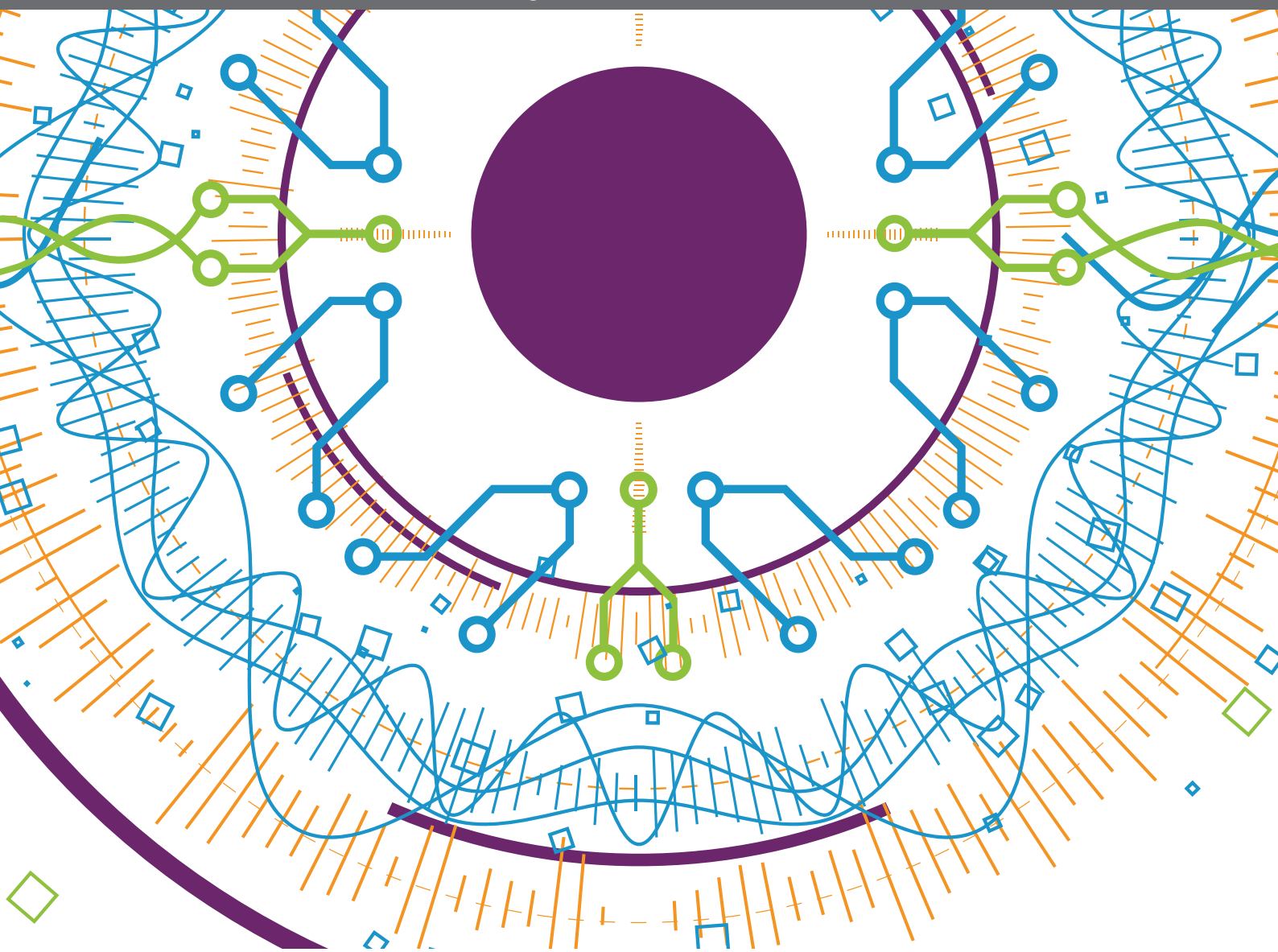
UNLEASHING INNOVATION ON PRECISION PUBLIC HEALTH: HIGHLIGHTS FROM THE MCBIOS & MAQC 2021 JOINT CONFERENCE

EDITED BY: Ramin Homayouni, Huixiao Hong, Prashanti Manda, Bindu Nanduri and Inimary Toby

PUBLISHED IN: Frontiers in Artificial Intelligence, Frontiers in Genetics and Frontiers in Big Data



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88976-539-3
DOI 10.3389/978-2-88976-539-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

UNLEASHING INNOVATION ON PRECISION PUBLIC HEALTH: HIGHLIGHTS FROM THE MCBIOS & MAQC 2021 JOINT CONFERENCE

Topic Editors:
Ramin Homayouni, Oakland University William Beaumont School of Medicine, United States
Huixiao Hong, United States Food and Drug Administration, United States
Prashanti Manda, University of North Carolina at Greensboro, United States
Bindu Nanduri, Mississippi State University, United States
Inimary Toby, University of Dallas, United States

Citation: Homayouni, R., Hong, H., Manda, P., Nanduri, B., Toby, I., eds. (2022). Unleashing Innovation on Precision Public Health: Highlights from the MCBIOS & MAQC 2021 Joint Conference. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-539-3

Table of Contents

frontiers in Artificial Intelligence

EDITORIAL
published: 25 February 2022
doi: 10.3389/frai.2022.859700

Editorial: Unleashing Innovation on Precision Public Health–Highlights From the MCBIOS and MAQC 2021 Joint Conference

Ramin Homayouni[1]*, *Huixiao Hong*[2], *Prashanti Manda*[3], *Bindu Nanduri*[4] and *Inimary T. Toby*[5]

[1] Oakland University William Beaumont School of Medicine, Rochester, MI, United States, [2] National Center for Toxicological Research, United States Food and Drug Administration, Jefferson, AR, United States, [3] University of North Carolina at Greensboro, Greensboro, NC, United States, [4] College of Veterinary Medicine, Mississippi State University, Mississippi State, MS, United States, [5] Department of Biology, University of Dallas, Irving, TX, United States

Keywords: machine learning, genomics, adverse drug effects, alternatives to animal testing, artificial intelligence

Editorial on the Research Topic

Editorial: Unleashing Innovation on Precision Public Health–Highlights From the MCBIOS and MAQC 2021 Joint Conference

This Research Topic is a product of the 17th annual conference of the Midsouth Computational Biology and Bioinformatics Society (MCBIOS), which has a broad membership of scientists and trainees with research interests in genomics, medicine, drug discovery and therapeutics. The topic includes a total of eight papers, four of which appear in *Frontiers in Artificial Intelligence* (including three original research articles and one review), three in *Frontiers in Big Data* (including one original research article, one Technology and Code article, and one brief research report), and one in *Frontiers in Genetics Computational Genomics* (Data Report). The papers can be categorized into two general themes of genomics and machine learning applications, as described below.

GENOMICS

Genomic data are generated in a complicated multi-step process that can impact the reproducibility of the results. In addition, many methods and software tools are available to analyze genomic data, which often yield different results from the same data. To address these challenges, Ma et al. developed a software infrastructure called NPARS (NGS post-pipeline accuracy and reproducibility system) that encapsulates genomic datasets in a portable database container, which can then be analyzed by well-established open-source application programming interfaces (APIs). They demonstrated the usefulness of NPARS in improving accuracy and reproducibility of different analysis methods on large and complex genomic data sets. In addition, the infrastructure provides a more convenient means to collaborate between groups.

Wang et al. enhanced the loci2path software for performing eQTL enrichment to identify enriched tissue specific pathways. The improved version includes additional pathways from PID, Reactome, and WikiPathways. The study uses over 13 million eQTLS from the Genotype Tissue Expression (GTEx) resource for 49 tissue types. Biological pathways that are likely to be involved in ten critical traits such as Alzheimer's disease, schizophrenia, and non-small cell lung cancer were identified. The software was shown to be valuable at uncovering new biological mechanisms of important traits.

OPEN ACCESS

Edited and reviewed by:
Thomas Hartung,
Johns Hopkins University,
United States

***Correspondence:**
Ramin Homayouni
rhomayouni@oakland.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *21 January 2022*
Accepted: *31 January 2022*
Published: *25 February 2022*

Citation:
Homayouni R, Hong H, Manda P, Nanduri B and Toby IT (2022) Editorial: Unleashing Innovation on Precision Public Health–Highlights From the MCBIOS and MAQC 2021 Joint Conference. Front. Artif. Intell. 5:859700. doi: 10.3389/frai.2022.859700

Quintanillla et al. developed a comprehensive database for genes and variants specifically related to Acute Respiratory Distress Syndrome (ARDS). The ARDS-DB framework provides gene and variant information and associated metadata derived from primary level curation of experimentally verified studies. The advantage of a dedicated gene database for deeper analysis of ARDS is that it provides the user with a centralized location to retrieve pertinent information. ARDS DB is freely available *via* an open-source repository and represents a major step toward filling a gap in computational resources for bench biologists and clinicians.

MACHINE LEARNING APPLICATIONS

Scientific data are growing and expanding at an overwhelming pace, making it challenging for scientists to organize, analyze and extract value from the vast amount of data. There is an urgent need for efficient and reliable methods and tools to mine signatures out of large datasets. Using an unsupervised machine learning approach, Nguyen et al. developed a software tool called SEAS (Statistical Enrichment Analysis of Samples) for mining biological sample information from genomic data. SEAS is available as a standalone or web version with a user-friendly graphical user interface. It can extract metadata and analyze numerical and categorical data to compute sample similarities and to cluster samples (e.g., patients). The authors demonstrated the utility of SEAS on publicly available data sets from The Cancer Genome Atlas (TCGA).

Li et al. present the development and implementation of DeepCarc, which uses a deep learning framework to predict carcinogenicity of small molecules. DeepCarc was developed using data in the National Center for Toxicological Research liver cancer database (NCTRlcdb) and tested against data in DrugBank and Tox21. DeepCarc model outperformed five machine learning classifiers, two state-of-the-art ensemble methods, and four molecule-based deep learning models. The DeepCarc model is designed to be an alternative method to test carcinogenicity and to alleviate the time-consuming and labor-intensive process of evaluating carcinogenic potency in experimental animal systems. DeepCarc is freely available for use and can be accessed *via* the following link: (https://github.com/TingLi2016/DeepCarc).

Application of machine learning to histopathological images is becoming common in both academic and commercial domains. There is still a need to detect and classify different immune cell types in the tumor immune microenvironment (TIME), which play crucial roles in determining cancer progression, metastasis, and response to treatment. Lee et al. provide a review of published models and applications in the three different scales of histopathology analyses: whole slide image (WSI)-level, region of interest (ROI)-level, and cell-level. In addition they provide a simplified framework for the development of a cell-type classifier using weakly labeled datasets generated from immunolabeled slides. The pros and cons for each method is highlighted and the future direction for histopathological image analysis is discussed.

Automated analysis of drug labels for "indication and usage" can be useful for clinical decision making, regulatory management as well as drug repositioning. Bhatt et al. developed a five-category Drug Indication Classification and Encyclopedia (DICE) based on >7,000 sentences from FDA approved human prescription drug labels. In addition, they developed nine different AI-based classifiers, including 4-word embeddings-based Bidirectional long short-term memory (BiLSTM) models and five transformer-based language models. The model performance was comprehensively assessed based on a test set and an independent validation set.

Adverse drug reactions (ADRs) such as drug-induced liver injury (DILI) are described in three sections, "Adverse Reactions", "Warnings and Precautions" and "Boxed Warning", in FDA drug labeling documents. Because of the complexity of the language and lack of standardization, Wu et al. explored using deep learning based language modeling approach to classify DILI from drug labels. A Bidirectional Encoder Representations from Transformers (BERT) model was trained for binary DILI classification of FDA-approved drug labeling documents and was externally validated using EMA-approved drug labeling documents.

Taken together, the papers selected for this Research Topic provide examples of cutting-edge approaches for standardizing analysis of large datasets and demonstrate the utility of applying machine learning methods to extract valuable insights from such data sources.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

PM is supported by a grant from the Division of Biological Infrastructure at the National Science Foundation (# 1942727). BN is partially supported by grant # P20GM103646 (Center for Biomedical Research Excellence in Pathogen Host Interactions) from the National Institute for General Medical Sciences. IT is supported by a grant from The Nancy Cain and Jeffrey A. Marcus Science Endowment in Honor of President Donald A. Cowan (University of Dallas).

Author Disclaimer: This editorial reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 02 August 2021
doi: 10.3389/frai.2021.711467

DICE: A Drug Indication Classification and Encyclopedia for AI-Based Indication Extraction

Arjun Bhatt[1,2,3], *Ruth Roberts*[4,5], *Xi Chen*[1], *Ting Li*[1], *Skylar Connor*[1], *Qais Hatim*[6], *Mike Mikailov*[7], *Weida Tong*[1]* *and Zhichao Liu*[1]*

[1]*Division of Bioinformatics & Biostatistics, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR, United States,* [2]*Dartmouth College, Hanover, NH, United States,* [3]*Brody School of Medicine, East Carolina University School of Medicine, Greenville, NC, United States,* [4]*ApconiX Ltd, Alderley Edge, United Kingdom,* [5]*Department of Biosciences, University of Birmingham, Birmingham, United Kingdom,* [6]*Office of Translational Sciences, Center for Drug Evaluation and Research, US FDA, Silver Spring, MD, United States,* [7]*Office of Science and Engineering Labs, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD, United States*

Drug labeling contains an 'INDICATIONS AND USAGE' that provides vital information to support clinical decision making and regulatory management. Effective extraction of drug indication information from free-text based resources could facilitate drug repositioning projects and help collect real-world evidence in support of secondary use of approved medicines. To enable AI-powered language models for the extraction of drug indication information, we used manual reading and curation to develop a **D**rug **I**ndication **C**lassification and **E**ncyclopedia (DICE) based on FDA approved human prescription drug labeling. A DICE scheme with 7,231 sentences categorized into five classes (indications, contradictions, side effects, usage instructions, and clinical observations) was developed. To further elucidate the utility of the DICE, we developed nine different AI-based classifiers for the prediction of indications based on the developed DICE to comprehensively assess their performance. We found that the transformer-based language models yielded an average MCC of 0.887, outperforming the word embedding-based Bidirectional long short-term memory (BiLSTM) models (0.862) with a 2.82% improvement on the test set. The best classifiers were also used to extract drug indication information in DrugBank and achieved a high enrichment rate (>0.930) for this task. We found that domain-specific training could provide more explainable models without performance sacrifices and better generalization for external validation datasets. Altogether, the proposed DICE could be a standard resource for the development and evaluation of task-specific AI-powered, natural language processing (NLP) models.

Keywords: natural language processing, deep learning, artificial intelligence, transformers, drug indication

OPEN ACCESS

Edited by:
Ramin Homayouni, Oakland University William Beaumont School of Medicine, United States

Reviewed by:
Arriel Benis, Holon Institute of Technology, Israel
Alexander Sedykh, Sciome LLC, United States

***Correspondence:**
Weida Tong
weida.tong@fda.hhs.gov
Zhichao Liu
Zhichao.liu@fda.hhs.gov

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *18 May 2021*
Accepted: *19 July 2021*
Published: *02 August 2021*

Citation:
Bhatt A, Roberts R, Chen X, Li T, Connor S, Hatim Q, Mikailov M, Tong W and Liu Z (2021) DICE: A Drug Indication Classification and Encyclopedia for AI-Based Indication Extraction. Front. Artif. Intell. 4:711467. doi: 10.3389/frai.2021.711467

INTRODUCTION

Drug labeling contains an 'INDICATIONS AND USAGE' section that provides vital information to support clinical decision making and regulatory management. The primary role of drug indications is to enable health care practitioners to readily identify appropriate therapies for patients and support clinical decision making (Sohn and Liu, 2014). The information on drug indication is part of the required information in FDA approved drug labeling and guides the content and format of labeling

of human prescription drugs and biological products [21 CFR 201.57(c) (2)]. Drug indications also provide guidance for facilitating clinical knowledge management and play an essential role in enabling the secondary use of electronic medical records (EMRs) for clinical-based translational research. Besides the primary drug indication approved for the drug, information on off-label uses and repurposing opportunities, or alternative uses of drugs, are common within biomedical-related data resources such as scientific literature, patents, public health forums, and pharmacological, biomedical, or drug labeling databases (Salmasian et al., 2015; Delavan et al., 2018). Furthermore, indication information extraction is also a regulatory requirement for creating the highlights section of the Physician Labeling Rule (PLR) labeling, which provides concise information for public health practitioners, patients and drug reviewers (https://www.fda.gov/drugs/laws-acts-and-rules/prescription-drug-labeling-resources). Thus, developing an effective approach to facilitate the mining of drug indication information from free text-based resources is an important task for biomedical natural language processing (NLP).

Some attempts to extract drug indications from free text-based documents have been undertaken, mainly based on the combination of named entity recognition (NER) approaches with conventional machine learning algorithms (Fung et al., 2013; Khare et al., 2014; Khare et al., 2015). One example is a two-step strategy for drug indication extraction proposed by Khare et al. 2014. Here, disease terminology is extracted from over 500 drug labels using a MetaMap tool with the Unified Medical Language System (UMLS)-based disease lexicon as the control vocabulary (Aronson, 2001). Then, a binary support vector machine (SVM) classifier is implemented to distinguish drug indication from other information such as adverse drug reactions, yielding an 86.3% F1 measure (the measure of a model's accuracy) for the indication extraction task, representing a 17% improvement over baseline approaches. With advances in AI-powered NLP, new approaches have been developed, which may provide additional performance improvements in the task of drug indication extraction. Artificial intelligence (AI)-powered language models such as transformers have achieved greater improvement compared to other approaches in various NLP tasks (Vaswani et al., 2017a; Devlin et al., 2018). Several biomedical-based BERT models (i.e., BioBERT, SciBERT, and clinicalBERT) have been developed for domain-specific tasks such as biomedical named entity recognition (NER) (Beltagy et al., 2019; Huang et al., 2019; Lee et al., 2020). Disease entity recognition corpora, such as the NCBI disease corpus, have become widely established sources for developing AI-based NER approaches (Doğan et al., 2014). However, the lack of large corpora for disease information classification hampers AI-based NLP development, and efforts to address this gap are urgently needed (Khare et al., 2015).

Based on guidance for industry on the 'INDICATIONS AND USAGE' section of Labeling for Human Prescription Drug and Biological Products, content should be concise but unambiguous. The information in the 'INDICATIONS AND USAGE' section should readily allow the identification of approved indication(s) and reflect current scientific evidence. Furthermore, indication terminology should be standardized, clinically relevant, scientifically valid, and easily understandable. Also, this information should be consistent within/across drug and therapeutic classes to aid the indexing of indications in electronic drug databases and medical information systems. Drug indication information often comprises mixed information such as age group, subpopulations, classifications such as adjunctive or concomitant therapy, specific tests/diagnoses, and other disease conditions or clinical manifestations. Thus, drug labeling is a great resource for drug indication classification, facilitating the development of AI-based NLP models, and further improving drug indication information extraction.

We developed a five-category **D**rug **I**ndication **C**lassification and **E**ncyclopedia (DICE) based on FDA approved human prescription drug labeling to facilitate the development of AI-based NLP approaches for enhanced drug indication extraction from free text-based document resources. The DICE scheme categorizes the >7,000 sentences in the 'INDICATIONS AND USAGE' section into five classes, including indication, contraindication, side effect, usage instruction, and clinical observations. To verify the utility of DICE, we developed nine different AI-based classifiers, including 4-word embeddings-based Bidirectional long short-term memory (BiLSTM) models and five transformer-based language models. The model performances were comprehensively assessed based on a test set and an independent validation set. Some critical questions such as the benefit of domain-specific training for AI-based NLP were also investigated. Furthermore, the model explainability was discussed for real-world applications.

MATERIALS AND METHODS

Figure 1 illustrates the workflow of the study. The study consisted of two components: DICE development, and AI-powered indication classification model development based on DICE.

Drug Indication Classification and Encyclopedia Development

To curate an indication classification corpus, we employed US Food and Drug Administration (FDA)-approved drug labeling. Drug labeling, also known as the package insert or prescribing information, accompanies every FDA approved medicine as required under the US Code of Federal Regulations (21 CFR 201.56). Drug labeling is submitted by the manufacturer and approved by FDA and includes a rich source of information on safe and effective drug usages. There are more than 80 sections embedded in a drug labeling document (Fang et al., 2020). Among the labeling sections, the INDICATIONS AND USAGE section aims to enable health care practitioners to readily identify appropriate therapies for patients by clearly communicating the drug's approved indication(s) (https://www.fda.gov/media/114443/download). The 'INDICATIONS AND USAGE' section mainly contains information such as

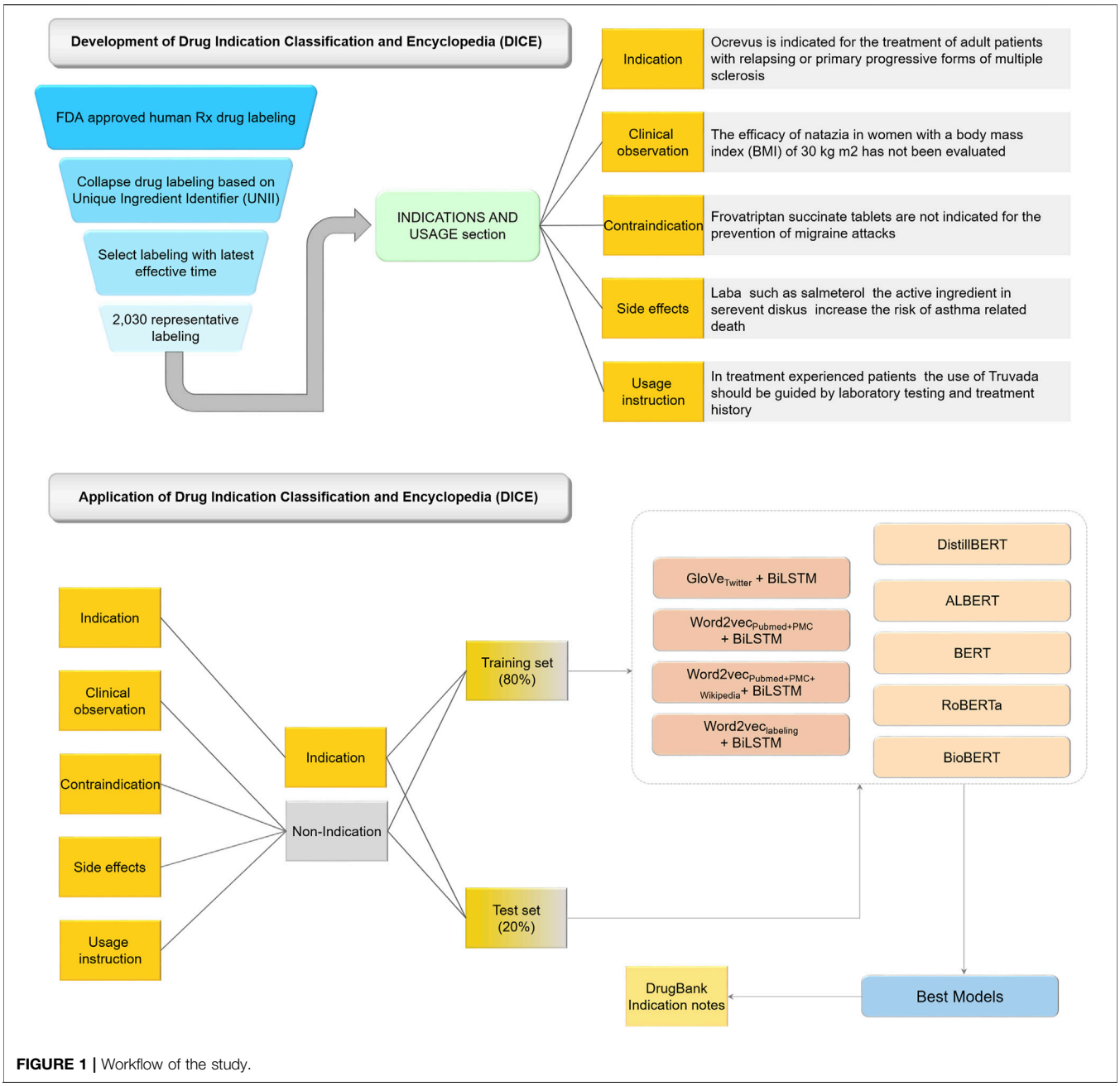


FIGURE 1 | Workflow of the study.

"1) The disease, condition, or manifestation of the disease or condition (e.g., symptoms) being treated, prevented, mitigated, cured, or diagnosed;" and "2) When applicable, other information necessary to describe the approved indication (e.g., descriptors of the population to be treated, adjunctive or concomitant therapy, or specific tests needed for patient selection)." Since the 'INDICATIONS AND USAGE' section contains such a variety of information, it is imperative to develop an indication classification corpus for automatic indication extraction.

Specifically, we extracted a list of FDA approved drug labels by using a search query "human Rx" under labeling type in the FDALabel databases (version 2.5, https://nctr-crs.fda.gov/fdalabel/ui/search) (Mehta et al., 2020). Consequently, we obtained queried results with summary information of human prescription (Rx) drug labeling. To obtain 'INDICATIONS AND USAGE' sections for a unique list of human prescription drug labels, we implemented the following strategy: 1) collapse the labeling with the same Unique Ingredient Identifier (UNII); 2) select the labeling with latest effective time as the representative one (i.e., XML file) for each collapsed labeling; 3) extract 'INDICATIONS AND USAGE' section from XML file based on Logical Observation Identifiers Names and Codes (LOINC) for Human Prescription Drug and Biological Product Labeling (https://www.fda.gov/industry/structured-product-labeling-

resources/section-headings-loinc). The LOINC code for INDICATIONS AND USAGE section is "34067-9" (**Figure 1**).

To manually annotate the information in the 'INDICATIONS AND USAGE' sections into different categories, we developed a five-class indication classification scheme (**Figure 1**). We split the extracted 'INDICATIONS AND USAGE' sections into sentences. Each of the 7,231 sentences were placed into one of five categories, namely indication, contraindication, side effect, usage instruction, and clinical observations. Assignments were based on predefined keywords and *a priori* knowledge. Three experienced, expert pharmacologists carried out the manual annotations independently and a consensus assignment was selected for indication information classification.

AI-Powered Indication Extraction Model Development

For the purposes of indication extraction, the extracted 7,231 sentences assigned with the category 'indication' were considered as positives, and sentences assigned to any of the other four categories were considered negative. The 7,231 curated sentences were divided into the training and test sets with an approximate ratio of 80:20. Consequently, we obtained 5,785 sentences and 1,446 sentences for the training and test sets, respectively. Two types of deep learning models were developed, including word embedding-based BiLSTM models, and transformer-based language models.

Preprocessing

We implemented the following procedure to preprocess the sentences: 1) the sentences were tokenized; with stripping of punctuation, digits, and words with less than two characters; 2) stop word removal; and 3) lemmatization.

Word Embeddings

Word embedding is a set of language modeling and learning techniques in NLP to map words or phrases from a vocabulary to a numeric vector representation. In this study, we used two types of word embeddings including Word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b) and Glove (Pennington et al., 2014).

Word2vec is a shallow neural network framework (i.e., continuous bag-of-words (CBOW) and continuous skip-gram) used to estimate continuous vector representations of words from large text corpora (Mikolov et al., 2013a; Mikolov et al., 2013b). The generated word embeddings position words with common contexts close to one another. Word2vec has been used widely in NLP tasks such as semantic relationship extraction (Chen et al., 2018), text classification (Jang et al., 2019), and sentiment analysis (Rezaeinia et al., 2019). In this study, we use three pretrained domain-specific word2vec models, including Word2vec with PubMed and PMC (i.e., Word2vecPubmed + PMC), word2vec with PubMed, PMC, and Wikipedia (i.e., Word2vecPubmed + PMC + Wikipedia), and word2vec with FDA approved human prescription labeling (Word2veclabeling). The Word2vecPubmed + PMC and Word2vecPubmed + PMC + Wikipedia (200-dimension vector models) were downloaded from https://bio.nlplab.org/(Moen and Ananiadou, 2013). We developed the pre-trained Word2veclabeling by using the human labeling documents described above. The in-house implementation of word2vec was consistent with the PubMed corpus; briefly, the implementation used the skip-gram model with a window size of 5, hierarchical SoftMax training, and a word subsampling threshold of 0.001 to create 200-dimensional vectors. The training was conducted using the Python Gensim package (version 0.6.0).

We also employed another well-known word embedding technique (i.e., GloVe 200-dimension vectors), which, when applied to aggregated global word-word co-occurrence statistics from a corpus, generate word vector representation (Pennington et al., 2014). Specifically, the pretrained GloVe model with 2 billion Twitter corpus was employed as the general domain specific word embedding (i.e., GloVeTwitter); this corpus can be downloaded from https://nlp.stanford.edu/projects/glove/.

Bidirectional Long Short-Term Memory

To better understand the framework and theory behind the BiLSTM, we provide a simple introduction on the Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM). An RNN is a set of artificial neural networks for sequential and time-series data. Unlike conventional neural networks, RNNs adopt recurrent hidden states to store previous inputs and leverage sequential information of the previous inputs to estimate the next element in the sequence. In theory, RNNs are able to leverage previous sequential information for arbitrarily long sequences. In practice, however, due to RNNs memory limitations called "vanishing gradients", the length of the sequential information is limited to only a few steps back (Hochreiter et al., 2001).

Hochereiter and Schimdhuber (Hochreiter and Schmidhuber, 1997) proposed the LSTM model, which is a gated RNN intended to solve the "vanishing gradients" problem and greatly expand RNNs applications for long sequence data (Gers et al., 1999). The LSTM cell consists of four components (i.e., input gate, memory cell, forget gate, and output gate) to remember information over a longer period of time and thus enable reading, writing, and deleting information from the cell's memory. The forget gate makes the decision of preserving/removing the existing information, the input gate specifies the extent to which the new information will be added into the memory, and the output gate controls whether the existing value in the cell contributes to the output (Siami-Namini et al., 2019). The deep-BiLSTMs are an extension of the described LSTM model above, in which two LSTMs are applied to the input sequence (i.e., forward layer) and reverse of the input sequence (i.e., backward layer) (Schuster and Paliwal, 1997). Applying the LSTM twice leads to the enhanced learning of long-term dependencies and thus improves the accuracy of the model.

Supplementary Figure S1 illustrates the proposed BiLSTM model infrastructure for indication classification. The processed sentences were vectorized by the different word embedding techniques described above; the now vectorized sentences were

then fed into bidirectional LSTM layers and a dense layer, followed by a flattened layer and a dense layer. The output layer is a probabilistic value of sentences belonging to the indication information category. Specifically, we used a learning rate of 0.001, Rectified Linear Units (ReLU) activation, and an Adagrad Optimizer. The optimizer was chosen due to its suitability for training on sparse data and its ability to perform more informed gradient-based learning.

Transformer-Based Language Models

To further investigate the performance of advance AI-powered NLP approaches on indication classification, we employed the Bidirectional Encoder Representations from Transformers (BERT) (Vaswani et al., 2017a; Devlin et al., 2018) and its derivatives including a distilled version of BERT (DistilBERT) (Sanh et al., 2019), A Lite BERT (ALBERT) (Lan et al., 2019), a Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019), and a pre-trained biomedical BERT (BioBERT) (Lee et al., 2020).

BERT is a transformer that learns contextual bidirectional representations from an unlabeled, large corpus of documents by using two training strategies: Masked Language Model (MLM) and Next Sentence Prediction (NSP) (Vaswani et al., 2017b; Devlin et al., 2018). In the MLM, a randomly selected 15% of words in a sequence are replaced with a [MASK] token, and the model aims to estimate masked words, based on the context provided by unmasked words. In the NSP, the model aims to utilize the pairs of sentences as inputs and predict the sequence order in the original documents. The BERT model has achieved state-of-the-art performance on diverse sets of NLP tasks (e.g., text classification, named entity recognition) while requiring only minimal task-specific architectural modification (i.e., fine-tuned layers).

Two condensed BERT models, DistillBERT and ALBERT, were proposed to overcome the obstacle of long training times. DistilBERT uses a technique called distillation, which approximates the BERT from the large neural network to a smaller one. By learning from the distilled version of BERT, DistilBERT retained about 97% performance while using only half as many parameters as the original BERT (Sanh et al., 2019). One of the key optimization functions used for posterior approximation in DistilBERT is Kulback Leiber (K-L) divergence to condense the network size while maintaining performance. ALBERT is a light version of BERT, which employs two techniques to reduce the parameters, including Factorized Embedding Parametrization and Cross-layer Parameter Sharing (Lan et al., 2019). Additionally, a self-supervised objective is proposed for sentence order prediction to further improve performance, addressing the suboptimal performance of the NSP task from BERT.

RoBERTa is an updated version of BERT that improves the pretrained optimization process (Liu et al., 2019). First of all, RoBERTa uses a much larger set of training data (161 GB) for pretraining to increase the model's generalization ability. Secondly, instead of the static masking pattern used in the MLM model, RoBERTa introduced a dynamic masking pattern to avoid same training mask for each training instance. Also, the RoBERTa model developed training objectives to enhance NSP model performance. Moreover, RoBERTa trained on longer sequences than BERT to further improve performance.

It was observed that generic pretrained transformer models may not work very well in conjunction with specific domain data. To fill this gap, BioBERT, a domain-specific BERT model, was proposed by training the BERTbase model on large biomedical corpus including PubMed abstracts and PMC full text (Lee et al., 2020). The BioBERT model outperformed BERTbase on some domain-specific tasks such as biomedical named entity recognition (NER), and bio-Questions and answering with a 0.51–9.61% absolute improvement.

BERT-like models are designed as pre-trained deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. They are then fine-tuned with an additional output layer to create models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. The fine-tuned base models of transformers were used in this study for the binary classification task for indication recognition. An important difference is that these models used their native tokenizers, which utilized sub-word tokenization (e.g., WordPiece) where larger words may be broken down to map to token(s), compared to the cruder tokenization implemented with the simpler model.

Model Performance Evaluation

To train the model and measure model performance, we employed area under the receiver operating characteristic (ROC) curve analysis, which demonstrates the performance of the classification model by plotting the true predictive rate (TPR) against the false positive rate (FDR). We calculated the area under the ROC curve (AUC) for each model described above. We also used seven other performance metrics including Matthews correlation coefficient (MCC), accuracy, sensitivity, specificity, precision, negative predictive value (NPV), and F1-score for further evaluation of model performance by using the following confusion matrix and formulas

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \quad (1)$$

$$accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (2)$$

$$sensitivity = \frac{TP}{TP+FN} \quad (3)$$

$$specificity = \frac{TN}{TN+FP} \quad (4)$$

$$PPV = \frac{TP}{TP+FP} \quad (5)$$

$$NPV = \frac{TN}{TN+FN} \quad (6)$$

$$F1 = \frac{2TP}{2TP+FP+FN} \quad (7)$$

		Actual class	
		Indication (positive)	Non-indication (negative)
Predicted class	Indication (positive)	True positive (TP)	False positive (FP)
	Non-indication (negative)	False negative (FN)	True negative (TN)

External Validation

To further investigate real-world applications of the developed indication classification model, we applied the best-performing models to indication descriptions in the DrugBank database. The indication information in DrugBank is a relatively concise description of the indication and usage of approved or investigational drugs (Wishart et al., 2018). Specifically, the DrugBank (version 5.1, downloaded on April 02, 2021) XML file was downloaded via https://go.drugbank.com/releases/latest. We developed an in-house script to extract the drug indication information from DrugBank XML file. Consequently, a list of 3,976 indication descriptions in DrugBank were extracted to further verify our developed model.

Visualization

To investigate the discrimination powers of different word embeddings or sentence embeddings yielded from the transformers models used in this study, we employed t-distributed stochastic neighbor embedding (t-SNE) (Hinton and Roweis, 2002). t-SNE is a non-linear dimension reduction method. With t-SNE, the algorithm calculates the similarity in both high dimensional space and low dimensional space. Next, the similarity difference in both spaces is minimized using an optimization method such as gradient descend.

Data and Code Availability

We developed a GitHub webpage (https://github.com/arjun-bhatt/TransformersIndicationExtraction) to share the source code and curated drug indication corpus. Specifically, all the code script is developed under Python 3.6. The BiLSTM model is based on tensorflow version 1.12.3. The transformers models were based on Huggingface package version 3.02 and its backend is tensorflow version 2.3.0 and PyTorch version 1.5.1. t-SNE was implemented by using Python Scikit-learn package version 0.23.2.

RESULTS

Drug Indication Classification and Encyclopedia

Figure 2A illustrates the distribution of 7,321 sentences in the proposed drug indication classification scheme. To curate a high-quality drug indication classification, three pharmacologists manually read the sentences and assigned them into five predefined categories including indication, non-indication miscellaneous, contraindication, side effect, and usage instruction. Based on consensus manual annotation results, the 7,231 sentences were categorized into 4,297 indication, 1,673 clinical observations, 701 contraindication, 492 usage instructions, and 68 side effects (**supplementary Table S1**). **Figure 2B** depicts the most frequent words in each indication classification category using word clouds. For example, the top five key words in the Indication category were "indicated", "treatment", "patients", "therapy", and "disease", respectively (**supplementary Table S2**). To develop an indication recognition classifier, we used the 4,297 indication as positives, and 2,934 combined sentences from the other categories as negatives, yielding a ratio between positives and negatives of 1.46. Then, we randomly split the 7,231 into a training set (80%) and a test set (20%). Accordingly, the training set (i.e., 5,785 sentences) consisted of 3,452 positives and 2,333 negatives (P/N ratio = 0.596), and the test set (i.e., 1,446 sentences) consisted of 845 positives and 601 negatives (P/N ratio = 0.584).

Word Embedding-Based Bidirectional Long Short-Term Memory Models

To develop BiLSTM models for indication classification, we used four types of word embeddings, including Word2vecPubmed + PMC, Word2vecPubmed + PMC + Wikipedia, Word2veclabbeling, and GloVeTwitter. To illustrate the potential benefit of domain-specific embedding, we randomly selected four different domain-specific words (i.e., aspirin, heart, azithromycin, and cancer) to get their top ten most similar words (**Figure 2**). **Figure 3** illustrates the clusters of similar words based on the t-SNE analysis. The Word2veclabbelin, Word2vecPubmed + PMC, and Word2vecPubmed + PMC + Wikipedia models could cluster similar words for the queried words more closely than GloVeTwitter models, highlighting the benefit of domain-specific word embedding for semantic relationship extraction in biomedical applications. We found that the performance of BiLSTM models with domain-specific word embeddings (i.e., MCC = 0.878 for Word2vecPubmed + PMC + Wikipedia > MCC = 0.864 for Word2vecPubmed + PMC > MCC = 0.857 for Word2veclabbeling) was slightly better than that of the BiLSTM model with general domain-based word embedding (MCC = 0.849 for GloVeTwitter). Furthermore, the other 7 performance metrics including accuracies, AUCs, F-scores, sensitivity, specificity, NPV and PPV of domain-specific embedding-based LSTMs were consistently better than general domain embedding-based BiLSTM, indicating domain-specific embedding-based BiLSTMs could extract the indication-related information more accurate (**Table 1**).

Transformers-Based Models Outperformed the Word Embedding-Based Bidirectional Long Short-Term Memory Model

To further explore the possibility of improving the binary indication classification model performance, we implemented

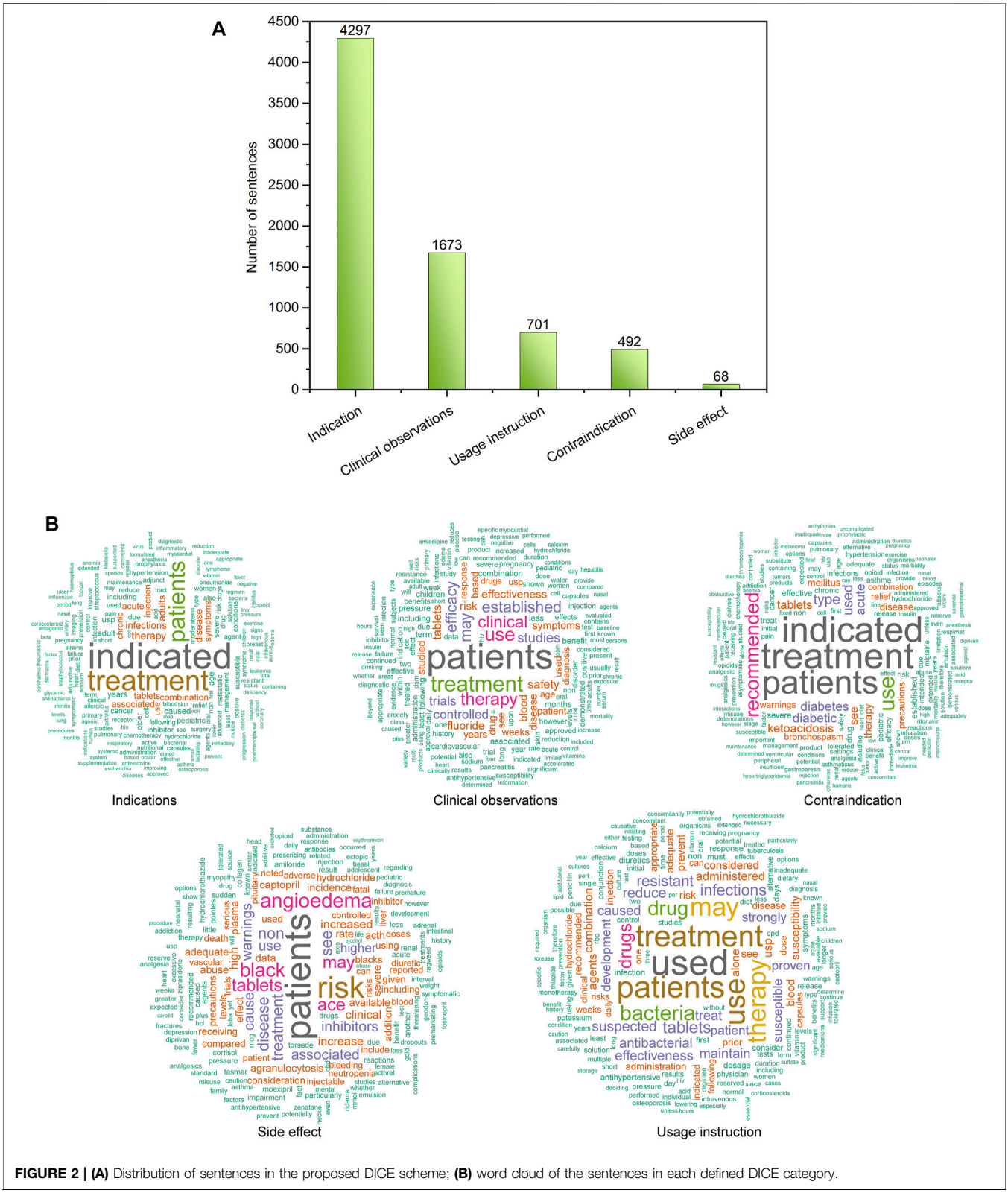


FIGURE 2 | (A) Distribution of sentences in the proposed DICE scheme; **(B)** word cloud of the sentences in each defined DICE category.

five different fine-turned BERT-like transformer models, including BERT, DistillBERT, ALBERT, RoBERTa, and BioBERT (**Table 1**). First, all transformer-based models except DistillBERT outperformed word embedding-based BiLSTMs. Second, RoBERTa, BioBERT, and BERT yielded better performance (MCC = 0.921, 0.917, and 0.899, respectively)

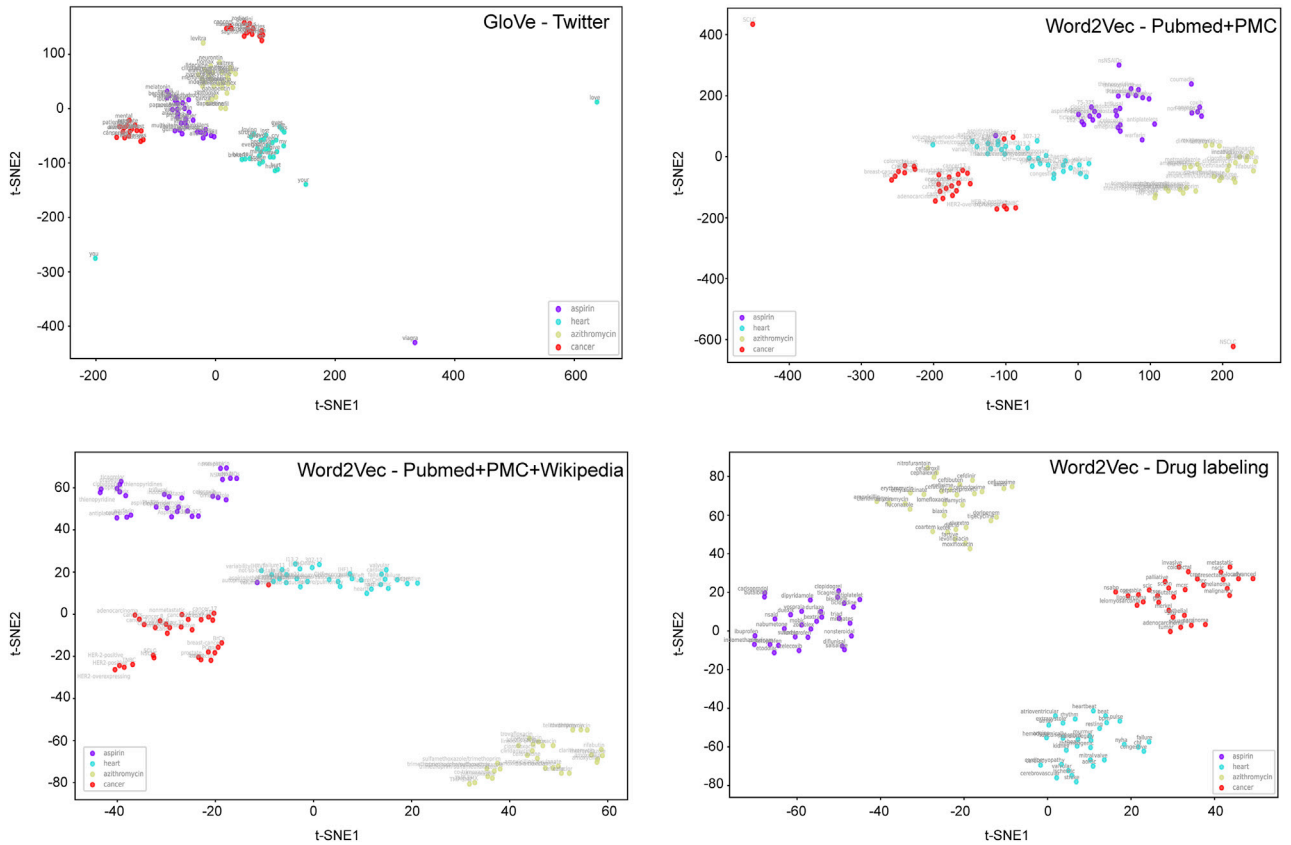


FIGURE 3 | t-SNE analysis of different word embedding models on the queried words.

TABLE 1 | Model performances of nine different AI-based models for indication classification on test set*.

Models	MCC	ACC	AUC	F-score	Sensitivity	Specificity	NPV	PPV
Bidirectional long short-term memory (BiLSTM)								
GloVe (twitter)	0.849	0.925	0.981	0.935	0.908	0.950	0.875	0.964
Word2vc (Drug Labeling)	0.857	0.929	0.977	0.940	0.916	0.950	0.883	0.965
Word2vec (PubMed+ PMC)	0.864	0.934	0.977	0.944	0.925	0.946	0.893	0.963
Word2vec (PubMed + PMC+ Wikipedia)	0.878	0.941	0.982	0.950	0.945	0.935	0.921	0.955
BERT and its derivates								
DistilBERT	0.820	0.911	0.970	0.922	0.896	0.933	0.862	0.950
ALBERT	0.877	0.941	0.978	0.950	0.964	0.907	0.946	0.937
BERT	0.899	0.951	0.985	0.958	0.949	0.954	0.927	0.968
BioBERT	0.917	0.960	0.987	0.966	0.972	0.943	0.959	0.960
RoBERTa	0.921	0.962	0.987	0.968	0.962	0.962	0.945	0.974

**Positive predictive value (PPV) and negative predictive value (NPV).*

than the condensed transformers including ALBERT and DistilBERT (MCC = 0.877 for ALBERT and MCC = 0.820 for DistilBERT). Third, domain-specific word embedding-based BiLSTM (i.e., MCC = 0.878 for Word2vecPubmed + PMC + Wikipedia) outperformed the condensed BERT models (i.e., MCC = 0.820 for DistilBERT), highlighting the improvement of model performance based on the large size of the domain-specific corpus, even with the relatively shallow deep learning model. Fourth, the performance of domain-specific BERT (i.e., BioBERT) was comparable to that of RoBERTa, which is trained on top of a large general corpus and with more aggressive hyperparameters.

We further employed a t-SNE analysis to visualize the contribution of hidden states of transformers on classification performance (**Figure 4**). We observed the obvious margin for discriminating positives from negatives based on the hidden layer information of most of the transformer models. It is interesting that the positives and negatives samples were closer for the

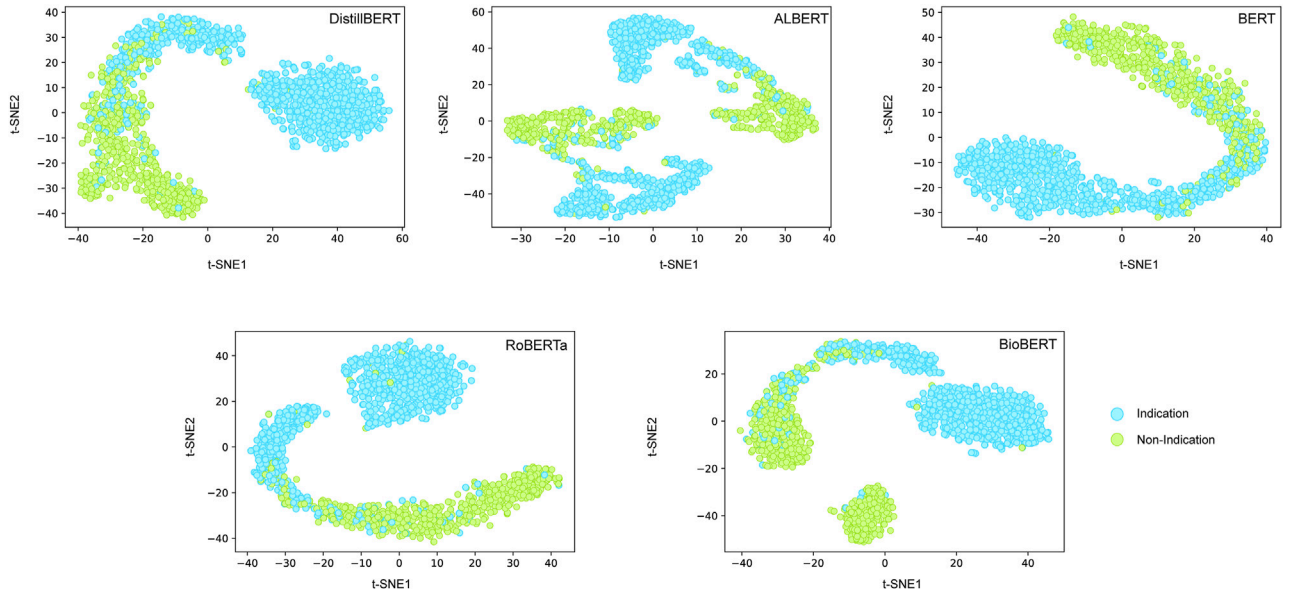


FIGURE 4 | t-SNE analysis of average hidden states of transformers on test set.

BioBERT model, which may be the reason for the unexpectedly small contribution of domain-specific training for the test set.

