column corresponds to the bold miRNA names in the Sentence column.

not be found in the abstracts. These eight associations were reviewed further. Two associations (from Wu 2020 [49]) could not appear in our result set as the corresponding publication journal 'Acta Medica Mediterranea' is not part of the Medline. Furthermore, in the publication by Ravanidis et al. [50], the two associations were not mentioned in the abstract, but only in the full-text. Finally, in the publication by Cressati et al. [51], miR-153 and miR-223 were mentioned in the abstract and these associations were correctly recognized, but in the review, they are listed as miR-153-3p and miR-223-5p. Only two associations were not found by the automated extraction system although they were mentioned in the abstract. These were missed because the corresponding miRNAs were not recognized. In summary, this analysis shows that 19 associations were described in the Medline articles, of which our system recognized 17 associations. This corresponds to a recall of 89%.

This evaluation shows that even after the sequential execution of automated NER, entity linkage, and association recognition, which have their own error rate that adds up in the overall result, the performance of the automated extraction system is remarkable and therefore very well suited to support systematic reviews such as that published for PD by Guévremont *et al*. [48].

Discussion

In this work, we presented a workflow for automatically extracting miRNA–disease associations from vast unstructured literature. The workflow is based on a large language model fine-tuned on a new corpus generated using a distant supervision technique. Due to the pretraining of large language model (for e.g. BioMegatron) on a huge corpora and the integrated self-attention mechanism, the model can exploit semantic and syntactic aspects of sentences and incorporates local contextual features of the included entities to extract relations with high accuracy. We used the workflow to extract miRNA–disease associations from Medline abstracts and analyzed the extracted set for AD, PD, and epilepsy. Compared to the existing curated database DisGeNet, where miRNA–disease associations were provided until 2020, we extracted a high number of new associations from Medline abstracts for the years 2020–23. An independent evaluation of the newly extracted PD–miRNA associations showed that we achieved high precision and high recall with this extraction workflow.

A current limitation of the corpus, and thus of the extraction workflow, is that the associations are encoded and recognized at sentence level. As authors may describe miRNA–disease relations beyond sentences in their publications, the workflow may miss these relations. Nevertheless, the evaluation showed that at least for PD, the PD–miRNA relations are usually expressed in the same sentence in abstracts. We missed associations due to false negative disease and miRNA recognition or because the relations were only expressed in the full-text tables. Strategies such as active learning might help to significantly reduce the curation effort to extend the corpus required for training a model that can perform extraction from tables included in full-text documents.

Although the large language models that are specifically designed for the biomedical domain produced great results in our work, incorporating the extensive prior knowledge on miRNAs and diseases directly in large language models might help to improve the results even further. Studies have shown that the process of knowledge fusion is able to overcome the limitations of individual sources by focusing on diverse knowledge. Information such miRNA sequences, disease embeddings obtained from ontologies (such as Disease ontology, MONDO) can be merged with the embeddings from large language models. Also, embeddings obtained through training of graph neural networks on sources such as DisGeNet can be further employed to improve

the models. In addition, it might be interesting to combine the literature-based models with new prediction models learning feature embeddings for miRNAs and diseases through graph machine learning [52].

In summary, by automatically generating a training corpus using distance learning methods and training a model based on a state-of-the-art large language model, we have demonstrated the promising performance of our trained workflow. Our evaluation results based on PD-miRNA associations strongly suggest that our workflow can provide useful support for extracting miRNA–disease relations.

Conclusion

In this work, we proposed a well-performing large language model approach for the identification of miRNA–disease relations from biomedical literature. The approach consists of modules that can perform the detection of miRNA and disease mentions, as well as the identification of their relationship. In order to extend the miRNA–disease training corpora, we applied the distant supervision technique using multiple publicly available databases. In our experiments with multiple state-of-the-art large language models, BioMegatron performed the best for the extraction of miRNA–disease associations. A high number of new associations could be identified with a high level of precision of recall and precision, when applying the whole machinery to infer associations from biomedical literature between 2020 and 2023.

The creation and use of dedicated databases that can contain many types of relations is considered best practice in biomedical research and up-to-date information is in high demand. However, to keep these databases up to date with the current scientific advancements is a major challenge. The solution is often to establish collaborations with researchers and institutions to provide regular updates. However, this requires a huge amount of human effort. This creates a demand for automated data mining techniques that should always be employed to extract relevant information from scientific literature and update the databases accordingly. With the three different case studies on neurodegenerative diseases such as AD, for which we identified and discussed novel relations that are yet missing in databases such as DisGeNet, we demonstrated the applicability and feasibility of our workflow for retrieving novel, hidden relations from literature.

