Fachhochschule Graubünden
University of Applied Sciences

Churer Schriften
zur Informationswissenschaft

Herausgegeben von
Wolfgang Semar, Bernard Bekavac, Ivo Macek, Armando Schär

Arbeitsbereich
Master of Science in Business Administration,
Major Information and Data Management

**Schrift 154**

# «Drug Repurposing»

Wie können unstrukturierte Textdaten für die Ermittlung neuer «Drug Repurposing» Kandidaten nutzbar gemacht werden und wie können sie Datenbanken ergänzen?

**Curdin Marxer**

Chur 2022

# Churer Schriften zur Informationswissenschaft

Herausgegeben von Wolfgang Semar,
Bernard Bekavac, Ivo Macek, Armando Schär

Schrift 154


# «Drug Repurposing»

Wie können unstrukturierte Textdaten für die Ermittlung neuer «Drug Repurposing» Kandidaten nutzbar gemacht werden und wie können sie Datenbanken ergänzen?


## Curdin Marxer



Diese Publikation entstand im Rahmen einer Thesis zum Master of Science FHGR in Business Administration, Major Information and Data Management.

Referent: Prof. Dr. Heiko Rölke

Korreferent: Prof. Dr. Franjo Pehar

**Verlag:** Fachhochschule Graubünden

**ISSN:** 1660-945X

**Ort, Datum:** Chur, November 2022

## Abstract

Das Konzept der Identifikation und Entwicklung neuer Verwendungszwecke bereits bekannter Arzneimittel und Wirkstoffe wird als "Drug Repurposing" bezeichnet. Computergestützte Methoden können durch Datenanalyse und Datenverarbeitung das "Drug Repurposing" unterstützen. Diese Daten sind neben einer standardisierten Form aus Datenbanken, auch als unstrukturierte Textdaten verfügbar.

Im Rahmen der vorliegenden Masterthesis wurde der forschungsleitenden Frage *"Wie können unstrukturierte Textdaten für die Ermittlung neuer "Drug Repurposing" Kandidaten nutzbar gemacht werden und wie können sie Datenbanken ergänzen?"* nachgegangen.

Um das Potenzial von biomedizinischen Textdaten für das "Drug Repurposing" zu untersuchen, wurden zwei unterschiedliche Methoden mit mehreren Variationen entwickelt, welche durch den Einsatz der biomedizinischen Named Entity Recognition (NER) neue "Drug Repurposing" Kandidaten aus Textdaten identifizieren konnten.

Für die Evaluation beider Methoden wurden im Rahmen eines Fallbeispiel neue alternativ anwendbare Wirkstoffe und Medikamente für die Behandlung des Glioblastoms aus 10'000 klinischen und biomedizinischen Dokumenten bestimmt. Dabei zeigten die Ergebnisse beider Methoden das Potenzial von unstrukturierten Textdaten zur Verschaffung eines kompakten und wahrheitsgetreuen Überblicks potenziell relevanter Wirkstoffe und Arzneimittel.

## Inhaltsverzeichnis

1 Einleitung ..... 1
2 Medikamentenentwicklung und die Rolle von "Drug Repurposing" ..... 5
2.1 Klinische Studien und Hürden der Medikamentenzulassung ..... 5
2.2 Das Potenzial von "Drug Repurposing" ..... 9
2.3 "Drug Repurposing": Definitionen und Prozesse ..... 10
3 "Computational Methods" und die Rolle von Datenbanken ..... 13
3.1 "Computational Methods" ..... 13
3.2 Rolle von medizinischen Datenbanken ..... 13
3.3 "Computational methods" und "Repurposing" Vorgehensweisen ..... 15
3.3.1 "target-based" ..... 16
3.3.2 "side-effect-based" ..... 17
3.3.3 "expression-based" ..... 17
3.3.4 "similarity-based" ..... 19
3.3.5 Weitere Vorgehensweisen und Methoden ..... 21
3.4 "Machine Learning" in der Medikamentenforschung und im "Drug Repurposing" 23
3.4.1 "Supervised" Methoden im "Drug Repurposing" ..... 25
3.4.2 "Unsupervised" Methoden im "Drug Repurposing" ..... 29
3.4.3 Moderne ML-Methoden im "Drug Repurposing" und Limitationen ..... 30
3.5 Das Potenzial unstrukturierter Textdaten und verfügbare Analysemethoden ..... 31
3.5.1 NLP in der Biomedizin ..... 31
3.5.2 "ABC-Modell": Ermittlung von biomedizinischen Assoziationen in Textdaten 36
4 Forschungsleitende Fragestellung & formulierte Unterfragen ..... 39
4.1 Grundeigenschaften der Workflows & generierte Unterfragen ..... 40
4.2 Zusätzliche Fragestellungen ..... 40
5 Forschungsmethodik und -design ..... 43
5.1 Grundlagen und Allgemeines ..... 43
5.1.1 Verwendete vorinstallierte Systempakete und Bibliotheken ..... 44
5.2 Übersicht und Auswahl biomedizinischer NLP-Werkzeuge und NER- Systeme in Python ..... 44
5.2.1 NER-Systeme für die Identifikation von UMLS und MeSH-Konzepten ..... 45
5.2.2 "Rule-based" NER-Systeme mit vortrainierten Modellen ..... 47

5.2.3 NER-Systeme mit "Transfer-Learning" Kapazitäten ..... 52
5.2.4 Vergleich der NER-Systeme und Auswahl für die praktische Umsetzung ..... 53
5.3 Auswahl der Textdaten und Datenbanken ..... 56
5.3.1 Extraktion, Bereitstellung und Normalisierung der Textdaten ..... 56
5.4 Methode 1: Kookkurrenz Analyse basierend auf dem "GBA-Prinzip" ..... 57
5.4.1 Kernelemente der Methode ..... 59
5.5 Methode 2: Assoziationsketten basierend auf Swanson's "ABC-Modell" ..... 61
5.5.1 Kernelemente der Methode ..... 62
5.6 Bestimmung des Fallbeispiels ..... 65
5.6.1 Auswahl der Fallbeispieldaten ..... 66
5.7 Evaluation der Ergebnisse ..... 68
5.7.1 Formulierte Bewertungskriterien und Bewertungsprozess ..... 68
6 Ergebnisse der Forschungsarbeit & Diskussion ..... 71
6.1 Form der Ergebnisse ..... 71
6.2 Methode 1: Analyse der Ergebnisse ..... 72
6.2.1 Methode 1: Vergleich der Variationen ..... 74
6.2.2 Methode 1: Zusätzliche Beobachtungen und Erkenntnisse ..... 76
6.3 Methode 2: Analyse der Ergebnisse ..... 77
6.3.1 Methode 2: Vergleich der Assoziationsketten ..... 78
6.4 Vergleich der Ergebnisse der Methode 1 und Methode 2 ..... 81
6.5 Beantwortung der forschungsleitenden Frage ..... 83
7 Fazit & Reflexion ..... 85
7.1 Festgestellte Grenzen der Forschungsarbeit & Fazit ..... 85
7.2 Weiterer Forschungsbedarf & Empfehlungen ..... 87
7.3 Persönliche Reflexion & Schlusswort ..... 89
8 Quellen & Literaturverzeichnis ..... 91

## Abbildungsverzeichnis

Abbildung 1: Abgebildeter Prozess der Medikamentenentwicklung mit der Anzahl Wirkstoffkandidaten und den zugehörigen Erfolgswahrscheinlichkeiten, Zykluszeiten und prozentualen Gesamtkostenanteilen jedes Entwicklungsschrittes (Sun et al., 2022, S. 1) ..... 8
Abbildung 2: Verfügbare biomedizinische Datenbanken und kategorische Einteilung nach Tanoli et al. (2021, S. 1658) ..... 15
Abbildung 3: Übersicht verfügbarer "computational“ Vorgehensweisen und Strategien zur Ermittlung von Krankheit-Medikament Zusammenhängen (Dudley et al., 2011, S. 305) ..... 16
Abbildung 4: "Guilt by Association" Schema für die Entdeckung neuer Anwendungen von Arzneimitteln (Chiang & Butte, 2009, S. 509) ..... 20
Abbildung 5: Beispiel eines simplen Entscheidungsbaumes zur Klassifikation eines Medikamentes auf Basis von Genexpressionsprofilen (Zhao & So, 2019, S. 224) ..... 27
Abbildung 6: "Swanson's ABC Modell" – A zeigt das geschlossene, B zeigt das offene Entdeckungsmodell (Andronis et al., 2011, S. 359) ..... 36
Abbildung 7: Vergleichstabelle der Rechenzeiten verfügbarer biomedizinischer NER-Modelle (Neumann et al., 2019, S. 320) ..... 50
Abbildung 8: F1-Scores von "HunFlair" im Vergleich zu anderen NER-Systemen und Modellen (Weber et al., 2021, S. 3) ..... 54
Abbildung 9: F1-Scores von "Stanza" im Vergleich zu "BioBERT" und "scispaCy" (Zhang et al., 2021, S. 1898) ..... 55
Abbildung 10: Spezialisierte biomedizinische NER-Modelle von "scispaCy" mit verfügbaren Labels (Neumann et al., 2019) ..... 55
Abbildung 11: Ausschnitt der aus UMLS extrahierten MeSH-Ontologie zum Oberbegriff "Neuroectodermal Tumors" mit allen untergeordneten Konzepten (nlm.nih.gov, 2022b) ..... 67
Abbildung 12: Zusammengefasste Ergebnisse der Methode 1 (eigene Grafik) ..... 73
Abbildung 13: Zusammengefasste Ergebnisse der Methode 2 (eigene Grafik) ..... 78

## Tabellenverzeichnis

Tabelle 1: Inhalte und Ziele der klinischen Phasen nach interpharma.ch (2022b) .....7
Tabelle 2: Übersicht zu spezialisierten Korpussen mit Fachgebieten und verfügbaren Entitätstypen (eigene Tabelle).....49
Tabelle 3: Variationen der Methode 1 mit den jeweilig verwendeten Labels der "scispaCy" NER-Modelle (eigene Tabelle) .....60
Tabelle 4: Formulierte "ABC" Assoziationsketten und verwendete Datenbanken (eigene Tabelle) .....64
Tabelle 5: Für die Evaluation selbstdefinierte Kennzahlen mit Beschreibung (eigene Tabelle) .....70
Tabelle 6: Methode 1 Variationen: Kennzahlen bezüglich Ergebnisse (eigene Tabelle) .....75
Tabelle 7: Methode 2 Assoziationsketten: Kennzahlen bezüglich Ergebnisse (eigene Tabelle) .....79

## Abkürzungsverzeichnis

| | |
|---|---|
| ADME | Absorption, Distribution, Metabolism and Excretion |
| BERT | Bidirectional Encoder Representation from Transformers |
| CART | Classification and Regression Trees |
| CMap | Connectivity Map |
| DNN | Deep Neural Networks |
| DTI | Drug–Target Interaction |
| GBA | Guilt by Association |
| GDSC | Genomics of Drug Sensitivity in Cancer |
| GPU | Graphics Processing Unit |
| ML | Machine Learing |
| MEDLINE | Medical Literature Analysis and Retrieval System Online |
| MeSH | Medical Subject Headings |
| NER | Named Entity Recognition |
| NLM | National Library of Medicine |
| NLP | Natural Language Processing |
| NLTK | Natural Language Toolkit |
| NME | New Molecular Entity |
| RF | Random Forest |
| ROI | Return of Investment |
| SVM | Support Vector Machine |
| TF-IDF | Term Frequency-Inverse Document Frequency |
| UMLS | Unified Medical Language System |
| USA | United States of America |

# 1 Einleitung

Im Jahr 2021 wurden durch die Zulassungsbehörde Swissmedic insgesamt 45 neue Humanarzneimittel in der Schweiz zugelassen (swissmedic.ch, 2022, S. 2). Obwohl die Schweiz mit Roche und Novartis zwei der grössten Pharmaunternehmen der Welt mit einem kombinierten Weltmarktanteil von ca. 9% und kombinierten Jahresumsatz von fast 100 Milliarden US-Dollar beherbergt, gelangten von diesen beiden Unternehmen im Zeitraum 2008 bis 2018 im Durchschnitt jährlich nur knapp 19 Medikamente den Schweizer Markt (interpharma.ch, 2019, S. 39, 56). Im Direktvergleich wurden in den USA im gleichen Zeitraum auch durchschnittlich nur 28 neue Medikamente pro Jahr zugelassen (statista.com, 2020, S. 28). Dieser Umstand, weshalb jährlich nur so wenige neue Medikamente den öffentlichen Markt erreichen, hat viele verschiedene Gründe (Dudley et al., 2011, S. 303).

Ein Grund dafür sind die heutzutage vorherrschenden konservativen Arzneimittelentwicklungsstrategien der Pharmaunternehmen. Die meistverwendeten Strategien zielen geschlossen auf die Entdeckung einer neuen therapeutisch wirksamen Substanz in einem vorbestimmten und damit stark eingegrenzten Einsatzgebiet (wie bspw. Onkologie oder Virologie) ab, bei der jeweils die neue Substanz eine eingeschränkte, spezifische und gezielte Wirkung haben soll. Damit werden schon anhand des vorformulierten Ziels mögliche alternative Wirkungen oder Resultate ausgeschlossen. Ein weiterer Grund liegt darin, dass bei einer dennoch erfolgreichen Entdeckung einer solchen Substanz, eine lange, langsame, teure und riskante experimentelle klinische Phase für den Test dieses potenziellen Medikamentes folgt (Dudley et al., 2011, S. 303). Die Durchfallquote neuer Medikamente in der klinischen Phase ist seit Jahren sehr hoch und stellt damit eine der grössten Herausforderung für die Pharmaunternehmen dar (Chiang & Butte, 2009, S. 509; van Vleet et al., 2019, S. 1). So können beispielsweise unerwartet auftretende Nebenwirkungen eines Medikamentes in der klinischen Phase den zuvor aufwändigen Entwicklungsprozess vollständig zum Halt bringen (Dudley et al., 2011, S. 303). Darüber hinaus stellen auch die international verschiedenen Prozesse der Arzneimittelzulassung eine weitere Hürde der Medikamentenentwicklung mit den damit verbundenen Kosten in Form von Geld oder Zeit dar. Erst kürzlich im Mai 2022 äusserte sich "Interpharma Schweiz", der Verband der forschenden pharmazeutischen Firmen in der Schweiz, kritisch gegenüber Swissmedic, da die Zeitspanne der Marktzulassung bis zur Aufnahme eines neuen Medikamentes in die Spezialitätenliste aktuell durchschnittlich 217 Tage beträgt. Mit durchschnittlich 42 Tagen war diese Zeitspanne im Jahr 2015 deutlich kürzer.

Dieser Umstand ist vor allem für Patienten fatal, welche unter Zeitdruck auf neue innovative Behandlungsmöglichkeiten angewiesen sind (interpharma.ch, 2022a, S. 1). Hinzu kommen sich stetig verändernde Zulassungsrichtlinien der unterschiedlichen Behörden, welche die Zulassungen von neuen Medikamenten erschweren und verlangsamen (Pushpakom et al., 2019, S. 41). Alle diese kostentreibenden Faktoren haben dazu geführt, dass schätzungsweise der "Return on Investment" (ROI) bei der Entwicklung eines neuen Medikamentes negativ ausfällt (Beachy et al., 2014, S. 32). Investitionen in die Medikamentenentwicklung sind damit aus der profitorientierten Sicht unattraktiv.

Eine vielversprechende alternative Lösung für diese beträchtlichen Hindernisse und Schwierigkeiten bei der Entwicklung neuer Arzneimittel zur Behandlung von Krankheiten, stellt die Wiederverwendung eines bereits entwickelten Arzneimittels dar (Issa et al., 2021, S. 132). Dieses Konzept der Identifikation und Entwicklung neuer Verwendungszwecke bereits bekannter Arzneimittel und Wirkstoffe wird als "**Drug Repurposing**" bezeichnet (Ashburn & Thor, 2004, S. 673). Dabei setzt sich "Drug Repurposing" das fundamentale Ziel (neue) Krankheiten zu finden, welche mit bereits bekannten oder klinisch getesteten Medikamenten behandelt werden könnten oder umgekehrt formuliert bereits entwickelte Wirkstoffe zu finden, welche für eine bisher unbekannten Behandlung einer Krankheit geeignet wären. Dabei kann das bereits vorhandene Wissen zu den Medikamenten zu deren Sicherheit, Zielmolekülen und Wirkmechanismen durch die bereits getätigte Forschung genutzt werden (Wang et al., 2019, S. 74). Die Entdeckung völlig neuer Indikationen solcher Medikamente, wie bspw. anhand ihrer Nebenwirkungen, ist dabei sehr lukrativ. Bei bereits bekannten Arzneimitteln können in einem solchen Fall auf viele sonst nötige Entwicklungsschritte für die Marktzulassung verzichtet werden (Tanoli et al., 2021, S. 1657). Viele berühmte Beispiele und Erfolgsgeschichten des "Drug Repurposing" umfassen Medikamente wie “Minoxidil“ (ursprünglich für die Behandlung von Bluthochdruck getestet, heute eingesetzt gegen Haarausfall), "Viagra" (ursprünglich getestet als Mittel gegen "Angina pectoris“, ein anfallsartiger Schmerz in der Brust verursacht durch Herzkrankheiten, heute universalbekanntes Mittel gegen Erektionsstörungen sowie Lungenhochdruck) und "Avastin" (ursprünglich nur für Behandlungen von nicht-kleinzelligem Lungenkrebs und metastasierendem Dickdarmkrebs, später auch für die Behandlung von metastasierendem Brustkrebs zugelassen) (Dudley et al., 2011, S. 303).

Aktuell werden verstärkt digitale Werkzeuge für das "Drug Repurposing" verwendet, welche die grossen vorhandenen Wissensbestände von medizinischen Daten effizient nutzen können (Alaimo & Pulvirenti, 2019, S. 100; Jin & Wong, 2014, S. 638). Dieses

Wissen, überwiegend standardisiert festgehalten in spezifischen Datenbanken, liegt jedoch auch vermehrt in unstrukturierten Textdaten vor. Die in den letzten Jahren stetig gewachsene Menge an medizinischen Daten sowie die erhöhte Anzahl unterschiedlicher verfügbarer und verwendeter Repositorien hat zum Problem geführt, dass sich die Daten dieser unterschiedlichen Repositorien bzw. Datenbanken in ihrer Qualität als auch Zuverlässigkeit deutlich unterscheiden (Neumann et al., 2019, S. 319; Tanoli et al., 2021, S. 1657). So sehen sich Forscherinnen und Forscher vermehrt mit der Herausforderung der Auswahl einer für den Anwendungsfall geeigneten Datenbank konfrontiert. Parallel verbirgt sich eine riesige Menge von exklusivem medizinischem Wissen in verschiedenen Arten von unstrukturierten Textdaten wie bspw. in klinischen Berichten, wissenschaftlichen Forschungsdokumenten oder Fachzeitschriften (Andronis et al., 2011, S. 364). Die technologischen Fortschritte im Bereich des Text-Mining und des Nature Language Processing (NLP) ermöglichen es Forscherinnen und Forschern heutzutage, vorhandene Assoziationsbeziehungen zwischen vielen Arten von biomedizinischen Entitäten, wie Genen, Arzneimitteln und Krankheiten innerhalb solcher Textdaten festzustellen (Andronis et al., 2011, S. 364).

Basierend auf der Warnung von Neumann et al. (2019, S. 319) zu den wachsenden Dateninkonsistenzen der verschiedenen Datenbanken sollen im Rahmen des Forschungsvorhabens dieser Arbeit unterschiedliche Methoden entwickelt und getestet werden, welche das "Drug Repurposing" für die Voraussage neuer potenzieller Kandidaten von Medikamenten oder chemischen Wirkstoffen mithilfe unstrukturierter Textdaten unterstützen können. Dafür sollen im Rahmen eines Fallbeispiels mehrere Prozesse bzw. Workflows, basierend auf einer wissenschaftlich orientierten Auswahl einer "besten" Datenbank und aus einer ausgewählten Methode für die praktischen Nutzung von Textdaten, entwickelt werden. Die forschungsleitende Frage im Rahmen dieser Arbeit lautet also wie folgt: *"Wie können unstrukturierte Textdaten für die Ermittlung neuer "Drug Repurposing" Kandidaten nutzbar gemacht werden und wie können sie Datenbanken ergänzen?"*

Zu Beginn wird im zweiten Kapitel das allgemeine Themengebiet der Medikamentenentwicklung und die Rolle des "Drug Repurposing" anhand mehrerer Facetten vorgestellt.

In Kapitel 3 wird der Forschungsstand und Forschungsbereich der "Computational Methods" für das "Drug Repurposing" detailliert dargestellt und die Rolle von Datenbanken sowie unstrukturierten biomedizinischen Textdaten erläutert.

In Kapitel 4 wird die Forschungsarbeit anhand der durch die Fachliteratur formulierte forschungsleitende Fragestellung sowie die zugehörigen Unterfragen und zusätzlichen Fragestellungen vorgestellt.

In Kapitel 5 wird das Forschungsdesign, die entwickelte Methodik sowie das Auswertungsvorhaben detailliert erläutert und begründet.

In Kapitel 6 werden die Ergebnisse und Erkenntnisse der Forschungsarbeit anhand der formulierten forschungsleiteten Frage, der Unterfragen sowie der zusätzlichen Fragestellungen beantwortet und diskutiert.

Abschliessend werden in Kapitel 7 ein Fazit zur gesamten Forschungsarbeit und den festgestellten Forschungsgrenzen sowie Empfehlungen für weitere Forschungsvorhaben erläutert. Am Schluss wird mit einer kritischen Selbstreflexion und einem persönlichen Schlusswort auf die Erarbeitung der Masterthesis zurückgeblickt.

# 2 Medikamentenentwicklung und die Rolle von "Drug Repurposing"

Wie zuvor erläutert verbirgt sich hinter der Entwicklung und Zulassung eines neuen Arzneimittels ein zeitintensiver, kostspieliger und risikoreicher Prozess. Dabei stellt die sogenannte experimentale Testphase eines Wirkstoffes die grösste Hürde dar (Dudley et al., 2011, S. 303). Sehr viele Medikamente scheitern in oder sogar vor der experimentellen Testphase des Wirkstoffes. Es gibt dabei drei unterschiedliche Kategorien der experimentellen Wirkstoffentwicklung (cadfem-medical.com, 2022):

- **"in vivo":** Umfassen Experimente im menschlichen Körper bzw. Experimente in einem klinischen Umfeld an Lebewesen
- **"in vitro":** Umfassen Experimente ausserhalb des menschlichen Körpers oder Lebewesen bspw. in Reagenzgläsern
- **"in silico":** Umfassen Experimente auch ausserhalb des Körpers, jedoch innerhalb einer virtuellen Umgebung am Computer mit Methoden wie bspw. Simulation

Bevor sich ein Medikament in klinische Studien bzw. die klinische Phase begeben kann, muss dieses in der sogenannten präklinischen Phase bezüglich Toxizität getestet werden. Besonders im Rahmen von "in vivo" Tierversuchen kommt der Fall sehr oft vor, dass, wenn eine sehr hohe Toxizität eines Medikaments festgestellt wird, der Entwicklungsprozess des neuen Wirkstoffes vollständig abgebrochen wird (van Vleet et al., 2019, S. 1). Somit stellt die Arzneimittelsicherheit ein kritischer Faktor für den Erfolg der Medikamentenentwicklung dar (Hodos et al., 2016, S. 196). Im Falle einer festgestellten hohen Toxizität des Wirkstoffes bei Tieren, besteht aber die potenzielle Chance, dass die initiale Beurteilung der Toxizität sich nicht eins zu eins auf den toxischen Effekt des Wirkstoffes beim Menschen übertragen lässt. Im Fall eines Entwicklungsstopps können weitere "in vitro" und "in silico" Verfahren solche Medikamente und dessen verbundene Entwicklungskosten retten (van Vleet et al., 2019, S. 1). So wird bei einer Weiterentwicklung von alten aufgegebenen Arzneistoffen oder medizinischen Produkten oft von "drug rescue" gesprochen (Langedijk et al., 2015, S. 1032; Tanoli et al., 2021, S. 1657).

## 2.1 Klinische Studien und Hürden der Medikamentenzulassung

Nach einem erfolgreichen präklinischen Test zur Toxizität eines neuen Arzneimittels, folgt im nächsten Entwicklungsschritt die klinische Studie. Dieser Übergang von der präklinischen Phase in die klinische Phase stellt für jedes Pharmaunternehmen oder jede

akademische Institution meist ein bedeutender Fortschritt in der Entwicklung dar, da die Arzneimittelkandidaten in der vorhergehenden präklinischen Phase mit intensivem Aufwand optimiert und getestet wurden (Sun et al., 2022, S. 1).

Wie interpharma.ch (2022b) für die Situation in der Schweiz zusammenfasst, muss jedoch bevor die klinische Studie für ein neues Arzneimittel beginnen darf, eine Zustimmung der zuständigen nationalen Zulassungsbehörde (in der Schweiz ist dies Swissmedic) und der zugehörigen Ethikkommission, bestehend aus erfahrenen Personen der Bereiche Medizin, Theologie, Recht und der Öffentlichkeit, vorliegen. Für die Zustimmung einer klinischen Studie wägen sie gestützt auf die vorhandenen Voruntersuchungen ab, ob und unter welchen Auflagen zum Schutz der Studienteilnehmenden die klinische Studie aus ethischer, medizinischer und rechtlicher Sicht durchgeführt werden darf. Erst nach einer erfolgten Zustimmung und der Erfüllung aller Auflagen der Zulassungsbehörde kann die klinische Studie gestartet werden.

Klinische Studien werden grundsätzlich in drei individuell gestaltete Phasen I, II und III strukturiert und verfolgen dabei folgende Ziele (interpharma.ch, 2022b):

| **Inhalte und Ziele** | |
|---|---|
| **Phase I** | In der Phase I werden geringe Mengen des neuen Wirkstoffes an eine kleine Anzahl gesunder Testprobanden verabreicht. Da diese Testprobanden gesund sind, setzt sich diese Phase das Ziel, die Vorhersagen aus den "in vivo" – Tierversuchen oder "in silico"-Simulationen zur toxischen Verträglichkeit, Aufnahme, Verteilung und Umwandlung sowie Ausscheidung des Wirkstoffes zu bestätigen. Gleichzeitig sollen die maximal verträgliche und geeignete Dosierung des Wirkstoffes ermittelt werden und unerwartet auftretende Nebenwirkungen erfasst werden (Sun et al., 2022, S. 4–5). |
| **Phase II** | In der nächsten Phase II wird der Wirkstoff in einer erweiterten Zusammenarbeit mit Kliniken, wie bspw. Universitätsspitälern oder anderen medizinischen Einrichtungen, erstmals an erkrankten Patientinnen und Patienten getestet. Dabei weisen die behandelnden Ärztinnen und Ärzte an der Zielkrankheit erkrankte Personen auf die Möglichkeit der Teilnahme an der Studie hin und betreuen diese. Im Durchschnitt nehmen in dieser Phase 100 bis 500 Erkrankte an der Studie teil. Parallel zur Phase I soll die maximale Verträglichkeit sowie die optimale Dosierung des Wirkstoffes ermittelt werden. Gleichzeitig werden erstmals Kontrollgruppen eingesetzt. Dabei wird parallel in einer gleich grossen Gruppe ein Placebo eingesetzt. Die Einteilung in diese zwei Gruppen erfolgt meistens per Zufallsprinzip und wird von den Patientinnen und Patienten sowie auch in sogenannten Doppelblindstudien auch vor den Ärztinnen und Ärzten geheim gehalten. Durch diese Massnahmen soll vermieden werden, dass Hoffnungen oder Befürchtungen der Studienteilnehmenden die Wirkung des Medikaments beeinflussen. |

| | |
|---|---|
| **Phase III** | In der letzten klinischen Phase III wird das neue Medikament an mehr als tausend Patientinnen und Patienten getestet, um nun auch die Verträglichkeit und Wirksamkeit bei sehr vielen unterschiedlichen Menschengruppen zu testen. Vorwiegend wird in diesen Studien intensiv beobachtet, ob ein Medikament bei unterschiedlichen Menschen mit unterschiedlichen Merkmalen anders oder gar nicht wirkt. Meistens können die Unterschiede der Wirkung eines Medikamentes auf die individuellen Eigenschaften des Patienten wie bspw. Blutwerte oder genetische Besonderheiten zurückgeführt werden. Die Bestimmung solcher Indikatoren, welche die Wirkung von Medikamenten beeinflussen, werden Biomarker genannt. Solche Biomarker können auch vorweg in der präklinischen Phase an Tieren erkannt oder in Computersimulationen identifiziert werden (Sun et al., 2022, S. 1). Parallel zu dieser Überprüfung der Wirksamkeit in der breiten Masse werden auch die Wechselwirkungen des Medikaments mit anderen Wirkstoffen untersucht.<br><br>Am Ende dieser klinischen Phase soll die Erfüllung der am Entwicklungsstart gesetzten "Endpunkte" geprüft werden, welche die gewünschten Effekte der Wirkung des Medikamentes beschreiben. |

Tabelle 1: Inhalte und Ziele der klinischen Phasen nach interpharma.ch (2022b)

Nur wenn ein Medikament in allen Phasen positive Testergebnisse zu Sicherheit, Wirksamkeit und Qualität vorzuweisen hat, ist es für eine potenzielle Markteinführung geeignet (Dudley et al., 2011, S. 303; Pushpakom et al., 2019, S. 41; van Vleet et al., 2019, S. 1). Laut "CMR International", einer reputablen Quelle zu Metriken und Trendanalysen der Pharmaindustrie, lag die Wahrscheinlichkeit der erfolgreichen Zulassung eines vollkommen neuen Wirkstoffes, welcher die klinischen Studien während des Zeitraumes 2015 bis 2017 erreichte, nur bei 7% (Dowden & Munro, 2019, S. 495). Anders formuliert bedeutet dies, dass durchschnittlich 9 von 10 neuen Arzneimittelkandidaten nach Eintritt in klinische Studien während der Phasen I, II, III und der finalen Arzneimittelzulassung scheitern. Wenn zusätzlich zu dieser Statistik auch noch die Arzneimittelkandidaten, welche sich während dieses Zeitraumes in der präklinischen Phase befanden, hinzugezählt werden, steigt die Misserfolgsrate der erfolgreichen Entdeckung und Zulassung eines Arzneimittels sogar über 90% (Dowden & Munro, 2019, S. 495). Sun et al. (2022, S. 1) stellen in Abbildung 1 den klassischen Prozess der Medikamentenentwicklung einer NME ("new molecular entity") mit den zugehörigen Erfolgswahrscheinlichkeiten der jeweiligen Schritte dar:


Abbildung 1: Abgebildeter Prozess der Medikamentenentwicklung mit der Anzahl Wirkstoffkandidaten und den zugehörigen Erfolgswahrscheinlichkeiten, Zykluszeiten und prozentualen Gesamtkostenanteilen jedes Entwicklungsschrittes (Sun et al., 2022, S. 1)

In einer erweiterten Analyse klinischer Daten des Zeitraums 2010 bis 2017 wurden verschiedene Gründe erfasst und klassifiziert, welche für den Misserfolg in den klinischen Phasen führen können (Dowden & Munro, 2019, S. 496; Lipinski, 2000, S. 248; Sun et al., 2022, S. 1):

- fehlende klinische Wirksamkeit des Arzneimittels
- zu hohe Toxizitätswerte
- schlechte arzneimittelähnliche Eigenschaften in den Bereichen der Absorption, Verteilung, Metabolismus und Ausscheidung (auch abgekürzt als ADME)
- fehlendes kommerzielles Potenzial
- schlechte gesamtstrategische Planung

Ganze 40 bis 50% aller Misserfolge von Medikamenten in klinischen Studien konnten im Rahmen dieser Analyse auf den ersten Grund, der mangelnden Wirksamkeit und Effektivität des Arzneimittels, zurückgeführt werden. Dies erklärt die zuvor erwähnte Tatsache, dass die Medikamentenentwickler vor allem viel Energie in die Verbesserung der Wirksamkeit eines Medikamentes vor Eintritt in die klinische Studie investieren (Sun et al., 2022, S. 2). Hohe Toxizitätswerte von Medikamenten lassen sich generell entweder durch die unterschätzten bzw. unbeabsichtigten Off- oder On-Target-Effekte der molekularen

Ziele, auch "drug targets" genannt, erklären (Sun et al., 2022, S. 3). Auch bei einer erfolgreichen Zulassung eines Medikamentes nach Bestehen aller klinischen Tests, gibt es weiterhin oft Schwierigkeiten bei der eins-zu-eins Übertragung der vorwiegend in der präklinischen Phase an Tieren getesteten Wirkungen des Medikamentes auf den Menschen, dadurch allfällige Nebenwirkungen generell oder bei gezielten menschlichen Teilpopulationen (bspw. solche mit Vorkrankheiten) nicht rechtzeitig erkannt werden (van Vleet et al., 2019, S. 1). Dieser Mangel in der systematischen Erfassung aller potenziell möglichen oder zusätzlichen Indikationen eines Medikamentes in der klinischen Phase, hat zur Folge, dass nicht antizipierte Nebenwirkungen auch nach einer Markteinführung auftreten. Aus diesem Grund kommen teure Rückrufaktionen von bereits vermarkteten Medikamenten häufig vor (Dudley et al., 2011, S. 303; van Vleet et al., 2019, S. 1).

## 2.2 Das Potenzial von "Drug Repurposing"

Aufgrund der hohen Misserfolgsraten von neuen Medikamenten in klinischen Studien, stellt "Drug Repurposing" eine grosse Chance dar, um aufgegebene Wirkstoffe über das Konzept des "drug rescue" wieder potenziell nutzbar zu machen (Langedijk et al., 2015, S. 1032; Tanoli et al., 2021, S. 1657). Das durch die vielen gescheiterten Entwicklungsvorhaben bereits vorhandene Wissen zu Medikamenten in Bezug auf ihre Sicherheit, Zielmoleküle und Wirkmechanismen, kann im Rahmen eines “Repurposing” Vorhabens verwendet werden (Wang et al., 2019, S. 74). So stellt beispielsweise die Nutzung von Nebenwirkungen als neue Indikationen für Medikamente eine sehr lukrative Vorgehensweise dar, da in solchen Fällen auf einige sonst nötige Entwicklungsschritte für die Marktzulassung verzichtet werden kann, weil Nebenwirkungen meistens erst ab der I. klinischen Phase erfasst werden (Tanoli et al., 2021, S. 1657).

Wie Hurle et al. (2013, S. 335) dagegen betonen, ist es für das "Drug Repurposing" weiterhin sehr wichtig, dass man sich bei der Auswahl eines "Repurposing" Kandidaten entweder in Form eines teil-erfolgreich getesteten oder auch eines aufgegebenen Wirkstoffes darauf fokussiert, welcher potenzielle Kandidat den Patienten am schnellsten zur Verfügung stehen kann. Dafür sollten im Idealfall die ausgewählten Wirkstoffe, die nötigsten Sicherheitstest schon bestanden haben. Das Nutzen-Risiko-Verhältnis spielt daher im "Drug Repurposing" eine grosse Rolle und dieses kann je nach dem Nutzungsszenario des Medikamentes variieren. So können beispielsweise Medikamente, deren Nutzung aufgrund von besonders gefährlichen Nebenwirkungen bei chronischen Krankheiten wie bspw. Diabetes oder Asthma nicht vertretbar sind, jedoch aber für die Behandlung von

lebensbedrohenden Krankheiten wie z.B. Krebs eine grosse Nutzungschance darstellen. "Drug Repurposing" ist demzufolge ein besonders attraktives Konzept für die Behandlung seltener Krankheiten, angesichts wissenschaftlicher sowie kommerzieller Gründe (Hodos et al., 2016, S. 186; Pushpakom et al., 2019, S. 44). Zur erweiterten Optimierung des Nutzen-Risiko-Verhältnisses von Medikamenten ist es demnach von grösster Bedeutung, dass während des Arzneientwicklungsprozesses die feine Balance zwischen der Dosismenge, Wirksamkeit und Toxizität gefunden wird (Sun et al., 2022, S. 11).