Indication Information Extraction for DrugBank Indication Notes

Working towards a real-world application, we applied the top performance models to extract the indication-related sentences in DrugBank indication description notes. The Drugbank indication description notes are concise information for drug indications without other information such as contraindications, side effects, and specific population. We considered all drug indication notes as positives. Therefore, we could calculate the enrichment rate that measures the number of indication information sentences correctly recognized by the developed models. The enrichment rates were ranked as RoBERTa (0.952) > BioBERT (0.936) > BERT (0.930), which is consistent with previous results based on test sets (**Figure 5**). Based on the model performances of both the test set and external validation set, BioBERT and RoBERTa could provide more robust performance and better generalization ability for different data resources.

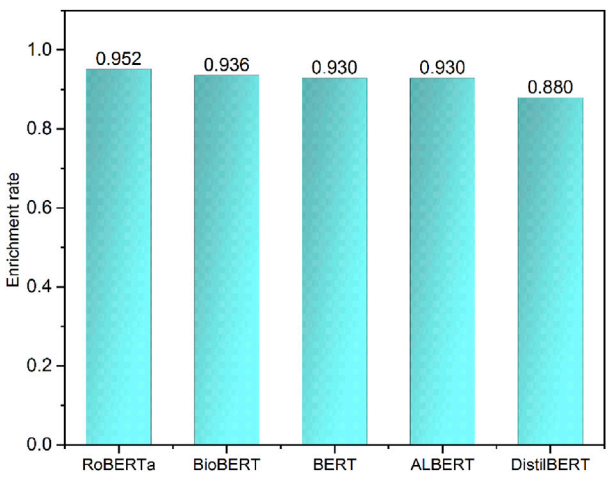


FIGURE 5 | Enrichment rate of fine-tuned transformers on Drugbank indication notes.

DISCUSSION

Drug indications provide key medical information to support clinical decision making and promote the appropriate use of medicines. Furthermore, drug indication information is also considered a fundamental resource to assist in the standardization of medical coding and to potentially eliminate medical errors (Fung et al., 2013). AI-powered NLP models have successfully been applied to various biomedical-related tasks such as biomedical entity recognition, text classification and questioning and answering. However, a standard corpus for domain-specific tasks is urgently needed to advance the development of AI algorithms. In this study, we developed a five-tier based Drug Indication Classification and Encyclopedia (DICE) based on FDA approved drug labels with a consensus manual curation strategy, to facilitate automatic indication information extraction from free text with AI-powered NLP approaches. To verify the utility of the proposed DICE, we conducted a comprehensive comparison of nine deep learning-based NLP models consisting of word embedding-based BiLSTMs and BERT family models. Encouragingly, the top models such as RoBERTa and BioBERT outperformed others

with MCCs greater than 0.910 and accuracy greater than 0.960 on test sets, and enrichment rates greater than 0.930 on DrugBank indication notes, demonstrating the great potential of the DICE with AI for automatic indication information identification.

There have been a few attempts to curate the standard corpus of drug indication information for NLP development. However, the sample size is limited (e.g., ~150 drug labels) (Khare et al., 2015). Here, we used the entire list of FDA approved human prescription drugs to develop the DICE with a five-tier classification scheme. The DICE scheme took into account the FDA guidance requirement for 'INDICATION AND USAGE' section drafting (https://www.fda.gov/regulatory-information/search-fda-guidance-documents/indications-and-usage-section-labeling-human-prescription-drug-and-biological-products-content-and). The potential utility of DICE can be divided into two aspects: 1) The DICE could serve as a standard biomedical classification corpus for deep learning-based NLP algorithm development; and 2) the DICE could also be utilized for indication information extraction model development towards real world applications such as off-label use and potential drug repurposing opportunities derived from free-text resources (e.g., PubMed, EMR, patent, and social media).

The benefits of the domain-specific training on different biomedical applications have been discussed elsewhere (Beltagy et al., 2019; Huang et al., 2019; Lee et al., 2020). The domain-specific word embedding-based BiLSTM yielded better prediction performance than those built from general domain corpora. Furthermore, the explainability of domain-specific word embedding was superior as demonstrated by t-SNE analysis. We did not observe any significant improvement of domain-specific transformers (i.e., BioBERT) compared to the original BERTbase and RoBERTabase on the test set, indicating the performance of transformers may be task-specific and data specific. Furthermore, further training of domain-specific transformers (e.g., BioBERT, SciBERT, and ClinicalBERT) on FDA approved drug labeling data may be a potential direction to pursue even better performance, however, it is out of scope of the current study.

Advances in AI in NLP and increased computational power have allowed various transformer-based language models to be developed and successfully used in different downstream tasks (Devlin et al., 2018; Brown et al., 2020). As proof-of-concept of the utility of the developed DICE, we selected the transformers based on the BERT architecture. Other transformer-based models such as Generative Pre-trained Transformer (GPT) 2/3 (Brown et al., 2020), an autoregressive language model, have demonstrated high performance in different NLP tasks, especially in text generation and reading comprehension, which may be worth further investigation for potential performance improvements, even in the indication information classification task. However, the balance between performance, computational cost, and data size must be considered. Based on model results of the test set and DrugBank data set, the condensed models such as DistillBERT and ALBERT could also largely maintain the prediction performance with a more economical usage of computational resources.

The current version of DICE and associated AI-based language models were based on the English language. Further evaluation of other languages will be a great addition to expand the utility of the developed DICE corpus. First, the proposed data curation process of the DICE corpus is reproducible and could be migrated to the documents in other languages. Accordingly, the associated AI-based language models could be developed for drug indication information extraction in other languages. Second, tremendous efforts have been made to language translation powered by AI in the biomedical domain (Liu et al., 2021). For example, Liu et al. proposed a novel cross-lingual biomedical entity linking model among ten typologically diverse languages, which could translate the domain-specific terminology between the languages. By combining the developed biomedical entity linking model, the proposed indication extraction models could be utilized in other languages. However, further investigation and evaluation are strongly recommended.

It is worthwhile to consider some additional studies to further investigate the utility of DICE in different medical applications. First, the model performance of different models was only evaluated based on one test set. Considering the lack of annotated data (i.e., ground truth) in the other resources, we only employed DrugBank indication notes as positives to verify the proposed models for a real-world application. Some extra verifications are strongly recommended for expanding the utility of the developed DICE and accompanying models. Second, the developed DICE and classification models could serve as the first step to extract indication information. The other biomedical entity recognition approaches (e.g., UMLS MetaMap (Aronson, 2001) or BioBERT (Lee et al., 2020)) could be applied to extract disease-related terms for further applications. Third, in the current study, the AI-based indication extraction models are binary-based. Considering the unbalanced distribution of the five defined categories (4,297 indication, 1,673 clinical observations, 701 contraindication, 492 usage instructions, and 68 side effects), we suggest further investigations on the performance of the multi-class models. Lastly, while the developed DICE is a five-tier indication classification scheme, we only investigated its utility for automatic indication information extraction through its usage as a binary classifier. Evaluation for potential utility for testing multiple-class model performance is suggested.

Automatic drug indication extraction is of great importance for different biomedical applications. To fill this gap, we developed the DICE to facilitate AI-based algorithm development and verification. We hope our developed DICE will be considered as a standard drug indication classification corpus, providing the opportunity for other biomedical NLP researchers to promote AI-powered indication extraction in different real-world applications.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

ZL conceived and designed the study. AB and ZL performed data analysis. ZL wrote the manuscript. RR, QH, MM, AB, ZL, and WT revised the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

AB is grateful to the support received from the Summer Student Research Program (SSRP) at the National Center for Toxicological Research (NCTR). TL, XC, SC are grateful to the U.S. Food and Drug Administration (FDA)/NCTR for postdoctoral support through the Oak Ridge Institute for Science and Education (ORISE).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.711467/full#supplementary-material.

REFERENCES

Aronson, A. R. (2001). Effective Mapping of Biomedical Text to the UMLS Metathesaurus: the MetaMap Program. *Proc. AMIA Symp.*, 17–21.

Beltagy, I., Lo, K., and Cohan, A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text*. arXiv preprint arXiv:1903.10676.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). *Language Models Are Few-Shot Learners*, 14165. arXiv preprint arXiv: 2005.

Chen, Z., He, Z., Liu, X., and Bian, J. (2018). Evaluating Semantic Relations in Neural Word Embeddings with Biomedical and General Domain Knowledge Bases. *BMC Med. Inform. Decis. Mak* 18, 65. doi:10.1186/s12911-018-0630-x

Delavan, B., Roberts, R., Huang, R., Bao, W., Tong, W., and Liu, Z. (2018). Computational Drug Repositioning for Rare Diseases in the Era of Precision Medicine. *Drug Discov. Today* 23, 382–394. doi:10.1016/j.drudis.2017.10.009

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). *Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805.

Doğan, R. I., Leaman, R., and Lu, Z. (2014). NCBI Disease Corpus: a Resource for Disease Name Recognition and Concept Normalization. *J. Biomed. Inform.* 47, 1–10. doi:10.1016/j.jbi.2013.12.006

Fang, H., Harris, S., Liu, Z., Thakkar, S., Yang, J., Ingle, T., et al. (2020). FDALabel for Drug Repurposing Studies and beyond. *Nat. Biotechnol.* 38, 1378–1379. doi:10.1038/s41587-020-00751-0

Fung, K. W., Jao, C. S., and Demner-Fushman, D. (2013). Extracting Drug Indication Information from Structured Product Labels Using Natural Language Processing. *J. Am. Med. Inform. Assoc.* 20, 482–488. doi:10.1136/amiajnl-2012-001291

Gers, F. A., Schmidhuber, J., and Cummins, F. (1999). *Learning to Forget: Continual Prediction with LSTM*.

Hinton, G., and Roweis, S. T. (2002). *Stochastic Neighbor Embedding*. Citeseer: NIPS, 833–840.

Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. (2001). "Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks* (IEEE Press).

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Huang, K., Altosaar, J., and Ranganath, R. (2019). *Clinicalbert: Modeling Clinical Notes and Predicting Hospital Readmission*. arXiv preprint arXiv:1904.05342.

Jang, B., Kim, I., and Kim, J. W. (2019). Word2vec Convolutional Neural Networks for Classification of News Articles and Tweets. *PLOS ONE* 14, e0220976. doi:10.1371/journal.pone.0220976

Khare, R., Burger, J. D., Aberdeen, J. S., Tresner-Kirsch, D. W., Corrales, T. J., Hirchman, L., et al. (2015). Scaling Drug Indication Curation through Crowdsourcing. *Database (Oxford)* 2015, 2015. doi:10.1093/database/bav016

Khare, R., Wei, C.-H., and Lu, Z. (2014). "Automatic extraction of drug indications from FDA drug labels," in AMIA Annual Symposium proceedings. AMIA Symposium. 2014, 787–895. doi:10.1016/b978-0-323-16916-5.00013-4

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). *Albert: A Lite Bert for Self-Supervised Learning of Language Representations*. arXiv preprint arXiv:1909.11942.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et al. (2020). BioBERT: a Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* 36, 1234–1240. doi:10.1093/bioinformatics/btz682

Liu, F., Vulić, I., Korhonen, A., and Collier, N. (2021). *Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking*. arXiv preprint arXiv:2105.14398.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). *Roberta: A Robustly Optimized Bert Pretraining Approach*. arXiv preprint arXiv: 1907.11692.

Mehta, D., Uber, R., Ingle, T., Li, C., Liu, Z., Thakkar, S., et al. (2020). Study of Pharmacogenomic Information in FDA-Approved Drug Labeling to Facilitate Application of Precision Medicine. *Drug Discov. Today* 25, 813–820. doi:10.1016/j.drudis.2020.01.023

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). *Distributed Representations of Words and Phrases and Their Compositionality*. arXiv preprint arXiv:1310.4546.

Moen, S., and Ananiadou, T. S. S. (2013). Distributional Semantics Resources for Biomedical Text Processing. *Proc. LBM*, 39–44.

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: Global Vectors for Word Representation," in Proceedings of the 2014 conference on empirical methods in natural language processing. Doha, Qatar: (EMNLP), 1532–1543. doi:10.3115/v1/d14-1162

Rezaeinia, S. M., Rahmani, R., Ghodsi, A., and Veisi, H. (2019). Sentiment Analysis Based on Improved Pre-trained Word Embeddings. *Expert Syst. Appl.* 117, 139–147. doi:10.1016/j.eswa.2018.08.044

Salmasian, H., Tran, T. H., Chase, H. S., and Friedman, C. (2015). Medication-indication Knowledge Bases: a Systematic Review and Critical Appraisal. *J. Am. Med. Inform. Assoc.* 22, 1261–1270. doi:10.1093/jamia/ocv129

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). *DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter*. arXiv preprint arXiv: 1910.01108.

Schuster, M., and Paliwal, K. K. (1997). Bidirectional Recurrent Neural Networks. *IEEE Trans. Signal. Process.* 45, 2673–2681. doi:10.1109/78.650093

Siami-Namini, S., Tavakoli, N., and Namin, A. S. (2019). "The Performance of LSTM and BiLSTM in Forecasting Time Series," in 2019 IEEE International Conference on Big Data (Big Data). IEEE, 3285–3292. doi:10.1109/bigdata47090.2019.9005997

Sohn, S., and Liu, H. (2014). Mitteilungen der DGKJ. *Monatsschr Kinderheilkd* 162, 1046–1055. doi:10.1007/s00112-014-3201-y

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). *Attention Is All You Need*. arXiv preprint arXiv:1706.03762.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, 5998–6008.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi:10.1093/nar/gkx1037

Disclaimer: The views presented in this article do not necessarily reflect those of the U.S. Food and Drug Administration. Any mention of commercial products is for clarification and is not intended as an endorsement.

Conflict of Interest: RR is co-founder and co-director of ApconiX, an integrated toxicology and ion channel company that provides expert advice on non-clinical aspects of drug discovery and drug development to academia, industry, and not-for-profit organizations.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Artificial Intelligence

REVIEW
published: 10 September 2021
doi: 10.3389/frai.2021.754641

Deep Learning of Histopathology Images at the Single Cell Level

*Kyubum Lee[1†], John H. Lockhart[2†], Mengyu Xie[1], Ritu Chaudhary[3], Robbert J. C. Slebos[3], Elsa R. Flores[2,4], Christine H. Chung[3,5] and Aik Choon Tan[1,5]**

[1]Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, United States, [2]Department of Molecular Oncology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, United States, [3]Department of Head and Neck-Endocrine Oncology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, United States, [4]Cancer Biology and Evolution Program, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, United States, [5]Molecular Medicine Program, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, United States

The tumor immune microenvironment (TIME) encompasses many heterogeneous cell types that engage in extensive crosstalk among the cancer, immune, and stromal components. The spatial organization of these different cell types in TIME could be used as biomarkers for predicting drug responses, prognosis and metastasis. Recently, deep learning approaches have been widely used for digital histopathology images for cancer diagnoses and prognoses. Furthermore, some recent approaches have attempted to integrate spatial and molecular omics data to better characterize the TIME. In this review we focus on machine learning-based digital histopathology image analysis methods for characterizing tumor ecosystem. In this review, we will consider three different scales of histopathological analyses that machine learning can operate within: whole slide image (WSI)-level, region of interest (ROI)-level, and cell-level. We will systematically review the various machine learning methods in these three scales with a focus on cell-level analysis. We will provide a perspective of workflow on generating cell-level training data sets using immunohistochemistry markers to "weakly-label" the cell types. We will describe some common steps in the workflow of preparing the data, as well as some limitations of this approach. Finally, we will discuss future opportunities of integrating molecular omics data with digital histopathology images for characterizing tumor ecosystem.

Keywords: histopathology image analysis, deep learning, image data labeling, cell type classification, tumor immune microenvironment, tumor heterogeneity

OPEN ACCESS

Edited by:
Inimary Toby,
University of Dallas, United States

Reviewed by:
Sungjoon Park,
Korea University, South Korea
Zongliang Yue,
University of Alabama at Birmingham, United States

****Correspondence:***
Aik Choon Tan
aikchoon.tan@moffitt.org

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *06 August 2021*
Accepted: *27 August 2021*
Published: *10 September 2021*

Citation:
Lee K, Lockhart JH, Xie M, Chaudhary R, Slebos RJC, Flores ER, Chung CH and Tan AC (2021) Deep Learning of Histopathology Images at the Single Cell Level. Front. Artif. Intell. 4:754641. doi: 10.3389/frai.2021.754641

INTRODUCTION

In clinical settings, histopathology images are a critical source of primary data for pathologists to perform cancer diagnostic. For some cancer types, clinicians may decide treatment strategies based on histopathology images coupled with molecular assay data. With the widespread adoption of digital slide scanners in both clinical and preclinical settings, it is becoming increasingly common to digitize histology slides into high-resolutions images. Digital pathology, which is the process of digitizing histopathology images, creates a new "treasure trove of image data" for machine learning (ML). Machine learning can be utilized for various image analysis tasks that are routinely performed during histological analyses including detection, segmentation, and classification. Some commercial image analysis software already incorporates machine learning algorithms to assist researchers and clinicians in quantifying and segmenting histopathological images. These tools have greatly reduced the laborious and tedious manual work in image analysis and can reduce inter-observer variability in reaching diagnostic consensus (Tizhoosh et al., 2021).

Machine Learning which focuses on methods to construct computer programs that learn from data with respect to some class of tasks and a performance measure, has been widely applied in several challenging problems in bioinformatics due to the algorithm's ability to extract complex relationships from high-dimensional data. Conventional machine learning methods (e.g. random forest, support vector machines) were limited by their ability to extract features from raw data, and in many cases, a feature selection step is needed to reduce dimensionality of the data. In addition, efforts have been invested in careful feature engineering and domain knowledge to construct informative features to train the model. However, some engineered features are difficult to interpret biologically and have limited utility in biomedical applications.

In early 2000, several breakthroughs including new types of algorithms (e.g., deep learning), availability of large datasets (e.g., open access and large digitized images), and advancements in computing power (e.g., graphical processing units) have reenergized the machine learning developments and applications in real-world problems (LeCun et al., 2015). Deep Learning (DL) is a family of new machine learning models composed of multiple processing layers that learn representations of data with multiple levels of abstraction without feature engineering. The ability of deep learning to discover intricate structure in large data sets powered by a backpropagation algorithm allows the machine to change its internal parameters to compute a representation in each layer from the previous layer. The "deep" in deep learning representing the number of layers used in the model to deconvolute the feature representation of the raw data. These methods have dramatically improved the state-of-the-art in multiple domains ranging from speech and text recognition to object detections in biomedical applications (Esteva et al., 2017; Lee et al., 2018; McKinney et al., 2020; Nagpal et al., 2020; Liu et al., 2021).

The field of cancer pathology is proving to be a supremely suitable proving ground for the development of machine learning models, in no small part due to the construction of publicly available, curated whole slide image (WSI) datasets from initiatives like the Cancer Genome Atlas (TCGA), Clinical Proteomic Tumor Analysis Consortium (CPTAC), and the Cancer Image Archive (TCIA) (Clark et al., 2013; Prior et al., 2013). The datasets contained within these repositories often include other related data, such as clinical characteristics, patient outcomes, molecular analyses, and other imaging modalities, in addition to the WSIs. These data can be utilized as target features, such as predicted progression-free survival duration, or even integrated into the machine learning model for higher dimensional analysis. The numerous types of cancer collected by these repositories allow researchers to focus their applications as narrowly or broadly as they desire, from single subtypes (e.g., lung adenocarcinoma) to pan-cancer analyses.

However, the majority of machine learning applications in this field rely on supervised learning methods based on clinical parameters or pathologists' annotations to generate training datasets. Within supervised learning approaches, there exists several distinct resolutions of annotation required to generate a high-quality training dataset depending on the scale of the analysis. The ultimate goal of machine learning applications for histopathology is to generate clinically beneficial output, but this may be achieved in a wide variety of ways. For example, both a model designed to flag regions of concern for a pathologist to review in detail and a tool that identifies cancer patients that are likely to respond to immunotherapies by classification of immune cell types are likely to improve clinical outcomes, but these two models will require very different training datasets. In this review, we will consider three different scales of histopathological analyses that machine learning can operate within: WSI-level, region of interest (ROI)-level, and cell-level.

Many reviews have been published in describing the methods and applications of deep learning in pathological image analysis [see (Janowczyk and Madabhushi 2016; Dimitriou et al., 2019; Serag et al., 2019; Roohi et al., 2020; van der Laak et al., 2021)], however, none of these publications discussed or reviewed on the topic of training datasets preparation for ML/DL, which is the most crucial step in developing a useful model in histopathological image analysis. In addition, it is becoming clear that the tumor immune microenvironment (TIME) plays crucial role in determining cancer progression, metastasis, and response to treatment. Therefore, it is important to detect and classify the different immune cell types of the TIME in histopathological images. However, it is impractical to manually curate and annotate these individual cell types for training the model. To address this knowledge gap, we will discuss the basics of applying machine learning models for histopathological analysis within a cancer pathology setting, review currently published models and applications in the three different scales of histopathology analyses, and provide a simplified framework for the development of a cell-type classifier using weakly labeled datasets generated from immunolabeled slides. We aim for this review to be an approachable introduction to histopathological applications of machine learning/deep learning for clinicians, biologists, and data scientists, thereby encouraging further development of this interdisciplinary field.

MACHINE LEARNING-BASED HISTOPATHOLOGY IMAGE ANALYSIS AND DATA GENERATION METHODS

Histopathology Image Data Preparation for Training Machine Learning Models

Regardless of the feature scale of the ML analysis, it is also important to understand how the input data is prepared. Histopathology analysis is most commonly performed using sections of tissue collected during biopsy or after surgical resection. These tissues are typically preserved by formalin fixation and paraffin embedding (FFPE) to preserve their morphology, and subsequently sliced into thin sections on glass slides for further processing. The tissue sections are commonly subjected to chemical staining to highlight specific tissue or cellular features, such as nuclei or proteins of interest. Normal hematoxylin staining produces intense purple staining the nuclei of cells while eosin is used to counterstain the remaining cytoplasm of cells a vivid pink (**Figure 1**).

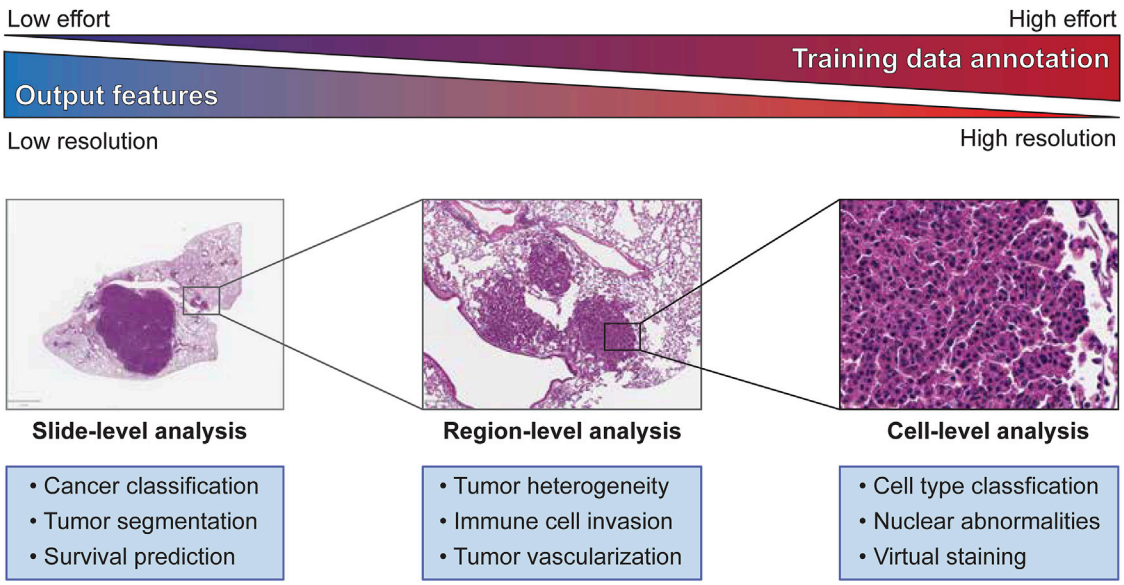


FIGURE 1 | Slide-, Region-, and Cell-level analyses on tumor histopathology images and its characteristics.

Hematoxylin and eosin (H&E) staining is employed in nearly every histological workup and is therefore the most common type of histological image used as inputs for machine learning models. **Figure 2** shows the overview of the histopathology slide preparation and machine learning model construction.

Training a machine learning model for image segmentation requires a large amount of high-quality, labeled images as a training dataset. Therefore, building an effective training dataset requires a careful balancing of data quantity, data quality, and cost. In comparison to many other fields that utilize computer vision, the amount of publicly available histopathology data suitable for training a machine learning model is quite limited. As previously mentioned, there is a growing number of datasets from consortia like TCGA, but many published studies also rely on in-house datasets for training and testing their machine learning models. This scarcity of data is compounded by the need for trained experts capable of producing accurate annotations that capture the defining features of the model's target classes. The difficulty of these annotations in terms of both the rater's expertise and the effort required to create increases sharply between whole slide-level, region-level, and cell-level analyses (**Figure 1**). Constructing the best training dataset will require annotations at the same level as the target outputs of the machine learning model (i.e., cell-level analysis performs best with cell-level annotation), but approaches like crowd-sourcing annotations may work as an alternative to expert annotation by sacrificing annotation quality for quantity (Amgad et al., 2019). For some applications, lower resolution annotations, such as annotating a region as a single class for training a cell-level classifier or labeling of cell centroids, may also be used to generate "weakly-labeled" annotations for model training. Alternatively, histology slides can be immunolabeled to identify specific cells or features of interest that can then be used to weakly label an adjacent, registered H&E image, an approach we will discuss later in this review. While these challenges may incline researchers to annotate at the lowest usable resolution to expedite model training, it should be noted that many analyses require the abstraction of higher resolution classifications to a lower level to produce biologically meaningful results (e.g., calculating the density of a cell type within a region after classification of individual cells).

Despite the complexities of training dataset generation, the number of applications and tools that utilize machine learning models in cancer pathology has grown rapidly over the last few years. This pace seems likely to continue and perhaps even increase as more datasets are made publicly available through cancer consortia and computational challenges like BreastPathQ (Petrick et al., 2021) or CAMELYON (Litjens et al., 2018). In the following three sections, we will cover published examples of machine learning applications and discuss the strengths, limitations, and considerations of each analysis scale level. We also summarized the published examples in **Table 1**.

Image-Level Analysis Methods

One of the immediate applications of deep learning in cancer research is to generate a cancer diagnosis from WSI analysis. Zhang et al. (2019) developed an artificial intelligence system to effectively automate WSI analysis to assist pathologists in cancer diagnosis. The AI system consists of three main neural networks: the scanner network (s-net), the diagnose network (d-net) and the aggregator network (a-net). In brief, the s-net ingests the WSI as inputs and automatically detects tumor regions in the images. Convolutional neural networks (CNN) was used in the s-net to manage tumor detection and cellular-level characterization. Once the tumor regions were detected, it further segments into ROIs for diagnostics. The d-net takes the ROI from s-net as inputs, and further characterizes by extracting pathological features and showing feature-aware network attention to explain the

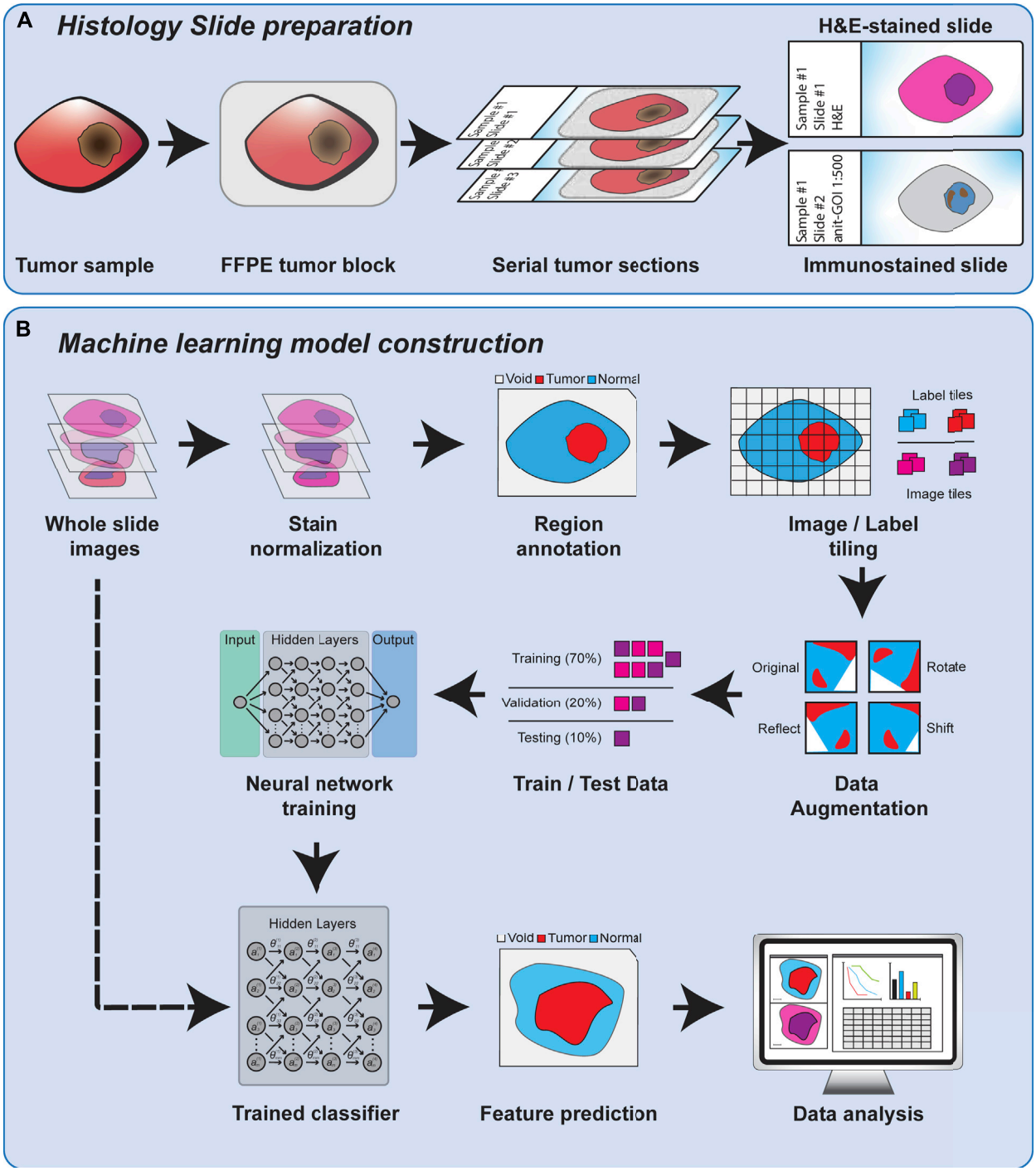


FIGURE 2 | An overview of histology slide preparation and machine learning model construction process.

network for interpretations. The d-net is developed by using fully connected recurrent neural networks (RNN). Finally, the a-net integrates all the characterized features and provides the final diagnosis. The a-net is implemented as a three-layer fully connected neural network that takes the features and predict the final labels. The authors also showed that their model could simultaneously generate pathological reports from the d-net. The authors trained this method using 913 H&E WSI slides obtained from TCGA Bladder Cancer and in-house slides. The authors showed that the prediction from the model matches the diagnoses from 17 pathologists. This study shows that machine learning/deep learning could be used to assist pathologists in cancer diagnostic from WSI.

TABLE 1 | Selected deep learning-based histopathology image analysis studies.

Publication	Input image type	Training annotations	Deep learning architecture	Prediction output	Other functions	Training dataset size	Level
Zhang et al. (2019)	Tiled images from whole slide (H&E)	Pathologist marking tumor and normal areas, input description text	CNN Attention module RNN	Probability of being tumor at Pixel level Attention area Text description	Text query of images using pathological terms	913 whole slides	WSI
TOAD Lu et al. (2021)	Whole slide	Tumor of origin Sex Primary or metastatic	CNN Attention module	Primary or metastatic Tumor of origin Attention areas		22,833 whole slides from 18 cancer types +6,499 test slides	WSI
Kalra et al. (2020)	Whole slide (H&E, frozen section)	Primary diagnosis	DenseNet	Relevant images ranked by similarity	Images features extracted as barcodes	29,120 whole slides from 32 cancer types (TCGA)	WSI
HE2RNA Schmauch et al. (2020)	Whole slide images	RNA expression for training	multilayer perceptron	RNA expression	Spatial mapping of gene expression	8,725 samples from 28 cancer types	WSI
Saltz et al. (2018)	Whole slide images (H&E)	Pathologist marking regions of lymphocytes and necrosis	CNN	TIL maps (Computational Staining)		5,455 images from 13 cancer types (TCGA)	ROI
Le et al. (2020)	Whole slide images (H&E)	Pathologist marking tumor regions, TIL annotations from Saltz et al. study	34-layer ResNet, 16-layer VGG, and Inception v4	Tumor probability and TIL probability heatmaps		Cancer detection: 393 breast cancer images from SEER and TCGA Lymphocyte detection: 1090 invasive breast cancer from TCGA	ROI
Lockhart et al. (2021)	Whole slide images (H&E)	Pathologist marking normal vs. tumor region (grade 1–4)	CNN(ResNet18)	Normal lung tissue/airways Lung adenocarcinoma of different grades (1–4)		In house images from mouse models	ROI
ConvPath Wang et al. (2019)	Whole slide images (H&E)	Pathologist labeled ROI (tumor, stroma, lymphocytes)	CNN	"spatial map" of tumor cells, stromal cells and lymphocytes (limited to lung adenocarcinoma)		TCGA-LUAD(1337) NLST(345) Beijing(102) SPORE(130)	ROI/ Cell level
CRImage Failmezger et al. (2020)	Whole slide images (H&E)	Single cell annotations, OS, omics data	Topological tumor graphs (TTG), unsupervised deep learning framework (CNx)	Cell level classified and mapped slices for further analysis and hypothesis generating		400 SKCM (TCGA)	Cell level

Cancer of unknown primary (CUP) represents a group of cancers in which the primary anatomical site of tumor origin cannot be determined. Unsurprisingly, CUP poses challenges to determine the appropriate treatments and clinical care. To address this challenge, Lu et al. (2021) developed a deep learning-based algorithm known as Tumor Origin Assessment via Deep Learning (TOAD), with the goal to predict the tissue of origin of the primary tumor using routinely acquired histology images. Histology slides from patients were automatically segmented and divided into thousands of small image patches and fed into a convolutional neural network (CNN) with fixed pretrained parameters. The CNN serves as the encoder to extract a compact, descriptive feature vector from each image patch. TOAD uses an attention-based multiple-instance learning algorithm to learn to rank all of the tissue regions in the WSI using the feature vectors, and aggregates this information across the whole slide based on their relative importance, and assigning more weights to regions perceived to have high diagnostic value. The authors also included patient's gender as an additional feature to further guide the classification of CUP. Based on this multi-branched network architecture and the multi-task learning objective, TOAD is able to predict both the tumor origin and distinguish primary from metastatic tumors. The authors trained TOAD on 22,833 WSIs spanning across 18 common origins of primary cancer, and tested TOAD on 6,499 WSIs with known primary tumor origins. TOAD achieved a top-1 accuracy of 0.83 and a top-3 accuracy of 0.96. Further testing TOAD on external test set of 682 samples

showed that it achieved top-1 accuracy of 0.80 and a top-3 accuracy of 0.93. Finally, the authors tested TOAD on 317 cases of CUP, and found that their model predicted in concordance for 61% of cases and a top-3 agreement of 82%. The authors suggested that this model could be used to assist pathologists to perform CUP assignment, as well as other difficult cases of metastatic tumor assignment.

In another study, Kalra et al. (2020) developed a pan-cancer diagnostic consensus through searching histopathology images using machine learning (ML) approaches. The authors first indexed ~30,000 WSI of 32 cancer types from TCGA using Yottixel, an image search engine previously developed by the authors (Kalra et al., 2020b). To index the WSI, the authors generated "bunch of barcodes" (BoB) index for each WSI instead of small patches of images. Because the dimensional reduction from patches of images to BoB, this indexing step accelerate the retrieval process and overcome the computation and storage of huge image files. The authors used DenseNet to extract the image patch and convert into a vector, and the BoB essentially is the binary form of the deep feature vector representations of each image patch. The authors illustrated the application of this machine learning approach on the TCGA WSI data, and showed that their method could retrieve relevant images with high accuracy (>90% on several cancer types). This study demonstrates that an alternative approach to query WSI in a database to retrieve relevant set of WSIs for potential cancer diagnosis, and particularly useful for rare cancer types.

Region of Interest-Level Analysis Methods

In addition to classification and searching tasks at whole slide level, deep learning approaches are also able to provide insights at more granular level, or give more emphasis to particular regions of interest (ROI) that are most informative. ROIs may be defined as geographic regions (e.g., central, marginal areas), or areas that are biologically divergent (e.g., tumor vs. stromal area, areas of different tumor grades), or areas that are enriched for specific cell types such as lymphocytes. A variety of ML tools have been developed to identify and analyze these ROIs and can provide unique insights into the biological differences between ROIs.

Spatial patterns of tumor-infiltrating lymphocytes (TILs) have shown significant value to cancer diagnosis and prognosis, however manually recognizing of those patterns requires tremendous efforts. Aiming to reduce the manual efforts and scale-up analysis capacity, Saltz et al. (2018) constructed a pipeline that mapped TILs to 5,455 H&E stained images from 13 TCGA tumor types. Their pipeline comprises two CNN modules (a lymphocyte-infiltrated classification CNN-lymphocyte CNN-and a necrosis segmentation CNN), that were trained on pathologist-annotated images of lymphocytes and necrosis. The training process also involves pathologists' feedback to improve performance. The CNNs combined outputs were used to produce TIL probability map that was then subjected to threshold adjustments to obtain the final TIL map. During testing, this pipeline achieved 0.95 area under the receiver operating characteristic curve (AUROC) which outperformed VGG16 network (0.92). Moreover, the authors compared the extracted TIL structure patterns with the molecular based estimation (i.e., CIBERSORT) and found it achieved ~0.45 correlation coefficient in best performed cancer types (e.g., BLCA, SKCM) and ~0.1–0.2 correlation coefficient in worst performed cancer types (e.g., UVM, PAAD).

Another group followed up the above-mentioned study and modified the deep learning architecture to especially focus on breast cancer cases. 198 high-resolution WSIs from the Surveillance, Epidemiology, and End Results (SEER) dataset and 195 annotated TCGA breast cancer WSIs were utilized for the cancer detection task, and 1,090 breast cancer WSIs annotated from Saltz study were used for TIL classification task. The authors adapted and compared three different architectures including 16-layer VGG (VGG16), the 34-layer ResNet (ResNet34), and the Inception-v4 network using accuracy, F1 score, and AUC as performance metrics. Overall, the ResNet34 was the best performer in both cancer detection task and lymphocyte detection task, even surpassing the Saltz study's accuracy in the case of breast cancer. Using their ResNet34 model, Le et al. (2020) showed that their estimated TIL infiltration was a significant survival predictor.

In addition to assessing lymphocyte infiltration, users are also interested in evaluating tumor progression and heterogeneity based on the distribution of cancer cells of different grades in the whole slide. To this end, the Flores laboratory has developed a deep learning system-Grading of Lung Adenocarcinoma with Simultaneous Segmentation by an Artificial Intelligence (GLASS-AI) (Lockhart et al., 2021), based on preclinical lung adenocarcinoma models. A ResNet18-based CNN was trained to classify and map the normal lung tissue, normal airways, and the different grades (1–4) of lung adenocarcinoma in WSI of mouse lungs. The modal not only achieved a micro-F1 score of 0.81 on a pixel-by-pixel basis, but also uncovered a high degree of intratumor heterogeneity that was not reported by the pathologists. We are currently utilizing this pipeline in conjunction with spatial transcriptomic analysis and IHC to conduct mechanism investigations to reveal new therapeutic targets and prognostic markers.

Cell-Level Analysis Methods

Understanding the spatial organization of different cell types in the tumor microenvironment (TME) provides information on cancer progression, metastasis, and response to treatment. Currently, this information could be provided by extensive immunolabeling of specific cell types or performing spatial transcriptomics, though this technology is still in its infancy. To compensate this, researchers have been actively developed innovative approaches to extract cell-level information from images.

To provide a deeper understanding about the spatial information of cells involved in stromal-immune interface, Failmezger et al. (2020) developed CRImage, a computational pathology pipeline used to classify cells in the H&E-stained specimens into stromal, immune or cancer cells. The authors performed the analysis on 400 melanoma specimens obtained from TCGA. The authors compared the estimated proportions of these cell types with independent measures of tumor purity, estimation of lymphocyte density by expert raters, computed

immune cell types and pathway analyses. Using a set of independent single-cell annotations, the authors showed that the classifier to achieve 84.9% balanced accuracy (81.9% recall, 90.9% precision). By comparing the gene expression profiles of these samples, the authors demonstrated that samples with high lymphocyte percentage were enriched for immune-related pathways, validating the CRImage approach.

In another study, Wang et al. (2020) developed Histology-based Digital-Staining (HDS), a deep learning-based computation model, to segment the tumor, stroma, lymphocyte, macrophage, karyorrhexis, and red blood cell nuclei from standard H&E-stained pathology images. They applied HDS in lung adenocarcinoma H&E images to classify cell nuclei and extracted 48 cell spatial organization-related features that characterize the TME. Based on these features, they developed an accurate prognostic model that can predict high-risk group in the National Lung Screening Trial dataset, and further validated the model in the TCGA lung adenocarcinoma dataset. More importantly, they showed that these image-derived TME features significantly correlated with the gene expression of biological pathways. For example, transcriptional activation of both the T-cell receptor and programmed cell death protein 1 pathways positively correlated with the density of detected lymphocytes in tumor tissues, while expression of the extracellular matrix organization pathway positively correlated with the density of stromal cells. Taken together, they demonstrated that by applying HSD at cell-level analysis in H&E images, spatial organization of different cell types could be identified and associated with the gene expression of biological pathways.

AN OVERVIEW OF GENERATING WEAKLY CELL-LEVEL ANNOTATION USING IHC STAINED IMAGES

IHC-Based Cell-Level Annotation

Building a dataset using image-level, or region-level annotation is comparably easier than cell-level annotation. Unlike image-level annotation, which is generally reduced to a simple classification task, both region-level and cell-level annotation require the addition of a segmentation step alongside classification. Indeed, the principal difference between region- and cell-level classification is a matter of scale, though this difference is several orders of magnitude in size. Considering that a WSI can easily contain several million cells, completely annotating enough images to train the classifier is almost impossible even for experts. The difficulty in producing cell-level annotations for training data has compelled most groups to use region-level analysis of WSIs to capture cell-level features, such as presence of tumor infiltrating lymphocytes (Saltz et al., 2018). These analyses use region-level annotation and consider that the cells in the annotated region have the same cell types. These approaches can still be informative but considering that tumor is very heterogeneous, and there are multiple types of cells coexists even in a small single region, labeling all the cells in the same region will cause many mislabeled cell-level annotations.

As discussed in *Histopathology Image Data Preparation for Training Machine Learning Models* of this review, H&E staining is used frequently for basic examination of tissue and cell morphology. Immunolabeling may also be performed to obtain additional information from samples such as cells' subtypes. For example, pan-cytokeratin staining (PCK) is commonly used to stain for tumor cells, and some antibodies such as CD3 (T-cell), CD8 (T-cell), and CD20 (B-cell) are used for characterizing immune cells in the sample (**Figure 3**). Conventional immunohistochemistry (IHC), multiplexed IHC (mIHC), or multiplexed IF (mIF) images can be used for labeling multiple cells in histopathology images. When preparing sections from biopsies or resected tumors, slides will be prepared from a series of adjacent sections. This allows pathologists to easily compare regions of a H&E and adjacent immune-stained slides. Even though the adjacent samples are still showing similar spatial characteristics, they are not identical to the other samples. To perform machine learning data analysis and interpretation, it is critical to align these differently stained histopathology images together. Most conventional image analyses software can perform a reasonable job in aligning these different slices of images and provide a final aligned image to study tumor heterogeneity and tumor immune microenvironment.

To train machine learning classifier for cell-level annotation, images must first be annotated with individual cells' boundaries and the cells' subtypes. As explained earlier, cell-level annotation process requires a tremendous manual effort. To make this cell-level annotation more approachable and scalable, we introduce a semi-automated method of generating cell-level annotation using adjacent IHC-stained images as the labeled dataset for machine learning methods. In this example, we are using H&E tumor cell images with multiple IHC stained images to generate a labeled training dataset in Common Objects in Context (COCO) format for Mask R-CNN (He et al., 2017). After training on this automatically labeled training data, the machine learning classifier can predict cell types in H&E image without further IHC or IF stained images with high accuracy. In the following sections we will provide an explanation of the steps required in this approach, beginning with acquisition of the WSIs.

Digital Image Acquisition From Prepared Slides

Once the histology slides have been prepared using H&E or IHC, they must then be converted to a digital format for analysis by the machine learning model. While such images can be taken using a simple brightfield microscope, the quality and characteristics of the images can vary considerably from one slide to the next. Computer driven microscopes offer a better solution for reproducible image acquisition and can even be used to generate WSIs through stitching of image tiles. However, unless carefully calibrated the resulting WSI may contain significant stitching artifacts that can confuse a machine learning classifier. Instead, it is highly advisable to use a dedicated slide scanner, such as Leica's Aperio platform, to generate WSIs of stained slides. The H&E and mIHC images

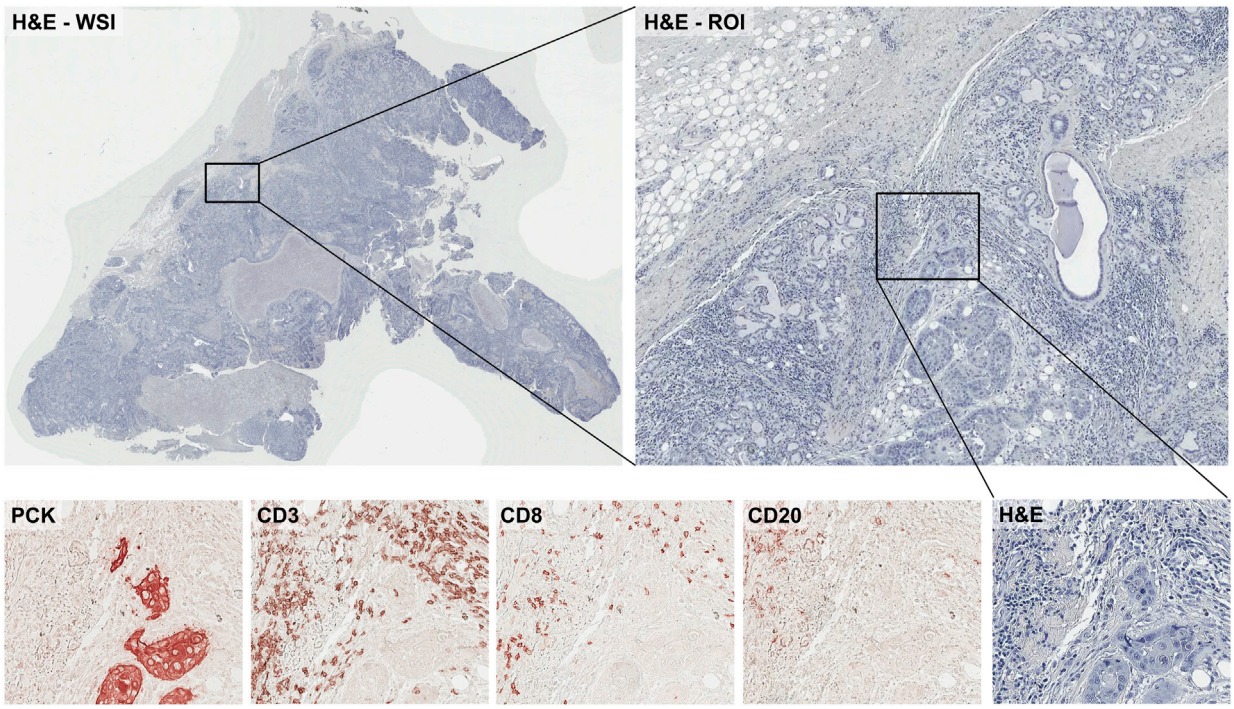


FIGURE 3 | Example of the IHC-stained histopathology images with different antibodies.

used in our example process were all captured on a Leica Aperio AT2 digital whole slide scanner at 20x magnification. This system produces a ".svs" file that contains an image of the slide label, a macro image of the entire slide, and a multi-resolution tiled ".tiff" of the WSI. The structure output files will vary among different slide scanning systems, but most rely on multi-resolution tiled TIFF files to store the WSI. For easier image modification and processing, svs files need to be converted to tiff or png formats.

While the example provided here relies on IHC stained slides, immunofluorescent (IF) labeling can also be used to generate cell-labeled slides. Most of the considerations described above for IHC apply to these approaches, but IF can generally provide a higher number of labels on a single slide. Systems like the Akoya Vectra or the Leica Aperio Versa are capable of scanning whole slides with up to 7-plex labeling. Spectral overlap of the fluorescent reporters presents a unique challenge in multiplexed IF, which was recently reviewed by Shakya et al. (2020). The plexity of both IHC and IF can be increased by sequential staining/de-staining for target proteins. Higher numbers of simultaneous labels (40+) can be achieved by newer techniques such as the Akoya CODEX or multiplexed ion beam imaging (MIBI) that use oligo- or metal-isotope-labeled antibodies, though these techniques require additional components or a specialized secondary ion mass spectroscopy system, respectively. It is also possible to label cells of interest by nucleic acid *in situ* hybridization (ISH). ISH and its fluorescent counterpart (FISH) are generally more laborious than immunolabeling and may not perform well on older samples due to the more sensitive nature of nucleic acids compared to proteins, but image acquisition is performed in the same manner as IHC or IF.

Image Registration and Normalization

After WSI images are stained with different antibodies are captured, a number of preprocessing steps should be followed to ensure optimal performance of the machine learning classifier, including stain normalization and slide registration. The adherence to a defined protocol for H&E and IHC staining is crucial to minimize the variability between batches of slides, but some level of variability will still occur due to imperfections in tissue sectioning or changes in tissue composition. Staining variability can be compensated for after image acquisition by normalization to a standard using several suggested methods (Macenko et al., 2009; Khan et al., 2014; Bejnordi et al., 2016; Alsubaie et al., 2017; Roy et al., 2018; Anghel et al., 2019). Stain normalization is especially important when incorporating external datasets due to potential differences in staining protocols and image capture systems.

Depending on what sampling/staining method is used, different steps of preprocessing are needed to generate aligned or registered image dataset. Conventional IHC uses adjacent slices of the samples for each staining, and they are not identical. Also, during the sampling, placing, staining, and processing, each section might be moved, rotated, mirrored, or simply imperfectly placed onto the slide which may causes images to be poorly registered (Wang et al., 2014). Since mIHC is stained for the same identical sample for multiple times, the images are comparably more aligned compared to conventional IHC

datasets. However, during the staining-washing-placing steps in each repetitive staining, the samples can be slightly displaced or even washed off. Because of that, mIHC images may still not perfectly align and should be registered as well. mIF datasets, in contrast, does not require repetitive washing and staining steps and should therefore be easier to register to adjacent slides.

Image registration is common step in image processing using multiple images especially in biomedical image analysis such as radiology. It is a process of overlaying two or more images from different sources or different time of the same object to align geometrically (Zitová and Flusser 2003). In radiology, this method is used for overlaying images from different sensors, different equipment, or different time (Fox et al., 2008; Tohka and Toga, 2015). This image registration is sometimes done manually using image viewer software when there are not many images, but for multiple image files, automated methods can be used. ImageJ (Rueden et al., 2017) and Fiji (Schindelin et al., 2012) have multiple registration plug-ins such as "Feature Extraction SIFT/MOPS", or "TrakEM2". MatLab also has several applications including "Registration Estimator App" and "Intensity-Based Automatic Image Registration". SimpleITK (Lowekamp et al., 2013) is an open-source image analysis toolkit that supports multiple platform such as Python, R, Java, C#, C++, and Ruby that provides powerful machine learning-based registration options for image registration process. In most of the automated methods, images are aligned globally, which means at a smaller scale parts of the images might be not very well aligned. In this case, additional registration steps can be done after deciding ROIs or split the samples in smaller patches.

