Exploring the potential of Open Text Data for Drug Repositioning: A Case Study in Glioblastoma Therapy

Curdin Marxer and **Heiko Rölke**
DAViS / Pulvermühlestrasse 57
University of Applied / 7000 Chur
Sciences of the Grisons / Switzerland
curdin.marxer@fhgr.ch
heiko.roelke@fhgr.ch

Alex Alfieri and **Marc-Eric Halatsch**
Dep. of Neurosurgery / Brauerstrasse 15
Cantonal Hospital / 8400 Winterthur
of Winterthur / Switzerland
alex.alfieri@ksw.ch
marc-eric.halatsch@ksw.ch

Abstract

New drugs are risky and costly to develop. "Drug repositioning" or "drug repurposing" describes the well-known practice in identifying new uses for already existing drugs or active compounds. Using a case study, this paper describes ongoing research about the exploration of the potential in using NLP techniques on publicly available data sources to identify drugs for glioblastoma therapy not documented in established standardized databases.

1 Introduction

Developing and discovering new drugs is risky, costly and takes a long time. Many factors such as poor drug-properties like high drug toxicity, lack of effectiveness in their originally intended purpose (Dowden and Munro, 2019; Hodos et al., 2016) as well as bad absorption, distribution, metabolism, or excretion (ADME) (Lipinski, 2000) contribute to a rate below 10% for a new drug to successfully enter approved world-wide markets. Identifying and developing new uses for already known drugs or active compounds can be summarized under the concept often referred to as "drug repositioning" or "drug repurposing" (Ashburn and Thor, 2004). Drug development for rare diseases like some variants of cancer is especially devoid of commercial interest (Alaimo and Pulvirenti, 2019), for this reason the concept of "drug repositioning" plays a key role in battling rare diseases. Utilizing the results of previous research and existing knowledge about drugs, e.g. on their target molecules and mechanisms of action or their safety, completely new indications for drugs or active compounds can be discovered (Wang et al., 2019). Especially the exploitation of side-effects for already known drugs can be very lucrative, because many otherwise necessary costly development steps for market approval can be dispensed (Tanoli et al., 2021). As a consequence of the high failure rates for market entry of new medical compounds, a high quantity of drugs and chemical compounds are left abandoned. By utilizing "drug repositioning" their development and trial costs can be "rescued" (Langedijk et al., 2015). One example of such a successful rescued drug is "azido-thymidine", originally developed to treat cancer and shelved after determined inert. This drug was later rescued for the treatment of HIV and its prevention of vertical transmission (Reed, 2016). Knowledge on these drugs, mostly recorded in standardized databases, is also progressively available in unstructured text data. As a resulting problem of the constantly expanding volume of medical data in recent years as well as the rise in number of different repositories or databases, data from these databases or repositories differ significantly in terms of quality and reliability (Neumann et al., 2019; Tanoli et al., 2021). This results in a challenge for researchers in choosing the adequate database(s) containing the required information. Additionally, copious amounts of exclusive medical knowledge are hidden in scientific research documents or clinical reports as unstructured text data. As solution to these challenges, advances in the field of Natural Language Processing (NLP) enable researchers to identify possible relationships between many types of biomedical entities, such as drugs, diseases and genes within unstructured textual data to predict new candidates for repositioning (Alaimo and Pulvirenti, 2019; Andronis et al., 2011).

Glioblastoma (GBM), one of the most malignant types of cancer, was selected for our case study due to high clinical relevance. In a recent phase I clinical study, an innovative treatment for GBM termed CUSP9v3 was favorably tested for safety and tolerability (Halatsch et al., 2021). CUSP9v3 comprises a regimen of 9 repurposed non-oncological drugs combined with metronomic temozolomide. It demonstrated tumour growth inhibition ability and exemplifies the successful discovery of drugs for repurposing.