Dennoch ist nicht jedes "Repurposing" Vorhaben von vornherein erfolgreich. Die meisten "Repurposing" Kandidaten scheitern ebenso, wie neue Wirkstoffe in der klinischen Studie. Vorwiegend allerdings nicht aufgrund von überschreitenden Toxizitätswerten, da das zugehörige Sicherheitsprofil des Medikaments grösstenteils schon durch die früheren Studien der vorhergegangenen Entwicklungsphase bekannt ist, sondern (wie die Mehrheit der Wirkstoffe) aufgrund der mangelnden Wirksamkeit (Pushpakom et al., 2019, S. 51). Des Weiteren gibt es auch nach der klinischen Studie weitere Gründe wie Patentfragen, ordnungsrechtliche Erwägungen oder organisatorische Hürden, welche ein "Repurposing" Vorhaben verlangsamen und sogar stoppen können (Pushpakom et al., 2019, S. 51).

Obwohl viele Ärzte sowie Pharma- und Biotechnologieunternehmen über eigene Möglichkeiten und die nötigen Vorkenntnisse verfügen, um ein "Drug Repurposing" Vorhaben in verfügbare klinische Studien zu integrieren oder fortzuführen, werden solche Gegebenheiten vorwiegend zufällig oder sehr selten genutzt. Dies basiert auf dem Grundproblem, dass diese "alternativen" therapeutischen Indikationen eines Medikamentes (abgezielte Wirkung) für die prospektive Prüfung schon zuvor bekannt sein müssen (Chiang & Butte, 2009, S. 507).

## 2.3 "Drug Repurposing": Definitionen und Prozesse

In der englischen Fachliteratur werden für das Konzept viele verschiedene Begriffe wie "repositioning", "repurposing", "redirecting", "reprofiling" oder "rediscovery" als Synonyme verwendet (Langedijk et al., 2015, S. 1028).

> *"The term drug repositioning is frequently used in the literature and has several synonyms such as drug repurposing, which have been used interchangeably. No common definition of drug repositioning or indeed for other similar terms has been found in the literature. Moreover, the definitions differed significantly in their wording used for the features, often leading to essential differences in their meaning. In the future, incentives might be established to stimulate drug repositioning and related activities that – from a legal or regulatory perspective – require clear terminology and a consistent definition."*
> *(Langedijk et al., 2015, S. 1033)*

Wie Langedijk et al. betonen, gibt es bisher noch keine klare einheitliche Nomenklatur rund um das Konzept des "Drug Repurposing". Um sich diesem Problem anzunehmen, analysierten Langedijk et al. (2015) mit quantitativen und qualitativen Methoden die Fachliteratur auf Unterschiede und Gemeinsamkeiten der jeweils unterschiedlich verwendeten Begriffe und Definitionen. Dabei konnten sie vier gemeinsame Kernelemente identifizieren (Langedijk et al., 2015, S. 1030–1032):

- **"Use":** Alle identifizierten Definitionen und Begriffe bezogen sich auf einen Nutzungsfall bzw. ein Nutzungsbedürfnis. Dies konnte bspw. ein beschriebenes klinisches Szenario oder ein therapeutischer Anwendungsfall sein. Trotz dieser unterschiedlichen Spezifizierungen bezogen sich alle Fälle auf die Behandlung einer Krankheit.
- "**Product":** Es stand immer ein Produkt bzw. eine Substanz im Mittelpunkt. Meist wurde der sehr allgemeine Begriff "drug" verwendet, welcher sich aber je unterschiedlich entweder auf ein fertiges medizinisches Produkt bzw. Medikament oder nur auf eine pharmakologisch aktive Substanz bzw. Wirkstoff bezog.
- **"Action":** Das Kernelement "Action" umschrieb, welchen Hauptzweck und welches grundlegende Vorgehen die verwendeten Definitionen beschrieben. Dabei wurden drei kategorisierten Zwecke erfasst:
    - das Identifizieren neuer Anwendungsmöglichkeiten durch Screening von pharmazeutischen Wirkstoffen, um für diese neue Verwendungen zu entdecken oder vorzuschlagen
    - die explizite Verwendung von Arzneimitteln oder Medikamenten für neue Anwendungen

    - die Entwicklung und Forschung neuer Anwendungen und Prozesse, wie bspw. die Forschung an einem Arzneimittel für dessen Marktzulassung
- **"Concept":** Unter "Concept" wurden Strategien, Ansätze sowie vor allem Prozesse wie bspw. detaillierte und praxisbezogene Workflows für eine vollständige "Drug Repurposing" Pipeline beschrieben.

Wie zuvor erwähnt, basierte die Mehrzahl der erfolgreichen Entdeckungen und Umsetzungsfälle des "Drug Repurposing" bisher ungeplant und zufällig. Sobald in klinischen Studien zufällig festgestellt wurde, dass ein Medikament einen bisher unbekannten, aber therapeutisch bedeutsamen Off-Target-Effekt besass, wurde diese Tatsache generell direkt kommerziell ausgenutzt (Pushpakom et al., 2019, S. 41).

Um dem Fehlen von "Systematik" im Bereich des "Drug Repurposing" entgegenzuwirken, hat die Anzahl entwickelter Methoden für das "Drug Repurposing" im Verlauf der Jahre deutlich zugenommen. Viele verschiedene Studien haben sich das Ziel gesetzt, eine effiziente Pipeline auf der Basis von verfügbarer Information zu Arzneimitteln und Krankheiten mit integrierten praktischen Methoden für das "Drug Repurposing" zu entwickeln (Jin & Wong, 2014, S. 638). Der Bereich der "computational methods" bietet besonders für die hypothetische Voraussagung von möglichen "Repurposing"-Kandidaten relativ schnelle, automatische und mechanisch-unvoreingenommene Methoden mit hohem Potenzial (Hurle et al., 2013, S. 335; Pushpakom et al., 2019, S. 44). Aufgrund der systematischen Prozesse können zudem auch die vorausgesagten Kandidaten statistisch bewertet und nachvollziehbar priorisiert werden. Dies kann "Repurposing" Vorhaben beschleunigen und qualitativ verbessern (Hurle et al., 2013, S. 335). Somit ermöglichen diese Methoden eine nachvollziehbare Auswahl von "Repurposing" Kandidaten, welche objektiv kosteneffizienter oder anhand ihrer Eigenschaften (Toxizität, klinischer Testfortschritt, Wirksamkeit, etc.) besser geeignet als andere sind (Alaimo & Pulvirenti, 2019, S. 98).

# 3 "Computational Methods" und die Rolle von Datenbanken

## 3.1 "Computational Methods"

Besonders "in silico" Technologien bieten eine in der Pharmaindustrie schon bewährte Lösung der in Kapitel 2.2 genannten Schwierigkeiten bei der Medikamentenentwicklung, alle möglichen Indikationen eines Wirkstoffes vor der klinischen Phase zu bestimmen. Diese Technologien können die Entdeckung und Evaluation neuer therapeutischer Indikationen von alten oder neuen Wirkstoffen unterstützen (Alaimo & Pulvirenti, 2019, S. 100). Mit den verfügbaren Werkzeugen der Bioinformatik, wie bspw. Computermodellierung oder Computersimulationen, können komplexe Analysen von biowissenschaftlichen Anwendungen intuitiv, präzise und reproduzierbar durchgeführt werden (cadfem-medical.com, 2022). Diese sogenannten "computational methods" bzw. computergestützten Methoden können im Gegensatz zu traditionellen Labormethoden einen riesigen Wissensbestand in Form von Daten nutzen, um neue zusätzliche Kontexte und Inhalte zu bisher unbekannten Mechanismen zu entdecken (Alaimo & Pulvirenti, 2019, S. 100; Jin & Wong, 2014, S. 638). Gleichzeitig gibt es eine hohe Anzahl von aufgegebenen Wirkstoffen, welche im Rahmen der traditionellen "Drug Repurposing" Methoden meist ungenutzt bleiben, jedoch bei "computational methods" ohne allfällige Risiken für die Analysen und Simulationen hinzugezogen werden können (Jin & Wong, 2014, S. 638). "Computational methods" sind heutzutage besonders wichtig bei der Suche von Wirkstoffen zur Behandlung von seltenen Krankheiten, da die zugehörige Entwicklung neuer Moleküle oder Arzneimittelkomponenten wirtschaftlich nicht tragbar ist (Alaimo & Pulvirenti, 2019, S. 100).

## 3.2 Rolle von medizinischen Datenbanken

Um neue Anwendungsmöglichkeiten eines Arzneimittels zu finden, müssen zuerst neue Zusammenhänge zwischen Arzneimitteln und Krankheiten gefunden werden, welche durch entsprechende Daten aus dem Medizinbereich gestützt werden (Wang et al., 2019, S. 74). Dabei spielen aktuell Datenbanken eine zentrale Rolle, weil sie eine unabdingbare Basis für die meisten modernen computergestützten Verfahren des "Drug Repurposing" darstellen (Masoudi-Sobhanzadeh et al., 2020, S. 1087). Daten zu Genexpressionen, zu Wechselwirkungen zwischen Medikamenten und Wirkstoffzielen, zu Proteinnetzwerken, zu elektronischen Gesundheitsakten, Berichte über klinische Studien und dazugehörige Berichte über unerwünschte Nebenwirkungen sind inzwischen in standardisierter Form

in zahlreichen Datenbanken verfügbar (Alaimo & Pulvirenti, 2019, S. 97). Diese Daten sind hingegen meistens komplex, hochdimensional sowie verrauscht und stellen bei deren Nutzung eine grosse Herausforderung, aber auch ein grosses Potenzial dar. "Computational Methods" bzw. computergestützte Methoden bieten die Möglichkeit, solche Daten richtig einzuordnen und passend zu verarbeiten. Mit ihnen kann einerseits die Medikamentenforschung beschleunigt werden, anderseits können auch neue Erkenntnisse über Wirkstoffmechanismen, Nebenwirkungen und Wechselwirkungen gewonnen werden (Hodos et al., 2016, S. 187). Vor allem können diese Methoden mit ihren verbesserten Fähigkeiten zur Modellierung sowie Vorhersage von Arzneimittelnebenwirkungen und schädlichen Wirkungen, die Effizienz der Medikamentenforschung entscheidend verbessern. Dies, indem sie frühzeitig unerwünschte Eigenschaften von Wirkstoffen wie bspw. Toxizität erkennen und damit weitere allfällige Investitionen von Ressourcen in einen untauglichen Wirkstoff verhindern (Hodos et al., 2016, S. 196).

Da jedoch die Grössen sowie Komplexitäten von Screening-Verfahren und Datenbanken immer weiter zunehmen, wachsen auch die Herausforderungen bei der Verwaltung, Analyse und Interpretation dieser verfügbaren Daten (van Vleet et al., 2019, S. 15). Das daraus verursachte Problem der Informationsüberflutung wird in den biomedizinischen Wissenschaften zusätzlich dadurch verstärkt, indem die meisten Forschungspapiere stetig neue und spezifische Entdeckungen zu den Wechselwirkungen zwischen einer Vielzahl an Genen, Wirkstoffen und Proteinen beinhalten (Neumann et al., 2019, S. 319).

Dies führt zu einer hohen Anzahl verschiedener verfügbarer Repositorien bzw. Datenbanken, deren Daten sich in ihrer Qualität als auch Zuverlässigkeit unterscheiden (Tanoli et al., 2021, S. 1657). Die daraus resultierende mangelnde Gesamtübersicht der allgemein verfügbaren Daten stellt Forscherinnen und Forscher vor die Herausforderung, eine für den Anwendungsfall "richtige" Datenbank auszuwählen und überhaupt die geeignete Information zu finden. Tanoli et al. (2021) bieten eine zeitlich aktuelle Übersicht zu 102 verfügbaren Datenbanken im Themenbereich des "Drug Repurposing" und kategorisieren diese in vier nicht-exklusive Hauptkategorien und 17 Unterkategorien. Diese vier erfassten Hauptkategorien lauten:

- **"Chemical databases"**
- **"Biomolecular databases"**
- **"Drug-target interaction databases"**
- **"Disease databases"**

In Abbildung 2 wird ein Ausschnitt der von Tanoli et al. (2021) durchgeführten Kategorisierung der verfügbaren biomedizinischen Datenbanken dargestellt:

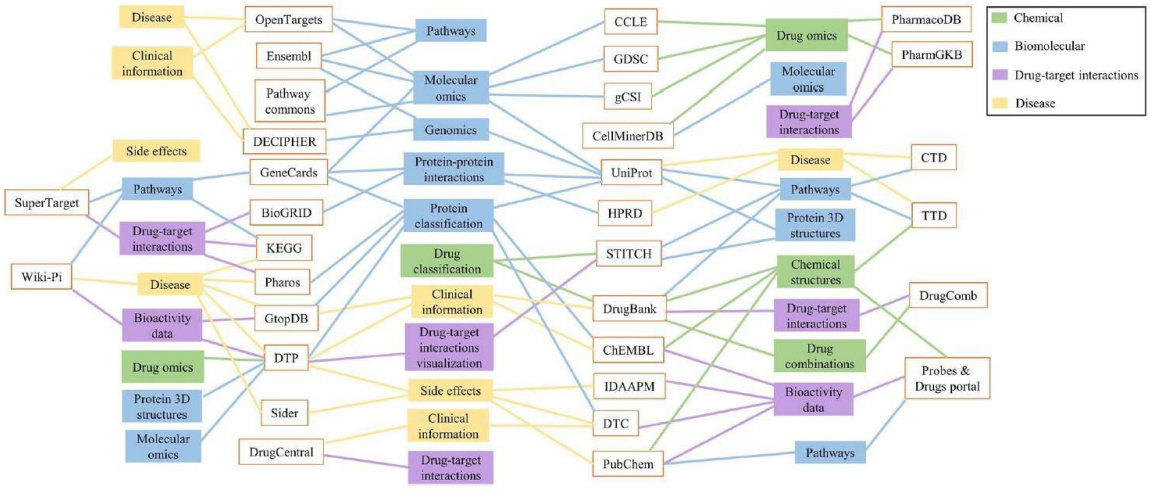


Abbildung 2: Verfügbare biomedizinische Datenbanken und kategorische Einteilung nach Tanoli et al. (2021, S. 1658)

Für die Einteilung der jeweiligen Datenbanken ist demnach entscheidend, welchen thematischen Umfang die verfügbaren Relationspaare der jeweiligen Datenbanken besitzen.

Weitergehend setzten sich Tanoli et al. (2021, S. 1671) das Ziel, für jede der 17 einzelnen Unterkategorien eine Empfehlung für deren Auswahl als "beste" Datenbank auszusprechen. Die Empfehlungen basierten dabei auf statistischen Eigenschaften zu Qualität, Verfügbarkeit, Datenredundanzen, Attribute, Vielfalt der Datentypen und Datenbanknutzung (Anzahl Zitationen).

## 3.3 "Computational methods" und "Repurposing" Vorgehensweisen

In den letzten Jahren wurden viele verschiedene Arten von Prozessen und Algorithmen im Bereich des "in silico – Drug Repurposing" entwickelt (Jin & Wong, 2014, S. 638). Diese Methoden setzten sich meistens das Ziel, die bekannten Prozesse zu systematisieren und zu imitieren, welche in der Vergangenheit zu den zufälligen (erfolgreichen) Entdeckungen des "Drug Repurposing" geführt haben (Alaimo & Pulvirenti, 2019, S. 100). Die neueren "computational methods" können das breite Wissen in Datenbanken dazu nutzen, die Entdeckung von neuen "Repurposing" Kandidaten zu beschleunigen (Alaimo & Pulvirenti, 2019, S. 97).

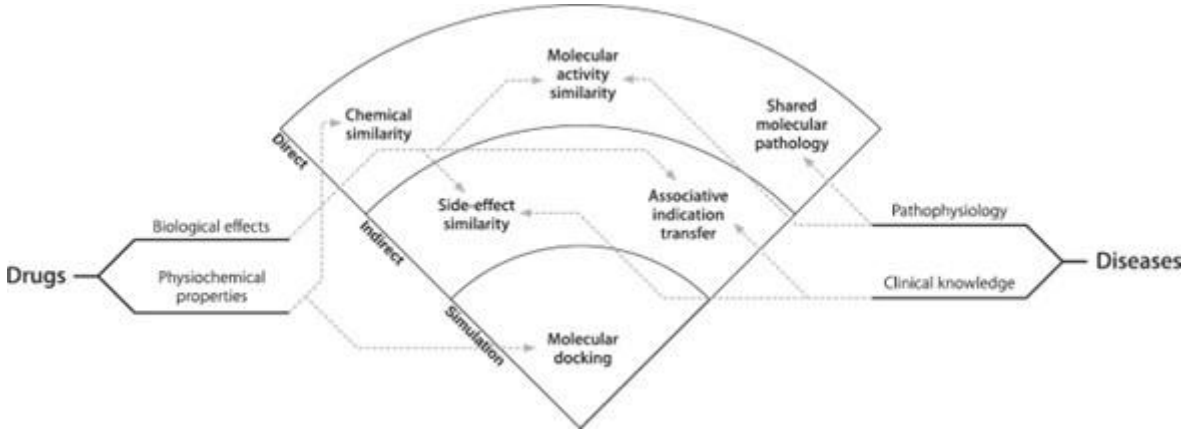


Abbildung 3: Übersicht verfügbarer "computational“ Vorgehensweisen und Strategien zur Ermittlung von Krankheit-Medikament Zusammenhängen (Dudley et al., 2011, S. 305)

Abbildung 3 zeigt die nach Dudley et al. (2011) verfügbaren Vorgehensweisen und Strategien für die Suche von Zusammenhängen zwischen Medikamenten und Krankheiten. Die gestrichelten Pfeile stellen dabei die verbundenen Informationsaspekte zwischen Medikamenten und Krankheiten dar, welche für die Suche von Zusammenhängen anhand Ähnlichkeiten genutzt werden können. Alaimo und Pulvirenti (2019, S. 100) fassen die in Abbildung 3 dargestellten Vorgehensweisen und Strategien, welche mit digitalen Werkzeugen verfügbar sind, in vier unterschiedliche theoretisch basierte Kategorien des Drug Repurposing" zusammen; **"target-based"**, **"side-effect-based"**, **"expression-based"** und **"similarity- based"**. Diese Vorgehensweisen nutzen die theoretischen Grundlagen der Genomik, Systembiologie, Netzwerkbiologie, Chemo- und Bioinformatik, um im besten Fall optimale "new target – known drug"-Paare zu identifizieren (Choudhury et al., 2022, S. 2). Die von Alaimo und Pulvirenti (2019) vorgestellte Kategorisierung schliesst aber die Möglichkeit nicht aus, dass die jeweiligen Vorgehensweisen auch kombiniert genutzt werden können. Die Auswahl der "richtigen" Vorgehensweise für das "Drug Repurposing" ist grundsätzlich von einer bestimmenden Fallbasis abhängig, welche dadurch bestimmt wird, wieviel und welches spezifizierte Wissen zu den Krankheiten oder Arzneimitteln bereits vorhanden ist (Jin & Wong, 2014, S. 639). Im Folgenden werden die Vor- und Nachteile dieser individuellen Vorgehensweisen behandelt.

### 3.3.1 "target-based"

Die "target-based" Vorgehensweisen nutzen die biologische Rolle der molekularen Zielstrukturen (drug-targets) zur Behandlung einer Krankheit durch Medikamente (Hodos et al., 2016, S. 198). Wenn bei einer Krankheit eine Liste von molekularen Zielstrukturen

vorhanden ist, können "drug-target interactions" Datenbanken genutzt werden, um anhand der Überschneidungen dieser Zielstrukturen neue Kandidaten von Arzneimitteln zu ermitteln (Alaimo & Pulvirenti, 2019, S. 100). Solche "targets" bzw. Zielstrukturen können z.B. Gene oder deren Produkte in Form von kleineren biomolekularen Strukturen wie bspw. Rezeptoren (Proteine) auf ausgewählten Zellen darstellen. Gleichzeitig kann eine sogenannte "drug-target interaction" (DTI) mit einer Vielzahl von experimentellen Techniken wie bspw. Bindungassays gemessen werden (Hodos et al., 2016, S.188). Dabei wird untersucht, ob und wie stark sich die Wirkstoffe an die Zielstrukturen binden (Issa et al., 2021, S. 132–133).

### 3.3.2 "side-effect-based"

"Side-effect-based" Verfahren basieren auf der Idee, dass therapeutisch beobachtete Nebenwirkungen von Medikamenten auch Hinweise auf neue alternative Anwendungsmöglichkeiten liefern können. Nach Definition sind Nebenwirkungen phänotypische Symptome, die von einem Medikament erzeugt werden, wenn der Wirkstoff sich an ein Off-Target bindet und dabei andere Signal- oder Stoffwechselwege stört (Balasundaram et al., 2019, S. 137). Die Identifizierung solcher neuen Nebenwirkungsassoziationen zu bereits zugelassenen Wirkstoffen stellt eine wichtige Aufgabe im Pharmabereich dar und wird unter dem Begriff der Pharmakovigilanz zusammengefasst (Hodos et al., 2016, S. 196). In einem Beispiel konnte man bei Patienten, welche wegen "benigner Prostatahyperplasie" (gutartige Vegrösserung der Prostata) mit "Finasterid" behandelt wurden, einen unerwarteten Haarwuchs feststellen. So wurde neu "Finasterid" auch als Mittel gegen "Androgenetische Alopezie" (Haarausfall) eingesetzt (Alaimo & Pulvirenti, 2019, S. 100). Diese Vorgehensweise ist jedoch auf bereits stattgefundene klinische Testphasen bzw. klinische Daten angewiesen und eignet sich damit nicht unbedingt für Wirkstoffe, welche sich in einer frühen Entwicklungsphase befinden oder vor der klinischen Phase aufgegeben wurden.

### 3.3.3 "expression-based"

Hinter den "expression-based" Vorgehensweisen verbergen sich die Schlüsselkonzepte der sogenannten "signature reversion" und des "signature matching". Dabei wird ein neuer "Repurposing" Kandidat dann erkannt, wenn das zugehörige Arzneimittel-Krankheits-Paar antikorrelierte Gen-Expressionsprofile aufweist. Dies bedeutet, wenn ein Gen infolge einer Krankheit gestört ist, könnte ein Medikament mit positiver Wirkung auf dieses

Gen ein neues potenzielles therapeutisches Mittel sein (Alaimo & Pulvirenti, 2019, S. 102; Issa et al., 2021, S. 136; Jin & Wong, 2014, S. 641). Alle Störungen, welche bspw. durch eine Einnahme eines Medikamentes oder durch eine Krankheit verursacht werden, können als Gen-Expressionsprofil für das Medikament oder die Krankheit dargestellt werden. Dabei wird jedem betroffenen Gen eine Zahl zugeordnet, die den Grad der Hoch- oder Herunterregulierung relativ zu einem Kontrollpunkt (z.B. die Differenz aller durchschnittlicher Expressionswerte) angibt (Hodos et al., 2016, S. 190). So ermöglichen Genexpressionsprofile quantitative molekulare Vergleiche zwischen den durch Krankheiten und den durch Medikamente verursachten "Störungen". Daher bieten Genexpressionsdaten eine hochdimensionale Anzeige der Zellzustände und biologischer Störungen, die sich aus der Arzneimittelbehandlung oder als Symptome einer Krankheit ergeben (Hodos et al., 2016, S. 199). Ein Vorteil dieser sogenannten Transkriptomdaten besteht darin, dass diese Art von Daten für nahezu jede chemische Verbindung oder Krankheit generiert werden kann, unabhängig vom Zulassungsstatus des Wirkstoffes sowie von Arzneimittel- oder Krankheitsmechanismen (Hodos et al., 2016, S. 199).

Die "Connectivity Map" (CMap) von Lamb et al. (2006) ist eine der ältesten und bedeutendsten Quellen solcher Transkriptomdaten. Mit "CMap" haben Lamb et al. (2006) als Pioniere in der Forschung versucht, Medikamente und Krankheiten anhand von Genexpressionen untereinander zu verbinden und als Datenbank zur Verfügung zu stellen (Wang et al., 2019, S. 74). Als Datenbank ist "CMap" öffentlich verfügbar und ist dadurch eine der meistgenutzten Datenquellen für transkriptomische Veränderungen, die von rund tausend Molekülen oder Chemikalien induziert werden (Zhao & So, 2019, S. 234). Das Projekt ist heute auch unter dem Namen "CLUE" bekannt und ist unter clue.io (2022) abrufbar.

Da diese Vorgehensweise auf den Expressionsprofilen von Krankheiten und Medikamenten basiert, kann sie im Rahmen eines "Repurposing" Vorhabens trotz wenig vorhandenem Vorwissen, eine unvoreingenommene Sicht auf die bekannten betroffenen Gene und Genome liefern und damit unbeabsichtigte Nebenwirkungen als Ergebnisse von subjektiv-geprägten Daten verhindern. Wenn ein Medikament oder eine Krankheit aber keine starken Effekte auf die betroffenen Gene ausübt, entstehen vermehrt verrauschte Expressionsprofile, was zu einer höheren Anzahl "false positives" und "false negatives" bei der Ermittlung von "Repurposing" Kandidaten auf Basis dieser Expressionsprofile führt (Alaimo & Pulvirenti, 2019, S. 102). Ein weiteres teilweise gravierendes Defizit dieser Vorgehensweise liegt in der Methode der "signature reversion". Das Prinzip der "signature

reversion" kann beispielsweise versagen, wenn das erfasste Krankheitsprofil fälschlicherweise als Ursache des Krankheitszustands interpretiert wird, anstelle als mögliches Symptom dieser Krankheit.

In einem Anwendungsfall würde eine Umkehrung eines solchen Profils durch den Einsatz eines Medikaments keinen therapeutischen Effekt haben (Hodos et al., 2016, S. 200).

Transkriptomdaten sind jedoch meist hochdimensional, d.h. die Anzahl Merkmale kann die Anzahl Beobachtungen übersteigen. Generell ist es bei solchen Daten sinnvoll davon auszugehen, dass nur ein Teil der jeweiligen im Expressionsprofil erfassten Gene eine signifikante Bedeutung für die Behandlung der betroffenen Krankheit hat (Zhao & So, 2019, S. 222). Für optimierte Rechenzeiten bietet es sich daher an, die Daten dementsprechend einzuschränken.

### 3.3.4 "similarity-based"

Die "similarity-based" Verfahren nutzen hingegen die Idee, dass wenn zwei verschiedene Krankheiten mindestens ein Medikament für die jeweilige Behandlung gemeinsam haben, vielleicht auch weitere Medikamente, welche bisher nur für die individuelle Behandlung der jeweiligen Krankheit genutzt wurden, eventuell auch für die gemeinsame Behandlung in Frage kommen könnten. Als Weiterführung dieser Grundidee geht die Vorgehensweise auch davon aus, dass sich die Ähnlichkeit zwischen zwei verschiedenen Medikamenten anhand der Übereinstimmungen im chemischen Aufbau, der molekularen Zielstrukturen und Nebenwirkungen bestimmen lässt. Die Ähnlichkeiten von Krankheiten lässt sich dagegen durch Ontologien oder geteilte Behandlungsprofile feststellen (Alaimo & Pulvirenti, 2019, S. 102; Chiang & Butte, 2009, S. 508).

Diese Überlegungen basieren auf dem **"Guilt by Association" (GBA)** Prinzip. Das Prinzip wird durch das Schema von Chiang und Butte (2009) in Abbildung 4 dargestellt:

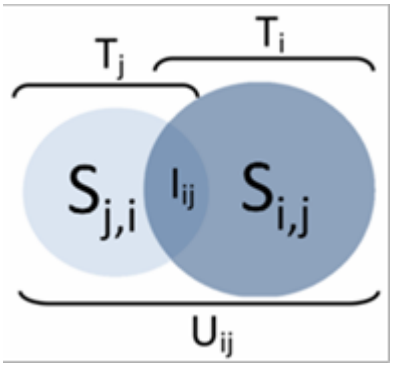


Abbildung 4: "Guilt by Association" Schema für die Entdeckung neuer Anwendungen von Arzneimitteln (Chiang & Butte, 2009, S. 509)

Gegeben seien zwei Krankheiten i und j mit ihren zugehörigen Behandlungsprofilen $T_i$ und $T_j$ mit einer gemeinsamen Menge von Arzneimitteln $I_{ij}$ für die Behandlung von i und j. So stellt $S_{i,j}$ die Menge vorgeschlagener neuer Arzneimittelkandidaten für die Behandlung von i auf der Grundlage des gemeinsamen Behandlungsprofils mit j dar, welche auch durch die Differenz der vereinigten Behandlungsprofile $U_{ij}$ und $T_i$ beschrieben werden kann. Umgekehrt kann die Menge vorgeschlagener neuer Arzneimittelkandidaten für die Behandlung von j ($S_{j,i}$) auf Basis des Behandlungsprofils für i durch die Subtraktion $T_j$ von $U_{ij}$ beschrieben werden (Chiang & Butte, 2009, S. 509).

Unter Anwendung dieses GBA-Prinzips können so bspw. Ähnlichkeiten zwischen Medikamenten anhand der gemeinsam auftretenden Nebenwirkungen bestimmt werden. Obwohl es kontraintuitiv erscheinen mag, Nebenwirkungen als Indikator für Ähnlichkeit zu verwenden, können Nebenwirkungen dennoch als potenzielle phänotypische Biomarker für Krankheitsbehandlungen dienen. Hurle et al. (2013, S. 338) stellten z.B. die Hypothese auf, dass wenn zwei Medikamente die gleiche selten vorkommende Nebenwirkung besitzen, sie auch einen gemeinsamen zugrundeliegenden Wirkmechanismus teilen, welcher die Nebenwirkung mit einer Krankheitsbehandlung verbindet. Solche Assoziationsstrategien basierend auf Nebenwirkungen, welche in Texten zu klinischen Studien festgehalten wurden, können so genutzt werden, um neue potenzielle "Repurposing" Kandidaten vorzuschlagen (Alaimo & Pulvirenti, 2019, S. 107; Hurle et al., 2013, S. 340). Hier soll noch erwähnt werden, dass im direkten Vergleich zu den "side-effect-based" Methoden die Nebenwirkungen nicht genau auf ihre betroffenen Signal- oder Stoffwechselwege untersucht werden, sondern nur einzig allein das Auftreten einer Nebenwirkung als untersuchungsrelevant gilt.

### 3.3.5 Weitere Vorgehensweisen und Methoden

Die vorgestellten Vorgehensweisen können auch erweitert fortgeführt oder miteinander kombiniert werden. Anbei wird eine Auswahl der bedeutendsten alternativen und "hybriden" Vorgehensweisen vorgestellt, welche Aspekte und Daten aus verschiedenen Basis-Vorgehensweisen nutzen.

#### 3.3.5.1 "Pathway- und network-based"

"Pathway-based" Vorgehensweisen nutzen die den "drug targets" zugehörigen Stoffwechselwege, Signalwege und Information aus Proteininteraktionsnetzwerken, um Ähnlichkeiten oder Zusammenhänge zwischen Medikamenten und Krankheiten vorherzusagen (Park, 2019, S. 60). "Network-based" Methoden nutzen die Daten zu gemeinsamen "drug-targets", Signalwegen, Krankheitsphänotypen, Medikamentenindikationen und Nebenwirkungen, um die Verbindungen zwischen Medikamenten und Krankheiten festzustellen (Jin & Wong, 2014, S. 641–642).

Der wichtigste Vorteil dieser beiden Methoden besteht darin, dass sie besonders dafür geeignet sind, allgemeine Signalnetzwerke von einer grossen Anzahl von Proteinen (bzw. "drug-targets") auf ein spezifisches Netzwerk mit wenigen Proteinen einzugrenzen (Jin & Wong, 2014, S. 641–642; Wang et al., 2019, S. 74).

#### 3.3.5.2 "Targeted mechanisms-based"

"Target mechanism-based" Vorgehensweisen integrieren die Daten zu Signalwegen, Proteininteraktionsnetzwerken und "Omics"-Daten, um neue Wirkmechanismen von Medikamenten zu entdecken (Park, 2019, S. 60). "Omics"-Daten sind eine allumfassende Sammlung von Daten aus allen Bereichen der (biologischen) Wissenschaft, welche im Englischen mit dem Kürzel "-omics" enden, wie bspw. Genomik, Transkriptomik, Proteomik, Epigenomik, Mikrobiomik oder Metabolomik (Osier et al., 2017, S. 19). Der Sammelbegriff "Omics" wird aufgrund der damit implizierten vorherrschenden hohen Datenmengen, gerne als Synonym des "Big Data" Begriffes im Bereich der Bioinformatik verwendet (Osier et al., 2017, S. 19).

Ein besonderer Vorteil dieser Vorgehensweise liegt darin, dass diese Methoden meist nicht nur exklusiv darauf abzielen, die Mechanismen im Zusammenhang mit den zu untersuchenden Krankheiten oder Medikamenten zu erfassen. Diese Methoden identifizieren sogar auch die weitgehend unklaren genauen Mechanismen, die im direkten Zusammenhang mit der Ursache oder der Medikamentenbehandlung von bestimmten Krankheiten

stehen (Jin & Wong, 2014, S. 642). So ist beispielsweise die Medikamentenresistenz bei Krebstherapien ein ungelöstes und akutes Problem. Obwohl viele Kranke zu Beginn gut auf ein Medikament ansprechen, entwickeln die Krebszellen oft nach mehreren Behandlungsmonaten eine Resistenz gegen das verwendete Medikament. Dies erfordert zusätzliches Wissen über die Wirkmechanismen von solchen Arzneimitteln, sodass man eventuell alternative "drug-targets" finden und damit dessen Wirkung verbessern könnte. "Targeted mechanisms-based" Vorgehensweisen können solchen Herausforderungen gut entgegenwirken (Jin & Wong, 2014, S. 642).

#### 3.3.5.3 "Knowledge based"

Die "Knowledge based" Strategie nutzt alle Methoden der "target-based", "pathway-based" und "target mechanism based" Vorgehensweisen gleichzeitig (Park, 2019, S. 60). Damit kann auch diese Vorgehensweise vorhandenes Wissen für die Vorhersage von bisher unbekannten Mechanismen nutzen und somit z.B. unbekannte "drug-targets" von Medikamenten, unbekannte Ähnlichkeiten zwischen Medikamenten und neue Biomarker für Krankheiten bestimmen (Jin & Wong, 2014, S. 641).

In einem Forschungsprojekt erstellten Zhu et al. (2020) einen auf dieser Strategie gestützten "Knowledge Graphen" für den Bereich des "Drug Repurposing". Als Datenquelle verwendeten sie sechs verschiedene Datenbanken unterschiedlicher Kategorien nach Tanoli et al. (2021): "PharmGKB" (drug omics, drug-target interaction), "TTD" (disease, pathways, chemical structures), "KEGG" (disease, pathways, drug-target interaction), "DrugBank" (drug-target interaction, clinical information, drug classification, chemical structures, drug combinations), "SIDER" (disease, side effects) und "Drug Indications Database" (unkategorisiert). Ein "Knowledge Graph" ist ein semantisches Netzwerk, welches die Beziehungen zwischen Entitäten aufzeigt (Zhu et al., 2020, S. 2739). Der Hauptzweck eines "Knowledge Graphen" besteht darin eine Konnektivität zwischen den bekannten Fakten in einer vereinten Wissensbasis darzustellen. Durch die Verknüpfung von Wissensfragmenten, die aus verschiedenen Wissensdatenbanken gesammelt wurden, wird ein umfassenderes und zentralisiertes Repositorium erstellt. Anhand der Akkumulationen der erstellten Verbindungen soll daraufhin neues Wissen generiert werden (Zhu et al., 2020, S. 2739).

