SPECIFIC AIMS

Nearly 6 million COVID19 cases and over 362,000 deaths have been reported worldwide as of this writing. No effective drugs have yet been approved to treat the disease. The situation has hit the minority population especially hard. Current experimental drugs such as Hydroxychloroquine and Remdesivir are not very efficacious. Ongoing strategies include quick repurposing of existing drugs. However, most of these efforts were based on non-systematic mining of research on related drugs/diseases. ***We hypothesize that a systematic survey of PubMed with the aid of AI-driven text mining technology would afford a better yield for potential drugs against COVID19*** based on our recent successes of using it for IBC (inflammatory breast cancer) and P53-interacting kinases. Here, we propose to adopt the AI-driven text mining technology *Word2Vec* (first developed by Google's AI team for general text analysis) and a machine learning tool called Support Vector Machine (SVM) to analyze the biomedical literature to discover hidden associations among known drugs, targets and diseases, and to identify new candidates for COVID19. This goal will be achieved via three Specific Aims as follows.

Specific Aim 1 - **Collecting, organizing and analyzing PubMed abstracts with the Word2Vec technology.** Research abstracts containing drugs, diseases, targets and pathways will be downloaded, with publications on COVID19 labeled for future identification. This textual corpus will be preprocessed for the Word2Vec program. Word2Vec analysis will result in a numerical representation of all biological concepts (drugs, diseases and targets) as high dimensional vectors. The dimensionality and textual window parameters of Word2Vec will be optimized by visualizing the resulting word vectors on t-SNE generated scatter plots for drugs, diseases and their targets to ensure similar drugs, similar diseases are clustered close together, respectively. The optimized word vectors for drugs, diseases and targets will be used in Specific Aim 2 for AI/SVM-based model development and validation.

Specific Aim 2 - Developing AI/SVM-based models for predicting drug-disease and drug-target associations. Known drug-disease and drug-target pairs will be collected from DrugBank, KEGG and other chemical biology databases to expand our own existing collection of 900 drug-disease pairs and 9000 drug-target pairs. These two sets will be used as the training sets for AI/SVM-based model development. The SVM algorithm will be employed to build two sets of binary classification models to be able to distinguish true drug-disease (DD') pairs from false drug-disease (DD') pairs as well as true drug-target (DT) pairs from false drug-target (DT) pairs, respectively. The models will first be validated using test sets (20% of the left-out DD' pairs and DT pairs) to ensure the accuracy of models; they will then be used to predict known drugs against known diseases as well as known drugs against targets to generate a comprehensive relationship map, i.e., a predicted drug-disease-target (DD'T) network *knowledge graph*. This graph will be analyzed in Aim 3, and will be shared with the COVID19 research community.

Specific Aim 3 - Extracting and validating the predicted DD' pairs and DT pairs relevant to COVID19. The above AI/SVM-model predicted drug-disease (DD') pairs and drug-target (DT) pairs contain information and data regarding COVID19. We will extract these relationship pairs and validate them in the PubMed literature for viable mechanisms of action against COVID19 as well as other relevant descriptions in the literature including *clinicaltrials.gov*. For example, drug-disease and drug-target pairs related to sigma 1 receptor, cathepsin L, IL6 & IL10 modulators and JAK inhibitors (related to cytokine storm) will be further examined because of literature reports on their potential uses against COVID19. A diverse set of predicted candidates will also be examined to complement existing ones.

At the end of this *in silico* drug repurposing project, we will deliver a list of drugs (including biologics) for *repurposing & expansion* against COVID19 as well as a predicted drug-disease-target (DD'T) knowledge graph for the COVID19 research community to pursue further research of potential therapies if they so desire. Our future plans include acquiring compounds for testing against some of the receptors (e.g., sigma 1 receptor) or enzymes (e.g., JAK kinase) at the BRITE screening center (see Dr. Navarro's Letter of Support) to follow up on the proposed compounds. We also plan to submit proposed drugs to the NIH/NCATS Center (PAR-18-462), which solicits drug repurposing hypotheses based on computational algorithms.