2 Related Research

2.1 Strategies for Drug Repositioning

To discover new "new target - known drug" pairs, the majority of strategies for drug repositioning use the theoretical foundations of network biology, systems biology and genomics (Choudhury et al., 2022). Alaimo and Pulvirenti (2019) outline these strategies in four different categories based on theory: ***Target-based*** approaches focus on the biological role of molecular target structures (e.g. genes, gene products, receptors, etc.) in diseases by using overlapping drug-targets or drug-target interactions to identify new repositioning candidates. ***Side-effect-based*** methods observe side effects of already developed drugs for possible alternative therapeutic uses by exploiting unintentional off-targets. This strategy however requires already existing clinical drug data and is therefore not suitable for active substances are shelved before clinical phases. ***Expression-based*** strategies utilize the key concepts of "signature reversion" or "signature matching" for genes. Using these concepts, new repositioning candidates can be predicted through quantitative molecular comparisons using gene expression profiles. If an associated drug-disease pair has anti-correlated gene expression profiles, thereby if a gene is disrupted through a disease, a drug with positive effect on that gene could be a potential therapeutic agent (Hodos et al., 2016; Issa et al., 2021). ***Similarity-based*** strategies utilize the idea that if two different diseases share at least one drug for their respective treatment, the rest of their not shared drugs could also be considered in joint treatment. Expanding this idea, the similarity between two different drugs can be predicted based on the culmination of multiple similarities in molecular target structures, side effects and chemical structure. Conversely, the similarities between diseases can be determined through shared treatment profiles or their semantic distance in ontologies (Chiang and Butte, 2009). Much knowledge about drugs and chemical compounds provided in biomedical texts is often only described by using vague indications, which can be harnessed by exploiting the guilt-by-association (GBA) principle proposed by Chiang and Butte (2009) to identify new potential drug candidates for repositioning.

2.2 NLP in Drug Repositioning

To extract new information from unstructured text data, complex information extraction algorithms, like Named Entity Recognition (NER) make it possible to identify biomedical concepts or entities such as drugs, chemical compounds, diseases and genes. NER represents an important key method in order to be able to keep up with the constant growth of newly discovered and defined concepts and entities from literature, such as new drugs or experimental active substances (Gao et al., 2021). NER systems such as "ScispaCy" (Neumann et al., 2019), "SparkNLP" (Kocaman and Talby, 2021) or "Stanza" (Zhang et al., 2021) enable the extraction of bio-medical entities from texts. Their provided annotation models are usually built through time-consuming and data-dependent training using ML or DL techniques. Their models are ready-to-use and can be quickly deployed on new unstructured texts or can even be trained via Transfer Learning (TL) to further improve their accuracy and sensitivity. Similar recent examples of research utilizing NER for drug repositioning range from identifying low-cost therapeutics for cancer through scraping PubMed abstracts (Subramanian et al., 2019) to the use of Social Media Mining to extract possible repositioning candidates for LONG-COVID (Koss and Bohnet-Joschko, 2022).

3 Our Research and Results

In our previous research endeavours we aimed to test and evaluate different approaches and methods to predict new drug repositioning candidates using NLP on open and publicly available unstructured text data. Based on the concerns of Tanoli et al. (2021) on the steadily growing data inconsistencies between the various available databases, we analysed the potential of unstructured text data to combat these database inconsistencies by filling possible data gaps. As case study, we selected the rare and malicious cancer glioblastoma (GBM). Our goal was to identify and predict new unknown reposition-able therapeutic drugs, not (yet) included in established databases. We employed two different methods on publicly available clinical and medical text data from PubMed (nlm.nih.gov, 2022) and ClinicalTrials.gov (2022). Especially ClinicalTrials is a valuable source for new knowledge that is often not yet provided by databases, e.g. on unknown side effects of individual drugs or drug-drug interactions (Su, 2019). As NER-system we chose ScispaCy v0.5.1 (Neumann et al., 2019) which provides fast, easy-to-use and robust biomedical NER-models. Despite hav-

ScispaCy NER model	Variations and used labels			
	Biomedical Entities	***Genes, genomes, gene products***	***Diseases, symptoms, side-effects***	***Cell-types, lines, components***
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

Table 1: Combined NER models and used labels of all Method 1 variations

Association chain type A-B-C-D	Entity relation type and used databases	
	A-B	***B-C***
"disease-gene-drug"	disease-gene from OpenTargets (Ochoa et al., 2022)	
"disease-gene_variant-drug"	disease-gene_variant from DisGeNET (Piñero et al., 2019)	
"disease–symptom–drug"	disease-symptom from Human Phenotype Ontology (HPO) (Köhler et al., 2021)	
"disease-drug-sideeffect-drug"	disease-drug from DrugBank (Wishart et al., 2006)	drug-sideeffect from SIDER (Letunic, 2022)
"disease-drug-cell_lines-drug"	disease-drug from DrugBank (Wishart et al., 2006)	drug-cell_lines from Genomics of Drug Sensitivity in Cancer (GDSC) (Yang et al., 2013)