Der grösste Vorteil von "knowledge-based" Methoden besteht darin, dass sie eine riesige kombinierte Menge von vorhandenem Wissen in den Gesamtprozess des "Drug Repurposing" einbeziehen, um so die Genauigkeit der Vorherhersagen zu verbessern (Jin &

Wong, 2014, S. 641). Zudem können die resultierenden grafischen Visualisierungen von biomedizinischen Netzwerken die Entdeckung neuer potenzieller Verbindungen zwischen entfernten Konzepten unterstützen (Andronis et al., 2011, S. 358).

#### 3.3.5.4 "Chemische Ähnlichkeiten"

Auch die rein chemischen Ähnlichkeiten zwischen Medikamenten können als Vorgehensweise für das "Drug Repurposing" genutzt werden. Die physikalisch-chemischen Eigenschaften von Medikamenten wie chemische Struktur, Schmelzpunkt oder Hydrophobizität werden im Rahmen dieser Vorgehensweise quantifiziert, um Ähnlichkeiten zwischen Medikamenten festzustellen (Hodos et al., 2016, S. 188). Der quantitative Abstand bzw. die Ähnlichkeit zwischen zwei chemischen Strukturen ist leicht mit dem Tanimoto-Koeffizienten (Tc) zu bestimmen, welcher den Jaccard-Index zweier chemischer Fingerabdrücke darstellt (Hodos et al., 2016, S. 188).

Auch die dreidimensionale Geometrie von Atomen und ihre elektronische Struktur können in simulationsbasierten Analysen wie des "Molecular Docking" genutzt werden (Hodos et al., 2016, S. 188).

Weitere chemische Methoden betreten den überschneidenden Bereich der "target-based" Vorgehensweisen, bei deren die Wechselwirkungen zwischen Medikamenten und biologischen Zielen wie bspw. Bindungs- oder andere kinetische Aktivitäten untersucht und quantifiziert werden (Hodos et al., 2016, S. 188; Issa et al., 2021, S. 132).

## 3.4 "Machine Learning" in der Medikamentenforschung und im "Drug Repurposing"

Mit dem schnellen Aufstieg von "Machine Learning" (ML) Technologien in den letzten zehn Jahren, ist auch das Interesse an der Anwendung von ML-Methoden im Bereich der Medikamentenentdeckung/-entwicklung und des "Drug Repurposing" gestiegen (Zhao & So, 2019, S. 219). "Machine Learning", als Teilgebiet der "Künstlichen Intelligenz", umfasst eine Vielzahl von Methoden, mit denen Computer ohne menschliches Eingreifen autonom „lernen“ und Erkenntnisse aus Daten gewinnen können. Viele dieser neuen ML-Technologien spielen eine Schlüsselrolle in der Überwindung der Grenzen der traditionellen Methoden der Arzneientwicklung, indem sie die Möglichkeit bieten, besonders genaue Vorhersagen und Prognosen zu formulieren (Choudhury et al., 2022, S. 2). Diese Vorhersagen und Prognosen basieren meist auf den Ergebnissen und Resultaten von komplexen Analysen, welche mehrdimensionale Daten aus mehreren unterschiedlichen

Datenquellen gleichzeitig verwerten (Yang et al., 2019, S. 10520). Aus diesen Gründen werden ML-Methoden ein besonders grosses Potenzial zugeschrieben, um den gesamtumfassenden Entwicklungs-/Entdeckungsprozess neuer oder auch alternativer Medikamente im Rahmen des "Repurposing" effizienter zu gestalten (Choudhury et al., 2022, S. 13).

Aus diesen Gründen werden sehr viele ML-Methoden in die meisten Aspekte der frühen Phase der Medikamentenentwicklung integriert. Die Anwendungsbereiche umfassen die Identifizierung möglicher "drug-targets", die "Hit to lead"-Erkennung (die Identifizierung von versprechenden chemischen Verbindungen), die Vorhersage der Toxizität und der Pharmakokinetik (umfasst alle Prozesse zur Verträglichkeit, Aufnahme, Verteilung und Umwandlung sowie Ausscheidung des Wirkstoffes) und die Planung und Strukturierung klinischer Studien (Yang et al., 2019, S. 10572). Zudem liefern die Methoden des "Deep Learning" (Teilgebiet des ML) neue Erkenntnisse darüber, wie die Medikamente sich an die Zielstrukturen bzw. Zielmoleküle binden und wie ihre chemischen Eigenschaften und Strukturen mit phänotypischen Veränderungen zusammenhängen (Issa et al., 2021, S. 132–133). Nach Issa et al. (2021, S. 133) spielt besonders die Vorhersage der Bindungsaffinität eines Wirkstoffes für die Priorisierung mehrerer alternativ vorhandener Kandidaten eine grosse Rolle. Dabei wird mithilfe von Simulationen die jeweilige Bindungsstärke zwischen den vielen kleinen Molekülen eines Wirkstoffes und den Zielen von Interesse bzw. "drug targets" errechnet. Die dabei für jede chemische Verbindung ermittelte Bindungsaffinität wird in Form einer Punktzahl dazu verwendet, um chemische Verbindungen für die Priorisierung in eine Rangordnung zu bringen. Dabei gilt folgende allgemeine Regel: Je grösser die vorhergesagte Bindungsaffinität, desto grösser ist auch die Wahrscheinlichkeit einer realen Kleinmolekül- Ziel-Wechselwirkung in klinischen Tests. Dieses Konzept, auch "Molekulares Docking" genannt, ist eine der populärsten chemisch-strukturbasierten Techniken zur Vorhersage und Priorisierung von "drug-target"-Wechselwirkungen.

> *"An important advantage is that ML algorithms are abundant and in rapid development, and any existing or new algorithms can be applied to repositioning without much modification."*
> *(Zhao & So, 2019, S. 220)*

Parallel zur Medikamentenentwicklung können diese Methoden auch im Rahmen des "Drug Repurposing" effektiv genutzt werden, um in kurzer Zeit eine hohe Anzahl vorhandener Arzneimittel auf die Bindungsaffinität alternativ klinisch relevanter Ziele ("drug-targets") zu untersuchen und zu evaluieren (Issa et al., 2021, S. 133–134).

Die meisten der bisher veröffentlichten Studien im Forschungsbereich des ML-gestützten "Drug Repurposing" konzentrieren sich auf die Verwendung von unterschiedlichen Lernalgorithmen, wie z.B. "support vector machine" (SVM) und "random forest" (RF), um sogenannte "supervised predictors" zu entwickeln, welche mithilfe bekannter Assoziationen bzw. Zusammenhängen zwischen Medikamenten, "drug-targets" und Krankheiten zu einem Modell trainiert wurden (Yang et al., 2019, S. 10565). Diese Zusammenhänge werden auf Basis der bekannten Vorgehensweisen aus Kapitel 3.3 identifiziert, wie bspw. anhand Genexpressions-Ähnlichkeiten oder Korrelationen zwischen Medikamentennebenwirkungen. Die Lernalgorithmen des ML können besonders durch die hohe Anzahl verfügbarer Daten zu kleinen chemischen Strukturen, biologischen Prozessen sowie krankheitsbezogenen Phänotypen (genetische Information) profitieren (Yang et al., 2019, S. 10565).

Alle verfügbaren Methoden im ML-Bereich können in die zwei Hauptkategorien "**supervised**" und "**unsupervised**" eingeteilt werden (Zhao & So, 2019, S. 219–220).

### 3.4.1 "Supervised" Methoden im "Drug Repurposing"

"Supervised" ML-Lernverfahren und Methoden stellen Modelle zur Vorhersage oder Schätzung durch einer oder mehreren Eingaben zur Verfügung. Sie werden als überwachte Methoden bezeichnet, da das Lernverfahren durch bereits bekannte Ausgabewerte ("gelabelte Daten") "überwacht" wird (Zhao & So, 2019, S. 220). Speziell im Bereich des "Drug Repurposing" handelt es sich meist um klassische Klassifikationsprobleme, bei denen man mithilfe von ML-Methoden versucht, die Wirkstoffkandidaten in Kategorien bzw. Klassen wie bspw. "tauglich" oder "untauglich" für Krankheit x zu klassifizieren (Zhao & So, 2019, S. 220). Für jede Anwendung eines sogenannten ML-Problems muss im Voraus entschieden werden, welche Merkmale bzw. Art der Daten ausgewählt werden sollen und aus welchen möglichen Ergebnissen bzw. Klassifikationen das ML-Modell aufgebaut werden soll. Dabei sind meistens weitere Vorbereitungsschritte nötig, wie bspw. die Datenvorverarbeitung oder allgemeine Qualitätskontrollverfahren. Sobald der ML-Algorithmus ausgewählt wird, kann dieser auf die Trainingsdaten angewendet und das Modell durch das "hyperparameter tuning" auf Basis eines Validierungssets optimiert

werden. Das finale Modell wird dann auf die Testdaten angewendet und anhand dessen Vorhersageleistung bezüglich der gelabelten Daten unter Aufsicht bewertet (Zhao & So, 2019, S. 220).

#### 3.4.1.1 Lineare ML-Modelle

Lineare ML-Modelle wie bspw. die lineare Regression können in vielen Analysenbereichen einfach und intuitiv angewendet werden. Im Fall der Daten aus biologischen Systemen, sind die in den Daten vorhandenen wahren Zusammenhänge aber vielfach nichtlinear. So wirken verschiedene Gene auf komplexe und nicht auf eine lineare Art, z.B. wie Gene die Wirksamkeit eines Medikamentes beeinflussen. Dennoch werden lineare Modelle als Benchmark für die Entwicklung anspruchsvollerer ML-Modelle angesehen (Zhao & So, 2019, S. 221). Lineare Modelle sind zudem rechnerisch schnell und können leicht in verschiedenen Programmiersprachen oder Statistikprogrammen implementiert werden (Zhao & So, 2019, S. 223).

Als Beispiel entwickelten Buza et al. (2020) mit "MOLIERE" ein modifiziertes lineares Regressions-Modell, welches anhand biomedizinischer Daten "drug-target" Wechselwirkungen akkurat voraussagte und damit traditionelle "state-of-the-art" Methoden übertraf.

#### 3.4.1.2 Entscheidungsbaumverfahren

"Classification and Regression Trees" (CART), auf Deutsch "Entscheidungsbäume" genannt, sind eine weitere wichtige Art der ML-Modelle für die Klassifikation und Regression. Die beiden beliebtesten Arten von baumbasierten Modellen sind "random forest" (RF) von Breiman (2001) und "gradient boosting machines" von Friedman (2001), welche im Allgemeinen einfache "CARTs" in ihrer Leistung übertreffen (Zhao & So, 2019, S. 223).

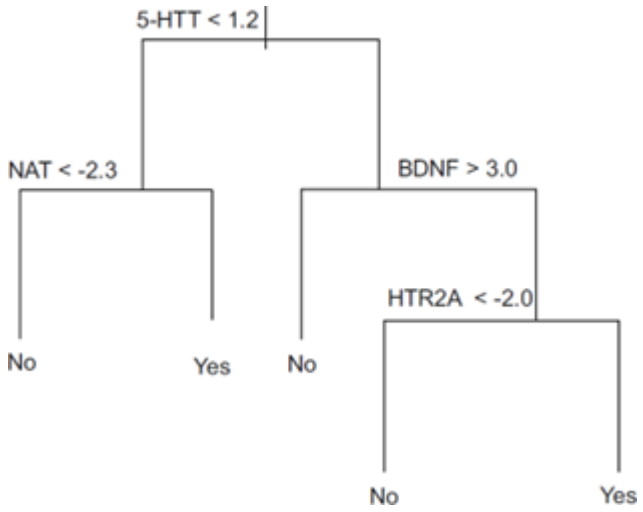


Abbildung 5: Beispiel eines simplen Entscheidungsbaumes zur Klassifikation eines Medikamentes auf Basis von Genexpressionsprofilen (Zhao & So, 2019, S. 224)

Ein einzelner Entscheidungsbaum leidet jedoch unter einer hohen statistischen Varianz der verfügbaren Daten, was zu schlechten Vorhersageleistungen führt. Durch das „Kombinieren“ vieler kleiner Bäume, welche mit verschiedenen Teilmengen von Trainingsdaten trainiert werden, kann die Vorhersageleistung des Gesamtmodells verbessert werden. RF und "gradient boosting machines" nutzen dieses Konzept (Zhao & So, 2019, S. 225). Im Vergleich zu den linearen Modellen sind Entscheidungsbäume zusätzlich auch gegen Ausreisser innerhalb der Daten robust und können komplexe nichtlineare Zusammenhänge gut erfassen (Zhao & So, 2019, S. 226). Als Praxisbeispiel haben Zhao und So (2019, S. 231) einen Random-Forest-Klassifikator modelliert, welcher die präzise Vorhersage ermöglichen soll, ob ein Medikament zur Behandlung von Depression auf Basis der Expressionsprofile der Krankheit verwendet werden kann. In Abbildung 5 wird ein Ausschnitt der verwendeten Entscheidungsbäume als Beispiel dargestellt.

#### 3.4.1.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) ist ein "Maximum-Margin-Classifier" der versucht, eine Menge von Objekten in verschiedene Klassen mit einem grossen Abstand zwischen diesen Klassen aufzuteilen (Cortes & Vapnik, 1995, S. 290). Mithilfe der Verwendung des sogenannten „Kernel-Tricks“ kann die SVM den Vektorraum von niedrigen Dimensionen in einen höherdimensionalen Raum überführen, wodurch Probleme, die in niedrigen Dimensionen nicht gelöst werden können, lösbar gemacht werden (Zhao & So, 2019, S. 226). SVM sind gegen Ausreisser innerhalb der Daten sehr robust und können auch nichtlineare Zusammenhänge gut modellieren (Zhao & So, 2019, S. 227).

Im Rahmen des "Drug Repurposing" sind SVM in der Lage, Daten zu chemisch molekularen Strukturen, Medikamentenwechselwirkungen und Expressionsprofilen ohne Probleme verarbeiten (Zhao & So, 2019, S. 227). Zum Beispiel entwickelten Napolitano et al. (2013) einen mehrklassigen SVM-Klassifikator, welcher die therapeutischen Klassen von Arzneimitteln hervorsagen konnte. Dieser Klassifikator wurde zuvor auf Basis von Daten zu Genexpressionen, chemischen Strukturen und "drug-targets" trainiert.

Da SVM-Modelle aber meist sehr schwierig zu interpretieren und die Bedeutungen einzelner Teilmerkmale schlecht quantifizierbar sind, stellt dies für das "Drug Repurposing" eine Einschränkung dar. Dies, weil ein Hauptinteresse bei "Drug Repurposing" Vorgehen hauptsächlich darin liegt festzustellen, welche genetischen oder biologischen Faktoren zu den resultierenden Behandlungseffekten eines Medikamentes beitragen. Der Einblick in diese Faktoren wird bei SVM-Modellen erschwert (Zhao & So, 2019, S. 227).

#### 3.4.1.4 "Supervised" Deep Learning mit Deep Neural Networks (DNN)

Auch "Deep Learning" als Teilgebiet des ML wird zunehmend im Bereich der Bioinformatik eingesetzt (Zhao & So, 2019, S. 228). "Deep Neural Networks" (DNN), auf Deutsch tiefe neuronale Netze genannt, können komplexe Zusammenhänge durch die Verwendung mehrerer "hidden layers" verarbeiten. In der aktuellen Forschung werden sie vorwiegend für die Vorhersage verschiedener Medikamenteneigenschaften, wie der Wirksamkeit, des Anwendungsfeldes oder der Toxizität verwendet, nachdem die DNN zuvor anhand Transkriptomdaten trainiert wurden (Aliper et al., 2016, S. 2525).

Die Architektur und Gestaltung eines DNN mit einer geeigneten Abstraktion für den Anwendungsfall ist dagegen mit einem hohen Aufwand verbunden. Zudem erfordern DNN in der Regel relativ grosse Datenmengen, um überhaupt gute Vorhersagen treffen zu können. Der Grund dafür ist die Schwierigkeit (bzw. Problem), dass die Anzahl Parameter bzw. "layers" in einem DNN meistens sehr hoch ist und zusätzlich "Overfitting" besonders schnell eintreten kann (Zhao & So, 2019, S. 228).

In einem Praxisbeispiel haben Preuer et al. (2018) ein DNN entwickelt, um die Vorhersage von synergetischen Wirkungseffekte zwischen Medikamenten in der Krebsbehandlung zu ermöglichen. Damit konnten unterschiedliche Kombinationstherapien vorgeschlagen werden.

### 3.4.2 "Unsupervised" Methoden im "Drug Repurposing"

Andererseits können auch die "unsupervised" Methoden des ML verwendet werden, um Zusammenhänge oder Muster zwischen Medikamenten oder Krankheiten zu identifizieren. Diese Methoden nutzen im Gegensatz zu den "supervised"-Methoden keine „gelabelten“ Daten (Zhao & So, 2019, S. 220). "Unsupervised" Methoden haben im Bereich des "Drug Repurposing" einige Vorteile, da sie durch den Verzicht von "gelabelten" Daten eventuell verborgene Muster und Zusammenhänge in Daten erkennen können (Yang et al., 2019, S. 10566). Hier werden parallel zum Bereich der "supervised" Methoden besonders Cluster und Klassifikationsalgorithmen verwendet, um Ähnlichkeiten zwischen Medikamenten und Krankheiten festzustellen (Yang et al., 2019, S. 10566).

In einem Projekt mit dem Namen "MANTRA" von Carrella et al. (2014) wurde so zum Beispiel mithilfe "unsupervised" ML-Methoden auf Basis von Daten aus der "Connectivity Map" (CMap) ein Medikamentennetzwerk generiert, bei dem alle Medikamente als Knotenpunkte und alle signifikanten Verbindungen zwischen Medikamenten als gewichtete Kanten dargestellt wurden. Dieses Netzwerk wurde anschliessend durch einen Clustering-Algorithmus durchsucht, um Gruppen von ähnlichen bzw. "verwandten" Medikamenten anhand ihrer Expressionsprofile zu bestimmen. Das Ziel war Gruppen mit gemeinsamen Wechselwirkungen oder gemeinsamen "drug-targets" zu identifizieren.

Mit "DrugNet" von Martínez et al. (2015) wurde ein ähnliches Netzwerk konstruiert, welches wiederum auf Daten zu "drug-targets" bzw. Zielstrukturen basierte. Auch mit diesem Netzwerk war es möglich, Ähnlichkeiten zwischen Krankheiten und Medikamenten zu errechnen. Zusätzlich konnten mit "DrugNet" neue potenzielle "Repurposing" Kandidaten identifiziert werden, indem mögliche Zusammenhänge zwischen Krankheiten und Medikamenten vorhergesagt wurden. Das Netzwerk war zusätzlich sogar in der Lage, vollkommen neue Zusammenhänge zwischen Entitäten zu bestimmen, welche bisher noch nicht in der Literatur oder in klinischen Studien dokumentiert waren (Martínez et al., 2015, S. 48).

Die Genauigkeiten von "unsupervised" ML-Modellen sind im Direktvergleich zu den "supervised" ML-Modellen in der Regel eher schlecht (Yang et al., 2019, S. 10566). Dies meist aufgrund der fehlenden Möglichkeit, eine Kontrolle mithilfe "gelabelter" Daten über die Medikamente und "drug-targets" durchzuführen. Gleichzeitig sind "unsupervised" ML-Modelle auf höhere heterogene Datenmengen angewiesen, um diese für Vorhersagen ausreichend balanciert trainieren zu können (Yang et al., 2019, S. 10566). Nichtdestotrotz

haben die "unsupervised" Methoden aus genau diesen gleichen Gründen das Potenzial, vollkommen neue Medikament-Krankheit Assoziationen zu entdecken, welche nach dem aktuellen Verständnis in der Pharmakologie noch unbekannt sind. Dies wird dadurch ermöglicht, da sie die chemischen und biochemischen Daten aus einer "unbiased" d.h. unvoreingenommenen Perspektive analysieren (Yang et al., 2019, S. 10566–10567).

### 3.4.3 Moderne ML-Methoden im "Drug Repurposing" und Limitationen

Die grosse Mehrheit der in den oberen Abschnitten beschriebenen ML-Modelle und vorgestellten Praxisbeispielen beruhte auf Ähnlichkeiten, d.h. die Modelle formulierten ihre Vorhersagen über die Anpassung von Koeffizienten, welche sie zuvor bezüglich Überschneidungen/Ähnlichkeiten der "drug-targets", der Expressionsprofile oder Wechselwirkungen von Krankheiten und Medikamenten errechnet hatten (Hodos et al., 2016, S. 195). Jedoch gibt es immer mehr ML-Projekte wie bspw. das zuvor erwähnte "MANTRA", bei welchen ähnlichkeitsbasierte Netzwerke gebaut werden und dabei Ähnlichkeitsmatrizen zwischen Arzneimitteln und Krankheiten verwendet werden (Hodos et al., 2016, S. 195).

Im Allgemeinen sieht man in den ML-Methoden das Potenzial, die übergreifenden Prozesse der Medikamentenentdeckung/-entwicklung und des "Drug Repurposing" effizienter zu gestalten (Choudhury et al., 2022, S. 13). Obwohl vorwiegend klassische ML-Methoden aufgrund begrenzter Datensatzmengen verwendet werden, erweisen sich moderne ML-Methoden als disruptive Technologien, indem sie mit ihren Methoden neue Arten und Möglichkeiten zeigen, wie verschiedene traditionelle Teilaufgaben in der Medikamentenentwicklung alternativ durchgeführt werden können (Choudhury et al., 2022, S. 13). Aktuell können ML-Methoden die Forschung umfassend dabei unterstützen, neue "drug-targets" zu identifizieren, 3D-Strukturen von Zielproteinen aus der Sequenz vorherzusagen, das Screening grosser Mengen kleiner wirkstoffähnlicher Moleküle zu vereinfachen, neue Liganden durch generative Algorithmen vorzuschlagen, retrosynthetische Wege für die Synthese zu empfehlen, Robotersysteme zur Synthetisierung von Verbindungen physikalisch zu steuern und Resultate von klinischen Studien zu prognostizieren (Choudhury et al., 2022, S. 13).

Somit sind ML-Methoden besonders für das "Drug Repurposing" sehr wertvoll. Aufgrund der hohen Kosten und den hohen Risiken, welche mit der Entwicklung neuer Medikamente verbunden sind, besitzt die Entdeckung neuer Indikationen für bestehende Medikamente ein deutlich besseres Risiko-Ertrags-Verhältnis. Gleichzeitig bieten sie die

Möglichkeit, neue Behandlungen für bisher vorwiegend aus kommerziellen Gründen vernachlässigte Behandlungsgruppen zu entdecken (Yang et al., 2019, S. 10573).

Die grösste bestehende Herausforderung und Hürde der ML-Methoden bleibt aber nach wie vor die Beschaffung genügender, qualitativ hochwertiger und problemspezifischer Daten (Yang et al., 2019, S. 10572–10573).

## 3.5 Das Potenzial unstrukturierter Textdaten und verfügbare Analysemethoden

Neben den Datenbanken verbirgt sich eine riesige Menge an medizinischem Wissen in verschiedenen Formen von unstrukturierten Textdaten wie bspw. in klinischen Berichten, wissenschaftlichen Forschungsdokumenten oder Fachzeitschriften (Andronis et al., 2011, S. 364). Allerdings liegt in diesen Textdaten das Wissen über Arzneimittel meist sehr spezifisch oder fallspezifisch vor, z.B. in Form von Studien über den Einsatz eines bestimmten Arzneimittels bei ausgewählten Personen zur Behandlung einer spezifischen Krankheit oder Kondition. Durch die Nutzung des GBA-Prinzips und von anderen verfügbaren Methoden des "Natural Language Processing" (NLP) können diese Textdaten für Analysezwecke dennoch ohne Probleme verwendet werden (Alaimo & Pulvirenti, 2019, S. 102).

### 3.5.1 NLP in der Biomedizin

Der technologische Fortschritt im Bereich des "Natural Language Processing" (NLP) ermöglichte Forscherinnen und Forschern im Bereich der Biomedizin, unstrukturierte Daten in Form von Textdaten wie z.B. Kliniknotizen, wissenschaftliche Publikationen sowie auch Webforum-Beiträge, in ein maschinenlesbares Format zu übersetzen, welches anschliessend mit ML und Deep-Learning-Methoden genutzt werden kann (Issa et al., 2021, S. 137).

Gekoppelt mit den Fortschritten im Bereich des "Machine Learning" hat die Anzahl verfügbarer Methoden für die Ermittlung und Vorhersage von potenziellen "Drug Repurposing" Kandidaten stark zugenommen, wobei sich auch die Methoden in ihren Genauigkeiten zu präzisen Vorhersagen deutlich verbessert haben (Mayers et al., 2019, S. 2). Mit dem auch immer weiter steigenden Fokus von neuen Forschungsprojekten auf "knowledge-based" Vorgehensweisen, welche sich als Ziel setzen, die "drug-drug" und "disease-disease" Ähnlichkeiten mit den "drug-disease" Assoziationen bzw.

Zusammenhängen zu kombinieren, kann eine grosse Menge an qualitativ hochwertigem Wissen in maschinelle Lernprozesse und deren Modelle integriert werden (Cheng et al., 2012, S. 10; Mayers et al., 2019, S. 2). NLP-Methoden ermöglichen es, solche Wissensnetzwerke anhand von Assoziationen aus unstrukturierten Textdaten zu generieren und nehmen damit eine wichtige Rolle in der zukünftigen Entwicklung von neuen maschinellen Vorhersagesystemen von "Drug Repurposing" Kandidaten ein (Mayers et al., 2019, S. 2).

Das allgemein übergeordnete Ziel, neues Wissen in Textdaten zu entdecken, wird durch die zahlreichen NLP-Verfahren auf lexikalischer oder grammatikalischer Ebene unterstützt. Diese reichen von eher simplen Verfahren, wie bspw. der Tokenisierung oder des Part-of-Speech-Tagging, bis hin zu komplexen Informationsextraktionsalgorithmen wie die **"Named Entity Recognition" (NER)**. Diese ermöglicht es, biomedizinische Konzepte bzw. Entitäten wie Gene, chemische Verbindungen, Medikamente und Krankheiten in Textdaten zu identifizieren sowie die zuvor erwähnten Assoziationen und Zusammenhänge zwischen diesen Konzepten zu erfassen (Gonzalez et al., 2016, S. 34). Neben solchen Assoziationen und Zusammenhängen können die Techniken des Text-Mining auch neue Muster oder Trends in Textdaten erkennen (Gonzalez et al., 2016, S. 34). Weitere komplexere Methoden umfassen das "Term-Matching", die Sentiment-Analyse zur Stimmungserkennung von Texten sowie erweiterte Beziehungsextraktionsalgorithmen, wie bspw. das System von Song et al. (2018) zur Extraktion von Assoziationen auf Basis von Literaturverweisen (Gonzalez et al., 2016, S. 34). Als weiteres Praxisbeispiel wurde in einem Forschungsprojekt von Li et al. (2018) auf Basis von klinischen Textdaten und NLP-Methoden ein "Deep-Learning" Modell trainiert, welches bei Anwendung "state-of-the-art" Genauigkeiten bei der Identifizierung von Nebenwirkungen in Textdaten verzeichnen konnte (Li et al., 2018, S. 10).

Da NLP- und Deep-Learning-Algorithmen durch die stetig wachsenden Datenmengen immer ausgefeilter werden, wird dieser Forschungsbereich zusätzlich durch die zunehmende Beteiligung grosser privater Tech-Unternehmen wie Google oder Amazon im Bereich des biomedizinischen "Machine Learning" begünstigt. Beide Unternehmen bieten schon eigene Deep-Learning-Plattformen mit "Google AutoML" und "Amazon Comprehend Medical" als kommerzielle Lösungen mit biomedizinischen NLP-Methoden an (Issa et al., 2021, S. 137). Durch die mittels der Tech-Unternehmen ermöglichte Nutzung von hohen Rechenressourcen im Bereich des "Supercomputing", wird die Entwicklung

neuer möglicher NLP- und Deep-Learning-Methoden für das "Drug-Repurposing" ohne Zweifel beschleunigt (Issa et al., 2021, S. 137).

#### 3.5.1.1 Biomedizinische "Named Entitiy Recognition" (NER)

NER bildet das Herzstück und der Grundbaustein der automatisierten maschinellen Extraktion von Wissen aus Textdaten und befasst sich mit der Aufgabe eindeutige Verweise auf Konzepte bzw. Entitäten wie Gene, Medikamente und Krankheiten zu finden und diese mit ihrem Ort (des Vorkommens) und Typ zu kennzeichnen (Gonzalez et al., 2016, S. 35; Śniegula et al., 2019, S. 261). Alternativ wird NER auch als "Entity Tagging" oder "Concept Extraction" bezeichnet (Gonzalez et al., 2016, S. 35). Speziell im biomedizinischen Forschungsbereich stellt NER eine wichtige Hilfsstütze dar, um mit dem stetigen Wachstum neuer entdeckter und definierter Konzepte und Entitäten aus der Fachliteratur, wie bspw. experimentellen Wirkstoffen oder Arzneimitteln, mithalten zu können (Gao et al., 2021, S. 1; Śniegula et al., 2019, S. 264).

Im biomedizinischen Bereich gilt NER im Allgemeinen als schwieriger als in anderen Bereichen wie z.B. Medienberichten (Gonzalez et al., 2016, S. 35). Dies liegt an den vorhandenen Inkonsistenzen zu den Bezeichnungen der bekannten Entitäten, welche sich anhand fehlender standardisierter Abkürzungen oder der Existenz mehrerer Bezeichnungen bzw. Synonyme für eine Entität ergeben (Gonzalez et al., 2016, S. 35). Als explizites Beispiel dieser Problematik besitzt bspw. das Protein bzw. Gen-Produkt "CXCR4" nach Quelle der US-amerikanischen "National Library of Medicine" (NLM) insgesamt 17 alternative Abkürzungen und Aliase (ncbi.nlm.nih.gov, 2022a): FB22, HM89, LAP3, LCR1, NPYR, WHIM, CD184, LAP-3, LESTR, NPY3R, NPYRL, WHIMS, HSY3RR, NPYY3R, WHIMS1 und D2S201E.

Wie Gao et al. (2021, S. 2) daher betonen, führt dieses Problem dazu, dass die Verallgemeinerbarkeit und die Genauigkeiten von NER-Verfahren sehr stark von den verwendeten Modellen und der Menge der verfügbaren "gelabelten" Textdaten innerhalb der sogenannten Text-"Korpussen" abhängen. In der biomedizinischen NER sind solche annotierten Daten je nach ausgewähltem Korpus, oft nur auf eine bestimmte Art von Entität wie Medikamenten, chemischen Substanzen, Krankheiten oder Gene beschränkt. Infolgedessen können vorhandene NER-Werkzeuge in ihrem technischen oder fachspezifischen Anwendungsbereich und Umfang eingeschränkt sein, indem sie nur eine sehr begrenzte Menge von Entitätstypen identifizieren können.

Die Entwicklung effektiver biomedizinischer NER-Systeme gestaltet sich daher meist als herausfordernd, weil solche annotierte Trainingsdaten als Korpusse nur in sehr begrenzten Mengen und Formen verfügbar sind. Auch die Generierung eigener bzw. neuer biomedizinischer Annotationen stellt grundsätzlich keine Alternative dar, da teures Expertenwissen erforderlich ist, um einen akzeptablen "Goldstandard" zu erreichen (Gao et al., 2021, S. 2). NER-Systeme, welche auf solchen annotierten Text-Korpusse oder anderem vordefiniertem Wissen aufgebaut werden, werden als "rule-based" Systeme bezeichnet. Die Mehrheit der existierenden NER-Systeme sind jedoch hybride Systeme und kombinieren dabei jeweils ML- und "rule-based"-Ansätze (Pradhan et al., 2015, S. 151). Zusätzlich erfordert der Einsatz von "rule-based" Methoden meist trotzdem die Erstellung von weiteren spezifischen Regeln durch Experten, da neben der Verwendung der Korpusse, zusätzliche Aufgaben wie bspw. das "Term-Matching" implementiert werden müssen (Gao et al., 2021, S. 1). Bei "Term-Matching" handelt es sich um den Abgleich der in Textdaten erkannten Entitäten mit einer bestehenden Konzeptdatenbank. Dabei werden alle einzelnen Entitäten eindeutig und standardisiert einem Konzept zugeordnet. Die bedeutendste biomedizinische Konzeptdatenbank ist "Unified Medical Language System" (UMLS) von Bodenreider (2004), welche von der US-amerikanischen "National Library of Medicine" zur Verfügung gestellt wird. UMLS umfasst mehr als eine Million Konzepte und mehr als 4 Millionen Konzeptnamen, welche die Beziehungen zwischen diesen Konzepten beschreiben. Obwohl die Konzepte aus verschiedenen Datenquellen mit verschiedenen Vokabularen und Terminologien stammen, stehen diese in UMLS standardisiert zur Verfügung (Andronis et al., 2011, S. 361– 362). MeSH (Medical Subject Headings) stellt das kontrollierte Vokabular von UMLS als Thesaurus dar und wird in den meisten biomedizinischen NER-Systemen für das standardisierte "Term-Matching" bzw. "Konzept-Matching" von Entitäten verwendet (ncbi.nlm.nih.gov, 2022b).

#### 3.5.1.2 Herausforderungen biomedizinischer NLP-Methoden

Eine übergreifende Herausforderung im Bereich des biomedizinischen "Literature Mining" besteht darin, so viele zur Verfügung stehende Wissensressourcen und Wissensnetzwerke wie möglich in eine NLP-Pipeline zu integrieren (Gonzalez et al., 2016, S. 35). Viele der für die Forschung relevanten Daten bspw. aus klinischen Berichten sind begrenzt, vorwiegend aufgrund der Einschränkungen bezüglich der Privatsphäre und Vertraulichkeit der betroffenen Patientinnen und Patienten (Śniegula et al., 2019, S. 261).

Śniegula et al. (2019, S. 261) weisen zusätzlich darauf hin, dass biomedizinische Texte meist in einem ungewöhnlichen Schreibstil und ungewöhnlicher Sprache verfasst

werden. So werden unverhältnismässig viele unvollständige Sätze mit Sonderzeichen, Bindestrichen, Abkürzungen und Akronymen benutzt, welche zudem von Rechtschreibfehlern übersät sind. Wie oben erwähnt, wächst darüber hinaus das biomedizinische Wissen und die damit verbundene Anzahl biomedizinischer Konzepte durch die hohe Anzahl durchgeführter Forschungen und klinischer Studien stetig. Dies macht es für die NER-Systeme besonders schwierig, ihre "rule-based" Korpusse und damit ihre Funktionalität auf dem neuesten Stand zu halten (Gao et al., 2021, S. 2; Śniegula et al., 2019, S. 261). Gleichzeitig stehen die meisten NER-Systeme nur in der englischen und chinesischen Sprache zur Verfügung. Für andere Sprachen fehlen annotierte Korpusse sowie die technologisch nötigen Anpassungen für die korrekte maschinelle Verarbeitung der Grammatiken (Śniegula et al., 2019, S. 261).