RESEARCH STRATEGY

Current experimental drugs such as Hydroxychloroquine and Remdesivir are not sufficiently efficacious - some with severe side effects [1,2]. Ongoing strategies for finding other therapies include quick repurposing of existing drugs, such as known viral entry inhibitors [3], viral replication inhibitors [4], cytokine storm modulators [5], and even anti-HIV drugs [6]. Most of these efforts were based on non-systematic literature mining of past research on related drugs/diseases. We propose to adopt the AI-driven text mining technology called Word2Vec [7] to analyze the biomedical literature to derive multi-dimensional vector representations of concepts (drugs, diseases and targets) that maintain the similarity relationships among drugs that have similar uses, targets that are involved in the same pathways as well as diseases that have similar symptoms or phenotypes. This special function of Word2Vec allows us to discover hidden associations among known drugs, targets and diseases and their cross-associations. We will employ a machine learning tool called Support Vector Machine (SVM) [8,9] to capture the associations in order to predict new candidate drugs/targets for COVID19. We name our approach LWAS for *Literature-Wide Association Analysis.* The overall workflow of this project is depicted in **Figure 1.**

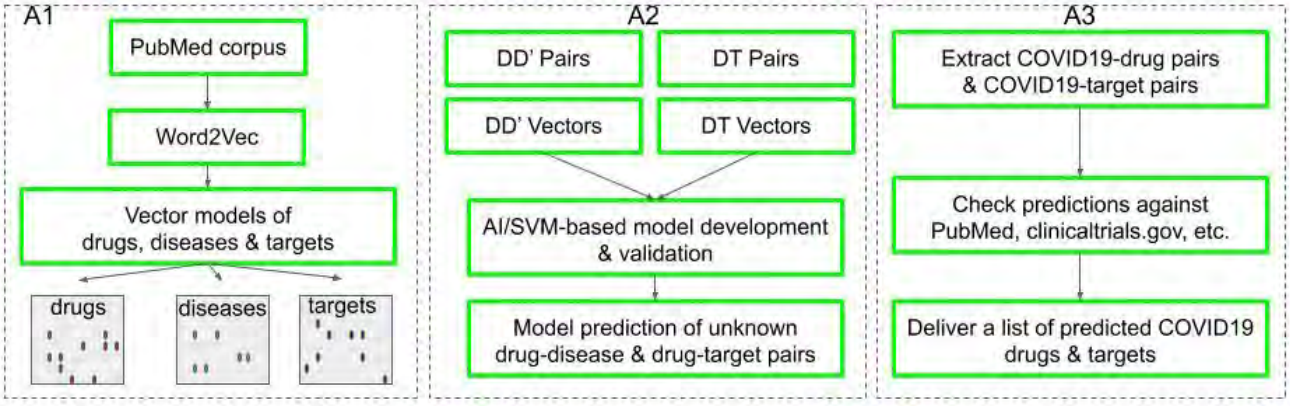


Figure 1. Overall workflow of this project (LWAS) from A1, A2 to A3.

Research Strategy (Specific Aim 1) - Collecting, organizing and analyzing PubMed abstracts with the Word2Vec technology.

Computational Methods and Design

Step 1 - Obtain textual corpus: search the PubMed literature for research abstracts that involve human diseases, drugs and biological targets (cf. Figure 1, A1). Abstracts will include recent publications with keywords COVID19, SARS-CoV-2, nCoV and coronavirus, etc. All abstracts related to COVID19 will be labeled as COVID19 for future analysis. PubMed ID (PMID) will also be included in the download for future tracking to original papers. Download will be saved as XML format.

Step 2 - Prepare Word2Vec input: preprocess the above XML file to remove symbols that are not needed for Word2Vec analysis. All words will be converted to upper case to ensure consistency across all abstracts. PMID will always be attached to the abstracts for book-keeping.