Cell Boundary Detection and IHC Intensity Level Acquisition

For the cell boundary information, many (semi-)automated cell boundary recognition programs can be used. CellProfiler is one of the most well-known histopathology image analysis program (McQuin et al., 2018). CellProfiler supports multiple types of histopathology images as input, and users can build their own pre/post-processing pipeline of the image quantification. Using CellProfiler, cell boundaries can be segmented by setting up some simple parameters such as the typical diameters of the nuclei (or cells), segmentation thresholding methods (e.g., Otsu, Minimum Cross-entropy), smoothing thresholds etc.

No cell segmentation program is perfectly accurate, and depending on the image dataset that are used for this process, the parameters may need to be modified. For example, the cell sizes in the images can vary depending on the zoom level of images, the sampled tissue parts of the organs and species. Finding good parameters to detect the cells with higher accuracy is an important step. For example, the cell smoothing threshold parameter is too high, multiple different cells can be recognized in a single cell. However, if the threshold is too low, a single cell is seldom recognized as multiple cells. These parameters are required to be optimized based on the target image dataset. This might be repetitive and time-consuming part of the image processing, however, finding optimal parameters for each input dataset are important for accurately finding the cell boundaries from the image. When the cell boundaries are detected, CellProfiler outputs the results in a mask image file. This mask image file is showing the detected cells in different colors, and the background as black.

CellProfiler also has the tools to measure staining intensities of each recognized cells from multiple aligned IHC images. For example, after CellProfiler detects cells boundaries in the main input image (e.g., H&E), the staining intensities of the same location are extracted from other IHC images (e.g., PCK, CD14). The intensities in aligned IHC images are obtained in a table with csv format. This table contains the x/y coordinates of the detected cells, and their cell marker intensities obtained from different IHC images. After semi-automated image processing steps, the information of the detected cell boundaries and the IHC intensities of the cells could be saved in mask image files and a csv file, respectively.

Combine the CellProfiler Information and Decide Cell Class Subtypes

The cell boundary information, and the intensities of each markers in the cells are obtained as two different types of files from the previous step. In this step, the intensity levels need to be converted to cell class labels. Depending on a prior knowledge of domain experts, the rule for cell class labeling can be made. For example, a cell with a high intensity level in PCK stained image will be labeled as a tumor cell. During this labeling rule generation, some markers will require a threshold values to divide positive and negative class cells. Many of the marker values can be shown bimodal distribution in histograms. The local minimum value in between the two modes can be used as the cutoff point for detecting positive/negative staining intensities (usually this cutoff point resides between 0.3 and 0.5). Some of the staining markers may not have bimodal distributions, but have different distributions (e.g., unimodal or multi-modal). In these cases, additional steps may require to establish a biologically-informed decision process based on the advice of domain experts. Following the application of the cell class decision rules, each cell in the csv files will be mapped to a single class label (e.g., tumor cell, immune cell, stromal cell).

Reformatting the Annotation Dataset Into Common Objects in Context challenge Format

After the cell subtypes are decided, the cell mask image (cell boundary information) and the csv information (cell subtype information) need to be combined and reformatted. The most commonly used annotation file format is called COCO format, which is used in Microsoft's Common Objects in Context challenge (COCO)[1] (Lin, Maire et al., 2014). The COCO dataset is one of the most popular object detection datasets with multiple different objects' images and their annotations.

[1]https://cocodataset.org

The annotation file format contains image names, object boundary locations, objects' class name, and the objects' class label of all the detected objects. Most of the current machine learning-based object detection methods[2,3,4] are using COCO format as the input of the annotation information, and there are multiple publicly available tools for generating, loading and modifying such as COCO API and FiftyOne.

Since the output of CellProfiler is not in COCO format but the mask images, the files cannot be used directly by most of the known image analysis methods. To convert the CellProfiler outputs to COCO format JSON file, the following processing steps are needed. First, the cell mask images must be converted into the list of x/y coordinates that represents the boundary of each cell. FiftyOne which is a Python open-source tool for image dataset building supports the input of mask images and can convert them into coordinates in the image. After the cells' boundaries are mapped to cell subtype classes, this information needs to be saved as COCO JSON format.

Through this step, the cell mask image and cell class information are converted to a single JSON format that includes cell boundary coordinates, cell classes, and the input image file information.

Divide Images Into Patches/Train + Test Dataset Generation

Now, the input images with cell-level annotation dataset are ready for use. However, depending on the input image size, the images need to be split into smaller patches or tiles. Most of the deep learning-based image segmentation methods require GPUs with high memory because of the amount of computations in the complex neural network structure. Depending on the available GPU memories of the machine, the size of the input images needs to be modified. Most of the popular deep learning methods take images from 128X128 pixels to 1024X1024 pixels, depending on the parameter settings of the code or GPU memory of the machine. WSI images are several orders of magnitude larger than an individual image patch, and even most of the ROI sections are still bigger than this range. Splitting images is easily accomplished using many of publicly available tools such as Pillow in Python; however, COCO annotation file needs to be re-generated by calculating the new coordinates of the cell boundaries in the split images. Since the cells on the edge of the image are not easily segmented by machine learning methods, it is recommended that the image splitting allows some overlaps between the image patches.

Also, as it is common in machine learning and data science field, before training the dataset with machine learning methods, the dataset needs to be separated into training, testing, and sometimes validating datasets. It is important that the training/testing data separation needs to be done in WSI level, not patch level–which means the patches from the same WSI must not co-exist in both train/test dataset. Adjacent patches from the same WSI can be overlapped, or shares many properties such as shapes, colors, and patterns, which can cause a boosted accuracy scores in the test dataset because the machine learning classifier already have seen very similar (adjacent or overlapped) cells or tissues.

Training Machine Learning Classifier

Once the dataset is ready, it is time to train a machine learning classifier. There are many machine learning classifiers[5,6,7] designed for image segmentation and classification for the COCO challenge, and by changing some parameter settings most of them can be used for this histopathology cell subtype segmentation task.

Deep learning-based machine learning classifiers usually require a large amount of training dataset. If the training dataset is not big enough, a "warm-start" method (pre-training/fine tuning) is highly recommended. For the popular machine learning models, there are pre-trained weights that are trained with big datasets are publicly available. For example, Wang et al. used a pre-trained Mask R-CNN model with COCO dataset and public balloon image dataset, and fine-tuned with their histopathology image dataset for cell segmentation task (Wang et al., 2020). This fine-tuning method (pre-train with big dataset/fine-tune with final dataset) is widely used to overcome the limitation of the lack of training dataset, especially for deep learning model training that requires big training dataset. For Mask R-CNN method, several pretrained weights are publicly available.[8,9]

As the goal of the tasks are not the same between balloon detection and cell segmentation, many of the parameters in Mask R-CNN codes needs to be updated to optimize the machine learning classifier to detect cells more accurately before training with the target dataset. For example, in histopathology images there are almost certainly a higher number of the target objects per image (e.g., nuclei or cells) than in other image sets. Therefore, the maximum detection threshold needs to be changed to higher number and the target sizes should be set to smaller, respectively. Also, target objects are taking up much more space in the histopathology images compared to other segmentation tasks, so changing ROI positive ratio will be helpful.

For cell-type detection, class imbalance problem is a known issue in the histopathology analysis. For example, in tumor sample slides, most of the cells are tumors, but only a limited number of cells are immune cells. As class imbalance is a well-known problem in machine learning field, there are several suggestions to solve this problem including over/under sampling, using different cost/weight schema during the training, (Almeida et al., 2014; Johnson and Khoshgoftaar 2019). If an insufficient number of images are included in the training dataset, image augmentation can be employed to

[2]https://github.com/matterport/Mask_RCNN
[3]https://github.com/facebookresearch/Detectron
[4]https://github.com/endernewton/tf-faster-rcnn

[5]https://github.com/matterport/Mask_RCNN
[6]https://github.com/facebookresearch/Detectron
[7]https://github.com/endernewton/tf-faster-rcnn
[8]https://github.com/matterport/Mask_RCNN/releases/download/v2.0/mask_rcnn_coco.h5
[9]https://github.com/matterport/Mask_RCNN/releases/download/v2.1/mask_rcnn_balloon.h5

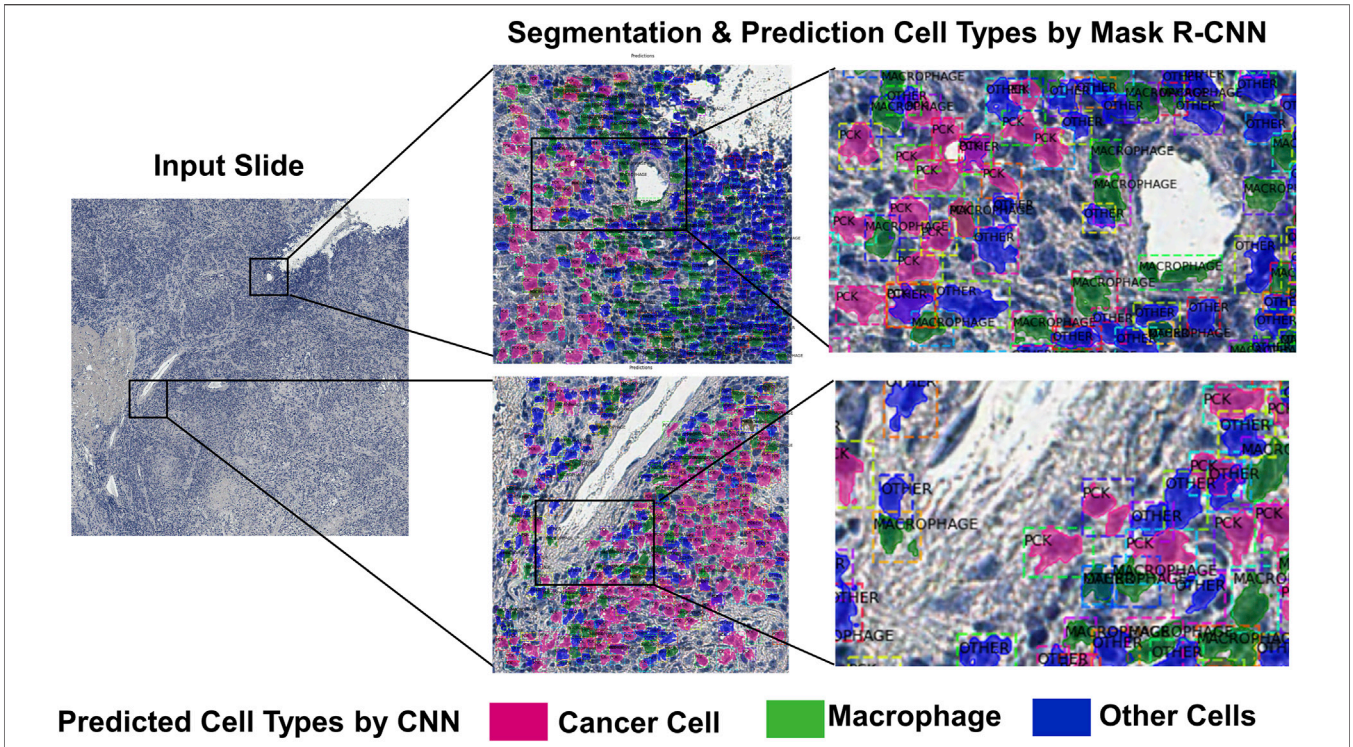


FIGURE 4 | An example of the deep learning-based cell subtype prediction results.

synthetically increase the dataset size. Image augmentation generates more images for training dataset by image alteration (e.g., rotation, flip, blur, crop, pad, or adding noise) and is widely used for deep learning methods (Shorten and Khoshgoftaar 2019). In the case of Mask R-CNN, an image augmentation step is built into the pipeline.

In histopathology image method training, there are several things to consider for getting good accuracy of cell segmentation and subtype prediction. Since deep learning methods require very high number of calculation and high memory during training the classifier, it is almost impossible to train classifiers without high performance hardware with GPUs and high memory. Depending on the performance of the hardware and time limitations, the training parameters (such as learning rates, epochs, the layers to be fine-tuned, etc.) require tuning to optimize the performance.

Evaluation of the Results

After obtaining the prediction results in the test dataset, the results need to be evaluated. To evaluate how accurately the machine learning classifier can find cells' boundaries and their subtypes, the Intersection over Union (IoU) metric, also known as Jaccard index can be used (Girshick 2015; Ren et al., 2015; He et al., 2017). For each class of subtypes, the intersection of the ground truth area and the predicted area (Rezatofighi et al., 2019). Usually, IoU is calculated based on the bounding box of the segmentation. When IoU is calculated in an image-level, it can be calculated for each class and averaged to see the final image segmentation accuracy for all the classes. This score is called mean-IoU (mIoU). When IoU is used for binary decision of a single object detection, if the IoU is higher than 0.5–0.7, it is considered that the object is correctly detected.

$$IoU = \frac{\#TP}{\#TP + \#FP + \#FN} = \frac{Area\ of\ Ground\ Truth \cap Area\ of\ Predicted}{Area\ of\ Ground\ Truth \cup Area\ of\ Predicted}. \quad (1)$$

Ultimately, validation and verification of the prediction by a pathologist remains the gold-standard for histopathology image analysis.

Visualization and Obtaining Biological Insights

As prediction of the cell subtypes are performed on the image patches, the predictions need to be stitched and rebuilt into the original image. As the classifier predicts the cell boundaries and subtypes, it is possible to overlap the predicted cell segments on the original images with different colors depending on the class. (**Figure 4**). This can give clinicians the information of tumor heterogeneity or immune infiltration in the sample.

There are several methods to quantify the tumor heterogeneity in the output image. Ripley's K function (Ripley 1988) that is a function of calculating spatial point pattern by finding number of other points within a circle of radius, is used for several histopathology image analysis (Mattfeldt et al., 2009; Yuan et al., 2012; Carstens et al., 2017). The Shannon diversity index (Shannon 1948) is also widely used for calculating the diversity of molecules or distribution of species. Graf and

Zavodszky suggested a method for calculating molecular entropy and heterogeneity diversity metrics based on the Shannon diversity index for quantifying tumor heterogeneity (Graf and Zavodszky 2017). These methods can be calculated at the patch-level or for image-level analysis using sliding window.

Pros and Cons of IHC-Based Weakly Cell-Level Annotation

Using registered H&E and IHC slides can provide a good alternative to cell-level manual annotations. This technique allows for easy production of a large dataset of "weakly labeled" images with a much better annotation resolution than single class region annotation for cell-level machine learning classifiers. Indeed, for applications with rare or non-contiguous organization, such as identification of immune cell types within tumours, region-level annotations are unlikely to produce a usable training set. However, cell-level registration has its own limitations on the accuracy of the resulting labels. In the simplest case, cell-level registration with a single marker on each slide from typical IHC, it may be necessary to register multiple slides of various distances to a single H&E image. It stands to reason that the greater the distance between the two registered slides, the less accurate the registration will be. Multiplexed or sequential immunolabeling, such as the mIHC approach used in our example, can mitigate this loss of accuracy. If only a few labels are needed to construct the training dataset it would be possible to generate a nearly perfect cell-level registration by digitizing the H&E slide, de-staining, and then re-staining using IHC (Hinton et al., 2019).

There are several other difficulties to consider while designing the machine learning models. Cell-level histopathology analysis can be prone to cell class imbalance problem, because some of the specific types of cells might be very rare compared to the other tumor or normal cells. There are several methods to handle class imbalance in machine learning (Johnson and Khoshgoftaar 2019), however, manually checking the distribution of the class labels are recommended. Manual validation process is also recommended to remove some low-quality images or images with noises. For example, sometimes during the staining or washing step of the sample preparation, some samples are missing, or air bubble can be formed in the image. These noises can cause bias in the training steps, and manual validation of images is recommended to remove these low-quality samples.

FUTURE DIRECTION OF THE HISTOPATHOLOGY IMAGE ANALYSIS AT SINGLE-CELL LEVEL WITH DEEP LEARNING

As shown in several publications, deep learning-based histopathology analysis is becoming a useful tool for tumor image data analysis. Most of the publication are focusing on classification of the sample in image, region or cell level, however, this deep learning-based histopathology techniques can be expanded into more broad topics of research.

Machine learning-based image analysis can be used to predict transcriptomics information from the image. Schmauch et al. (2020) demonstrated a machine learning model trained with bulk RNAseq data to predict spatial gene expression from H&E WSI images. With the recent development of spatial transcriptomics techniques such as LCM-seq or 10X Visium Spatial Gene Expression, it is possible to obtain the gene expression of the small region in the sample with the histopathology image together. These spatial transcriptomics data with paired histopathology images can be a great training dataset for gene expression prediction from histopathology image of the sample. With this dataset of spatial gene expression and images with deep learning methods make it available to predict gene expression of the specific region from its histopathology image. He et al. (2020) used spatial genomics dataset of 23 breast cancer patients to predict gene expression from the histopathology of small regions. The results are promising though limited because of the small size of the dataset. After collecting more dataset from these spatial genomics datasets with images, the gene expression prediction will certainly be improved. The scope of these prediction will be eventually combined with single cell RNA-seq dataset for cell-level prediction. Currently, the technical and cost limitation of single cell technology, it is almost impossible to generate enough dataset for cell-level annotations for deep learning dataset, however, in the future after the limitation is removed, cell-level expression prediction will be possible.

As histopathology images include a lot of information, and it can be used for predicting further information. Histopathology images may include rich information such as tumor grade, tumor subtype, immune infiltration, and so on. This information may give clinicians some idea for finding the personalized treatment for each patient. As an example, Wang et al. (2020) predicted overall survival of the patients of the samples using deep learning-method.

As explained earlier, small training dataset problem can be mitigated by training machine learning models with a big dataset that is similar to the target dataset followed by fine tuning with the target dataset. For this purpose of pretraining deep learning methods, public dataset of histopathology images will be very helpful for researchers who are having lack of training dataset issues. Komura et al. introduced several WSI datasets that are publicly available (Komura and Ishikawa 2018), and The Cancer Image Archive (TCIA) is also a good place to find the similar image datasets to pretrain the model. One major problem of these publicly available image dataset is that the images are generated and processed in many ways, which makes it harder for combining them to construct a big training dataset. Some standardized steps of generating and pre-processing image datasets will make the public image dataset more valuable and easier to use for machine learning model training.

In conclusion, building a machine learning model for histology image analysis requires a significant investment of time and effort on both the computational and the biological side. A basic understanding of the biological and data science principles that underpin these methods is key to establishing a productive multi-disciplinary team of researchers for this promising and rapidly growing field. In addition, as the amount of WSIs and associated molecular data becoming widely available to researchers, the development and application of computational approaches will become more robust and reproducible. We are optimistic that these

computational approaches will play an important role to uncover the insights contained in these histopathology image datasets.

AUTHOR CONTRIBUTIONS

KL, JHL, ACT, and MX wrote the manuscript. ACT, KL, and JHL designed and conducted the study. RJCS, RC, and JHL obtained and processed the image data. ACT, CHC, and ERF supervised the project. All the authors revised and approved the final manuscript.

FUNDING

This work was partly supported by the National Institutes of Health (NIH) under Award Numbers P30CA076292, P01CA250984, R01DE030508 and the James and Esther King Biomedical Research Grant (21K04). KL and JHL are postdoctoral fellows of the ICADS T32 postdoctoral training program at the Moffitt Cancer Center (NCI T32 CA233399). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

REFERENCES

Almeida, H., Meurs, M.-J., Kosseim, L., Butler, G., and Tsang, A. (2014). Machine Learning for Biomedical Literature Triage. *PLoS One* 9 (12), e115892. doi:10.1371/journal.pone.0115892

Alsubaie, N., Trahearn, N., Raza, S. E. A., Snead, D., and Rajpoot, N. M. (2017). Stain Deconvolution Using Statistical Analysis of Multi-Resolution Stain Colour Representation. *PLOS ONE* 12 (1), e0169875. doi:10.1371/journal.pone.0169875

Amgad, M., Elfandy, H., Hussein, H., Atteya, L. A., Elsebaie, M. A. T., Abo Elnasr, L. S., et al. (2019). Structured Crowdsourcing Enables Convolutional Segmentation of Histology Images. *Bioinformatics* 35 (18), 3461–3467. doi:10.1093/bioinformatics/btz083

Anghel, A., Stanisavljevic, M., Andani, S., Papandreou, N., Rüschoff, J. H., Wild, P., et al. (2019). A High-Performance System for Robust Stain Normalization of Whole-Slide Images in Histopathology. *Front. Med. (Lausanne)* 6, 193. doi:10.3389/fmed.2019.00193

Carstens, J. L., Correa de Sampaio, P., Yang, D., Barua, S., Wang, H., Rao, A., et al. (2017). Spatial Computation of Intratumoral T Cells Correlates with Survival of Patients with Pancreatic Cancer. *Nat. Commun.* 8 (1), 15095. doi:10.1038/ncomms15095

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., et al. (2013). The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J. Digit Imaging* 26 (6), 1045–1057. doi:10.1007/s10278-013-9622-7

Dimitriou, N., Arandjelović, O., and Caie, P. D. (2019). Deep Learning for Whole Slide Image Analysis: An Overview. *Front. Med.* 6, 264. doi:10.3389/fmed.2019.00264

Ehteshami Bejnordi, B., Litjens, G., Timofeeva, N., Otte-Höller, I., Homeyer, A., Karssemeijer, N., et al. (2016). Stain Specific Standardization of Whole-Slide Histopathological Images. *IEEE Trans. Med. Imaging* 35 (2), 404–415. doi:10.1109/tmi.2015.2476509

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature* 542 (7639), 115–118. doi:10.1038/nature21056

Failmezger, H., Muralidhar, S., Rullan, A., de Andrea, C. E., Sahai, E., and Yuan, Y. (2020). Topological Tumor Graphs: A Graph-Based Spatial Model to Infer Stromal Recruitment for Immunosuppression in Melanoma Histology. *Cancer Res.* 80 (5), 1199–1209. doi:10.1158/0008-5472.can-19-2268

Fox, T., Elder, E., Crocker, I., Paulino, A. C., and Philadelphia, B. S. Teh. (2008). "Image Registration and Fusion Techniques," in *Radiotherapy Treatment Planning* (Elsevier), 35–51. doi:10.1016/b978-1-4160-3224-3.50006-2

Girshick, R. (2015). "Fast R-CNN." arXiv:1504.08083. doi:10.1109/iccv.2015.169

Graf, J. F., and Zavodszky, M. I. (2017). Characterizing the Heterogeneity of Tumor Tissues from Spatially Resolved Molecular Measures. *PLOS ONE* 12 (11), e0188878. doi:10.1371/journal.pone.0188878

He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, A., et al. (2020). Integrating Spatial Gene Expression and Breast Tumour Morphology via Deep Learning. *Nat. Biomed. Eng.* 4 (8), 827–834. doi:10.1038/s41551-020-0578-x

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). *Mask R-CNN*. arXiv: 1703.06870.

Hinton, J. P., Dvorak, K., Roberts, E., French, W. J., Grubbs, J. C., Cress, A. E., et al. (2019). A Method to Reuse Archived H&E Stained Histology Slides for a Multiplex Protein Biomarker Analysis. *Methods Protoc.* 2 (4). doi:10.3390/mps2040086

Janowczyk, A., and Madabhushi, A. (2016). Deep Learning for Digital Pathology Image Analysis: A Comprehensive Tutorial with Selected Use Cases. *J. Pathol. Inform.* 7 (1), 29. doi:10.4103/2153-3539.186902

Johnson, J. M., and Khoshgoftaar, T. M. (2019). Survey on Deep Learning with Class Imbalance. *J. Big Data* 6 (1), 27. doi:10.1186/s40537-019-0192-5

Kalra, S., Tizhoosh, H. R., Shah, S., Choi, C., Damaskinos, S., Safarpoor, A., et al. (2020). Pan-cancer Diagnostic Consensus through Searching Archival Histopathology Images Using Artificial Intelligence. *NPJ Digit Med.* 3, 31–15. doi:10.1038/s41746-020-0238-2

Kalra, S., Tizhoosh, H. R., Choi, C., Shah, S., Diamandis, P., Campbell, C. J. V., et al. (2020). Yottixel - an Image Search Engine for Large Archives of Histopathology Whole Slide Images. *Med. Image Anal.* 65, 101757. doi:10.1016/j.media.2020.101757

Khan, A. M., Rajpoot, N., Treanor, D., and Magee, D. (2014). A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-specific Color Deconvolution. *IEEE Trans. Biomed. Eng.* 61 (6), 1729–1738. doi:10.1109/tbme.2014.2303294

Komura, D., and Ishikawa, S. (2018). Machine Learning Methods for Histopathological Image Analysis. *Comput. Struct. Biotechnol. J.* 16, 34–42. doi:10.1016/j.csbj.2018.01.001

Le, H., Gupta, R., Hou, L., Abousamra, S., Fassler, D., Torre-Healy, L., et al. (2020). Utilizing Automated Breast Cancer Detection to Identify Spatial Distributions of Tumor-Infiltrating Lymphocytes in Invasive Breast Cancer. *Am. J. Pathol.* 190 (7), 1491–1504. doi:10.1016/j.ajpath.2020.03.012

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep Learning. *Nature* 521 (7553), 436–444. doi:10.1038/nature14539

Lee, K., Kim, B., Choi, Y., Kim, S., Shin, W., Lee, S., et al. (2018). Deep Learning of Mutation-Gene-Drug Relations from the Literature. *BMC Bioinformatics* 19 (1), 21. doi:10.1186/s12859-018-2029-1

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., et al. (2014). "Microsoft COCO: Common Objects in Context." arXiv:1405.0312. doi:10.1007/978-3-319-10602-1_48

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., et al. (2018). 1399 H&E-stained sentinel Lymph Node Sections of Breast Cancer Patients: the CAMELYON Dataset. *GigaScience* 7 (6). doi:10.1093/gigascience/giy065

Liu, R., Rizzo, S., Whipple, S., Pal, N., Pineda, A. L., Lu, M., et al. (2021). Evaluating Eligibility Criteria of Oncology Trials Using Real-World Data and AI. *Nature* 592 (7855), 629–633. doi:10.1038/s41586-021-03430-5

Lockhart, J. H., Ackerman, H. D., Lee, K., Abdalah, M., Davis, A., Montey, N., et al. (2021). Abstract PO-082: Automated Tumor Segmentation, Grading, and Analysis of Tumor Heterogeneity in Preclinical Models of Lung Adenocarcinoma. *Clin. Cancer Res.* 27 (5 Suppl. ment). PO-082-PO-082. doi:10.1158/1557-3265.adi21-po-082

Lowekamp, B. C., Chen, D. T., Ibáñez, L. D., and Blezek, D. (2013). The Design of SimpleITK. *Front. Neuroinform* 7 (45), 45. doi:10.3389/fninf.2013.00045

Lu, M. Y., Chen, T. Y., Williamson, D. F. K., Zhao, M., Shady, M., Lipkova, J., et al. (2021). AI-based Pathology Predicts Origins for Cancers of Unknown Primary. *Nature* 594 (7861), 106–110. doi:10.1038/s41586-021-03512-4

Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Xiaojun, G., et al. (2009). A Method for Normalizing Histology Slides for Quantitative Analysis. In 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. Boston, MA: IEEE. doi:10.1109/ISBI.2009.5193250

Mattfeldt, T., Eckel, S., Fleischer, F., and Schmidt, V. (2009). Statistical Analysis of Labelling Patterns of Mammary Carcinoma Cell Nuclei on Histological Sections. *J. Microsc.* 235 (1), 106–118. doi:10.1111/j.1365-2818.2009.03187.x

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., et al. (2020). International Evaluation of an AI System for Breast Cancer Screening. *Nature* 577 (7788), 89–94. doi:10.1038/s41586-019-1799-6

McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B. A., Karhohs, K. W., et al. (2018). CellProfiler 3.0: Next-Generation Image Processing for Biology. *Plos Biol.* 16 (7), e2005970. doi:10.1371/journal.pbio.2005970

Nagpal, K., Foote, D., Tan, F., Liu, Y., Chen, P.-H. C., Steiner, D. F., et al. (2020). Development and Validation of a Deep Learning Algorithm for Gleason Grading of Prostate Cancer from Biopsy Specimens. *JAMA Oncol.* 6, 1372–1380. doi:10.1001/jamaoncol.2020.2485

Petrick, N., Akbar, S., Cha, K. H., Nofech-Mozes, S., Sahiner, B., Gavrielides, M. A., et al. (2021). SPIE-AAPM-NCI BreastPathQ challenge: an Image Analysis challenge for Quantitative Tumor Cellularity Assessment in Breast Cancer Histology Images Following Neoadjuvant Treatment. *J. Med. Imaging (Bellingham)* 8 (3), 034501. doi:10.1117/1.jmi.8.3.034501

Prior, F. W., Clark, K., Commean, P., Freymann, J., Jaffe, C., Kirby, J., et al. (2013). TCIA: An Information Resource to Enable Open Science. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2013, 1282–1285. doi:10.1109/EMBC.2013.6609742

Ren, S., He, K., Girshick, R., and Sun, J. (2015). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." arXiv:1506.01497.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. (2019). *Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression*, 09630. arXiv:1902.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Roohi, A., Faust, K., Djuric, U., and Diamandis, P. (2020). Unsupervised Machine Learning in Pathology. *Surg. Pathol. Clin.* 13, 349–358. doi:10.1016/j.path.2020.01.002

Roy, S., kumar Jain, A., Lal, S., and Kini, J. (2018). A Study about Color Normalization Methods for Histopathology Images. *Micron* 114, 42–61. doi:10.1016/j.micron.2018.07.005

Rueden, C. T., Schindelin, J., Hiner, M. C., DeZonia, B. E., Walter, A. E., Arena, E. T., et al. (2017). ImageJ2: ImageJ for the Next Generation of Scientific Image Data. *BMC Bioinformatics* 18 (1), 529. doi:10.1186/s12859-017-1934-z

Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., et al. (2018). Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep* 23 (1), 181–e7. doi:10.1016/j.celrep.2018.03.086

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., et al. (2012). Fiji: an Open-Source Platform for Biological-Image Analysis. *Nat. Methods* 9 (7), 676–682. doi:10.1038/nmeth.2019

Schmauch, B., Romagnoni, A., Pronier, E., Saillard, C., Maillé, P., Calderaro, J., et al. (2020). A Deep Learning Model to Predict RNA-Seq Expression of Tumours from Whole Slide Images. *Nat. Commun.* 11 (1), 3877. doi:10.1038/s41467-020-17678-4

Serag, A., Ion-Margineanu, A., Qureshi, H., McMillan, R., Saint Martin, M.-J., Diamond, J., et al. (2019). Translational AI and Deep Learning in Diagnostic Pathology. *Front. Med.* 6, 185. doi:10.3389/fmed.2019.00185

Shakya, R., Nguyen, T. H., Waterhouse, N., and Khanna, R. (2020). Immune Contexture Analysis in Immuno-Oncology: Applications and Challenges of Multiplex Fluorescent Immunohistochemistry. *Clin. Transl Immunol.* 9 (10), e1183. doi:10.1002/cti2.1183

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 27 (3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Shorten, C., and Khoshgoftaar, T. M. (2019). A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* 6 (1), 60. doi:10.1186/s40537-019-0197-0

Tizhoosh, H. R., Diamandis, P., Campbell, C. J. V., Safarpoor, A., Kalra, S., Maleki, D., et al. (2021). Searching Images for Consensus: Can AI Remove Observer Variability in Pathology? *Am. J. Pathol.*

Tohka, J., and Toga, A. W. (2015). *Rigid-Body Registration. Brain Mapping.* Waltham: Academic Press, 301–305. doi:10.1016/b978-0-12-397025-1.00299-2

van der Laak, J., Litjens, G., and Ciompi, F. (2021). Deep Learning in Histopathology: the Path to the Clinic. *Nat. Med.* 27 (5), 775–784. doi:10.1038/s41591-021-01343-4

Wang, C.-W., Ka, S.-M., and Chen, A. (2014). Robust Image Registration of Biological Microscopic Images. *Sci. Rep.* 4 (1), 6050. doi:10.1038/srep06050

Wang, S., Rong, R., Yang, D. M., Fujimoto, J., Yan, S., Cai, L., et al. (2020). Computational Staining of Pathology Images to Study the Tumor Microenvironment in Lung Cancer. *Cancer Res.* 80 (10), 2056–2066. doi:10.1158/0008-5472.can-19-1629

Wang, S., Wang, T., Yang, L., Yang, D. M., Fujimoto, J., Yi, F., et al. (2019). ConvPath: A Software Tool for Lung Adenocarcinoma Digital Pathological Image Analysis Aided by a Convolutional Neural Network. *EBioMedicine* 50, 103–110. doi:10.1016/j.ebiom.2019.10.033

Yuan, Y., Failmezger, H., Rueda, O. M., Ali, H. R., Gräf, S., Chin, S.-F., et al. (2012). Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Genomic Profiling. *Sci. Translational Med.* 4 (157), 157ra143. doi:10.1126/scitranslmed.3004330

Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., et al. (2019). Pathologist-level Interpretable Whole-Slide Cancer Diagnosis with Deep Learning. *Nat. Mach Intell.* 1 (5), 236–245. doi:10.1038/s42256-019-0052-1

Zitová, B., and Flusser, J. (2003). Image Registration Methods: a Survey. *Image Vis. Comput.* 21 (11), 977–1000. doi:10.1016/s0262-8856(03)00137-9

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

TECHNOLOGY AND CODE
published: 16 September 2021
doi: 10.3389/fdata.2021.725276

Statistical Enrichment Analysis of Samples: A General-Purpose Tool to Annotate Metadata Neighborhoods of Biological Samples

*Thanh M. Nguyen[1], Samuel Bharti[2], Zongliang Yue[1], Christopher D. Willey[3†] and Jake Y. Chen[1]**

[1]Informatics Institute, School of Medicine, The University of Alabama at Birmingham, Birmingham, AL, United States, [2]Centre for Computational Biology and Bioinformatics, Amity Institute of Biotechnology, Amity University, Noida, India, [3]Department of Radiation Oncology, School of Medicine, The University of Alabama at Birmingham, Birmingham, AL, United States

OPEN ACCESS

Edited by:
Huixiao Hong,
United States Food and Drug Administration, United States

Reviewed by:
Ehsan Ullah,
Qatar Computing Research Institute, Qatar
Shailesh Tripathi,
Tampere University of Technology, Finland

***Correspondence:**
Jake Y. Chen
jakechen@uab.edu

†ORCID:
Christopher D. Willey
orcid.org/0000-0001-9953-0279

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: 15 June 2021
Accepted: 06 September 2021
Published: 16 September 2021

Citation:
Nguyen TM, Bharti S, Yue Z, Willey CD and Chen JY (2021) Statistical Enrichment Analysis of Samples: A General-Purpose Tool to Annotate Metadata Neighborhoods of Biological Samples. Front. Big Data 4:725276. doi: 10.3389/fdata.2021.725276

Unsupervised learning techniques, such as clustering and embedding, have been increasingly popular to cluster biomedical samples from high-dimensional biomedical data. Extracting clinical data or sample meta-data shared in common among biomedical samples of a given biological condition remains a major challenge. Here, we describe a powerful analytical method called Statistical Enrichment Analysis of Samples (SEAS) for interpreting clustered or embedded sample data from omics studies. The method derives its power by focusing on sample sets, i.e., groups of biological samples that were constructed for various purposes, e.g., manual curation of samples sharing specific characteristics or automated clusters generated by embedding sample omic profiles from multi-dimensional omics space. The samples in the sample set share common clinical measurements, which we refer to as "clinotypes," such as age group, gender, treatment status, or survival days. We demonstrate how SEAS yields insights into biological data sets using glioblastoma (GBM) samples. Notably, when analyzing the combined The Cancer Genome Atlas (TCGA)—patient-derived xenograft (PDX) data, SEAS allows approximating the different clinical outcomes of radiotherapy-treated PDX samples, which has not been solved by other tools. The result shows that SEAS may support the clinical decision. The SEAS tool is publicly available as a freely available software package at https://aimed-lab.shinyapps.io/SEAS/.

Keywords: sample enrichment analysis, clinotype, SEAS, glioblastoma multiforme, patient-derived xenograft, patient-derived xenograft

INTRODUCTION

Systematic software platforms to organize large metadata and clinical data [also called "clinotype" (Nguyen et al., 2021)] is essential in biomedical research (Burgun and Bodenreider, 2008; Ohmann and Kuchinke, 2009). These software platforms, such as (Ta et al., 2018; Kim et al., 2019; Hume et al., 2020), have two key objectives. First, it allows the biomedical researcher to perform manual cohort selection quickly. Here, the researcher inputs the filtering query and gets the data from all patients meeting the filtering criteria. Second, it allows quick data exploration, including data visualization and simple aggregated analysis. Here, the researcher may view the basic characteristic of the selected

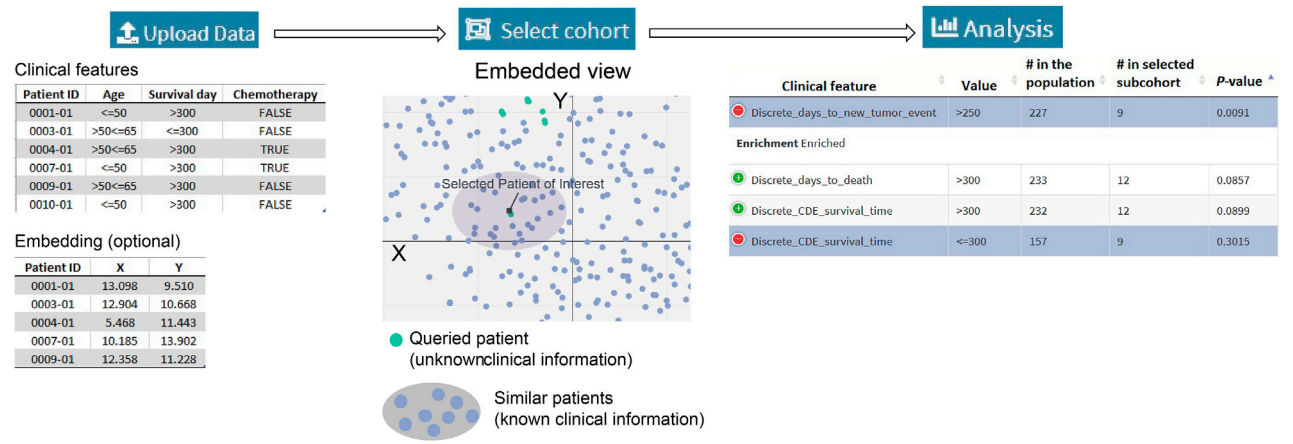


FIGURE 1 | Overview of data processing and analysis.

subcohort, find potential clinical bias, and adjust the filtering criteria to obtain a better subcohort. Integrating Biology and the Bedside (Murphy et al., 2010) is a typical example of a clinical metadata software system. Some systems and techniques may offer more in-depth and specific analysis. For example, Weng et al. (2017) implemented a machine-learning based system to estimate the patients' cardiovascular risk from the routine checkup records. Fang et al. (Fang et al., 2014) implemented a visual analytic system to view patient's geographical demographic and disease comorbidities.

On the other hand, the state-of-the-art clinical data software still has three limitations. First, the simple aggregated analysis has not been well-developed for categorical clinical attributes. Therefore, the researcher may not easily find whether a specific categorical attribute is explicit for the selected cohort compared to the whole population. Second, methods to quantify and visualize patients' similarities have not been implemented. Therefore, the existing clinical software is likely ineffective in clinical support scenarios such as "finding the clinical outcome data about previous patients that are the most similar to the under-treatment patients". Third, the existing software does not support patient clustering. Therefore, they may not automatically recommend subcohort to the researcher. This feature could provide new insights to biomedical research; for example, a tool that quickly shows two clusters in a treatment-selected cohort may enable a new hypothesis about the treatment outcome.

This work introduces Statistical Enrichment Analysis of Samples (https://aimed-lab.shinyapps.io/SEAS/), a software tool with both online and standalone versions to tackle the above limitations. SEAS graphical user interface is user-friendly, where the user interacts by uploading datafile, primarily uses mouse operations, and requires a very limited amount of typing. Furthermore, SEAS implements methods to analyze numerical and categorical data, compute patient similarity, and automatically cluster the patients. For the demo, we use SEAS to analyzing the glioblastoma multiforme (GBM) patients' clinical metadata in The Cancer Genome Atlas Program (TCGA) (Verhaak et al., 2010) and estimate the clinical outcome of patient-derived xenograft (PDX) models data.

SEAS FUNCTIONS

Figure 1A summarizes a SEAS session. The required input is the clinical metadata that is organized in one table. The user may choose to let SEAS automatically compute and represent the patients' similarity in a 2D embedding space or optionally upload another patients' scatterplot. Here, each plot represents a patient, and the distance among the plots should represent patients' similarities. Then, the user may manually enter a subcohort, automatically let SEAS select a subcohort, or semi-automatically choose a subcohort. After selecting a subcohort, SEAS performs clinical feature enrichment analysis (CFEA) and reports all enriched features in the selected subcohort.

Automatically Compute Patients' Similarity and Embedding

In this step, the categorical clinical attributes are digitized as in (Zaki et al., 2014). For example, if the categorical attribute X has three discrete values: low, normal, and high, it can be decomposed into three binary attributes: is_X_low, is_X_normal, is_X_high. If a patient has a "high" categorical value for X, then the patient's digital representation is (0, 0, 1). On the other hand, the numerical attributes are normalized using the z-score approach.

After digitizing the clinical attributes, SEAS applies the embedding method (**Figures 2–7**) to represent the patients in a 2D space. By default, SEAS uses the umap (McInnes et al., 2018) algorithm. Alternatively, the user may also select tSNE (Hinton and Roweis, 2002) for embedding. SEAS computes patients' similarities using the 2D embedded coordinate.

Automatically Select a Subcohort

In SEAS, the user can manually define a subcohort by typing the list of patient IDs (**Figure 2**). Besides, the user may use SEAS to

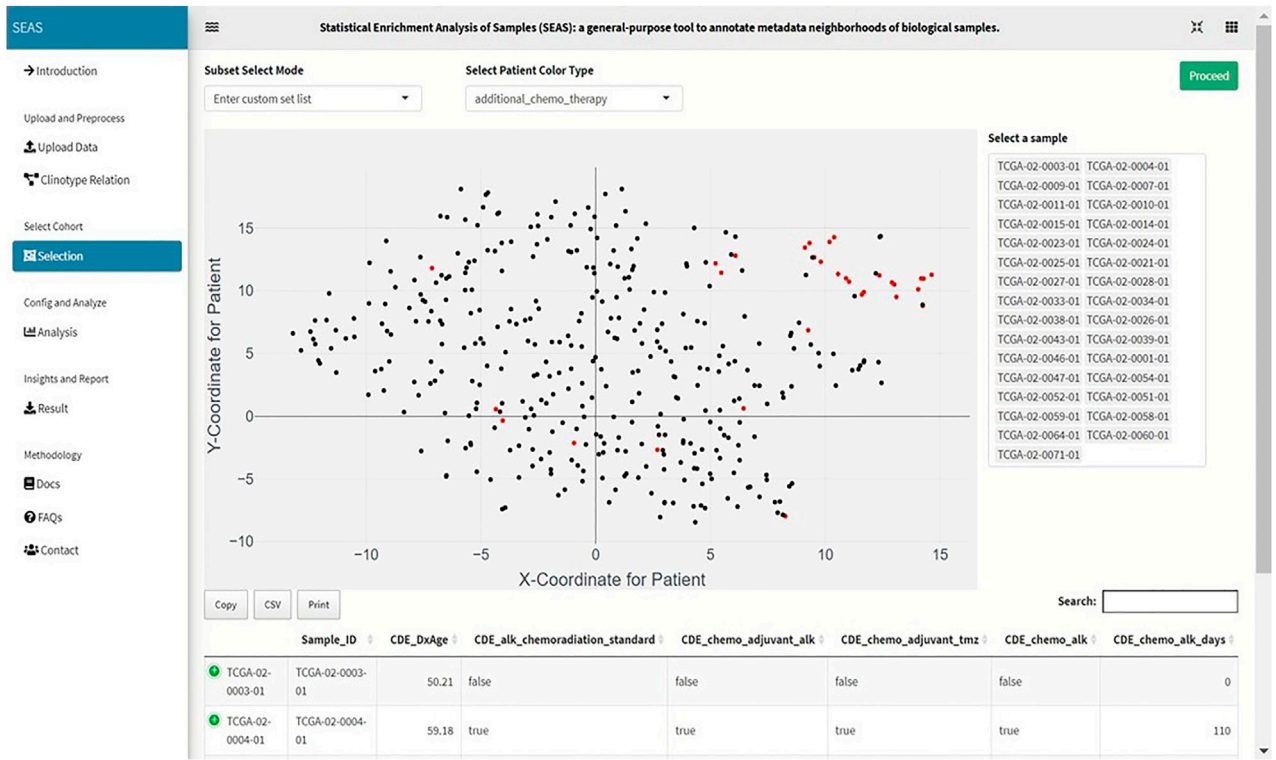


FIGURE 2 | Screenshot showing that SEAS visualizes the TCGA-GBM patients using embedding, and the user manually selects the subcohort.

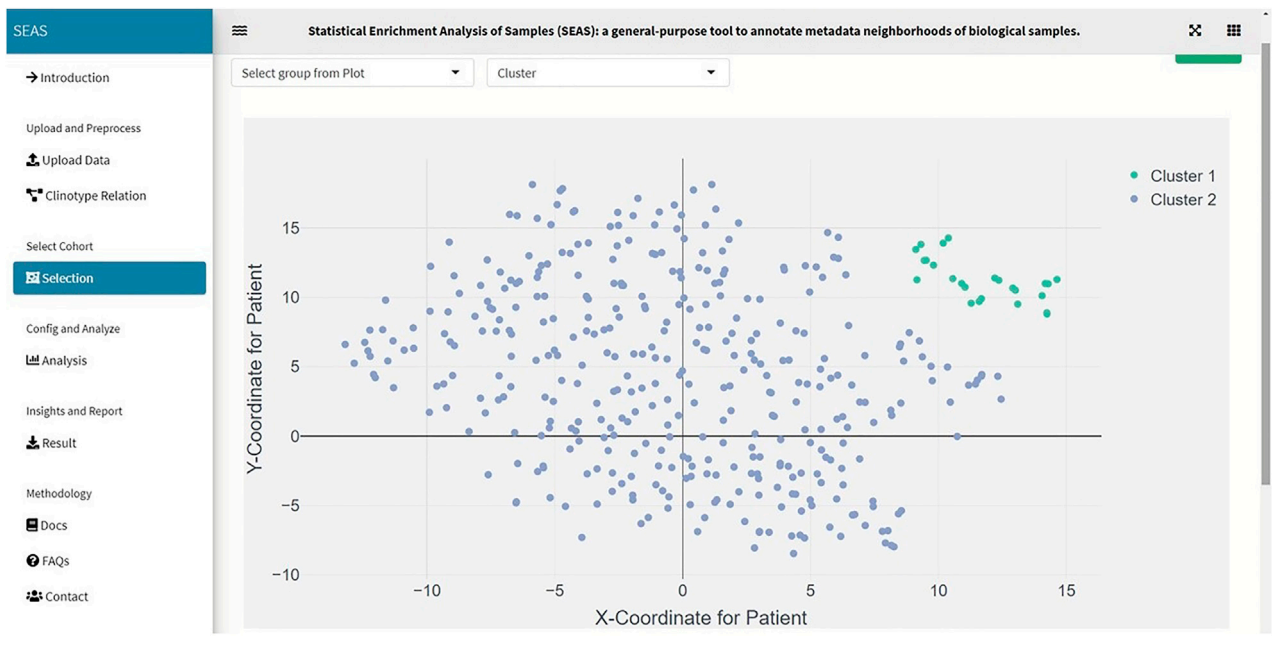


FIGURE 3 | SEAS identifies a subcohort by clustering the TCGA-GBM patients (green dots on the top-right of the embedding scatterplot).

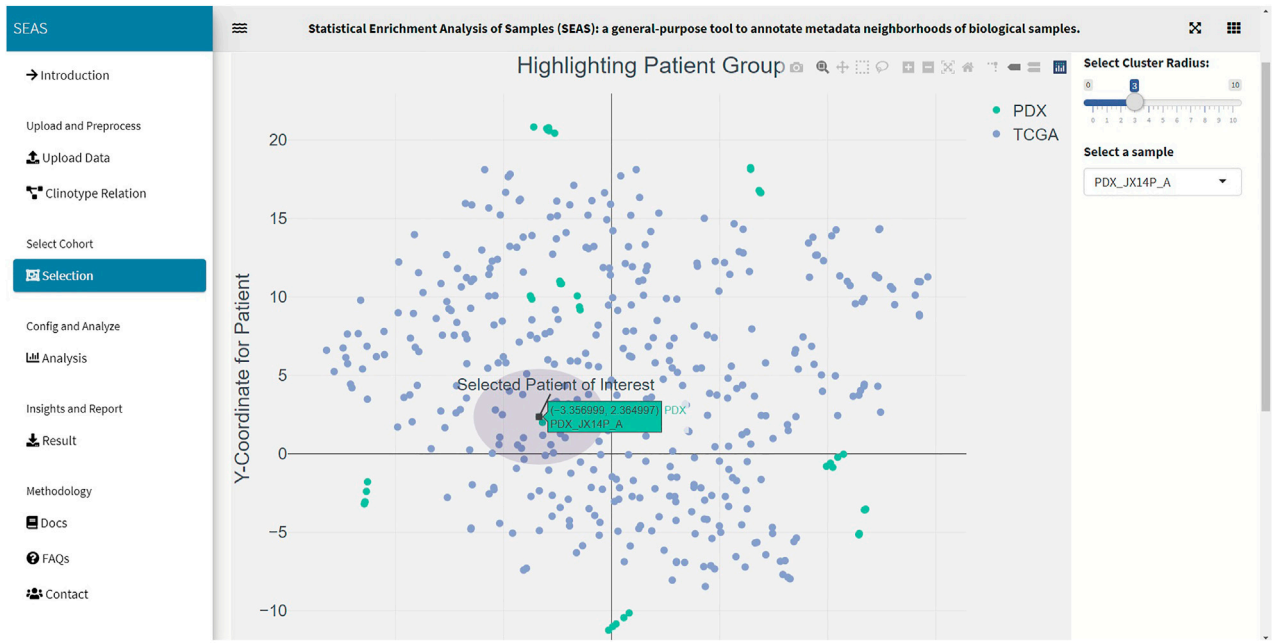


FIGURE 4 | SEAS identifies a subcohort by a circle region around PDX JX14P_A datapoint.