Mit GERNERMED von Frei und Kramer (2021) wurde erstmals ein öffentliches "rule-based" und "Deep-Learning"-gestütztes NLP-System in deutscher Sprache mit NER-Modellen zur Identifizierung biomedizinischer Entitäten in klinischen Textdaten veröffentlicht. Dieses neue NLP-System konnte auf separaten Testdaten einen übergreifenden F1-Score von

81.54 verzeichnen (Frei & Kramer, 2021, S. 8). Im Direktvergleich mit dem englischsprachigen NER-Modell "SciBERT" auf Basis eines ähnlich spezialisierten Korpus "BC5CDR", verzeichnete "SciBERT" einen F1-Score von 90.01, was den sprach- und damit datenabhängigen Qualitätsunterschied zwischen den NLP-Systemen und den darin verfügbaren NER-Modellen verdeutlicht (Beltagy et al., 2019, S. 3616).

#### 3.5.1.3 Weitere Potenziale von Textdaten

Trotz der bemerkenswerten Fortschritte bei der Extraktion von Entitäten aus Textdaten, beschränkt sich die Verwendung dieser Methoden derzeit hauptsächlich darauf, die Forscherinnen und Forscher in ihren Entscheidungsprozessen zu unterstützen. Dabei könnten die Methoden des Text-Mining auch direkt vermehrt für das "Drug Repurposing" und der Entdeckung vollkommen neuer Wirkstoffe eingesetzt werden (Gonzalez et al., 2016, S. 37).

Besonders das Gebiet der Pharmakogenomik, welches sich mit den Einflüssen von Erbanlagen auf die Wirkung von Arzneimitteln beschäftigt, profitierte erheblich von den aktuellen Fortschritten von Text-Mining in der Biomedizin. Dabei konnten NLP-Methoden auf Basis von klinischen Textdaten die genetische Grundlage individueller Arzneimittelreaktionen vorhersagen, indem sie die Assoziationen bzw. Zusammenhänge zwischen

Arzneimitteln, Genen und Krankheiten erfassten und analysierten (Gonzalez et al., 2016, S. 37).

Ein weiterer Nutzungsbedarf an den Methoden des Text-Mining lässt sich auch besonders im Gebiet der Toxikologie feststellen, wo durch den Einsatz von NLP-Methoden Vorhersagen zu den chemisch-biologischen Wechselwirkungen von Chemikalien bzw. Medikamenten ermöglicht werden. Durch diese Vorhersagen können hohe Toxizitäten von Medikamenten frühzeitig erkannt und somit Misserfolge in klinischen Studien verhindert werden (Gonzalez et al., 2016, S. 38).

### 3.5.2 "ABC-Modell": Ermittlung von biomedizinischen Assoziationen in Textdaten

Wie in Kapitel 3.5.1 beschrieben, entwickeln sich die verfügbaren NLP-Methoden stetig weiter und sind aktuell schon in der Lage, Beziehungen zwischen biomedizinischen Konzepten und Entitäten in unstrukturierten Textdaten zu erfassen. Dies ermöglicht die Generierung von Assoziationsnetzwerken zwischen biomedizinischen Entitäten wie Genen, Symptomen, Medikamenten und Krankheiten (Andronis et al., 2011, S. 364; Yang et al., 2017, S. 496). Ein bekanntes Modell, um Beziehungen zwischen Entitäten festzustellen, ist das **"ABC-Modell"** von Swanson (1986):

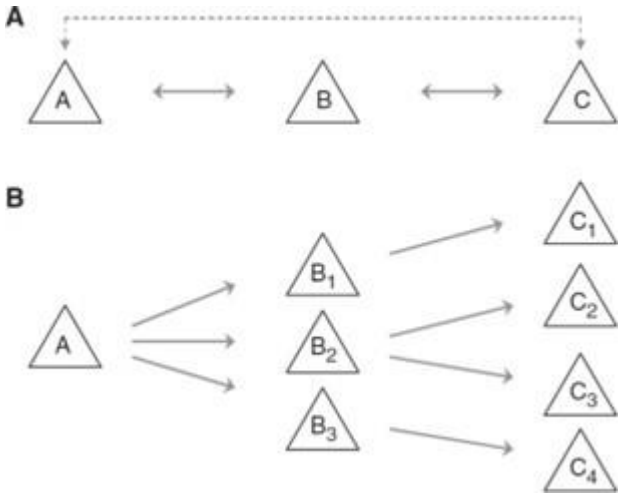


Abbildung 6: "Swanson's ABC Modell" – A zeigt das geschlossene, B zeigt das offene Entdeckungsmodell (Andronis et al., 2011, S. 359)

Das "ABC-Modell" basiert auf dem Konzept der Transitiven Relation. Im Anwendungsbereich des "Drug Repurposing" Vorhabens kann das Modell folgendermassen erläutert werden:

Angenommen, es ist anhand einer Datenquelle bzw. Datenbank bekannt, dass eine Krankheit C ein bestimmtes Argument B hat (wie z.B. Krankheit C wird durch eine Downregulierung des Gens B verursacht). Gleichzeitig ist bekannt, dass das Medikament A eine gewisse Wirkung auf B hat (wie z.B. Medikament A kann die Genexpression von Gen B wiederherstellen). Daraus lässt sich mit einer transitiven Schlussfolgerung ableiten, dass A einen möglichen Einfluss auf C hat. Damit stellt das Medikament A ein potenzieller "Repurposing" Kandidat für die Behandlung von Krankheit C dar. Hierzu bietet die Anzahl vorhandener Beziehungen zwischen den jeweils ausgewählten Konzepten A, B und C ein natürlich messbares Indiz, wie stark die implizite Beziehung zwischen A und C sein kann (Alaimo & Pulvirenti, 2019, S. 107; Andronis et al., 2011, S. 358).

Wie in Abbildung 6 dargestellt, wird zwischen dem offenen und dem geschlossenen ABC-Entdeckungsmodell unterschieden. Beim geschlossenen Entdeckungsmodell bzw. Entdeckungsverfahren sind die Ausgangskonzepte (z.B. Entitäten) A & C bereits bekannt und die implizite Beziehung zwischen A & C soll als Hauptziel anhand der jeweiligen expliziten Beziehungen zum Konzept B genau untersucht und interpretiert werden. Dies ist das ursprünglich konzipierte Modell von Swanson (1986). Hingegen ist beim offenen Entdeckungsmodell bzw. Entdeckungsverfahren nur das Konzept A (z.B. eine Krankheit) als Startpunkt bekannt. Nun wird nach Argumenten B gesucht, welche mit Konzept A zusammenhängen (z.B. Krankheitssymptom). In einem zweiten Schritt werden Konzepte C (z.B. Medikamente) gesucht, welche einen direkten Zusammenhang mit den Argumenten B besitzen (wie z.B. Bekämpfung dieser Krankheitssymptome) (Andronis et al., 2011, S. 359).

Das offene Entdeckungsmodell und der Aufbau von Assoziationsketten bieten damit eine Möglichkeit, mithilfe von NLP-Methoden "Repurposing" Kandidaten in Textdaten zu ermitteln.

# 4 Forschungsleitende Fragestellung & formulierte Unterfragen

Im Rahmen des Literature Reviews wurde das Thema "Drug Repurposing" mit besonderem Fokus auf die "in silico" bzw. "computational methods" vorgestellt. Dabei wurden die Fähigkeiten der verfügbaren Werkzeuge des "Natural Language Processing" zur Extraktion von Zusammenhängen bzw. Assoziationen aus unstrukturierten Textdaten präsentiert.

Bezogen auf von Tanoli et al. (2021) beschriebenen Herausforderung der „richtigen“ Datenbankauswahl und dem in Kapitel 3.5 geschilderten grossen Potenzial unstrukturierter Textdaten als Wissensbestände, sollen als Forschungsziel neue "Best-Practice" Workflows als Methoden für die Ermittlung neuer "Drug Repurposing" Kandidaten entwickelt, getestet und verglichen werden. Die Workflows sollen dabei feststellen, welches und wieviel Wissen in unstrukturierten biomedizinischen Textdaten vorhanden ist und wie dieses Wissen als Ergänzung von Datenbanken extrahiert und genutzt werden kann. Daraus ergibt sich folgende forschungsleitende Frage:

*"Wie können unstrukturierte Textdaten für die Ermittlung neuer "Drug Repurposing" Kandidaten nutzbar gemacht werden und wie können sie Datenbanken ergänzen?"*

Die Workflows sollen auf den vorgestellten Vorgehensweisen, Prinzipen und theoretischen Modellen basieren und unterschiedlich aufgebaut sein. Damit sollen die Workflows die jeweilig unterschiedlichen Vor- und Nachteile der theoretischen Modelle, Vorgehensweisen und verwendeten Konzepte vergleichbar machen. Gleichzeit sollen die benutzten Datenbanken und Textdatenquellen begründet und die Methoden des "Natural Language Processing" zur Analyse und Verarbeitung der unstrukturierten Textdaten verwendet werden. Mit dem gemeinsamen Ziel neue "Repurposing“ Kandidaten zu identifizieren, sollen sie im Rahmen eines Anwendungsfalls getestet werden. Dazu soll eine Krankheit als fixierter Startpunkt für beide Workflows ausgewählt werden.

Die abschliessende Evaluation der Leistung bzw. Zuverlässigkeit der unterschiedlichen Methoden soll anhand einer nachträglichen Kontrolle der ermittelten "Repurposing" Kandidaten (als Krankheit-Medikament Paare) durchgeführt werden. Dabei sollen die ermittelten Kandidaten mit dem schon vorhandenen Wissen in spezialisierten Medikamentendatenbanken (wie z.B. go.drugbank.com oder drugcentral.org) abgeglichen werden. So kann zum Beispiel ein "richtiger" Kandidat als Medikament, durch die Existenz einer schon durchgeführten oder bekannten klinischen Studie für die ausgewählte Krankheit bestätigt werden. Im Kapitel 5 wird das Forschungsdesign detaillierter erläutert.

## 4.1 Grundeigenschaften der Workflows & generierte Unterfragen

Für das formulierte Forschungsvorhaben sollen grundsätzlich zwei unterschiedliche Grundarten von Workflows entwickelt werden. Die erste Art Workflow, im weiteren Rahmen dieser Arbeit "Methode 1" genannt, soll sich mit der Unterfrage befassen:

> **Unterfrage 1:** Wie können "Repurposing" Kandidaten aus unstrukturierten Textdaten ohne die Verwendung von Vorwissen aus Datenbanken zu Medikamenten oder Krankheiten ermittelt werden?

Mit Unterfrage 1 soll das alleinige Potenzial von unstrukturierten biomedizinischen Textdaten als Wissensbestände untersucht werden, indem in einem ungerichteten Ansatz nur die verfügbaren Werkzeuge des biomedizinischen NLP genutzt werden, um neue "Repurposing" Kandidaten zu ermitteln. Dabei soll auf die Verwendung von vorhandenem spezifischem Vorwissen zu den zu untersuchenden Krankheiten aus Datenbanken verzichtet werden.

Die zweite Art Workflow, "Methode 2" genannt, soll sich mit folgender Unterfrage 2 befassen:

> **Unterfrage 2:** Wie können "Repurposing" Kandidaten aus unstrukturierten Textdaten unter Verwendung von Vorwissen aus Datenbanken zu Medikamenten oder Krankheiten ermittelt werden?

Mit der Unterfrage 2 soll das Potenzial von unstrukturierten biomedizinischen Textdaten als ergänzende Wissensbestände untersucht werden. Dabei soll in einem gerichteten Ansatz, neben den verfügbaren Werkzeugen des biomedizinischen NLP, zusätzlich Vorwissen aus den "besten" verfügbaren Datenbanken eingesetzt werden, um neue "Repurposing" Kandidaten in den Textdaten zu bestimmen.

Bei Methode 1 sowie bei Methode 2 steht dabei das gleiche biomedizinische Wissen zur Verfügung, welches in das ausgewählte NER-System integriert wurde.

## 4.2 Zusätzliche Fragestellungen

Im Rahmen der praktischen Untersuchung der Unterfragen wurden weitere zusätzliche Fragestellungen in Bezug zu den erstellten Workflows und dem vorhandenen theoretischen Wissen aus der Forschung formuliert.

Alaimo und Pulvirenti (2019, S. 102) beschrieben anhand eigener Forschungsresultate, dass mit "Repurposing" Vorgehensweisen basierend auf übereinstimmenden Genen oder Genomen zwar eine hohe Anzahl potenzieller Kandidaten erfasst werden, diese jedoch aber aus vielen "false positive" Treffern bestehen. Dagegen würden Übereinstimmungen von Nebenwirkungen treffsichere, aber eine niedrigere Anzahl Kandidaten bestimmen. Auf Basis dieser Beobachtung sollen die Ergebnisse der Methode 1 in Anbetracht der folgenden Frage evaluiert und untersucht werden:

**ZF 1:** Wie unterscheiden sich die Ergebnisse der Methode 1 abhängig der ausgewählten Entitätstypen für die Kookkurrenz Analyse zur Bestimmung der Ähnlichkeiten zwischen den Dokumenten?

In der Umsetzung des zweiten Workflows mit Methode 2 sollen offene Assoziationsketten auf Basis des "ABC"-Modells von Swanson (1986) gebildet werden, um neue "Repurposing" Kandidaten zu ermitteln (Kapitel 5.5). Bei der Bildung der Assoziationsketten sollen unterschiedliche "A-B" Startpaare in Form von unterschiedlichen Relationstypen verwendet werden. Auf Basis der gleichen Forschungsbeobachtung von Alaimo und Pulvirenti (2019, S. 102) wie bei der **ZF 1,** sollen die potenziellen Unterschiede der Vorgehensweisen anhand der Ergebnisse der Methode 2 evaluiert und untersucht werden.

**ZF 2:** Wie unterscheiden sich die Ergebnisse der verschiedenen Arten der Assoziationsketten?

Abschliessend betonen Yang et al. (2017, S. 489) in ihrer Forschungsarbeit, dass ungerichtete Kookkurrenz Analysen, wie in Methode 1 verwendet, eine hohe Anzahl "false positive" Kandidaten bestimmen würden. Auf Basis dieser Aussage sollen die Ergebnisse der Methoden 1 & 2 verglichen und beobachtete Probleme dokumentiert werden:

**ZF 3:** Wie unterscheiden sich die Ergebnisse der ungerichtete Methode 1 und der gerichteten Methode 2 (mit Vorwissen) und welche Probleme liessen sich beobachten?

# 5 Forschungsmethodik und -design

## 5.1 Grundlagen und Allgemeines

Im Rahmen dieser Forschungsarbeit sollen zwei automatisierte "Drug Repurposing" Workflow-Arten als Methoden entwickelt werden, welche die Analyse von unstrukturierten biomedizinischen Textdaten zur Ermittlung und Ausgabe von "Repurposing" Kandidaten ermöglichen sollen. Für die technische Umsetzung dieser Workflows soll als technische Grundlage, die dem Autor bekannte Programmiersprache "Python" verwendet werden. "Python" ist eine universelle, interpretierte, objektorientierte höhere Programmiersprache und verfügt über eine grosse Auswahl von fortgeschrittenen Datenstrukturen, welche kombiniert mit dynamischer Typisierung eine attraktive Lösung für die schnelle Anwendungsentwicklung bietet. Die einfache und leicht zu erlernende Syntax von "Python" soll die Lesbarkeit von Programmcode erhöhen und dabei den Aufwand der Programmwartung reduzieren. "Python" unterstützt die Verwendung von Modulen und Paketen in Form von Programmierbibliotheken, was auch die allgemeine Wiederverwendung von Programmiercode fördert (van Rossum, 2022b). Aufgrund dieser Vorteile und der persönlichen Präferenzen des Autors wird "Python" als Basis Programmiersprache für die technische Umsetzung ausgewählt. Dazu soll zusätzlich "Anaconda 3" als zugehörige Distribution verwendet werden, deren Fokus auf den Bereich der "Data Science" für die Verarbeitung von grösseren Datenmengen, Vorhersagenanalysen und wissenschaftlichem Rechnen liegt (docs.anaconda.com, 2022). Die Verwendung von Anaconda ermöglicht zusätzlich die Nutzung des Produktes "Jupyter Notebook", welches eine Programmierumgebung kombiniert mit erweiterten Dokumentationsmöglichkeiten bietet (jupyter-notebook.readthedocs.io, 2022). Aufgrund der in Kapitel 5.2.4 nachfolgenden Erläuterungen zur Auswahl der zu verwendenden technischen Werkzeuge wird für Entwicklung die Python Version "3.8.5" verwendet.

Als zusätzliche Unterstützung für ein besseres Verständnis zu biomedizinischen Konzepten soll die Konzeptdatenbank UMLS (Unified Medical Language System) zur erweiterten Übersicht über den Fachbereich genutzt werden. Dazu soll vom Autor die Version "2022AA" von UMLS durch die NLM lizensiert und heruntergeladen werden (nlm.nih.gov, 2022b).

Die Programmierung soll auf einem Computer mit Betriebssystem "Windows 10 Pro" mit folgenden Hardwarespezifikationen umgesetzt und ausgeführt werden:

- x-64 Architektur
- Intel® Core™ i7-8550U CPU @ 1.80Ghz Prozessor mit 4 Kernen
- 16 GB installierter physischer Arbeitsspeicher, 37.9 GB gesamter virtueller Speicher

Der Rechner besitzt dazu umfassende Installationen der "Visual Studio Build Tools 2019" wie "Windows 10 SDK" und "Windows 11 SDK", welche für das korrekte Interpretieren und Kompilieren vereinzelter Python-Bibliotheken auf Windows als Voraussetzung gelten.

### 5.1.1 Verwendete vorinstallierte Systempakete und Bibliotheken

Für die Optimierung der Rechenzeiten soll eine virtuelle Umgebung "repu" erstellt werden, welche als Jupyter Kernel in Jupyter Notebook hinzugefügt werden kann. Anaconda als Distribution ermöglicht die parallele Nutzung der Paketverwaltungsprogramme "PIP" (pip.pypa.io, 2022) und "conda" (docs.conda.io, 2022), sodass bei potenziellen inkompatiblen Biobliotheken für ein Paketverwaltungssystem, auf das andere System als Alternative zugegriffen werden kann.

Im Rahmen dieser Forschungsarbeit werden diverse Basispakete und Basisbibliotheken der "Anaconda 3" Distribution verwendet; wie bspw. "Pandas" (eine Bibliothek von Datenstrukturen zur einfachen Verarbeitung und Analyse von Daten) und "NumPy" (eine Bibliothek, welche die einfache Handhabung von Vektoren, Matrizen oder generell grossen mehr- dimensionalen Arrays ermöglicht).

Eine detaillierte Liste aller verwendeten Bibliotheken ist im Anhang dieser Arbeit beigefügt.

## 5.2 Übersicht und Auswahl biomedizinischer NLP-Werkzeuge und NER-Systeme in Python

Im Hauptfokus der praktischen Umsetzung beider Workflows steht die Auswahl der zu verwendeten NLP-Werkzeuge in Form von weiteren zusätzlichen Paketinstallationen. Für die Normalisierung der zu untersuchenden Textdaten soll die Bibliothek "Natural Language Toolkit" (NLTK) mit dem verfügbaren Tokenizer und der englischen Stoppwörterliste genutzt werden (nltk.org, 2022). Als Stütze und Orientierung der Implementierung sollen als Grundlage Teile des empfohlenen Programmiercode von Sarkar (2019) für die erweiterte Textanalyse verwendet werden.

Im Kern der Workflows steht jedoch die Auswahl eines NER-Systems, welches eine einfache und schnelle Identifizierung und Kategorisierung von biomedizinischen Entitäten innerhalb der zu analysierenden Textdaten ermöglichen soll, um so die formulierten Forschungs- und Unterfragen beantworten zu können. Im folgenden Kapitel wird eine kleine kategorisierte Auswahl von NER-Systemen vorgestellt, welche für die Umsetzung in "Python" zur Verfügung stehen. Anhand der jeweiligen Eigenschaften sowie Vor- und Nachteile soll ein für die praktische Umsetzung geeignetes NER-System ausgewählt werden.

### 5.2.1 NER-Systeme für die Identifikation von UMLS und MeSH-Konzepten

Neben den vielen NER-Systemen, welche als Hauptzweck die direkte Identifizierung und Einteilung biomedizinischer Entitäten in übergeordnete Konzepte wie bspw. Medikamente, chemische Substanzen, Krankheiten oder Genen verfolgen, gibt es auch NER-Systeme, welche sich dagegen das primäre Ziel setzen, die vorhandenen biomedizinischen Konzepte so genau wie möglich durch das Abgleichen mit vorhandenen Konzeptdatenbanken zu identifizieren. Diese NER-Systeme des "Term-Matching" bzw. "Konzept-Matching" versuchen in ihrer Funktionsweise, die Textdaten als Stringteile unter Verwendung von statistischen Massen so präzise wie möglich zu bspw. UMLS-Konzepten oder MeSH-Vokabular zuzuordnen (Soldaini & Goharian, 2016, S. 1).

Die meisten Entwicklerinnen und Entwickler von NER-Systemen konzentrieren sich generell darauf, die Genauigkeit ("Precision") und die Sensitivität ("Recall") der übergreifenden Extraktion von biomedizinischen Entitäten erhöhen. Dies geschieht aber meist durch zeitintensives und datenabhängiges Training der Modelle durch ML-Methoden. Dagegen fokussieren NER-Systeme wie "QuickUMLS" oder "pyMeSHSim" sich primär auf ihre Effizienz, indem sie durch das Abgleichen einen schnellen Zugriff auf vorhandenes biomedizinisches "state-of-the-art" Konzeptwissen ermöglichen (Soldaini & Goharian, 2016, S. 1). Nach einer erfolgreichen Identifikation der einzelnen Konzepte, müssen diese je abhängig des geplanten Forschungsvorhabens, noch zu ihren übergeordneten Konzepten für eine Klassifikation zugeordnet und zusammengefasst werden. Dies kann zusätzliche Fach- und Syntaxkenntnisse zu den verwendeten Konzeptdatenbanken wie bspw. UMLS voraussetzen.

Gleichzeitig sind solche "Term-Matching" Algorithmen verstärkt von der innerhalb der Textdaten verwendeten Sprache und Form abhängig. Bei einer Analyse von Textdaten in informeller Sprache können die Konzepte in Form von informellen Begriffen oder Synonymen schlechter zugeordnet werden, falls diese alternativen Konzeptbezeichnungen

nicht in der Konzeptdatenbank hinterlegt sind (Soldaini & Goharian, 2016, S. 4). Dies verdeutlicht die Abhängigkeit solcher Systeme zu den jeweilig vernetzten Konzeptdatenbanken, wenn keine unterstützenden Regeln für das Matching selbstständig in das NER-System implementiert werden.

#### 5.2.1.1 QuickUMLS

Soldaini und Goharian (2016) entwickelten mit "QuickUMLS" ein solches NER-System, das auf einem annähernden Abgleich von Begriffen aus der englischsprachigen Teilmenge von UMLS beruht, um damit biomedizinische UMLS-Konzepte in unstrukturierten Texten identifizieren zu können. Im Rahmen der Funktionsweise werden die erkannten Begriffe innerhalb der Nominalphrasen durch ihr Lemma ersetzt, bevor sie mit den Konzepten aus UMLS abgeglichen werden. Nach dem Abgleichen jedes Lemmas wird eine Liste von Konzepten als mögliche Treffer mit einem jeweilig zugehörigen statistischen Wahrscheinlichkeitsmass zu deren Treffsicherheit als "Score" generiert. Abschliessend wird das Konzept aus UMLS mit dem höchsten "Score" diesem Begriff zugeordnet. Mit dem Fokus auf die Effizienz konnte das NER-System bei einem Textdokument bestehend aus ca. 1'000 "Tokens", die zugehörigen biomedizinischen Entitäten über den Abgleich mit der UMLS-Teilmenge innerhalb von 500–1000ms zuordnen (Soldaini & Goharian, 2016, S. 1).

In einem Direktvergleich zu zwei ähnlichen NER-Systemen "cTAKES" und "MetaMap", welche auch über die gleiche Funktionalität des "UMLS-Konzept-Matching" verfügen, verzeichnete "QuickUMLS" eine höhere Sensitivität, dennoch aber eine niedrigere Genauigkeit als beide Konkurrenzsysteme. Jedoch im Rahmen der zeitlichen Effizienz war "QuickUMLS" von 2.5- bis 135mal schneller als "MetaMap" oder "cTAKES" (Soldaini & Goharian, 2016, S. 3).

#### 5.2.1.2 pyMeSHSim

Nach einem ähnlichen Funktionsprinzip haben Luo et al. (2020) mit "pyMeSHSim" ein NER-System publiziert, welches ein Konzept-Matching der unstrukturierten Textdaten mit dem kontrollierten Vokabular MeSH (als zusätzliche Eingrenzung von UMLS-Konzepten) ermöglicht. Die zusätzliche Eingrenzung der Konzepte auf das MeSH-Vokabular bietet den Vorteil der besseren Vernetzung zu den zu untersuchenden Textdaten, da MeSH auch als Indexsystem der bibliografischen Textdatenbanken "PubMed/MEDLINE" sowie anderer NLM-Datenbanken verwendet wird. Somit können bei einer Verwendung von Textdaten aus PubMed, die verfügbaren MeSH-Konzepte in Form der Indizes der

jeweiligen Dokumente zusätzlich für eine erweiterte Vernetzung der im Text identifizierten Konzepte eingesetzt werden (Luo et al., 2020, S. 2). Zusätzlich ist "pyMeSHSim" in der Lage, MeSH-Begriffe aus biomedizinischen Textdaten direkt zu parsen und die semantische Ähnlichkeit zwischen den MeSH-Konzeptpaaren zu bestimmen. Ähnlich zur Leistung von "QuickUMLS" konnte "pyMeSHSim" eine hohe Sensitivität (≥ 0.94), allerdings eine mittlere Genauigkeit (≥ 0.56) erzielen (Luo et al., 2020, S. 10).

Der Mangel der Genauigkeit solcher Systeme lässt sich darauf zurückzuführen, dass vor allem die Annotationen zu Symptomen sowie Nebenwirkungen von Medikamenten in solchen Datenbanken mehrheitlich unvollständig sind. Dies führt bei den Resultaten des NER-Systems zu einer hohen Anzahl "false positives" und senkt damit ihre Genauigkeiten (Soldaini & Goharian, 2016, S. 4).

### 5.2.2 "Rule-based" NER-Systeme mit vortrainierten Modellen

NER-Systeme, welche auf vorhanden annotierten Text-Korpussen, orthografischen Merkmalen, Ontologien oder anderem vordefiniertem Wissen basieren, werden als "rule-based" NER-Systeme bezeichnet. "Rule-based" NER-Systemen benötigen für ihren Einsatz keine zusätzlichen annotierten Textdaten und können auf neue unstrukturierte Textdaten mit ihren vortrainierten Modellen angewendet werden (Jansen, 2021). Damit sind sie für neue Nutzerinnen und Nutzer meist einfach zu verstehen und simpel zu handhaben. Wie in Kapitel 3.5.1.2 bereits erwähnt, liegt das grösste Problem beim Einsatz solcher Systeme in ihrer Abhängigkeit zur zeitlichen Aktualität der bereitgestellten Modelle. Das biomedizinische Wissen und die damit verbundene Anzahl biomedizinischer Konzepte wächst durch die hohe Anzahl durchgeführter Forschungen und klinischen Studien fortgehend. Um die Leistung und Qualität solcher NER-Systeme zu garantieren, müssen sie ständig mit neuen zusätzlichen gelabelten Daten ergänzt werden (Gao et al., 2021, S. 2; Śniegula et al., 2019, S. 261). Damit sind die Nutzerinnen und Nutzer solcher NER-Systeme besonders von den Entwicklerinnen und Entwicklern abhängig.

Kim et al. (2003) stellten mit "GENIA", einen der am häufigsten genutzten annotierten Text-Korpusse zu biomedizinischen Konzepten zur Verfügung. "GENIA" wird häufig von Forscherinnen und Forschern sowohl als Basiskorpus, als auch als unterstützendes Wörterbuch für die Entwicklung neuer "rule-based" NER-Systeme verwendet (Śniegula et al., 2019, S. 262). "GENIA" wird auch aufgrund seiner Popularität und grossen Datenmenge, als beliebter Testdatensatz für die Evaluation und den Leistungs-Vergleich neuer NER-Systeme verwendet. Der GENIA-Korpus ist aktuell in Version 3.0 verfügbar und besteht

auf der Basis von 2'000 MEDLINE-Abstracts, aus mehr als 400'000 Wörtern und fast 100'000 zugehörigen biomedizinischen Annotationen (Kim et al., 2003).

Neben "GENIA" gibt weitere annotierte Text-Korpusse, welche für "rule-based" Systeme zur Verfügung stehen. Es gibt dabei auch spezialisierte Korpusse, welche sich mit ihren annotierten Daten auf Fachgebiete der Medizin wie bspw. Onkologie oder auf bestimmte biomedizinische Entitätstypen wie z.B. Gene oder Genprodukte spezialisieren. In Tabelle 2 werden als kleine Auswahl vier spezialisierte Korpusse zu ihren spezialisierten Fachgebieten und ausgewählten Entitätstypen vorgestellt:

| Korpus | Fachgebiet und verfügbare Entitätstypen |
|---|---|
| **CRAFT** | Der "Colorado Richly Annotated Full Text" ("CRAFT") Korpus von Cohen et al. (2017) wurde auf der Grundlage von Volltext-Artikeln aus biomedizinischen Journals gebildet. Im Rahmen dieses Annotationsprojektes wollten die Autorinnen und Autoren sich auf das Wissen innerhalb Volltexte fokussieren, welches eventuell aufgrund textlicher oder struktureller Unterschiede nicht in Abstracts vorhanden ist (Cohen et al., 2017, S. 1379). Die Annotationen der Entitäten von "CRAFT" wurden in ihrer ersten Edition unter der Vereinigung von zehn verschiedenen öffentlich verfügbaren Ontologien durchgeführt; der "Chemical Entities of Biological Interest ontology", "Cell Ontology", "Gene Ontology Biological Process", "Gene Ontology Cellular Component", "Gene Ontology Molecular Function", "Molecular Process Ontology", "NCBI Taxonomy", "Protein Ontology", "Sequence Ontology" und der "Uberon anatomical ontology".<br>Damit ist "CRAFT" besonders auf die Erfassung von Zelltypen, Chemikalien, Proteinen und Genen spezialisiert (Neumann et al., 2019, S. 322). |
| **JNLPBA** | Kim et al. (2004) veröffentlichten als Folgeprojekt von "GENIA" den spezialisierten Korpus "JNLPBA" ("Joint Workshop on Natural Language Processing in Biomedicine and its Applications"), welcher die vereinzelten zugehörigen Entitäten der 36 unterschiedlichen "GENIA" Klassen bzw. Entitätstypen, auf nur insgesamt fünf Klassen reduzierte.<br>Diese fünf Klassen umfassen Proteine, DNA, RNA, Zelllinien und Zelltypen (Kim et al., 2004, S. 71). |
| **BC5CDR** | Mit "BC5CDR" ("BioCreative V CDR task") erstellten Li et al. (2016) einen spezialisierten Korpus für die Erfassung von chemischen Wirkstoffen (bzw. Medikamenten) und Krankheiten. Der resultierte BC5CDR-Korpus basiert auf 1'500 PubMed-Artikeln mit insgesamt 4'409 annotierten Chemikalien, 5'818 Krankheiten und 3'116 annotierte Wechselwirkungen zwischen Chemikalien und Krankheiten. Dabei wurde für die Annotation "MeSH" als kontrolliertes Vokabular und zusätzliche Stütze verwendet. |
| **BIONLPCG13** | Pyysalo et al. (2015) konnten in einem mehrjährigen kollaborativen Prozess durch die ergänzende Kombination des sogenannten "Cancer Genetics" (CG) Korpus und des "Pathway Curation" (PC) Korpus, einen auf Krebs spezialisierten Korpus "BIONLPCG13" aufbauen.<br>Der "CG" Korpus konzentrierte sich auf die Annotation von Entitäten zu physiologischen und pathologischen Prozessen bei Krebskrankheiten. Im "PC" Korpus hingegen wurden Entitäten zu "target-reactions" annotiert, welche für die Analyse von biomolekularen Signalwegen auf Basis von vorhandenen Ontologien relevant sein könnten (Pyysalo et al., 2015, S. 1). Als Datengrundlage dienten für beide Korpusse PubMed-Abstracts, auf deren Basis "CG" insgesamt über 17'000 Annotationen in 600 Abstracts und "PC" über 12'000 |

| | |
|---|---|
| | Annotationen in 525 Abstracts erfasste (Pyysalo et al., 2015, 17).<br>Somit stellt "BIONLPCG13" einen spezialisierten Korpus für NER-Systeme für den Einsatz in der Krebsgenetik oder allgemeinen Krebsforschung dar. |

Tabelle 2: Übersicht zu spezialisierten Korpussen mit Fachgebieten und verfügbaren Entitätstypen (eigene Tabelle)

#### 5.2.2.1 scispaCy

Neumann et al. (2019) stellten mit "scispaCy" ein öffentliches und spezialisiertes "rule based" NER-System für die Verarbeitung von biomedizinischen Textdaten zur Verfügung, welches als Erweiterung der robusten "spaCy"-Bibliothek erstellt wurde. Die Python-basierende "spaCy"-Bibliothek von Ines Montani et al. (2022) stellt eine Vielzahl praktischer NLP-Werkzeuge für die Textverarbeitung in mehreren Sprachen zur Verfügung. Die verfügbaren Modelle bieten aufgrund ihrer Geschwindigkeit, Robustheit und "state-of-the-art" nahen Leistung eine einfache und praktische Lösung dar (Neumann et al., 2019, S. 320). Besonders das NER-System von "spaCy" schnitt in einer Vergleichsstudie von vier etablierten Open-Source-NER-Systemen ("Stanford NER", "spaCy", "Alias-i LingPipe" und "NLTK") zur Genauigkeit, Sensitivität und den Rechenzeiten am zweitbesten hinter "Stanford NER" ab (Jiang et al., 2016, S. 24; Śniegula et al., 2019, S. 263). "scispaCy" soll eine robuste, effiziente und leistungsfähige Erweiterung von "spaCy" darstellen und wurde daher mit den vorhandenen spaCy3-Modellen für das "POS-Tagging" und "dependency parsing" auf neuen biomedizinischen Textdaten trainiert, um den "Tokenizer" mit neuen Regeln zu biomedizinischen Konzepten zu ergänzen (Neumann et al., 2019, S. 320). Dabei wurde das NER-System auch bspw. mit den verfügbaren Daten des "GENIA"-Korpus trainiert (Neumann et al., 2019, S. 321). Die Basisinstallation von "scispaCy" enthält die zwei Kernmodelle "en_core_sci_sm" (kleines Paket mit kleinerem Vokabular) und "en_core_sci_md" (mittelgrosses Paket mit mittelgrossem Vokabular und Wortvektoren), welche die Erfassung von biomedizinischen Entitäten als zusammengefasstes Label in Textdaten ermöglicht.

| Software Package | Processing Times Per Abstract (ms) | Sentence (ms) |
|---|---|---|
| NLP4J (java) | 19 | 2 |
| Genia Tagger (c++) | 73 | 3 |
| Biaffine (TF) | 272 | 29 |
| Biaffine (TF + 12 CPUs) | 72 | 7 |
| jPTDP (Dynet) | 905 | 97 |
| Dexter v2.1.0 | 208 | 84 |
| MetaMapLite v3.6.2 | 293 | 89 |
| **en_core_sci_sm** | 32 | 4 |
| **en_core_sci_md** | 33 | 4 |

Abbildung 7: Vergleichstabelle der Rechenzeiten verfügbarer biomedizinischer NER-Modelle (Neumann et al., 2019, S. 320)

In Abbildung 7 werden die nötigen Rechenzeiten einer Anwendung der beiden Kernmodelle von "scispaCy" mit den Rechenzeiten von alternativen NER-Systemen verglichen, die in anderen Programmiersprachen wie bspw. "C++" oder "Java" bereitgestellt wurden. Obwohl "scispaCy" nicht die Geschwindigkeiten von NER-Systemen wie "NLP4J" erreichen kann, hat die Verwendung von "scispaCy" dennoch den Vorteil, dass aufgrund der Python-Umgebung auf weitere hochleistungsstarke Bibliotheken für die schnelle Datenverarbeitung für ML (z.B. "NumPy") oder für die Textverarbeitung (z.B. "NLTK") zur Verfügung stehen (Neumann et al., 2019, S. 320).