Step 3 - Train and optimize Word2Vec vector models. The preprocessed XML file will be used to train the Word2Vec program to derive vectors for drugs, diseases and targets. We will extract FDA-approved drugs (their names) from DrugBank [10], disease names from KEGG [11] and target names from KEGG as well. Other disease ontology databases will be utilized for diseases as well [12]. This analysis will result in a numerical representation of all biological concepts (drugs, diseases and targets) as high dimensional vectors. We will explore the dimensionality parameter (=10, 30, 50, 70, 90, ...150) and contextual window size (=5,10) to ensure optimal models which will be examined in Step 4. The Word2Vec program was from the *Gensim* package [13, 14].

Step 4 - Visualization by t-SNE map. For each of the Word2Vec vector models developed in Step 3, we will conduct a t-SNE [15] analysis using the *SciKit Learn package* [16] to obtain a 2D scatter plot for drugs, targets and diseases, respectively (cf. Figure 1, A1). We will examine the plots to ensure similar drugs, similar diseases and similar targets are clustered together, respectively. Based on the clustering results, we will choose the vector model(s) for Specific Aim 2 as the basis for AI/SVM-model development and validation.

Preliminary Data

Drug repurposing is about crosslinking related as well as seemingly unrelated diseases, targets and drugs together and making unexpected findings. To do so, we have to employ systematic & automatic analysis of ALL data. Our group has been studying this in the past three years and several manuscripts were submitted /finalized. For example, 3.8M PubMed abstracts were resulted from a search with keyword "cancer" covering research from 1787 to 2019. Shown below are t-SNE generated scatter plots for drugs and diseases. Another study of 649 kinases for their potential to interact with tumor suppressor P53 based on Word2Vec and kNN (a k-Nearest Neighbor machine learning technique) has enriched the discovery rate by 400% [17].

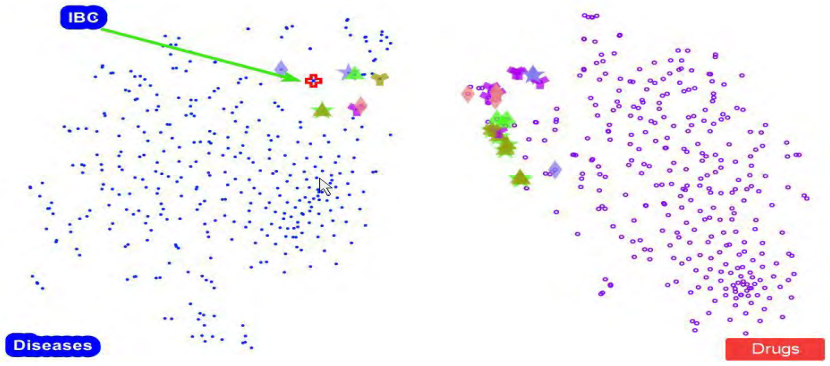


Figure 2 t-SNE generated scatter plots for diseases (left) and corresponding drugs (right). Similar diseases are clustered together, so are their corresponding drugs.

We plan to employ the same protocol to analyze all PubMed literature that covers research from 1787 to 2020 (incl. COVID19 research papers) to generate the vectors for drugs, diseases and targets.

<u>Research Strategy Specific Aim 2</u> - Developing AI/SVM-models for predicting drug-disease and drug-target associations.

Computational Methods and Design

Step 1 - Known drug-disease and drug-targets pairs will be collected from DrugBank [10], KEGG [11] to expand our own existing collection of 900 drug-disease pairs and 9000 drug-target pairs.

Step 2a - for each of the drug-disease (DD') pairs, we concatenate the drug vector (Dvector) with disease vector (D'vector) to form the pair's attributes vector (Dvector, D'vector). For binary classification, known DD' pairs will be given a value of 1. Equal number of negative samples will be generated on-the-fly by randomly combining a drug with a disease. This way, a training set is created for DD' pairs.

Step 2b - for each of the drug-target pairs, we concatenate the drug vector (Dvector) with the pairing target vector (Tvector) to form the pair's attribute vector (Dvector, Tvector). Similar to 2a, known DT pairs will be labeled as 1. Equal number of negative samples will be generated by randomly combining drug-target pairs which are not in the original ones. A training set is created for DT pairs.