Table 2: Chains of association, entity relation and used databases of Method 2 variations

ing poorer performance compared to other NER-systems, ScispaCy offers four different specialized NER models with a wide subject-specific biomedical scope. By using the controlled vocabulary thesaurus MeSH (Medical Subject Headings) we selected the broader concept to GBM "Neuroectodermal tumours" as the narrowing search term for our text data extraction. On July 12, 2022, 6,741 clinical studies from ClinicalTrials and the most relevant 3,259 abstracts from PubMeD were extracted using our search terms. We normalized the extracted text data using the stop word lists from "NLTK" (Bird et al., 2009) and by removing line breaks, multiple spaces, full stops, colons or commas using a regular expression pattern. For further supervision and improvements of the NER Tagger, a supplementary stop word list was curated.

3.1 Our Methods and Evaluation

For our first method we employed the "GBA principle" by Chiang and Butte (2009). Applying this "similarity-based" approach we carried out a co-occurrence analysis while utilizing NER. All biomedical entities not identified by the NER-models were removed from each document so that only entities such as drugs, diseases, symptoms, genes, etc. were used for the co-occurrence analysis. For this step we used different NER models from ScispaCy in multiple variations. After calculating and merging the texts of the most similar document pairs, we applied Chiang and Butte's "GBA principle" to predict our "drug repositioning" candidates by utilizing ScispaCy and its specialized NER models "BC5CDR" (Li et al., 2016) (for diseases, chemicals and drugs) and the model "BIONLP13CG" (Pyysalo et al., 2015) (for simple chemicals). Based on the theory of overlapping treatment profiles with regard to the "GBA principle", the assumption now applies for the merged texts that every drug tagged by the NER tagged is a potential drug for treatment or a "repositioning candidate" for each tagged disease. Related to the mentioned repurposing strategies, we tested four variations for this method utilizing the models and labels in combination shown in Table 1.

Our second method combined existing knowledge from state-of-the-art public available databases with the integrated knowledge of NER systems. By utilizing the open discovery process according to the ABC model by Swanson (1986), (A-B) starting association pairs were extracted from databases using available relations between biomedical entities. By using NER, we determined the biomedical entity pairs (B-C) in unstructured text data in order to predict new repositioning candidates with the transitive relation (A-C). While using this method, entity types of A, B and C as well as the length of the utilized association chains were varied in order to explore the potential the different repositioning approaches. After designating the biomedical entity types of A&B as starting pairs, we chose the most suitable database on the recommendations based on previous research by Tanoli et al. (2021). All available B-type entities of the (A-B) relations were extracted as search terms for a full-text search to determine all documents that contain at least one of these B-terms. At the final step, while using the available specialized NER models, all sought-after entities of type C

were extracted from the hit documents. Depending on the type and length of the association chain used, C represented repositioning candidate for A as a result. Table 2 shows all variations of chains of association we tested in our research.

For our case-study, we evaluated the approximate quality of our predicted repositioning candidates from Method 1 & 2 using the database DrugBank (Wishart et al., 2006) as core reference. DrugBank provides the most comprehensive state-of-the-art collection of drugs and chemical substances with reference to their possible uses and their current status in clinical studies (Jin et al., 2021). For our evaluation, all 346 individual drugs or chemical compounds, which are associated with GBM for therapy, were extracted on the 11th of January 2023 (go.drugbank.com, 2023). To further improve our results, all extracted drug repositioning candidates from both methods were matched against the external Unified Medical Language System (UMLS) (Bodenreider, 2004) knowledge base via the available ScispaCy concept matching pipeline "Entity Linker". We devised three categories to approximately evaluate the quality of our results: For a drug repositioning candidate to be classified as valid, it had to be either a chemical element or compound, generic or brand name of a drug, a possible treatment method like *TT-Fields*, experimental vaccine or drug-specific antibody. The ***"Invalid"*** category comprised of all candidates which were deemed as invalid, e.g. unspecific generic terms like "antibody" or "acid". All candidates which were also confirmed to be possible therapeutics by DrugBank, either by being approved drugs or drugs in current ongoing drug trials for GBM, were assigned to the category ***"Known in DrugBank"***. Other valid candidates, which were not found in DrugBank, were allocated to ***"Unknown candidates"*** and represent the body of knowledge which could supplement the database. The size of the "Invalid" category demonstrates the general ability to extract valid drugs or chemical compounds via our overall methodological efforts. "Known in DrugBank" together with "Unknown candidates" embody all possible repurposing candidates for GBM. To estimate the quality of our drug candidates for repositioning, the ratio between the categories "Known in DrugBank" and "Unknown candidates" can be used an approximate indication on how realistically truthful our results are. The smaller the ratio from "Known in DrugBank" to "Unknown candidates", presumably the higher the number of false positive repurposing candidates in the "Unknown candidates" category.