Neben den Basis-Korpussen bietet "scispaCy" auch weitere auf verschiedenen Korpussen basierte Modelle für das allgemeine sowie auch für das spezialisierte biomedizinische NER-Tagging an (Neumann et al., 2019):

- **"en_core_sci_lg"**: Ein Erweiterungsmodell der Kernmodelle für die verallgemeinerte Erfassung von biomedizinischen Entitäten, bestehend aus ca. 785'000 Wörtern und 600'000 Wortvektoren.
- **"en_core_sci_scibert"**: Ein Modell für die verallgemeinerte Erfassung von biomedizinischen Entitäten, welches aus einem Vokabular aus ca. 785.000 Wörtern und einem Transformer-Modell zu "SciBert" besteht. "SciBert" ist dabei ein Modell, welches mit Volltextdaten aus dem Korpus von semanticscholar.org (1.14 Millionen Dokumente mit insgesamt 3.1 Milliarden "Tokens" bzw. Wörtern) trainiert wurde (Beltagy et al., 2019). Bei dessen Einsatz wird ein leistungsfähiger GPU für die nötige Rechenarbeit empfohlen.
- **"en_ner_craft_md"**: Ein spezialisiertes Modell basierend auf dem "CRAFT" Korpus, welches die Erfassung von Zelltypen, Chemikalien, Proteinen und Genen ermöglicht.

- **"en_ner_jnlpba_md"**: Ein spezialisiertes Modell basierend auf dem "JNLPDA" Korpus, welches sich auf die Erfassung von Proteinen, DNA, RNA, Zelllinien und Zelltypen spezialisiert.
- **"en_ner_bc5cdr_md"**: Ein spezialisiertes Modell basierend auf dem "BC5CDR" Korpus, welches die Extraktion von Chemikalien bzw. Medikamenten und Krankheiten ermöglicht.
- **"en_ner_bionlp13cg_md"**: Ein spezialisiertes Modell basierend auf dem "BIONLPCG13" Korpus, welches die Erfassung von insgesamt 17 verschieden Entitätstypen für den spezialisierten Bereich der Krebsforschung möglich macht**.**

Mit dem "EntityLinker" bietet "scispaCy" eine zusätzliche Funktionskomponente an, mit welcher erkannte Entitäten mit Konzeptdatenbanken (z.B. "UMLS" und "MeSH") abgeglichen werden können. Dabei können die zutreffenden Konzepte anhand eines "Matching-Scores" ausgegeben werden (Neumann et al., 2019, S. 321). Mit dieser Funktionalität sollen zusätzliche Möglichkeiten zur verbesserten Kontrolle und zum besseren Verständnis der Resultate geboten werden.

#### 5.2.2.2 Stanza (StanfordNLP)

"Stanza" (zuvor als "StanfordNLP" bekannt) von Zhang et al. (2021) ist eine öffentlich zugängliche Sammlung präziser und effizienter NLP-Werkzeuge, welche in vielen verschiedenen Sprachen verfügbar sind. "Stanza" stellt seit 2021 mit seinem NER-System mit neuralem Pipeline-Design, eine grosse Auswahl von biomedizinischen sowie auch klinischen Modellen zur Verfügung (Zhang et al., 2021, S. 1893). Das "rule-based" NER-System von "Stanza" wurde auf Basis der Korpusse "CRAFT" und "GENIA" vortrainiert.

Für den biomedizinischen Bereich bietet "Stanza" ganze 8 verschiedene spezialisierte Modelle auf Basis der öffentlich verfügbaren annotierten Daten der Korpusse "AnatEM", "BC5CDR", "BC4CHEMD", "BioNLP13CG", "JNLPBA", "Linnaeus", "NCBI-Disease" und "S800" an. Diese umfassen Daten zur anatomischen Analyse, Chemikalien, Genetik, Krankheiten und Zellbiologie (Zhang et al., 2021, S. 1895).

Darüber hinaus stellt "Stanza" auch speziell für den klinischen Bereich, d.h. zur gezielten Analyse von Texten zu klinischen Studien oder ärztlichen Berichten, zwei individuelle Modelle zur Verfügung. Das erste dieser Modelle "i2b2" kann Entitäten anhand der Labels "problem", "test" und "treatment" aus klinischen Textdaten extrahieren. Das zweite Modell "radiology" ermöglicht dagegen die spezialisierte Entitätsextraktion von

radiologischen Berichten zu den Labels "anatomy", "observation", "anatomy modifier", "observation modifier" und "uncertainty" (Zhang et al., 2021, S. 1895).

Sowohl die biomedizinischen als auch die klinischen Modelle von "Stanza" verzeichneten sehr präzise Resultate und bieten zugleich eine wettbewerbsfähige "state-of-the-art" Leistung zu anderen öffentlich verfügbaren NER-Systemen. Im einem von Zhang et al. (2021, S. 1898) selbstdurchgeführten Direktvergleich mit "scispaCy" konnte "Stanza" bei der Verwendung der spezialisierten biomedizinischen Modelle, welche auf den gemeinsamen Korpussen basierten, eine durchschnittlich +4.72% höhere Genauigkeit verzeichnen.

#### 5.2.2.3 HunFlair

"Flair" von Akbik et al. (2019) ist ein weiteres öffentlich verfügbares NLP-System, welches bereits in hunderten von Forschungsprojekten und industriellen Anwendungen im Einsatz ist. Flair zählt dabei zu den populären "Deep-Learning" Frameworks des NLP und bietet für die Sprachen Deutsch, Englisch, Holländisch und Spanisch eigene "state-of-the-art" NER- Modelle.

Weber et al. (2021) entwickelten mit "HunFlair" ein "rule-based" NER-System, welches auf der Basis des biomedizinischen Korpus "HUNER" neu umgestaltet und trainiert wurde. Für den Korpus "HUNER" wurden insgesamt 34 verschiedene Korpusse mit biomedizinischen Volltexten und Abstracts zu wissenschaftlichen Artikeln und Patentschriften unter Anwendung von ML zusammengeführt und die darin enthaltenen biomedizinischen Konzepte anhand der fünf verschiedenen Labels "chemicals", "cell lines", "diseases", "genes" und "species" annotiert (Weber et al., 2020, S. 297). Mit "HunFlair" wurde dieser Korpus in das "Flair"-NLP-Framework übertragen und steht nun als NER-System zur Erfassung von Zelllinien, Chemikalien, Krankheiten, Genen und Spezies mit hoher Genauigkeit zur Verfügung (Weber et al., 2021, S. 1).

### 5.2.3 NER-Systeme mit "Transfer-Learning" Kapazitäten

Neben der Verwendung von direkt funktionsbereiten NER-Systemen wie bspw. "QuickUMLS" oder "scispaCy" besteht auch die Möglichkeit, ein neues eigenes NER-System durch die Methoden des "Transfer-Learning" zu erstellen. Dabei dient meist ein NER-System als Vorlage und wird anschliessend mithilfe neuer Daten durch ML selbst weitertrainiert und weiterentwickelt. Die Mehrheit der bisher vorgestellten NER-Systeme

wurden anhand dieser Methoden bereitgestellt und konnten ihre jeweilige Leistung im Vergleich zur Vorlage deutlich steigern (Jansen, 2021).

Eine der wichtigsten automatisierten Techniken des "Transfer-Learning" im Bereich des NLP ist "Bidirectional Encoder Representation from Transformers" (BERT), welche von Google durch Devlin et al. (2018) publiziert und entwickelt wurde. "BERT" stellt mit seiner "Transformer Neural Network" Architektur, eine effektive "Deep Learning" Methode zum Trainieren von NLP-Modellen für bspw. Sprachübersetzungen oder NER dar (Asghari et al., 2022, S. 184).

> *"According to BERT developers, training using the Base version of BERT with 12 Layers, 768 Hidden sizes, and 12 Self Attention heads (i.e., 110 million Total parameters) is accomplished in four days using four cloud Tensor Processing Units (TPUs). …. This model is a general language model and not a specific domain model, such as health care, which opens an area of research and application to train domain-specific BERT." (Asghari et al., 2022, S. 184–185)*

Wie Asghari et al. (2022) betonen, eignet sich das "BERT" Modell hervorragend für das Trainieren von biomedizinischen NER-Modellen. Jedoch ist die Verwendung von "BERT" mit hohen Rechenkapazitäten verbunden. Eine der massgeblich prominenten "BERT"-trainierten biomedizinischen NER-Modelle ist "BioBERT", welches unter Verwendung von Textdaten aus dem allgemeinen Bereich und anschliessend von Textdaten aus dem biomedizinischen Bereich wie bspw. PubMed trainiert wurde (Jansen, 2021; Lee et al., 2020). Der Trainingsprozess dieses NER-Modells nahm dabei ganze 23 Tage unter der Verwendung von acht "NVIDIA V100 GPU's" mit jeweils 32GB Speicher in Anspruch (Lee et al., 2020, S. 1237).

Viele vortrainierte "BERT"-NER-Modelle stehen in einer Form auch den direkt einsatzbereiten "rule-based" NER-Systemen zur Verfügung wie z.B. "SciBERT" in "scispaCy".

### 5.2.4 Vergleich der NER-Systeme und Auswahl für die praktische Umsetzung

Aus den drei zuvor vorgestellten Kategorien von NER-Systemen soll nun ein NER-System für das Forschungsvorhaben ausgewählt werden, welches für die Beantwortung der formulierten Forschungs- und Unterfragen ideal geeignet ist.

Die Entwicklung eines neuen eigenen NER-Systems mithilfe der Methoden des "Transfer Learnings" stellt leider keine Option dar, da sowohl die nötigen Fachkompetenzen als auch die nötigen Rechenressourcen dem Autor nicht zur Verfügung stehen.

Die Nutzung eines NER-Systems für die Extraktion und das Matching von UMLS und MeSH-Konzepten, wie "QuickUMLS" oder "pyMeSHSim", bietet zwar eine zuverlässige Lösung zur Identifizierung von biomedizinischen Entitäten innerhalb von Textdaten, erfordert jedoch einen deutlich hohen Aufwand für die Nachbereitung der Resultate. So müssten die nach UMLS oder MeSH extrahierten Konzepte nachträglich in eigene übergeordnete Gruppen als Klassen zusammenfasst werden, damit die Analyse anhand der formulierten Unterfragen durchgeführt werden kann.

Die "rule-based" NER-Systeme stellen mit ihren vortrainierten verfügbaren allgemeinen oder spezialisierten Modellen die beste Lösung für das geplante Forschungsvorhaben dar. Im Rahmen dieser Arbeit wurden drei verschiedene öffentlich verfügbare "rule-based" NER-Systeme vorgestellt, welche nach den Aussagen der jeweiligen Entwicklerinnen und Entwickler über "state-of-the-art" biomedizinische NER-Modelle verfügen. "HunFlair" bietet ein einziges, dennoch sehr starkes spezialisiertes Modell für die Extraktion von "chemicals", "cell lines", "diseases", "genes" und "species" an. Wie in Abbildung 8 dargestellt verzeichnet "HunFlair" auf den Korpussen "JNLPBA", "BC5CDR" und "NCBI" als Testdaten, bessere Resultate in Form von F1-Scores als "scispaCy" (Weber et al., 2021, S. 3):

| | JNLPBA (Gene) | BC5CDR | NCBI |
|---|---|---|---|
| SciBERT | 77.28 | 90.01 | 88.57 |
| BioBERT v1.1 | 77.49 | 89.76 | **89.71** |
| CollaboNET | **78.58** | 87.68 | 88.60 |
| SciSpacy | – | 83.92 | 81.56 |
| HunFlair | 77.60 | 89.65 | 88.65 |
| HunFlair (vanilla) | 77.78 | **90.57** | 87.47 |

Abbildung 8: F1-Scores von "HunFlair" im Vergleich zu anderen NER-Systemen und Modellen (Weber et al., 2021, S. 3)

"Stanza" stellt dagegen acht spezialisierte biomedizinische NER-Modelle und sogar zwei klinische Modelle bereit. In Abbildung 9 verglichen Zhang et al. (2021, S. 1898) die Qualität von "Stanza" anhand F1-Scores auf Basis bekannter Korpusse als Testdaten mit "BioBERT" und "scispaCy":

| Category | Dataset | Domain (# of Entities) | Stanza | BioBERT | scispaCy |
|---|---|---|---|---|---|
| Bio | AnatEM | Anatomy (1) | 88.18 | – | 84.14 |
| | BC5CDR | Chemical, Disease (2) | 88.08 | – | 83.92 |
| | BC4CHEMD | Chemical (1) | 89.65 | 92.36 | 84.55 |
| | BioNLP13CG | Cancer Genetics (16) | 84.34 | – | 77.60 |
| | JNLPBA | Protein, DNA, RNA, Cell line, Cell type (5) | 76.09 | 77.49 | 73.21 |
| | Linnaeus | Species (1) | 88.27 | 88.24 | 81.74 |
| | NCBI-Disease | Disease (1) | 87.49 | 89.71 | 81.65 |
| | S800 | Species (1) | 76.35 | 74.06 | – |
| Clinical | i2b2 | Problem, Test, Treatment (3) | 88.13 | 86.73 | – |
| | Radiology | Radiology Report (5) | 84.80 | – | – |

Abbildung 9: F1-Scores von "Stanza" im Vergleich zu "BioBERT" und "scispaCy" (Zhang et al., 2021, S. 1898)

Diese Faktoren machen "Stanza" zum Favoriten des geplanten Forschungsvorhaben. Bei einem Testeinsatz auf dem vom Autor eingesetzten Computer verursachte "Stanza" jedoch enorm hohe Rechenzeiten, vermutlich verursacht durch das neuronale Pipeline-Design und dem Mangel eines leistungsstarken GPUs, was zu einem erzwungenen Verzicht auf "Stanza" für das praktische Vorhaben der Forschungsarbeit führte. Die dabei verwendete virtuelle Umgebung und die getesteten Rechenzeiten von "Stanza" sind im Anhang dieser Arbeit dokumentiert. Mit dem von Neumann et al. (2019) betonten Fokus auf schnelle und robuste NER-Modelle, fällt die **endgültige Wahl des NER-Systems auf "scispaCy"**. Im Vergleich zu "HunFlair" bietet "scispaCy" trotz schlechterer Tagging-Resultate, vier verschiedene spezialisierte NER-Modelle, welche eine breitere Analyse der formulierten Unterfragen im Vergleich zu "HunFlair" ermöglicht.

Für das Forschungsvorhaben soll somit das erweiterte allgemeine Basis-Modell "en_core_sci_lg" und die in Abbildung 10 dargestellten spezialisierten biomedizinischen NER-Modelle verwendet werden:

| model | F1 | Entity Types |
|---|---|---|
| en_ner_craft_md | 78.35 | GGP, SO, TAXON, CHEBI, GO, CL |
| en_ner_jnlpba_md | 70.89 | DNA, CELL_TYPE, CELL_LINE, RNA, PROTEIN |
| en_ner_bc5cdr_md | 84.70 | DISEASE, CHEMICAL |
| en_ner_bionlp13cg_md | 76.79 | AMINO_ACID, ANATOMICAL_SYSTEM, CANCER, CELL, CELLULAR_COMPONENT, DEVELOPING_ANATOMICAL_STRUCTURE, GENE_OR_GENE_PRODUCT, IMMATERIAL_ANATOMICAL_ENTITY, MULTI-TISSUE_STRUCTURE, ORGAN, ORGANISM, ORGANISM_SUBDIVISION, ORGANISM_SUBSTANCE, PATHOLOGICAL_FORMATION, SIMPLE_CHEMICAL, TISSUE |

Abbildung 10: Spezialisierte biomedizinische NER-Modelle von "scispaCy" mit verfügbaren Labels (Neumann et al., 2019)

Auf Basis dieser finalen Auswahl muss für die Verwendung von "scispaCy" auf die Python-Version "3.8.5" zurückgegriffen werden, da nach aktuellem Stand dieser Arbeit in der Windows-Umgebung einige Distributionen ("wheels") zu den benötigten Bibliotheken bei einer Verwendung höherer Python-Versionen fehlen.

## 5.3 Auswahl der Textdaten und Datenbanken

Als Textdaten sollen (frei)-verfügbare Volltexte oder Abstracts aus PubMed und ClinicalTrials.gov genutzt werden. PubMed ist der auf Grundlage von MeSH indexierte öffentlich zugängliche Anteil der bibliografischen Datenbank MEDLINE. In PubMed stehen mehr als 34 Millionen Abstracts zu Publikationen der Bereiche Medizin und Biowissenschaft zur Verfügung (nlm.nih.gov, 2022a). Als fokussierte Quelle von Textdaten soll jedoch primär die Textdatenbank ClinicalTrials.gov genutzt werden, welche sowohl privat und öffentlich finanzierte klinische Studien beherbergt. Die Textdaten dieser Studien stehen jeweils in Form einer gekürzten sowie einer detaillierten Zusammenfassung zur Verfügung (ClinicalTrials.gov, 2022). ClinicalTrials stellt besonders für das noch nicht in anderen Datenbanken oder in der Fachliteratur erfasste Wissen (wie z.B. zu Nebenwirkungen von Medikamenten) eine wertvolle Quelle dar (Su, 2019, S. 61).

Die Auswahl einer spezialisierten Datenbank für die Extraktion von biomedizinischem Vorwissen wird im späteren Abschnitt zur Methode 2 genauer erläutert. Dabei sollen die Empfehlungen von Tanoli et al. (2021) besondere Beachtung finden, welche auf statistischen Eigenschaften zu Qualität, Verfügbarkeit, Datenredundanz, Attribute, Vielfalt der Datentypen und der Datenbanknutzung (Anzahl Zitationen) basieren. Die Kategorie der Datenbank soll allerdings durch das formulierte Ziel der Methodenvariation bestimmt werden.

### 5.3.1 Extraktion, Bereitstellung und Normalisierung der Textdaten

Aus der öffentlichen Datenbank von ClinicalTrials können grosse Mengen an klinischen Studienberichten (nach Durchführung einer gezielten Suche) in Blöcken von je maximal 10'000 xml-Dateien heruntergeladen werden. Die für diese Arbeit relevanten Textdaten jeder einzelnen Studie sind dabei jeweils unter dem xml-tag "detailed-description" in Form einer detaillierten Zusammenfassung oder unter dem xml-tag "brief_description" in Form einer kurzen Zusammenfassung bzw. Abstract extrahierbar. Nach dem Download der Studien aus ClinicalTrials sollen die gewünschten Textdaten der lokal gespeicherten

xml-Dateien, mithilfe der Python-Bibliothek "BeautifulSoup" in das Programm eingelesen werden (Richardson, 2007).

Die Abstracts aus PubMed können dagegen auch nach einer durchgeführten Suche zur Eingrenzung, mit der Funktion "save-PubMed" als "pubmed-…-set" in Form einer einzelnen txt-Datei, bestehend aus maximal 10'000 Abstracts, heruntergeladen werden. Diese txt-Datei ist nach einer speziellen String-Struktur aufgebaut, sodass für die Extraktion der Abstracts als einzelne Dokumente ein regulärer Ausdruck durch das Python Paket "regex" eingesetzt werden soll (van Rossum, 2022a).

Für die Normalisierung und Standardisierung der extrahierten Textdaten sollen Stoppwörter anhand der verfügbaren englischen Stoppwortlisten von "NLTK" entfernt und die Texte von unerwünschten Zeichen wie Punkte, Doppelpunkte oder Kommas mithilfe eines regulären Ausdruckmusters bereinigt werden. Neben der Entfernung von Zeilenumbrüchen und mehrfachen Leerzeichen werden keine weiteren Normalisierungen der Texte durchgeführt. Auf die in anderen Anwendungsfällen empfohlene Konvertierung aller Buchstaben in die Kleinschreibung, wird aufgrund des Einsatzes eines NER-Systems verzichtet. Das NER-System konnte bei Tests, unabhängig des ausgewählten NER-Modells, mit dem Verzicht auf die Konvertierung der Textdaten in die Kleinschreibung eine höhere Anzahl Entitäten erfassen. Um die Qualität der Ergebnisse der verwendeten NER-Tagger zu verbessern, wurden anhand der Ergebnisse selbständig ungewollte und unpassende Entitäten als zusätzliche "Stoppwörter" bzw. "Filterwörter" für das NER-Tagging gesammelt. Diese zusätzlichen "Filterwörter" sollten für die fortgehende Forschungsarbeit bei der Extraktion von biomedizinischen Entitäten zukünftig nicht mehr erfasst werden.

## 5.4 Methode 1: Kookkurrenz Analyse basierend auf dem "GBA-Prinzip"

Die Methode 1 setzt sich das Ziel, die **Unterfrage 1** zu untersuchen und eine diesbezügliche praktische Umsetzung zu entwickeln. Mit der Methode 1 soll das alleinige Potenzial von unstrukturierten biomedizinischen Textdaten als Wissensbestände überprüft und messbar gemacht werden, indem in einem ungerichteten Ansatz nur das in die NER-Systeme integrierte biomedizinische Wissen verwendet werden soll.

Beruhend auf dem "GBA-Prinzip" von Chiang und Butte (2009) zur Ermittlung von Ähnlichkeiten von Medikamenten und Krankheiten auf Basis der überschneidenden Behandlungsprofilen, soll basierend auf dieser "similarity based" Vorgehensweise, mithilfe NLP-Methoden eine Kookkurrenz Analyse mit unstrukturierten biomedizinischen Textdaten

durchgeführt werden. Die Kookkurrenz Analyse soll dabei die Textdaten als individuelle Einzeldokumente (PubMed-Abstract oder klinische Studie) analysieren. Die Kookkurrenz Analyse verfolgt das Ziel, die semantische Nähe zwischen diesen Dokumenten auf Basis von Mustern zum gemeinsamen Auftreten von Wörtern zu bestimmen. Die Analyse basiert dabei auf der Annahme, dass Terme bzw. Wörter voneinander abhängig sind, wenn sie auffällig häufig gemeinsam in den gleichen Dokumenten auftreten (Kroeger, 2009, S. 20).

Um dem biomedizinischen NER-Tagging die bedeutendste Rolle zuzuweisen, sollen alle nicht identifizierten biomedizinischen Entitäten aus den Dokumenten entfernt werden, sodass nur Entitäten, wie z.B. Medikamente, Krankheiten, Symptome, Gene, etc., für die Kookkurrenz Analyse verwendet werden. Für diesen Schritt sollen geeignete NER-Modelle von "scispaCy" zu den ausgewählten Entitätstypen verwendet werden. Mit dem übergeordneten Ziel der Untersuchung und Beantwortung der Unterfrage 1, soll mithilfe der Methode 1 geprüft werden, welches und wieviel Wissen (in Form von "Drug Repurposing" Kandidaten) sich mit der alleinigen Verwendung eines biomedizinischen NER-Systems aus unstrukturierten Textdaten extrahieren lässt.

Die durch das NER-Tagging gefilterten Dokumente werden anschliessend als "Bag of Words"-Vektoren mit ihren "TF-IDF"-Werten (term frequency – inverse document frequency) dargestellt (Harris, 1954). Abschliessend soll die Kosinus-Ähnlichkeit als statistisches Mass dienen, um die Ähnlichkeiten zwischen den einzelnen Dokumenten in Form von Abständen zu ermitteln. Anhand der errechneten Ähnlichkeitsabständen zwischen den jeweiligen Dokumenten untereinander sollen die ähnlichsten Dokumentenpaare erfasst werden. Dies soll in zwei Schritten durchgeführt werden. Mit der Beschränkung auf die von den NER-Modellen erkannten Entitäten ist der Fall, dass zwei Dokumente nun anhand ihrer gefilterten Inhalte identisch sind, nicht selten. Diese identischen Paare werden erfasst und gespeichert. Im zweiten Schritt werden anschliessend auch die jeweiligen ähnlichsten Dokumentenpaare, eingeschränkt durch eine maximale Obergrenze der errechneten Dokumentenabständen, erfasst und gespeichert.

Im letzten Schritt des Workflows soll das "GBA-Prinzip" von Chiang und Butte (2009) auf die erfassten Dokumentenpaare für die Ermittlung von "Drug Repurposing" Kandidaten direkt angewendet werden. Dazu werden die jeweiligen Inhalte der bestimmten Dokumentenpaare, in ihrer ungefilterten Form paarweise als String zusammengeführt. Anschliessend werden mithilfe "scispaCy" und dem spezialisierten NER-Modell "BC5CDR" zu Krankheiten und Chemikalien bzw. Medikamenten (und dem Modell "BIONLP13CG"

als Unterstützung), alle Entitäten dieser Typen aus den jeweiligen zusammengeführten Dokumentenpaaren extrahiert. Auf Basis der von Chiang und Butte (2009) formulierten Theorie der überschneidenden Behandlungsprofilen bezüglich des "GBA-Prinzips" gilt nun die Annahme, dass jedes durch den NER-Tagger erfasste Medikament, ein potenzielles Behandlungsmittel bzw. ein "Repurposing Kandidat" für jede einzelne erfasste Krankheit darstellt. Diese neu identifizierten Assoziationen zwischen den Medikamenten und Krankheiten werden abschliessend mithilfe der Dictionary-Datenstruktur gesammelt und gespeichert, indem jede erfasste Krankheit ein "key" und jedes zugehörige Medikament ein "value" dieses "key" darstellt. Dieses Dictionary soll abschliessend auch als JSON-Datei lokal abgespeichert werden.

### 5.4.1 Kernelemente der Methode

Innerhalb dieses Teilkapitels sollen die jeweiligen Bezüge der Methode 1 zu den zusätzlichen Fragestellungen erläutert und Kernelemente der technischen Implementation hervorgehoben werden.

#### 5.4.1.1 Variationen der Entitätsselektion

Um biomedizinischen Entitäten die grösste Bedeutung der Kookkurrenz Analyse zuzuweisen, werden mithilfe verschiedener ausgewählter NER-Modelle von "scispaCy" diese Entitäten dokumentenweise extrahiert und als jeweilige gefilterte Dokumente für die Analyse verwendet. Für diese Entitätsselektion sollen für die Methode 1 vier verschiedene Variationen bereitgestellt werden, welche sich an den unterschiedlichen "Drug Repurposing" Vorgehensweisen orientieren. Dabei sollen die verfügbaren NER-Modelle von "scispaCy" kombiniert werden, indem die modellübergreifenden Labels zu den gleichen Fachgebieten ergänzend genutzt werden. Damit soll das in den unterschiedlichen Modellen vorhandene Fachwissen zu den einzelnen spezialisierten Fachbereichen optimal eingesetzt werden.

Für die Untersuchung und Beantwortung der zusätzlichen Fragestellung **ZF 1** wurden vier unterschiedliche Variationen bezüglich der Selektion der Entitäten gebildet:

- **Biomedizinische Entitäten (verallgemeinert):** Hier werden alle vom erweiterten Basismodell "en_core_sci_lg" erkannten verallgemeinerten biomedizinischen Entitäten extrahiert.
- **Gene, Genome und Genprodukte:** Hier werden alle Entitäten im Bereich der Genetik und Genprodukte durch die kombinierte Verwendung der Modelle

"en_ner_craft_md" und "en_ner_bionlp13cg_md" extrahiert. Diese Variation orientiert sich anhand der Bedeutung von überschneidenden Genen, welche auch vertieft bei den "target-based" Vorgehensweisen im Fokus liegen.

- **Krankheiten, Symptome bzw. Nebenwirkungen:** Da fachspezifisch kein NER- Modell in der Lage ist, zwischen Krankheiten, Nebenwirkungen und Symptomen zu differenzieren, sollen dennoch orientiert an der "side-effect-based" Vorgehensweise mithilfe des Modells "en_ner_bc5cdr_md" alle Entitäten zu Krankheiten, welche aber auch Symptome oder Nebenwirkungen darstellen können, extrahiert werden. Dabei soll vorweg das Modell "en_ner_bionlp13cg_md" als Unterstützung verwendet werden, indem mit diesem Modell vorab "ungewollte" Entitäten in Form von Krankheiten als Stoppwörter identifiziert werden.
- **Zellen, Zellkomponenten und Zelllinien:** In der letzten Variation sollen alle Entitäten im Bereich der Zellbiologie unter der kombinierten Verwendung der Modelle "en_ner_craft_md", "en_ner_jnlpba_md" und "en_ner_bionlp13cg_md" extrahiert werden. Orientiert an der "target-based" Vorgehensweise stehen dabei die Überschneidungen der Medikamentenzielen als bspw. Zellen oder Zelllinien im Fokus.

In Tabelle 3 werden die für jede Variation verwendeten Modelle und zugehörigen Labels dargestellt:

| **scispaCy NER-Modell** | **Variation und verwendete Labels** | | | |
|---|---|---|---|---|
| | *Allgemeine biomedizinische Entitäten* | *Gene, Genome und Genprodukte* | *Krankheiten, Symptome bzw. Nebenwirkungen* | *Zellen, Zellkomponenten und Zelllinien* |
| en_core_sci_lg | "ENTITY" | – | – | – |
| en_ner_craft_md | – | "GO", "SO", "GGP" | – | "CL" |
| en_ner_jnlpba_md | – | – | – | "CELL_TYPE", "CELL_LINE" |
| en_ner_bc5cdr_md | – | – | "DISEASE" | – |
| en_ner_bionlp13cg_md | – | "GENE_OR_GENE_PRODUCT" | "CANCER", "PATHOLOGICAL_FORMATION" | "CELL", "CELLULAR_COMPONENT" |

Tabelle 3: Variationen der Methode 1 mit den jeweilig verwendeten Labels der "scispaCy" NER-Modelle (eigene Tabelle)

#### 5.4.1.2 Berechnung der Ähnlichkeiten und Dokumentenabstände

Um die Ähnlichkeit zwischen den Dokumenten zu bestimmen, wird die Kosinus-Ähnlichkeit anhand den "Bag of words" Dokumentenvektoren berechnet. Da alle Dokumenten-

vektoren keine negativen Werte einnehmen können, bewegen sich somit die Werte der Kosinus-Ähnlichkeit zwischen 0 für die minimale Ähnlichkeit und 1 für die maximale Ähnlichkeit. In der praktischen Umsetzung wurden daher für die Bestimmung von paarweisen Ähnlichkeitsabständen der zugehörige Kosinus-Abstand verwendet, welcher die 1-x invertierte Kosinus-Ähnlichkeit darstellt, sodass 0 den minimalen Ähnlichkeitsabstand und 1 den maximalen Ähnlichkeitsabstand einnimmt. Die detaillierte Implementierung befindet sich im Anhang dieser Arbeit.

#### 5.4.1.3 Bestimmung der identischen und ähnlichsten Dokumentenpaare

Mit den errechneten Kosinus-Abständen können nun die identischen und ähnlichsten Dokumentenpaare ermittelt werden. Für die Bestimmung der identischen Dokumente soll die generierte Ähnlichkeitsabstandsmatrix verwendet werden, bei der die jeweiligen Zeilen i und Spalten j die vereinzelten Dokumente und die darin gespeicherten Einträge, die jeweiligen paarweisen Ähnlichkeitsabstände darstellen. Dies bedeutet, dass bei jedem vorhandenen Eintrag des Wertes 0, das Dokument i und das Dokument j ein identisches Dokumentenpaar bilden. Die jeweiligen Einträge der Eigenpaare, d.h. wenn i=j gilt, sollen bei der praktischen Umsetzung auf den maximalen Abstand 1 gesetzt werden. Abschliessend werden die identischen Dokumentenpaare i, j der zugehörigen Einträge mit Abstand 0, als Tupel der Dokumentenindexe in eine Liste ("idents") gespeichert.

Für die Bestimmung der "ähnlichsten" Paare werden anhand der gleichen Ähnlichkeitsabstandsmatrix die minimalen Werteinträge (>0) für jedes Einzeldokument gesucht. Zusätzlich wird eine maximale Obergrenze der minimalen Ähnlichkeitsabstände festgelegt, um damit die "ähnlichsten" Dokumentenpaare einzugrenzen. Die verbliebenen Dokumentenpaare i, j werden anschliessend als Tupel der Dokumentenindexe in eine Liste ("sims") gespeichert.

## 5.5 Methode 2: Assoziationsketten basierend auf Swanson's "ABC-Modell"

Mit Methode 2 soll **Unterfrage 2** untersucht und dabei ein praktischer Workflow für dessen mögliche Beantwortung entwickelt werden. Die Unterfrage 2 erforscht das Potenzial von unstrukturierten biomedizinischen Textdaten als ergänzende Wissensbestände zu Datenbanken. Dabei soll in einem gerichteten Ansatz neben dem in das NER-System integrierte biomedizinische Wissen, auch Fach- bzw. Vorwissen aus den "besten" verfügbaren Datenbanken eingesetzt werden, um neue "Repurposing" Kandidaten zu ermitteln.

Basierend auf dem offenen "ABC-Entdeckungsmodell" von Swanson (1986) sollen (A-B)-Startassoziationspaare anhand verfügbarer Relationen zwischen biomedizinischen Entitäten aus Datenbanken extrahiert werden. Anschliessend sollen in unstrukturierten Textdaten, die zugehörigen biomedizinischen Entitätspaare (B-C) ermittelt werden, um somit mit der transitiven Relation (A-C) neue "Repurposing" Kandidaten zu bestimmen. Im Rahmen dieser Methode sollen dabei bestimmten Entitätstypen von jeweils A, B und C sowie die Länge der Assoziationskette variiert werden, um somit unterschiedliche Arten von Vorgehensweisen testen zu können.

In der praktischen Umsetzung dieser Methode werden nach einer Auswahl der biomedizinischen Entitätstypen von A & B als Startpaare, auf Basis der Empfehlungen der Fachliteratur, wie bspw. von Tanoli et al. (2021), eine bestgeeignete Datenbank bestimmt. Aus dieser Datenbank werden anschliessend die verfügbaren B-Entitäten der (A-B) Relationen als Suchterme für eine Volltextsuche extrahiert. Mithilfe dieser B-Terme werden mithilfe einer Volltextsuche die betreffenden Dokumente bestimmt, welche mindestens einen dieser B-Terme enthalten. Abschliessend werden mit den verfügbaren spezialisierten NER-Modellen, die anhand der Assoziationskette vorbestimmten, gesuchten Entitätstypen C aus den Trefferdokumenten extrahiert. Abhängig zur Art oder Form der verwendeten Assoziationskette, stellt so bspw. C ein "Repurposing" Kandidat für A als Ergebnis dar.

### 5.5.1 Kernelemente der Methode

Innerhalb dieses Teilkapitels sollen die jeweiligen Bezüge der Methode 2 zu den zusätzlichen Fragestellungen erläutert und Kernelemente der technischen Implementation hervorgehoben werden.

#### 5.5.1.1 Arten der Assoziationsketten und Auswahl der Datenbanken

Für die Untersuchung und Beantwortung der zusätzlichen Fragestellung **ZF 2** wurden auf Basis der Fachliteratur vier unterschiedliche Arten von Assoziationsketten gebildet. Um eine Vergleichbarkeit der Assoziationsketten deren Ergebnisse zu ermöglichen, sollen alle Ketten vom gleichen Entitätstyp A ausgehen und mit dem gleichen Entitätstyp C als Ergebnis abschliessen. Als Entitätstyp A soll daher eine Krankheit ausgewählt und für den Entitätstyp C sollen Medikamente als "Repurposing" Kandidaten für A ermittelt werden. Diese bestimmte Struktur der Assoziationsketten ermöglicht somit auch die Vergleich-

barkeit der Ergebnisse der Methode 1 mit den Ergebnissen der Methode 2, sodass auch die **ZF 3** im Rahmen der Forschungsarbeit untersucht und überprüft werden kann.