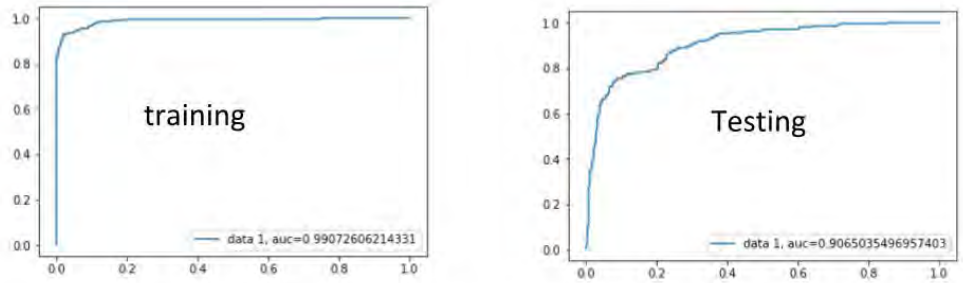
Step 3 - The two training sets above will be used for AI/SVM model development for DD' pair prediction and DT pair prediction, respectively. The DD' training set will have 50% positives ("1") and 50% negative samples ("-1"). The DT set will also have 50% positive ("1") and 50% negative samples ("-1"). Each full set will be then split into a 80% subset as training and a 20% subset for testing.

Step 4 - for DD' models, ROC (Receiver Operating Characteristic) curves will be generated for the training and testing sets. This standard characteristic curve is an indicator of the model performance. For DT set models, similar ROC curves will be generated to indicate the model performance as well.

Step 5 - We will predict known drugs against all known diseases including COVID19 using the validated DD' model resulting from Step 4. We will also predict known drugs against all targets with the DT models. These two predictions will essentially generate a <u>knowledge graph</u> linking drugs, diseases and targets.

Preliminary Data

We have employed the above protocol to analyze our past collection of 900 drug-disease (DD') pairs and 9000 drug-target (DT) pairs. Examples of the ROC curves for DD' pairs and DT pairs are shown in Figure 3a and 3b.



Drug-disease pairs: training & testing ROC curves, respectively

Figure 3a. ROC curves for training and testing sets of drug-disease pairs.

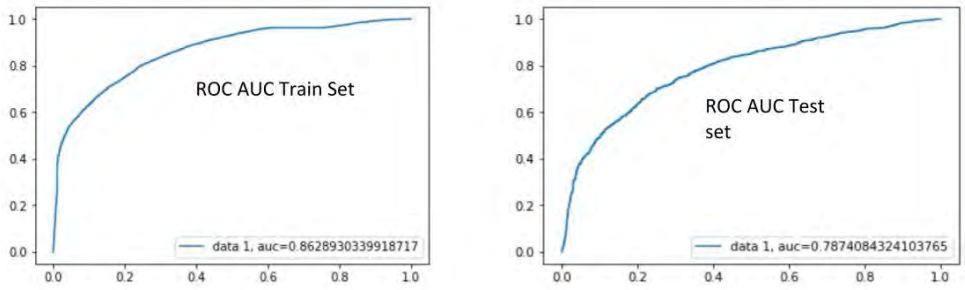


Figure 3b. ROC curves for training and testing sets of drug-target pairs.

An AUC (Area Under Curve) of >99% for training and one >90% for testing were obtained for DD' pairs, indicating that excellent models had been developed in our past studies of DD' pairs. Similarly, an AUC of 86% was obtained for DT pair training set and 79% was obtained for DT pair testing set, which indicates a fairly good classification model.

We plan to employ the same protocols to analyze expanded DD' (drug-disease) pairs and expanded DT (drug-target) pairs to train and test new models for our current project. These models will then predict unknown drug-disease pairs and unknown drug-target pairs to generate an expanded, predicted drug-disease-target knowledge graph.

<u>Research Strategy Specific Aim 3</u> - Extracting and validating the predicted DD' pairs and DT pairs relevant to COVID19.

Computational Method and Design

Step 1 - extract predicted drug-disease pairs and drug-target pairs related to COVID19.

Step 2 - search in PubMed, clinicaltrials.gov, DrugBank[10] and PubChem[18] to validate each predicted pair in terms of potential mechanisms of action. If potential mechanisms in terms of target suitability/ pathway involvement in COVID19 are found, the drugs or targets will be retained as potential candidates.