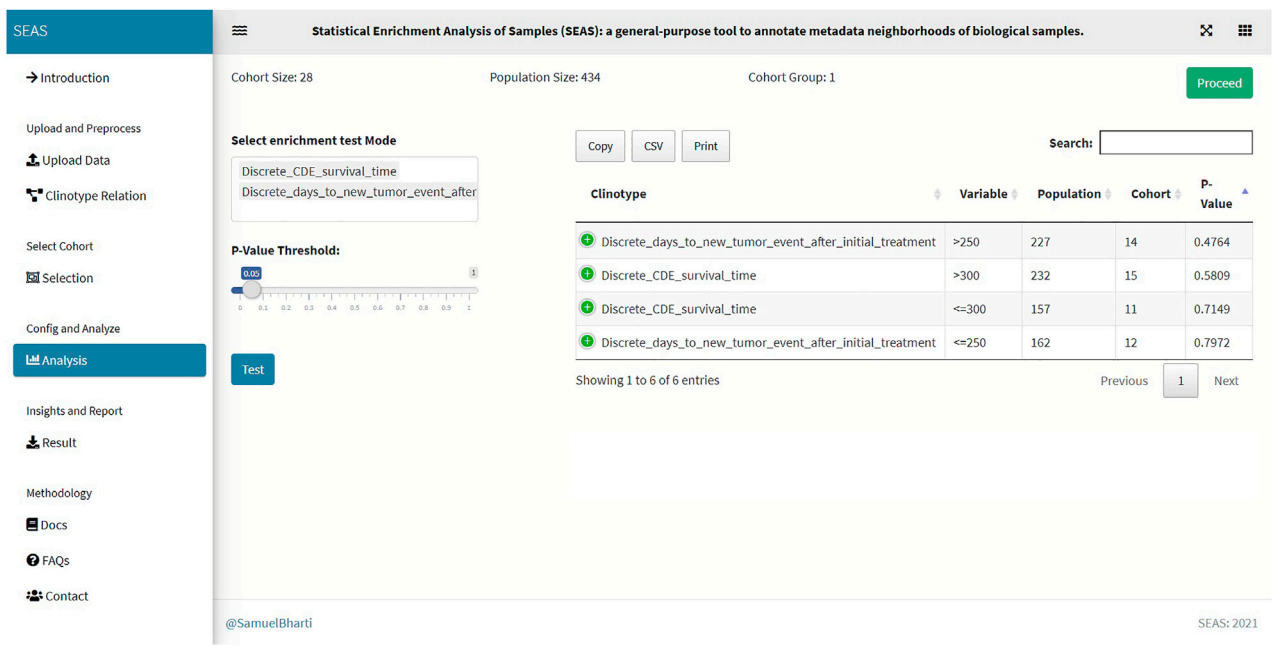


FIGURE 5 | SEAS identifies enriched clinical features for the subcohort in **Figure 4**.

automatically select a subcohort in two ways. In the fully automatic approach, SEAS applies clustering algorithms to divide the patient data into multiple groups. Then, the user selects a group as a subcohort. This approach is preferred because the clustering results can provide the threshold to discretize the numerical attributes into categorical attributes, resulting in the next step. By default, SEAS uses the density-based clustering algorithm (Ester et al., 1996, **Figure 3**). In the semi-automatic approach (**Figures 4**, **6**), the user selects a patient ID, a radius of "similarity area" in the 2D embedding

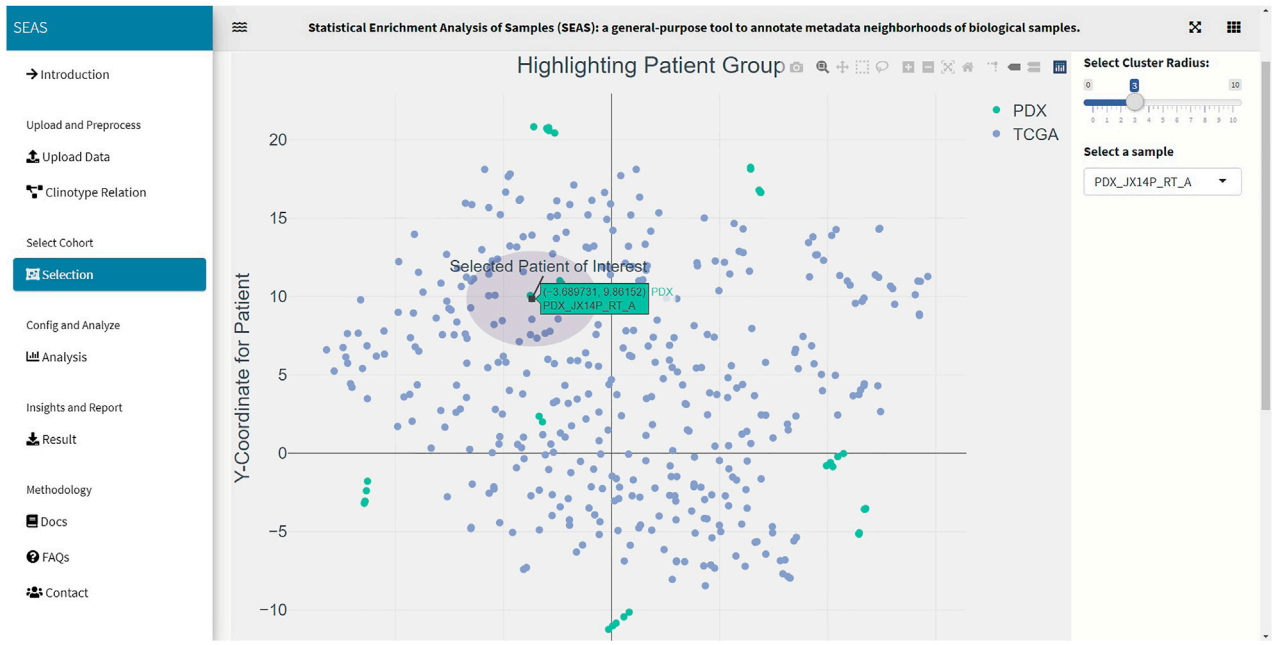


FIGURE 6 | SEAS identifies a subcohort by a circle region around PDX JX14P_RT_A datapoint.

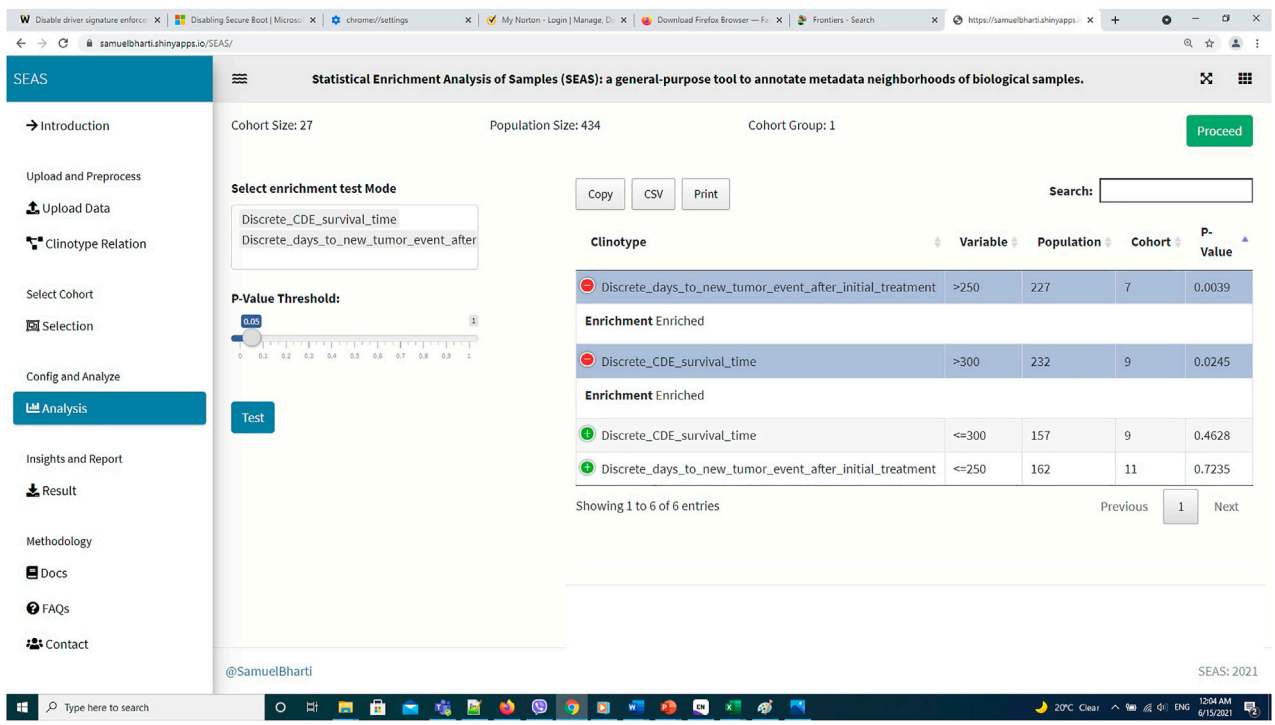


FIGURE 7 | SEAS identifies enriched clinical features for the subcohort in **Figure 6**.

space. All patients in the circle area are centered by the selected patient ID, and the radius becomes the selected subcohort.

Analyze Clinical Feature Enrichment

Besides implementing Wilcoxon-ranksum (Mann and Whitney, 1947) and test between the selected cohort and the whole

population for numerical attributes, SEAS defines the CFEA that can be applied for both numerical and categorical attributes. Here, we denoted a patient population **S** and a set of all clinical attributes **C**. Given any cohort **s** in **S**, the main question is which attributes are representative or enriched in **s**. For a categorical attribute, SEAS applies the hypergeometric test, which compares the proportions of patients having the attribute between **s** and **S**. This approach is well-known in gene set enrichment analysis (Falcon and Gentleman, 2008). Here, the null hypothesis is the proportion of patients having attribute C in **s** and **S** is the same. This is analog to the null hypothesis in the Wilcoxon-ranksum (Mann and Whitney, 1947) test, where the median of attribute **C** in **s** and **S** is the same. To apply in numerical data, the numerical attributes are discretized. For example, in our GBM case study, "CDE_survival_time" (survival day), which is a numerical attribute, is discretized into "Discrete_CDE_survival_time <300 days" and "Discrete_CDE_survival_time ≥300 days." As mentioned in the previous section, clustering the patient and using the cluster to determine the numerical thresholds is a good approach. SEAS reports all enriched clinical attributes and their *p*-values and the Bonferroni adjusted *p*-value (for false discovery rate control) (Sedgwick, 2014), as in **Figure 5**.

Implementing the Software

The SEAS web version is built primarily by bs4Dash (https://cran.r-project.org/web/packages/bs4Dash/index.html) and R-shiny (https://shiny.rstudio.com/) packages. Both packages run based on R and can be hosted inside well-known web programming languages: HTML, CSS, and javascript. In addition, the data processing and statistical methods are also implemented in R.

Demo Using TCGA-GBM Dataset

We acquired and preprocessed TCGA-GBM dataset, which consists of 389 patients, according to the pipeline in Jia et al. (2018). The dataset had both the genetic and the clinical sections. Among 108 clinical attributes, 22 categorical and seven numerical ones were used to compute patient similarity and embedding (**Supplementary Data S1**). Also, we used 45 GBM tumor-samples hosted in patient-derived xenograft (PDX) models (Willey et al., 2020). In these samples, the patients were treated by radiation therapy (RT), but did not have clinical information. Besides the automatic embedding using the clinical data, we manually applied tSNE (Hinton and Roweis, 2002) on the combined TCGA-GBM and PDX genetic data as another 2D representation. We checked the quality of the embedding by the close positions of the PDX JX14P_A/JX14P_B sample pair and the PDX JX14P_RT_A/JX14P_RT_B sample pair. These pairs are replicates of the same patient tumor JX14P (before radiation therapy) and JX14P_RT (after radiation therapy—RT), as shown in **Supplementary Figure S1**.

In this case study, to estimate the clinical outcome of an unknown PDX sample, we select a TCGA-GBM subcohort surrounding the PDX sample (**Figures 4**, **6**) and performed SEAS in the selected TCGA subcohort. In **Figures 4**, **5**, SEAS shows no enriched clinical feature for sample PDX JX14P_A. Here, the average survival time among the surrounding TCGA patients was 339 days. In **Figures 6**, **7**, feature "Discrete_CDE_survival_time >300", which means that the patients who survive for more than 300 days, are enriched among the TCGA samples surrounding the PDX JX14P_RT_A sample. Here, the average survival time for these patients was 434 days. This result suggests radiation therapy may improve the clinical condition of the JX14P patient. Thus, SEAS analysis suggests two opposite clinical outcomes for GBM patients even when being treated by the same therapy. The finding could be helpful in further clinical decisions regarding the selected patients.

Other Notes About Similarity Measures and Embedding Options

Similarity Measures

In SEAS, we used the embedded coordinates to compute the Euclidean distance between two patient datapoints

$$d(i, j) = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \quad (1)$$

Here, i and j denotes two patients, $d(i, j)$ denotes the distance between i and j, (x_i, y_i) denotes the embedded coordinate for patient i, and (x_j, y_j) denotes the embedded coordinate for patient j. We did not use any other similarity measure because we assume that the good embedding results already reflect the patient-wise similarity. In case the user's defined similarity could not be reflected by SEAS, the user can manually enter the list of similar patients to perform the enrichment analysis.

Embedding Options

By default, if the user does not supply the embedding input, SEAS may use umap (McInnes et al., 2018) or tSNE (Hinton and Roweis, 2002) to embed the patient from the clinical features. The embedding algorithms, as in (Konopka, 2020), require a pairwise distance or similarity matrix. At this release, SEAS supports the Euclidean distance (default), cosine similarity, and Jaccard index. Besides, the user is encouraged to supply an embedding file for more in-depth analysis. For example, in our GBM case study, the patient pairwise similarity and embedding are computed by the gene expression data instead of the clinical feature. The PDX have gene expression data but do not have clinical attributes; therefore, they could not be embedded correctly with SEAS default option. When the clinical data is insufficient to compute good embedding results, we highly recommend the user to use other tools to compute the embedding prior to using SEAS.

DISCUSSION AND CONCLUSION

To summarize, we developed the user-friendly and online version of SEAS. The tool can provide new and significant insights into clinical data research and may support the clinical decision. In the future, we expect to develop the add-on version of SEAS, which can be integrated into I2B2 clinical data management system.

One limitation in this SEAS first release is that we have not implemented techniques handling missing values in the patients'

clinical data. To lower the impact of this limitation, we chose the enrichment methods, such as the hypergeometric test, that do not require a very large data size. In our GBM case study, the population consists of 389 patients, which is a moderate size. However, it is sufficient to perform the statistical test even if the missing data rate for one clinical attribute is 10%. On the other hand, we encourage the user to use the non-clinical data to embed the patients; therefore, the missing clinical data may not impact the quality of SEAS results. In fact, our GBM case study shows an approach to infer unknown clinical attributes in PDX data by SEAS analysis of TCGA-GBM data.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/aimed-uab/SEAS.

AUTHOR CONTRIBUTIONS

TN acquired, processed, and analyzed the combined TCGA-GBM/GBM-PDX data, and wrote the manuscript draft. SB implemented the SEAS software. ZY processed the GBM-PDX data. CW prepared and provided the GBM-PDX data. JC conceptualized the SEAS framework and designed the analytical experiment. All authors read, revised, and approved the manuscript.

FUNDING

The work is partly supported by the National Institute of Health Center for Clinical and Translational Science grant award (No. U54TR002731) to the University of Alabama at Birmingham (UAB) where JC is a co-investigator, a research start-up fund provided by the UAB Informatics Institute to JC, and the National Cancer Institute grant award (No. U01CA223976) in which CW is a principal investigator and JC is a co-investigator.

ACKNOWLEDGMENTS

All authors thank the following general technical support that made case studies included for this work possible: Christian Stackhouse for helping CW generate the GBM-PDX model, Jelai Wang for managing the data management and data analysis computing framework, and Christian Stackhouse and Lara Lanov for executing the RNA-seq analysis pipelines.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2021.725276/full#supplementary-material

Supplementary Figure S1 | Locations of datapoint pairs PDX JX14P_A / JX14P_B sample pair (top) and the PDX JX14P_RT_A / JX14P_RT_B (bottom).

REFERENCES

Burgun, A., and Bodenreider, O. (2008). Accessing and Integrating Data and Knowledge for Biomedical Research. *Yearb. Med. Inform.*, 91–101.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. In KDD conference, Munich, Germany, 226–231. Available at: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf.

Falcon, S., and Gentleman, R. (2008). *Hypergeometric Testing Used for Gene Set Enrichment Analysis, Bioconductor Case Studies*. Springer, 207–220. doi:10.1007/978-0-387-77240-0_14

Fang, S., Palakal, M., Xia, Y., Grannis, S. J., and Williams, J. L. (2014). *Health-Terrain: Visualizing Large Scale Health Data*. Indianapolis, IN: INDIANA UNIV INDIANAPOLIS.

Hinton, G., and Roweis, S. T. (2002). *Stochastic Neighbor Embedding*. Citeseer: NIPS, 833–840.

Hume, S., Sarnikar, S., and Noteboom, C. (2020). Enhancing Traceability in Clinical Research Data through a Metadata Framework. *Methods Inf. Med.* 59, 75–85. doi:10.1055/s-0040-1714393

Jia, D., Li, S., Li, D., Xue, H., Yang, D., and Liu, Y. (2018). Mining TCGA Database for Genes of Prognostic Value in Glioblastoma Microenvironment. *Aging* 10, 592–605. doi:10.18632/aging.101415

Kim, H. H., Park, Y. R., Lee, K. H., Song, Y. S., and Kim, J. H. (2019). Clinical MetaData Ontology: a Simple Classification Scheme for Data Elements of Clinical Data Based on Semantics. *BMC Med. Inform. Decis. Mak* 19, 166. doi:10.1186/s12911-019-0877-x

Konopka, T. (2020). *Package 'umap' Version 0.2.7.0*. CRAN. Available at: https://cran.r-project.org/web/packages/umap/index.html.

Mann, H. B., and Whitney, D. R. (1947). On a Test of whether One of Two Random Variables Is Stochastically Larger Than the Other. *Ann. Math. Statist.* 18, 50–60. doi:10.1214/aoms/1177730491

McInnes, L., Healy, J., and Melville, J. (2018), Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426.

Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., et al. (2010). Serving the enterprise and beyond with Informatics for Integrating Biology and the Bedside (I2b2). *J. Am. Med. Inform. Assoc.* 17, 124–130. doi:10.1136/jamia.2009.000893

Nguyen, T., Zhang, T., Fox, G., Zeng, S., Cao, N., Pan, C., et al. (2021). Linking Clinotypes to Phenotypes and Genotypes from Laboratory Test Results in Comprehensive Physical Exams. *BMC Med. Inform. Decis. Mak* 21, 51. doi:10.1186/s12911-021-01387-z

Ohmann, C., and Kuchinke, W. (2009). Future Developments of Medical Informatics from the Viewpoint of Networked Clinical Research. Interoperability and Integration. *Methods Inf. Med.* 48, 45–54.

Sedgwick, P. (2014). Multiple Hypothesis Testing and Bonferroni's Correction. *BMJ* 349, g6284. doi:10.1136/bmj.g6284

Ta, C. N., Dumontier, M., Hripcsak, G., Tatonetti, N. P., and Weng, C. (2018). Columbia Open Health Data, Clinical Concept Prevalence and Co-occurrence from Electronic Health Records. *Sci. Data* 5, 180273. doi:10.1038/sdata.2018.273

Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi:10.1016/j.ccr.2009.12.020

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., and Qureshi, N. (2017). Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data? *PLoS One* 12, e0174944. doi:10.1371/journal.pone.0174944

Willey, C. D., Stackhouse, C. T., Rowland, J. R., Langford, C. P., Anderson, J. C., Ianov, L., et al. (2020). Multi-omic Exploration of Inherent and Acquired Radiation Resistance of Glioblastoma Patient-Derived Xenografts. *Int. J. Radiat. Oncology*Biology*Physics* 108, S40. doi:10.1016/j.ijrobp.2020.07.2148

Zaki, M. J., Meira, W., Jr, and Meira, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Big Data

BRIEF RESEARCH REPORT
published: 27 September 2021
doi: 10.3389/fdata.2021.725095

NPARS—A Novel Approach to Address Accuracy and Reproducibility in Genomic Data Science

*Li Ma[1,2†], Erich A. Peterson[1†], Ik Jae Shin[1], Jason Muesse[1], Katy Marino[1], Matthew A. Steliga[1] and Donald J. Johann Jr[1]**

[1]Winthrop P. Rockefeller Cancer Institute, University of Arkansas for Medical Sciences, Little Rock, AR, United States, [2]Department of Information Science, University of Arkansas at Little Rock, Little Rock, AR, United States

Background: Accuracy and reproducibility are vital in science and presents a significant challenge in the emerging discipline of data science, especially when the data are scientifically complex and massive in size. Further complicating matters, in the field of genomic-based science high-throughput sequencing technologies generate considerable amounts of data that needs to be stored, manipulated, and analyzed using a plethora of software tools. Researchers are rarely able to reproduce published genomic studies.

Results: Presented is a novel approach which facilitates accuracy and reproducibility for large genomic research data sets. All data needed is loaded into a portable local database, which serves as an interface for well-known software frameworks. These include python-based Jupyter Notebooks and the use of RStudio projects and R markdown. All software is encapsulated using Docker containers and managed by Git, simplifying software configuration management.

Conclusion: Accuracy and reproducibility in science is of a paramount importance. For the biomedical sciences, advances in high throughput technologies, molecular biology and quantitative methods are providing unprecedented insights into disease mechanisms. With these insights come the associated challenge of scientific data that is complex and massive in size. This makes collaboration, verification, validation, and reproducibility of findings difficult. To address these challenges the NGS post-pipeline accuracy and reproducibility system (NPARS) was developed. NPARS is a robust software infrastructure and methodology that can encapsulate data, code, and reporting for large genomic studies. This paper demonstrates the successful use of NPARS on large and complex genomic data sets across different computational platforms.

Keywords: genomics, data science, reproducibility, accuracy, analytic validity

OPEN ACCESS

Edited by:
Huixiao Hong, United States Food and Drug Administration, United States

Reviewed by:
Jung Hun Oh, Memorial Sloan Kettering Cancer Center, United States
Sheeba Samuel, Friedrich Schiller University Jena, Germany

****Correspondence:***
Donald J. Johann Jr
djjohann@uams.edu

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *14 June 2021*
Accepted: *07 September 2021*
Published: *27 September 2021*

Citation:
Ma L, Peterson EA, Shin IJ, Muesse J, Marino K, Steliga MA and Johann DJ (2021) NPARS—A Novel Approach to Address Accuracy and Reproducibility in Genomic Data Science. Front. Big Data 4:725095. doi: 10.3389/fdata.2021.725095

INTRODUCTION

The intersection of data science, analytics, and precision medicine are now having an increasingly important role in the formation and delivery of health care, especially in cancer where the treatment regimens are complex and becoming more individualized (Ginsburg and Phillips, 2018). The National Research Council defined precision medicine as the ability to guide health care toward the most effective treatment for a given patient, improving quality and reducing the need for

unnecessary diagnostic testing and therapies (National Research Council, 2011). Our understanding of the genomic basis of disease (cancer) is being transformed by the combination of next generation sequencing (NGS) and state-of-the-art computational data analysis, which are empowering the entry of innovative molecular assays into the clinic, and further enabling precision medicine (Berger and Mardis, 2018). Precision medicine is data science driven (Ginsburg and Phillips, 2018).

Data science is a nascent, cross-disciplinary field that can be viewed as an amalgamation of classic disciplines. These include, but are not limited to: statistics, applied mathematics, and computer science, and importantly is focused on finding non-obvious and useful patterns from large datasets (Kelleher and Tierney, 2018). Data science seeks to find patterns and discriminators in order to support actionable decision making (Cao, 2017a; He and Lin, 2020). How can an insight be actionable? Except for domain-specific factors, the *predictive power* of an insight makes itself actionable (Dhar, 2013). A central tenet in science that distinctly extends into data science is *accuracy,* which is the quality or state of being correct or precise. It is also defined as simply the ratio of correctly predicted observations to the total observations, and is utilized to measure predictive power.

Data science is enabling new and different understandings and reshaping several traditional fields (e.g., microbiology and microbiome, supply chain management, astronomy) into heavily data-driven disciplines (Borne, 2010; Hazen et al., 2014; Bolyen et al., 2019). The term "*Data Science*" is becoming increasingly associated with data sets massive in size, but there are additional challenges in this rapidly evolving field. Some factors considered to contribute to the challenges include: 1) *data complexity*, which refers to complicated data circumstances and characteristics, including the quality of data, largeness of scale, high dimensionality, and extreme imbalance; 2) the development of effective algorithms and, common task infrastructures and learning paradigms needed to handle various aspects of data; 3) the appropriate design of experiments; 4) proper translation mechanisms in order to present and visualize analytical results; 5) *domain complexities*, which refers to expert knowledge, hypotheses, meta-knowledge, etc., in the particular subject matter field (Cao, 2017b).

There is a known reproducibility problem in science. This was investigated and quantified by a survey conducted by the journal Nature involving over 1,500 scientists (Baker, 2016). The survey results reported that over 70% of researchers have tried and failed to reproduce another scientist's results and, more than half have failed to reproduce their own experiments. The survey also uncovered ambiguity concerning the exact definition of reproducibility and, this definition may be different depending on the scientific field.

In data science, *reproducibility* is generally defined as the ability to re-compute data analytic results, with an observed dataset and requisite information regarding the analysis tools (Peng, 2015). Given reproducibility, independent researchers can build up evidence for or in contradiction to a scientific hypothesis (Peng, 2011; Aarts et al., 2015). Some studies have suggested a large number of practical rules or methods for enhancing reproducibility in research (Sandve et al., 2013; Rupprecht et al., 2020). Nonetheless, in several fields, non-reproducibility is still an obstacle towards the better understanding of datasets, further blocking the path to new scientific discoveries (Mobley et al., 2013; Iqbal et al., 2016; Goodman et al., 2018; Wen et al., 2018). In addition, the current situation has forced us to face an awkward truth, that is, while our ability to generate data has grown dramatically, our ability to thoroughly understand data outputs has not developed at the same rate (Peng, 2015). Only if an analytical result is reproducible, can its accuracy be determined. The accuracy itself is based on evaluating the average performance of a series of analytical results from the same dataset. Then can we say such an analytical result is valid and has *analytical validity*. In other words, analytic validity can tell us how well the predictive power of an insight can be. Accuracy and reproducibility are cornerstones of analytical validity.

As more realize the implications and challenges presented by reproducibility in the field of biology, outstanding bioinformatics tools have been developed for improving the situation. To conquer the heterogeneities in bioinformatics tools, Bioconda (Grüning et al., 2018a) integrates more than 3,000 Conda tools. Docker based Dugong (Menegidio et al., 2018) automates the installation of more than 3,500 bioinformatics tools. Pachyderm (Novella et al., 2019) has been developed for managing complicated analyses including multiple stages and multiple tools. For specific studies, reproducible pipelines have been introduced: PiGx (Wurmus et al., 2018) has been created for reproducible genomics analysis, whereas, QIIME 2 (Bolyen et al., 2019) has been released for reproducible, interactive, scalable, and extensible microbiome data science. Finally, many researchers have utilized the web-based platform Galaxy (Jalili et al., 2020) to facilitate collaborative and reproducible (Grüning et al., 2018b) biomedical analyses.

In genomic data science, to address reproducibility, improve scientific accuracy, and enhance collaboration, we present a robust software infrastructure and methodology that can encapsulate data, code, and reporting for large genomic studies. Our system is specifically focused on post-NGS pipeline (downstream) analysis, since it is at this juncture where collaborative endeavors arise focused on gleaning biological insights into studies employing one or more large and complex omics data sets. While the aforementioned tools each offer some methods for tackling the collaborative and reproducibility problems associated with pipeline software, none offer all the features and flexibility in our area of inquiry; post-pipeline (downstream) analysis collaboration and reproducibility. As an example, Galaxy is able to provide collaboration and reproducibility of downstream analyses, however, its ability to execute arbitrary code *via* a programming language of the researcher's choice—if possible—can be quite burdensome.

Our system is named NGS Post-pipeline Accuracy and Reproducibility System (NPARS) and its core technologies are graphically illustrated in **Figure 1**. NPARS is different from other approaches. Specifically, it is the first to focus on the challenges

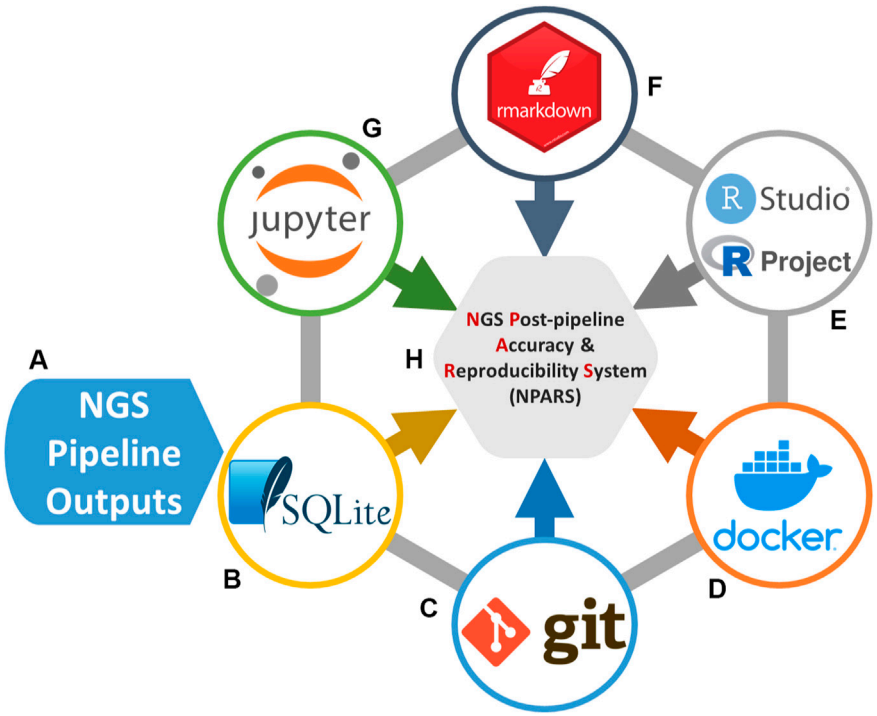


FIGURE 1 | Software technologies used for the NGS Post-pipeline Accuracy and Reproducibility System (NPARS) infrastructure creation. The six core technologies used are shown. **(A)** Study results from a genomics pipeline or repository are extracted and prepared for insertion into a SQLite database. **(B)** SQLite stores all genomic study outputs along with salient study metadata. **(C)** Git provides version control of the Dockerfiles (Docker image specification, i.e., analysis environment) and analysis source code. **(D)** Docker wraps the development environmental information into a container, simplifying software configuration management and, the initialization of a reproducible analysis environment. **(E)** RStudio, provides an integrated development environment for the R programming language and R Projects that are utilized, which provide an efficient way to organize software development activities. **(F)** RMarkdown generates self-documenting analytical reports into HTML files. **(G)** Jupyter Notebooks, are utilized as a development and visualization environment for Python-based projects and reports.

associated with the accuracy, reproducibility, as well as, providing a more convenient manner of collaboration with colleagues. This is achieved by the ability of NPARS to encapsulate large and complex genomic datasets into a portable database container, which may then be analyzed by well-established APIs (Python/Jupyter Notebook, R/Rmd). The infrastructure first loads all data needed for subsequent analyses into a local lightweight (SQLite, 2021) database. The data is then captured within the database along with salient metadata into a schema, which can then be accessed *via* well-known open-source application programming interfaces. These include the use of Jupyter Notebooks (Python) (Kluyver et al., 2016; Python Software Foundation, 2021), RProjects and RMarkdown (R) (Allaire et al., 2021; R-Project, 2021) with an aim to generate self-documenting source code, and results in portable formats. All software may be managed using Docker (Merkel, 2014) containers and Git (Git, 2021) (version control), simplifying configuration management.

METHODS

Synthetic Data

Synthetic data was used in this study. All synthetic data was derived from actual human tumor tissue data sets (e.g., FastQ files). RNA-seq synthetic data was produced by RSEM (Li and Dewey, 2011). DNA-based synthetic data was produced through aggregation and averaging from a pool of human tumor samples. All FastQ files were initially created from BCL files using bcl2fastq2 v2.18.0.12 (bcl2fastq2 and bcl2fastq, 2021) and when needed or indicated, adapter trimming was performed during the conversion. FastQC v0.11.4 (FastQC, 2021) was used to assess the quality of all FastQ files.

RNA Sequencing Pipeline

Transcriptome Reconstruction and Gene-Level Count Qualification

STAR v2.5.3a (Dobin et al., 2013) was used to align each sample's paired-end reads to the Ensembl Homo Sapiens reference genome build GRCh37.75, using STAR's "2-pass" method. Quality control and assessment of resulting BAM files was performed using QualiMap v2.2.1 (García-Alcalde et al., 2012) and STAR output metrics. Picard v2.0.1 (Picard, 2021) was used to add read group information. The marking of duplicate reads and sorting of aligned files was also performed using Sambamba v0.6.5 (Tarasov et al., 2015).

Each sample's BAM file was initially processed using StringTie v1.3.3b (Pertea et al., 2015), along with Ensembl gene annotations to guide transcriptome reconstruction with novel transcript

discovery enabled. Each patient's samples (i.e., study cohort) transcriptome was merged using StringTie's merge mode. Finally, the cohort's BAM files were processed using the newly created merged transcriptome. The StringTie option to output "Ballgown-ready" files was enabled.

Ballgown-ready files containing transcript coverage data was "rolled-up" to the gene-level and the R v4.0.3 (R-Project, 2021) library IsoformSwitchAnalyzeR v1.13.05 (Vitting-Seerup and Sandelin, 2019) was used to disambiguate novel findings from StringTie output. Unnormalized count data was extracted from IsoformSwitchAnalyzeR and used for downstream analysis.

RNA Expressed Mutation Calling and Gene Fusion Detection

RNA variants were called using the Broad Institute's GATK Best Practices for RNA-seq variant calling (Calling Variants in RNAseq, 2021). These steps include the following: STAR was used to align reads to the Ensembl Homo Sapiens reference genome (build GRCh37.75), using the recommended "2-pass" approach. Duplicates were marked and the aligned reads sorted with Sambamba. Next, the tool SplitNCigarReads [GATK v3.9 (McKenna et al., 2010; DePristo et al., 2011)] was used to split reads into exon segments, clip reads which overhang intronic regions, and assign a default MAPQ score of 60 to all reads. Variants were called using the HaplotypeCaller tool (GATK). Gene fusions were detected by passing FastQ files directory to STAR-Fusion v1.4.0 (Haas et al., 2019).

DNA Sequencing Pipeline

Targeted Mutational Panel

FastQ files were submitted to the QIAGEN Data Analysis Center (QIAGEN, 2021) in a tumor/normal configuration and processed using the smCounter2 (Xu et al., 2018) pipeline. The aforementioned pipeline generates aligned reads in BAM format and variants detected in VCF format. Quality control and assessment of resulting BAM files was performed using QualiMap.

Low-Pass Whole Genome Copy Number Variation

Each sample's FastQ paired-end files were aligned to the Ensembl Homo Sapiens reference genome (build GRCh37.75) using BWA v0.7.12 (Li and Durbin, 2009). Quality control and assessment of BAM files was performed with QualiMap. BAM files were post-processed to mark duplicates and sort aligned reads (Sambamba). Copy number data was computational inferred using the R library ichorCNA v0.2.0 (Adalsteinsson et al., 2017).

Post-pipeline Reproducible Data Science Software Infrastructure

NPARS was implemented using the following software packages: Python v2.7.5/3.7.1; Jupyter Notebooks v6.3.0; IPython v7.22.0 (Pérez and Granger, 2007); R v4.1.0; RStudio v1.4.1717 (RStudio, 2020); RMarkdown v.2.7; SQLite v3.35; Docker v20.10.3; and Git v2.26.2 (Git, 2021).

RESULTS

NPARS Overview and Workflow

Figure 2 illustrates an overview and workflow for NPARS. First, the data associated with the study of interest is identified. This may be performed from either a central database/repository or directly from pipeline output files as shown in **subfigure (A)**. Next, custom Python scripts are used to perform extraction and transform operations on the pipeline outputs and associated metadata **(B)**. The result is to produce a set of standardized/structured output files, i.e., well-formatted comma-separated files **(C)**. A Python script **(D)** imports the structured output files into the local SQLite database containing a well-defined schema to hold the data. The SQLite database **(E)**, is a light-weighted and easily portable database, and is utilized to store the study's data and metadata in a well-organized manner. Well known and regarded APIs (RProject and R-Markdown, Jupyter notebooks) are utilized to interface **(F)** to the SQLite database for analysis type activities.

Docker images are utilized to "spin-up" containers, which contain installations of an analysis environment **(G)**. For example, a Docker image containing an R/RStudio environment was created, which includes the necessary libraries (e.g., RMarkdown, DESeq2, etc.) to perform exploratory data analysis (EDA) and differential gene expression on a given study of interest. Python utilizing Jupyter Notebooks is another example analysis environment. Other analysis environments can be easily "Dockerized", or encapsulate the analysis environment within a Docker image in order to offer the desired functionality. NPARS can also be run without Docker.

Docker image specifications are checked into a Git repository in the Dockerfile format, to allow images to be easily shared and to provide version control of the analysis environments and their dependencies. This greatly aids the ultimate goal of NPARS, which is reproducible output **(H)**. Version controlled analysis source code, can interface directly with a SQLite database *via* well-defined, open-source interfaces provided by the software framework of choice. For example, the R library RSQLite (RSQLite, 2021) may be used to directly query the data to be analyzed from the SQLite database. Finally, given the SQLite database along with access to the Git repository containing the Docker specification and source code, any collaborator may generate a reproducible, complete analysis environment, as well as, analysis results from self-documenting RMarkdown or Jupyter Notebooks.

Database Schema

The SQLite database utilized by the NPARS is displayed in **Figure 3** and contains several groups of major tables. The entity relationship model illustrates the metadata and genomics study data within the context of the database schema. The *Study Meta Data* table (**subfigure A**) provides an essential repository of metadata, as well as means of central connection to the other database tables *via* a combination of

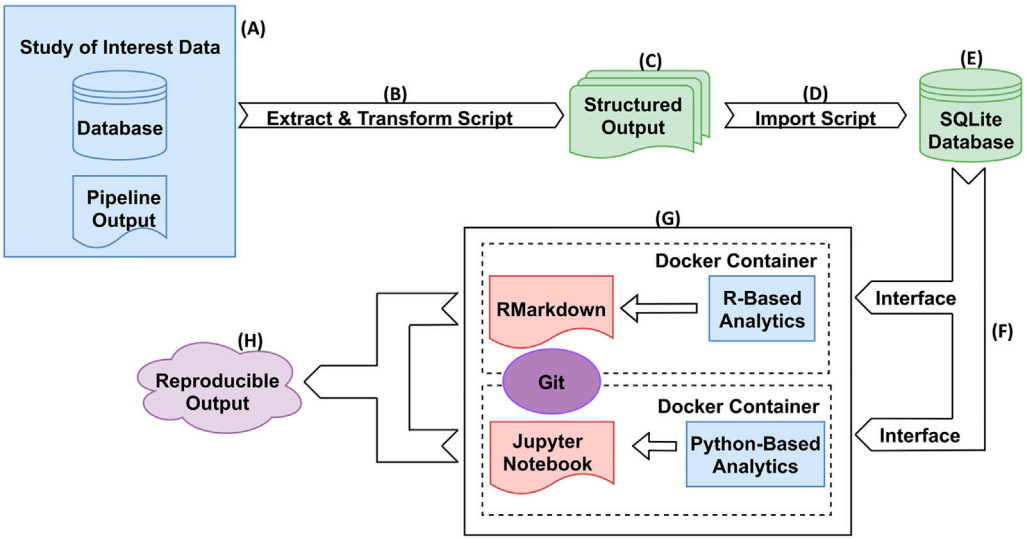


FIGURE 2 | NPARS Overview and Workflow. **(A)** Genomic pipeline output for a particular study of interest is identified. This output can be stored in a database(s) and/or in output files. **(B)** A Python script extracts the identified study results and transforms them into well-defined structured output files. **(C)** The structured output files contain all data and metadata to be imported into the SQLite database. **(D)** A Python script imports the structured output files into the local SQLite database, which already has a well-defined schema to hold the data. **(E)** The SQLite database stores the scientific study data and metadata in a well-organized manner. **(F)** The only interface between the user and the data, is through the particular SQLite API for that development environment. For example, R provides the RSQLite library that provides access to the data. **(G)** Each analysis environment is an abstraction (container) within a Docker container and the source code for it is checked into Git. Self-documenting coding technologies such as R/RMarkdown and Python/Jupyter Notebooks, are used to perform the desired analyses. **(H)** Reproducible reports/analyses are generated, that are both portable and reproducible.

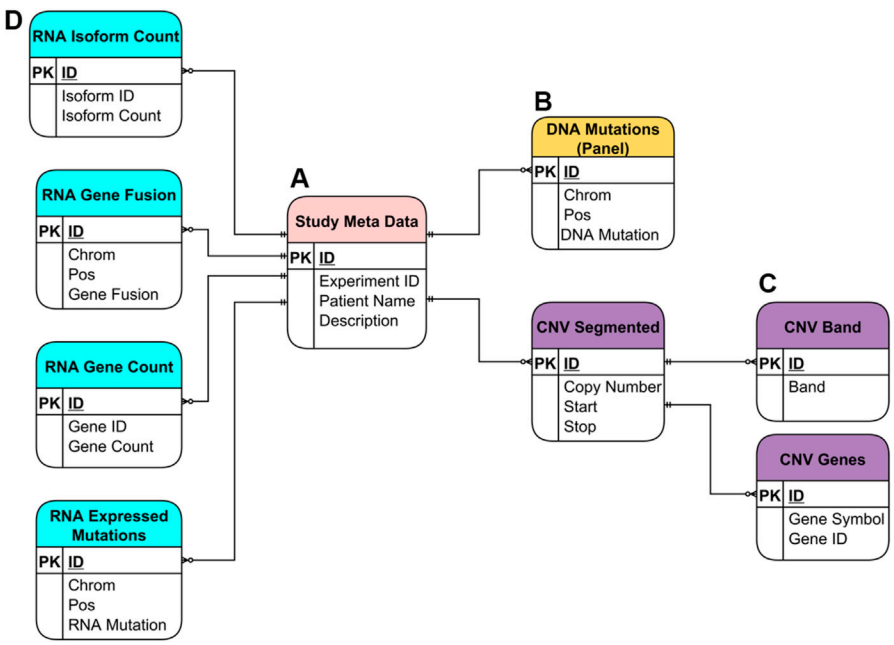


FIGURE 3 | Entity Relationship (ER) Model for the SQLite database utilized in NPARS. Metadata and genomics study data are shown within the context of the database schema. **(A)** Study metadata table ("Study Meta Data"), provides a central repository of metadata, and means of connection to the rest of the tables *via* primary and foreign keys. **(B)** DNA mutations table ["DNA Mutations (Panel)"] contains mutational data from a targeted DNA NGS panel. **(C)** Three tables store copy number variation (CNV) data ("CNV Segmented"), where each CNV segment is a range of chromosome bases of similar copy number value. Each CNV segment is associated with possibly many genes within it ("CNV Genes"), and with possibly many cytobands ("CNV Band"). **(D)** The four tables which hold RNA-based study data: isoform count ("RNA Isoform Count"), gene fusions ("RNA Gene Fusion"), gene count ("RNA Gene Count") and, expressed mutations ("RNA Expressed Mutations").

primary and secondary keys. The *DNA Mutations* table (**B**) contains NGS mutational data from a targeted panel.

The remaining tables house six different types of genomic data results. Tables that contain the copy number variation data derived from DNA using an ultra-low-pass whole genome sequencing approach are shown in purple (**C**). As part of the ultra-low-pass approach, copy number data is segmented into chromosomal regions of similar copy number status (*CNV Segmented*) and, each segment/locus is annotated *via* one-to-many relationships with associated genes, (*CNV Genes*) and, associated cytobands (*CNV Band*). Genomic study results include a variety of RNA-based results, which are shown in light blue (**D**). These include isoform count data (*RNA Isoform Count*), gene fusion data (*RNA Gene Fusion*), gene count data (*RNA Gene Count*) which is essentially "rolled up" isoform count data and, expressed mutations data (*RNA Expressed Mutations*).

Data Analyses

NPARS can generate a wide variety of plots and tables for the purposes of EDA and/or other user-specific analyses, such as finding differentially expressed genes (DEGs). Here we disseminate some examples of reproducible analyses results that were performed on the samples (a total of 21 different NGS experiments yielding large and complex multi-omic datasets), which were described in the Methods section. EDA is an approach for analyzing datasets, summarizing, and showing their main statistical properties in graphics or other data visualization algorithms (Tukey, 1977). **Supplementary Figure S1**, displays a few examples used in NPARS for RNA-seq data. **Subfigure A** shows violin and box plots displaying the distribution of read counts for the replicates of three classes of samples colored blue, green and maroon. In this example each sample class contains three replicates. Next a principal component analysis plot **(B)** of the samples begins to explore the data. The three tissue types used in this study are circled and color coded. The two principal components explain 72% of the variation. **(C)** A hierarchical clustering analysis (HCA) with heatmap of mean normalized counts, showing the top 20 most variable genes on the *y*-axis, and the three tissue types along with their three replicates colored and listed along the *x*-axis. It is known that tissue types T2 and T3 are biologically similar. Tissue type T1 is known to be biologically different from T2 and T3, and this is reflected in the dendrogram.

In addition to traditional EDA plots, the R library RCircos v.1.2.1 (Zhang et al., 2013) was used in NPARS to visualize multiple NGS studies in a single plot (**Supplementary Figure S2**). From the outermost ring inward this figure is composed of: **i.** human chromosomal ideogram, **ii.** DNA panel mutations (tumor vs. germline), **iii**. RNA expressed mutations from the full transcriptome, **iv**. whole genome DNA copy number variations (tumor vs. germline) colored according to the legend symbols that denote amplification, normal, or deletion, **v**. RNA gene expression (TPM) and, **vi**. RNA gene fusions.

Differential gene expression (DGE) analysis takes normalized RNA-based read count data and performs a statistical analysis, to find quantitative changes in expression levels between different experimental groups. A DGE analysis report is generated by NPARS, and an abbreviated example output is shown in **Supplementary Tables S1, 2**. This information was produced as part of a RSQLite query. The novel gene findings report (**Supplementary Table S1**) is discussed. Subtable **A** shows columns for the following: **i.** predicted novel gene (ID), **ii**. locus, **iii**. gene name corresponding to the nearest annotated gene **iv.** log2 fold change (case over control), **v.** *p*-value, and **vi**. adjusted *p*-value. Subtable **B** displays: **i.** predicted novel gene (ID), **ii.** Case sample mean normalized count (*via* replicates), **iii.** Case sample standard deviation (replicates), **iv.** control sample mean normalized count (replicates) and, **v**. control sample standard deviation (replicates).

Supplementary Table S2 illustrates an abbreviated example report for annotated gene findings. Subtable **A** shows columns for the following: **i.** annotated gene (ID), **ii.** gene symbol, **iii**. locus, **iv.** strand information, **v.** log2 fold change (case over control), **vi.** *p*-value and, **vii**. adjusted *p*-value. Subtable **B** shows columns for the following: **i.** annotated gene (ID), **ii.** Case sample mean normalized count (*via* replicates), **iii**. Case sample standard deviation (replicates), **iv.** control sample mean normalized count (replicates) and, **v.** control sample standard deviation (replicates).

An example of an abbreviated copy number variation (CNV) report derived from an ultra-low-pass whole genome (tumor/germline) NGS approach and processed by the ichor package, was generated by NPARS and is displayed in **Supplementary Table S3**. The table is produced as part of a RSQLite query and shows columns for the following: **i.** gene symbol, **ii.** annotated gene (ID) per Ensembl, **iii**. Chromosome number, **iv.** Chromosomal segment start position, **v.** chromosomal segment end position, **vi.** median logR, where logR = log2 (T1/Germline), **vii.** subclone status, meaning is the amplication or deletion event part of a subclone per the ichor package **viii.** copy number, **ix.** copy number type and, **x.** cytoband. This report shows a small example of salient CNV findings from a small selection of genes.

A Python/Jupyter Notebook utilizing a library from scikit-learn (Pedregosa et al., 2011) was used to generate the *clustergram* plot in **Supplementary Figure S3** by NPARS. This approach is used as part of finding the optimal number of clusters for a K-Means analysis. RNA-seq data normalized across three sample types using DESeq2 were used in this example. The *x*-axis displays the number of clusters (k) during an iteration of k-means clustering analysis, and the *y*-axis displays the PCA weighted mean of the clusters. Each point (red dot) represents the center of a cluster and, the size of each point represents the amount of information contained in each cluster. The thickness of lines (blue) connecting points represent observations potentially moving between clusters. In this example per the clustergram plot the optimal number of clusters should be 2 or 3.

To further investigate the optimal number of clusters for K-Means, *silhouette coefficient plots* (Zhou and Gao, 2014) were performed using the Python/Jupyter Notebook code employing scikit-learn and shown in **Supplementary Figure S4**. Shown are a series of silhouette plots, which graphically evaluate a variety k-means cluster configurations (2 through 7) along with corresponding silhouette coefficients and threshold value (dotted red vertical line). The value of a silhouette coefficient (*x*-axis) ranges from -1 to 1, the higher the value indicates greater cohesion within the cluster and greater

separation between clusters. A negative value indicates a possible improper cluster assignment and, a zero value indicates the object assignment is between clusters. The higher the coefficient value, the more separated and clearly identifiable is the particular cluster. The thickness of each cluster silhouette (*y*-axis, associated with the cluster label) indicates the cluster size. **(A)** Silhouette analysis for k-means clustering on sample data with 2 clusters. **(B)** Silhouette analysis for k-means clustering on sample data with 3 clusters. In this case the new cluster (cluster label 2) has a zero coefficient value meaning it is not significant. **(C)** Silhouette analysis for k-means clustering on sample data with 4 clusters. This plot shows cluster labels 2 and 3 are not significant. **(D)** Silhouette analysis for k-means clustering on sample data with 5 clusters. **(E)** Silhouette analysis for k-means clustering on sample data with 6 clusters. **(F)** Silhouette analysis for k-means clustering on sample data with 7 clusters. According to the plots, the optimal cluster number should be 2. A confluence of evidence based on this evaluation and the previous (clustergram) is indicating the optimal k-means cluster value may be 2. Datasets used to generate the plots are the same simulated data which were used to generate the clustergram plot (**Supplementary Figure S3**). A Jupyter/Python Notebook was used to perform this analysis.

Based on the prior results from the *clustergram* and *silhouette coefficient plots,* k-means was run twice, once with two clusters, and then three clusters. **Supplementary Figure S5** contains results obtained from the Python/Jupyter Notebook code for this analysis, with k-means and two clusters (**A**), and three clusters (**B**). The same RNA-seq data processed by DESeq2 was used. The plot shapes indicate the cluster membership labels: 0, 1, 2 and, the colors represent the tissue types, T1 (Tissue 1, blue), T2 (Tissue 2, orange), T3 (Tissue 3, green). A small red circle is used to highlight the primary difference between the two plots, namely, a new cluster is formed from T1. Analyzing plots **A** and **B**, it appears that two clusters may more efficiently group the data versus three clusters and, supports the results of the *silhouette plots* (**Supplementary Figure S4**) and, is also in agreement with the *clustergram* plot (**Supplementary Figure S3**).

DISCUSSION

The next evolution in oncology research and cancer care are being driven by data science (Yu and Kibbe, 2021). So, it is of paramount importance to address current accuracy and reproducibility issues. In the field of genomic data science, accuracy and reproducibility remains a considerable challenge due to the sheer size, complexity, and dynamic nature plus relative inventiveness of the quantitative biology approaches. The accuracy and reproducibility challenge does not just block the path to new scientific discoveries, more importantly, it may lead to a scenario where critical findings used for medical decision making are found to be incorrect (Huang and Gottardo, 2013). NPARS has been developed to meet the unmet need of improving accuracy and reproducibility in genomic data science. Currently, a limitation of our system is the requirement of the user to put their data into a standardized format for import into NPARS. These steps are not automated.

An accuracy and reproducibility test of NPARS was performed by running the R/RMarkdown and Python Jupyter Notebook code with the SQLite database on two different systems, 1) Windows 10-based system and, 2) system utilizing the Ubuntu Linux distribution. The results demonstrated the use of NPARS on two different systems produced identical outputs and this is summarized in **Table 1**. Here, the term "Passed" means the observed and expected outputs were identical on the respective systems. The R/RMarkdown outputs were first compared. The RCircos graphic (**Supplementary Figure S2**), which summarizes and integrates seven genomics studies into a single graphical plot was visually inspected from the Windows and Linux systems and found to be identical. **Supplementary Tables S1A,B, 2A,B, 3** were also identical. All EDA graphics from **Supplementary Figure S2** were compared by visual inspection and found to be identical. For the analyses performed by Python/Jupyter Notebook, the *clustergram* (**Supplementary Figure S3**), *silhouette coefficient* plots (**Supplementary Figure S4**) and k-means graphics (**Supplementary Figure S5**) were regenerated on each system, compared by close visual inspection and found to be identical.