Für jede Art der Assoziationskette soll für die Extraktion der jeweiligen Startrelationen (A-B), die "beste" Datenbank anhand der Empfehlungen aus der Fachliteratur ausgewählt werden. In Tabelle 4 werden die verwendeten Assoziationsketten und Datenbanken beschrieben:

| **Assoziationskette** | **Begründete Datenbankauswahl** | **Beschreibung** |
|---|---|---|
| **"disease-gene-drug"** | Die Datenbank "OpenTargets" (opentargets.org) ist eine der umfassendsten Datenbanken für Gen, Protein- und Krankheitsassoziationen (Tanoli et al., 2021, S. 1672). "OpenTargets" beherbergt dabei mehr als 28'000 Proteine als Genprodukte, 10'000 verschiedene Krankheiten und damit insgesamt mehr als drei Millionen Assoziationsrelationen zwischen Gen/Proteinen und Krankheiten (Tanoli et al., 2021, S. 1669). | Im Rahmen dieser Assoziationskette wird der theoretische Ansatz von Yang et al. (2017) verwendet, dass bei einer Behandlung einer Krankheit mit einem Medikament bestimmte Gene angesprochen werden.<br>Andere Medikamente, welche auch eine (positive) Wirkung auf diese Gene haben, kommen somit als "Repurposing" Kandidaten in Frage (Alaimo & Pulvirenti, 2019, S. 102; Jin & Wong, 2014, S. 641). So soll zu Beginn eine Krankheit bestimmt und die davon beeinträchtigten Gene als Startpaare (A-B) aus "OpenTargets" extrahiert werden. |
| **"disease-gene_variant-drug"** | "DisGeNET" (disgenet.org) stellt neben "OpenTargets" eine der bedeutendsten Datenbanken für Gen, Protein- und Krankheitsassoziationen dar (Tanoli et al., 2021, S. 1672). Im Vergleich zu "OpenTargets" stellt "DisGe-NET" zusätzlich mehr als 210'498 Assoziationen zwischen Gen/Protein-Varianten und Krankheiten zur Verfügung (Tanoli et al., 2021, S. 1669). | Als Erweiterung des von Yang et al. (2017) beschriebenen Ansatzes zur Überschneidung von Genen, soll mit einer weiteren Assoziationskette anhand der Überschneidungen von Gen- und Proteinvarianten neue "Repurposing" Kandidaten bestimmt werden. Dabei sollen nach Auswahl einer Krankheit, die zugehörigen bekannten Genvarianten aus "DisGeNET" als Startpaare (A-B) extrahiert werden. |
| **"disease-symptom-drug"** | Die von Köhler et al. (2021) veröffentlichte "Human Phenotype Ontology" Datenbank, beherbergt eine allumfassende Sammlung phänotypischer Anomalien, die bei menschlichen Krankheiten dokumentiert wurden und wird als Datenbank als weltweiter Top-Standard bezeichnet (Köhler et al., 2021, S. 1). | Diese Assoziationskette basiert auf der Annahme von Vogt et al. (2014, S. 52–53), dass zwei Krankheiten ähnlich sind, wenn sie sich die gleichen Symptome teilen. Auf Basis dieser Annahme sollen nach Auswahl einer Krankheit, alle dazu zugehörigen Symptome als (A-B) Startpaare aus der Datenbank "Human Phenotype Ontology" (hpo.jax.org) extrahiert werden. |
| **"(disease)-drug-sideeffect-drug"** | "DrugBank" (go.drugbank.com) ist eine auf Medikamenten und Wirkstoffen spezialisierte Datenbank und stellt Wissen zu den chemischen, pharmakologischen und pharmazeutischen Eigenschaften, kombiniert mit dem "drug-target" Wissen zu | Zhang et al. (2013) haben im Rahmen eines Forschungsprojektes festgestellt, dass die Ermittlung von potenziellen "Drug Repurposing" Kandidaten anhand überschneidender Nebenwirkungen zu präziseren Resultaten führt, als anhand gemeinsamer chemischen oder genetischen |

| | | |
|---|---|---|
| | z.B. Krankheiten, zugehörigen Sequenzen, Strukturen und Signalwegen zur Verfügung. Sie bietet Wissen zu insgesamt mehr als 12'000 Chemikalien und Medikamenten und stellt die umfassendste und meistzitierte Datenbank für diesen Fachbereich dar (Tanoli et al., 2021, S. 1667).<br><br>"SIDER" von Letunic (2022) ist eine der einzigen, jedoch meistausführlichen und meistzitierten Datenbanken für die Extraktion von Medikament-Nebenwirkung Assoziationen. Sie verfügt über Daten zu Nebenwirkungen von insgesamt mehr als 1'400 Medikamenten (Tanoli et al., 2021, S. 1670). | Zielstrukturen.<br>Aufgrund der theoretischen Grundlage von Assoziationsbeziehungen kann die Assoziationskette beliebig mit zusätzlichen Entitätsbeziehungen erweitert werden. Auf Basis der Forschungsresultate von Zhang et al. (2013) und orientiert an der "side-effect-based" Vorgehensweise soll eine erweiterte Assoziationskette der Form (A-B-C-D) gebildet werden. Für die Gewährleistung der Vergleichbarkeit der Ergebnisse soll auch für die erweiterte Kette eine Krankheit A als Startpunkt ausgewählt werden. Anschliessend sollen zwei Datenbanken für die Extraktion der Relationen "disease-drug" (A-B) und "drug-sideeffect" (B-C) eingesetzt werden. Die extrahierten Entitäten C werden anschliessend als Suchterme verwendet, um Dokumente mit Entitäten D als "Repurposing" Kandidaten für A zu ermitteln.<br>Für die Extraktion der Relationen (A-B) soll die Datenbank "DrugBank" verwendet werden. Anschliessend werden aus "SIDER" die Nebenwirkungen C aller Medikamente B erfasst. |
| **"(disease)-drug-cell_lines-drug"** | Neben "DrugBank" soll die Datenbank "GDSC" ("Genomics of Drug Sensitivity in Cancer", cancerrxgene.org) für die Extraktion von Medikament-Zelllinien Assoziationen verwendet werden.<br>Speziell für den Fachbereich der Onkologie stellt "GDSC" die umfangreichste Datenbank mit mehr als 1'000 Krebs-Zelllinien und Daten aus ca. 75'000 Experimenten dar (Tanoli et al., 2021, S. 1673). | In einem theoretisch unabhängigen offenen Ansatz sollen Ähnlichkeiten von Medikamenten daran erfasst werden, indem die Überschneidungen zu Zelllinien gesucht werden. Dies beruht auf der Annahme, dass anhand der in klinischen Studien gleichen beobachteten Zelllinien bei Medikamenten potenzielle Ähnlichkeiten zwischen den Medikamenten und deren Wirkung bestimmt werden können.<br>Parallel zum Aufbau der Kette "(disease)-drug-sideeffect-drug" werden für die betroffenen Medikamente B von "DrugBank" aus der Datenbank "GDSC" die getesteten Zelllinien C als Suchterme erfasst. |

Tabelle 4: Formulierte "ABC" Assoziationsketten und verwendete Datenbanken (eigene Tabelle)

Die formulierten Assoziationsketten und dessen unterschiedliche Strukturen sollen auf Basis eines Fallbeispiels und den daraus resultierenden Ergebnissen getestet und verglichen werden.

#### 5.5.1.2 Einlesen und Verwendung der Suchterme für die Dokumentensuche

Die Mehrheit der verwendeten Datenbanken stellen nach einer eingrenzenden Suche durch A, die gesuchten (A-B) Relationen für die Extraktion der B-Entitäten als tsv-Downloaddatei zur Verfügung. Nur bei der "Human Phenotype Ontology" müssen die

gesuchten Entitäten B manuell als Suchterme extrahiert werden. Für das Einlesen der tsv-Dateien wird die Bibliothek "Pandas" verwendet, welches ermöglicht heruntergeladenen Datensätze als "Pandas"-Dataframe Datentyp zu nutzen. Anhand dieser Dataframes werden jeweils die gewünschten Entitäten als Suchterme extrahiert und als separate Strings in eine Liste gespeichert. Je nach Art der Assoziationskette werden bei Bedarf diese Terme um ihre jeweils korrespondierende kleingeschriebene Form ergänzt. Die Liste der Suchterme soll abschliessend in einem letzten Schritt von allfälligen Duplikaten oder ungewollten Termen bereinigt werden.

Die Volltextsuche wird anschliessend mit einer einfachen for-Schleife über alle vorhandenen Dokumente durchgeführt, mit dieser in jedem Dokument die gesammelten Terme entweder als Stringteile oder als Einzelwörter (mithilfe der ".split()" Methode) gesucht werden. Jedes Dokument, welches mindestens einen Suchterm enthält, wird als Trefferdokument in eine neue Trefferliste hinzugefügt. Gleichzeitig werden parallel als zusätzliche Kontrollstütze alle gefundenen Terme in eine separate Liste gespeichert, anhand dieser unter Aufsicht ungewollte Suchterme entfernt werden können.

#### 5.5.1.3 Bestimmung der "Repurposing" Kandidaten

Für die Bestimmung der "Repurposing" Kandidaten werden nun in allen Dokumenten, die sich in der Trefferliste befinden, mithilfe "scispaCy" und dem spezialisierten NER-Modell "BC5CDR" (und dem Modell "BIONLP13CG" als Unterstützung) alle Chemikalien bzw. Medikamente als erkannte Entitäten C extrahiert und mit den zuvor gefundenen zugehörigen Termen für das Dokument als Assoziationspaare (B-C) in ein Dictionary gespeichert. Die Terme stellen dabei die "keys" und die Chemikalien bzw. Medikamente die zugehörigen "values" dar. Abschliessend werden alle im gesamten Dictionary vorhandenen "values" in eine Liste gespeichert. Diese "values" als Entitäten C (oder D) stellen als Ergebnis "Repurposing" Kandidaten für die zuvor ausgewählte Krankheit A dar.

## 5.6 Bestimmung des Fallbeispiels

Wie kurz in Kapitel 3.3.5.2 angedeutet, hat der Bereich der Onkologie besonders mit dem ungelösten und akuten Problem der Medikamentenresistenz zu kämpfen (Jin & Wong, 2014, S. 642). Der Einsatz von Arzneimitteln für die Krebstherapie stellt eine besondere Herausforderung dar, da schon nach wenigen Einsätzen eines Medikamentes die Krebszellen eine Resistenz gegen dieses Medikament entwickeln (Rodrigues et al., 2022, S. 1). Aus diesem Grund sind Krebs-Patientinnen und Patienten auf zusätzliche bzw.

alternative Medikamente angewiesen. Die Medikamentenentwicklung im Bereich der Onkologie leidet besonders unter langen Entwicklungszeiten bis zur Markteinführung und den damit verbundenen exorbitant hohen Kosten (Issa et al., 2021, S. 132).

"Drug Repurposing" stellt für den Krebsbereich mit zahlreichen Vorteilen wie bspw. einer höheren Effizienz, einem geringeren zeitlichen und finanziellen Aufwand sowie einem geringeren Ausfallrisiko in den klinischen Studien, eine signifikante Alternative zur Medikamentenentwicklung dar (Rodrigues et al., 2022, S. 2). Auch das in Kapitel 2.2 vorgestellte "Nutzen-Risiko-Verhältnis" ist bei "Repurposing" Kandidaten für die Behandlung von Krebs deutlich besser, als bei anderen eher herkömmlichen Krankheiten (Hodos et al., 2016, S. 186; Pushpakom et al., 2019, S. 44).

Nach Rodrigues et al. (2022, S. 3) können "Repurposing" Kandidaten bei Krebstherapien als Monotherapie (Behandlung einer Krankheit mit genau einem einzigen Medikament), als Mittel zur Chemoprävention (Mittel zur Vorbeugung gegen Krebs) oder als kombinierte synergetische Verstärkungsmittel bereits bekannter Krebsmedikamente (d.h. Kombinationstherapie) genutzt werden. Darüber hinaus können sie auch für die Regulierung von Nebenwirkungen anderer Medikamente verwendet werden und auch für die "Adjuvante Therapie" eingesetzt werden, welche das Ziel hat, das Wiederauftreten von Tumoren zu verhindern.

All diese Gründe machen das "Drug Repurposing" zu einem sehr wichtigen Konzept und auch grossen Chance für die Entwicklung neuer Krebstherapien. Mit einem zusätzlichen Bezug auf eine klinische Studie von Halatsch et al. (2021), bei welcher neun "Repurposing" Kandidaten als mögliche synergetische Verstärkungsmittel des bereits etablierten Medikamentes "Temozolomid" für die Behandlung des wiederkehrenden Glioblastoms (bösartiger hirneigener Tumor) getestet wurden, soll das Glioblastom als Krankheit und Fallbeispiel für die Forschungsarbeit ausgewählt werden. So sollen mit dem Einsatz der Methoden 1 & 2, "Repurposing" Kandidaten für die Behandlung des Glioblastom ermittelt werden und die formulierten Unterfragen und zusätzlichen Fragestellungen anhand der ermittelten "Repurposing" Kandidaten als Ergebnisse beantwortet und diskutiert werden.

### 5.6.1 Auswahl der Fallbeispieldaten

Um die jeweiligen Methoden messbar zu machen und grosse Datenmengen zu verhindern, sollen die verwendeten Textdaten aus PubMed und ClinicalTrials anhand eines für das Glioblastom übergeordneten MeSH-Begriffes ausgewählt werden.

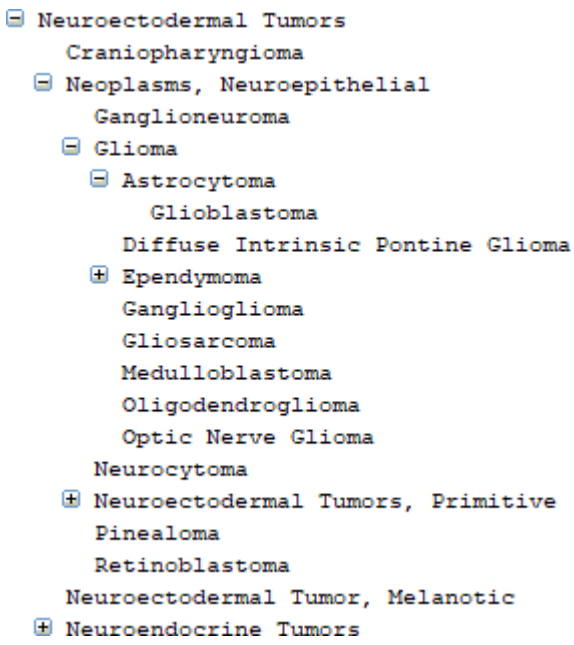


Abbildung 11: Ausschnitt der aus UMLS extrahierten MeSH-Ontologie zum Oberbegriff "Neuroectodermal Tumors" mit allen untergeordneten Konzepten (nlm.nih.gov, 2022b)

Wie in Abbildung 11 dargestellt, wird als eingrenzender Suchbegriff der in MeSH dem Glioblastom übergeordnete Begriff **"Neuroectodermal Tumors"** ausgewählt. Dieser umfasst neben dem Glioblastom alle ausgehenden Tumore des zentralen und peripheren Nervensystems wie z.B. Hautkrebs oder Augenkrebs. Diese Selektion basiert auf der Überlegung, dass die in der Ontologie nahen Konzepte sich inhärente Eigenschaften teilen, wie z.B. im Fallbeispiel, dass alle untergeordneten Tumore sich aus Nervenzellen entwickeln.

Am 12.7.2022 wurden anhand des Suchbegriffes "Neuroectodermal Tumors" aus ClinicalTrials insgesamt 6'608 klinische Studien extrahiert. Zur Erweiterung und zur Vervollständigung wurden anschliessend alle Studien zu den untergeordneten Konzepten "Glial Tumor Malignant" und "Glioblastoma" extrahiert, um potenzielle Studien zu ergänzen, welche anhand der Metadaten, unvollständig in der Datenbank erfasst wurden. So wurden die Textdaten aus ClinicalTrials auf insgesamt 6'741 Studien ergänzt. Zum gleichen Datum wurden aus PubMed für den Suchbegriff "Neuroectodermal Tumors" die Abstracts der 10'000 relevantesten Ergebnisse extrahiert. Als finaler Testdatensatz wurden bei der Implementierung und der Ermittlung der Resultate alle **6'741 Studien aus ClinicalTrials** mit den ersten **3'259 Abstracts aus PubMed** ergänzt. Somit umfasste der Testdatensatz insgesamt 10'000 Dokumente.

Im Rahmen der Methode 2 wurde für die Extraktion der (A-B) Startassoziationen aus den verschiedenen Datenbanken, der Suchbegriff "Glioblastoma" verwendet.

## 5.7 Evaluation der Ergebnisse

Sowohl Methode 1 als auch Methode 2 verfolgen das Kernziel, mit ihren Resultaten neue "Repurposing" Kandidaten in Form von Medikamenten bzw. chemischen Substanzen für ausgewählte Krankheiten zu bestimmen. Anhand des Fallbeispiels ist beim gerichteten Ansatz der Methode 2 diese Krankheit mit "Glioblastoma" vorbestimmt. Bei der Methode 1 und allen zugehörigen Variationen, wird dagegen aufgrund des ungerichteten Ansatzes ein Dictionary zu allen extrahierten Krankheiten und zugehörigen "Repurposing" Kandidaten generiert. Um nun eine Vergleichbarkeit mit den Ergebnissen der Methode 2 zu ermöglichen, sollen bei allen verwendeten Variationen der Methode 1, nur die Einträge der "keys" "glioblastoma" und "glioblastoma multiforme" extrahiert werden. Dadurch können im Rahmen des Fallbeispiels alle Ergebnisse untereinander verglichen werden.

Für die Evaluation und Analyse aller Ergebnisse soll die Datenbank **"DrugBank"** als Abgleichgrundlage und Hilfestütze dienen. Als "state-of-the-art" Datenbank stellt sie die umfassendste Sammlung von Medikamenten und chemische Substanzen mit Bezügen zu ihren Einsatzmöglichkeiten sowie ihren aktuellen Status in klinischen Studien dar (Jin et al., 2021, S. 6). Für den Abgleich aller ermittelten Ergebnisse mit "DrugBank" wurde für das ausgewählte Fallbeispiel der Eintrag für "Glioblastoma Multiforme (GBM)" verwendet, dieser auch als Synonym von "Glioblastoma" gilt (go.drugbank.com, 2022b).

Da jedoch in "DrugBank" die jeweiligen Einzelwirkstoffe und -medikamente als Einzeleinträge vorhanden sind, d.h. nicht zusammen mit ihren möglichen Kombinationen in Rahmen einer Kombinationstherapie, sollen daher auch die extrahierten Resultate beider Methoden auf einzelne Wirkstoffe und Medikamente reduziert werden. In diesem Prozess werden Duplikate der erfassten Einzelwirkstoffe und Einzelmedikamente aus der Liste aller "Repurposing" Kandidaten entfernt, um somit das Verhältnis zu neuen individuellen Chemikalien, Wirkstoffen und Medikamenten als potenziell unbekannte "Repurposing" Kandidaten im Rahmen des Forschungsvorhaben besser hervorzuheben.

### 5.7.1 Formulierte Bewertungskriterien und Bewertungsprozess

Aufgrund der komplexen Art der vorhandenen Resultate ist eine repräsentative Bewertung der einzelnen "Repurposing" Kandidaten anhand ihrer Qualität fast unmöglich, da im Rahmen dieser Arbeit weder die Ressourcen noch das fachspezifische Wissen vorhanden sind, um jeden einzelnen extrahierten Kandidaten zu dessen Wirkung bei der Behandlung des Glioblastoms testen zu können. Um zumindest annähernde Aussagen zu

deren Qualität machen zu können, soll deshalb das in "DrugBank" vorhandene Wissen als Benchmark und Orientierung genutzt werden. So wurden für die Analyse und Auswertung drei selbstdefinierte Faktoren als annähernde Kennzahlen gebildet:

| Kennzahl | Beschreibung |
|---|---|
| **Trefferrate (TR)** | Die Trefferrate stellt den Anteil der durch das NER-Tagging gültig erfassten Medikamenten, Wirkstoffen oder Therapiemöglichkeiten dar, d.h. den Anteil aller gültig identifizierten biomedizinischer Entitäten. Dieser soll die allgemeine NER-Tagging-Leistung oder die Leistung der Methode generell auf Basis aller ermittelten Resultate darstellen.<br><br>Die Trefferrate soll daher für jede Methode und Variante anhand folgenden Verhältnisses berechnet werden:<br><br>$Trefferrate = \frac{Anzahl\ gültiger\ Medikamente, Chemikalien\ und\ Therapien}{Gesamtanzahl\ ermittelter\ Kandidaten}$<br><br>Dabei stellt die Gesamtanzahl ermittelter Kandidaten, die Gesamtanzahl aller durch die Methoden ermittelten Kandidaten innerhalb der jeweilig finalen Liste dar und die Anzahl der "gültiger" Kandidaten alle Elemente dieser Liste, welche tatsächlich unabhängig ihrer vermeintlichen Qualität ein Einzelmedikament, eine Chemikalie oder eine bekannte Therapie (z.B. "TT Fields", eine Behandlungsmethode mit dieser schwache elektrische Felder erzeugt werden, welche durch die Haut der Kopfhaut pulsieren und damit die Zellteilung von Krebszellen verhindern) darstellen. Die Trefferrate stellt gleichzeitig auch die gemeinsame Summe des Qualitätsfaktors und des Potenzialfaktors der ausgewählten Methode dar. |
| **Qualitätsfaktor (QR)** | Der Qualitätsfaktor soll eine orientierbare Kennzahl für die annähernde Messung der Qualität der ermittelten Ergebnisse als "Repurposing" Kandidaten darstellen. Dieser gibt den Anteil der ermittelten Wirkstoffe, Medikamente und Therapiemöglichkeiten dar, welche für das "Glioblastom" in "DrugBank" entweder als bestätigte Medikamente, Therapien oder als bereits bekannte "Repurposing" Kandidaten in der Rubrik "Drug Trials" gekennzeichnet sind.<br><br>$Qualitätsfaktor =$<br><br>$\frac{Anzahl\ in\ DrugBank\ verzeichnete\ Medikamente, Chemikalien\ und\ Theapien}{Gesamtanzahl\ ermittelter\ Kandidaten}$<br><br>Für die für das Fallbeispiel ausgewählte Krankheit "Glioblastoma", sind in "DrugBank" insgesamt 320 verschiedene Wirkstoffe bzw. Medikamente verzeichnet. Davon sind vier als offizielle und bestätigte Behandlungsmedikamente gekennzeichnet: "Carmustin", "Irinotecan", "Nimotuzumab" und "Temozolomid"<br><br>Die restlichen 316 Wirkstoffe sind als offizielle potenzielle Kandidaten in der Rubrik "Drug Trials" verzeichnet. |

| | |
|---|---|
| **Potenzialfaktor (PF)** | Der Potenzial-Faktor soll das Potenzial der unstrukturierten Textdaten als ergänzende Wissensbestände erfassen, indem er den Anteil, der von "Drug-Bank" unbekannten potenziellen "Repurposing" Kandidaten wiedergibt. Im ausgewählten Fallbeispiel sind dies alle gefundenen Arzneimittel, Wirkstoffe oder chemischen Verbindungen, welche für das Glioblastom in "DrugBank" nicht verzeichnet sind.<br>$$Potenzialfaktor = \frac{Anzahl\ unbekannter\ Medikamente, Chemikalien\ und\ Therapien}{Gesamtanzahl\ ermittelter\ Kandidaten}$$ |

Tabelle 5: Für die Evaluation selbstdefinierte Kennzahlen mit Beschreibung (eigene Tabelle)

Anhand dieser drei Kennzahlen sollen die Ergebnisse beider Methoden sowie deren Unterarten und Varianten vergleichbar gemacht werden. Dabei stellt die gemeinsame Summe des Qualitätsfaktors und des Potenzialfaktors die Trefferrate der ausgewählten Methode dar.

Für den Bewertungsprozesses wurden nach einem Testprobelauf der Analyse folgende Regeln zur Einteilung und Kategorisierung der einzelnen Kandidaten formuliert:

- Synonyme, Abkürzungen sowie "Brand"-Namen (wie bspw. "Avastin", der von Roche bestimmte Produktname für den Wirkstoff "Bevacizumab") eines Medikamentes oder Wirkstoffes gelten als Treffer.
- Trotz der zuvor verwendeten Bereinigungsmassnahmen der Kandidatenliste, können dennoch Kandidaten in Form einer Multitherapie wie z.B. "bleomycin/vincristine/lomustine/dacarbazine" vorkommen. In einem solchen Fall wird der Kandidat als unbekannter Kandidat klassifiziert, falls mindestens einer dieser Wirkstoffe in "DrugBank" nicht verzeichnet ist. Im zuvor erwähnten Beispiel sind "Lomustine" und "Vincristine" in "DrugBank" als Kandidaten für Glioblastoma verzeichnet. "Bleomycin" und "Dacarbizine" jedoch nicht, demnach wird dieser Kandidat als "unbekannt" klassifiziert.
- Unspezifische Oberbegriffe, wie bspw. "acid", "alcohol", "malate" oder "antibody", werden nicht als gültige Treffer gezählt.

# 6 Ergebnisse der Forschungsarbeit & Diskussion

Auf Basis der forschungsleitenden Fragestellung wurden zwei verschiedene Methoden entwickelt, um biomedizinische Textdaten nutzbar zu machen und damit neue "Repurposing" Kandidaten zu ermitteln. In diesem Kapitel werden die formulierten Unterfragen, die zusätzlichen Fragestellungen und die forschungsleitende Frage beantwortet sowie diskutiert. Gleichzeitig werden die wichtigsten Beobachtungen und Erkenntnisse der Forschungsarbeit dokumentiert.

## 6.1 Form der Ergebnisse

Bei der Anwendung aller Methoden konnten zusammengefasst folgende Arten von Medikamenten, Wirkstoffen oder Therapiemöglichkeiten als "Repurposing" Kandidaten bestimmt werden:

- **Chemische Elemente:** Es konnten in "DrugBank" verzeichnete chemische Elemente wie Kalzium, Indium, Gadolinium sowie auch radioaktive Isotopen von Elementen wie z.B. Lutetium-177 erfolgreich ermittelt werden. Gleichzeitig konnten auch für "DrugBank" unbekannte Elemente wie Eisen oder Platin erfasst werden, welche in diversen Kombinationstherapien verwendet wurden.
- **Impfungen:** Es konnten zahlreiche experimentelle Impfstoffe für die Behandlung von Glioblastoma erfasst werden, welche zu einem Grossteil nicht in "DrugBank" verzeichnet waren. Ein Beispiel ist der Impfstoff "DNX-2401" ("Tasadenoturev") zur Behandlung von "Recurrent Glioblastoma", welcher nach aktuellem Stand die zweite klinische Phase abgeschlossen hat.
- **Chemische Verbindungen:** Die meisten Kandidaten wurden in Form von chemischen Verbindungen, wie z.B. Säuren, Salzen, synthetischen Derivaten, etc., erfasst. Bestätigte Kandidaten wie bspw. "5-Aminolävulinsäure", "Natriumchlorid" und "O6-Benzylguanine" konnten erfolgreich identifiziert werden.
- **Diverse therapeutische Mittel:** Auch diverse therapeutische Mittel, welche kein Medikament oder eine chemische Verbindung darstellen konnten als Kandidaten bestimmt werden, wie bspw. das elektronische Therapiegerät "TT Fields".

Neben diesen vermehrt bestätigten "Repurposing" Kandidaten wurden auch folgende Wirkstoffe und Behandlungsmittel extrahiert:

- **Hormone:** Auch Hormone, wie bspw. Östrogen oder Steroidhormone, konnten als Kandidaten identifiziert werden. Dabei wurden allerdings auch zahlreiche Kontrazeptiva erfasst, welche in einigen Studien in einen krebsfördernden Zusammenhang mit Gliomen bei Erwachsenen gebracht wurden (Felini et al., 2009).
- **Fluoreszierende Chemikalien**: Neben chemischen Wirkstoffen wurden auch fluoreszierende Chemikalien erfasst, welche im Rahmen der Krebstherapien als fluoreszierende Stoffe für Screening-Methoden eingesetzt werden, um die durch Wirkstoffe verursachten biologischen Prozesse in lebenden Organismen besser sichtbar zu machen. Ein Beispiel eines solchen erfassten Mittels ist "7-Aminoactinomycin" (7-AAD).

Dagegen wurden die meisten ungültigen Kandidaten in Form von Mengeneinheiten und zeitlich bedingten Dosisangaben (z.B. "10mg/kg", "21-day"), von unklaren Abkürzungen oder Synonymen von potenziellen Wirkstoffen (bspw. "iac"), von zu allgemeinen Oberbegriffen (z.B. "acid", "agent", "anti-cancer") und von Krankheiten oder Symptomen (z.B. "glioblastoma", "twitch") erfasst.

## 6.2 Methode 1: Analyse der Ergebnisse

Die Methode 1 setzte sich das Ziel folgende Unterfrage zu beantworten:

> **Unterfrage 1:** Wie können "Repurposing" Kandidaten aus unstrukturierten Textdaten ohne die Verwendung von Vorwissen aus Datenbanken zu Medikamenten oder Krankheiten ermittelt werden?

Mit vier verschiedenen Methodenvarianten zu unterschiedlichen Vorgehensweisen konnten mithilfe einer Kookkurrenz Analyse von Textdaten aus ClinicalTrials und PubMed neue "Repurposing" Kandidaten in Form von Medikamenten und Wirkstoffen erfolgreich ermittelt werden. Dabei wurde für die Bestimmung der "Repurposing" Kandidaten kein zusätzliches fachspezifisches Wissen aus Datenbanken herangezogen, sondern nur exklusiv das in "scispaCy" und dessen Korpussen integrierte biomedizinische Wissen verwendet. Dennoch wurden die jeweiligen Ergebnisse unter einer Aufsicht ("Supervision") generiert, indem für den Einsatz der jeweiligen unterschiedlichen NER-Modelle ergänzende "Stoppwortlisten" zur Entfernung von ursprünglich erfassten, dennoch unerwünschten Entitäten erstellt wurden. Damit konnten für die Kookkurrenz Analyse unerwünschte Entitäten wie bspw. "repeats", "antibody" oder "stop" ausgeschlossen werden. In Abbildung 12 werden die zusammengefassten Ergebnisse aller vier Arten der Kookkurrenz Analyse dargestellt:

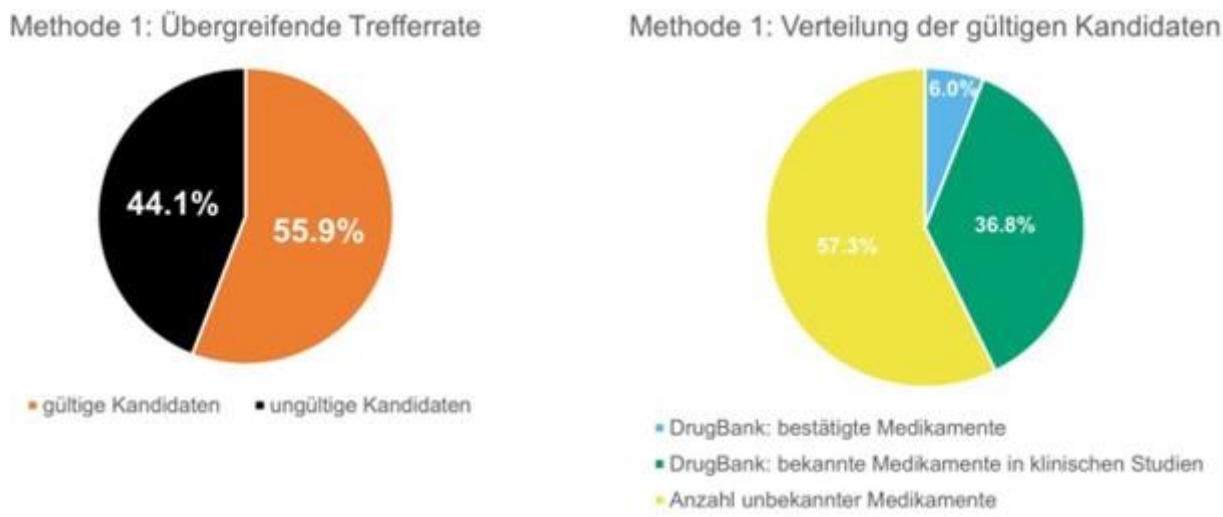


Abbildung 12: Zusammengefasste Ergebnisse der Methode 1 (eigene Grafik)

Mit 55.9% scheint die Trefferrate eher tief zu wirken, jedoch muss hier beachtet werden, dass im Rahmen der einzelnen Bereinigungsschritte der aus dem Dictionary extrahierten Entitäten, Duplikate entfernt und Kombinationstherapien aufgeteilt sowie auf Einzelmedikamente und Einzelwirkstoffe reduziert wurden. Durch diese Schritte wird die allgemeine Trefferrate der Ergebnisse vermindert. Nichtsdestotrotz bietet die Methode mit einer Trefferrate über 50% und damit mit mehr als der Hälfte aller ermittelten Kandidaten, eine effektive und gültige Auswahl (fast) individueller Einzelwirkstoffe und Einzelmedikamente als "Repurposing" Kandidaten zur potenziellen Behandlung des Glioblastoms. Die Anteile der Trefferrate in Form des Qualitätsfaktors mit 23.9% und des Potenzialfaktor mit 32.0% zeigen, dass fast ein Viertel aller ursprünglich ermittelten potenziellen "Repurposing" Kandidaten auch in "DrugBank" als bestätigte "Repurposing" Kandidaten für das Glioblastom gelten. Damit stellen die restlichen überwiegenden 32.0% auch ein tendenziell repräsentatives Potenzial der Textdaten zur Entdeckung neuer möglicher "Repurposing" Kandidaten für das Glioblastom dar.

Als weiterer möglicher Vergleichspunkt der vermeintlichen Qualität der Ergebnisse wurde mithilfe des NER-Taggings, jedes Einzeldokument nach dem "GBA-Prinzip" von Chiang und Butte (2009) analysiert und alle gefundenen Krankheit-Medikament Assoziationen nach der gleichen Systematik wie in Methode 1, in ein Dictionary gespeichert sowie abschliessend alle erfassten Kandidaten für "glioblastoma" und "glioblastoma multiforme" in eine bereinigte Liste extrahiert. Aus dieser finalen Liste von insgesamt 4'605 Treffern wurden als Stichprobe 1'000 Treffer ausgewählt und nach der gleichen Evaluierungsmethode durch den Abgleich mit "DrugBank" ausgewertet. Dabei wurde für diese 1'000 Treffer eine Trefferrate von 46.9%, bestehend aus einem Qualitätsfaktor von 16.3% und

Potenzialfaktor von 30.6%, festgestellt. Wie zwar zuvor erwähnt, dass die Trefferrate durch die Entfernung von Duplikaten bei einer höheren Gesamtanzahl ermittelter Kandidaten tendenziell abnimmt, können die übergreifenden Ergebnisse der Methode 1 einen mit 7.3 Prozentpunkten höheren Qualitätsfaktor und damit einen signifikanten Unterschied *(Normalapproximation, zweiseitig α=0.05, p=0.00, N=3906)* vorweisen. Damit kann die Vermutung aufgestellt werden, dass die Kookkurrenz Analyse unmittelbar qualitativ bessere Ergebnisse ermitteln kann.