Variation	Total number of extracted entities as candidates	3 most extracted candidates as "Known in DrugBank"	3 most extracted candidates as "Unknown candidates"
"Biomedical Entities"	2734	"temozolomide" "carmustine" "cyclophosphamide"	"arsenic trioxide" "arsenic" "selumetinib"
"Genes, genomes, gene products"	6865	"Camptothecin-11" "Avastin" "Temodar"	"cisplatin" "anthracyclines" "Maleic acid"
"Diseases, symptoms, side-effects"	3898	"temozolomide" "bevacizumab" "nivolumab"	"adrenal cortex hormones" "tremelimumab" "ict-107"
"Cell-types, lines, components"	22025	"temozolomide" "irinotecan" "vincristine"	"cisplatin" "tremelimumab" "vasopressin"

Table 3: Results of all variations of Method 1

Variation	Total number of extracted entities as candidates	3 most extracted candidates as "Known in DrugBank"	3 most extracted candidates as "Unknown candidates"
"disease-gene-drug"	2226	"temozolomide" "erlotinib" "vincristine"	"cisplatin" "octreotide" "dacarbazine"
"disease-gene_variant-drug"	8	No results	"amifostine" "cisplatin" "glutathione"
"disease-symptom-drug"	975	"temozolomide" "vincristine" "etoposide"	"steroids" "2,6-dinitrotoluene" "cisplatin"
"disease-drug-sideeffect-drug"	7021	"Camptothecin-11" "Avastin" "Temodar"	"cisplatin" "ifosfamide" "melphalan"
"disease-drug-cell_line-drug"	47	"temozolomide" "docetaxel" "interferon alfa-2b"	"cisplatin" "baccatin III" "calcitonin"

Table 4: Results of all variations of Method 2

3.2 Results & Discussion

In our results for GBM, all valid extracted candidates for repositioning are either chemical elements, chemical compounds, experimental vaccines, hormones or other various therapeutics. Table 3 and Table 4 show an excerpt of the results for each tested method and variation, with the total number of extracted entities with the three most occurring repurposing candidates known and unknown to DrugBank.

In the summarized results of all used variations for Method 1, 43.9% of extracted entities are categorized as "Known in DrugBank", 18.2% as "Unknown candidates" and 38.0% as "Invalid". For Method 2 and its variations, 39.4% are allocated to the "Known in DrugBank", 25.2% to the "Unknown candidates" and 35.5% to the "Invalid" category. The smallest difference in the results of both methods is observed in the "Invalid" ratio which suggests that our utilized NER-models perform similar in accuracy. The biggest difference is noticed between the ratios of "Unknown candi-

dates", which could imply a lesser quality of the provided drug repositioning candidates of Method 2, but also a higher proportion of previously unknown candidates with a possible high potential to combat existing inconsistencies in DrugBank. In summary, both methods prove to be able to identify new drug repositioning candidates while still upholding a representative amount of candidates known to DrugBank. The successful extraction of recent trial therapeutics for GBM, e.g. *tasadenoturev* (dnatrix.com, 2022), shows the great potential of unstructured text data for filling potential gaps in databases. Contrary, some candidates known to DrugBank are missing in our results, e.g. the anti-tumor agent *abemaciclib*.

4 Limitations & further Research

One great limitation of our research is that we used mostly unspecified association relationships between the entities from text data, with the exception of the start association pairs of Method 2, to predict our repositioning candidates. Most associations are not analyzed based on their exact semantic connections, such as their possible causalities as well as their positive or negative relationships. Thus, many of the candidates identified can also have an effect in promoting the tumor being false positives. Also, ScispaCy only provides a limited number of specialized NER models with lower accuracy than other available models from e.g. "Stanza" (Zhang et al., 2021) or "SparkNLP" (Kocaman and Talby, 2021).