Automated techniques for information extraction need to be regularly revised to keep up with the pace of development in NLP. Recent large language models such as ChatGPT, BARD, some of which are unfortunately not yet available for scientific experimentation, open up new avenues for solving challenges. Future studies are required to find out exactly how these models can be utilized to not only extract a single type of relation but also to solve many complex bioNLP challenges at once.

Acknowledgements

We thank André Gemünd for their support regarding the computational infrastructure of SCAI. We thank Jürgen Klein for their support in preprocessing the PubMed Central documents.

Author contributions

S.M. and J.F. were involved in conceptualization; S.M. were involved in methodology; S.M. and L.K. were involved in data curation, formal analysis, visualization, investigation, and validation; S.M. and J.F. were involved in supervision; S.M. and L.K. were involved in writing the original draft; S.M., L.K., H.F., J.F., and M.H. were involved in writing, review and editing.

Supplementary data

Supplementary data is available at *Database* online.

Conflict of interest

None declared.

Funding

This work was funded through the project Integrative Data Semantics for Neurodegenerative research (IDSN), which was supported by the German Federal Ministry of Education and Research (BMBF) as part of the program 'i:DSem–Integrative Data Semantics in the Systems Medicine', project number 031L0029 [A-C].

Data availability

We provide our code at https://github.com/SCAI-BIO/mirna-disease-association-detection. Our database is located at https://zenodo.org/records/10523046

References

1. Rupaimoole R, Slack FJ. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat Rev Drug Discov* 2017;**16**:203–22. https://doi.org/10.1038/nrd.2016.246
2. Takamizawa J, Konishi H, Yanagisawa K *et al*. Reduced expression of the let-7 MicroRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res* 2004;**64**:3753–56. https://doi.org/10.1158/0008-5472.CAN-04-0637
3. Lin CW, Chang YL, Chang YC *et al*. MicroRNA-135b promotes lung cancer metastasis by regulating multiple targets in the Hippo pathway and LZTS1. *Nat Commun* 2013;**4**:1877. https://doi.org/10.1038/ncomms2876
4. Rupani H, Sanchez-Elsner T, Howarth P. MicroRNAs and respiratory diseases. *Eur Respir J* 2013;**41**:695–705. https://doi.org/10.1183/09031936.00212011
5. Kumar S, Orlov E, Gowda P *et al*. Synaptosome microRNAs regulate synapse functions in Alzheimer's disease. *NPJ Genom Med* 2022;**7**:47. https://doi.org/10.1038/s41525-022-00319-8
6. Takousis P, Sadlon A, Schulz J *et al*. Differential expression of microRNAs in Alzheimer's disease brain, blood, and cerebrospinal fluid. *Alzheimers Dement* 2019;**15**:1468–77. https://doi.org/10.1016/j.jalz.2019.06.4952
7. Hébert SS, Delay C. MicroRNAs and Alzheimer's disease mouse models: current insights and future research avenues. *Int J Alzheimer's Dis* 2011;**2011**:894938. https://doi.org/10.4061/2011/894938
8. Bagewadi S, Bobić T, Hofmann-Apitius M *et al*. Detecting miRNA mentions and relations in biomedical literature. *F1000Res* 2015;**3**:205. https://doi.org/10.12688/f1000research.4591.3