TABLE 1 | NPARS Accuracy and Reproducibility Testing Summary.

Analysis test	System #1, Windows 10	System #2, Linux/Ubuntu	Comparative results (system #1 vs. System #2)
RCircos, **Supplementary Figure S2**	Passed	Passed	Identical
DESeq2 Novel Genes, **Supplementary Table S1**	Passed	Passed	Identical
DeSeq2 Annotated Genes, **Supplementary Table S2**	Passed	Passed	Identical
Copy Number Analysis, **Supplementary Table S3**	Passed	Passed	Identical
Violin Plots, **Supplementary Figure S1**	Passed	Passed	Identical
Box Plots, **Supplementary Figure S1**	Passed	Passed	Identical
PCA Plot, **Supplementary Figure S1**	Passed	Passed	Identical
HCA Plot, **Supplementary Figure S1**	Passed	Passed	Identical
Clustergram, **Supplementary Figure S3**	Passed	Passed	Identical
Silhouette Coefficient Plots, **Supplementary Figure S4**	Passed	Passed	Identical
K-means Plots, **Supplementary Figure S5**	Passed	Passed	Identical

The first column, "Analysis Test" lists the name of each test along with corresponding supplemental figure or table information. The columns "System #1, Windows-10" and "System #2, Linux/Ubuntu" lists the results of each test run on these respective systems. The column titled "Comparative Results (System #1 vs. System #2) reports the comparative results outcome. The term "Passed" means the observed and expected outputs were the same on the respective systems.

The innovative and evolving landscape of oncology research and cancer care are dependent on accurate, reproducible, and robust data science. High-throughput instrumentation are generating increasingly massive and complex genomic data sets, and continue to create opportunities and challenges in the dynamic field of genomic data science. This makes collaboration, verification, validation, and reproducibility of findings difficult. To address these challenges NPARS was developed. NPARS is the first system to focus on NGS downstream analysis accuracy, reproducibility, and enhancing collaboration, by effectively capturing large and complex genomic datasets into a portable database container and exposing it to well-established APIs. In this paper we have profiled and demonstrated NPARS, which is a robust software infrastructure and methodology that can encapsulate both data, code, and reporting for large genomic studies. This study demonstrates the successful use of NPARS on large and complex genomic data sets across different computational platforms and begins to address the prevailing challenges of accuracy and reproducibility in genomic data science.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://gitlab.com/erichpeterson/npars-analysis.

AUTHOR CONTRIBUTIONS

DJ conceived the project. DJ, LM, EP devised the experiments. LM and EP performed the software implementation. IS performed laboratory experiments. MS, JM, KM coordinated laboratory experiments. DJ, EP, LM wrote the manuscript. All authors read and approved the manuscript.

FUNDING

The authors would like to acknowledge the financial support of the United States Department of Health and Human Services, Food and Drug Administration, contract HHSF223201610111C through the Arkansas Research Alliance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2021.725095/full#supplementary-material

REFERENCES

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., et al. (2015). Estimating the Reproducibility of Psychological Science. *Science* 349 (6251):aac4716. doi:10.1126/science.aac4716

Adalsteinsson, V. A., Ha, G., Freeman, S. S., Choudhury, A. D., Stover, D. G., Parsons, H. A., et al. (2017). Scalable Whole-Exome Sequencing of Cell-free DNA Reveals High Concordance with Metastatic Tumors. *Nat. Commun.* 8 (1), 1324. doi:10.1038/s41467-017-00965-y

Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., et al. (2021). *Rmarkdown: Dynamic Documents for R.*

Baker, M. (2016). 1,500 Scientists Lift the Lid on Reproducibility. *Nature* 533, 452–454. doi:10.1038/533452a

bcl2fastq2 and bcl2fastq (2021). bcl2fastq2 and Bcl2fastq Conversion Software Downloads. Available at: https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software/downloads.html

Berger, M. F., and Mardis, E. R. (2018). The Emerging Clinical Relevance of Genomics in Cancer Medicine. *Nat. Rev. Clin. Oncol.* 15 (6), 353–365. doi:10.1038/s41571-018-0002-6

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi:10.1038/s41587-019-0209-9

Borne, K. D. (2010). Astroinformatics: Data-Oriented Astronomy Research and Education. *Earth Sci. Inform.* 3, 5–17. doi:10.1007/s12145-010-0055-2

Calling Variants in RNAseq (2021). Calling Variants in RNAseq: Methods and Workflows. Available at: https://www.broadinstitute.org/gatk/guide/article?id=3891

Cao, L. (2017). Data Science : A Comprehensive Overview. *ACM Comput. Surv.* 50 (3). 1–42. doi:10.1145/3076253

Cao, L. (2017). Data Science. *Commun. ACM* 60, 59–68. doi:10.1145/3015456

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data. *Nat. Genet.* 43 (5), 491–498. doi:10.1038/ng.806

Dhar, V. N. Y. U. (2013). Data Science and Prediction. *Commun. ACM* 56 (12): 64–73. doi:10.1145/2500499

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinformatics* 29 (1), 15–21. doi:10.1093/bioinformatics/bts635

FastQC (2021). A Quality Control Tool for High Throughput Sequence Data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., et al. (2012). Qualimap: Evaluating Next-Generation Sequencing Alignment Data. *Bioinformatics* 28 (20), 2678–2679. doi:10.1093/bioinformatics/bts503

Ginsburg, G. S., and Phillips, K. A. (2018). Precision Medicine: From Science to Value. *Health Aff.* 37 (5), 694–701. doi:10.1377/hlthaff.2017.1624

Git (2021). Git. Available at: https://git-scm.com/

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. (2018). What Does Research Reproducibility Mean? *Sci. Transl Med.* 8, 341ps12–102. doi:10.1126/scitranslmed.aaf5027

Grüning, B., Chilton, J., Köster, J., Dale, R., Soranzo, N., van den Beek, M., et al. (2018). Practical Computational Reproducibility in the Life Sciences. *Cel Syst.* 6 (6), 631–635. doi:10.1016/j.cels.2018.03.014

Grüning, B., Dale, R., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., et al. (2018). Bioconda: Sustainable and Comprehensive Software Distribution for the Life Sciences. *Nat. Methods* 15 (7), 475–476. doi:10.1038/s41592-018-0046-7

Haas, B. J., Dobin, A., Li, B., Stransky, N., Pochet, N., and Regev, A. (2019). Accuracy Assessment of Fusion Transcript Detection *via* Read-Mapping and De Novo Fusion Transcript Assembly-Based Methods. *Genome Biol.* 20 (1), 213. doi:10.1186/s13059-019-1842-9

Hazen, B. T., Boone, C. A., Ezell, J. D., and Jones-Farmer, L. A. (2014). Data Quality for Data Science, Predictive Analytics, and Big Data in Supply Chain Management: An Introduction to the Problem and Suggestions for Research and Applications. *Int. J. Prod. Econ.* 154, 72–80. doi:10.1016/j.ijpe.2014.04.018

He, X., and Lin, X. (2020). Challenges and Opportunities in Statistics and Data Science: Ten Research Areas. *Harv. Data Sci. Rev.* doi:10.1162/99608f92.95388fcb

Huang, Y., and Gottardo, R. (2013). Comparability and Reproducibility of Biomedical Data. *Brief. Bioinform.* 14, 391–401. doi:10.1093/bib/bbs078

Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., and Ioannidis, J. P. (2016). Reproducible Research Practices and Transparency across the Biomedical Literature. *Plos Biol.* 14, e1002333–13. doi:10.1371/journal.pbio.1002333

Jalili, V., Afgan, E., Gu, Q., Clements, D., Blankenberg, D., Goecks, J., et al. (2020). The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2020 Update. *Nucleic Acids Res.* 48 (W1), W395–W402. doi:10.1093/nar/gkaa434

Kelleher, J. D., and Tierney, B. (2018). *Data Science*. Cambridge, MIT Press.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., et al. (2016). Jupyter Notebooks—A Publishing Format for Reproducible Computational Workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016. Göttingen, Germany: Electronic Publishing, 87–90.

Li, B., and Dewey, C. N. (2011). RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome. *BMC bioinformatics* 12 (1), 1–16. doi:10.1186/1471-2105-12-323

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data. *Genome Res.* 20 (9), 1297–1303. doi:10.1101/gr.107524.110

Menegidio, F. B., Jabes, D. L., Costa de Oliveira, R., and Nunes, L. R. (2018). Dugong: a Docker Image, Based on Ubuntu Linux, Focused on Reproducibility and Replicability for Bioinformatics Analyses. *Bioinformatics* 34 (3), 514–515. doi:10.1093/bioinformatics/btx554

Merkel, D. (2014). Docker : Lightweight Linux Containers for Consistent Development and Deployment Docker: a Little Background under the Hood. *Linux J.* 2014, 2–7.

Mobley, A., Linder, S. K., Braeuer, R., Ellis, L. M., and Zwelling, L. (2013). A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic. *PLoS ONE* 8, e63221–6. doi:10.1371/journal.pone.0063221

National Research Council (2011). *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease.* Washington (DC), National Academies Press (US).

Novella, J. A., Emami Khoonsari, P., Herman, S., Whitenack, D., Capuccini, M., Burman, J., et al. (2019). Container-based Bioinformatics with Pachyderm. *Bioinformatics* 35 (5), 839–846. doi:10.1093/bioinformatics/bty699

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *J. machine Learn. Res.* 12, 2825–2830.

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science* 334, 1226–1227. doi:10.1126/science.1213847

Peng, R. (2015). The Reproducibility Crisis in Science: A Statistical Counterattack. *Significance* 12, 30–32. doi:10.1111/j.1740-9713.2015.00827.x

Pérez, F., and Granger, B. E. (2007). IPython: a System for Interactive Scientific Computing. *Comput. Sci. Eng.* 9 (3), 21–29. doi:10.1109/mcse.2007.53

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads. *Nat. Biotechnol.* 33 (3), 290–295. doi:10.1038/nbt.3122

Picard (2021). Picard. Available at: http://broadinstitute.github.io/picard/

Python Software Foundation (2021). Python Software Foundation. Available at: http://www.python.org

QIAGEN (2021). *QIAGEN Data Analysis Center*.

R-Project (2021). R: A Language and Environment for Statistical Computing. Available at: http://www.r-project.org/.

RSQLite (2021). SQLite' Interface for R. Available at: https://cran.r-project.org/web/packages/RSQLite/index.html.

RStudio (2020). "Integrated Development for R," in *RStudio, PBC.* (Boston, MA: RStudio Team).

Rupprecht, L., Davis, J. C., Arnold, C., Gur, Y., and Bhagwat, D. (2020). Improving Reproducibility of Data Science Pipelines through Transparent Provenance Capture. *Proc. VLDB Endow.* 13, 3354–3368. doi:10.14778/3415478.3415556

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. *Plos Comput. Biol.* 9, e1003285–4. doi:10.1371/journal.pcbi.1003285

SQLite (2021). SQLite. Available at: https://www.sqlite.org/index.html

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., and Prins, P. (2015). Sambamba: Fast Processing of NGS Alignment Formats. *Bioinformatics* 31 (12), 2032–2034. doi:10.1093/bioinformatics/btv098

Tukey, J. W. (1977). *Exploratory Data Analysis, Vol. 2*. Reading: Mass.

Vitting-Seerup, K., and Sandelin, A. (2019). IsoformSwitchAnalyzeR: Analysis of Changes in Genome-wide Patterns of Alternative Splicing and its Functional Consequences. *Bioinformatics* 35 (21), 4469–4471. doi:10.1093/bioinformatics/btz247

Wen, H., Wang, H.-Y., He, X., and Wu, C.-I. (2018). On the Low Reproducibility of Cancer Studies. *Natl. Sci. Rev.* 5, 619–624. doi:10.1093/nsr/nwy021

Wurmus, R., Uyar, B., Osberg, B., Franke, V., Gosdschan, A., Wreczycka, K., et al. (2018). PiGx: Reproducible Genomics Analysis Pipelines with GNU Guix. *Gigascience* 7 (12). doi:10.1093/gigascience/giy123

Xu, C., Gu, X., Padmanabhan, R., Wu, Z., Peng, Q., DiCarlo, J., et al. (2018). smCounter2: an Accurate Low-Frequency Variant Caller for Targeted Sequencing Data with Unique Molecular Identifiers. *Bioinformatics.* 35(8): 1299–1309. doi:10.1093/bioinformatics/bty790

Yu, P., and Kibbe, W. (2021). Cancer Data Science and Computational Medicine. *JCO Clin. Cancer Inform.* 5, 487–489. doi:10.1200/cci.21.00006

Zhang, H., Meltzer, P., and Davis, S. (2013). RCircos: an R Package for Circos 2D Track Plots. *BMC Bioinformatics* 14, 244. doi:10.1186/1471-2105-14-244

Zhou, H. B., and Gao, J. T. (2014). Automatic Method for Determining Cluster Number Based on Silhouette Coefficient. *Adv. Mater. Res.* 951, 227–230.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Big Data

ORIGINAL RESEARCH
published: 03 November 2021
doi: 10.3389/fdata.2021.719737

Systematic Exploration in Tissue-Pathway Associations of Complex Traits Using Comprehensive eQTLs Catalog

*Boqi Wang[1], James Yang[2], Steven Qiu[3], Yongsheng Bai[4] and Zhaohui S. Qin[5]**

[1]Emory University, Atlanta, GA, United States, [2]Carmel High School, Carmel, IN, United States, [3]James Martin High School, Arlington, TX, United States, [4]Next-Gen Intelligent Science Training, Ann Arbor, MI, United States, [5]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, United States

The collection of expression quantitative trait loci (eQTLs) is an important resource to study complex traits through understanding where and how transcriptional regulations are controlled by genetic variations in the non-coding regions of the genome. Previous studies have focused on associating eQTLs with traits to identify the roles of trait-related eQTLs and their corresponding target genes involved in trait determination. Since most genes function as a part of pathways in a systematic manner, it is crucial to explore the pathways' involvements in complex traits to test potentially novel hypotheses and to reveal underlying mechanisms of disease pathogenesis. In this study, we expanded and applied loci2path software to perform large-scale eQTLs enrichment [i.e., eQTLs' target genes (eGenes) enrichment] analysis at pathway level to identify the tissue-specific enriched pathways within trait-related genomic intervals. By utilizing 13,791,909 eQTLs cataloged in the Genotype-Tissue Expression (GTEx) V8 data for 49 tissue types, 2,893 pathway sets reported from MSigDB, and query regions derived from the Phenotype-Genotype Integrator (PheGenI) catalog, we identified intriguing biological pathways that are likely to be involved in ten traits [Alzheimer's disease (AD), body mass index, Parkinson's disease (PD), schizophrenia, amyotrophic lateral sclerosis, non-small cell lung cancer (NSCLC), stroke, blood pressure, autism spectrum disorder, and myocardial infarction]. Furthermore, we extracted the most significant pathways for AD, such as BioCarta D4-GDI pathway and WikiPathways sulfation biotransformation reaction and viral acute myocarditis pathways, to study specific genes within pathways. Our data presented new hypotheses in AD pathogenesis supported by previous studies, like the increased level of caspase-3 in the amygdala that cleaves GDP dissociation inhibitor and binds to beta-amyloid, leading to increased apoptosis and neuronal loss. Our findings also revealed potential pathogenesis mechanisms for PD, schizophrenia, NSCLC, blood pressure, autism

OPEN ACCESS

Edited by:
Prashanti Manda, University of North Carolina at Greensboro, United States

Reviewed by:
Kui Zhang, Michigan Technological University, United States
Tianhua Niu, Tulane University, United States

***Correspondence:**
Zhaohui S. Qin
zhaohui.qin@emory.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *03 June 2021*
Accepted: *13 October 2021*
Published: *03 November 2021*

Citation:
Wang B, Yang J, Qiu S, Bai Y and Qin ZS (2021) Systematic Exploration in Tissue-Pathway Associations of Complex Traits Using Comprehensive eQTLs Catalog. Front. Big Data 4:719737. doi: 10.3389/fdata.2021.719737

Abbreviations: AD, Alzheimer's disease; ARHGDIB, Rho GDP dissociation inhibitor beta; Aß, amyloid-beta; DCM, dilated cardiomyopathy; eGene, target gene of an expression quantitative trait locus; eQTL, expression quantitative trait locus; GCase, glucocerebrosidase; HLA-C, major histocompatibility complex, Class I, C; HLA-DQB1, major histocompatibility complex, Class II, DQ Beta 1; HLA-DRB1, major histocompatibility complex, Class II, DR Beta 1; ICM, ischemic cardiomyopathy; IL-17, interleukin-17; IL-37, interleukin-37; ILC, innate lymphoid cell; KMO, kynurenine 3-monooxygenase; NSCLC, non-small cell lung cancer; PD, Parkinson's disease; PP2A, protein phosphatase 2A; PPP2R5A, protein phosphatase 2 regulatory subunit B'alpha; ROS, reactive oxygen species.

spectrum disorder, and myocardial infarction, which were consistent with past studies. Our results indicated that loci2path's eQTLs enrichment test was valuable in unveiling novel biological mechanisms of complex traits. The discovered mechanisms of disease pathogenesis and traits require further in-depth analysis and experimental validation.

Keywords: eQTLs, gene pathway sets, gene set enrichment analyses, tissue-pathway association, complex traits

1 INTRODUCTION

Expression quantitative trait loci (eQTLs) have been one of the major focuses in determining the genetic variants that affect gene expressions locating in non-coding regions of the genome. eQTLs' nature of influencing expression levels of their target genes (eGenes) makes them powerful at studying transcription regulation (Li et al., 2010). The traditional usage of genomic physical proximity to connect genetic loci with their corresponding eGenes has been proven somewhat ineffective since it has been demonstrated that only about 25% of eQTLs have their physically closest genes to be their eGenes (Zhu et al., 2016; Xu et al., 2020). Further, eQTLs have become an increasingly popular tool for researchers to identify specific genes for diseases and traits.

Researchers often use eQTLs associations to link expression traits to genotypes of genetic variants located in genomic intervals. Multiple studies have been conducted on connecting eQTLs and various traits including Alzheimer's disease (AD) to determine the roles trait-related eQTLs and their corresponding eGenes play in pathogenesis (Hormozdiari et al., 2016; Zhao et al., 2019; Sieberts et al., 2020). Though many interesting findings have been discussed, the observed eQTLs patterns in cerebral and cerebellar brain regions require further investigations with respect to their potential functions, but so far, to our knowledge, no systematic in-depth studies have been performed to explore the roles of such eQTLs in etiologies of neurodegenerative diseases such as AD (Zhao et al., 2019; Sieberts et al., 2020). Another common practice is to use eQTLs mapping to link an expression trait to genetic variants in certain genomic regions, which holds promise in elucidating gene regulations and predicting gene networks associated with complex phenotypes (Li et al., 2010). By using eQTLs mapping methods, we can generate a comprehensive connection map of eQTLs and their eGenes' enriched pathways to help us develop a more thorough understanding of eQTLs' involvement in gene regulation, thus providing insights in discovering hidden biological mechanisms (Gilad et al., 2008). In addition, eQTLs studies can also help reveal the architecture of gene regulation, which in combination with results from previous genetic association studies of human traits may help predict regulatory roles for genetic variants previously associated with particular human phenotypes (Gilad et al., 2008). Therefore, it is crucial to explore the associations between eQTLs and genes at the pathway level in complex traits to develop a systematic review of such associations and infer mechanisms of pathogenesis.

The objective of this study was to perform large-scale eQTLs enrichment tests at the pathway level and determine the tissue-specific enriched pathways for trait-related genomic intervals based on the Bioconductor package loci2path (Xu et al., 2020). There are two key advantages of using loci2path than other existing methods: first, we do not depend on physical proximity to provide a link between an eQTL and its target gene, which could be unreliable; second, eQTLs enable us to produce the regulatory annotation for specific tissue types (Xu et al., 2020). For a specific genomic interval containing multiple eQTLs, if eQTLs enrichment analysis indicates that their corresponding eGenes are participating in the same biological pathway, this could imply a potential relationship between that specific pathway and the genomic interval of interest. The tissue-specific eQTLs sets also can demonstrate in what specific tissues would such enrichment be observed, which could help us generate new hypotheses on the biological mechanisms of disease pathogenesis.

In this study, we used the computer program loci2path to perform eQTLs enrichment analysis for genomic regions of ten traits [AD, body mass index, Parkinson's disease (PD), schizophrenia, amyotrophic lateral sclerosis, non-small cell lung cancer (NSCLC), stroke, blood pressure, autism spectrum disorder, and myocardial infarction]. We have updated the loci2path to utilize the most current data sets of query regions, eQTLs sets, and pathway sets. We used the entire multi-tissue eQTLs data from the GTEx V8 data release that contains 13,791,909 eQTLs with 32,958 unique eGenes for 49 tissue types. In addition to BioCarta and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway sets that were included in the original loci2path (Xu et al., 2020), we have added pathway sets from three new pathway databases, i.e., Pathway Interaction Database (PID), Reactome, and WikiPathways to generate more comprehensive results.

2 MATERIALS AND METHODS

2.1 Extension of the loci2path

In this study, we extended the Bioconductor package loci2path (Xu et al., 2020) that runs on an R-based platform, and then applied the extended loci2path to perform eQTLs enrichment analyses at pathway level based on different pathway databases to identify enriched pathways for genomic intervals of multiple traits. The advantage of loci2path is that this computer program uses eQTLs information to directly link to their eGenes, rather than using genome proximity, because an eQTL and its corresponding eGene are not always located near each other. For each gene set, the loci2path will first identify eGenes based on the eQTLs set in the given genomic intervals and then evaluate the significance of these eGenes' enrichment within a gene set. The eQTLs enrichment program really refers to their corresponding eGenes' enrichment because multiple eQTLs could target the same eGenes due to linkage disequilibrium. p-values calculated using Fisher's exact test for an eQTLs set could be computed for each pathway to evaluate the enrichment significance, and those pathways with greater enrichments were indicated by smaller p-values. The results

TABLE 1 | The numbers of genomic intervals selected that contain known GWAS variants for each of the ten complex traits.

Trait	Number of genomic intervals
Alzheimer's Disease	319
Body Mass Index	2,052
Parkinson's Disease	199
Schizophrenia	1,296
Amyotrophic Lateral Sclerosis	342
Non-Small Cell Lung Cancer	120
Stroke	939
Blood Pressure	3,123
Autism Spectrum Disorder	570
Myocardial Infarction	934

were filtered with a p-value of 10^{-4}, which was chosen after multiple trials to balance the number of most significant tissue-pathway combinations and specificity, and used to construct heatmaps for further analysis. We have tried other p-values and obtained similar outcomes.

2.2 Datasets

2.2.1 GTEx eQTLs

For this study, we used the full set of multi-tissue QTL data from the GTEx V8 data release as the input data of eQTLs sets, consisting of 49 tissue types (GTEx Consortium, 2020). The data were downloaded from GTEx through this link: https://storage.googleapis.com/gtex_analysis_v8/multi_tissue_qtl_data/GTEx_Analysis_v8.metasoft.txt.gz. eQTLs sets for each tissue were filtered with a p-value threshold of 10^{-4}. Each gene's entrez ID and gene name were obtained by using the given gene's ensemble gene ID and the Bioconductor package biomaRt.

2.2.2 MSigDB Pathways

A total of 2,893 pathways from BioCarta, KEGG, PID, Reactome, and WikiPathways gene sets were used in this study as the input data of gene sets. The data were downloaded from the MSigDB website: http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp.

2.2.3 Phenotype-Genotype Integrator Query Regions

The list of known trait-associated variants was obtained from National Center for Biotechnology Information (NCBI) via PheGenI website: https://www.ncbi.nlm.nih.gov/gap/phegeni. For a given genetic variant, the genomic region is defined as a flanking 50 kb on each of left and right sides of that variant, which spans 100 kb. Overlapped regions were merged. A total of 9,894 genomic intervals were used in this study, and the numbers of genomic regions for each trait are demonstrated in **Table 1**.

3 RESULTS

3.1 Overview

The objectives of this study were to identify significantly enriched pathways for eQTLs sets of specific tissues at trait-related genomic intervals to generate potentially novel hypotheses of trait determination. A workflow of the study is presented in **Figure 1**, showing that the input data were query regions, and the internal process involved usages of eQTLs sets and gene pathway sets, and the output results were enriched pathways and the corresponding tissues sorted by multiplicity-adjusted enrichment p-values. We used loci2path to conduct eQTLs enrichment analyses by computing the p-values of Fisher's exact test adjusted by Benjamini & Hochberg correction method (Benjamini & Hochberg, 1995), and then converting such results into a heatmap. The heatmap was displayed where each row represents a tissue type, and each column represents a gene pathway. The strong significant enrichments were indicated by red cells, and the weak insignificant enrichments were indicated by blue cells. Other data including eQTLs in pathways, eQTLs in tissues, and hit genes generated by loci2path were used to construct tables. Various adjusted p-values of genes through Fisher's exact test were used as thresholds to filter out the most significant pathway-tissue combinations for each trait. Specific genes that pathways hit in the eQTLs sets were extracted for further analysis. Additional heatmaps and result tables for traits can be found in **Supplementary Figures**. The results of three of the ten traits, i.e., body mass index, amyotrophic lateral sclerosis, and stroke were not presented, because the outputs obtained from eQTLs enrichment tests at the pathway level for these traits were insignificant, and no further analyses could be performed on them.

3.2 Adding PID, Reactome, and WikiPathways to loci2path

We have extended the loci2path (Xu et al., 2020) by adding gene pathway sets of PID, Reactome, and WikiPathways to loci2path′s

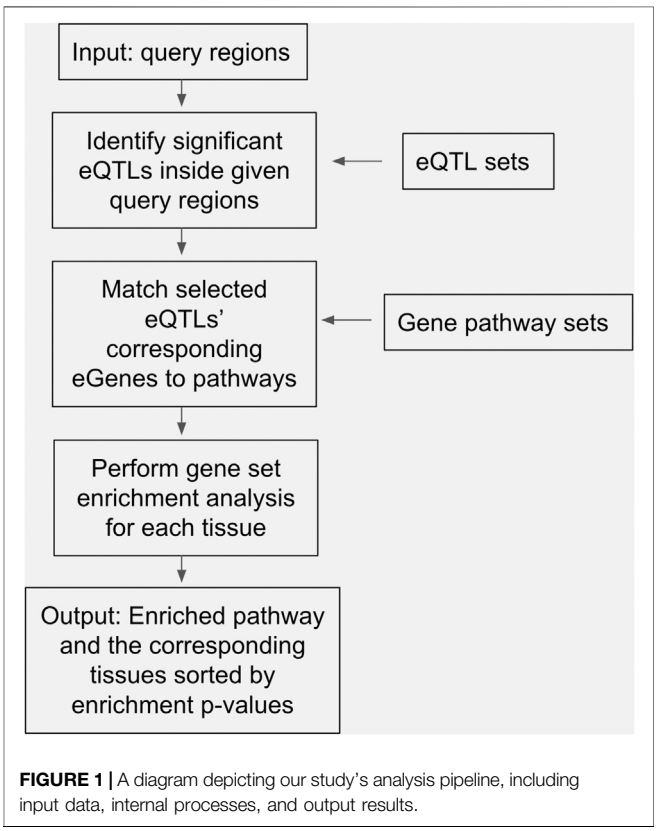


FIGURE 1 | A diagram depicting our study's analysis pipeline, including input data, internal processes, and output results.

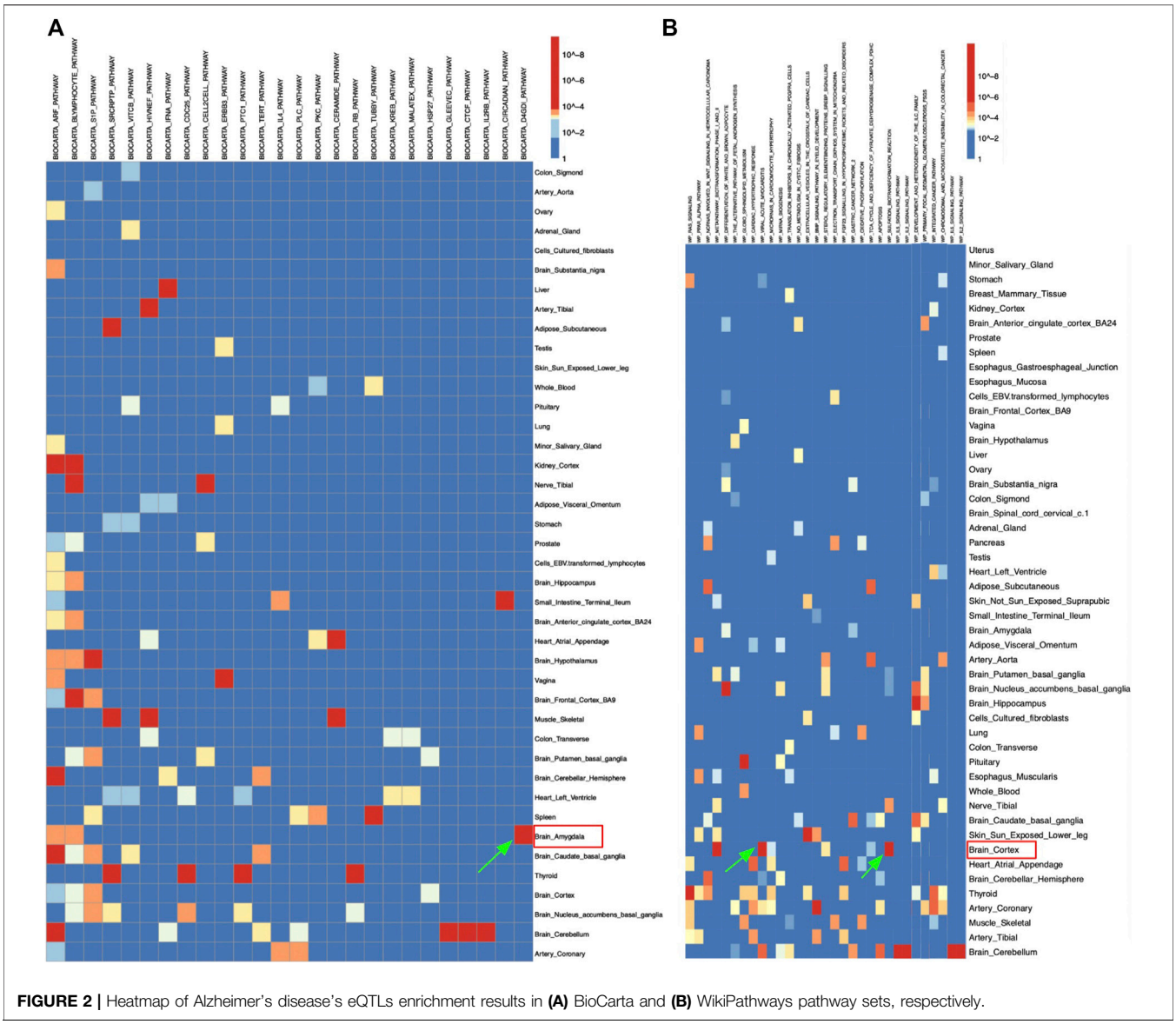


FIGURE 2 | Heatmap of Alzheimer's disease's eQTLs enrichment results in **(A)** BioCarta and **(B)** WikiPathways pathway sets, respectively.

pathway collection. The data of pathway links and NCBI entrez gene IDs were retrieved from the MSigDB website: http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp. The data were separated into two text documents with one containing gene links and the other containing the pathway's gene entrez IDs using a self-written R program (**Supplementary Data Sheet S1**). The documents were added into the loci2path Bioconductor package at loci2path-master/inst/extdata/geneSet, which could be called by the loci2path-running program to match significant eQTLs at the new gene pathway sets.

3.3 Alzheimer's Disease

Currently, there are three major pathology divisions for AD: protein accumulation, neuron loss, and reactive process (Duyckaerts et al., 2009). Past studies have shown that the extracellular accumulation and deposition of amyloid-beta (Aβ) protein induce the appearance of senile plaques and create an abnormal neuron environment, which causes cognitive disabilities (Sadigh-Eteghad et al., 2015; Cheignon et al., 2018). Such accumulation of Aβ not only enhances the interaction between amyloid-forming protein and neuronal membrane and increases membrane permeability through hypothetical mechanisms like amyloid-forming protein's channel-like conductance, but also contributes to the increase in the reactive oxygen species production and thus the disruption of neuronal membrane integrity (Butterfield and Lashuel, 2010; Cheignon et al., 2018).

Figure 2A demonstrated the eQTLs enrichment of AD-related genomic intervals in the BioCarta pathway set. There was a distinct significant enrichment of the D4-GDI pathway in the brain amygdala (**Figure 2A**). Significant eQTLs enrichment results from the amygdala tissue were extracted for further analysis. The table has demonstrated that most pathways' gene hit in brain amygdala tissue was Rho GDP dissociation inhibitor

TABLE 2 | P-values Obtained from Fisher's Exact Test of Significant eQTLs Enrichment for Alzheimer's Disease in BioCarta Pathway Set for Brain Amygdala Tissue

Pathway	Gene hit	Genomic location	Fisher's exact test *p*-value[a]
D4-GDI	*ARHGDIB*	12p12.3	0.020
Blymphocyte	*CR1*	1q32.3	0.023
ARF	*POLR1A*	2p11.2	0.028
Caspase	*ARHGDIB*	12p12.3	0.037
TNFR1	*ARHGDIB*	12p12.3	0.048
FAS	*ARHGDIB*	12p12.3	0.050
HIVNEF	*ARHGDIB*	12p12.3	0.091

[a]*Fisher's exact test* p-*value represents the adjusted* p-*value for genes in the pathway using Fisher's exact test that are adjusted by Benjamini & Hochberg correction method.*

beta (*ARHGDIB*) gene (**Table 2**). The D4-GDI pathway had the lowest *p*-value of genes, which was consistent with the data in **Figure 2A** where the D4-GDI pathway was only enriched in amygdala tissue (**Table 2**; **Figure 2A**). D4-GDI represents the negative regulator of Ras-related Rho GTPases, and its removal is crucial to induce apoptosis since Rho GTPases increase the cytoskeletal and membrane modification related to apoptosis (Coleman and Olson, 2002). As an enzyme that cleaves D4-GDI, caspase-3 was found to be positively correlated with mild cognitive deficiency in early AD pathology (Gastard et al., 2003). Clinical research suggested that Aβ could sequester caspase-3 via direct interaction and induce neuronal apoptosis via caspase-3 activation, thus strengthening AD development (Chang et al., 2016). One possible hypothesis was that an increased level of caspase-3 in the amygdala leads to increased apoptosis and neuronal loss and thus contributes to the memory loss symptom of AD.

Similarly, **Figure 2B** showed significant enrichment of sulfation biotransformation reaction and viral acute myocarditis pathways in brain cortex, IL2 and IL5 signaling pathways in brain cerebellum, and development and heterogeneity of the innate lymphoid cell (ILC) pathway in brain hippocampus for the WikiPathways set (**Figure 2B**). The significant enrichment of viral acute myocarditis pathway in the brain cortex suggested that the correlation observed between heart failure and AD was due to not only the majority of patients' age, but also genetic factors (**Figure 2B**) (Li et al., 2006). Such findings were consistent with a previous study where the viral myocarditis pathway from other pathway sets was identified to be significantly associated with AD (Liu et al., 2014). One population study also found a higher than 80% risk of developing AD for patients with heart failures when major confounders like vascular comorbidities were controlled (Qiu et al., 2006). The significant enrichment in the sulfation biotransformation reaction pathway could also be explained by previous findings (**Figure 2B**). One research suggested an increased frequency of reduced metabolism and impaired sulfation of xenobiotics among AD patients (McFadden, 1996). A clinical study showed that sulfated curcumin can bind to copper and iron ions that are enriched in the brain cortex of AD patients and induce Aβ peptide formation, thus indicating that impaired sulfation ability would increase risk of AD (Baum and Ng, 2004). One possible connection between acute viral myocarditis and AD is kynurenine 3-monooxygenase (KMO), which is a key regulatory enzyme in the kynurenine metabolism pathway that converts kynurenine to 3-hydroxykynurenine (Kubo et al., 2017). Studies have shown that the absence of KMO increased the production of kynurenine pathway metabolite, which lowered the synthesis of chemokine and thus resulted in the decrease of mortality of viral acute myocarditis by encephalomyocarditis virus in mice (Kubo et al., 2017). Interestingly, another study pointed out that JM6, a KMO inhibitor, was found to be able to prevent memory deficiency and synaptic loss in AD mouse models through the increase of the neuroprotective kynurenine metabolite kynurenic acid (Zwilling et al., 2011). Such interaction may imply a hidden mechanism in AD's pathogenesis that increases KMO production and thus decreases levels of neuroprotective kynurenine metabolite and enhances AD symptoms, which explains AD's connection to acute viral myocarditis.

3.4 Parkinson's Disease

One key sign of PD is the accumulation of α-synuclein and the formation of Lewy bodies in brainstem, limbic system, and cortical areas (Alecu and Bennett, 2019). Pathological hallmarks also include the loss of dopaminergic neurons from the substantia nigra and Lewy bodies in surviving cells of affected brains, which leads to reduced voluntary movements (Gegg et al., 2012).

As demonstrated in the **Supplementary Figure S1A**, the enrichment of the KEGG sphingolipid metabolism pathway was observed to be highly and uniquely significant in amygdala tissue, which indicates a correlation between sphingolipid metabolism and PD. This is consistent with previous studies since the metabolism of sphingolipid glucosylceramide catalyzed by glucocerebrosidase (GCase) was found to be deficient in PD patients (Gegg et al., 2012). The deficiency of GCase that catalyzes sphingolipid metabolism has reached up to 40% at amygdala for PD patients compared to normal patients, which is likely to cause α-synuclein accumulation as GCase mRNA level decreased in cells with exogenous α-synuclein (Gegg et al., 2012). One possible explanation for the decreasing GCase could be a mutation at glucosylceramidase-beta gene that encodes this lysosomal enzyme. Similarly, the lysosomal-associated membrane protein 2A and heat shock cognate 70 from lysosome had significantly lower expression levels in amygdala of brains with PD compared to brains with AD or normal brains (Alvarez-Erviti et al., 2010). The chaperone-mediated autophagy strongly depends on these two proteins, and the downregulation of lysosomal-associated membrane protein 2A has increased the mean half-life of α-synuclein from 46.5 to 65 h, suggesting a direct link between this protein and PD (Alvarez-Erviti et al., 2010). Since wild-type α-synuclein was mostly degraded by chaperone-mediated autophagy, it is valid to hypothesize that impaired lysosomal functions could initiate the accumulation of α-synuclein and thus lead to PD.

3.5 Schizophrenia

As demonstrated, most significantly enriched pathways in all 49 tissues were immune-related pathways including allograft rejection, graft vs. host disease, and antigen processing and presentation pathways (**Table 3**). The significantly enriched KEGG allograft rejection pathway in different tissues shared the major histocompatibility complex, Class I, C (*HLA-C*) gene (**Figure 3A**; **Table 3**). *HLA-C* has been shown to be strongly associated with

TABLE 3 | Adjusted *p*-values of the Ten Most Significant eQTLs for Schizophrenia from 49 tissues.

Tissue	Pathway	Gene hits	Genomic locations	Fisher's exact test *p*-value[a]
Breast Mammary Tissue	KEGG allograft rejection	*CD80;HLA-E;HLA-G;HLA-C;HLA-DQB1;HLA-DRB5;HLA-DOB;HLA-DQA2;HLA-DRB1;HLA-DQA1;HLA-DRA;HLA-B*	3q13.33, 6p22.1, 6p21.33, 6p21.32	2.59E-12
	KEGG graft versus host disease	*CD80;HLA-E;HLA-G;HLA-C;HLA-DQB1;HLA-DRB5;HLA-DOB;HLA-DQA2;HLA-DRB1;HLA-DQA1;HLA-DRA;HLA-B*	3q13.33, 6p22.1, 6p21.33, 6p21.32	1.05E-11
	KEGG type I diabetes mellitus	*CD80;HLA-E;HLA-G;HLA-C;HLA-DQB1;HLA-DRB5;HLA-DOB;HLA-DQA2;HLA-DRB1;HLA-DQA1;HLA-DRA;HLA-B*	3q13.33, 6p22.1, 6p21.33, 6p21.32	1.99E-11
Esophagus Gastroesphageal Junction	KEGG type I diabetes mellitus	*CD80;HLA-E;HLA-G;HLA-A;HLA-C;HLA-DQB1;HLA-DRB5;HLA-DQA2;HLA-DMA;HLA-DRA;HLA-DRB1;HLA-DQA1;HLA-B;LTA*	3q13.33, 6p22.1, 6p21.33, 6p21.32	1.59E-14
	KEGG allograft rejection	*CD80;HLA-E;HLA-G;HLA-A;HLA-C;HLA-DQB1;HLA-DRB5;HLA-DQA2;HLA-DMA;HLA-DRA;HLA-DRB1;HLA-DQA1;HLA-B*	3q13.33, 6p22.1, 6p21.33, 6p21.32	6.37E-14
	KEGG graft versus host disease	*CD80;HLA-E;HLA-G;HLA-A;HLA-C;HLA-DQB1;HLA-DRB5;HLA-DQA2;HLA-DMA;HLA-DRA;HLA-DRB1;HLA-DQA1;HLA-B*	3q13.33, 6p22.1, 6p21.33, 6p21.32	3.00E-13
	KEGG antigen processing and presentation	*CTSS;HLA-E;HLA-G;HLA-A;HLA-C;HLA-DQB1;HLA-DRB5;HLA-DQA2;HLA-DMA;HLA-DRA;HLA-DRB1;HLA-DQA1;TAP2;TAPBP;HLA-B;LTA*	6p22.1, 6p21.33, 6p21.32, 1q21.3	2.52E-12
	KEGG autoimmune thyroid disease	*CD80;HLA-E;HLA-G;HLA-A;HLA-C;HLA-DQB1;HLA-DRB5;HLA-DQA2;HLA-DMA;HLA-DRA;HLA-DRB1;HLA-DQA1;HLA-B*	3q13.33, 6p22.1, 6p21.33, 6p21.32	9.45E-12
Muscle Skeletal	KEGG allograft rejection	*CD80;HLA-E;HLA-C;HLA-G;HLA-DQB1;HLA-DRB5;HLA-DMA;HLA-DRA; HLA-DQA2;HLA-DRB1;HLA-DQA1;HLA-A, HLA-B*	3q13.33, 6p22.1, 6p21.33, 6p21.32	5.14E-12
	KEGG graft versus host disease	*CD80;HLA-E;HLA-C;HLA-G;HLA-DQB1;HLA-DRB5;HLA-DMA;HLA-DRA; HLA-DQA2;HLA-DRB1;HLA-DQA1;HLA-A, HLA-B*	3q13.33, 6p22.1, 6p21.33, 6p21.32	2.37E-11

[a]*Fisher's exact test* p-*value represents the adjusted* p-*value for genes in the pathway using Fisher's exact test that are adjusted by Benjamini & Hochberg correction method.*

schizophrenia by multiple past studies. HLA-C*01:02 was positively associated with schizophrenia, while HLA-C*07:01 was negatively associated with schizophrenia (Andreassen et al., 2015; Corvin, 2012). One study suggested that in the absence of glutamic acid at the 74th position of the mature protein encoded by the major histocompatibility complex, Class II, DR Beta 1(*HLA-DRB1*) gene, the amino acid methionine at the 99th position of *HLA-C* may contribute to individuals' susceptibility to schizophrenia, in which the glutamic acid in *HLA-DRB1* has a protective function against the disease (Seshasubramanian et al., 2020). Interestingly, *HLA-DRB1* was hit by the majority of tissues enriched with the KEGG allograft rejection pathway (**Table 3**). Similarly, the major histocompatibility complex, Class II, DQ Beta 1 (*HLA-DQB1*) gene was also shared by most tissues with such a pathway, a molecule that presents peptides derived from extracellular proteins and is expressed in antigen expression cells (**Table 3**). DQB1*05:01:01 was also positively associated with schizophrenia and the predominant haplotype in the schizophrenia population, while decreased frequency of DQB1*02:01 was found among schizophrenia patients (Katrinli et al., 2019; Seshasubramanian et al., 2020). No studies have been conducted on specific mechanisms of *HLA-C*, *HLA-DRB1*, and *HLA-DQB1*'s interventions in schizophrenia pathogenesis, but their interaction is much likely to contribute to the disease.

In the PID pathway set, the FOXO pathway was significantly enriched in the eQTLs set of thyroid tissue, which suggested a potential correlation between the forehead box transcription factor O family and schizophrenia at thyroid (**Figure 3B**). The mRNA expression level of FOXO pathway genes including *FOXO1* and *FOXO3A* were significantly lower in patients with acute schizophrenia (Gu et al., 2021).

3.6 Non-Small Cell Lung Cancer

The BioCarta IL1R pathway was shown to be significantly enriched in the testis tissue for NSCLC (**Supplementary Figure S2**). The IL1R pathway involves signal transduction through interleukin-1. One study found that interleukin-17 (IL-17) was involved in angiogenesis in a variety of inflammatory associated cancers, although it still remains unclear how IL-17 contributes to the process (Pan et al., 2015). It is also known that interleukin-37 (IL-37), a new member of the interleukin-1 family, plays an immunosuppressive role in a variety of inflammatory disorders. A study recently found that IL-37 demonstrates a protective role in cancer development possibly through tumor angiogenesis and that it could serve as a promising therapeutic target for NSCLC (Ge et al., 2016).

In **Supplementary Figure S2**, the PCG1A pathway was enriched in the kidney. The PCG1A pathway involves the regulation of peroxisome proliferator-activated receptor gamma coactivator-1 alpha (PGC-1a), which is a tissue-specific coactivator that enhances the activity of many nuclear receptors and coordinates transcriptional programs

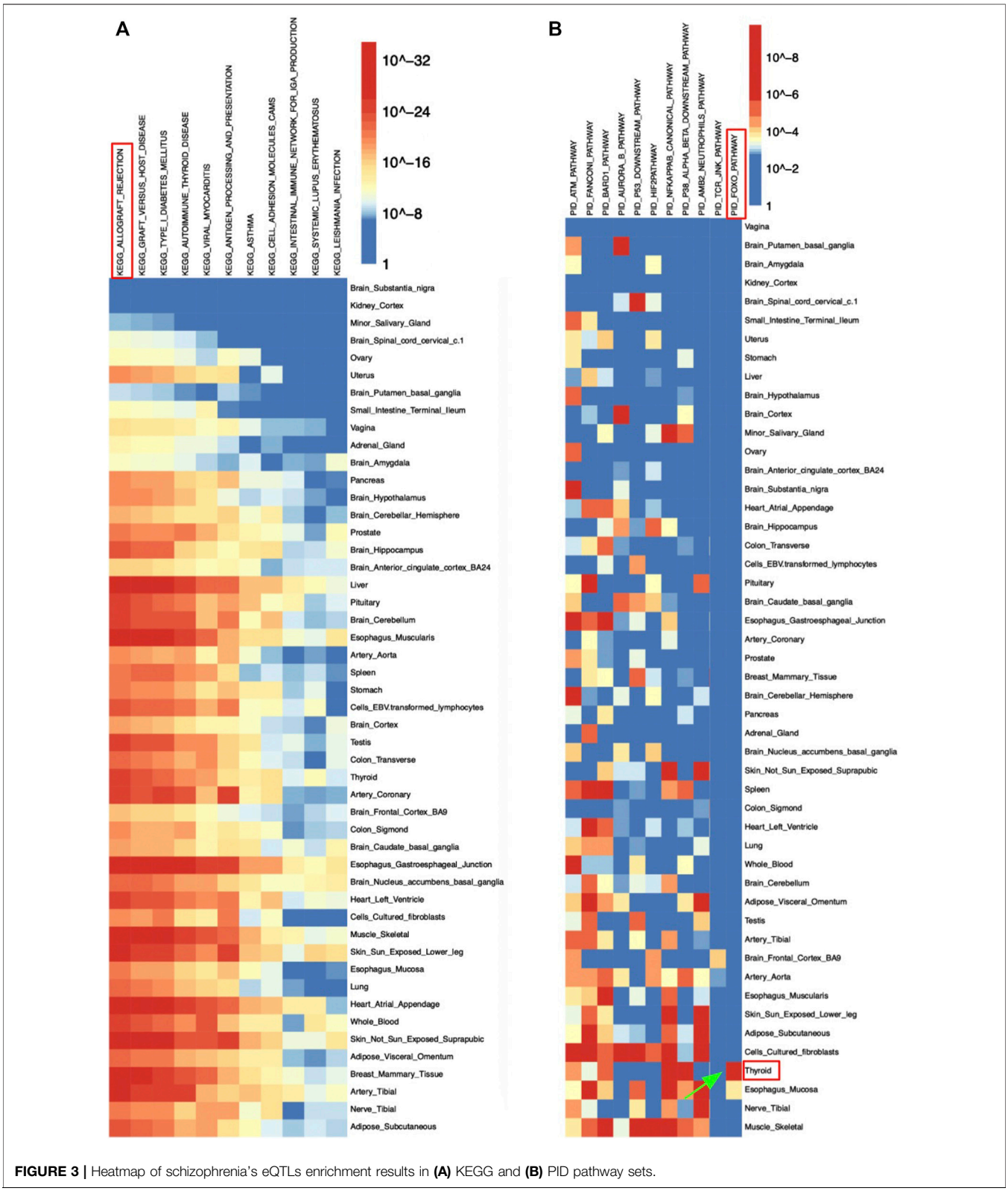


FIGURE 3 | Heatmap of schizophrenia's eQTLs enrichment results in **(A)** KEGG and **(B)** PID pathway sets.

important for energy metabolism and homeostasis. In NSCLC patients, there are a variety of cases where the cells show therapeutic resistance. As a result, a plethora of studies focus on drug resistance mechanisms, but not many have focused on the metabolic flexibility of drug-resistant NSCLC. In one study, it was found that during the development

TABLE 4 | Adjusted *p*-values of Five Most Significant eQTLs for Blood Pressure in KEGG and WikiPathways Sets for Heart Atrial Appendage Tissue.