Die übergreifenden Ergebnisse der Methode 1 zeigen, dass man auch mithilfe NLP-Methoden und durch die Bestimmung von Ähnlichkeiten zwischen Dokumenten von biomedizinischen Textdaten repräsentative "Repurposing" Kandidaten ermitteln kann. Dies auch ohne grosses Vor- und Fachwissen zu Medikamenten oder Krankheiten. Durch den ungerichteten Ansatz der Methode 1 können gleichzeitig auch "Repurposing" Kandidaten ausserhalb des ausgewählten Fallbeispiels, bzw. für andere Krankheiten neben "Glioblastoma", aus den Ergebnissen ermittelt werden.

### 6.2.1 Methode 1: Vergleich der Variationen

**ZF 1:** Wie unterscheiden sich die Ergebnisse der Methode 1 abhängig der ausgewählten Entitätstypen für die Kookkurrenz Analyse zur Bestimmung der Ähnlichkeiten zwischen Dokumenten?

Auf Basis der ersten zusätzlichen Fragestellung wurden vier unterschiedliche Variationen der Methode 1 verwendet und evaluiert. Dabei wurden durch die Nutzung verschiedener NER-Modelle unterschiedliche Arten von biomedizinischen Entitäten für die Kookkurrenz Analyse aus den Textdaten extrahiert. Hierbei soll zusätzlich beachtet werden, dass das fokussierte Ziel der Forschung darstellte, individuelle "Best-Practice" Variationen zu entwickeln. Dabei wurde für fast jede Variation eine eigene individuelle Stoppwortliste zur Optimierung der jeweiligen Ergebnisse erstellt. Daher kann der Vergleich der Resultate aufgrund der individuell unterschiedlichen Optimierungen nicht als rein objektiv betrachtet werden:

| | Allgemeine biomedizinische Entitäten | Gene, Genome und Genprodukte | Krankheiten, Symptome bzw. Nebenwirkungen | Zellen, Zell-komponenten und Zelllinien |
|---|---|---|---|---|
| **Maximaler paar-weiser Dokumen-tenabstand** | < 0.35 | < 0.2 | < 0.2 | < 0.2 |
| **Gesamtanzahl ermittelter Kandidaten** | 470 | 746 | 692 | 998 |
| **Trefferrate** | 51.28% | 54.42% | 56.94% | 58.42% |
| **Qualitätsfaktor** | 23.83% | 26.27% | 24.28% | 21.84% |
| **Potenzialfaktor** | 27.45% | 28.15% | 32.66% | 36.57% |

Tabelle 6: Methode 1 Variationen: Kennzahlen bezüglich Ergebnisse (eigene Tabelle)

Wie in Tabelle 6 dargestellt, liessen sich für jede Variation der Methode 1 bei den Verhältnis Kennzahlen Trefferrate, Qualitätsfaktor und Potenzialfaktor keine extremen und deutliche Ausreisser zu den übergreifenden Mittelwerten (TR 55.9%, QF 23.9%, PF 32%) feststellen.

Im paarweisen Vergleich der einzelnen Variationen können die grössten Unterschiede zwischen den Ergebnissen der Variation *"Allgemeine biomedizinische Entitäten"* und der Variation *"Zellen, Zellkomponenten und Zelllinien"* festgestellt werden. Die Variation *"Allgemeine biomedizinische Entitäten"* wurde als einzige Variation nicht mit einer Stoppwortliste optimiert und konnte trotz des erhöhten maximalen Dokumentenabstandes, die wenigsten Kandidaten ermitteln. Zusätzlich verzeichnete sie die niedrigste Trefferrate mit 51.28%. Dennoch blieb der Qualitätsfaktor im gleichen Wertbereich wie bei den anderen Variationen, was dafürspricht, dass sich die Qualität der Ergebnisse trotz des erhöhten maximalen Dokumentenabstandes kaum verschlechtert hat. Als grösster Gegensatz konnte die Variation *"Zellen, Zellkomponenten und Zelllinien"* die höchste Anzahl an Kandidaten ermitteln und dabei gleichzeitig die höchste Trefferrate mit 58.42% sowie den höchsten Potenzialfaktor mit 36.57% verzeichnen. Diese Unterschiede zwischen der höchsten und der niedrigsten Trefferrate *(Normalapproximation, zweiseitig α=0.05, p=0.01, N=1468)* sowie des höchsten und des niedrigsten Potenzialfaktors *(Normalapproximation, zweiseitig α=0.05, p=0.00, N=1468)* sind signifikant. Zusätzlich konnte diese Variation den höchsten überwiegenden Potenzialfaktor im Verhältnis zum Qualitätsfaktor vorweisen, was ein Indiz für eine mögliche hohe Anzahl von "false positives" darstellen könnte.

Für einen zusätzlichen Vergleichspunkt wurde eine Kookkurrenz Analyse mit den ungefilterten normalisierten Textdaten durchgeführt. Dabei konnten mit einem maximalen Dokumentenabstand von < 0.35 insgesamt 383 Kandidaten mit einer zugehörigen Trefferrate von 50.65%, einem Qualitätsfaktor von 25.07% und einem Potenzialfaktor von 25.09% erfasst werden. Im Rahmen der individuellen Gestaltung der Methode 1 lässt sich so vermuten, dass durch die Nutzung eines spezialisierten NER-Taggings zur Filterung für die Kookkurrenz-Analyse, die allgemeine Ermittlung von gültigen Einzelkandidaten leicht verfeinert werden kann. So können vermutlich durch die Verwendung von niedrigeren maximalen Dokumentenabständen treffsichere "Repurposing" Kandidaten ermittelt werden.

In der Gesamtbetrachtung aller Variationen, trotz der unterschiedlich orientierten Vorgehensweisen wie bspw. "target-based" und "side-effect-based", liessen sich anhand der Ergebnisse bis auf wenig einzelne signifikanter Anteilsunterschiede keine bedeutenden Ergebnisunterschiede erkennen.

### 6.2.2 Methode 1: Zusätzliche Beobachtungen und Erkenntnisse

Wie zuvor erwähnt, wurden fast alle Variationen unter individueller Aufsicht mithilfe Stoppwortlisten optimiert. Aufgrund der Gestaltung des Workflows konnten durch die genaue Untersuchung der identischen Dokumente, ungewollte Entitäten gut vorab erkannt und als neues Stoppwort hinzugefügt werden. Dieser Aufbau ermöglicht damit auch für zukünftige Analysevorhaben eine Aufsicht der Zwischenergebnisse der verwendeten NER-Modelle. Für die Variation *"Gene, Genome und Genprodukte"* wurden insgesamt 95 Stoppwörter, bestehend aus vermeintlich falschen Entitäten wie bspw. "repeat" oder auch unspezifisch allgemeinen Entitäten wie "DNA" gesammelt. Bei der Variation *"Zellen, Zellkomponenten und Zelllinien"* wurden nach gleichem Prinzip, die zu unspezifischen Entitäten wie bspw. "cell" oder "tumor cell" zu insgesamt 67 Stoppwörtern gesammelt. Mit erfassten 15'086 Stoppwörtern wurde die umfassendste Aufsicht bei der Variation *"Krankheiten, Symptome bzw. Nebenwirkungen"* angewendet, bei der die identifizierten Entitätsarten "CANCER" und "PATHOLOGICAL_FORMATION" als Stoppwörter erfasst wurden, um die für die Kookkurrenz Analyse erfassten Anteile von Krankheits- und Krankheitssymptoms-Entitäten reduzieren zu können und damit einen Fokus auf Nebenwirkungen zu ermöglichen.

Die grösste Herausforderung und gleichzeitig grösste Limitation der Methode 1 stellt die geeignete Auswahl der verfügbaren NER-Modelle und zugehörigen Labels dar. Viele der

verfügbaren Labels, wie bspw. "GENE_OR_GENPRODUCT" oder "DISEASE", stellen eine maximale Grenze der möglichen spezifizierten Klassifizierung von Entitäten dar, sodass in diesem Beispiel nicht zwischen Genen oder Proteinen sowie Krankheiten oder Symptomen unterschieden werden kann. Dies ist dagegen auch dem Umstand geschuldet, dass viele Konzepte in der Biomedizin nicht trennscharf und oft miteinander verbunden sind. So werden z.B. Alkylanzien wie "Carboplatin", d.h. Wirkstoffe bei denen Alkylgruppen in die DNA eingeführt werden, auch unter dem Label der "GENE_OR_GENPRODUCT" erfasst, obwohl es im Hauptkontext ein Medikament bzw. Wirkstoff darstellt.

## 6.3 Methode 2: Analyse der Ergebnisse

**Unterfrage 2:** Wie können "Repurposing" Kandidaten aus unstrukturierten Textdaten unter Verwendung von Vorwissen aus Datenbanken zu Medikamenten oder Krankheiten ermittelt werden?

Für die Beantwortung der Unterfrage 2 wurden fünf unterschiedliche Assoziationsketten für die Ermittlung von "Repurposing" Kandidaten aus Textdaten entwickelt und getestet. Dabei wurden unterschiedliche Arten von Vorwissen für das Fallbeispiel in Form von extrahierten "A-B" Relationen aus ausgewählten Datenbanken als Start-Assoziationspaare genutzt, um im Rahmen eines gezielten Ansatzes neue "Repurposing" Kandidaten entdecken zu können. Die Arten von Assoziationsketten unterschieden sich dabei in ihrem Aufbau sowie auch in ihrer Länge. Jedoch mussten sie für die Gewährleistung der Vergleichbarkeit aller Resultate, als gemeinsamen Startpunkt die für das Fallbeispiel ausgewählte Krankheit "Glioblastom" verwenden und als gemeinsamen Endpunkt Medikamente und Wirkstoffe erfassen. In Abbildung 13 werden die zusammengefassten Ergebnisse aller fünf verschiedenen Arten der gebildeten Assoziationsketten der Methode 2 dargestellt:

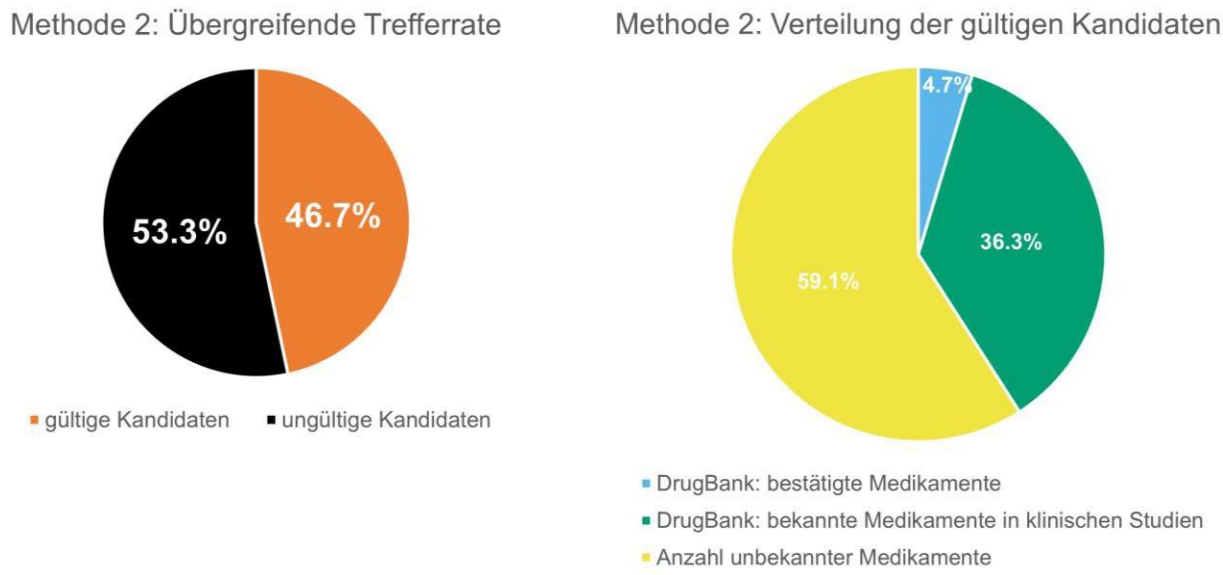


Abbildung 13: Zusammengefasste Ergebnisse der Methode 2 (eigene Grafik)

Trotz eines gezielten Ansatzes konnten die Assoziationsketten der Methode 2 mit einer übergreifenden Trefferrate von nur 46.7% und damit mehr als der Hälfte aller ermittelten Kandidaten als ungültige Treffer, ein eher ernüchterndes Resultat vorweisen. Dabei bestand die übergreifende Trefferrate aus einem Qualitätsfaktor von 19.1% und einem Potenzialfaktor von 27.6%. Die Verteilung der gültigen Kandidaten zeigt, dass auch vermehrt neue und in "DrugBank" unbekannte Medikamente identifiziert wurden.

Die Ergebnisse der Methode 2 zeigen dennoch, dass man auch mithilfe einer einfachen und schnellen Methode "Repurposing" Kandidaten aus Textdaten extrahieren kann. Weitere Erkenntnisse der Methode 2 werden in Kapitel 6.4 im Direktvergleich mit der Methode 1 genauer erläutert.

### 6.3.1 Methode 2: Vergleich der Assoziationsketten

**ZF 2:** Wie unterscheiden sich die Ergebnisse der verschiedenen Arten der Assoziationsketten?

Für die zweite zusätzliche Fragestellung wurden fünf unterschiedliche Arten von Assoziationsketten für das Fallbeispiel getestet und evaluiert. Für dies wurden die Arten auf Basis der unterschiedlichen Vorgehensweisen und theoretischen Ansätzen aufgebaut. Dabei wurden für jede Kette von der Fachliteratur empfohlene Datenbanken für die Extraktion der Startassoziationen (A-B) oder (B-C) ausgewählt.

| | "disease-gene-drug" | "disease-gene_variant–drug" | "disease-symptom-drug" | "disease-drug-sideeffect-drug" | "disease-drug-cell_line-drug" |
|---|---|---|---|---|---|
| **Anzahl aus Datenbanken extrahierte Assoziationspaare als Suchterme** | *118* | *291* | *24* | *200* | *68* |
| **Gesamtanzahl ermittelter Kandidaten** | 867 | 6 | 441 | 1267 | 36 |
| **Trefferrate** | **49.71%** | **16.67%** | **38.32%** | **47.67%** | **47.22%** |
| **Qualitätsfaktor** | 23.64% | 0.00% | 14.06% | 17.60% | 27.78% |
| **Potenzialfaktor** | 26.07% | 16.67% | 24.26% | 30.07% | 19.44% |

Tabelle 7: Methode 2 Assoziationsketten: Kennzahlen bezüglich Ergebnisse (eigene Tabelle)

Die Tabelle 7 bietet eine Übersicht zu den Kennzahlen zu den jeweilig unterschiedlichen Ergebnissen der getesteten Assoziationsketten. Dabei lassen sich besonders zwei Ausreisser erkennen. Sowohl die Assoziationskette zu den Genvarianten von Glioblastoma als auch die verlängerte Assoziationskette zu den Zelllinien der bestätigten Medikamente für die Behandlung von Glioblastoma konnten wenige bzw. keine Kandidaten ermitteln. Dies lässt vermuten, dass entweder diese Entitätsarten als Konzepte sehr selten in den klinischen Berichten oder PubMed-Abstracts erwähnt werden oder dass sie verstärkt durch eine informelle Sprache beschrieben werden. Dies verdeutlicht eine grosse Schwäche der Methode 2, dass sie verstärkt von einer bestimmten Syntax der Textdaten sowie der extrahierten Suchterme abhängig ist.

Aufgrund der geringen Treffermengen der beiden zuvor erwähnten Ketten konnten diese nicht repräsentativ mit den anderen Ketten verglichen werden. Bei der Betrachtung der Ergebnisse der restlichen drei Assoziationsketten liessen sich teilweise bei den Verhältnis Kennzahlen Trefferrate, Qualitätsfaktor und Potenzialfaktor leichte Ausreisser zu den übergreifenden Mittelwerten (TR 46.7%, QF 19.1%, PF 27.6%) feststellen. Von diesen konnte die Kette *"disease-symptom-drug"* die tiefste Trefferrate mit 38.32% (inklusive den niedrigsten Qualitätsfaktor und Potenzialfaktor) und damit den grössten Unterschied mit 8.38 Prozentpunkten zur übergreifenden Trefferrate der Methode 2 vorweisen. Im

Vergleich mit den Trefferraten der beiden anderen Ketten konnten daher auch signifikante Unterschiede *(Normalapproximation, zweiseitig α=0.05, p=0.00/0.00, N=1708/1308)* festgestellt werden.

Zwischen den Ketten *"disease-gene-drug"* und *"disease-drug-sideeffect-drug"* liess sich einzig ein signifikanter Unterschied *(Normalapproximation, zweiseitig α=0.05, p=0.00, N=2134)* zwischen den Qualitätsfaktoren feststellen. Dabei fiel besonders bei der Kette *"disease-drug-sideeffect-drug"* das deutlich überwiegende Verhältnis des Potenzialfaktors zum Qualitätsfaktor mit 30.07% zu 17.6% auf. Diese Beobachtung widerspricht den von Zhang et al. (2013, S. 1574) ermittelten Ergebnissen ihres entwickelten Vorhersagesystems, bei diesem sie durch die Erfassung von Überschneidungen von Medikamentennebenwirkungen den geringsten Anteil von "false positive – Repurposing Kandidaten" verzeichneten. Da aber im Rahmen dieser Forschungsarbeit die Qualitätsprüfung der Kandidaten nur durch den Abgleich der Medikamentenkandidaten mit "DrugBank" in Form des Qualitätsfaktors ermöglicht wird, kann dieser Widerspruch nur bedingt objektiv interpretiert werden.

Neben der Volltextsuche der erfassten Suchterme als Stringteile, wurde als zusätzlicher Vergleichspunkt der Ergebnisse der Assoziationskette *"disease-gene-drug"*, die Volltextsuche der erfassten Suchterme als Wörter oder Terme durchgeführt und die dadurch erfassten Teilergebnisse ausgewertet. Dabei wurden durch die Dokumententreffer, anstelle den 867 Kandidaten nur insgesamt 405 Kandidaten erfasst. Dabei konnten die Kandidaten eine Trefferrate von 56.54% vorweisen, welche 6.83 Prozentpunkte über der Trefferrate der Ergebnisse der Volltextsuche auf Basis der Stringteile lagen. Dies legt die Vermutung nahe, dass mit einer Volltextsuche anhand Terme oder Wörter, potenziell ungewollte Dokumententreffer verhindert und damit die Anzahl ungültiger Kandidaten reduziert werden kann.

Abschliessend konnte für die Methode 2 beobachtet werden, dass vorwiegend Gene, Nebenwirkungen und Symptome als Suchterme für die Volltextsuche in den ausgewählten Textdaten geeignet sind. Dabei wurde einzig bei der Kette *"disease-drug-sideeffect-drug"* die Suche auf Basis von Termen und Wörtern, anstelle von Stringteilen, zur qualitativen Einschränkung und potenziellen Optimierung der Ergebnisse durchgeführt.

## 6.4 Vergleich der Ergebnisse der Methode 1 und Methode 2

**ZF 3:** Wie unterscheiden sich die Ergebnisse der ungerichtete Methode 1 und der gerichteten Methode 2 (mit Vorwissen) und welche Probleme liessen sich beobachten?

Der Vergleich der visualisierten übergreifenden Ergebnisse beider Methoden durch die Abbildungen 12 & 13 zeigt, dass die Trefferraten beider Methoden sich mit einer Differenz von 9.2 Prozentpunkten signifikant unterscheiden *(Normalapproximation, zweiseitig α=0.05, p=0.00, N=5534)*. Jedoch bei den jeweiligen Verteilungen der gültigen Kandidaten konnten auch bei den grössten Anteilsdifferenzen (bei Anteil unbekannter Kandidaten) keine signifikanten Unterschiede erkannt werden *(Normalapproximation, zweiseitig α=0.05, p=0.34, N=2846)*. Dies zeigt, dass trotz der unterschiedlichen Trefferraten, sich die Komposition der gültigen Kandidaten sehr ähnlich oder gar identisch gestaltet.

Beide Methoden besassen allerdings diverse eigene Vor- und Nachteile. Durch die ungerichtete Art der Methode 1 müssen zwanghaft, um gezielte Resultate für ein Fallbeispiel ermitteln zu können, die richtigen und geeigneten Textdaten ausgewählt werden. Demnach ist es auch bei zukünftigen Forschungsvorhaben, welche das Ziel haben "Repurposing" Kandidaten zu einer spezifizierten Krankheit zu finden, unter Verwendung dieser Methode wichtig, dass auch Textdaten mit genügend Direktbezügen zur spezifizierten Krankheit in die zu untersuchenden Textdaten inkludiert werden. Diese nötige Vorzusammenstellung der Textdaten ist bei Methode 2 nicht für eine optimale Leistung erforderlich. Bei der Methode 2 muss dieser zusätzliche Bezug zum Fallbeispiel aufgrund der zuvor extrahierten Assoziationspaare (A-B) nicht mehr hergestellt werden und die Methode kann somit auf jeden beliebigen Textdatensatz direkt angewandt werden.

Gleichzeitig muss auch betont werden, dass besonders für die Verbesserung der Kookkurrenz Analyse bei der Methode 1, eine menschlich-gesteuerte Aufsicht ("Supervision") durch die Erstellung ergänzender Stoppwortlisten zur jeweiligen Optimierung der Ergebnisse stattfand. Für Methode 2 wurden meist lediglich die Suchterme mit ihren korrespondierenden Kleinbuchstabenform ergänzt oder bei manchen Ketten anhand ihrer von der Datenbank vorbestimmten Relevanz oder Qualität eingeschränkt. So kann eventuell die höhere übergreifende Trefferrate der Methode 1 im Vergleich zur Methode 2, auch durch diese zusätzlichen Eingriffe zur Optimierung erklärt werden. Daher können auch je nach Argumentation, die Resultate der Forschungsarbeit aus Methode 1 als nicht vollständig "ungerichtet" bezeichnet werden.

Als grösstes Problem und Herausforderung der Methode 2 konnte dessen starke Abhängigkeit zur Form der zu untersuchenden Textdaten festgestellt werden. Durch die Volltextsuche von vorbestimmten Suchtermen vorwiegend in formeller Sprache, können bspw. informelle Synonyme für diese Suchterme nicht erkannt werden. Gleichzeitig liess sich anhand der Resultate feststellen, dass spezifische Arten von Suchtermen wie z.B. Zelllinien oder Genvarianten nicht in den ausgesuchten Textdaten vorhanden waren. Daher können die Suchterme, je nach Form oder Schreibstil unterschiedlich gut für die zu untersuchenden Textdaten geeignet sein. Die Methode 1 kann hingegen aufgrund der Funktionsweise einer Kookkurrenz Analyse, sich an die vorhandenen Formen und Schreibstile der ausgewählten Textdaten besser anpassen.

Die einzige Art der menschlich-gesteuerten Aufsicht, welche die Methode 2 bietet, geschieht im Rahmen der Auswahl der Suchterme. Dabei konnten je nach Datenbank die Extraktion der (A-B) Startrelationen anhand zusätzlicher Kriterien, wie bspw. bei den Nebenwirkungen anhand Prozentwerte oder expliziten Beschreibungen wie "common", eingeschränkt werden. Als Test einer solchen uneingeschränkten Extraktion von Suchtermen, wurde für die Assoziationskette *"disease-gene-drug"* alle zu "Glioblastoma" assoziierten Gene, unabhängig ihrer vermeintlichen Qualität, als B-Suchterme aus der NLM-Datenbank extrahiert. Dabei konnten mithilfe dieser Suchterme insgesamt 2'278 Kandidaten ermittelt werden, im Gegensatz zu den 867 Kandidaten auf Basis der qualitativen Auswahl der Gene aus "OpenTargets". Durch die Verwendung solcher Einschränkungen kann vermutlich der Anteil "false positive – Repurposing Kandidaten" vermindert werden.

Abschliessend lässt sich anhand des Vergleiches der Resultate beider Methoden feststellen, dass das im Rahmen der Methode 2 verwendete Vorwissen aus einer "state-of-the-art" Datenbank zu keinen qualitativ besseren Ergebnissen führte. Dies verdeutlicht umso mehr die komplexe Art von Textdaten als Wissensbestände, da man auf das in den Textdaten vorhandene Wissen weder leicht zugreifen noch einfach für einen Workflow ergänzend nutzen kann. Die Methode 1 bietet mit einem zwar rechenintensiven Prozess die Möglichkeit, den gesamten Textdatensatz offen zu potenziellen Ähnlichkeiten zwischen allen vorhandenen Krankheiten innerhalb der Dokumente zu untersuchen. Mit Methode 2 können dagegen in einem ressourcensparenden Prozess effektiv und schnell neue "Repurposing" Kandidaten aus beliebigen Textdaten erfasst werden.

## 6.5 Beantwortung der forschungsleitenden Frage

"Wie können unstrukturierte Textdaten für die Ermittlung neuer "Drug Repurposing" Kandidaten nutzbar gemacht werden und wie können sie Datenbanken ergänzen?"

Für die Beantwortung der forschungsleitenden Frage wurden im Rahmen dieser Forschungsarbeit zwei unterschiedliche Methoden als Workflows entwickelt, welche erfolgreich "Drug Repurposing" Kandidaten für das ausgewählte Fallbeispiel ermitteln konnten. Dabei standen bei beiden Methoden, die verfügbaren biomedizinischen NER-Systeme und Modelle als Kernelemente im Mittelpunkt, welche die Ermittlung von biomedizinischen Entitäten aus den unstrukturierten Textdaten überhaupt ermöglichten. Durch die Nutzung eines schnell einsetzbaren "rule-based" NER-System mit "scispaCy" und den verfügbaren NER-Modellen konnten nachvollziehbare Auswahlen von "Repurposing" Kandidaten für das Glioblastom bestimmt werden. Im Rahmen der Methode 1 wurden für die Ermittlung neuer "Drug Repurposing" Kandidaten, die Ähnlichkeiten zwischen einzelnen Textdokumenten auf Basis der Wort-Vorkommen und Häufigkeiten ausgenutzt. Mit Methode 2 wurden durch die Nutzung von Vorwissen aus Datenbanken Suchterme extrahiert und damit gezielt Dokumente identifiziert, welche auf Basis dieses Vorwissens mit der ausgewählten Krankheit in Verbindung stehen. Aus den Inhalten dieser Dokumente wurden anschliessend Medikamente und Wirkstoffe als "Drug Repurposing" Kandidaten extrahiert.

Beide Methoden zeigten, dass durch eine geeignete Verarbeitung und Analyse von unstrukturierten Textdaten, den Datenbanken unbekannte "Drug Repurposing" Kandidaten ermittelt werden können. Somit haben unstrukturierte Textdaten das Potenzial Datenbanken zu ergänzen.

# 7 Fazit & Reflexion

Zum Abschluss soll in diesem Kapitel über die Masterthesis reflektiert und ein finales Fazit zur Bearbeitung der zusammengefassten Forschungsarbeit gegeben werden.

## 7.1 Festgestellte Grenzen der Forschungsarbeit & Fazit

Anhand der Ergebnisse der Forschungsarbeit konnte festgestellt werden, dass auch unstrukturierte Textdaten neben den vorhandenen Datenbanken, eine geeignete Wissensquelle für "Drug Repurposing" Vorhaben darstellen. Dabei konnte besonders der selbstformulierte Potenzialfaktor aus den Ergebnissen zeigen, dass aus den Textdaten mehrheitlich den Datenbanken bisher unbekannte "Drug Repurposing" Kandidaten ermittelt werden konnten. Dennoch konnten im Rahmen der Forschungsarbeit mehrere Grenzen und Einschränkungen festgestellt werden.

Während der Auswertung aller Resultate konnten die von Neumann et al. (2019, S. 319) erwähnten Dateninkonsistenzen der verschiedenen Datenbanken auch durch den Autor selbst festgestellt werden. So sind zum Beispiel, im Direktvergleich der Datenbanken "DrugBank" und "DrugCentral 2021" verschiedene bestätigte Medikamente für das Glioblastom verzeichnet. In "DrugCentral" sind insgesamt neun Medikamente verzeichnet, in "DrugBank" sind von diesen drei als bestätigte Medikamente und vier als "Drug Trials" Kandidaten dokumentiert. Die restlichen zwei Medikamente "Artesunate" und "Procarbazine" fehlen. Umgekehrt fehlt dagegen in "DrugCentral", das von "DrugBank" bestätigte Medikament "Nimotuzumab". Auch bei einer zusätzlichen Betrachtung der in "DrugBank" verzeichneten Medikamente, sind viele Wirkstoffe als sogenannte "Stub" Einträge vorhanden, bei diesen sie nur teilweise und nicht vollständig annotiert sind. Des Weiteren wurden auch im Rahmen der Auswertung weitere Wirkstoffe wie z.B. "Cisplatin", "Cixutumumab" und "Paracetamol" evaluiert, welche in keiner der beiden Datenbanken aufgeführt waren, jedoch aber nach eigenen Recherchen schon mehrfach in klinischen Tests zur Behandlung des Glioblastoms eingesetzt wurden. Diese Tatsachen zeigen umso mehr die Herausforderung des biomedizinischen Bereiches mit den stetig wachsenden Datenmengen mithalten zu können.

Wie allerdings auch bei der Extraktion aus den Datenbanken für Methode 2 beschrieben, gibt es viele unterschiedliche Ansätze der Datenbanken bezüglich der Sammlung und Erfassung der Daten. Die meisten Datenbanken legen einen verstärkten Wert auf die sorgfältige Kuration der bereitgestellten Daten, dadurch sie nicht jede neue Entdeckung

direkt in die Datenbank integrieren und damit Einträge verhindern können, welche in Vorhaben zu "false positive" "Repurposing" Kandidaten führen. Andere Datenbanken, wie bspw. die NLM bei der Genauswahl für "Glioblastoma", bieten eine uneingeschränkte Sammlung aller Gene, welche jemals mit der Krankheit assoziiert wurden.

Die grösste Grenze der Forschungsarbeit stellte aber die Möglichkeit einer objektiven qualitativen Bewertung der ermittelten "Repurposing" Kandidaten dar. Zu den ermittelten Kandidaten konnte keine abschliessende Empfehlung gegeben werden, welche spezifischen Kandidaten das grösste Potenzial besitzen. Das war auch der Tatsache geschuldet, dass für Bestimmung aller Kandidaten vorwiegend bis auf die Startassoziationspaare der Methode 2, unspezifizierte Assoziationsbeziehungen aus den Textdaten verwendet wurden. Diese Assoziationsbeziehungen wurden dabei nicht anhand ihrer genauen Zusammenhänge analysiert wie z.B. anhand ihrer möglichen Kausalitäten oder ihrer positiven oder negativen Zusammenhänge. Somit könnten auch viele der in der Forschungsarbeit ermittelten Kandidaten eine negative therapeutische Wirkung zum ausgewählten Fallbeispiel "Glioblastoma" haben und somit z.B. auch Wirkstoffe darstellen, welche die Zellteilung der betroffenen Krebszellen beschleunigen können.

Zusätzlich stellte vor allem die hohe Anzahl von verschiedenen Synonymen für die Identifizierung der einzelnen Medikamente eine Herausforderung für die Auswertung dar. Neben der wissenschaftlichen Bezeichnung eines Wirkstoffes wie z.B. für "Bevacizumab", wurden teilweise entweder der Name des kommerziellen Produktes "Avastin" oder gar auch alternative Codes oder Identifizierungsnummern wie "ABP-215" verwendet. Aufgrund der hohen ermittelten Kandidatenmengen durch die jeweiligen Methoden und des damit verbundenen Aufwands der jeweiligen Einzelerfassung der individuellen Wirkstoffe im Rahmen jeder einzelnen Auswertung, wurden Synonyme jeweils als Treffer gezählt. Darüber hinaus wurden im Rahmen der Kandidatenbereitstellung alle möglicherweise einzigartigen Kombinationstherapien in jeweils Einzelwirkstoffe umgewandelt. Damit fand für alle Kandidaten, welche in Form einer Kombinationstherapie erfasst wurden, ein Informationsverlust durch die Trennung der zuvor vorhandenen Assoziationen zwischen diesen Kandidaten statt. Als Beispiel eines solchen Informationsverlustes wurde im Rahmen der unbereinigten Kandidatenliste die Kombinationstherapie, bestehend aus "Capecitabine" und "Temozolomide" ("CAPTEM") erfasst, welche beide nach der Bereinigung nur noch als Einzelmedikamente in der Liste vorhanden waren und somit dieser zuvor vorhandene Zusammenhang verloren ging.

Weitere Begrenzungen dieser Forschungsarbeit bezogen sich auf die Auswahl des NER-Systems und der zu untersuchenden Textdaten. Durch die Auswahl von "scispaCy" standen eine begrenzte, aber dennoch akzeptable Anzahl an verwendbaren NER-Modellen zur Verfügung. Dabei mussten die jeweils zugehörigen Genauigkeiten dieser Modelle, welche teilweise im Direktvergleich zu anderen NER-Systemen schlechter waren, in Kauf genommen werden. Gleichzeitig wurde auch die Auswahl der Textdaten für das spezifizierte Fallbeispiel durch Zugriffsrechte oder der allgemeinen verfügbaren Form beschränkt. Klinische Textdaten stehen mehrheitlich als kurze Zusammenfassungen und nicht als ausführliche Volltexte zur Verfügung und sind somit in ihrem Schreibstil und ihrer möglichen Ausführungsform eingeschränkt. Daneben würde jedoch auch die Auswahl von Volltextdaten, durch die stark erhöhten nötigen Rechenzeiten eine weitere Herausforderung darstellen.

Trotz all dieser Einschränkungen kann als abschliessendes Fazit betont werden, dass mithilfe der entwickelten Methoden ein kompakter Überblick zu allen potenziell relevanten Wirkstoffen und Arzneimitteln rund um die ausgewählte Krankheit gebildet werden konnte. Viele der erfassten Wirkstoffe waren nicht in den vorhandenen "state-of-the-art" Datenbanken verzeichnet und konnten in der anschliessenden Auswertung erfolgreich als relevante experimentelle Wirkstoffe aus klinischen Studien identifiziert werden. Dies zeigt das allgemeine Potenzial der Analyse von Textdaten zur Verschaffung eines erweiterten Überblickes für das "Drug Repurposing" zu potenziell verfügbaren Medikamenten und Wirkstoffen, welche durch die Dateninkonsistenzen von Datenbanken möglicherweise unbeachtet und unentdeckt bleiben.

## 7.2 Weiterer Forschungsbedarf & Empfehlungen

Für Weiterentwicklung bzw. Anpassung der erstellten Methoden der Forschungsarbeit, gibt es viele verschiedene Richtungen mit viel Potenzial. Der Mangel an der Aussagekraft zur Qualität der unterschiedlichen ermittelten Kandidaten, könnte durch die Nutzung der Häufigkeiten der jeweiligen Medikamenten- und Wirkstoffvorkommen in Form eines Priorisierungssystems umgesetzt werden. Dazu müssten anhand leichter Anpassungen der Workflows, die erfassten Duplikate zukünftig nicht mehr entfernt werden und in einem erweiterten Ansatz das verfügbare UMLS-Konzeptmatching von "scispaCy" genutzt werden, um somit jede erfasste Entität einem UMLS-Konzept zuzuordnen. Damit könnten auch ansatzweise vorhandene Synonyme oder Abkürzungen, einheitlichen Konzepten zugeordnet werden. Abschliessend nach der Zuordnung aller Entitäten, könnten so die

erfassten Häufigkeiten der jeweiligen UMLS-Konzepte eine angenäherte qualitative Auskunft zur Relevanz der jeweils erfassten "Repurposing" Kandidaten bieten und somit für eine Priorisierung der ermittelten Kandidaten herangezogen werden. Besonders solche Priorisierungssysteme können den übergreifenden Erfolg von klinischen Medikamententests, durch die Ermöglichung einer Vorauswahl geeigneter Testkandidaten, unterstützen (Sun et al., 2022, S. 5).