9. Li G, Ross KE, Arighi CN *et al*. miRTex: a text mining system for miRNA-gene relation extraction. *PLoS Comput Biol* 2015;**11**:1–24. https://doi.org/10.1371/journal.pcbi.1004391
10. Gupta S, Ross KE, Tudor CO *et al*. miRiaD: a text mining tool for detecting associations of microRNAs with diseases. *J Biomed Semant* 2016;**7**:9. https://doi.org/10.1186/s13326-015-0044-y
11. Bravo À, Piñero J, Queralt-Rosinach N *et al*. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinf* 2015;**16**:55.. https://doi.org/10.1186/s12859-015-0472-9
12. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J *et al*. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* 2020;**48**:D845–55. https://doi.org/10.1093/nar/gkz1021
13. Devlin J, Chang MW, Lee K *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, pp. 4171–86. Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/N19-1423
14. Brown TB, Mann B, Ryder N *et al*. Language models are few-shot learners. arXiv preprint arXiv:200514165. 2020.
15. Vaswani A, Shazeer N, Parmar N *et al*. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–10. Red Hook, NY, USA: Curran Associates Inc., 2017. NIPS'17.
16. Lee J, Yoon W, Kim S *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–40. https://doi.org/10.1093/bioinformatics/btz682
17. Shin HC, Zhang Y, Bakhturina E *et al*. BioMegatron: larger biomedical domain language model. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 4700–06. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.emnlp-main.379
18. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv:190405342 [cs]. 2019.
19. Jiang Z, Shuang L, and Huang D. A general protein-protein interaction extraction architecture based on word representation and feature selection. *Int J Data Min Bioinform* 2016;**14**:276–91. https://doi.org/10.1504/IJDMB.2016.074878
20. Zhu Y, Li L, Lu H *et al*. Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions. *J Biomed Informat* 2020;**106**:103451. https://doi.org/10.1016/j.jbi.2020.103451
21. Gurulingappa H, Klinger R, Hofmann-apitius M *et al*. An empirical evaluation of resources for the identification of disease and adverse effects in biomedical literature. In: *The seventh international conference on Language Resources and Evaluation (LREC)*, 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining, *Valletta, Malta*, May 2010, pp. 15–22, 2010. http://www.lrec-conf.org/proceedings/lrec2010/workshops/W8.pdf.
22. Li J, Zhou Y, Jiang X *et al*. Are synthetic clinical notes useful for real natural language processing tasks: a case study on clinical entity recognition. *J Am Med Inf Assoc* 2021;**28**:2193–201. https://doi.org/10.1093/jamia/ocab112
23. Lentzen M, Madan S, Lage-Rupprecht V *et al*. Critical assessment of transformer-based AI models for German clinical notes. *JAMIA Open* 2022;**5**:ooac087. https://doi.org/10.1093/jamiaopen/ooac087
24. Pattankar VV, and Priyanga P. Review on event extraction for BioNLP with a survey. In: *2023 International Conference for Advancement in Technology (ICONAT)* Goa, India, 24-26 January 2023, pp. 1–5. Goa, India: IEEE, 2023. https://doi.org/10.1109/ICONAT57137.2023.10080428
25. Shang Y, Li Y, Lin H *et al*. Enhancing biomedical text summarization using semantic relation extraction. *PLoS One* 2011;**6**:e23862. https://doi.org/10.1371/journal.pone.0023862
26. Bressem KK, Adams LC, Gaudin RA *et al*. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. *Bioinformatics* 2020;**36**:5255–61. https://doi.org/10.1093/bioinformatics/btaa668
27. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. *J Biomed Informat* 2014;**47**:1–10. https://doi.org/10.1016/j.jbi.2013.12.006
28. Li J, Sun Y, Johnson RJ *et al*. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)* 2016;**2016**:baw068. https://doi.org/10.1093/database/baw068
29. Ramshaw LA, and Marcus MP. Text chunking using transformation-based learning. In: Armstrong S, Church K, Isabelle P, Manzi S, Tzoukermann E, Yarowsky D, (eds.), *Natural Language Processing Using Very Large Corpora*, 1st edn. Dordrecht, Netherlands: Springer, 1999, 157–76. https://doi.org/10.1007/978-94-017-2390-9_10
30. Smirnova A, and Cudré-Mauroux P. Relation extraction using distant supervision: a survey. *ACM Comput Surv* 2018;**51**:106:1–106:35. https://doi.org/10.1145/3241741
31. Li Y, Qiu C, Tu J *et al*. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res* 2014;**42**:1–5. https://doi.org/10.1093/nar/gkt1023
32. Huang Z, Shi J, Gao Y *et al*. HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res* 2019;**47**:D1013–7. https://doi.org/10.1093/nar/gky1010
33. Jiang Q, Wang Y, Hao Y *et al*. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* 2009;**37**:D98–104. https://doi.org/10.1093/nar/gkn714
34. Hanisch D, Fundel K, Mevissen H *et al*. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinf* 2005;**6**:S14. https://doi.org/10.1186/1471-2105-6-S1-S14
35. Caruana R. Multitask learning. In: Thrun S, Pratt L (eds.), *Learning to Learn*. Boston, MA: Springer US, 1998, 95–133.
36. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 2019;**47**:D155–62. https://doi.org/10.1093/nar/gky1141
37. Wright D, Katsis Y, Mehta R *et al*. *NormCo: Deep Disease Normalization for Biomedical Knowledge Base Construction*. https://openreview.net/forum?id=BJerQWcp6Q (1 June 2024, date last accessed).
38. Bergstra J, Yamins D, and Cox D Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In: *International Conference on Machine Learning. PMLR*, Atlanta, Georgia, USA: Machine Learning Research Press, pp. 115–23. 2013.
39. Akiba T, Sano S, Yanase T *et al*. Optuna: a next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, August 4 - 8, 2019, pp. 2623–31, New York, NY, United States: Association for Computing Machinery, 2019.
40. Vasilevsky NA, Matentzoglu NA, Toro S *et al*. Mondo: unifying diseases for the world, by the world. *medRxiv* 2022;2022.04.13.22273750.
41. Crichton G, Pyysalo S, Chiu B *et al*. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinf* 2017;**18**:368. https://doi.org/10.1186/s12859-017-1776-8