Step 3a - examine specifically sigma 1 receptor related drugs; drugs related to cathepsin L; drugs related to IL6 & IL10 modulation and JAK inhibitors related to cytokine storm prevention.

Step 3b - investigate drugs predicted to be related to SARS-CoV (SARS) and SARS-CoV-2 (COVID19) in general as a diverse pool of potential candidates to complement existing ones.

Preliminary Data

In our previous analyses of DD' pairs, the SVM-model predicted the association between Paroxetine and peptic ulcer. Paroxetine was developed as an antidepressant. Literature showed that repeated paroxetine treatment significantly attenuates the stress-induced ulcerogenic process. Also predicted by SVM-model was Prednisone and Eosinophilic fasciitis - Prednisone is an anti-inflammatory glucocorticoid. Literature search showed that the concomitant use of methotrexate and prednisone was effective for managing patients with Eosinophilic fasciitis.

In this project, we plan to employ the same approach to elucidate novel COVID19-drug relationships and find potential therapeutic targets for developing COVID19 drugs.

REFERENCES

1: Gautret, P. et al. Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. Int J Antimicrob Agents. 2020 Mar 20: 105949.

2: Grein, J. et al. Compassionate Use of Remdesivir for Patients With Severe Covid-19. N Engl J Med. 2020; NEJMoa2007016. doi: 10.1056/NEJMoa2007016.

3: McKee, DL. et al. Candidate drugs against SARS-CoV-2 and COVID-19. Pharmacol Res. 2020 Apr 29 : 104859.

4: Caly, L. et al. The FDA-approved drug ivermectin inhibits the replication of SARS-CoV-2 in vitro. Antiviral Res. 2020 Jun; 178: 104787.

5: Pedersen, S.F. et al. SARS-CoV-2: A Storm Is Raging. J Clin Invest. 2020 May 1;130(5):2202-2205. doi: 10.1172/JCI137647.

6: Pawar, A.Y. Combating Devastating COVID-19 by Drug Repurposing Int J Antimicrob Agents. 2020 Apr 17 : 105984.

7: Mikolov T. et al. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 (2013).

8: Cortes, C. et al. Support-vector networks. Machine Learning. 20 (3): 273–297. CiteSeerX 10.1.1.15.9362. doi:10.1007/BF00994018 (1995).

9: Scikit Learn SVM implementation at https://scikit-learn.org/stable/modules/svm.html

10: Wishart, DS et al. Drugbank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 34 (Database issue):D668-72. 16381955, (2006).

11: Kanehisa, M. et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000)

12: Schriml, LM. et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 40(Database issue): D940–D946, (2012).

13: Pedregosa et al. Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830, (2011).

14: Word2Vec Gensim implementation at https://radimrehurek.com/gensim/models/word2vec.html

15: van der Maaten, L.J.P. et al. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, (2008).

16: t-SNE: scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

17: Cox, S. et al. A Semantic Similarity Based Methodology for Predicting Protein-Protein Interactions: Evaluation with P53-Interacting Kinases. Journal of Biomedical Informatics. Submitted, May, 2020.

18: Kim, S. et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 47(D1):D1102–D1109. doi:10.1093/nar/gky1033, (2019).

TIMELINE AND FUTURE PLANS

Tasks	July, August	September, October	November,December
Aim1	Word2Vec models		
Aim2		SVM models	
Aim3			Deliver COVID19 drugs

Deliverables: we will deliver a list of drugs and their connections to drug targets and COVID19. A knowledge map covering drugs-diseases-targets will also be shared as RDF triples so that the general community can display and search for predicted relationships.

Future Plan

Future plans include:

(a) acquiring predicted compounds for testing internally at BRITE Screening Center (cf. Dr. Navarro's LOS),

and

(b) applying for a COVID19 testing grant externally at NIH/NCATS via PAR-18-462, which solicits predicted drugs for COVID19 using computational algorithms.