Pathway	Gene hits	Genomic locations	Fisher's exact test *p*-value[a]
WikiPathways Ebola virus pathway on host	*MERTK;KPNA1;RFC1;ITGA2;HLA-G;HLA-A;HLA-C;HLA-B;HLA-E;HLA-DOA;HLA-DRB5;HLA-DQB2;HLA-DMA;HLA-DPA1;HLA-DRB1;HLA-DPB1;HLA-DQA2;HLA-F;HLA-DQB1;HLA-DOB;HLA-DQA1;HLA-DRA;RAC1;SCIN;CAV2;CAV1;CTSB;ITGB1;TPCN2;MFGE8;IQGAP1;NPC1;VPS16*	6p22.1, 6p21.33, 6p21.32, 2q13, 3q21.1, 4p14, 5q11.2, 7p22.1, 7p21.3, 7q31.2, 8p23.1, 10p11.22, 11q13.3, 15q26.1, 18q11.2, 20p13	3.64E-13
WikiPathways allograft rejection	*CASP9;CD55;CD86;CSCL8;PDGFRA;BHMT2;HLA-G;HLA-A; HLA-C;HLA-B;C4A;HLA-E;MICA;HLA-DOA;HLA-DRB5;HLA-DMA;HLA-DPA1;HLA-DRB1;HLA-DPB1; HLA-DQA2;HLA-F;HLA-DQB1;HLA-DOB;C4B;HLA-DQA1;HLA-DRA;LRRK2*	6p22.1, 6p21.33, 6p21.32, 1p36.21, 1q32.2, 3q13.33, 4q12, 5q14.1, 12q12	1.03E-12
KEGG allograft rejection	*CD86;HLA-G;HLA-A;HLA-C;HLA-B;HLA-E;HLA-DOA;HLA-DRB5;HLA-DMA;HLA-DPA1;HLA-DRB1;HLA-DPB1;HLA-DQA2;HLA-F;HLA-DQB1;HLA-DOB;HLA-DQA1;HLA-DRA*	6p22.1, 6p21.33, 6p21.32, 3q13.33	1.03E-12
KEGG viral myocarditis	*CASP9;CD55;CD86;HLA-G;HLA-A;HLA-C;HLA-B;HLA-E;HLA-DOA;HLA-DRB5;HLA-DMA;HLA-DPA1;HLA-DRB1;HLA-DPB1;HLA-DQA2;HLA-F;HLA-DQB1;HLA-DOB;HLA-DQA1;HLA-DRA;RAC1;CAV1;RAC3*	6p22.1, 6p21.33, 6p21.32, 7p22.1, 7q31.2, 1p36.21, 1q32.2, 3q13.33, 17q25.3	5.70E-12
KEGG graft versus host disease	*CD86;HLA-G;HLA-A;HLA-C;HLA-B;HLA-E;HLA-DOA;HLA-DRB5;HLA-DMA;HLA-DPA1;HLA-DRB1;HLA-DPB1;HLA-DQA2;HLA-F;HLA-DQB1;HLA-DOB;HLA-DQA1;HLA-DRA*	6p22.1, 6p21.33, 6p21.32, 3q13.33	9.82E-12

[a]*Fisher's exact test* p-*value represents the adjusted* p-*value for genes in the pathway using Fisher's exact test that are adjusted by Benjamini & Hochberg correction method.*

of resistance for tyrosine kinase inhibitors, NSCLC cells switched from glycolysis to oxidative phosphorylation through increasing activity of the mitochondria. Cells were treated with the MCT-1 inhibitor AZD3965 and there was a resulting significant decrease in cell proliferation and motility in TK1-sensitive and TK-resistant cells. A study recently found that IL-37 demonstrates a protective role in cancer development possibly through tumor angiogenesis and that it could serve as a promising therapeutic target for NSCLC (Huang et al., 2020).

3.7 Blood Pressure

For blood pressure, the majority of the pathways most significantly enriched in tissues were immune-related, and the atrial appendages tissue contained the most pathways with the most significant *p*-values (**Table 4**). The role of the immune system in the pathogenesis of hypertension has been firmly established by many laboratories. The KEGG viral myocarditis pathway and the tissue heart atrial appendage had one of the most significant *p*-values at 3.08E-14, the KEGG type I diabetes mellitus pathway was also significantly enriched at the atrial appendage tissue (**Table 4**).

Myocarditis is a cardiac disease associated with inflammation and injury of the myocardium. It results from various etiologies, but coxsackievirus is considered the dominant etiological agent. Infiltrating macrophages have been proven as a pivotal pathological inflammatory cell subset in coxsackievirus induced viral myocarditis, however, the mechanisms involving initiation and promotion are still unknown (Zhang et al., 2017).

Type 1 diabetes is the autoimmune destruction of the insulin producing beta-cells. High blood pressure is a common symptom of diabetes because the high levels of glucose in the blood damage the blood vessels and lead to hypertension. One study found that the left atrium mechanical functions were impaired in patients with type 1 diabetes (Acar et al., 2009).

3.8 Autism Spectrum Disorder

Few significant pathways were uniquely enriched in one or two tissues for autism spectrum disorder as shown in **Supplementary Figures S3–S7**. KEGG pathways of drug metabolism by cytochrome p450 and metabolism of xenobiotics by cytochrome p450 were found to be enriched in various tissues and most significantly in the liver tissue (**Supplementary Figure S4**; **Table 5**). Out of 29 most significant pathway-tissue combinations passing the *p*-value threshold of 10^{-4}, genes *GSTM3* and *GSTM5* were hit 24 times, followed by genes *GSTM1*, *GSTP1*, *GSTM4*, and *GSTM2* (**Supplementary Table S1**). The two most significantly enriched pathways, Reactome phase II conjugation of compounds and KEGG metabolism of xenobiotics by cytochrome p45 pathways, were in liver tissues, and they have both hit genes *GSTM2*, *GSTM3*, *GSTM4*, and *GSTM5*, which encode for multiple proteins from the glutathione S-transferase mu class (**Table 5**). The two pathways cover proteins functioning in pharmacological inactivation of chemicals and detoxification, and the mu class enzymes are known for their functions in detoxification of electrophilic compounds by conjugation with glutathione (Cheng et al., 2020). Therefore, such highly significant adjusted *p*-values suggested a key role glutathione S-transferase mu enzymes play in autism spectrum disorder (**Table 5**). Studies have shown that when exposed to chronic heavy metal and chemical xenobiotic pollution, patients with autism spectrum disorder demonstrated significantly higher total glutathione and oxidized glutathione in red blood cells (Faber

TABLE 5 | Adjusted *p*-values of 10 Most Significant eQTLs for autism spectrum disorder from 49 tissues.

Tissue	Pathway	Gene hits	Genomic locations	Fisher's exact test *p*-value[a]
Adipose Visceral Omentum	Reactome biological oxidations	*GSTM5;GSTM3;GSTM1;GSTM4;EPHX1; NCOA1;ABHD14B;UGT2A1;SULT1E1; SLC26A1;UGT2B7;UGT3A2;AIP;GSTP1; CES1;CYB5B;ALDH3A1*	1p13.3, 1q42.12, 2p23.3, 3p21.2, 4q13.3, 4p16.3, 3q13.2, 5p13.2, 11q13.2, 16q12.2, 16q22.1, 17p11.2	3.44E-06
Brain Anterior cingulate cortex BA24	WikiPathways photodynamic therapyinduced NFE2L2 NRF2 survival signaling	*GCLM;EPHX1;ABCC2;GSTP1;CES1;NQO1; SRXN1*	1q42.12, 11q13.2, 16q12.2, 1p22.1, 10q24.2, 16q22.1, 20p13	7.66E-06
Brain Caudate basal ganglia	KEGG steroid hormone biosynthesis	*SRD5A3;UGT2A1;UGT2B4;UGT2B15; SULT1E1;UGT2B28*	4q13.3, 4q12, 4q13.2	1.62E-05
Colon Transverse	KEGG metabolism of xenobiotics by cytochrome p450	*GSTM5;GSTM3;GSTM2;GSTM1;GSTM4; EPHX1;UGT2B4;GSTP1;ALDH3A1*	1p13.3, 1q42.12, 11q13.2, 17p11.2, 4q13.3	1.28E-05
Kidney Cortex	Reactome biological oxidations	*GSTM5;GSTM3;GSTM4;GSTM2;NCOA1; UGT2A1;UGT2B4;UGT2B15;SULT1E1; UGT2B28;UGT3A2*	1p13.3, 2p23.3, 4q13.3, 5p13.2, 4q13.2	1.23E-06
Liver	Reactome phase II conjugation of compounds	*GSTM5;GSTM3;GSTM4;GSTM2;UGT2A1; UGT2B4;UGT2B15;SULT1E1;UGT2B28; UGT3A2*	1p13.3, 4q13.3, 5p13.2, 4q13.2	1.73E-08
	KEGG metabolism of xenobiotics by cytochrome p450	*GSTM5;GSTM3;GSTM4;GSTM2;UGT2A1; UGT2B4;UGT2B15;UGT2B28*	1p13.3, 4q13.3, 4q13.2	7.30E-08
	KEGG metabolism of xenobiotics by cytochrome p450	*GSTM5;GSTM3;GSTM1;GSTM4;EPHX1; UGT2A1;UGT2B7;ALDH3B2;GSTP1; ALDH3A1*	1p13.3, 1q42.12, 4q13.3, 3q13.2, 11q13.2, 17p11.2	3.71E-06
Lung	KEGG pentose and glucuronate interconversion	*UGDH;UGT2B4;UGT2A1;DHDH*	4q13.3, 4p14, 19q13.33	7.16E-06
Skin Not Sun Exposed Suprapubic	KEGG drug metabolism cytochrome p450	*GSTM5;GSTM3;GSTM4;GSTM2;UGT2A1; UGT2B4;UGT2B15;UGT2B28*	1p13.3, 4q13.3, 4q13.2	9.15E-08

[a]*Fisher's exact test* p-*value represents the adjusted* p-*value for genes in the pathway using Fisher's exact test that are adjusted by Benjamini & Hochberg correction method.*

et al., 2019). The study also believed that the elevated glutathione was a compensatory mechanism to the exposure of a high xenobiotic environment (Faber et al., 2019). However, such a mechanism could not deal with oxidative stress as the reduced to oxidized glutathione ratio was lower in autistic patients, which indicates a crucial role glutathione plays in the xenobiotic detoxification among patients with autism spectrum disorder (Faber et al., 2019; Bjørklund et al., 2020).

3.9 Myocardial Infarction

Supplementary Figure S8 demonstrated the eQTLs enrichment in BioCarta and Reactome pathway sets of myocardial infarction-related genomic intervals. The AT1R pathway from the BioCarta pathway set was significantly enriched in brain cortex tissue (**Supplementary Figure S8A**), and the cell cycle pathway from the Reactome pathway set was enriched in whole blood tissue (**Supplementary Figure S8B**), respectively. *RAC1* gene was hit by the BioCarta AT1R pathway at the brain cortex tissue, and *PPP2R5A* gene was hit by the Reactome cell cycle pathway at the whole blood tissue (**Table 6**). In myocardial infarction, the RAC1 protein in the brain cortex tissue paired with the BioCarta AT1R pathway was enriched. The RAC1 protein belongs to the RAS superfamily of small GTP-binding proteins. Members of this superfamily appear to regulate a diverse array of cellular events, including the control of cell growth, cytoskeletal reorganization, and the activation of protein kinases. In terms of myocardial infarction, the RAC1 protein serves as a small GTP-binding protein that regulates NADPH oxidase. NADPH oxidase is a reactive oxygen species (ROS) that contributes to heart failure, such as myocardial infarction. Failing of the myocardium in patients with dilated cardiomyopathy (DCM) and ischemic cardiomyopathy (ICM) is characterized by an upregulation of NADPH oxidase–mediated ROS release associated with increased RAC1 activity (Maack et al., 2003).

Furthermore, the AT1R pathway is responsible for promoting hypertension, G protein-dependent signaling, transactivation of growth factor receptors, NADPH oxidase, and ROS signaling explaining why the *RAC1* gene was enriched by the AT1R pathway (Kawai et al., 2017). In addition to the *RAC1* gene, the *PPP2R5A* gene in the tissue whole blood paired with the Reactome cell cycle pathway was hit on. The *PPP2R5A* gene stands for protein phosphatase 2 regulatory subunit B'alpha. The gene serves as a subunit of the protein phosphatase 2A (PP2A) holoenzyme, which plays an essential role in regulating a diverse array of myocyte functions through dephosphorylation of target molecules. Functioning as an important phosphatase, the PP2A holoenzyme is critical for serving as a regulatory module within the heart, such that dysregulation of PP2A function may contribute to cardiac diseases. Alterations in PP2A activity are associated with heart failure and arrhythmia (Lubbers and

TABLE 6 | Adjusted *p*-values of Five Most Significant eQTLs for Myocardial Infarction in BioCarta and Reactome Pathway Sets from 49 tissues.

Tissue	Pathway	Gene hits	Genomic locations	Fisher's exact test *p*-value[a]
Brain Cortex	BioCarta AT1R pathway	*SHC1;AGT;RAC1;GNAQ;MAPK3*	1q21.3, 1q42.2, 7p22.1, 9q21.2, 16p11.2	0.00378
	BioCarta PYK2 pathway	*SHC1;MAPK14;RAC1;GNAQ;MAPK3*	1q21.3, 7p22.1, 9q21.2, 16p11.2, 6p21.31	0.00378
Brain Nucleus accumbens basal ganglia	Reactome glutathione conjugation	*GSTM2;GSTM5;GSTM1;HPGDS;GGCT;GSTO1; CNDP2*	1p13.3, 4q22.3, 7p14.3, 10q25.1, 18q22.3	3.71E-05
Lung	BioCarta ATRBRCA pathway	*RAD17;FANCE;FANCG;MRE11;FANCA*	5q13.2, 6p21.31, 9p13.3, 11q21, 16q24.3	0.00950
Ovary	BioCarta ATRBRCA pathway	*RAD17;FANCG;MRE11;FANCA*	5q13.2, 9p13.3, 11q21, 16q24.3	0.00385
Testis	Reactome signaling by Rho GTPases	*KDM1A;WASF2;YWHAQ;CENPC;RASGRF2; IQGAP2;H2BC1;H3C6;H2BC3;H2AC4;H2BC4; CENPQ;MAPK14;H3C12;RAC1;H2AZ2; ARHGEF35;ARHGEF10;DLC1;RHOBTB1;CFL1; KLC2;CTTN;RHOG;RHOJ;MAPK3;SKA1; SPC24;SRC*	1p36.12, 6q21, 2p25.1, 4q13.2, 7p22.1, 16p11.2, 6p21.31, 5q14.1, 5q13.3, 6p22.2, 6p12.3, 6p22.1, 7p13, 7q35, 8p23.3, 8p22, 10q21.2, 11q13.1, 11q13.2, 11q13.3, 11p15.4, 14q23.2, 18q21.1, 19p13.2, 20q11.23	8.38E-05
Whole Blood	Reactome cell cycle	*PPP2R5A;AHCTF1;LPIN1;VRK2;MZT2A;ANAPC4; CENPC;DHFR;H3C6;H4C3;H2BC5;CENPQ; TUBB2B;TUBB2A;CDKN1A;H4C12;H2BC14; POM121;MAD1L1;H2AZ2;POM121C;PRKAR2B; MCM4;RAB2A;DCTN3;CDKN2B;CDKN2A;SMC2; PPP2R2D;BANF1;RAB1B;MRE11;NUP98; ANKLE2;PSMC6;PPP2R5E;MAPK3;SPC24; CHMP4B;DSN1*	16p11.2, 11q21, 4q13.2, 6p22.2, 6p12.3, 7p13, 19p13.2, 1q32.3, 1q44, 2p25.1, 2p16.1, 2q21.1, 4p15.2, 5q14.1, 6p25.2, 6p21.2, 6p22.1, 7q11.23, 7p22.3, 7q22.3, 8q11.21, 8q12.1, 9p13.3, 9p21.3, 9q31.1, 10q26.3, 11q13.1, 11q13.2, 11p15.4, 12q24.33, 14q22.1, 14q23.2, 20q11.22, 20q11.23	1.61E-07
	Reactome Rho GTPase effectors	*WASF2;PPP2R5A;AHCTF1;CENPC;H3C6;H4C3; H2BC5;CENPQ;TUBB2B;TUBB2A;H4C12; H2BC14;MAD1L1;RAC1;H2AZ2;NCF1;CTTN; RHOG;NUP98;PPP2R5E;MAPK3;SPC24;DSN1*	6q21, 7p22.1, 16p11.2, 4q13.2, 6p22.2, 6p12.3, 7p13, 11q13.3, 11p15.4, 19p13.2, 1q32.3, 1q44, 6p25.2, 6p22.1, 7p22.3, 11p15.4, 14q23.2, 20q11.23, 7q11.23	3.08E-05
	Reactome signaling by Rho GTPases	*WASF2;PPP2R5A;AHCTF1;CENPC;ARAP2; H3C6;H4C3;H2BC5;CENPQ;TUBB2B;TUBB2A; H4C12;H2BC14;MAD1L1;RAC1;H2AZ2;NCF1; ARHGEF35;ARHGEF5;DLC1;CTTN;RHOG; NUP98;PPP2R5E;MAPK3;SPC24;DSN1*	6q21, 7p22.1, 16p11.2, 4q13.2, 6p22.2, 6p12.3, 7p13, 7q35, 8p22, 11q13.3, 11p15.4, 19p13.2, 1q32.3, 1q44, 6p25.2, 6p22.1, 7p22.3, 11p15.4, 14q23.2, 20q11.23, 7q11.23, 4p14, 7q35	8.11E-05
	BioCarta MAPK pathway	*MAP3K6;SHC1;MAP3K7;RIPK1;MAPK13;RAC1; MAP3K11;RPS6KA5;MAPK3*	1p36.11, 7p22.1, 1q21.3, 16p11.2, 6q15, 6p25.2, 6p21.31, 11q13.1, 14q32.11	0.00499

[a]*Fisher's exact test* p-*value represents the adjusted* p-*value for genes in the pathway using Fisher's exact test that are adjusted by Benjamini & Hochberg correction method.*

Mohler, 2016). The varying types of myocardial infarction make it difficult for researchers to pinpoint a cure. In recent years, scientists have recognized multiple types of myocardial infarction with different causes, yet the knowledge of its pathogenic mechanisms is still poorly understood and greatly lacking (DeFilippis et al., 2019). While the different causes of myocardial infarction can be difficult to pinpoint, we can start by identifying the pathways, tissues, genes that are related to the causes. The results have shown some genomic mechanisms contributing to myocardial infarction, whether it be the enrichment of the RAC1 protein leading to the regulation of NADPH oxidase causing heart failure, or the altered regulation in the PP2A gene leading to heart failure and arrhythmia. The importance of these findings is two-fold: first, these results could serve as a pipeline to benefit the scientific community through reducing repeated work, and second, the discovered specific pathway-tissue-gene results could help researchers to reveal pathogenesis mechanisms in myocardial infarction in hopes to lower its occurrence rates or raise the rates of survival.

4 DISCUSSION

We have extended the loci2path (Xu et al., 2020) by using the latest multi-tissue eQTLs data set from GTEx V8 release and adding PID, Reactome, and WikiPathways databases. The total numbers of eQTLs for each of 49 tissues we used in this study are shown in **Supplementary Table S2**. Our results of enrichment analysis have suggested multiple novel biological hypotheses of disease mechanisms for AD, PD, and schizophrenia. The proposed mechanisms of the increase of caspase-3 level in amygdala tissue and KMO production that may contribute to AD's memory loss symptoms by increasing apoptosis and neuronal loss and decreasing kynurenine metabolite levels were supported by multiple past studies. The impaired lysosomal functions of GCase, lysosomal-associated membrane protein 2A, and heat shock cognate 70 resulted from mutations in genes corresponding to these proteins may cause α-synuclein accumulation to begin and thus lead to PD. The interaction

among *HLA-C*, *HLA-DRB1*, and *HLA-DQB1* is likely to take part in schizophrenia's pathogenesis as well.

Our study has extensively evaluated multiple gene pathways' involvements in the ten traits and further investigated significant genes in each pathway that were hit in the given genomic query regions. The proposed hypotheses have opened new avenues to explore the underlying molecular mechanisms and thus could illuminate further investigations on these traits. We have also found many interesting associations between eQTLs and gene pathways at trait-associated variants of NSCLC, blood pressure, autism spectrum disorder, and myocardial infarction which provided valuable insights into our comprehensive understandings of them. Furthermore, our study has confirmed the advantages of using tissue-specific eQTLs enrichment analysis at pathway level, because our findings based on loci2path software were strongly supported by multiple previous studies (Xu et al., 2020). This has indicated that using eQTLs catalogs to find links between genomic loci and their corresponding eGenes is valid and should be vastly applied in future studies involving gene sets and traits.

There were several limitations in our study. Due to the nature of the statistical analysis, our findings from loci2path could not be considered as providing direct understandings of biological mechanisms underpinning these traits, and we were only able to generate hypotheses for trait determination. These hypotheses should be experimentally verified by conducting further in-depth functional studies by molecular biology laboratories. In addition, loci2path′s reliance on current eQTLs sets data from GTEx could also lead to biased results since the eQTLs sets data from brain tissues were significantly smaller than other tissues like tibial nerves, leg skin without sun exposure, and thyroid. This was caused by the limited sample sizes of brain tissues from GTEx, which may result in missing important biological pathways in brain tissues for neurodegenerative diseases due to inadequate statistical power. The imbalance of eQTLs sizes of various tissues could also bring false-positive results in tissues with more samples and generate coincidental enrichment of certain pathways at tissues not related to the traits. Therefore, results from loci2path need to be treated with extra care, and only the most significant tissue-pathway associations should be extracted for analysis with sufficient past evidence. The software itself also has rooms for improvement, like including new gene pathway sets and adding annotations on pathways uniquely enriched in a tissue.

Future studies on neurodegenerative diseases specifically should implement more data on brain tissues to increase the accuracy of loci2path. Other neurodegenerative diseases like bipolar disorder and attention deficit disorder could be added for a systematic analysis on their patterns to find potential patterns for commonality among this type of disease.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

BW has modified the loci2path program and generated and processed the raw data. BW, SQ, and JY did the analysis. ZQ and YB supervised the project and provided suggestions and guidance on directions. All authors participated in writing and revising.

ACKNOWLEDGMENTS

Funding for article processing charge is provided by the Halle Institute for Global Research at Emory University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2021.719737/full#supplementary-material

REFERENCES

Acar, G., Akcay, A., Sokmen, A., Ozkaya, M., Guler, E., Sokmen, G., et al. (2009). Assessment of Atrial Electromechanical Delay, Diastolic Functions, and Left Atrial Mechanical Functions in Patients with Type 1 Diabetes Mellitus. *J. Am. Soc. Echocardiography* 22 (6), 732–738. doi:10.1016/j.echo.2009.03.028

Alecu, I., and Bennett, S. A. L. (2019). Dysregulated Lipid Metabolism and its Role in α-Synucleinopathy in Parkinson's Disease. *Front. Neurosci.* 13, 328. doi:10.3389/fnins.2019.00328

Alvarez-Erviti, L., Rodriguez-Oroz, M. C., Cooper, J. M., Caballero, C., Ferrer, I., Obeso, J. A., et al. (2010). Chaperone-Mediated Autophagy Markers in Parkinson Disease Brains. *Arch. Neurol.* 67 (12), 1464–1472. doi:10.1001/archneurol.2010.198

Andreassen, O. A., Harbo, H. F., Harbo, H. F., Wang, Y., Thompson, W. K., Schork, A. J., et al. (2015). Genetic Pleiotropy between Multiple Sclerosis and Schizophrenia but Not Bipolar Disorder: Differential Involvement of Immune-Related Gene Loci. *Mol. Psychiatry* 20 (2), 207–214. doi:10.1038/mp.2013.195

Baum, L., and Ng, A. (2004). Curcumin Interaction with Copper and Iron Suggests One Possible Mechanism of Action in Alzheimer's Disease Animal Models. *Jad* 6 (4), 367–377. doi:10.3233/JAD-2004-6403

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodological)* 57 (1), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bjørklund, G., Tinkov, A. A., Hosnedlová, B., Kizek, R., Ajsuvakova, O. P., Chirumbolo, S., et al. (2020). The Role of Glutathione Redox Imbalance in Autism Spectrum Disorder: A Review. *Free Radic. Biol. Med.* 160, 149–162. doi:10.1016/j.freeradbiomed.2020.07.017

Butterfield, S. M., and Lashuel, H. A. (2010). Amyloidogenic Protein-Membrane Interactions: Mechanistic Insight from Model Systems. *Angew. Chem. Int. Edition* 49 (33), 5628–5654. doi:10.1002/anie.200906670

Chang, Y.-J., Linh, N. H., Shih, Y. H., Yu, H.-M., Li, M. S., and Chen, Y.-R. (2016). Alzheimer's Amyloid-β Sequesters Caspase-3 *In Vitro* via its C-Terminal Tail. *ACS Chem. Neurosci.* 7 (8), 1096–1106. doi:10.1021/acschemneuro.6b00049

Cheignon, C., Tomas, M., Bonnefont-Rousselot, D., Faller, P., Hureau, C., and Collin, F. (2018). Oxidative Stress and the Amyloid Beta Peptide in Alzheimer's Disease. *Redox Biol.* 14, 450–464. doi:10.1016/j.redox.2017.10.014

Cheng, C.-H., Ma, H.-L., Deng, Y.-Q., Feng, J., Chen, X.-L., and Guo, Z.-X. (2020). The Role of Mu-type Glutathione S-Transferase in the Mud Crab (Scylla Paramamosain) during Ammonia Stress. *Comp. Biochem. Physiol. C: Toxicol. Pharmacol.* 227, 108642. doi:10.1016/j.cbpc.2019.108642

Coleman, M. L., and Olson, M. F. (2002). Rho GTPase Signalling Pathways in the Morphological Changes Associated with Apoptosis. *Cell Death Differ* 9, 493–504. doi:10.1038/sj.cdd.4400987

Corvin, A. (2012). Irish Schizophrenia Genomics Consortium and the Wellcome Trust Case Control Consortium 2Genome-wide Association Study Implicates HLA-C*01:02 as a Risk Factor at the Major Histocompatibility Complex Locus in Schizophrenia. *Biol. Psychiatry* 72 (8), 620–628. doi:10.1016/j.biopsych.2012.05.035

DeFilippis, A. P., Chapman, A. R., Mills, N. L., de Lemos, J. A., Arbab-Zadeh, A., Newby, L. K., et al. (2019). Assessment and Treatment of Patients with Type 2 Myocardial Infarction and Acute Nonischemic Myocardial Injury. *Circulation* 140 (20), 1661–1678. doi:10.1161/CIRCULATIONAHA.119.040631

Duyckaerts, C., Delatour, B., and Potier, M.-C. (2009). Classification and Basic Pathology of Alzheimer Disease. *Acta Neuropathol.* 118, 5–36. doi:10.1007/s00401-009-0532-1

Faber, S., Fahrenholz, T., Wolle, M. M., Kern, J. C., Pamuku, M., Miller, L., et al. (2019). Chronic Exposure to Xenobiotic Pollution Leads to Significantly Higher Total Glutathione and Lower Reduced to Oxidized Glutathione Ratio in Red Blood Cells of Children with Autism. *Free Radic. Biol. Med.* 134, 666–677. doi:10.1016/j.freeradbiomed.2019.02.009

Gastard, M. C., Troncoso, J. C., and Koliatsos, V. E. (2003). Caspase Activation in the Limbic Cortex of Subjects with Early Alzheimer's Disease. *Ann. Neurol.* 54, 393–398. doi:10.1002/ana.10680

Ge, G., Wang, A., Yang, J., Chen, Y., Yang, J., Li, Y., et al. (2016). Interleukin-37 Suppresses Tumor Growth through Inhibition of Angiogenesis in Non-small Cell Lung Cancer. *J. Exp. Clin. Cancer Res.* 35, 13. doi:10.1186/s13046-016-0293-3

Gegg, M. E., Burke, D., Heales, S. J. R., Cooper, J. M., Hardy, J., Wood, N. W., et al. (2012). Glucocerebrosidase Deficiency in Substantia Nigra of Parkinson Disease Brains. *Ann. Neurol.* 72, 455–463. doi:10.1002/ana.23614

Gilad, Y., Rifkin, S. A., and Pritchard, J. K. (2008). Revealing the Architecture of Gene Regulation: the Promise of eQTL Studies. *Trends Genet.* 24 (8), 408–415. doi:10.1016/j.tig.2008.06.001

GTEx Consortium (2020). The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *Science* 369 (6509), 1318–1330. doi:10.1126/science.aaz1776

Gu, S., Cui, F., Yin, J., Fang, C., and Liu, L. (2021). Altered mRNA Expression Levels of Autophagy- and Apoptosis-Related Genes in the FOXO Pathway in Schizophrenia Patients Treated with Olanzapine. *Neurosci. Lett.* 746, 135669. doi:10.1016/j.neulet.2021.135669

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., et al. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* 99 (6), 1245–1260. doi:10.1016/j.ajhg.2016.10.003

Huang, C.-Y., Hsu, L.-H., Chen, C.-Y., Chang, G.-C., Chang, H.-W., Hung, Y.-M., et al. (2020). Inhibition of Alternative Cancer Cell Metabolism of EGFR Mutated Non-small Cell Lung Cancer Serves as a Potential Therapeutic Strategy. *Cancers* 12 (1), 181. doi:10.3390/cancers12010181

Katrinli, S., Lori, A., Kilaru, V., Carter, S., Powers, A., Gillespie, C. F., et al. (2019). Association of HLA Locus Alleles with Posttraumatic Stress Disorder. *Brain Behav. Immun.* 81, 655–658. doi:10.1016/j.bbi.2019.07.016

Kawai, T., Forrester, S. J., O'Brien, S., Baggett, A., Rizzo, V., and Eguchi, S. (2017). AT1 Receptor Signaling Pathways in the Cardiovascular System. *Pharmacol. Res.* 125, 4–13. doi:10.1016/j.phrs.2017.05.008

Kubo, H., Hoshi, M., Mouri, A., Tashita, C., Yamamoto, Y., Nabeshima, T., et al. (2017). Absence of Kynurenine 3-monooxygenase Reduces Mortality of Acute Viral Myocarditis in Mice. *Immunol. Lett.* 181, 94–100. doi:10.1016/j.imlet.2016.11.012

Li, D., Parks, S. B., Kushner, J. D., Nauman, D., Burgess, D., Ludwigsen, S., et al. (2006). Mutations of Presenilin Genes in Dilated Cardiomyopathy and Heart Failure. *Am. J. Hum. Genet.* 79, 1030–1039. doi:10.1086/509900

Li, S., Lu, Q., and Cui, Y. (2010). A Systems Biology Approach for Identifying Novel Pathway Regulators in eQTL Mapping. *J. Biopharm. Stat.* 20 (2), 373–400. doi:10.1080/10543400903572803

Liu, G., Yao, L., Liu, J., Jiang, Y., Ma, G., Chen, Z., et al. (2014). Cardiovascular Disease Contributes to Alzheimer's Disease: Evidence from Large-Scale Genome-wide Association Studies. *Neurobiol. Aging* 35 (4), 786–792. doi:10.1016/j.neurobiolaging.2013.10.084

Lubbers, E. R., and Mohler, P. J. (2016). Roles and Regulation of Protein Phosphatase 2A (PP2A) in the Heart. *J. Mol. Cell. Cardiol.* 101, 127–133. doi:10.1016/j.yjmcc.2016.11.003

Maack, C., Kartes, T., Kilter, H., Schäfers, H. J., Nickenig, G., Böhm, M., et al. (2003). Oxygen Free Radical Release in Human Failing Myocardium Is Associated with Increased Activity of Rac1-GTPase and Represents a Target for Statin Treatment. *Circulation* 108 (13), 1567–1574. doi:10.1161/01.cir.0000091084.46500.bb

McFadden, S. A. (1996). Phenotypic Variation in Xenobiotic Metabolism and Adverse Environmental Response: Focus on Sulfur-dependent Detoxification Pathways. *Toxicology* 111 (1–3), 43–65. doi:10.1016/0300-483X(96)03392-6

Pan, B., Shen, J., Cao, J., Zhou, Y., Shang, L., Jin, S., et al. (2015). Interleukin-17 Promotes Angiogenesis by Stimulating VEGF Production of Cancer Cells via the STAT3/GIV Signaling Pathway in Non-small-cell Lung Cancer. *Sci. Rep.* 5, 16053. doi:10.1038/srep16053

Qiu, C., Winblad, B., Marengoni, A., Klarin, I., Fastbom, J., and Fratiglioni, L. (2006). Heart Failure and Risk of Dementia and Alzheimer Disease: A Population-Based Cohort Study. *Arch. Intern. Med.* 166 (9), 1003–1008. doi:10.1001/archinte.166.9.1003

Sadigh-Eteghad, S., Sabermarouf, B., Majdi, A., Talebi, M., Farhoudi, M., and Mahmoudi, J. (2015). Amyloid-beta: a Crucial Factor in Alzheimer's Disease. *Med. Princ Pract.* 24 (1), 1–10. doi:10.1159/000369101

Seshasubramanian, V., Raghavan, V., SathishKannan, A. D., Naganathan, C., Ramachandran, A., Arasu, P., et al. (2020). Association of HLA-A, -B, -C, -DRB1 and -DQB1 Alleles at Amino Acid Level in Individuals with Schizophrenia: A Study from South India. *Int. J. Immunogenet.* 47 (6), 501–511. doi:10.1111/iji.12507

Sieberts, S. K., Perumal, T. M., Carrasquillo, M. M., Allen, M., Reddy, J. S., Hoffman, G. E., et al. (2020). Large eQTL Meta-Analysis Reveals Differing Patterns between Cerebral Cortical and Cerebellar Brain Regions. *Sci. Data* 7, 340. doi:10.1038/s41597-020-00642-8

Xu, T., Jin, P., and Qin, Z. S. (2020). Regulatory Annotation of Genomic Intervals Based on Tissue-specific Expression QTLs. *Bioinformatics* 36 (3), 690–697. doi:10.1093/bioinformatics/btz669

Zhang, H., Yue, Y., Sun, T., Wu, X., and Xiong, S. (2017). Transmissible Endoplasmic Reticulum Stress from Myocardiocytes to Macrophages Is Pivotal for the Pathogenesis of CVB3-Induced Viral Myocarditis. *Sci. Rep.* 7, 42162. doi:10.1038/srep42162

Zhao, T., Hu, Y., Zang, T., and Wang, Y. (2019). Integrate GWAS, eQTL, and mQTL Data to Identify Alzheimer's Disease-Related Genes. *Front. Genet.* 10, 1021. doi:10.3389/fgene.2019.01021

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of Summary Data from GWAS and eQTL Studies Predicts Complex Trait Gene Targets. *Nat. Genet.* 48, 481–487. doi:10.1038/ng.3538

Zwilling, D., Huang, S.-Y., Sathyasaikumar, K. V., Notarangelo, F. M., Guidetti, P., Wu, H.-Q., et al. (2011). Kynurenine 3-Monooxygenase Inhibition in Blood Ameliorates Neurodegeneration. *Cell* 145 (6), 863–874. doi:10.1016/j.cell.2011.05.020

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 18 November 2021
doi: 10.3389/frai.2021.757780

DeepCarc: Deep Learning-Powered Carcinogenicity Prediction Using Model-Level Representation

Ting Li[1,2], Weida Tong[1], Ruth Roberts[3,4], Zhichao Liu[1] and Shraddha Thakkar[5]**

[1]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, United States, [2]University of Arkansas at Little Rock and University of Arkansas for Medical Sciences Joint Bioinformatics Program, Little Rock, AR, United States, [3]ApconiX Ltd., Alderley Edge, United Kingdom, [4]Department of Biosciences, University of Birmingham, Birmingham, United Kingdom, [5]Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, United States*

Carcinogenicity testing plays an essential role in identifying carcinogens in environmental chemistry and drug development. However, it is a time-consuming and label-intensive process to evaluate the carcinogenic potency with conventional 2-years rodent animal studies. Thus, there is an urgent need for alternative approaches to providing reliable and robust assessments on carcinogenicity. In this study, we proposed a DeepCarc model to predict carcinogenicity for small molecules using deep learning-based model-level representations. The DeepCarc Model was developed using a data set of 692 compounds and evaluated on a test set containing 171 compounds in the National Center for Toxicological Research liver cancer database (NCTRlcdb). As a result, the proposed DeepCarc model yielded a Matthews correlation coefficient (MCC) of 0.432 for the test set, outperforming four advanced deep learning (DL) powered quantitative structure-activity relationship (QSAR) models with an average improvement rate of 37%. Furthermore, the DeepCarc model was also employed to screen the carcinogenicity potential of the compounds from both DrugBank and Tox21. Altogether, the proposed DeepCarc model could serve as an early detection tool (https://github.com/TingLi2016/DeepCarc) for carcinogenicity assessment.

Keywords: carcinogenicity, deep learning, QSAR, non-animal models, NCTRlcdb

OPEN ACCESS

Edited by:
Inimary Toby,
University of Dallas, United States

Reviewed by:
Ehsan Ullah,
Qatar Computing Research Institute, Qatar
Shailesh Tripathi,
Tampere University of Technology, Finland

****Correspondence:***
Zhichao Liu
Zhichao.Liu@fda.hhs.gov
Shraddha Thakkar
Shraddha.Thakkar@fda.hhs.gov

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *12 August 2021*
Accepted: *27 October 2021*
Published: *18 November 2021*

Citation:
Li T, Tong W, Roberts R, Liu Z and Thakkar S (2021) DeepCarc: Deep Learning-Powered Carcinogenicity Prediction Using Model-Level Representation. Front. Artif. Intell. 4:757780. doi: 10.3389/frai.2021.757780

INTRODUCTION

It is crucial to assess the carcinogenic potency for chemicals, an important factor that triggers regulatory actions for both new and existing chemicals. In 1995, the ICH′ Guideline on the Need for Carcinogenicity studies of Pharmaceuticals was introduced and outlined the need, study design, and interpretation for carcinogenicity studies. Essentially, since carcinogenicity studies are time-consuming and resource-intensive, they should only be performed when human exposure warrants the need for information from lifetime studies in animals to assess carcinogenic potential (ICHS1A, 1995) (Guideline, 1996). Generally, the experimental approach requires a long-term carcinogenicity study (104 weeks) in the rodent plus one other study that supplements the main study (ICHS1B, 1997) (Guideline, 1998), which can be a second-long term study or a shorter study (29 weeks) in a second species. This more concise study could use a transgenic mouse bioassay or a model based on initiation-promotion (ICHS1B, 1997) (Guideline, 1998).

Irrespective of the choices around carcinogenicity studies, each of these studies, on average, requires ~500 rodents and costs around $1.1 m. Moreover, there is evidence of flawed extrapolation for carcinogenicity. There have been many endeavors to address this issue, such as developing biomarkers for use in shorter-term studies as predictors of outcome (Yamamoto et al., 1998; Venkatachalam et al., 2001; Morton et al., 2002). However, these approaches still rely heavily on experimental animals and do not address the 3Rs (replacement, reduction, and refinement of animals in toxicology testing). Programs such as Horizon 2020, The Seventh Framework Programme 7 (FP7), Tox21, Horizon 2020 Precision Toxicology, and other public-private partnerships (Vinken et al., 2021) have offered innovative thinking on developing animal-free methodologies and offer improved translation to humans. These new approach methodologies combine in silico and *in vitro* approaches such as read-across (Shah et al., 2016), toxicogenomics (Yauk et al., 2020), and adverse outcome pathways (AOPs) (Yang et al., 2020).

Several studies have investigated the prediction of carcinogenic potency (Lee et al., 2003; Morales et al., 2006; Tanabe et al., 2010; Caiment et al., 2014; Toropova and Toropov, 2018). The use of the quantitative structure-activity relationship (QSAR) model has become increasingly important for risk assessment because it can provide a fast and economic evaluation of the toxicity of a molecule using only the chemical structure. Some of the QSAR models were developed for carcinogenicity assessment for particular chemical classes (i.e., aromatic amines, food-relevant phytochemicals, polycyclic aromatic hydrocarbon) (Franke et al., 2001; Benigni and Passerini, 2002; Franke et al., 2010; Glück et al., 2018; Li et al., 2019). Although the predictions of these models can vary with interpretation, the application of these models was limited to specific domains. Models for non-congeneric chemicals include various classes of chemicals, which are of great interest for regulatory use (Fjodorova et al., 2010; Zhang et al., 2016a; Zhang et al., 2017; Wang et al., 2020). For example, Zhang et al. (2016b) built a naïve Bayes classifier on 1,042 compounds with rat carcinogenicity and yielded an overall accuracy of 0.90 ± 0.008 and 0.68 ± 0.019 for the training set and external test set, respectively. Zhang et al. (2017) developed an ensemble XGBoost model using 1,003 compounds with rat carcinogenicity and reported an accuracy of 0.7, sensitivity of 0.65, and specificity of 0.77 in external validation. Wang et al. (2020) constructed a novel sparse data deep learning (DL) tool based on the 1003 compounds from Zhang's study (Zhang et al., 2017) and yielded an accuracy of 0.85, sensitivity of 0.82, and specificity of 0.88. These models covered a wide range of chemical classes. However, the annotation of carcinogenicity was only based on the rat in these studies. Since the animal carcinogenicity assessment was required to be conducted at least on two rodent species, it would give a more robust annotation by combining the carcinogenicity signal from both rats and mice. Therefore, we used the National Center for Toxicological Research liver cancer database (NCTRlcdb) (Young et al., 2004), which compressed the carcinogenicity information from both genders of rats and mice.

Deep learning (DL) has been successfully applied to predict complex endpoints, such as drug-induced liver injury (DILI) (Hwang et al., 2020; Li et al., 2020; Semenova et al., 2020) and cardiovascular toxicity (Wang et al., 2017; Maher et al., 2020; Rashed-Al-Mahfuz et al., 2021; Zeleznik et al., 2021). We proposed the DeepDILI model to incorporate model-level representations produced by five different machine learning algorithms into a neural network framework for DILI prediction (Li et al., 2021). The proposed DeepDILI outperformed the publicly available chemical-based DILI prediction models developed from different machine learning (ML) algorithms. However, the DeepDILI study only applied one arbitrary strategy for base classifier selection. The more sophisticated and automatic base classifier selection strategies that should be implemented may further improve the DeepDILI model architecture for other toxicity assessments.

In this paper, we proposed a DeepCarc model to predict carcinogenicity for small molecules using DL based model-level representations. The carcinogenicity annotation was obtained from the NCTRlcdb, incorporating the carcinogenicity signals from both rats and mice. In addition to the previous arbitrary base classifier selection strategy, we also explored a new strategy to select robust base classifiers based on the training set and development set performance. The developed DeepCarc model was comprehensively compared with the optimized 5 ML classifiers, two state-of-the-art ensemble classifiers, and four DL models. In addition, we also employed the DeepCarc model in prioritizing chemicals for carcinogenic potency in the DrugBank and Tox21 chemical databases.

MATERIALS AND METHODS

Data Preparation

To curate a list of compounds for DeepCarc model development, we employed the NCTRlcdb with liver-specific carcinogenicity (Young et al., 2004). The NCTRlcdb provided a single carcinogenicity call per compound, summarizing multiple records representing each gender, species, route of administration, and organ-specific toxicity from the Carcinogenic Potency Database (CPDB) (Gold et al., 1999). Additionally, NCTRlcdb removed inorganic compounds, mixtures, and organometallics from the CPDB to facilitate QSAR model development. In total, NCTRlcdb contained 999 compounds with seven carcinogenicity categories. We excluded compounds from four categories without clear carcinogenicity information, including associated, probable, equivocal, and no opinion. We only employed the compounds from the other three categories, including cancer-liver, cancer-other and negative. The compounds from cancer liver and cancer-other were considered as carcinogens, while compounds from negative were classified as non-carcinogens. More specifically, the non-carcinogens were the compounds without carcinogenic potency observed during

reasonably thorough, chronic long-term tests (Gold et al., 1991). Duplicate compounds were removed by comparing their InChI keys. The final data set consisted of 863 compounds, of which 561 were carcinogens and 302 were non-carcinogens (**Supplementary Table S1**).

To assign the chemical structures uniformly and avoid potential data bias, we applied the Kennard-Stone (KS) (Kennard and Stone, 1969) algorithm to split the whole data set (i.e., 863 compounds) into the training set, development set, and test set. Consequently, the training set included 554 compounds (360 carcinogens/194 non-carcinogens), the development set contained 138 compounds (90 carcinogens/48 non-carcinogens), and the test set consisted of 171 compounds (111 carcinogens/60 non-carcinogens). The structure description file (SDF) of compounds was downloaded from PubChem (https://pubchem.ncbi.nlm.nih.gov/pc_fetch/pc_fetch.cgi) for molecular descriptor calculation (Kim et al., 2021).

Chemical Representation

Three different types of descriptors were calculated for each compound: Mol2vec (Jaeger et al., 2018), Mold2 (Hong et al., 2008), and Molecular ACCess System (MACCS) (Durant et al., 2002) structural keys.

Mol2vec is an unsupervised ML approach trained on a corpus containing 19.9 million compounds to learn vector representations of molecular substructures (Jaeger et al., 2018). For chemical-related substructures, their vector representations point to similar directions in the high dimensional space. Compounds can be represented as vectors that add up from the vectors of the individual substructures. 300-dimensional vector representations were constructed for all compounds.

Mold2 (https://www.fda.gov/science-research/bioinformatics-tools/mold2) is a publicly available software for calculating 777 chemical-physical based 1D/2D descriptors from chemical structure (Hong et al., 2008). The Mold2 software enables a rapid calculation of these large and diverse descriptors. Compared with commercial software packages (Hong et al., 2008), it requires low computing resources to generate the Mold2 descriptors, which contain a similar amount of information.

MACCS is a substructure of keys-based fingerprints encoded as SMART patterns (Durant et al., 2002). Two versions are available, one with 960 structural keys and the other with 166 structure keys. The shorter one is more popular as it can be calculated by several software packages and includes most of the chemical features for drug discovery and virtual screening. A single binary bit value of the bit string indicates the presence or absence of a substructure in the compound.

Two steps of descriptor preprocessing were applied to these three chemical representations. First, we removed the descriptors with zero variance. Secondly, we only kept one descriptor if two descriptors had a pairwise correlation coefficient of more than 0.9. Consequently, 297 of 300 Mol2vec descriptors, 330 of 777 Mold2 descriptors, and 138 of 166 MACCS descriptors were kept for model development (**Supplementary Table S2**).

Discrimination Ability of Chemical Representations

To investigate whether the three chemical representations have a discrimination ability to distinguish between carcinogens and non-carcinogens, we calculated the pairwise compound similarity within carcinogens and non-carcinogens in training and development sets, respectively. We applied the Tanimoto coefficient to calculate the degree of similarity of any two compounds, as it is an appropriate choice for similarity calculation (Willett, 2006; Bajusz et al., 2015). All three chemical representations, Mol2vec, Mold2, and MACCS, were used to calculate the similarity. The Tanimoto coefficient $S_{A,B}$ of molecules A and B is calculated by **Eq. 1** for the continuous variables (e.g., Mol2vec and Mold2) and **Eq. 2** for dichotomous variables (e.g., MACCS).

$$S_{A,B} = \frac{\sum_{j=1}^{n} X_{jA} X_{jB}}{\sum_{j=1}^{n} \left(X_{jA}\right)^2 + \sum_{j=1}^{n} \left(X_{jB}\right)^2 - \sum_{j=1}^{n} X_{jA} X_{jB}} \quad (1)$$

$$S_{A,B} = \frac{c}{a + b - c} \quad (2)$$

Where X_{jA} is the value of the j th feature in molecule A, X_{jB} is the value of the j th feature in molecule B, a is the number of bits with value 1 in molecule A, b is the number of bits with value 1 in molecule B, and c is the number of bits with value 1 in both molecule A and B.

DeepCarc Model Development

DeepCarc model employed the same model architecture as DeepDILI (Li et al., 2021) by implementing a novel base classifier selection strategy (**Figure 1**). The input of NN is the probabilities output of the base classifiers (model-level representation). We hypothesized that no single learning algorithm could fit any modeling circumstance while different algorithms may provide complementary information. Therefore, the ensemble classifiers' performance can improve to some extent.

Base Classifier Development

Base classifiers were developed from five algorithms, including KNN, LR, SVM, RF, and XGBoost. The description of these five algorithms is as previously described (Cox, 1958; Cortes and Vapnik, 1995; Guo et al., 2003; Svetnik et al., 2003; Chen and Guestrin, 2016; Li et al., 2021). Comprehensive hyperparameter optimization was conducted for every algorithm using a bootstrap aggregating strategy (Breiman, 1996) (**Supplementary Table S3**). Specifically, 100 base classifiers were developed for each hyperparameter combination with randomly selected compounds from the training set (80%) and then validated on the development set. The best hyperparameter combination was obtained when the 100 base classifiers achieved the highest average Matthews correlation coefficient (MCC).

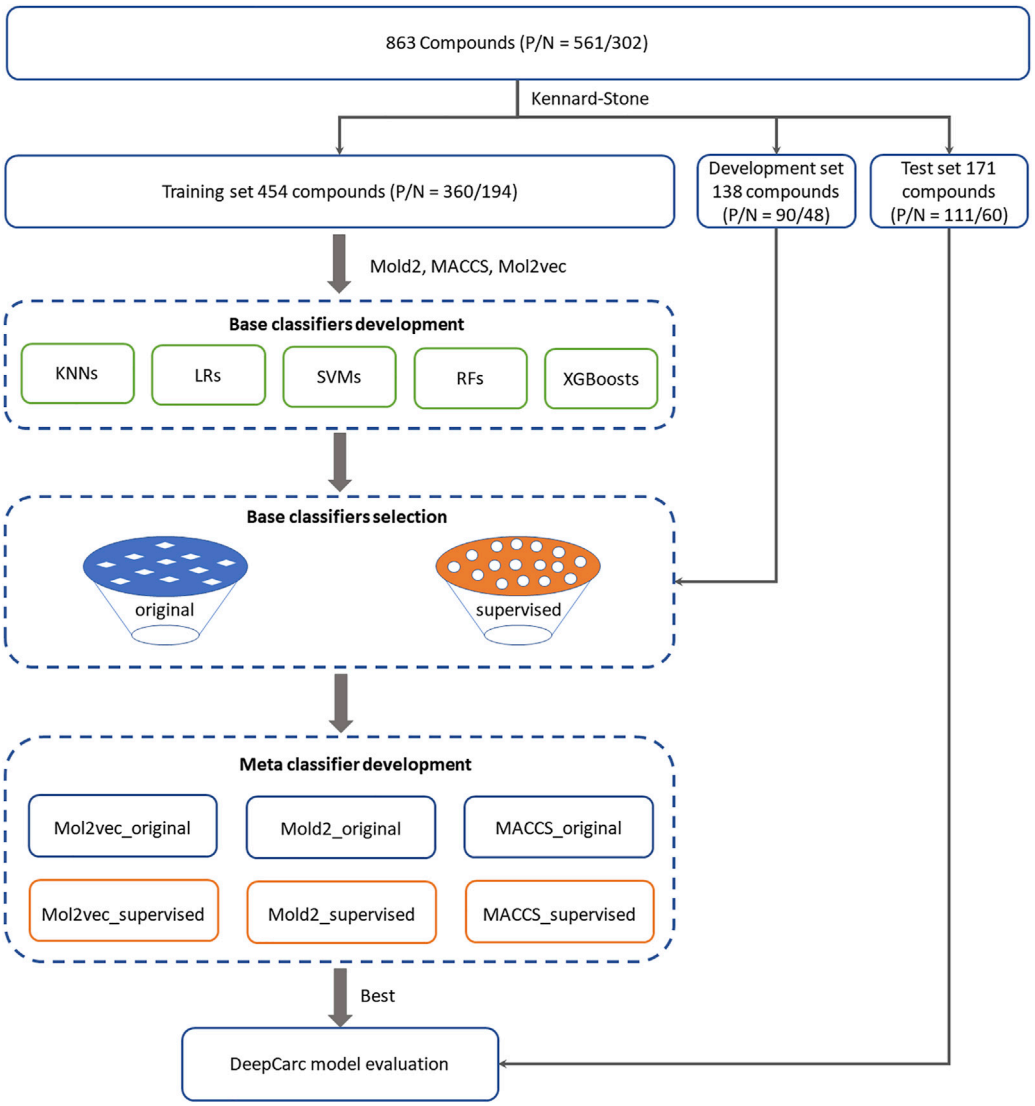


FIGURE 1 | Overall workflow for the DeepCarc model including: (1) Data preparation. 863 compounds were split into training (554 compounds), development (138 compounds), and test (171 compounds) sets based on the Kennard-stone algorithm. (2) Base classifiers development. Five algorithms were used to develop the base classifiers from three different chemical representations, including Mol2vec, Mold2, and MACCS. Two base classifiers selection strategies were employed to select the optimized classifiers for meta classifier development. (3) Meta classifier development. With three chemical representations and two selection methods, six groups of base classifiers, including Mol2vec_supervised, Mol2vec_original, Mold2_supervised, were used Mold2_original, MACCS_supervised, and MACCS_original. The probability prediction from selected base classifiers was used to train the neural network. (4) Model evaluation. The DeepCarc model was evaluated on the independent test set.