42. Davis AP, Wiegers TC, Roberts PM *et al.* A CTD-Pfizer collaboration: manual curation of 88 000 scientific articles text mined for drug-disease and drug-phenotype interactions. *Database* 2013;**2013**:bat080. https://doi.org/10.1093/database/bat080
43. Kühnel L, Fluck J. We are not ready yet: limitations of state-of-the-art disease named entity recognizers. *J Biomed Semant* 2022;**13**:26. https://doi.org/10.1186/s13326-022-00280-6
44. Chen Q, Deng N, Lu K *et al.* Elevated plasma miR-133b and miR-221-3p as biomarkers for early Parkinson's disease. *Sci Rep* 2021;**11**:15268. https://doi.org/10.1038/s41598-021-94734-z
45. Han CL, Liu YP, Guo CJ *et al.* The lncRNA H19 binding to let-7b promotes hippocampal glial cell activation and epileptic seizures by targeting Stat3 in a rat model of temporal lobe epilepsy. *Cell Prolif* 2020;**53**:e12856. https://doi.org/10.1111/cpr.12856
46. Szwed K, Szwed M, Kozakiewicz M *et al.* Circulating microRNAs and novel proteins as potential biomarkers of neurological complications after heart bypass surgery. *J Clin Med* 2021;**10**:3091. https://doi.org/10.3390/jcm10143091
47. Yang YL, Lin TK, Huang YH. MiR-29a inhibits MPP + -Induced cell death and inflammation in Parkinson's disease model in vitro by potential targeting of MAVS. *Eur J Pharmacol* 2022;**934**:175302. https://doi.org/10.1016/j.ejphar.2022.175302
48. Guévremont D, Roy J, Cutfield NJ *et al.* MicroRNAs in Parkinson's disease: a systematic review and diagnostic accuracy meta-analysis. *Sci Rep* 2023;**13**:16272. https://doi.org/10.1038/s41598-023-43096-9
49. Wu L, Zhao W, Kong F *et al.* Serum miR-9a and miR-133b, diagnostic markers for Parkinson's sisease, are up-regulated after Levodopa treatment. *Acta Med Mediterr* 2020;**36**:1857–1863. https://doi.org/10.19193/0393-6384_2020_3_291.
50. Ravanidis S *et al* Circulating Brain-enriched MicroRNAs for detection and discrimination of idiopathic and genetic Parkinson's disease. *Mov Disord* 2020;**35**:457–467. https://doi.org/10.1002/mds.27928
51. Cressatti M, Juwara L, Galindez JM *et al.* Salivary microR-153 and microR-223 Levels as Potential Diagnostic Biomarkers of Idiopathic Parkinson's Disease. *Mov Disord* 2020;**35**:468–477. https://doi.org/10.1002/mds.27935
52. Peng W, Che Z, Dai W *et al.* Predicting miRNA-disease associations from miRNA-gene-disease heterogeneous network with multi-relational graph convolutional network model. *IEEE/ACM Trans Comput Biol Bioinform* 2023;**20**:3363–75. https://doi.org/10.1109/TCBB.2022.3187739
53. van Mulligen EM, Fourrier-Reglat A, Gurwitz D *et al.* The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *J Biomed Informat* 2012;**45**:879–84. https://doi.org/10.1016/j.jbi.2012.04.004

Database, 2024, 00, baae066, DOI: https://doi.org/10.1093/database/baae066, Original article