The NCATS initiative (NOT-TR-20-012) states "*many existing experimental drugs, FDA approved drugs, and biologics already have been tested in humans, and detailed information is available about their pharmacology, formulation, and potential toxicity. By building upon previous research and development efforts, new uses for existing drugs or biologics can be advanced to testing in clinical trials more quickly than starting from scratch. If a new therapy receives regulatory approval, it can be efficiently integrated into clinical practice. Coronaviruses are a diverse family of viruses that cause a range of disease in humans and animals, and there are currently no approved coronavirus therapeutics. In January 2020, a novel coronavirus, SARS-CoV-2, was identified as the causative agent of an outbreak of viral pneumonia. Transmission characteristics and the associated morbidity and mortality are not completely understood, but there is clear evidence of human-to-human transmission. Many other aspects of the disease are poorly understood. Given this, there is an urgent public health need to better understand COVID-19 and find therapies to treat infections.*"

We believe that our approach is unique in that it surveys all PubMed publications from its beginning to current with the aid of a modern AI-driven text mining tool. It has the potential to discover hidden, non-obvious solutions to the problem at hand. The results of this pilot project would provide data to support our application to the NCATS's FOA.

BUDGET

A $50K total salary for 6 months to cover 2 scientists ($25K per scientist)
Plus,
$3K publication fee

Total cost: $53K

BUDGET JUSTIFICATION

Weifan Zheng, Ph.D. (PI, 20% effort) will be responsible for the overall system design, supervision and execution of the project. He will also be responsible for biomedical/ pharmaceutical analyses of all predicted drug-disease and drug-target relationships. He will be in charge of the final release of data and coordinating collaborations with colleagues at BRITE to prepare for drug testing. He will also be writing grants to NIH for testing predicted drugs against COVID19.

Chunming Jin, Ph.D. (Research Scientist, 100% effort) will be responsible for conducting text mining and machine learning modeling. He has worked on Word2Vec and Support Vector Machine modeling before, and will be the main hands-on execution of Aims 1 and 2.

Zheng Huang, M.S.E. (Programmer, 100%) will be responsible for writing and implementing visualization software Similarity Explorer that will be used to analyze drug-drug similarity, target-target similarity as well as disease-disease similarity. The tool will be instrumental for the final analysis of the results in Aim 3.

Publication of findings just in time to contribute to the developing COVID19 pandemic situation requires financial support (color figures or open access fees, etc.).

DATA SHARING

All data will be shared via open source means on the Internet as plain text files and RDF triples.

OTHERS

No animal studies involved.

No experimental reagents are needed.

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Weifan Zheng

eRA COMMONS USER NAME (credential, e.g., agency login): WEIFANZHENG

POSITION TITLE: Associate Professor, Pharmaceutical Sciences

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
Peking University, Peking	B.S.	07/1985	Chemistry (organic)
Nankai University, Tianjin	M.S.	07/1988	Computer & Chemistry
University of North Carolina, Chapel Hill	Ph.D.	12/1997	Medicinal Chemistry (Computational Cheminformatics)

A. Personal Statement

I have the expertise and complete dedication to carry out the proposed research project. I have a broad background in cheminformatics, with specific training in computational drug discovery (1997). As an investigator at GlaxoSmithKline (1998-2001), I had carried out various cheminformatics research and development work. As a result, I had been awarded a special Award in cheminformatics at GSK. As a Sr. Research Scientist at Lilly Research Labs (2001-2004), I had been involved in chemogenomics research to support the design and development of gene family related chemical libraries. I had been a co-investigator on the NIH supported project - Exploratory Cheminformatics Center at UNC-Chapel Hill from 2005-2008. I had been the PI of an NIH supported grant on integrated informatics system for orphan neurodegenerative diseases as well as grants from rare disease foundations and NIH. In sum, I have a broad experience in informatics research, both in pharmaceutical industry and academia. At NCCU, my lab has developed a suite of computational drug discovery tools including receptor-dependent QSAR technologies and a shape pharmacophore based virtual screening tool (Shape4).