Two base classifier selection strategies were proposed, named original strategy and supervised strategy:

1) The original strategy was the base classifier selection approach used in the DeepDILI model. Specifically, 100 classifiers generated by each of the five algorithms with the best hyperparameters were rank-ordered based on MCC values. Only the ones with their MCC in the range of 5–95% percentile were chosen as optimized base classifiers for the meta-classifier development.
2) In the supervised strategy, we developed 1,000 base classifiers for each algorithm with the best hyperparameter combination from the training set. For each algorithm, the performance of every base classifier and the average performance of these 1,000 models was evaluated on both the training set and development set. Only the base classifiers with MCC values higher than the average MCC of both the training set and the development set were selected as the optimized base classifiers. Then, the optimized base classifiers selected from the five algorithms were combined for the meta-classifier development.

Meta-Classifier Development

The meta-classifier NN aims to find the underlying relationship that transfers the optimized base classifiers' information to target through linear or non-linear mathematical expression. In this study, a three-layer NN was developed as the meta-classifier for carcinogenicity prediction. Specifically, the input of NN came from the probabilities output of the optimized base classifiers (model-level representation) on the development set, which means a compound was represented by a vector of probabilities output from the optimized base classifiers. The hidden layer included 10 nodes with rectified linear unit (Relu) activation, stochastic gradient descent optimization, batch normalization, and a dropout of 0.5. The output layer used the sigmoid function to project the hidden layer information to probabilistic values of carcinogenicity prediction. The meta-classifier method was employed to develop six DeepCarc candidate models from the combination of three chemical representations (Mol2vec, Mold2, and MACCS) and two base classifiers selection strategies (original and supervised). For example, the candidate DeepCarc model of Mol2vec_original indicates the base classifiers were developed with the chemical representation of Mol2vec and filtered by the original base classifier selection method.

DeepCarc Model Evaluation

The developed DeepCarc model performance was evaluated in the test set, including 171 compounds (111 carcinogens/60 non-carcinogens). The DeepCarc model was assessed by six performance metrics, including MCC, F1, accuracy, balanced accuracy (BA), sensitivity, and specificity, which were calculated using the following equations.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{3}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{4}$$

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{5}$$

$$BA = \frac{sensitivity + specificity}{2} \tag{6}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$specificity = \frac{TN}{TN + FP} \tag{8}$$

The TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. In addition, the area under the receiver operating characteristic (ROC) curve (AUC) was also computed, where the ROC curve presents the performance of the classification model by measuring the relationship between true positive rate (TPR) against false positive rate (FPR) (Fawcett, 2006).

To investigate whether the probabilistic values yielded by DeepCarc could prioritize the compounds regarding carcinogenetic potential, we employed the Chi-Square test in different probabilistic thresholds (i.e., probabilistic value cut-off values were from 0.1 to 0.9 with a step of 0.1). Meanwhile, we calculated the positive predictive value (PPV) and negative predictive value (NPV) to investigate the discrimination power of probabilistic values for true positive and true negatives carcinogens, as shown in the following formulas:

$$PPV = \frac{TP}{TP + FP} \tag{9}$$

$$NPV = \frac{TN}{TN + FN} \tag{10}$$

Comparative Analysis With Other Modeling Approaches

To further evaluate the proposed DeepCarc model, we compared DeepCarc with the optimized base classifiers developed from five algorithms, including KNN, LR, SVM, RF, and XGBoost. Furthermore, two ensemble methods, including the majority voting and average probability methods, were employed to justify the extra value of the proposed DeepCarc model over the conventional ensemble approaches. In the majority voting method, a consensus call of carcinogen/non-carcinogen was derived by the majority calls of the optimized base classifiers. In the average probability method, a new call was given to the non-carcinogen if the average probability of the optimized base classifiers was <0.5 and vice versa.

In addition, we compared the DeepCarc model against four other molecular-based DL models, including Text Convolutional neural network (CNN) from DeepChem (DC-TEXTCNN) (Wu et al., 2018), Chemistry Chainer-Neural Fingerprint (CH-NFP) (Duvenaud et al., 2015), Edge Attention-based Multi-relational Graph Convolutional Networks (EAGCNG) (Shang et al., 2018), and Convolutional Neural Network Fingerprint (CNF) (Tetko et al., 2019). The DC-TEXTCNN implemented the TEXTCNN based on chemical information, where the TEXTCNN was constructed to classify sentence tasks based on word representations. In the DC-TEXTCNN, the Simplified Molecular Input Line Entry System (SMILES) strings of molecules are the "sentence" input with the characters of the string represented as vectors. In the CH-NFP, the neural fingerprints are extracted from graphs of molecules and forwarded to a multilayer perceptron to make a classification prediction. The EAGCNG learns node features and attention weights in a graph convolutional network, where a molecular graph is represented by a real-valued attention matrix instead of a binary adjacency matrix. The CNF improves the molecule prediction by combining the synergy effect between CNN and the multiplicity of SMILES, which is used for feature extraction and data augmentation, respectively. These four DL models were developed from the Online Chemical Modeling Environment (OCHEM) website (https://ochem.eu/home/show.do). We used our training set and development set together to develop the models and then evaluated them on the independent test set.

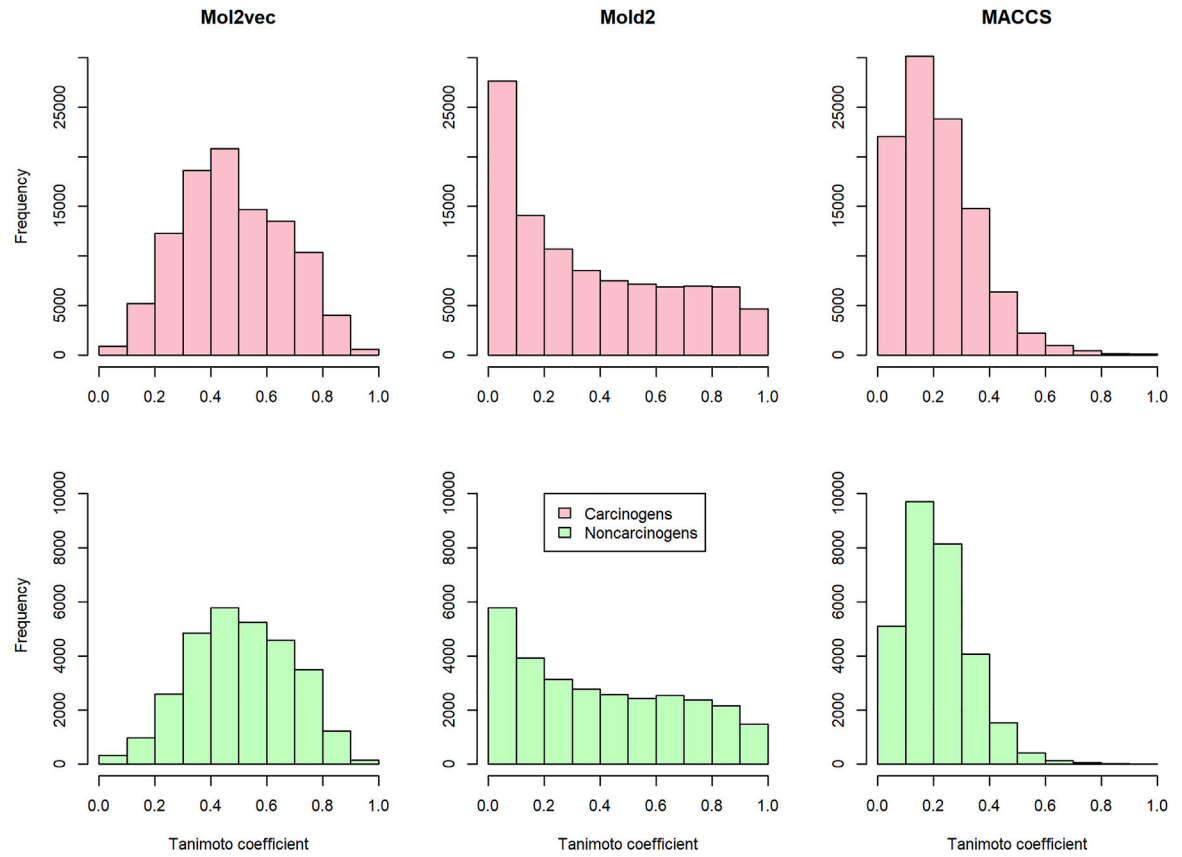


FIGURE 2 | The distribution of the pairwise Tanimoto coefficients calculated from Mol2vec, Mold2, and MACCS: The pink and green indicate that the pairwise Tanimoto coefficients were calculated from the carcinogenic molecules and noncarcinogenic molecules, respectively.

DeepCarc for Screening Carcinogenicity Potential of Compounds

The developed DeepCarc model was used as a screening tool for carcinogenicity risk detection in two external datasets, including DrugBank and Tox21. First, we collected 10,741 compounds from DrugBank database version 5.1.7 (Wishart et al., 2018), including approved and investigational drugs. After removing organometallics, heavy molecules, and the overlap compounds with our NCTRlcdb datasets, 9,814 investigated and approved drugs were kept (**Supplementary Table S4**). The output of predicted probabilistic values from the DeepCarc model was used to measure the carcinogenicity concern quantitatively. Second, we collected 8,410 compounds from the U.S. Tox21 program https://tripod.nih.gov/pub/tox21/, including food-additives, household cleaning products, medicines, and environmental hazard chemicals. The selection criteria of DrugBank were employed in the Tox21 dataset, and 7176 compounds were kept for screening by the DeepCarc model (**Supplementary Table S5**). We used the output of predicted probabilistic values from the DeepCarc model to quantitatively measure the carcinogenicity concern.

Code Availability

All the models introduced above were developed with the open-source Python (version 3.6.5). The Mol2vec descriptors were generated from the source code https://github.com/samoturk/mol2vec. The open-source cheminformatics toolkit RDKit37 (version: 2020.09.1) was employed to construct the MACCS fingerprints. The Keras library version 2.0 with TensorFlow version 1.14 as the backend was used to develop NN classifiers. The scikit-learn package version 0.22 (Pedregosa et al., 2011) was applied to develop models with these four algorithms of KNN, LR, SVM, and RF. The open-source XGBoost library implemented on Python (version 3.6.5) was used to build all the XGBoost models. The scripts of all the models in this study are available at https://github.com/TingLi2016/DeepCarc.

RESULTS

Discrimination Power of Chemical Representations

To investigate the discrimination power of different chemical representations, we calculated the pairwise compound similarity (i.e., Tanimoto coefficients) among the compounds belonging to carcinogens (i.e., 450 compounds in training and development set) and non-carcinogens (i.e., 242 compounds in training and development set) with each chemical representation, respectively

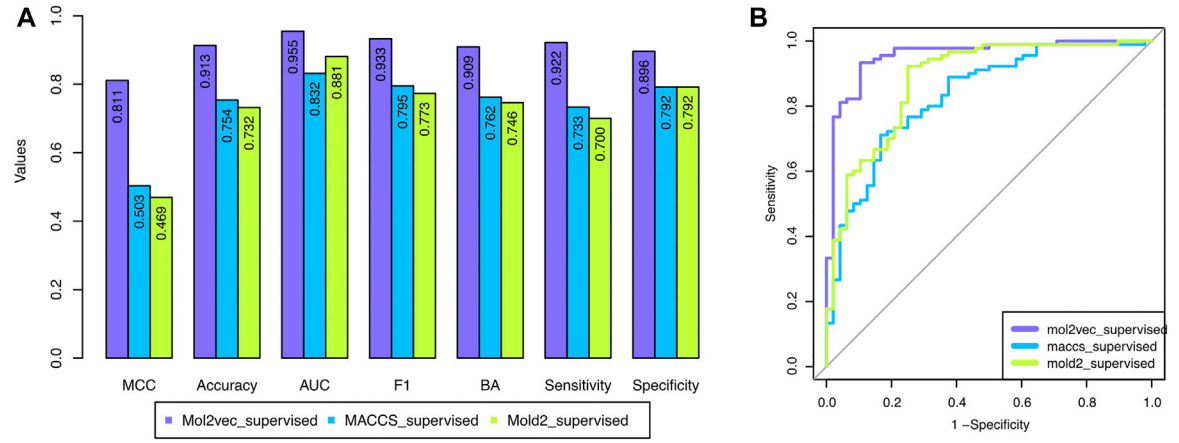


FIGURE 3 | The performance of the developed DeepCarc models based on the proposed supervised base classifier selection strategy with the three chemical representations: the three chemical representations included Mol2vec, Mold2, and MACCS. **(A)**: Seven performance metrics; **(B)**: Area under the ROC curve.

(**Figure 2**). Within each chemical representation (e.g., Mol2vec, Mold2, or MACCS), we observed a similar distribution of Tanimoto coefficients for carcinogens and non-carcinogens. For example, the average and standard deviation of Tanimoto coefficients were 0.479 ± 0.187 and 0.505 ± 0.182 for carcinogens and non-carcinogens based on Mol2vec chemical representation. Furthermore, the average and standard deviations of Tanimoto coefficients derived from Mold2 were 0.356 ± 0.297 and 0.401 ± 0.292 for carcinogens and non-carcinogens, whereas for MACCS they were 0.217 ± 0.143 and 0.214 ± 0.123. The Mol2vec tended to generate higher Tanimoto coefficients than Mold2 or MACCS, suggesting higher discrimination power of Mol2vec to cluster the compounds from the same category (i.e., carcinogens and non-carcinogens).

Mol2vec With Supervised Selection Outperformed Other Combinations

To overcome the shortcoming of the base classifier selection strategy, we proposed a supervised classifier selection strategy by considering the performance from both training and development sets (see *Material and Methods*). **Figure 3** depicted the development set performance using the proposed supervised base classifier selection strategy with the three chemical representations. The developed DeepCarc based on the Mol2vec with the proposed supervised base classifier selection strategy yielded the best performance across all the performance metrics (e.g., MCC = 0.811), which was much higher than that of Mold2 (i.e., MCC = 0.503) and MACCS (i.e., MCC = 0.469). Furthermore, the performance metrics of the DeepCarc model based on the proposed supervised base classifier selection strategy with Mol2vec were also much higher than those of the original strategy across all the performance metrics (**Supplementary Figure S1**). For example, the DeepCarc developed by the Mol2vec and supervised base classifier selection strategy had an improved rate of 18.57% compared to that of the original base classifier selection strategy (e.g., MCC = 0.684). Eventually, The DeepCarc model developed based on Mol2vec with the proposed supervised base classifier selection strategy consists of 296 RF, 285 LR, 277 KNN, 266 XGBoost, and 254 SVM which was considered as the optimized model for the following analysis.

DeepCarc Effectively Augmented the Performance of Selected Base Classifiers

To evaluate whether the DeepCarc model could benefit from complementary information provided by different conventional machine learning algorithms, we compared the optimized DeepCarc model to the selected base classifiers developed from 5 ML algorithms (**Table 1**). For each machine learning algorithm, the average and standard deviation of the seven-performance metrics of the selected base classifiers were calculated for the development set and test set, respectively. The DeepCarc yielded the highest values in all the performance metrics except sensitivity (i.e., MCC = 0.811, accuracy = 0.913, AUC = 0.955, F1 score = 0.933, Balanced accuracy = 0.909, sensitivity = 0.922 and specificity = 0.896) compared to the selected base classifiers. For example, the DeepCarc made approximately an improvement of 77–127% of MCC over the selected base classifiers in the development set. Although the selected base classifiers achieved high sensitivities, they yielded very imbalanced performance regarding sensitivity (e.g., 0.991 ± 0.007 for RF) and specificity (0.212 ± 0.035 for RF). The performance followed the same trend in the test set, where the DeepCarc model achieved the highest value in MCC (0.432), accuracy (0.754), AUC (0.776), F1 (0.828), BA (0.688), and specificity (0.467). For instance, the DeepCarc made approximately 127–184% improvement in MCC over the selected base classifiers. Furthermore, the DeepCarc provided the most balanced performance regarding sensitivity (0.910) and specificity (0.467), whereas the selected base classifiers generated extremely lower specificity. In other words, the selected base classifiers tended to predict all the samples in the test set as carcinogens.

TABLE 1 | The comparison between the base classifiers and DeepCarc performance on the development set and test set.

Data set	Model	MCC	Accuracy	AUC	F1	BA	Sensitivity	Specificity
Development set	DeepCarc	0.811	0.913	0.955	0.933	0.909	0.922	0.896
	XGBoost	0.458 ± 0.027	0.758 ± 0.011	0.785 ± 0.02	0.842 ± 0.006	0.659 ± 0.016	0.986 ± 0.007	0.331 ± 0.034
	LR	0.412 ± 0.024	0.746 ± 0.009	0.772 ± 0.012	0.830 ± 0.007	0.657 ± 0.016	0.95 ± 0.0260	0.364 ± 0.051
	SVM	0.408 ± 0.026	0.737 ± 0.010	0.754 ± 0.021	0.831 ± 0.005	0.626 ± 0.016	0.991 ± 0.012	0.261 ± 0.040
	KNN	0.372 ± 0.029	0.726 ± 0.009	0.694 ± 0.029	0.825 ± 0.005	0.612 ± 0.014	0.987 ± 0.010	0.236 ± 0.032
	RF	0.357 ± 0.032	0.720 ± 0.011	0.805 ± 0.018	0.822 ± 0.006	0.601 ± 0.016	0.991 ± 0.007	0.212 ± 0.035
Test set	DeepCarc	0.432	0.754	0.776	0.828	0.688	0.910	0.467
	XGBoost	0.187 ± 0.039	0.672 ± 0.007	0.715 ± 0.022	0.797 ± 0.004	0.536 ± 0.010	0.991 ± 0.003	0.081 ± 0.021
	LR	0.176 ± 0.033	0.670 ± 0.007	0.663 ± 0.017	0.794 ± 0.004	0.538 ± 0.011	0.981 ± 0.012	0.096 ± 0.028
	SVM	0.152 ± 0.039	0.665 ± 0.007	0.733 ± 0.020	0.793 ± 0.004	0.529 ± 0.009	0.986 ± 0.008	0.071 ± 0.020
	KNN	0.190 ± 0.037	0.672 ± 0.007	0.586 ± 0.031	0.797 ± 0.004	0.534 ± 0.009	0.993 ± 0.005	0.076 ± 0.019
	RF	0.163 ± 0.039	0.665 ± 0.006	0.700 ± 0.027	0.794 ± 0.003	0.524 ± 0.008	0.997 ± 0.004	0.051 ± 0.015

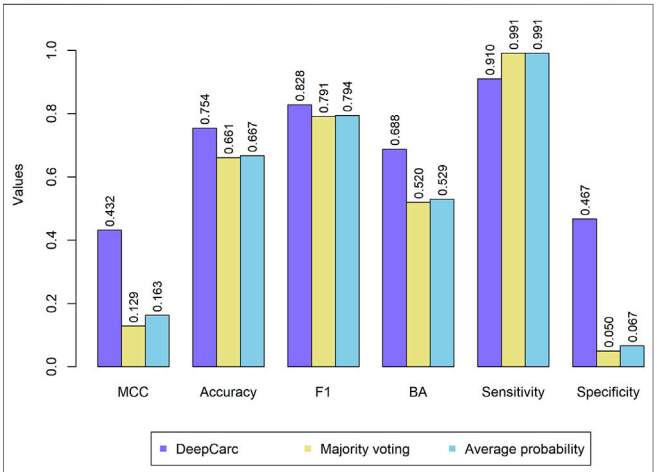


FIGURE 4 | Ensemble models performance on the test set.

DeepCarc Outperformed the State-of-the-Art Ensemble Classifiers

The comparison between DeepCarc and two state-of-the-art ensemble classifiers (i.e., majority voting and average probability) was also conducted on the test set (**Figure 4**). Consequently, the DeepCarc yielded better performance than the other two ensemble classifiers on MCC, accuracy, F1, BA, and specificity with an average improvement of 195.89, 13.55, 4.48, 31.17, and 698.29%, respectively. The majority voting and average probability generated the highest sensitivity (0.991 and 0.991, respectively), but with extremely low specificity (0.050 and 0.067, respectively), suggesting the proposed DeepCarc model could effectively optimize and combine the base classifiers.

DeepCarc With Model-Level Representation Outperformed Molecule Representation-Based Deep Learning Models

To confirm the model-level representation and the molecule-based representation in carcinogenicity prediction, we compared the DeepCarc model with four other publicly available DL models, including DC-TEXTCNN, CH-NFP, EAGCNG, and CNF (**Table 2**). The model performance of these four DL models varied. Among these four deep learning models, DC-TEXTCNN resulted in the highest performance in the MCC of 0.392, accuracy of 0.735, F1 of 0.829, and sensitivity of 0.982. CH-NFP yielded the highest AUC of 0.776 and BA of 0.639, while EAGCNG achieved the highest specificity of 0.400. The imbalanced performance in sensitivity and specificity were also observed in these four deep learning models. DeepCarc outperformed these four deep learning models on MCC, accuracy, AUC, BA, and specificity. For example, DeepCarc improved 10–134% in MCC over the other four deep learning models.

Predicted Probabilistic Values of the DeepCarc Model for Prioritizing Compounds on Their Carcinogenic Risk

To investigate the potential use of the DeepCarc model as the screening tool for prioritizing the carcinogenic risk, we employed the Chi-Square test to examine the correlation between carcinogen

TABLE 2 | The model performance of DeepCarc and four advanced DNN models on the test set.

Models	MCC	Accuracy	AUC	F1	BA	Sensitivity	Specificity
DeepCarc	0.432	0.754	0.776	0.828	0.688	0.910	0.467
DC-TEXTCNN	0.392	0.735	0.719	0.829	0.627	0.982	0.271
CH-NFP	0.353	0.725	0.776	0.814	0.639	0.928	0.350
EAGCNG	0.328	0.713	0.682	0.800	0.641	0.883	0.400
CNF	0.185	0.673	0.636	0.796	0.541	0.982	0.100

TABLE 3 | The relationship between predicted probabilistic values of DeepCarc and carcinogen risk.

Probabilistic threshold	DeepCarc prediction	Carcinogen		*p* Value	Positive predictive value	Negative predictive value
		Positive	Negative			
0.1	Predicted positive	110	56	5.188E-2	0.663	0.800
	Predicted negative	1	4			
0.2	Predicted positive	110	52	1.074E-3	0.679	0.889
	Predicted negative	1	8			
0.3	Predicted positive	110	44	1.51E-07	0.714	0.941
	Predicted negative	1	16			
0.4	Predicted positive	108	40	5.22E-08	0.730	0.870
	Predicted negative	3	20			
0.5	Predicted positive	101	32	4.22E-08	0.759	0.737
	Predicted negative	10	28			
0.6	Predicted positive	89	29	2.74E-05	0.754	0.585
	Predicted negative	22	31			
0.7	Predicted positive	81	22	7.18E-06	0.786	0.559
	Predicted negative	30	38			
0.8	Predicted positive	68	14	2.44E-06	0.829	0.517
	Predicted negative	43	46			
0.9	Predicted positive	47	6	9.85E-06	0.887	0.458
	Predicted negative	64	54			

potential and predicted probabilistic values (**Table 3**). The *p* values yielded from the Chi-Square test were all less than 0.05 in probabilistic threshold from 0.2 to 0.9 with a step of 0.1, showing the strong correlation between the predicted probabilistic values of DeepCarc and the carcinogen risk. Furthermore, with the threshold increased, the PPVs increased from 0.663 to 0.887, meaning 88.7% compounds predicted with probabilistic values greater or equal to 0.9 were carcinogens. Meanwhile, the NPVs decreased as the threshold increased. The NPV yielded the highest value of 0.941 with the classification threshold value of 0.3 on the test set, indicating 94.1% of compounds predicted with a probabilistic value less than 0.3 were non-carcinogens. Altogether, the predicted probabilistic values of the DeepCarc model could be used as the indicators for prioritizing compounds regarding their potential carcinogenic risk.

DeepCarc Is Employed to Screen DrugBank and Tox21 Compounds

The DeepCarc was used as a screening tool for identifying the carcinogenicity potential of the compounds from DrugBank (**Figure 5A**). The predicted probabilistic values ranging from 0 to 1 were split into 10 intervals with a size of 0.1. Of 9,814 compounds, there were 7,410 (i.e., 7410/9814 = 75.50%), 916 (9.33%), 440 (4.48%), 290 (2.95%), 188 (1.92%) compounds with their predicted probabilities belong to the intervals of (0, 0.1), (0.1, 0.2), (0.2, 0.3), (0.3, 0.4), and (0.4, 0.5), respectively, indicating low carcinogenicity concern. In total, 570 compounds (5.81%) were predicted with probabilistic values ≥0.5, indicating compounds with carcinogenicity risk. Of 570 compounds, there were 45 compounds (0.46%) with the predicted probability ≥0.9, indicating high carcinogenicity concern. The predicted probabilistic value of each drug is included in **Supplementary Table S4**.

The DeepCarc further screened the carcinogenicity potential of the compounds from the Tox21 (**Figure 5B**). Similarly, the predicted probabilistic values were separated into 10 intervals. Of the 7,176 compounds, there were 3731 (i.e., 3731/7176 = 51.99%), 787 (10.97%), 504 (7.02%), 442 (6.16%), 371 (5.17%) compounds with their predicted probabilities belong to the intervals of (0, 0.1), (0.1, 0.2), (0.2, 0.3), (0.3, 0.4), and (0.4, 0.5), respectively, indicating low carcinogenicity concern. The other 1341 (18.69%) compounds were predicted with probabilistic values ≥0.5, suggesting the compounds possessed carcinogenicity risk. There were 113 (1.57%) compounds with the predicted probabilistic value ≥0.9, suggesting high carcinogenicity concern (**Supplementary Table S5**).

DISCUSSION

Effectively evaluating the carcinogenicity of compounds is essential to improve the regulatory efficacy and promote public health. Performing a standard toxicity assay with two rodents (rats and mice) is expensive and time-consuming. Only a small proportion of compounds have been tested on carcinogenicity. Therefore, there is an urgent need for developing alternative methods to test carcinogenicity quickly and cost-effectively. A lot of computational models have been developed for prediction of carcinogenic potency. Some of these models can only be applied to specific chemical classes, and some were developed based only on rat's carcinogenicity assay results. We developed a DeepCarc model to fill the gap by combining model-level representation generated from five conventional ML classifiers into a DL framework with Mol2vec descriptor and supervised base classifier selection strategy. The proposed DeepCarc model outperformed the optimized 5 ML classifiers, two state-of-the-art ensemble methods, and four molecule-based deep learning models. The developed DeepCarc model is publicly available through https://github.com/TingLi2016/DeepCarc.

The DeepCarc model was developed from the NCTRlcdb, which includes 863 compounds, and the carcinogenicity

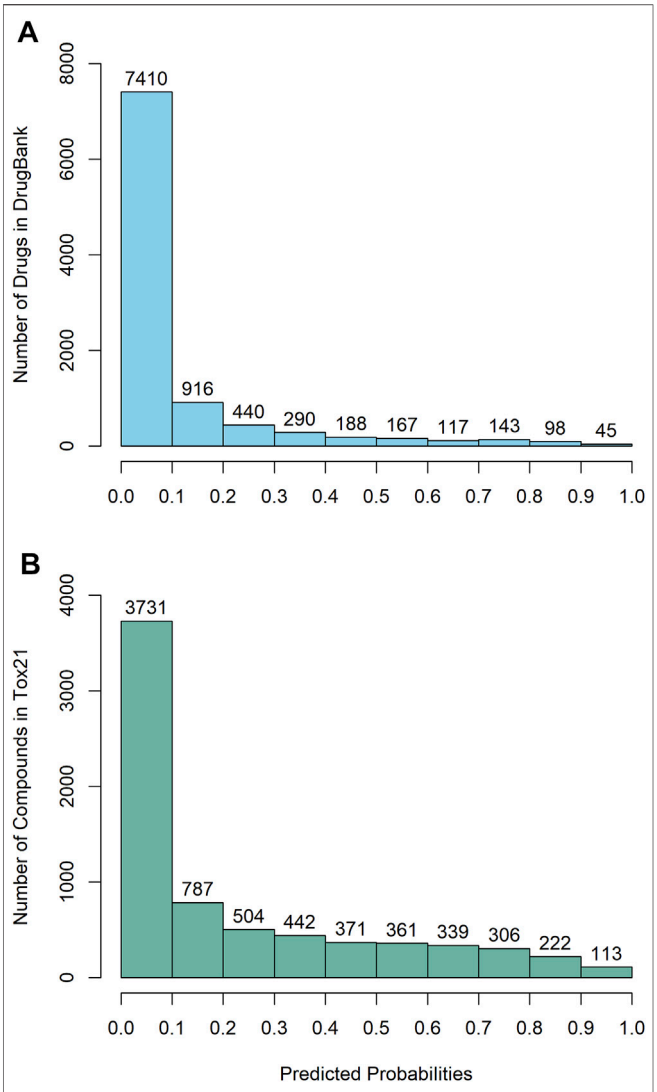


FIGURE 5 | The probability distribution of the DeepCarc prediction of the compounds from **(A)** DrugBank; **(B)** Tox21.

classification was built based on the carcinogenicity results of both rats and mice. The DeepCarc model was designed to predict the general carcinogens, which are non-organ specific. We investigated other reported machine learning-based prediction models with the NCTRlcdb data set (Liu et al., 2011; Tung, 2013; Tung, 2014; Beger et al., 2004). However, all the other reported prediction models aim to discriminate liver-specific carcinogens from others. Furthermore, samples used in these developed models varied from each other. One of the significant challenges of AI-based models towards real-world application is explainability. Here, we employed the Uniform Manifold Approximation and Projection (UMAP) to investigate the driving force of the proposed supervised base classifier selection strategy outperforming the original one (McInnes et al., 2018) (**Supplementary Figure S2**). The UMAP is a non-linear dimension reduction technique that captures the local relationships within the groups and the global relationships between different groups (Becht et al., 2019). We found that the supervised selection method had better discrimination power in distinguishing the carcinogens from non-carcinogens than the original selection method.

The DeepCarc model was compared with the other four DL carcinogenicity prediction models (DC-TEXTCNN, CH-NFP, EAGCNG, and CNF) using the chemical representation as a direct input. Different from the chemical descriptors used in the DeepCarc development, we explored three other different types of chemical representation, including SMILES strings (DC-TEXTCNN, and CNF), molecular graphs (CH-NFP), and molecular graphs with attention (EAGCNG). We also evaluated the impact on carcinogenicity prediction by enlarging the data set with the multiplicity of SMILES strings in the CNF model. DeepCarc outperformed these four DL models with the highest MCC of 0.432. The DC-TEXTCNN and CNF with SMILES strings as input had the highest sensitivity but lowest specificity. The CH-NFP and EAGCNG with the molecular graph as input reached higher specificity than the two DL models (DC-TEXTCNN and CNF) with SMILES string as input. Enlarging the data set by the multiplicity of SMILES string did not improve the performance in this carcinogenicity prediction.

Considering a large proportion of compounds in DrugBank and Tox21 without the carcinogenic test result, we employed the DeepCarc model to assess the carcinogenicity risk for the compounds from DrugBank and Tox21 to provide the information for further prioritizing the compounds for carcinogenicity assessment. We found that 1341 (1341/7176 = 18.69%) compounds were predicted with carcinogenicity risk in Tox21, which is much larger than 570 (570/9814 = 5.81%) drugs predicted with carcinogenicity risk in DrugBank. One of the possible reasons is that Tox21 includes environmental chemicals and household cleaning products, which are less likely to be evaluated by the carcinogenicity bioassay. However, there is a rigorous procedure to avoid carcinogens from getting marketed in drug development. A drug is required to take the 2-years carcinogenicity animal study if it will be used in treatment continuously for 6 months or more or with some special causes for concern, such as belonging to a class of the known carcinogens, showing evidence of precancerous changes in the chronic toxicity studies, and retaining in tissues for a long time (Rang and Hill, 2013). We conducted a literature survey to collect the compounds' carcinogenic potential details with very high and low probabilities. However, we found little information on the carcinogenic testing results of these compounds. For example, Osimertinib was predicted with the carcinogenic probability of 0.928 and a study reported that it induced autophagy and apoptosis via reactive oxygen species generation in non-small cell lung cancer cells (Tang et al., 2017).

To investigate the potential artifact yield in the data split process, we randomly split the total 863 chemicals were into the different training set, development set, and test data set for 10 times to develop DeepCarc models. The low specificity of the test set compared to the development set is consistently observed in every newly developed DeepCarc model (**Supplementary Figure S3**). Identifying compounds with

potential carcinogenic risks is very costly, time-consuming, and labor-intensive. A model with high sensitivity for detecting high carcinogenic risk compounds could be beneficial to narrow down a large number of compounds into a handled scale for further risk assessment. Considering the relatively low specificity and high sensitivity nature of the current DeepCarc model, we highly recommended positioning the model on screening of molecules in the early stage of development.

A low false-negative rate is one of the essential prerequisites to warrant the practical application of the prediction model in screening carcinogens. Therefore, we investigated the false-positives cases in our proposed DeepCarc model. There were 10 of 111 carcinogens predicted as non-carcinogens in the test set. The common structure analysis was employed for these 10 carcinogens. However, we did not find any common substructure, indicating only chemical information is insufficient to identify these carcinogens. Therefore, we recommend applying alternative approaches such as high-throughput *in vitro* toxicity assays (Li et al., 2017; Chiu et al., 2018) to further screen the non-carcinogens predicted by the DeepCarc to eliminate the false-negative cases in the real-world application.

The development of animal-free models is a new trend of modernized toxicity assessment. The 2-years bioassays in rats and mice are impossible to assess the carcinogenic potential of every compound efficiently and accurately. The DeepCarc model we developed could help prioritize potential carcinogens in the early stages of compounds development. Moreover, we hope our work will attract more interest to further exploring advanced artificial intelligence (AI) approaches for carcinogenic potency prediction.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

ZL and WT conceived and designed the study. TL and ZL performed data analysis. TL, ZL, and RR wrote the manuscript. RR, ST, ZL, and WT revised the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

TL is grateful to the National Center for Toxicological Research (NCTR) of the U.S. Food and Drug Administration (FDA) for postdoctoral support through the Oak Ridge Institute for Science and Education (ORISE). RR is grateful to the contract program with NCTR for the support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.757780/full#supplementary-material

REFERENCES

Bajusz, D., Rácz, A., and Héberger, K. (2015). Why Is Tanimoto index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform* 7, 20–13. doi:10.1186/s13321-015-0069-3

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., et al. (2019). Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nat. Biotechnol.* 37, 38–44. doi:10.1038/nbt.4314

Beger, R. D., Young, J. F., and Fang, H. (2004). Discriminant Function Analyses of Liver-specific Carcinogens. *J. Chem. Inf. Comput. Sci.* 44, 1107–1110. doi:10.1021/ci0342829

Benigni, R., and Passerini, L. (2002). Carcinogenicity of the Aromatic Amines: from Structure-Activity Relationships to Mechanisms of Action and Risk Assessment. *Mutat. Research/Reviews Mutat. Res.* 511, 191–206. doi:10.1016/s1383-5742(02)00008-x

Breiman, L. (1996). Bagging Predictors. *Mach Learn.* 24, 123–140. doi:10.1007/bf00058655

Caiment, F., Tsamou, M., Jennen, D., and Kleinjans, J. (2014). Assessing Compound Carcinogenicityin Vitrousing Connectivity Mapping. *Carcin* 35, 201–207. doi:10.1093/carcin/bgt278

Chen, T., and Guestrin, C. (2016). "Xgboost: A Scalable Tree Boosting System," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, San Francisco California USA, 13 August 2016 (IEEE), 785–794.

Chiu, W., Guyton, K. Z., Martin, M. T., Reif, D. M., and Rusyn, I. (2018). Use of High-Throughput *In Vitro* Toxicity Screening Data in Cancer hazard Evaluations by IARC Monograph Working Groups. *Altex* 35, 51–64. doi:10.14573/altex.1703231

Cortes, C., and Vapnik, V. (1995). Support-vector Networks. *Mach Learn.* 20, 273–297. doi:10.1007/bf00994018

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *J. R. Stat. Soc. Ser. B (Methodological)* 20, 215–232. doi:10.1111/j.2517-6161.1958.tb00292.x

Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002). Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* 42, 1273–1280. doi:10.1021/ci010132r

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., et al. (2015). Convolutional Networks on Graphs for Learning Molecular Fingerprints. arXiv preprint arXiv: 1509.09292.

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern recognition Lett.* 27, 861–874. doi:10.1016/j.patrec.2005.10.010

Fjodorova, N., Vračko, M., Tušar, M., Jezierska, A., Novič, M., Kühne, R., et al. (2010). Quantitative and Qualitative Models for Carcinogenicity Prediction for Non-Congeneric Chemicals Using CP ANN Method for Regulatory Uses. *Mol. Divers.* 14, 581–594. doi:10.1007/s11030-009-9190-4

Franke, R., Gruska, A., Bossa, C., and Benigni, R. (2010). QSARs of Aromatic Amines: Identification of Potent Carcinogens. *Mutat. Research/Fundamental Mol. Mech. Mutagenesis* 691, 27–40. doi:10.1016/j.mrfmmm.2010.06.009

Franke, R., Gruska, A., Giuliani, A., and Benigni, R. (2001). Prediction of Rodent Carcinogenicity of Aromatic Amines: A Quantitative Structure-Activity Relationships Model. *Carcinogenesis* 22, 1561–1571. doi:10.1093/carcin/22.9.1561

Glück, J., Buhrke, T., Frenzel, F., Braeuning, A., and Lampen, A. (2018). In Silico genotoxicity and Carcinogenicity Prediction for Food-Relevant Secondary Plant Metabolites. *Food Chem. Toxicol.* 116, 298–306. doi:10.1016/j.fct.2018.04.024

Gold, L. S., Manley, N. B., Slone, T. H., and Rohrbach, L. (1999). Supplement to the Carcinogenic Potency Database (CPDB): Results of Animal Bioassays Published in the General Literature in 1993 to 1994 and by the National Toxicology Program in 1995 to 1996. *Environ. Health Perspect.* 107, 527–600. doi:10.2307/3434550

Gold, L. S., Slone, T. H., Manley, N. B., Garfinkel, G. B., Hudes, E. S., Rohrbach, L., et al. (1991). The Carcinogenic Potency Database: Analyses of 4000 Chronic Animal Cancer Experiments Published in the General Literature and by the U.S. National Cancer Institute/National Toxicology Program. *Environ. Health Perspect.* 96, 11–15. doi:10.1289/ehp.919611

Guideline, I. (1996). "Guideline on the Need for Carcinogenicity Studies of Pharmaceuticals S1A," in International Conference on Harmonization1996.

Guideline, I. H. T. (1998). "Testing for Carcinogenicity of Pharmaceuticals S1B," in International Conference on Harmonization.

Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2003). "KNN Model-Based Approach in Classification," in OTM Confederated International Conferences" On the Move to Meaningful Internet Systems, Catania, Sicily, Italy, November 3-7, 2003 (Springer), 2888, 986–996. doi:10.1007/978-3-540-39964-3_62

Hong, H., Xie, Q., Ge, W., Qian, F., Fang, H., Shi, L., et al. (2008). Mold2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* 48, 1337–1344. doi:10.1021/ci800038f

Hwang, D., Jeon, M., and Kang, J. (2020). "A Drug-Induced Liver Injury Prediction Model Using Transcriptional Response Data with Graph Neural Network," in 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), Busan, Korea, Feb. 2020 (IEEE), 323–329. doi:10.1109/bigcomp48618.2020.00-54

Jaeger, S., Fulle, S., and Turk, S. (2018). Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* 58, 27–35. doi:10.1021/acs.jcim.7b00616

Kennard, R. W., and Stone, L. A. (1969). Computer Aided Design of Experiments. *Technometrics* 11, 137–148. doi:10.1080/00401706.1969.10490666

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2021). PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res.* 49, D1388–D1395. doi:10.1093/nar/gkaa971

Lee, M., Kwon, J., and Chung, M.-K. (2003). Enhanced Prediction of Potential Rodent Carcinogenicity by Utilizing Comet Assay and Apoptotic Assay in Combination. *Mutat. Research/Genetic Toxicol. Environ. Mutagenesis* 541, 9–19. doi:10.1016/s1383-5718(03)00175-x

Li, H.-H., Chen, R., Hyduke, D. R., Williams, A., Frötschl, R., Ellinger-Ziegelbauer, H., et al. (2017). Development and Validation of a High-Throughput Transcriptomic Biomarker to Address 21st century Genetic Toxicology Needs. *Proc. Natl. Acad. Sci. USA* 114, E10881–E10889. doi:10.1073/pnas.1714109114

Li, N., Qi, J., Wang, P., Zhang, X., Zhang, T., and Li, H. (2019). Quantitative Structure-Activity Relationship (QSAR) Study of Carcinogenicity of Polycyclic Aromatic Hydrocarbons (PAHs) in Atmospheric Particulate Matter by Random forest (RF). *Anal. Methods* 11, 1816–1821. doi:10.1039/c8ay02720j

Li, T., Tong, W., Roberts, R., Liu, Z., and Thakkar, S. (2020). Deep Learning on High-Throughput Transcriptomics to Predict Drug-Induced Liver Injury. *Front. Bioeng. Biotechnol.* 8, 562677. doi:10.3389/fbioe.2020.562677

Li, T., Tong, W., Roberts, R., Liu, Z., and Thakkar, S. (2021). DeepDILI: Deep Learning-Powered Drug-Induced Liver Injury Prediction Using Model-Level Representation. *Chem. Res. Toxicol.* 34, 550–565. doi:10.1021/acs.chemrestox.0c00374

Liu, Z., Kelly, R., Fang, H., Ding, D., and Tong, W. (2011). Comparative Analysis of Predictive Models for Nongenotoxic Hepatocarcinogenicity Using Both Toxicogenomics and Quantitative Structure-Activity Relationships. *Chem. Res. Toxicol.* 24, 1062–1070. doi:10.1021/tx2000637

Maher, G., Parker, D., Wilson, N., and Marsden, A. (2020). Neural Network Vessel Lumen Regression for Automated Lumen Cross-Section Segmentation in Cardiovascular Image-Based Modeling. *Cardiovasc. Eng. Tech.* 11, 621–635. doi:10.1007/s13239-020-00497-5

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426.

Morales, A. H., Pérez, M. Á. C., Combes, R. D., and González, M. P. (2006). Quantitative Structure Activity Relationship for the Computational Prediction of Nitrocompounds Carcinogenicity. *Toxicology* 220, 51–62. doi:10.1016/j.tox.2005.11.024

Morton, D., Alden, C. L., Roth, A. J., and Usui, T. (2002). The Tg rasH2 Mouse in Cancer hazard Identification. *Toxicol. Pathol.* 30, 139–146. doi:10.1080/01926230252824851

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). *Scikit-learn: Machine Learning in Python*. The Journal of Machine Learning Research, 12, 2825–2830.

Rang, H. P., and Hill, R. G. (2013). "Chapter 15-Assessing Drug Safety", *Drug Discovery and Development: Facts and Figures, Drug Discovery and Development* Editors RG Hill and HP Rang. 2nd edition. (Churchill Livingstone: Elsevier), 211–225. doi:10.1016/B978-0-7020-4299-7.00015-9 https://www.sciencedirect.com/science/article/pii/B9780702042997000159.

Rashed-Al-Mahfuz, M., Moni, M. A., Uddin, S., Alyami, S. A., Summers, M. A., and Eapen, V. (2021). A Deep Convolutional Neural Network Method to Detect Seizures and Characteristic Frequencies Using Epileptic Electroencephalogram (EEG) Data. *IEEE J. Transl. Eng. Health Med.* 9, 1–12. doi:10.1109/jtehm.2021.3050925

Semenova, E., Williams, D. P., Afzal, A. M., and Lazic, S. E. (2020). A Bayesian Neural Network for Toxicity Prediction. *Comput. Toxicol.* 16, 100133. doi:10.1016/j.comtox.2020.100133

Shah, I., Liu, J., Judson, R. S., Thomas, R. S., and Patlewicz, G. (2016). Systematically Evaluating Read-Across Prediction and Performance Using a Local Validity Approach Characterized by Chemical Structure and Bioactivity Information. *Regul. Toxicol. Pharmacol.* 79, 12–24. doi:10.1016/j.yrtph.2016.05.008

Shang, C., Liu, Q., Chen, K.-S., Sun, J., Lu, J., Yi, J., et al. (2018). Edge Attention-Based Multi-Relational Graph Convolutional Networks. arXiv e-prints, arXiv: 1802.04944.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* 43, 1947–1958. doi:10.1021/ci034160g

Tanabe, K., Lučić, B., Amić, D., Kurita, T., Kaihara, M., Onodera, N., et al. (2010). Prediction of Carcinogenicity for Diverse Chemicals Based on Substructure Grouping and SVM Modeling. *Mol. Divers.* 14, 789–802. doi:10.1007/s11030-010-9232-y

Tang, Z.-H., Cao, W.-X., Su, M.-X., Chen, X., and Lu, J.-J. (2017). Osimertinib Induces Autophagy and Apoptosis via Reactive Oxygen Species Generation in Non-small Cell Lung Cancer Cells. *Toxicol. Appl. Pharmacol.* 321, 18–26. doi:10.1016/j.taap.2017.02.017

Tetko, I. V., Karpov, P., Bruno, E., Kimber, T. B., and Godin, G. (2019). "Augmentation Is what You Need," in International Conference on Artificial Neural Networks, Munich, Germany, September 17-19, 2019 (Springer), 831–835. doi:10.1007/978-3-030-30493-5_79

Toropova, A. P., and Toropov, A. A. (2018). CORAL: QSAR Models for Carcinogenicity of Organic Compounds for Male and Female Rats. *Comput. Biol. Chem.* 72, 26–32. doi:10.1016/j.compbiolchem.2017.12.012

Tung, C.-W. (2014). "Acquiring Decision Rules for Predicting ames-negative Hepatocarcinogens Using Chemical-Chemical Interactions," in IAPR International Conference on Pattern Recognition in Bioinformatics, Stockholm, Sweden, August 21-23, 2014 (Springer), 1–9. doi:10.1007/978-3-319-09192-1_1

Tung, C.-W. (2013). "Prediction of Non-Genotoxic Hepatocarcinogenicity Using Chemical-Protein Interactions," in IAPR International Conference on Pattern Recognition in Bioinformatics, Nice, France, June 17-20, 2013 (Springer), 231–241. doi:10.1007/978-3-642-39159-0_21

Venkatachalam, S., Tyner, S., Pickering, C., Boley, S., Recio, L., French, J., et al. (2001). Is P53 Haploinsufficient for Tumor Suppression? Implications for the P53 +/- Mouse Model in Carcinogenicity Testing. *Toxicologic Path.* 29, 147–154. doi:10.1080/019262301753178555

Vinken, M., Benfenati, E., Busquet, F., Castell, J., Clevert, D.-A., De Kok, T. M., et al. (2021). Safer Chemicals Using Less Animals: Kick-Off of the European Ontox Project. *Toxicology* 458, 152846. doi:10.1016/j.tox.2021.152846

Wang, J., Ding, H., Bidgoli, F. A., Zhou, B., Iribarren, C., Molloi, S., et al. (2017). Detecting Cardiovascular Disease from Mammograms with Deep Learning. *IEEE Trans. Med. Imaging* 36, 1172–1181. doi:10.1109/tmi.2017.2655486

Wang, Y.-W., Huang, L., Jiang, S.-W., Li, K., Zou, J., and Yang, S.-Y. (2020). CapsCarcino: A Novel Sparse Data Deep Learning Tool for Predicting Carcinogens. *Food Chem. Toxicol.* 135, 110921. doi:10.1016/j.fct.2019.110921

Willett, P. (2006). Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discov. Today* 11, 1046–1053. doi:10.1016/j.drudis.2006.10.005

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi:10.1093/nar/gkx1037

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* 9, 513–530. doi:10.1039/c7sc02664a

Yamamoto, S., Urano, K., Koizumi, H., Wakana, S., Hioki, K., Mitsumori, K., et al. (1998). Validation of Transgenic Mice Carrying the Human Prototype C-Ha-Ras Gene as a Bioassay Model for Rapid Carcinogenicity Testing. *Environ. Health Perspect.* 106, 57–69. doi:10.2307/3433912

Yang, H., Lou, C., Li, W., Liu, G., and Tang, Y. (2020). Computational Approaches to Identify Structural Alerts and Their Applications in Environmental Toxicology and Drug Discovery. *Chem. Res. Toxicol.* 33, 1312–1322. doi:10.1021/acs.chemrestox.0c00006

Yauk, C. L., Harrill, A. H., Ellinger-Ziegelbauer, H., van der Laan, J. W., Moggs, J., Froetschl, R., et al. (2020). A Cross-Sector Call to Improve Carcinogenicity Risk Assessment through Use of Genomic Methodologies. *Regul. Toxicol. Pharmacol.* 110, 104526. doi:10.1016/j.yrtph.2019.104526

Young, J. F., Tong, W., Fang, H., Xie, Q., Pearce, B., Hashemi, R., et al. (2004). Building an Organ-Specific Carcinogenic Database for SAR Analyses. *J. Toxicol. Environ. Health A* 67, 1363–1389. doi:10.1080/15287390490471479

Zeleznik, R., Foldyna, B., Eslami, P., Weiss, J., Alexander, I., Taron, J., et al. (2021). Deep Convolutional Neural Networks to Predict Cardiovascular Risk from Computed Tomography. *Nat. Commun.* 12, 1–9. doi:10.1038/s41467-021-20966-2

Zhang, C., Cheng, F., Li, W., Liu, G., Lee, P. W., and Tang, Y. (2016). In silicoPrediction of Drug Induced Liver Toxicity Using Substructure Pattern Recognition Method. *Mol. Inf.* 35, 136–144. doi:10.1002/minf.201500055

Zhang, H., Cao, Z.-X., Li, M., Li, Y.-Z., and Peng, C. (2016). Novel Naïve Bayes Classification Models for Predicting the Carcinogenicity of Chemicals. *Food Chem. Toxicol.* 97, 141–149. doi:10.1016/j.fct.2016.09.005

Zhang, L., Ai, H., Chen, W., Yin, Z., Hu, H., Zhu, J., et al. (2017). CarcinoPred-EL: Novel Models for Predicting the Carcinogenicity of Chemicals Using Molecular Fingerprints and Ensemble Learning Methods. *Sci. Rep.* 7, 1–14. doi:10.1038/s41598-017-02365-0

Author Disclaimer: This manuscript reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration. Any mention of commercial products is for clarification only and is not intended as approval, endorsement, or recommendation.

Conflict of Interest: RR is co-founder and co-director of ApconiX, an integrated toxicology and ion channel company that provides expert advice on non-clinical aspects of drug discovery and drug development to academia, industry, and not-for-profit organizations.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 06 December 2021
doi: 10.3389/frai.2021.729834

BERT-Based Natural Language Processing of Drug Labeling Documents: A Case Study for Classifying Drug-Induced Liver Injury Risk

Yue Wu, Zhichao Liu, Leihong Wu, Minjun Chen and Weida Tong**

Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, United States Food and Drug Administration, Jefferson, AR, United States

OPEN ACCESS

Edited by:
Ruchir Shah,
Sciome LLC, United States

Reviewed by:
Arpit Tandon,
Sciome LLC, United States
Adyasha Maharana,
University of North Carolina at Chapel Hill, United States

****Correspondence:***
Weida Tong
Weida.Tong@fda.hhs.gov
Minjun Chen
Minjun.Chen@fda.hhs.gov

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *23 June 2021*
Accepted: *17 November 2021*
Published: *06 December 2021*

Citation:
Wu Y, Liu Z, Wu L, Chen M and Tong W (2021) BERT-Based Natural Language Processing of Drug Labeling Documents: A Case Study for Classifying Drug-Induced Liver Injury Risk. Front. Artif. Intell. 4:729834. doi: 10.3389/frai.2021.729834

Background & Aims: The United States Food and Drug Administration (FDA) regulates a broad range of consumer products, which account for about 25% of the United States market. The FDA regulatory activities often involve producing and reading of a large number of documents, which is time consuming and labor intensive. To support regulatory science at FDA, we evaluated artificial intelligence (AI)-based natural language processing (NLP) of regulatory documents for text classification and compared deep learning-based models with a conventional keywords-based model.

Methods: FDA drug labeling documents were used as a representative regulatory data source to classify drug-induced liver injury (DILI) risk by employing the state-of-the-art language model BERT. The resulting NLP-DILI classification model was statistically validated with both internal and external validation procedures and applied to the labeling data from the European Medicines Agency (EMA) for cross-agency application.