In our future reserach we will employ our methods on clinical or medical full-texts from PubMed Central in addition to clinical studies from ClinicalTrials using a much more expanded data set. To enhance the quality of our methods, more sophisticated NER models from SparkNLP will be considered. Furthermore, an additional evaluation of our results by experts will be included.

References

Salvatore Alaimo and Alfredo Pulvirenti. 2019. Network-based drug repositioning: Approaches, resources, and research directions. In Quentin Vanhaelen, editor, *Computational Methods for Drug Repurposing*, pages 97–113. Springer New York, New York, NY.

Christos Andronis, Anuj Sharma, Vassilis Virvilis, Spyros Deftereos, and Aris Persidis. 2011. Literature mining, ontologies and information visualization for drug repurposing. *Briefings in bioinformatics*, 12(4):357–368.

Ted T. Ashburn and Karl B. Thor. 2004. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews. Drug discovery*, 3(8):673–683.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(Database issue):D267–70.

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

Chinmayee Choudhury, N. Arul Murugan, and U. Deva Priyakumar. 2022. Structure-based drug repurposing: Traditional and advanced ai/ml-aided methods. *Drug discovery today*.

ClinicalTrials.gov. 2022. Home - clinicaltrials.gov.

dnatrix.com. 2022. Dnatrix is developing potent oncolytic immunotherapies for the treatment of cancer using genetically modified viruses.

Helen Dowden and Jamie Munro. 2019. Trends in clinical success rates and therapeutic focus. *Nature reviews. Drug discovery*, 18(7):495–496.

Shang Gao, Olivera Kotevska, Alexandre Sorokine, and J. Blair Christian. 2021. A pre-training and self-training approach for biomedical named entity recognition. *PloS one*, 16(2):1–23.

go.drugbank.com. 2023. Glioblastoma multiforme (gbm) | drugbank online.

Marc-Eric Halatsch, Richard E Kast, Georg Karpel-Massler, Benjamin Mayer, Oliver Zolk, Bernd Schmitz, Angelika Scheuerle, Ludwig Maier, Lars Bullinger, Regine Mayer-Steinacker, Carl Schmidt, Katharina Zeiler, Ziad Elshaer, Patricia Panther, Birgit Schmelzle, Anke Hallmen, Annika Dwucet, Markus D Siegelin, Mike-Andrew Westhoff, Kristine Beckers, Gauthier Bouche, and Tim Heiland. 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.

Rachel A. Hodos, Brian A. Kidd, Khader Shameer, Ben P. Readhead, and Joel T. Dudley. 2016. In silico methods for drug repurposing and pharmacology. *Wiley interdisciplinary reviews. Systems biology and medicine*, 8(3):186–210.

Naiem T. Issa, Vasileios Stathias, Stephan Schürer, and Sivanesan Dakshanamurthy. 2021. Machine and deep learning approaches for cancer drug repurposing. *Seminars in Cancer Biology*, 68:132–142.

Shuting Jin, Zhangming Niu, Changzhi Jiang, Wei Huang, Feng Xia, Xurui Jin, Xiangrong Liu, and Xiangxiang Zeng. 2021. Hetdr: Drug repositioning based on heterogeneous networks and text mining. *Patterns*, 2(8):100307.

Veysel Kocaman and David Talby. 2021. Spark nlp: Natural language understanding at scale. *Software Impacts*, 8:100058.

Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C. Carmody, David Lewis-Smith, Nicole A. Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M. Brower, Tiffany J. Callahan, Christopher G. Chute, Johanna L. Est, Peter D. Galer, Shiva Ganesan, Matthias Griese, Matthias Haimel, Julia Pazmandi, Marc Hanauer, Nomi L. Harris, Michael J. Hartnett, Maximilian Hastreiter, Fabian Hauck, Yongqun He, Tim Jeske, Hugh Kearney, Gerhard Kindle, Christoph Klein, Katrin Knoflach, Roland Krause, David Lagorce, Julie A. McMurry, Jillian A. Miller, Monica C. Munoz-Torres, Rebecca L. Peters, Christina K. Rapp, Ana M. Rath, Shahmir A. Rind, Avi Z. Rosenberg, Michael M. Segal, Markus G. Seidel, Damian Smedley, Tomer Talmy, Yarlalu Thomas, Samuel A. Wiafe, Julie Xian, Zafer Yüksel, Ingo Helbig, Christopher J. Mungall, Melissa A. Haendel, and Peter N. Robinson. 2021. The human phenotype ontology in 2021. *Nucleic acids research*, 49(D1):D1207–D1217.