B. Positions and Honors

1988 – 1993 Lecturer, Nankai University, Lab for Computing & Chemical Testing
1998 – 2001 Investigator, Cheminformatics Department, GlaxoSmithKline, Philadelphia
2001 – 2004 Sr. Computational Chemist, promoted to Sr. Research Scientist, Eli Lilly & Company
2005 – 2006 Research Associate Professor, School of Pharmacy, UNC-Chapel Hill
2006 – present Associate Professor, Pharmaceutical Sciences, BRITE Institute, NC Central Univ.
2006 – present Adjunct Associate Professor, Eshelman School of Pharmacy, UNC-Chapel Hill

C. Key Papers

1. Chemical Library & Compound Database Design & Chemogenomics

a. Dong X, Ebalunode JO, Yang SY, Zheng W. Receptor-based pharmacophore and pharmacophore key descriptors for virtual screening and QSAR modeling. Curr Comput Aided Drug Des. 2011 Sep 1;7(3):181-9.
b. Zheng W, Hung ST, Saunders JT, Seibel GL. Piccolo: A Tool for Combinatorial Library Design via Multi-criterion Optimization. In Pacific Symposium on Biocomputing, Eds. R. B. Altman, A. K. Dunker, World Scientific, Singapore, 2000, pp.588-599.

2. Research in QSAR (quantitative structure-activity relationship) methods
   a. Zheng W, Tropsha A. Novel variable selection quantitative structure -property relationship approach based on the k-nearest-neighbor principle. J. Chem. Inf. Comput. Sci. 2000, 40(1), 185-94.
   b. Dong X, Ebalunode JO, Cho SJ, Zheng W. A novel structure-based multimode QSAR method affords predictive models for phosphodiesterase inhibitors. J Chem Inf Model. 2010 Feb 22;50(2):240-50.
   c. Dong X, Hilliard SG, Zheng W. Structure-based quantitative structure--activity relationship modeling of estrogen receptor β-ligands. Future Med Chem. 2011 Jun;3(8):933-45.
   d. Dong, X and Zheng, W. Receptor-dependent QSAR technologies. Invited chapter, E-book series, Future Science Publishing, 2013.
3. Receptor-based Pharmacophore and Database Search
   a. Ebalunode JO, Ouyang Z, Liang J, Zheng, W. A novel approach to structure-based pharmacophore searching using computational geometry and shape matching techniques. J Chem Inf Model. 2008; 48(4):889-901.
   b. Ebalunode JO and Zheng W. Molecular Shape Technologies in Drug Discovery: Methods and Applications. Curr Top Med Chem. 2010;10(6):669-79
4. Cheminformatics approach to model siRNA gene silencing
   a. Ebalunode, JO and Zheng W. Cheminformatics Approach to Gene Silencing: Z Descriptors of Nucleotides and SVM Regression Afford Predictive Models for siRNA Potency. Molecular Informatics. 29(12), 871–881, 2010.
   b. Ebalunode JO, Jagun C and Zheng W. Informatics Approach to the Rational Design of siRNA Libraries. Invited chapter in: Methods Mol Biol. 2011; 672:341-58.
5. Text mining for drug repurposing and drug discovery
   Cox S.; Rai R.; Dong X.; Christopherson, L.; Zheng, W.*; Tropsha, A.*; Schmitt, C.* Semantic Similarity Approaches to the Prediction of P53-Kinase Interactions. (submitted, 2020)

D. Additional Information: Research Support

Ongoing

U01CA207160-01 (Co-Inv: Zheng, W.; PI: Alex Tropsha, UNC-Chapel Hill) 9/1/2016 – 8/31/2020
NIH $75K / year subcontract
Drug Repurposing for Cancer Therapy: From Man to Molecules to Man (based on text mining)

UNC ROI Grant (Co-I: Zheng; PI: Tropsha, UNC-Chapel Hill) 10/2017 – 9/2020
$75K / year
Data Science: Infohub for Rare Diseases (based on text mining and semantic databases)