Auf der anderen Seite könnten mit zusätzlich verfügbaren Rechenressourcen, umfangreichere und vor allem neuere "state-of-the-art" öffentlich verfügbare NER-Systeme wie "Stanza" oder auch kommerzielle Lösungen wie "Spark NLP for Healthcare" (johnsnowlabs.com), welche auch sogar "Stanza" überlegene spezialisierte biomedizinische NER-Modelle zur Verfügung stellen, zukünftig verwendet werden. Besonders mit den speziell für klinische Texte spezialisierten NER-Modellen, können zusätzliche Zusammenhänge und Inhalte aus den Daten extrahiert werden. Mit der allgemein erhöhten Leistung könnten damit auch parallel die Textdatenmengen vergrössert werden. Zusätzlich würden diese Lösungen mit integrierten ML-Pipelines die Möglichkeit bieten, eigene oder erweiterte Modelle auf eigenen Textdaten zu trainieren. Wie in Kapitel 5.2.3 vorgestellt, liegt vermutlich das grösste Zukunftspotenzial in hochoptimierten NER-Modellen, welche sich durch die Methoden des "Transfer Learning" in einem stets fortlaufenden Prozess mit neuem Wissen in Form von neuen Daten erweitern und optimieren sowie sich damit in ihrer Leistung stetig verbessern. Aufgrund der immer wachsenden unübersichtlichen Datenmengen im biomedizinischen Bereich sind zukünftig vor allem ML-Methoden gefragt, welche die "omics" Daten effektiv auswerten und in automatisierte lernende ML-Modelle integrieren können (Zhao & So, 2019, S. 235). Wie aber bei der Auswertung der Forschungsergebnisse festgestellt wurde, fehlen aktuell explizite Gold-Standards für eine mögliche objektive Bewertung und Kontrolle der entwickelten Voraussagesysteme für das "Drug Repurposing". Dies stellt besonders für die Entwicklung neuer ML-Modelle ein Problem dar, da die Modelle so nicht anhand eines Referenzdatensatzes in ihrer Qualität und möglichen Fehlern getestet sowie mit anderen ML-Modellen verglichen werden können (Yang et al., 2019, S. 10568).

Trotz der übergreifend wachsenden Datenmengen fehlen dennoch in manchen spezifischen Gebieten der Biomedizin, besonders aus dem klinischen Bereich, ausreichend problemspezifische und qualitativ hochwertige Daten (Issa et al., 2021, S. 139). Viele Daten zu von der Pharmaindustrie gesponserten klinischen Studien der Phase II-IV sind nicht öffentlich zugänglich, besonders Daten über Studien zu gescheiterten Wirkstoffen

(Pushpakom et al., 2019, S. 56). Viele klinische Studien sind dabei auch in vielen unterschiedlichen Sprachen weltweit dokumentiert, welche durch die aktuell vorwiegende sprachliche Einschränkung der biomedizinischen NLP-Werkzeuge auf Englisch und Chinesisch nicht genutzt werden können. So müssten zukünftig entweder für jede weitere Sprache eigene Systeme und Modelle entwickelt werden oder diese klinischen Texte über optimierte Übersetzungsalgorithmen in die englische Sprache überführt werden. Auch der besonders in der Medizin bedeutende Datenschutz der jeweiligen Studienteilnehmerinnen und Studienteilnehmern, trägt zur Verminderung der öffentlich zugänglichen Daten bei (Pushpakom et al., 2019, S. 55).

Wie im Rahmen dieser Arbeit festgestellt wurde, stellt die ergänzende Kombination von Wissen aus unstrukturierten Textdaten und den strukturierten Datenbanken für das "Drug Repurposing" eine grosse Herausforderung, dennoch aber eine Chance mit hohem Potenzial dar. Die Kombination dieser beiden Datendomänen ermöglicht es wahrheitsgetreue Beziehungen zwischen biomedizinischen Entitäten in den Textdaten zu bestätigen und jedoch auch gleichzeitig neue potenzielle Beziehungen anhand Textinhalten und Textmustern zu identifizieren (Gonzalez et al., 2016, S. 39).

Dennoch muss hier noch abschliessend erwähnt werden, dass im biomedizinischen Fachbereich viele Daten zu Medikamenten oder experimentellen Wirkstoffen zwar öffentlich zugänglich sind, aber die Medikamente und experimentellen Wirkstoffe selbst nicht (Pushpakom et al., 2019, S. 55). Trotz der grossen Chancen einer Wiederverwendung können viele dieser Wirkstoffe aus privatwirtschaftlichen oder patentrechtlichen Gründen nicht verwendet werden. So müssten gegebenenfalls branchenübergreifende Abkommen oder externe monetäre Subventionen zu einem möglichen Austausch dieser Stoffe geschaffen werden, um diese im Speziellen für die Behandlung von seltenen Krankheiten (wie bspw. Krebs) für "Drug Repurposing" Vorhaben überhaupt zugänglich zu machen (Pushpakom et al., 2019, S. 56).

## 7.3 Persönliche Reflexion & Schlusswort

Die für die Masterthesis bearbeiteten relevanten Fachbereiche der Biomedizin, Pharmakologie und der Bioinformatik erwiesen sich alle als sehr anspruchsvoll und komplex. Demnach verlangte die Forschungsarbeit eine zeitintensive Phase zur Einarbeitung in die verschiedenen Themen durch die Fachliteratur. Besonders die komplexen und teilweise unbekannten Zusammenhänge zwischen Medikamenten, Zellen, Genen, Proteinen und die zugehörigen Wechselwirkungen erforderten im Rahmen der Literatur-

recherche zusätzliche Aufmerksamkeit. Dennoch stoss der Autor auf eigene Grenzen zum Verständnis vieler Teilaspekte, besonders im Bereich der Pharmakokinetik. Trotzdem gestaltete sich das Einlesen in die Fachgebiete für den Autor als sehr interessant.

Bei der Nutzung der technischen Werkzeuge konnte auf sehr viel persönlichen Vorwissen aus dem Studium zurückgegriffen werden. Die grösste Herausforderung in diesem Bereich nahm die Installation der nötigen Bibliotheken für die verfügbare Maschine ein, da viele der getesteten Bibliotheken primär für Linux Systeme und nicht für Windows zur Verfügung standen.

Abschliessend möchte sich der Autor besonders für bei dem betreuenden Referenten Prof. Dr. Heiko Rölke für wertvolle und geschätzte Unterstützung in Form von ermöglichten Besprechungsterminen und des konstruktiven Feedbacks bedanken.

Rückblickend konnte die Masterthesis ohne grosse Hindernisse oder Zwischenfälle erarbeitet und verfasst werden.

# 8 Quellen & Literaturverzeichnis

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S. & Vollgraf, R. (2019). FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In W. Ammar, A. Louis & N. Mostafazadeh (Hrsg.), *Proceedings of the 2019 Conference of the North* (S. 54–59). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4010

Alaimo, S. & Pulvirenti, A. (2019). Network-Based Drug Repositioning: Approaches, Resources, and Research Directions. In Q. Vanhaelen (Hrsg.), *Computational Methods for Drug Repurposing* (S. 97–113). Springer New York.

Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P. & Zhavoronkov, A. (2016). Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Molecular pharmaceutics*, *13*(7), 2524–2530. https://doi.org/10.1021/acs.molpharmaceut.6b00248

Andronis, C., Sharma, A., Virvilis, V., Deftereos, S. & Persidis, A. (2011). Literature mining, ontologies and information visualization for drug repurposing. *Briefings in bioinformatics*, *12*(4), 357–368. https://doi.org/10.1093/bib/bbr005

Asghari, M., Sierra-Sosa, D. & Elmaghraby, A. S. (2022). BINER: A low-cost biomedical named entity recognition. *Information Sciences*, *602*, 184–200. https://doi.org/10.1016/j.ins.2022.04.037

Ashburn, T. T. & Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews. Drug discovery*, *3*(8), 673–683. https://doi.org/10.1038/nrd1468

Balasundaram, P., Kanagavelu, R., James, N., Maiti, S., Veerappapillai, S. & Karuppaswamy, R. (2019). Implementation of a Pipeline Using Disease-Disease Associations for Computational Drug Repurposing. In Q. Vanhaelen (Hrsg.), *Computational Methods for Drug Repurposing* (S. 129–148). Springer New York.

Beachy, S. H., Johnson, S. G., Olson, S. & Berger, A. C. (2014). *Drug Repurposing and Repositioning: Workshop Summary.* https://doi.org/10.17226/18731

Beltagy, I., Lo, K. & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. In K. Inui, J. Jiang, V. Ng & X. Wan (Hrsg.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (S. 3613–3618). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, *32*(Database issue), D267–70. https://doi.org/10.1093/nar/gkh061

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Buza, K., Peška, L. & Koller, J. (2020). Modified linear regression predicts drug-target interactions accurately. *PloS one*, *15*(4), e0230726. https://doi.org/10.1371/journal.pone.0230726

cadfem-medical.com. (2022). in silico Erklärung – Was bedeutet in silico eigentlich? https://cadfem-medical.com/de/glossar/in-silico/

cancerrxgene.org. (2022). *Home page – Cancerrxgene – Genomics of Drug Sensitivity in Cancer*. https://www.cancerrxgene.org/

Carrella, D., Napolitano, F., Rispoli, R., Miglietta, M., Carissimo, A., Cutillo, L., Sirci, F., Gregoretti, F. & Di Bernardo, D. (2014). Mantra 2.0: an online collaborative resource for drug mode of action and repurposing by network analysis. *Bioinformatics (Oxford, England)*, *30*(12), 1787–1788. https://doi.org/10.1093/bioinformat-ics/btu058

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J. & Tang, Y. (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS computational biology*, *8*(5), e1002503. https://doi.org/10.1371/journal.pcbi.1002503

Chiang, A. P. & Butte, A. J. (2009). Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clinical pharmacology and therapeutics*, *86*(5), 507–510. https://doi.org/10.1038/clpt.2009.103

Choudhury, C., Arul Murugan, N. & Priyakumar, U. D. (2022). Structure-based drug repurposing: Traditional and advanced AI/ML-aided methods. *Drug discovery today.* Vorab-Onlinepublikation. https://doi.org/10.1016/j.drudis.2022.03.006

ClinicalTrials.gov. (2022). *Home – ClinicalTrials.gov*. https://clinicaltrials.gov/ct2/home

clue.io. (2022). *About CLUE*. https://clue.io/about

Cohen, K. B., Verspoor, K., Fort, K., Funk, C., Bada, M., Palmer, M. & Hunter, L. E. (2017). The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation in the Biomedical Domain. In N. Ide & J. Pustejovsky (Hrsg.), *Springer eBook Collection Social Sciences. Handbook of Linguistic Annotation* (S. 1379–1394). Springer. https://doi.org/10.1007/978-94-024-0881-2_53

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018, 11. Oktober). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://arxiv.org/pdf/1810.04805

disgenet.org. (2022). *DisGeNET – a database of gene-disease associations*. https://www.disgenet.org/

docs.anaconda.com. (2022). *Anaconda Documentation – Anaconda documentation*. https://docs.anaconda.com/

docs.conda.io. (2022). *Conda – Conda documentation*. https://docs.conda.io/en/latest/

Dowden, H. & Munro, J. (2019). Trends in clinical success rates and therapeutic focus. *Nature reviews. Drug discovery*, *18*(7), 495–496. https://doi.org/10.1038/d41573-019-00074-z

drugcentral.org. (2022). *Drug Central 2021: Online drug Compendium-Database Update Oct 2021*. https://drugcentral.org/

Dudley, J. T., Deshpande, T. & Butte, A. J. (2011). Exploiting drug-disease relationships for computational drug repositioning. *Briefings in bioinformatics*, *12*(4), 303–311. https://doi.org/10.1093/bib/bbr013

Felini, M. J., Olshan, A. F., Schroeder, J. C., Carozza, S. E., Miike, R., Rice, T. & Wrensch, M. (2009). Reproductive factors and hormone use and risk of adult gliomas. *Cancer causes & control : CCC*, *20*(1), 87–96. https://doi.org/10.1007/s10552-008- 9220-z

Frei, J. & Kramer, F. (2021, 24. September). *GERNERMED – An Open German Medical NER Model*. https://arxiv.org/pdf/2109.12104

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gao, S., Kotevska, O., Sorokine, A. & Christian, J. B. (2021). A pre-training and self-training approach for biomedical named entity recognition. *PloS one*, *16*(2), 1–23. https://doi.org/10.1371/journal.pone.0246310

go.drugbank.com. (2022a). DrugBank Online | Database for Drug and Drug Target Info. https://go.drugbank.com/

go.drugbank.com. (2022b). *Glioblastoma Multiforme (GBM)* | *DrugBank Online*. https://go.drugbank.com/indications/DBCOND0046976#drugs

Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C. & Greene, C. S. (2016). Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. *Briefings in bioinformatics*, *17*(1), 33–42. https://doi.org/10.1093/bib/bbv087

Halatsch, M.-E., Kast, R. E., Karpel-Massler, G., Mayer, B., Zolk, O., Schmitz, B., Scheuerle, A., Maier, L., Bullinger, L., Mayer-Steinacker, R., Schmidt, C., Zeiler, K., Elshaer, Z., Panther, P., Schmelzle, B., Hallmen, A., Dwucet, A., Siegelin, M. D., Westhoff, M.-A., . . . Heiland, T. (2021). A phase Ib/IIa trial of 9 repurposed drugs combined with temozolomide for the treatment of recurrent glioblastoma: CUSP9v3. *Neuro-oncology advances, 3*(1), vdab075. https://doi.org/10.1093/noajnl/vdab075

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hodos, R. A., Kidd, B. A., Shameer, K., Readhead, B. P. & Dudley, J. T. (2016). In silico methods for drug repurposing and pharmacology. *Wiley interdisciplinary reviews. Systems biology and medicine, 8*(3), 186–210. https://doi.org/10.1002/wsbm.1337 hpo.jax.org. (2022). Human Phenotype Ontology. https://hpo.jax.org/app/

Hurle, M. R., Yang, L., Xie, Q., Rajpal, D. K., Sanseau, P. & Agarwal, P. (2013). Computational drug repositioning: from data to therapeutics. *Clinical pharmacology and therapeutics*, *93*(4), 335–341. https://doi.org/10.1038/clpt.2013.1

Ines Montani, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann, Jim Geovedi, Jim O'Regan, Maxim Samsonov, Duygu Altinok, György Orosz, Daniël de Kok, Søren Lind Kristiansen, Lj Miranda, Explosion Bot, Roman, Peter Baumgartner, Leander Fiedler, Richard Hudson, . . . Flusskind. (2022). *spaCy: Industrial-strength Natural Language Processing in Python [Computer software]*. Zenodo.

interpharma.ch. (2019). *Pharma-Markt Schweiz*. https://www.interpharma.ch/wp-content/uploads/2020/02/ly_iph.01.19.002_-_pharmamarkt_schweiz_2019_d_web-komprimiert.pdf

interpharma.ch. (2022a). Blogserie Patientenzugang, Teil 2: Der Zugang zu innovativen Medikamenten ist für Patientinnen und Patienten in der Schweiz massiv verzögert. https://www.interpharma.ch/blog/blogserie-patientenzugang-teil-2-der-zugang-zu-innovativen-medikamenten-ist-fuer-patientinnen-und-patienten-in-der-schweiz-massiv-verzoegert/

interpharma.ch. (2022b). *Klinische Phase*. https://www.interpharma.ch/themen/fuhrend-in-forschung-entwicklung/der-weg-eines-medikaments/klinische-phase-phase-i-ii-iii/

Issa, N. T., Stathias, V., Schürer, S. & Dakshanamurthy, S. (2021). Machine and deep learning approaches for cancer drug repurposing. *Seminars in Cancer Biology*, *68*, 132–142. https://doi.org/10.1016/j.semcancer.2019.12.011

Jansen, S. (4. Mai 2021). Who’s Who and What’s What: Advances in Biomedical Named Entity Recognition (BioNER). *Towards Data Science*. https://towardsdatascience.com/whos-who-and-what-s-what-advances-in-biomedical-named-entity-recognition-bioner-c42a3f63334c

Jiang, R., Banchs, R. E. & Li, H. (2016). Evaluating and Combining Name Entity Recognition Systems. In X. Duan, R. E. Banchs, M. Zhang, H. Li & A. Kumaran (Hrsg.), *Proceedings of the Sixth Named Entity Workshop* (S. 21–27). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2703

Jin, G. & Wong, S. T. C. (2014). Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug discovery today*, *19*(5), 637–644. https://doi.org/10.1016/j.drudis.2013.11.005

Jin, S., Niu, Z., Jiang, C., Huang, W., Xia, F., Jin, X., Liu, X. & Zeng, X. (2021). HeTDR: Drug repositioning based on heterogeneous networks and text mining. *Patterns*, *2*(8), 100307. https://doi.org/10.1016/j.patter.2021.100307

johnsnowlabs.com. (2022). Spark NLP for Healthcare | Award Winning Medical NLP | John Snow Labs.

jupyter-notebook.readthedocs.io. (2022). *The Jupyter Notebook — Jupyter Notebook 6.4.12 documentation.* https://jupyter-notebook.readthedocs.io/en/stable/

Kim, J.-D., Ohta, T., Tateisi, Y. & Tsujii, J. (2003). GENIA corpus--semantically annotated corpus for bioextmining. *Bioinformatics (Oxford, England)*, *19 Suppl 1*, i180-2. https://doi.org/10.1093/bioinformatics/btg1023

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y. & Collier, N. (2004). Introduction to the bio-entity recognition task at JNLPBA. In N. Collier, P. Ruch & A. Nazarenko (Hrsg.), *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications – JNLPBA '04* (S. 70). Association for Computational Linguistics. https://doi.org/10.3115/1567594.1567610

Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., Brower, A. M., Callahan, T. J., Chute, C. G., Est, J. L., Galer, P. D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., . . . Robinson, P. N. (2021). The Human Phenotype Ontology in 2021. *Nucleic acids research*, *49*(D1), D1207-D1217. https://doi.org/10.1093/nar/gkaa1043

Kroeger, P. (2009). *Analyzing grammar: An introduction* (5. Auflage). Cambridge Univ. Press.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S. & Golub, T. R. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science (New York, N.Y.)*, *313*(5795), 1929–1935. https://doi.org/10.1126/science.1132939

Langedijk, J., Mantel-Teeuwisse, A. K., Slijkerman, D. S. & Schutjens, M.-H. D. B. (2015). Drug repositioning and repurposing: terminology and definitions in literature. *Drug discovery today*, *20*(8), 1027–1034. https://doi.org/10.1016/j.drudis.2015.05.001

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H. & Kang, J. (2020). BioBERT: a pretrained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Letunic, I. (2022). *SIDER Side Effect Resource*. http://sideeffects.embl.de/

Li, F., Liu, W. & Yu, H. (2018). Extraction of Information Related to Adverse Drug Events from Electronic Health Record Notes: Design of an End-to-End Model Based on Deep Learning. *JMIR medical informatics*, *6*(4), e12159. https://doi.org/10.2196/12159

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C. & Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database : the journal of biological databases and curation*, *2016.* https://doi.org/10.1093/database/baw068

Lipinski, C. A. (2000). Drug-like properties and the causes of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods*, *44*(1), 235–249. https://doi.org/10.1016/S1056-8719(00)00107-6

Luo, Z.-H., Shi, M.-W., Yang, Z., Zhang, H.-Y. & Chen, Z.-X. (2020). pyMeSHSim: an integrative python package for biomedical named entity recognition, normalization, and comparison of MeSH terms. *BMC Bioinformatics*, *21*(1). https://doi.org/10.1186/s12859-020-03583-6

Martínez, V., Navarro, C., Cano, C., Fajardo, W. & Blanco, A. (2015). DrugNet: network-based drug-disease prioritization by integrating heterogeneous data. *Artificial intelligence in medicine*, *63*(1), 41–49. https://doi.org/10.1016/j.artmed.2014.11.003

Masoudi-Sobhanzadeh, Y., Omidi, Y., Amanlou, M. & Masoudi-Nejad, A. (2020). Drug databases and their contributions to drug repurposing. *Genomics*, *112*(2), 1087–1095. https://doi.org/10.1016/j.ygeno.2019.06.021

Mayers, M., Li, T. S., Queralt-Rosinach, N. & Su, A. I. (2019). Time-resolved evaluation of compound repositioning predictions on a text-mined knowledge network. *BMC bioinformatics*, *20*(1), 653. https://doi.org/10.1186/s12859-019-3297-0

Napolitano, F., Zhao, Y., Moreira, V. M., Tagliaferri, R., Kere, J., D'Amato, M. & Greco, D. (2013). Drug repositioning: a machine-learning approach through data integration. *Journal of cheminformatics, 5*(1), 30. https://doi.org/10.1186/1758-2946-5-30

ncbi.nlm.nih.gov. (2022a). *CXCR4 C-X-C motif chemokine receptor 4 [ Homo sapiens (human) ]*. https://www.ncbi.nlm.nih.gov/gene/7852

ncbi.nlm.nih.gov. (2022b). *MeSH – NCBI*. https://www.ncbi.nlm.nih.gov/mesh/

Neumann, M., King, D., Beltagy, I. & Ammar, W. (2019). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *Proceedings of the BioNLP 2019 workshop*, 319–327.

nlm.nih.gov. (2022a). *MEDLINE Overview*. https://www.nlm.nih.gov/medline/medline_overview.html

nlm.nih.gov. (2022b). *Unified Medical Language System (UMLS)*. https://www.nlm.nih.gov/research/umls/index.html

nltk.org. (2022). *NLTK : Natural Language Toolkit*. https://www.nltk.org/ opentargets.org. (2022). *Home – Open Targets*. https://www.opentargets.org/

Osier, N. D., Imes, C. C., Khalil, H., Zelazny, J., Johansson, A. E. & Conley, Y. P. (2017). Symptom Science: Repurposing Existing Omics Data. *Biological research for nursing*, *19*(1), 18–27. https://doi.org/10.1177/1099800416666716

Park, K. (2019). A review of computational drug repurposing. *Translational and clinical pharmacology*, *27*(2), 59–63. https://doi.org/10.12793/tcp.2019.27.2.59

pip.pypa.io. (2022). *pip documentation v22.2*. https://pip.pypa.io/en/stable/

Pradhan, S., Elhadad, N., South, B. R., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W. W. & Savova, G. (2015). Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association : JAMIA*, *22*(1), 143–154. https://doi.org/10.1136/amiajnl-2013-002544

Preuer, K., Lewis, R. P. I., Hochreiter, S., Bender, A., Bulusu, K. C. & Klambauer, G. (2018). DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics (Oxford, England)*, *34*(9), 1538–1546. https://doi.org/10.1093/bioinformatics/btx806

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., Norris, A., Sanseau, P., Cavalla, D. & Pirmohamed, M. (2019). Drug repurposing: progress, challenges and recommendations. *Nature reviews. Drug discovery*, *18*(1), 41–58. https://doi.org/10.1038/nrd.2018.168

Pyysalo, S., Ohta, T., Rak, R., Rowley, A., Chun, H.-W., Jung, S.-J., Choi, S.-P., Tsujii, J.'i. & Ananiadou, S. (2015). Overview of the Cancer Genetics and Pathway Curation tasks of BioNLP Shared Task 2013. *BMC Bioinformatics*, *16 Suppl 10*, S2. https://doi.org/10.1186/1471-2105-16-S10-S2

Richardson, L. (2007). *Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation*. https://beautiful-soup-4.readthedocs.io/en/latest/

Rodrigues, R., Duarte, D. & Vale, N. (2022). Drug Repurposing in Cancer Therapy: Influence of Patient's Genetic Background in Breast Cancer Treatment. *International journal of molecular sciences*, *23*(8). https://doi.org/10.3390/ijms23084280

Sarkar, D. (2019). *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing* (2nd ed.). Apress L. P. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5778375

Śniegula, A., Poniszewska-Marańda, A. & Chomątek, Ł. (2019). Study of Named Entity Recognition methods in biomedical field. *Procedia Computer Science*, *160*, 260–265. https://doi.org/10.1016/j.procs.2019.09.466

Soldaini, L. & Goharian, N. (2016). *QuickUMLS: a fast, unsupervised approach for medical concept extraction*. https://ir.cs.georgetown.edu/downloads/quickumls.pdf

Song, M., Kang, K. & Young An, J. (2018). Investigating drug-disease interactions in drug- symptom-disease triples via citation relations. *Journal of the Association for Information Science and Technology,* *69*(11), 1355–1368. https://doi.org/10.1002/asi.24060

statista.com. (2020). *Globale Pharmaindustrie: Statista-Dossier zum Thema globale Pharmaindustrie*. https://de.statista.com/statistik/studie/id/12440/dokument/globale-pharmaindustrie-statista-dossier/

Su, E. W. (2019). Drug Repositioning by Mining Adverse Event Data in ClinicalTrials.gov. In Q. Vanhaelen (Hrsg.), *Computational Methods for Drug Repurposing* (S. 61–72). Springer New York.

Sun, D., Gao, W., Hu, H. & Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B.* Vorab-Onlinepublikation. https://doi.org/10.1016/j.apsb.2022.02.002

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, *30*(1), 7–18. https://doi.org/10.1353/pbm.1986.0087

swissmedic.ch. (2022). *Zulassungen von Humanarzneimitteln mit neuem Wirkstoff und Indikationserweiterungen 2021*. https://www.swissmedic.ch/dam/swissmedic/de/dokumente/zulassung/zl_hmv_iv/zl-ham-nas-ie-2021.pdf.download.pdf/Swissmedic-NAS-IE-2021.pdf

Tanoli, Z., Seemab, U., Scherer, A., Wennerberg, K., Tang, J. & Vähä-Koskela, M. (2021). Exploration of databases and methods supporting drug repurposing: a comprehensive survey. *Briefings in bioinformatics*, *22*(2), 1656–1678. https://doi.org/10.1093/bib/bbaa003

van Rossum, G. (2022a). *re — Regular expression operations — Python 3.10.5 documentation*. https://docs.python.org/3/library/re.html

van Rossum, G. (2022b). *What is Python? Executive Summary*. https://www.python.org/doc/essays/blurb/

van Vleet, T. R., Liguori, M. J., Lynch, J. J., Rao, M. & Warder, S. (2019). Screening Strategies and Methods for Better Off-Target Liability Prediction and Identification of Small-Molecule Pharmaceuticals. *SLAS discovery : advancing life sciences R & D*, *24*(1), 1–24. https://doi.org/10.1177/2472555218799713

Vogt, I., Prinz, J. & Campillos, M. (2014). Molecularly and clinically related drugs and diseases are enriched in phenotypically similar drug-disease pairs. *Genome Medicine*, *6*(7), 52. https://doi.org/10.1186/s13073-014-0052-z

Wang, Y., Yella, J. & Jegga, A. G. (2019). Transcriptomic Data Mining and Repurposing for Computational Drug Discovery. In Q. Vanhaelen (Hrsg.), *Computational Methods for Drug Repurposing* (S. 73–95). Springer New York.

Weber, L., Münchmeyer, J., Rocktäschel, T., Habibi, M. & Leser, U. (2020). HUNER: improving biomedical NER with pretraining. *Bioinformatics (Oxford, England)*, *36*(1), 295–302. https://doi.org/10.1093/bioinformatics/btz528

Weber, L., Sänger, M., Münchmeyer, J., Habibi, M., Leser, U. & Akbik, A. (2021). HunFlair: An Easy-to-Use Tool for State-of-the-Art Biomedical Named Entity Recognition. *Bioinformatics (Oxford, England).* Vorab-Onlinepublikation. https://doi.org/10.1093/bioinformatics/btab042

Yang, H.-T., Ju, J.-H., Wong, Y.-T., Shmulevich, I. & Chiang, J.-H. (2017). Literature-based discovery of new candidates for drug repurposing. *Briefings in bioinformatics*, *18*(3), 488–497. https://doi.org/10.1093/bib/bbw030

Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. (2019). Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chemical reviews*, *119*(18), 10520–10594. https://doi.org/10.1021/acs.chemrev.8b00728

Zhang, P., Wang, F., Hu, J. & Sorrentino, R. (2013). Exploring the relationship between drug side-effects and therapeutic indications. *AMIA Annual Symposium Proceedings*, *2013*, 1568–1577.

Zhang, Y., Zhang, Y., Qi, P., Manning, C. D. & Langlotz, C. P. (2021). Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association : JAMIA*, *28*(9), 1892–1899. https://doi.org/10.1093/jamia/ocab090

Zhao, K. & So, H.-C. (2019). Using Drug Expression Profiles and Machine Learning Approach for Drug Repurposing. In Q. Vanhaelen (Hrsg.), *Computational Methods for Drug Repurposing* (S. 219–237). Springer New York.

Zhu, Y., Che, C., Jin, B., Zhang, N., Su, C. & Wang, F. (2020). Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. *Health informatics journal*, *26*(4), 2737–2750. https://doi.org/10.1177/1460458220937101

## Bisher erschienene Schriften

Ergebnisse von Forschungsprojekten erscheinen jeweils in Form von Arbeitsberichten in Reihen. Sonstige Publikationen erscheinen in Form von alleinstehenden Schriften.

Derzeit gibt es in den Churer Schriften zur Informationswissenschaft folgende Reihen:
Reihe Berufsmarktforschung

Weitere Publikationen

Churer Schriften zur Informationswissenschaft – Schrift 138
Herausgegeben von Wolfgang Semar
Mara Funaro
Ursachen für die geringe Verbreitung von Extreme Programming
Weshalb sich lediglich Praktiken der agilen Methode durchgesetzt haben
Chur, 2021
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 139
Herausgegeben von Wolfgang Semar
Debora Messerli
Nachhaltigkeitsprojekte in Bibliotheken
Massnahmenkatalog zur Vermittlung der UN-Agenda 2030 in Öffentlichen und Wissenschaftlichen Bibliotheken
Chur, 2021
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 140
Herausgegeben von Wolfgang Semar
Noemi Andres
Status quo des Social-Media-Einsatzes in Schweizer Tambouren-, Clairon- und Pfeifervereinen
Chur, 2021
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 141
Herausgegeben von Wolfgang Semar
Rachel Noëmi Thommen
Lärmmanagement an Deutschschweizer Hochschulbibliotheken
Evaluation der Wahrnehmung des Geräuschpegels von Studierenden in Hochschulbibliotheken und Einfluss von Covid-19
Chur, 2021
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 142
Herausgegeben von Wolfgang Semar
Daria Gloor
Berichterstattung von $CO_2$-Emissionen im Scope 3 des GHG Protocol
Eine Fallstudie zur Ableitung von digitalen Best Practices für Unternehmen zur Messung und Angabe von $CO_2$-Emissionen der Kriterien im Scope 3
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 143
Herausgegeben von Wolfgang Semar
Leonardo Personini
What role have academic libraries and librarians had in the fight against the COVID-19 pandemic?
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 144
Herausgegeben von Wolfgang Semar
Jasmin Suter
TikTok User sind einfacher manipulierbar
Einfluss von Videoplattformen auf das Verhalten in der Pre-Purchase Phase am Beispiel TikTok
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 145
Herausgegeben von Wolfgang Semar
Lea Bächli
Die Veränderungen der Angebote öffentlicher Bibliotheken in der Deutschschweiz durch die COVID-19-Pandemie
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 146
Herausgegeben von Wolfgang Semar
Jeffrey Santana de Jesus
Mithilfe von Digital Nudging mehr Privatsphäre in sozialen Netzwerken?
Digital Nudging in sozialen Netzwerken
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 147
Herausgegeben von Wolfgang Semar
Regina Eicher
Die Entwicklung inhaltlicher Sprachbegriffe für eine verbesserte Erschliessung von Kinder- und Jugendzeichnungen
Eine qualitative Inhaltsanalyse von 12 ausgewählten Märchen
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 148
Herausgegeben von Wolfgang Semar
Andrej Kilian
«Die Bibliotheksthematik hat sich in den letzten Jahren stark relativiert»
Interne Bibliotheken in der Deutschschweiz und in Liechtenstein – Versuch eines Einblicks
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 149
Herausgegeben von Wolfgang Semar
Sandra Freiburghaus
Untersuchung von Anzeige- und Reservationssystemen zur Lernplatzorganisation in Bibliotheken
Unter Betrachtung der Bedürfnisse und Erfahrungen der Institution
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 150
Herausgegeben von Wolfgang Semar
Lisa Heller
Zur Genese eines nationalen Bibliotheksprojekts: Swiss Library Service Platform (SLSP)
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 151
Herausgegeben von Wolfgang Semar
Antonin Friberg
Die Effektivität von Social Norms Nudging in der Customer Journey
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 152
Herausgegeben von Wolfgang Semar
Nicole Fässler
User Adoption bei der Einführung einer Kollaborations- und Kommunikationssoftware im Modern Workplace Umfeld
Chur, 2022
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 153
Herausgegeben von Wolfgang Semar
Antonin Friberg
Die Effektivität von Social Media Norms Nudging in der Customer Journey
Chur, 2022
ISSN 1660-945X

# Über die Informationswissenschaft der Fachhochschule Graubünden

Die Informationswissenschaft ist in der Schweiz noch ein relativ junger Lehr- und Forschungsbereich. International weist diese Disziplin aber vor allem im anglo-amerikanischen Bereich eine jahrzehntelange Tradition auf. Die klassischen Bezeichnungen dort sind Information Science, Library Science oder Information Studies. Die Grundfrage-stellung der Informationswissenschaft liegt in der Betrachtung der Rolle und des Um-gangs mit Information in allen ihren Ausprägungen und Medien sowohl in Wirtschaft und Gesellschaft. Die Informationswissenschaft wird in Chur integriert betrachtet.

Diese Sicht umfasst nicht nur die Teildisziplinen Bibliothekswissenschaft, Archivwissenschaft und Dokumentationswissenschaft. Auch neue Entwicklungen im Bereich Medienwirtschaft, Informations- und Wissensmanagement und Big Data werden gezielt aufgegriffen und im Lehr- und Forschungsprogramm berücksichtigt.

Der Studiengang Informationswissenschaft wird seit 1998 als Vollzeitstudiengang in Chur angeboten und seit 2002 als Teilzeit-Studiengang in Zürich. Seit 2010 rundet der Master of Science in Business Administration das Lehrangebot ab.

Das Forschungsfeld Informationswissenschaft vereinigt Cluster von Forschungs-, Entwicklungs- und Dienstleistungspotenzialen in unterschiedlichen Kompetenzzentren:

- Bibliothek und Digitalisierung von analogem Kulturgut
- Bildungsinformatik
- Data Analytics
- Digital Business and Usability Engineering
- Information Lifecycle Management
- Knowledge and User Research
- Practical Data Science
- Process Data, Visualization, and Machine Learning
- Scientific Computing

Diese Kompetenzzentren werden im Swiss Institute for Information Science (SII) zusammengefasst.

# Impressum


## Impressum

FHGR – Fachhochschule
Graubünden
Information Science
Pulvermühlestrasse 57
CH-7000 Chur

www.informationscience.ch
www.fhgr.ch
**ISSN 1660-945X**



## Institutsleitung

Prof. Dr. Ingo Barkow
Telefon: +41 81 286 24 61
Email: ingo.barkow@fhgr.ch

## Sekretariat

Telefon: +41 81 286 24 24
Fax: +41 81 286 24 00
Email: clarita.decurtins@fhgr.ch