Jonathan Koss and Sabine Bohnet-Joschko. 2022. Social media mining of long-covid self-medication reported by reddit users: Feasibility study to support drug repurposing. *JMIR Form Res*, 6(10):e39582.

Joris Langedijk, Aukje K. Mantel-Teeuwisse, Diederick S. Slijkerman, and Marie-Hélène D. B. Schutjens. 2015. Drug repositioning and repurposing: terminology and definitions in literature. *Drug discovery today*, 20(8):1027–1034.

Ivica Letunic. 2022. Sider side effect resource.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database : the journal of biological databases and curation*, 2016.

Christopher A. Lipinski. 2000. Drug-like properties and the causes of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods*, 44(1):235–249.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. *Proceedings of the BioNLP 2019 workshop*, pages 319–327.

nlm.nih.gov. 2022. Medline overview.

David Ochoa, Andrew Hercules, Miguel Carmona, Daniel Suveges, Jarrod Baker, Cinzia Malangone, Irene Lopez, Alfredo Miranda, Carlos Cruz-Castillo, Luca Fumis, Manuel Bernal-Llinares, Kirill Tsukanov, Helena Cornu, Konstantinos Tsirigos, Olesya Razuvayevskaya, Annalisa Buniello, Jeremy Schwartzentruber, Mohd Karim, Bruno Ariano, Ricardo Esteban Martinez Osorio, Javier Ferrer, Xiangyu Ge, Sandra Machlitt-Northen, Asier Gonzalez-Uriarte, Shyamasree Saha, Santosh Tirunagari, Chintan Mehta, Juan María Roldán-Romero, Stuart Horswell, Sarah Young, Maya Ghoussaini, David G Hulcoop, Ian Dunham, and Ellen M McDonagh. 2022. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Research*, 51(D1):D1353–D1359.

Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. 2019. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Jun'ichi Tsujii, and Sophia Ananiadou. 2015. Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013. *BMC Bioinformatics*, 16 Suppl 10:S2.

Michael D. Reed. 2016. The Rescue and Repurposing of Pharmaceuticals: Augmenting the Drug Development Paradigm. *The Journal of Pediatric Pharmacology and Therapeutics*, 21(1):4–6.

Eric Wen Su. 2019. Drug repositioning by mining adverse event data in clinicaltrials.gov. In Quentin Vanhaelen, editor, *Computational Methods for Drug Repurposing*, pages 61–72. Springer New York, New York, NY.

Shivashankar Subramanian, Ioana Baldini, Sushma Ravichandran, Dmitriy A. Katz-Rogozhnikov, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Kush R. Varshney, Annmarie Wang, Pradeep Mangalath, and Laura B. Kleiman. 2019. Drug repurposing for cancer: An nlp approach to identify low-cost therapies.

D. R. Swanson. 1986. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18.

Ziaurrehman Tanoli, Umair Seemab, Andreas Scherer, Krister Wennerberg, Jing Tang, and Markus Vähä-Koskela. 2021. Exploration of databases and methods supporting drug repurposing: a comprehensive survey. *Briefings in bioinformatics*, 22(2):1656–1678.

Yunguan Wang, Jaswanth Yella, and Anil G. Jegga. 2019. Transcriptomic data mining and repurposing for computational drug discovery. In Quentin

Vanhaelen, editor, *Computational Methods for Drug Repurposing*, pages 73–95. Springer New York, New York, NY.

David S. Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. 2006. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(Database issue):D668–72.

Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A. Smith, I. Richard Thompson, Sridhar Ramaswamy, P. Andrew Futreal, Daniel A. Haber, Michael R. Stratton, Cyril Benes, Ultan McDermott, and Mathew J. Garnett. 2013. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(Database issue):D955–61.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D. Manning, and Curtis P. Langlotz. 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.