Completed

MBRS-SC3 (PI: Zheng, W.) 10/01/2008 – 12/31/2013
National Institutes of Health $75K / year
Informatics Resource for neurodegenerative targets
CHDI Foundation Grant (PI: Zheng, W.) 7/01/2008 – 6/30/2010
$150K / year
Informatics System for Huntington's Disease.
P01GM055876-12S1 (PI: Eckenhoff; Co-investigator: Zheng, W) 09/30/2009 -08/31/2011
National Institutes of Health $90,910 / year
Interaction of Inhalational Anesthetics with Macromolecules

June 1, 2020

Weifan Zheng, Ph.D.
Department of Pharmaceutical Sciences
& BRITE Institute
North Carolina Central University
Durham, NC 27707

RE.: Your application for NCCU COVID19 grant

Dear Dr. Zheng:

I have read your grant proposal for a pilot research on COVID19 entitled “Drug Repurposing for COVID19 Using Data Mining and Machine Learning Technologies”. I am excited about the potential of employing AI-driven text mining technologies to survey and analyze the biomedical literature, PubMed, in order to discover hidden relationships among drugs, targets and diseases. This research has a great potential for drug repurposing in general, and in particular, it should afford a fast approach to explore the literature to discover existing drugs for the COVID19 pandemic. I look forward to learning about any recommended drugs and other preclinical compounds uncovered as part of this research. All of BRITE’s resources, especially our Drug Discovery core, will be available to support this project should assays need to be developed to test interesting compounds.

I enthusiastically support your timely proposal.

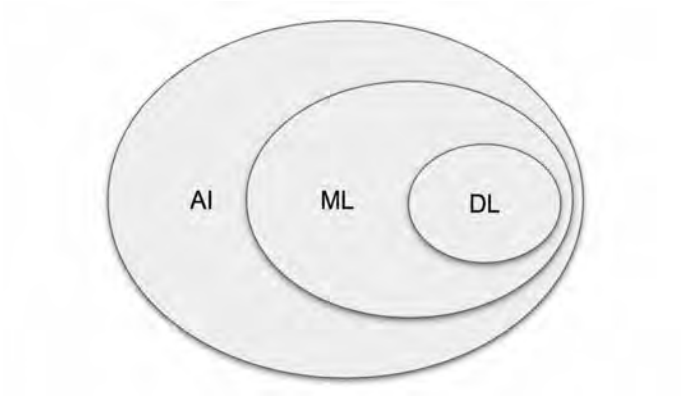
Sincerely yours,

Hernán Navarro, PhD
Director
Biomanufacturing Research Institute & Technology Enterprise (BRITE)
NC Central University
Durham NC 27707
Tel: 919-530-7001
Email: hnavarro@NCCU.EDU

Dr. Kumar:

Thank you for your comments regarding the use of the term ML as opposed to the term AI in our application, as well as the suggestion of doing topic-model analysis. Both are good points, which are addressed further as follows.

1. The relationship between AI (artificial intelligence) and ML (machine learning) can be summarized in the following figure, where AI refers to the broadest domain of computational algorithms that mimic human intelligence while ML covers a smaller scope of algorithms that include DL (deep learning) as well as SVM (support vector machine), etc. Therefore, though the term AI/SVM is technically correct, it is not as specific as ML/SVM. Thus, I agree to the terminology change - rather than AI/SVM, we will use ML/SVM instead in future revised applications.



2. Topic-model

This is a great suggestion. In fact, we have an ongoing initial development that does exactly that, where Word2Vec embedding vectors of all the terms in an abstract are simply summed up into a high dimensional vector for the abstract. This is called the Doc2Vec method [1]. This way, an abstract is characterized by a single vector. Similar abstracts can be related and grouped together as a “topic” based on similarity among the document-vectors. All the abstracts can be visualized on a t-SNE scatter plot or clustered with a hierarchical clustering method. Researchers can investigate interesting groups of abstracts/papers together to find topics such as cytokine storm, blood coagulation, ARDS, etc. as related to COVID19.

Another more sophisticated technique is to use LDA2Vec [2] designed to achieve the same goal of topic modeling, which will be explored in the future.

References

1. https://radimrehurek.com/gensim/models/doc2vec.html
2. https://towardsdatascience.com/lda2vec-word-embeddings-in-topic-models-4ee3fc4b2843