Results: The NLP-DILI model developed using FDA labeling documents and evaluated by cross-validations in this study showed remarkable performance in DILI classification with a recall of 1 and a precision of 0.78. When cross-agency data were used to validate the model, the performance remained comparable, demonstrating that the model was portable across agencies. Results also suggested that the model was able to capture the semantic meanings of sentences in drug labeling.

Conclusion: Deep learning-based NLP models performed well in DILI classification of drug labeling documents and learned the meanings of complex text in drug labeling. This proof-of-concept work demonstrated that using AI technologies to assist regulatory activities is a promising approach to modernize and advance regulatory science.

Keywords: regulatory science, drug labeling, natural language processing, BERT, drug induced liver injury, United States Food and Drug Administration, European medicines agency, named entity recognition

INTRODUCTION

The United States FDA regulates consumer products including foods, medications and tobacco, which account for about 25% of the United States market (US Food and Drug Administration, 2011a). The core responsibility of FDA is to ensure safe and effective products, while at the same time promote innovation to produce products of better quality (US Food and Drug Administration, 2010). Therefore, FDA must be equipped with the best available tools and methods to facilitate pre-market evaluation and post-market surveillance, which requires a strong field of regulatory science to develop standards and approaches that assess FDA-regulated products with reliable efficiency and consistency (US Food and Drug Administration, 2011a; Hamburg, 2011).

Currently, science and technology are rapidly evolving in the field of healthcare, introducing more complexity to the development and manufacture of new drugs, biologics and medical devices. Artificial intelligence (AI), especially, is a fast-growing area and has shown great potential in addressing the unmet medical and public health needs (Yu et al., 2018; Basile et al., 2019; Chan et al., 2019). A long-lasting challenge for FDA is to efficiently retrieve needed information from a huge number of documents received and regularly generated, such as approval documents, guidance, policies and meeting minutes. A significant amount of time must be spent on manually reading and searching information of interest, besides product evaluation and decision making. AI-based natural language processing (NLP) is a promising approach of speeding up this time-consuming and labor-intensive process.

In this study, we applied AI-based NLP to classify drug labeling documents as a proof-of-concept to demonstrate the utility of AI for regulatory applications. Drug labeling provides comprehensive summaries of medications as a reference for healthcare professionals in making prescribing decisions (Watson and Barash, 2009; McMahon and Preskorn, 2014). It is also an essential resource for FDA reviewers during drug evaluations, and the research community for pharmacovigilance and drug repositioning (Chen et al., 2011, 2016; Hoffman et al., 2016; Fang et al., 2020). There are over 130,000 drug labeling documents in the repository, of which 47,000 are labeling for prescription drugs and biologics (Fang et al., 2020). This represents large amounts of regulatory text data, making manually assessing all drug labeling documents prohibitory, if not impossible. Here, we developed an AI-based approach to classify drug-induced liver injury (DILI) risk indicated in drug labeling documents, which serves as a proxy to test the applicability of AI in facilitating text classification from regulatory documents.

Adverse drug reactions (ADRs) such as DILI are described in three sections, "Adverse Reactions", "Warnings and Precautions" and "Boxed Warning", in FDA drug labeling documents (US Food and Drug Administration, 2006; US Food and Drug Administration, 2011b). The "Warnings and Precautions" section contains the most comprehensive and complicated descriptions not limited to ADRs, but also includes other related aspects such as warnings to patients for signs and symptoms, clinical/laboratory monitoring plans and contraindications, for which sentences containing DILI-related terms do not necessarily suggest attributable DILI events (US Food and Drug Administration, 2011b). In contrast, the "Boxed Warning" section, specific to FDA labeling, contains concise highlights of the most serious ADRs from the "Warnings and Precautions" section (US Food and Drug Administration, 2011b), while the "Adverse Reactions" section more or less lists all possible ADRs (US Food and Drug Administration, 2006). The current manual classification approach largely relies on the use of pre-defined DILI terms to determine whether sentences in the three labeling sections indicate DILI (Chen et al., 2011; 2016). Considering that the terms used in the drug labeling are not well normalized to the international standards such as Medical Dictionary for Regulatory Activities (MedDRA) and Systematized Nomenclature of Medicine (SNOMED) and the complexity of language used for describing ADRs, interpretation and judgement by experts with relevant knowledge and experience are necessary. We used an AI-based approach to address these issues in the current study, as language models can capture the semantic meanings of sentences in free text rather than simple string matching (Radford et al., 2018). Specifically, the state-of-the-art language model, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), was trained for binary DILI classification of FDA-approved drug labeling documents and was externally validated using EMA-approved drug labeling documents. The deep learning-based model, hybrid deep learning-based model and keywords-based model developed in this study were compared for DILI risk classification on drug labeling documents.

MATERIALS AND METHODS

Data Sources for Drug Labeling

FDA drug labeling documents were retrieved from DailyMed (www.dailymed.nlm.nih.gov), a public database that contains up-to-date drug labeling approved by the FDA. Meanwhile, since the EMA issues standardized drug labeling for drugs approved through a centralized procedure, we used UK-marketed drugs as representatives of drugs authorized in Europe (European Medicines Agency, 2009). EMA drug labeling documents were collected from the EMC (www.medicines.org.uk), which maintains the EMA-approved drug labeling for drugs licensed in the United Kingdom.

Drug Selection Criteria

We selected prescription drugs based on three criteria, i) with a single active ingredient, ii) either oral or injection use, and iii) in the categories of NDA, ANDA or BLA, by querying the FDALabel database (https://nctr-crs.fda.gov/fdalabel/ui/search) which maintains over 130,000 drug labeling documents containing critical information pertinent to the safe and effective use of medications (Fang et al., 2020). Over-the-counter drugs were removed because of their different labeling format and requirements compared to prescription drugs. The DILIrank

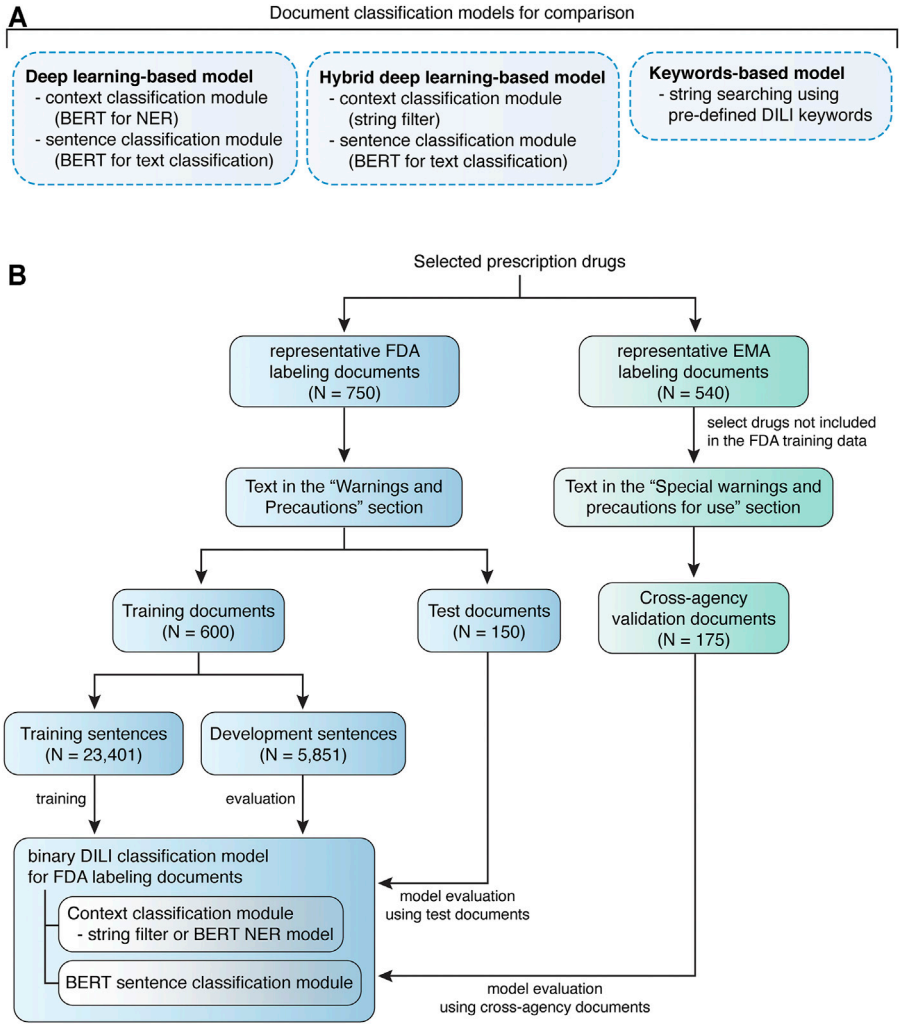


FIGURE 1 | Quorum flowchart describes the study design. **(A)** Drug labeling document classification models developed and compared in this study. **(B)** The study design of model training and evaluation using FDA labeling documents and model validation using EMA labeling documents.

dataset provides the DILI risk annotation for 1,036 drugs marketed in the United States as of 2010 (Chen et al., 2016). We retrieved the most recent drug labeling documents for the queried 750 representative prescription drugs from the DILIrank dataset. Among these drugs, 540 were also licensed in the United Kingdom market. The corresponding EMA drug labeling documents were collected and assessed for DILI risk using the same classification schema described in previously studies (Chen et al., 2011).

Datasets

We focused our analysis on the "Warnings and Precautions" section of FDA labeling documents, as the language for ADR descriptions in this section has the highest complexity compared with the other two sections (US Food and Drug Administration, 2011b). The corresponding section in the EMA labeling documents is the "Special warnings and precautions for use" section (European Medicines Agency, 2009). Texts were extracted from either the "Warnings and Precautions" section (FDA) or the "Special warnings and precautions for use" section (EMA), followed by formatting clearing and sentence tokenization (**Figures 1B**, **2**).

For model training on FDA labeling documents, the representative documents (N = 750) were stratified split into 80% training document dataset (N = 600) and 20% test document dataset (N = 150). Unique sentences (N = 29,252) were extracted from the training document dataset, among which DILI-positive (N = 540) or DILI-negative sentences (N = 28,712) were determined independently by two experts. All disagreements were resolved by discussion. To generate data with more balanced class labels, intermediate datasets were created to facilitate filtering of context prior to sentence classification, via Named Entity Recognition (NER). The unique sentences (*N* = 29,252) from training documents were annotated using the Inside-Outside-Beginning (IOB) style. The annotated sentences were randomly split into 80% training sentence dataset (N =

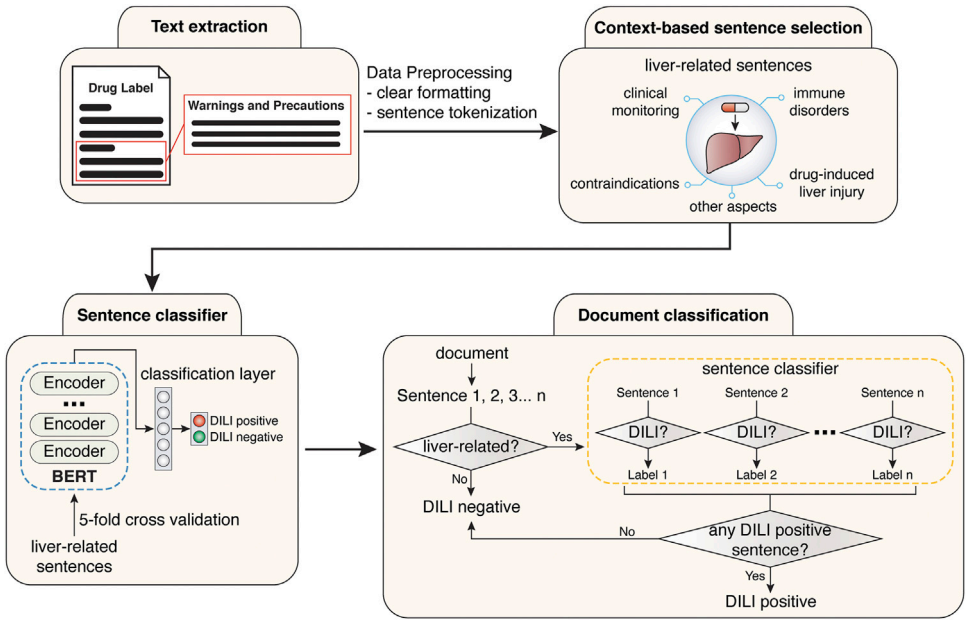


FIGURE 2 | Workflow for the training of sentence classification module and the development of final document classification model.

23,041) and 20% development sentence dataset (N = 5,851) for NER model training (**Figure 1B**). Sentences with tokens related to a liver context, 540 DILI-positive and 1,313 DILI-negative, were selected as liver-related sentences. To simplify the comparison between models, human validated liver-related sentences from the annotated sentences (*N* = 29,252) were used for developing the sentence classification module. Test document dataset was used to evaluated developed models, and cross-agency data, i.e., EMA labeling documents for drugs not included in the FDA training data, was used for external validation (**Figure 1B**).

Examples are given here to illustrate the datasets created for model training. Dataset for context classification included liver-related sentences such as "Hepatic toxicity including hepatic failure resulting in transplantation or death have been reported" and "Rozerem should not be used by patients with severe hepatic impairment" and sentences irrelevant to liver including "Treat all infections due to Group A beta-hemolytic streptococci for at least 10 days". The first two liver-related sentences were used for developing sentence classification models. The first sentence was considered as DILI-positive, while the second sentence is for contraindication information and thus considered as DILI-negative.

To further examine the portability of BERT-based models across agencies, we also developed models using EMA labeling documents as training data and validated the models using FDA labeling documents (**Supplementary Figure 1**). EMA labeling documents (N = 540) were stratified split into 80% training document dataset (N = 431) and 20% test document dataset (N = 109). Unique sentences (N = 14,915) were extracted from the training document dataset, including 232 DILI-positive and 14,683 DILI-negative sentences. Similarly, intermediate datasets were created to facilitate filtering of context prior to sentence classification, via NER. The unique sentences (*N* = 29,252) from training documents were annotated using the IOB style, and randomly split into 80% training sentence dataset (*N* = 11,931) and 20% development sentence dataset (*N* = 2,984) for NER model training (**Supplementary Figure 1**). Sentences with tokens related to a liver context, 232 DILI-positive and 927 DILI-negative, were selected as liver-related sentences. Human validated liver-related sentences from the annotated sentences (*N* = 14,915) were used for developing the sentence classification module. EMA test document dataset was used to evaluated developed models, and FDA labeling documents for drugs not included in the EMA training data, was used for external validation.

Models for Document Classification

In this study, deep learning-based (BERT for DILI classification), hybrid deep learning-based and keywords-based models were developed for classifying drug labeling documents based on whether they contain any sentence suggesting DILI risk (**Figure 1A**).

The deep learning-based and hybrid deep learning-based document classification models consisted of two working modules, a context classification module and a BERT sentence classification module (**Figures 1A**, **2**). These two models shared the same BERT sentence classification module but differed in the context classification module. For each input document, each sentence was passed into the two working modules sequentially (**Figure 2**). The first step was to determine whether the current sentence was related to the liver topic at the context classification module. If not, this sentence was DILI-negative. If yes, this sentence was then passed to the BERT sentence classification module to

TABLE 1 | Sentence count with or without pre-defined liver-related context.

	Without pre-defined context		In context of liver (string-filter)		In context of liver (BERT for NER)	
	FDA	EMA	FDA	EMA	FDA	EMA
DILI positive sentences	540	232	540	232	540	232
DILI negative sentences	28,712	14,915	961	764	1,313	927

determine whether it was DILI-positive or DILI-negative. After evaluating all the sentences in the input document, an array of predicted sentence labels was generated. If any DILI-positive sentences were found in the input document, the document was considered DILI-positive, otherwise as DILI-negative.

A keywords-based document classification model was also developed as a comparison to the deep learning-based and hybrid deep learning-based models (**Figure 1A**). Keywords for detecting DILI risk in the drug labeling were collected from three previous studies (Chen et al., 2011; Demner-Fushman et al., 2018; Suzuki et al., 2015) (**Supplementary Table 1**). Chen et al. summarized a list of DILI keywords for text-mining (via human reading) in the drug labeling, while Suzuki et al. selected a list of MedDRA PT terms for hepatocellular and cholestatic liver injury for text-mining in the WHO VigiBase™. These two lists covered most of the DILI terms, but the keywords commonly had multiple imperfect matches in the drug labeling documents. Thus, these keywords could not be used directly for computerized text-mining in the drug labeling documents. Demner-Fushman et al. normalized the ADR terms in 200 drug labeling documents to MedDRA Preferred Terms (PTs). By using the matching data in the Demner-Fushman et al. study, we generated a keyword list that covers DILI (Chen et al., 2011), liver injury (Suzuki et al., 2015) and hepatic ADRs (Demner-Fushman et al., 2018) terms used in drug labeling. The FDA and EMA test document sets were used to evaluate the performance of keywords-based document classification.

Development of the Context Classification Modules

Two types of context classification modules were created in this study. The first one is a string pattern matching-based context filter. The other one is an NER-based context classification model.

For the hybrid deep learning-based model, general string patterns were used to match sentences with any possible relation to liver, including indications, contraindications, ADRs, clinical monitoring, immune disorders, etc. (**Supplementary Table 2**). Most DILI-negative sentences irrelevant to liver were filtered out by applying such pre-defined context, yielding relatively balanced sentence datasets without losing any DILI-positive sentences (**Table 1**).

Meanwhile, a BERT-based NER model was developed as the context classification module in the deep learning-based model. The NER model was developed by using training sentence dataset and evaluated on development sentence dataset at each epoch of training. The hyperparameters used for model training are listed in **Supplementary Table 3**. This BERT-based context classification module was then evaluated by performing context classification on sentences extracted from test documents and cross-agency validation documents.

Development of the BERT-Based Sentence Classification Module

The liver-related sentences selected from training sentence dataset were used for developing a BERT (base, uncased) model for binary DILI classification as the sentence classification module, while the liver-related sentences selected from development sentence dataset were used to evaluate the performance of the BERT-based sentence classification module. The hyperparameters used for model training are listed in **Supplementary Table 4**. The sentence classification module was evaluated using shuffled five-fold cross-validations on the liver-related sentences for 100 times (**Supplementary Figure 2**). In comparison to developing a context-dependent sentence classification model, we also trained a sentence classification model using imbalance sentence datasets extracted from training documents. To address the dataset imbalance issue, we applied an oversampling method, i.e., randomly sampling based on class weights.

Permutation analysis was conducted to determine whether the models developed in this study perform at chance (Chen et al., 2013). Permutated datasets were generated by 100 times of resampling the liver-related training and test sentence datasets with randomly shuffled DILI classification labels (positive or negative). The performance of the resulting 100 models was compared with that from 100 repetitions of cross-validations with random sampling (**Supplementary Figure 2**). A two-sided *t*-test was used determine the statistical significance of the difference between the accuracy scores obtained from permutated data and original data.

Shapley Additive Explanations (SHAP) values (Lundberg and Lee, 2017) were used to quantify the contribution of each token to the prediction made by the model. Higher feature values (red) push the model prediction towards DILI-positive, while lower features (blue) values push the model prediction towards DILI-negative.

Implementation

The embedding layer and 12-layer encoder from BERT were adopted and connected with a dense layer for token or sentence classification. The deep learning-based model combines NER (token classification) and sentence classification modules. A document is broken down into sentences s_1, $s_2 \ldots s_i$. All sentences are passed into the NER module, where tokens

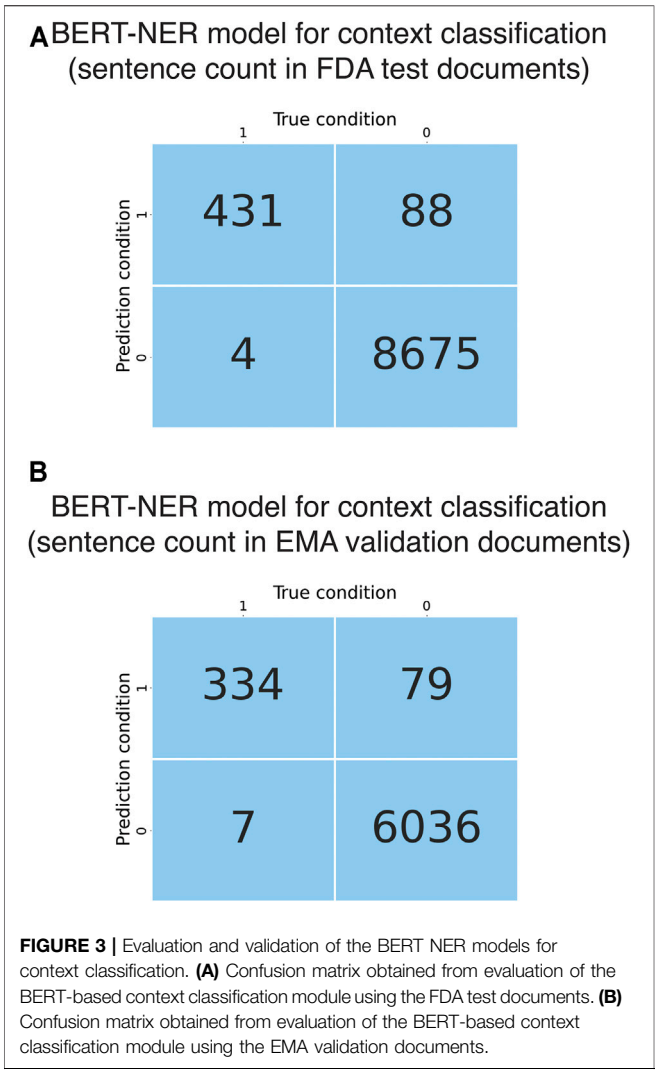


FIGURE 3 | Evaluation and validation of the BERT NER models for context classification. **(A)** Confusion matrix obtained from evaluation of the BERT-based context classification module using the FDA test documents. **(B)** Confusion matrix obtained from evaluation of the BERT-based context classification module using the EMA validation documents.

$[t_{11}, t_{12}...t_{1j}]$, $[t_{21}, t_{22}...t_{2j}]...[t_{i1}, t_{i2}...t_{ij}]$ are classified. If none of the tokens is associated with "Liver" (with $(argmax(t_{i1}) = y) \mid (argmax(t_{i2}) = y) \mid ... \mid (argmax(t_{ij}) = y)$ being False for any sentence s_i in a given document, where y equals the value of "Liver" tag.), then document label is returned as 0 (DILI negative). Otherwise, all selected liver related sentences are passed into sentence classification module. Document label is returned as 0 if none of the liver-related sentences is DILI positive ($(\sum_i argmax(s_i) = 0)$, else returned as 1 (DILI negative).

Evaluation Metrics

The NER-based context classification was evaluated at two levels. Recall, precision, and f1-score were reported at token level. Context classification at sentence level was evaluated by recall and precision. The BERT-based binary sentence classification was evaluated using accuracy, recall and precision. The test documents were used to assess the performance of the deep learning-based and hybrid deep learning-based models on document classification. Matthews correlation coefficient (MCC), recall and precision were used to evaluate the quality of binary DILI classification predicted by the models.

RESULTS

Development of the Deep Learning-Based Model for DILI Classification of Labeling Documents

The developed deep learning-based model had a BERT-based NER model as the context classification module and a BERT-based sentence classification module (**Figure 1A**). FDA test documents were used to evaluate the performance of the NER-based context classification module in selecting liver-related sentences. At token level, the context classification module showed excellent performance in recognizing liver-related words, with an F1 score of 0.98 ± 0.003, recall of 0.99 ± 0.002 and precision of 0.98 ± 0.008. When evaluated at sentence level, it had great sensitivity (0.99) as it was able to extract 431 of 435 liver-related sentences from the test documents (**Figure 3A**). The precision was 0.83 (0.83 ± 0.001 from cross-validations) due to that 88 false positives were generated. Considering the large number of non-liver sentences (N = 8,763) in the test documents, the context classification module performed well in predicting non-liver sentences as the false positive rate was 1%. Further, the context classification module was externally validated using EMA test documents. It detected 334 of 341 liver-related sentences while 79 false positives were predicted from 6,115 non-liver sentences, which was comparable to the results obtained using FDA test documents (**Figure 3B**).

The BERT-based sentence classification module is the same from the hybrid deep learning-based model, which was developed using liver-related sentences. This module showed an accuracy of 0.81 ± 0.02, recall of 0.82 ± 0.03 and precision of 0.82 ± 0.02. To confirm that the sentence classification module did not perform at chance, we conducted permutation tests. The sentence classification models trained on the permutated FDA training sentences exhibited a great decrease in average accuracy score, as compared to that obtained from cross-validations (0.56 versus 0.81, $p < 0.0001$) (**Supplementary Figure 2**). These results suggested that the observed accuracy scores of the sentence classification models were unlikely to be obtained by chance.

The performance of the deep learning-based model regarding document classification was evaluated using FDA test documents and externally validated using EMA validation documents (**Figure 1B**). The deep learning-based model also showed excellent performance in DILI prediction on drug labeling documents with an MCC of 0.84 (**Table 2**). It could detect all 40 of the DILI-positive documents in the FDA test set (**Figure 4A** and **Table 2**). Eleven false positives were found from a total of 110 DILI-negative documents, and thus the precision was 0.78. These results were consistent with that from model validation using cross-agency data (EMA validation documents), which had an MCC of 0.79, recall of 1 and precision of 0.71 (**Figure 4D** and **Table 2**).

In comparison with models trained on liver-related sentences, we also developed sentence classification models using all sentences from the training documents, which were extremely imbalanced

TABLE 2 | Model evaluation and validation using cross-agency data.

Model evaluation using FDA test documents			
Document classification models	Matthews correlation coefficient	Recall	Precision
Deep learning-based model	0.84	1.00	0.78
Hybrid deep learning-based model	0.87	1.00	0.82
Keywords-based model	0.60	0.90	0.58
Model validation using cross-agency data (EMA test documents)			
Document classification models	Matthews correlation coefficient	Recall	Precision
Deep learning-based model	0.79	1.00	0.71
Hybrid deep learning-based model	0.84	1.00	0.77
Keywords-based model	0.61	0.96	0.55

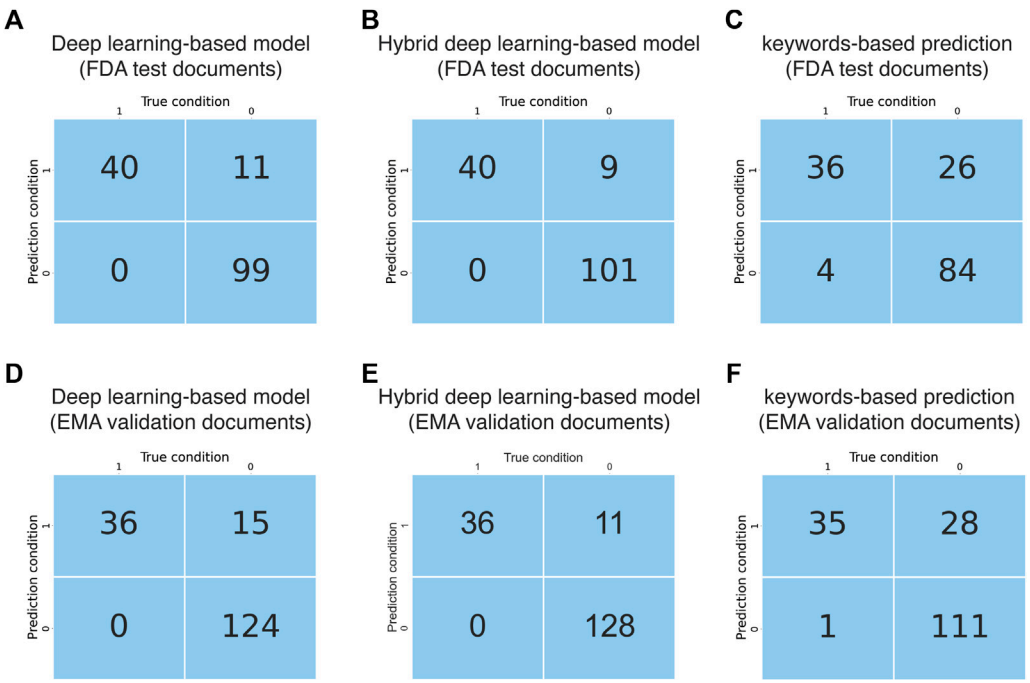


FIGURE 4 | Evaluation and validation of the document classification models. **(A)** Confusion matrix obtained from evaluation of the AI model using FDA test documents. **(B)** Confusion matrix obtained from evaluation of the hybrid deep learning-based model using FDA test documents. **(C)** Confusion matrix obtained from evaluation of the keywords-based model using FDA test documents. **(D)** Confusion matrix obtained from evaluation of the AI model using EMA validation documents. **(E)** Confusion matrix obtained from evaluation of the hybrid deep learning-based model using EMA validation documents. **(F)** Confusion matrix obtained from evaluation of the keywords-based model using EMA validation documents.

between DILI-positive and negative labels. We observed decreased recall (0.75 ± 0.08) and precision (0.76 ± 0.04) as compared to models developed using liver-related sentences. When oversampling was conducted by randomly sampling according to class weights, recall was increased to 0.80 ± 0.04 while precision dropped significantly to 0.68 ± 0.04. None of these models outperformed the deep learning-based model with NER-based intermediate module at sentence level. When evaluated at document level, the sentence classification model trained on all sentences predicted more false negative FDA documents (N = 4), causing decreased recall (0.90). Interestingly, precision (0.86) was higher than that obtained from the deep learning-based model, as less false positive documents were obtained (N = 6). Similarly, decreased recall (0.89) and increased precision (0.82) were observed when EMA documents were used as external validation data. Higher recall is preferred for the investigated topic in this study, i.e., ADR detection in drug labeling documents, because false positive documents are much easier to be detected during the phase of result interpretation or model validation, as compared to false negative documents.

Development of the Hybrid Deep Learning-Based Model for DILI Classification of Labeling Documents

The developed hybrid deep learning-based model had a string filter-based context classification module followed by a BERT-

based sentence classification module (**Figure 1A**). After context filtering of sentences from the "Warnings and Precautions" section of FDA training documents, 1,501 unique liver-related sentences were collected, of which 540 were DILI-positive while 961 were DILI-negative (**Table 1**). This sentence dataset was used for training the BERT-based sentence classification module. The developed sentence classification module reached high performance regarding DILI classification with accuracy scores of 0.81 ± 0.02 obtained from 100 repetitions of five-fold cross-validations.

The performance of this hybrid deep learning-based model regarding document classification was evaluated using FDA test documents and externally validated using EMA validation documents (**Figures 1B**, **2**). The hybrid deep learning-based model achieved excellent performance in DILI prediction on drug labeling documents with an MCC of 0.87 (**Table 2**). It had a high recall of 1, as it could detect all 40 of the DILI-positive documents in the FDA test set (**Figure 4B** and **Table 2**). Nine false positives were found which resulted in a precision of 0.82. These results were corroborated with that from model validation using cross-agency data (EMA test documents). The hybrid deep learning-based model had a consistent MCC of 0.84, recall of 1 and precision of 0.77 when predicting on the EMA validation documents (**Figure 4E** and **Table 2**).

Interestingly, we observed subtle differences between the deep learning-based and hybrid deep learning-based models in prediction DILI risk. The hybrid deep learning-based model was better at distinguishing liver injury statements in animal studies from human liver injury statements. Also, hepatosplenic T-cell lymphomas due to immunosuppressive treatment could confuse the deep learning-based model rather than the hybrid deep learning-based model. In contrast, the deep learning-based model performed better in detecting term variants/abbreviations, such as SGOT/AST for aspartate aminotransferase and SGPT/ALT for alanine aminotransferase. Although limited in number, the examples from the current data could provide some insight for future research.

Comparison of the Deep Learning-Based and Hybrid Deep Learning-Based Models With the Keyword-Based Model for DILI Classification of Labeling Documents

As a comparison to the deep learning-based and hybrid deep learning-based models, a keyword matching-based approach was also used to classify the FDA and EMA test documents. The keyword-based classification on FDA test documents showed a significantly lower MCC of 0.60, as compared to that from predictions made by the deep learning-based (0.84) and hybrid deep learning-based (0.87) models (**Table 2**). It produced a larger number of false positives (N = 26), thus the precision (0.58) was remarkably lower than the deep learning-based (0.78) and hybrid deep learning-based (0.82) models (**Figure 4C** and **Table 2**). Most of the false positives produced by keyword-based DILI classification, but not by the deep learning-based and hybrid deep learning-based models, were related to description of contraindications or precautions to special populations (e.g., patients with hepatic impairment) and hypersensitivity reactions (**Supplementary Table 5**). Also, four false negatives were generated by the keywords-based document classification model, but none by deep learning-based and hybrid deep learning-based models. Corroborated with the DILI classification results obtained from the FDA test documents, the keywords-based DILI classification on the EMA validation documents also showed poor performance in controlling the number of false positives, which generated a low precision of 0.55 (**Figure 4F** and **Table 2**). The MCC was calculated to be 0.61.

DISCUSSION

In this study we used an AI-based NLP approach to classify drug labeling documents according to the DILI risk suggested in the text from the "Warnings and Precautions" section. The motivation of this investigation was to address two questions that are important to both regulatory application and drug safety research, i) whether AI-based NLP tools can be used to classify a drug's DILI potential specified in the drug labeling documents, and ii) whether an AI-based model developed using FDA labeling documents was portable to the documents in other regulatory agencies with comparable performance. Therefore, we developed BERT-based deep learning models for DILI classification, which were rigorously evaluated in this study.

Our results showed that both the deep learning-based model and the hybrid deep learning-based model developed in this study had outstanding performance in predicting DILI risk encoded in the drug labeling documents, regardless of whether FDA labeling documents or EMA labeling documents were used for model training. This suggested that the deep learning-based models could capture the semantic meanings of sentences in the drug labeling documents, considering that the descriptions approved by the two agencies have some degree of difference in terms of language style and format. The contributions of word tokens to model predictions were explored to examine whether the model learned reasonable semantic meanings of the sentences in the drug labeling. SHAP values were used to quantify the contributions of each word token to the prediction made by the model. In the representative DILI-positive sentences (**Figures 5A,B**), DILI-related words such as "hepatic failure", "hepatotoxicity" and "hepatitis" showed positive contributions (red) and pushed the model prediction toward DILI-positive. In contrast, the word "hepatitis" did not have positive contributions when it was in the phrases "chronic hepatitis B" and "chronic hepatitis C". Collectively, these results suggested that the developed NLP models could capture the semantic relationships between words in a given sentence.

Notably, the deep learning-based NLP models developed using FDA labeling documents could also be used by other agencies such as EMA without a notable decrease in performance. Furthermore, we also developed a deep learning-based model and a hybrid deep learning-based model using EMA labeling documents (**Supplementary Figure 1**). The models trained on the EMA data showed comparable performance when evaluated using EMA test documents and the FDA validation documents (**Supplementary Table 6**), which confirmed the portability of the

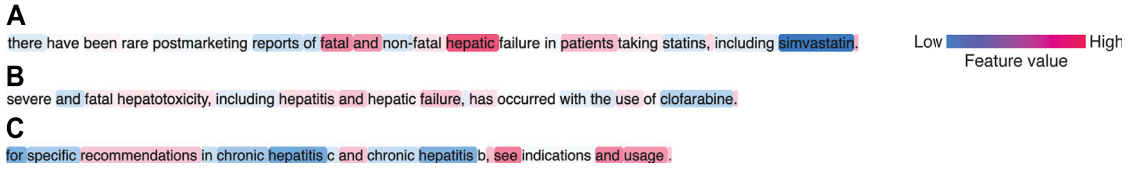


FIGURE 5 | Representative sentences showing contributions of word tokens to model predictions. **(A)** DILI-positive sentence due to fatal hepatic failure. **(B)** DILI-positive sentence due to hepatitis/hepatic failure. **(C)** DILI-negative sentence that provides indication information.

deep learning-based NLP models across agencies. This demonstrated a promising potential of using AI technology to facilitate regulatory activities including drug evaluation and pharmacovigilance.

To best resemble our human reading-based approach and allow for an interpretable classification, we chose a sentence classification strategy over directly using whole documents as input. Briefly, we wanted our final model to be able to select liver-related sentences and determine whether they suggest DILI risk. The determination of DILI risk of a document was not based on quantitative measurement of the number of DILI-positive sentences, but rather dependent on detection of at least one DILI-positive sentence. In this regard, the document classification model is sensitive to false positives. Both the FDA and EMA models developed in this study had low false positive rates (6–10%), suggesting that the models performed well in controlling false positives. Furthermore, the sentence classification strategy allowed us to easily track which sentences in a document were the basis for the document classification model to determine DILI potential. It also provided information regarding what type of sentences were ambiguous in DILI risk to the models. From a technical perspective, the current BERT pre-trained model has an input limit of 512 tokens. In order to process lengthy documents such as the "Warnings and Precautions" section containing hundreds to thousands of words, various solutions have been proposed, including i) text truncation and ii) text splitting combined with different pooling methods or Long Short-Term Memory networks (Adhikari et al., 2019a, 2019b; Sun et al., 2020). Such more complex model structures do not fit better the classification criteria for this study and complicate the model interpretation, as compared to a sentence classification-based model structure. Therefore, we used a hierarchical model structure to predict DILI risk on each individual sentence in a given drug labeling document and output a document classification label based on the combined sentence classification results. Moreover, since not all sentences should contribute to the DILI prediction, we used a context filter as a gating mechanism to select liver-related sentence for DILI prediction, which is similar to aspect-based sentiment analysis (Sun et al., 2019; Xu et al., 2019; Choi et al., 2020). The framework for creation of dataset and training of context classification model can be extended to other topics, e.g., cardiotoxicity, drug indication and drug-drug interactions. Outputs from context-classification can also be used for information retrieval pipelines.

Of note, sentence classification models trained on all sentences with skewed distributions did not have dramatically decreased performance than NER-sentence classification combined models. We observed 7 and 6% drop in recall and precision respectively at sentence level, and 10% decrease in recall but 8% increase in precision at document level. However, addition of an NER-based context classification module would be a better approach for the following reasons. First, all the BERT-based models developed in this study were designed to record sentences that were predicted as DILI-positive for human justification. Since the number of sentences suggesting adverse events is far less than that of sentences carrying no information of adverse events, it is much easier to find false positive documents as compared to false negative documents. Also, the false positive sentences collected from users could be used later for model improvement by further training or re-training. Therefore, higher recall is preferred. Second, inclusion of NER-based context classification module enables context-specific sentence classification, which is more flexible, especially in the case of classifying sentences belong to multiple contexts. For example, DILI can be associated with immune-mediated cutaneous ADRs such as Drug Reaction with Eosinophilia and Systemic symptoms, Stevens-Johnson syndrome and toxic epidermal necrolysis (Andrade et al., 2019). Sentences containing information across different contexts could be ambiguous to multiclass sentence classification models for detecting different types of ADRs. If binary sentence classification models were developed for detecting each type of ADRs, large number of negative samples would be used for model training repeatedly, which is not an efficient design. Moreover, NER-based context classification module is versatile and can provide additional functionalities including facilitating information retrieval.

Previous efforts in data mining of drug labeling documents primarily relied on the use of specific ADR terms (Chen et al., 2011; Demner-Fushman et al., 2018; Wu et al., 2019). International standards, MedDRA and SNOMED, have been used for searching ADR terms in drug labeling (Demner-Fushman et al., 2018; Wu et al., 2019). The ADR descriptions in drug labeling often do not follow these standards, which requires human effort in matching ADR terms in drug labeling with standards. Annotation resources have been reported to normalize the terms used in drug labeling (Demner-Fushman et al., 2018). However, providing annotations for such a large repository is not a trivial task. As shown in **Supplementary Table 3**, many standard terms such as MedDRA PTs have a number of matched terms in drug labeling.

For example, there have been at least 31 different terms in FDA labeling for the MedDRA PT "Alanine aminotransferase increased", and 34 for "Blood bilirubin increased". New variations in ADR terms are likely to be introduced into drug labeling in the future. Therefore, updating and maintaining such annotations are labor intensive. The deep learning-based model developed in the current study, with BERT-based NER and sentence classification combined, outperformed the keywords-based model by a large margin. Importantly, BERT-based models are not only easy to implement and extend but can also be further improved with better pretrained models in the future.

Furthermore, DILI classification of the labeling documents is a more complicated task than keywords matching. In some cases, a sentence containing hepatic ADR terms does not necessarily suggest DILI. For example, a sentence containing the term hepatitis could indicate antiviral treatment of hepatitis B viruses. It could also be contraindication information specifying that patients with hepatic deficiency due to hepatitis should not take the drug. All these cases are present in the complex descriptions from the "Warnings and Precautions" section. Therefore, human interpretation has been necessary to determine DILI-positive sentences in drug labeling documents (Chen et al., 2011; 2016).

Over the past few years, transformers models have changed the landscape of NLP (Wolf et al., 2020). The BERT model used in this study enables bidirectional text learning by using masks (Devlin et al., 2019). Notably, the multi-headed attention architecture leverages the use of deep neural networks to capture the relationships between words within a sentence and across sentences (Vaswani et al., 2017; Devlin et al., 2019). These two important features allow the BERT model to learn the semantic meanings of a sentence or sentences effectively and efficiently. Thus, we chose BERT as our first attempt to develop AI-based NLP tools, which do not rely on keywords dictionaries but rather learn the meaning of text and perform tasks close to humans. Indeed, our results showed that model predictions were driven by the DILI-related words such as hepatic failure, hepatotoxicity and hepatitis in the representative DILI positive sentences. For the representative DILI positive sentences, model predictions were based on the detection of DILI-negative information including chronic hepatitis B/C, even though DILI-related words were also present in the sentence.

Additionally, we acknowledge the following limitations of this study. The dataset size is relatively small, especially for document-level classification results. This is by large due to that DILI is not a common adverse event, with an incidence of approximately 20 cases per 100,000 persons annually (Garcia-Cortes et al., 2020). There are limited number of drugs carrying warnings for DILI. The developed pipeline was evaluated on just a single topic, i.e., liver injury. Thus, it remains to be proven by future research that this framework is indeed extensible to other topics. The pre-trained BERT model was trained on corpuses using general language. Drug labeling, however, uses many domain-specific terms. Further in-domain training of the BERT model might improve the model performance. Also, we did not try other transformers models such as GPT-2 (Radford et al., 2019) and XLNet (Yang et al., 2019) for comparison. The main purpose of this work was to test the applicability of modern language models on regulatory documents, rather than select better models.

CONCLUSION

In the current study we demonstrated that AI-based NLP tools performed well in DILI classification of drug labeling documents from two different regulatory agencies, FDA and EMA. The deep learning-based and hybrid deep learning-based models outperformed the keywords-based models and were portable from one agency to the other without a notable decrease in performance. Our results suggest that AI models are able to learn the meaning of text and handle NLP tasks with good accuracy. This proof-of-concept work show that using AI technology to facilitate regulatory activities is a promising approach to modernize and advance regulatory science.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://nctr-crs.fda.gov/fdalabel/ui/search.

AUTHOR CONTRIBUTIONS

Conceptualization (WT); Data acquisition (YW and LW); Methodology (YW, WT, ZL, and MC); Data analysis (YW); Manuscript writing (YW, MC, and WT); Manuscript review and editing (YW, WT, and MC).

FUNDING

The research is internally funded by the project (E0767701) at National Center for Toxicological Research, United States Food and Drug Administration.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.729834/full#supplementary-material

REFERENCES

Adhikari, A., Ram, A., Tang, R., and Lin, J. (2019a). *DocBERT: BERT for Document Classification*. *arXiv* :1904.08398v3 [cs.CL].

Adhikari, A., Ram, A., Tang, R., and Lin, J. (2019b). "Rethinking Complex Neural Network Architectures for Document Classification," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Minneapolis, MN: Association for Computational Linguistics), 1, 4046–4051. doi:10.18653/v1/N19-1408

Andrade, R. J., Chalasani, N., Björnsson, E. S., Suzuki, A., Kullak-Ublick, G. A., Watkins, P. B., et al. (2019). Drug-induced Liver Injury. *Nat. Rev. Dis. Primers* 5 (1), 58. doi:10.1038/s41572-019-0105-0

Basile, A. O., Yahi, A., and Tatonetti, N. P. (2019). Artificial Intelligence for Drug Toxicity and Safety. *Trends Pharmacol. Sci.* 40, 624–635. doi:10.1016/j.tips.2019.07.005

Chan, H. C. S., Shan, H., Dahoun, T., Vogel, H., and Yuan, S. (2019). Advancing Drug Discovery via Artificial Intelligence. *Trends Pharmacol. Sci.* 40, 592–604. doi:10.1016/j.tips.2019.06.004

Chen, M., Hong, H., Fang, H., Kelly, R., Zhou, G., Borlak, J., et al. (2013). Quantitative Structure-Activity Relationship Models for Predicting Drug-Induced Liver Injury Based on FDA-Approved Drug Labeling Annotation and Using a Large Collection of Drugs. *Toxicol. Sci.* 136, 242–249. doi:10.1093/toxsci/kft189

Chen, M., Suzuki, A., Thakkar, S., Yu, K., Hu, C., and Tong, W. (2016). DILIrank: the Largest Reference Drug List Ranked by the Risk for Developing Drug-Induced Liver Injury in Humans. *Drug Discov. TodayToday* 21, 648–653. doi:10.1016/j.drudis.2016.02.015

Chen, M., Vijay, V., Shi, Q., Liu, Z., Fang, H., and Tong, W. (2011). FDA-approved Drug Labeling for the Study of Drug-Induced Liver Injury. *Drug Discov. Today* 16, 697–703. doi:10.1016/j.drudis.2011.05.007

Choi, G., Oh, S., and Kim, H. (2020). Improving Document-Level Sentiment Classification Using Importance of Sentences. *Entropy* 22, 1336. doi:10.3390/e22121336

Demner-Fushman, D., Shooshan, S. E., Rodriguez, L., Aronson, A. R., Lang, F., Rogers, W., et al. (2018). A Dataset of 200 Structured Product Labels Annotated for Adverse Drug Reactions. *Sci. Data* 5, 180001–180008. doi:10.1038/sdata.2018.1

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *arXiv* :1810.04805v2 [cs.CL].

European Medicines Agency (2009). A Guideline on Summary of Product Characteristics. Available at: http://ec.europa.eu/health/files/eudralex/vol-2/c/smpc_guideline_rev2_en.pdf.

Fang, H., Harris, S., Liu, Z., Thakkar, S., Yang, J., Ingle, T., et al. (2020). FDALabel for Drug Repurposing Studies and beyond. *Nat. Biotechnol.* 38, 1378–1379. doi:10.1038/s41587-020-00751-0

Garcia-Cortes, M., Robles-Diaz, M., Stephens, C., Ortega-Alonso, A., Lucena, M. I., and Andrade, R. J. (2020). Drug Induced Liver Injury: an Update. *Arch. Toxicol.* 94, 3381–3407. doi:10.1007/s00204-020-02885-1

Hamburg, M. A. (2011). Advancing Regulatory Science. *Science* 331, 987. doi:10.1126/science.1204432

Hoffman, K. B., Dimbil, M., Tatonetti, N. P., and Kyle, R. F. (2016). A Pharmacovigilance Signaling System Based on FDA Regulatory Action and Post-Marketing Adverse Event Reports. *Drug Saf.* 39, 561–575. doi:10.1007/s40264-016-0409-x

Lundberg, S. M., and Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. *arXiv* :1705.07874v2 [cs.AI].

McMahon, D., and Preskorn, S. H. (2014). The Package Insert. *J. Psychiatr. Pract.* 20, 284–290. doi:10.1097/01.pra.0000452565.83039.20

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). *Improving Language Understanding with Unsupervised Learning*. OpenAi: Technical report.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). *Language Models Are Unsupervised Multitask Learners*. OpenAi: Technical report.

Sun, C., Huang, L., and Qiu, X. (2019). *Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence*. *arXiv* :1903.09588 [cs.CL].

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2020). *How to Fine-Tune BERT for Text Classification*. arXiv *:1905.05583v3 [cs.CL]*.

Suzuki, A., Yuen, N. A., Ilic, K., Miller, R. T., Reese, M. J., Brown, H. R., et al. (2015). Comedications Alter Drug-Induced Liver Injury Reporting Frequency: Data Mining in the WHO VigiBase™. *Regul. Toxicol. Pharmacol.* 72, 481–490. doi:10.1016/j.yrtph.2015.05.004

US Food and Drug Administration (2011a). Advancing Regulatory Science at FDA: A Strategic Plan. Available at: https://www.fda.gov/media/81109/download.

US Food and Drug Administration (2010). Advancing Regulatory Science for Public Health. Available at: https://www.fda.gov/media/123792/download.

US Food and Drug Administration (2006). Adverse Reactions Section of Labeling for Human Prescription Drug and Biological Products — Content and Format. Available at: https://www.fda.gov/media/71836/download.

US Food and Drug Administration (2011b). Warnings and Precautions, Contraindications, and Boxed Warning Sections of Labeling for Human Prescription Drug and Biological Products — Content and Format. Available at: https://www.fda.gov/media/71866/download.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). *Attention Is All You Need*. arXiv :1706.03762v5 [cs.CL].

Watson, K. T., and Barash, P. G. (2009). The New Food and Drug Administration Drug Package Insert: Implications for Patient Safety and Clinical Care. *Anesth. Analgesia* 108, 211–218. doi:10.1213/ane.0b013e31818c1b27

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). *Transformers: State-Of-The-Art Natural Language Processing*. arXiv :1910.03771v5 [cs.CL].

Wu, L., Ingle, T., Liu, Z., Zhao-Wong, A., Harris, S., Thakkar, S., et al. (2019). Study of Serious Adverse Drug Reactions Using FDA-Approved Drug Labeling and MedDRA. *BMC Bioinformatics* 20, 97. doi:10.1186/s12859-019-2628-5

Xu, H., Liu, B., Shu, L., and Yu, P. S. (2019). *BERT Post-Training for Review Reading Comprehension and Aspect-Based Sentiment Analysis*. arXiv :1904.02232v2 [cs.CL].

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. arXiv :1906.08237 [cs.CL].

Yu, K.-H., Beam, A. L., and Kohane, I. S. (2018). Artificial Intelligence in Healthcare. *Nat. Biomed. Eng.* 2, 719–731. doi:10.1038/s41551-018-0305-z

Author Disclaimer: The views presented in this article do not necessarily reflect those of the United States Food and Drug Administration. Any mention of commercial products is for clarification and is not intended as an endorsement.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, orclaim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

DATA REPORT
published: 09 December 2021
doi: 10.3389/fgene.2021.750568

A Data Report on the Curation and Development of a Database of Genes for Acute Respiratory Distress Syndrome

*Erick Quintanilla, Kimberly Diwa, Ashley Nguyen, Lavang Vu and Inimary T. Toby**

University of Dallas, Department of Biology, Irving, TX, United States

Keywords: acute respiratory distress (ARDS), genes, variants, database (DB), chromosome location, biological insights and machine learning

INTRODUCTION

Acute respiratory distress syndrome (ARDS) is a syndrome of hypoxic respiratory failure characterized by diffuse pulmonary infiltrates and accumulation of protein-rich pulmonary edema that cause reduction in lung compliance alveolar collapse and ventilation-perfusion mismatch (Katzenstein et al., 1976; Ware and Matthay, 2000; Rubenfeld et al., 2005; Matute-Bello et al., 2008; Phua et al., 2009; Force et al., 2012). ARDS affects approximately 190,600 patients per year in the United States, with mortality up to 45% (Wellman et al., 2016). Despite improvements in intensive care during the last fifteen years, ARDS is still the major cause of mortality and morbidity in intensive care (Katzenstein et al., 1976; Ware and Matthay, 2000; Matute-Bello et al., 2008; Force et al., 2012; Wellman et al., 2016). In fact, ARDS therapy has seen limited progress since its initial description in 1967 and management is still largely supportive, with no established therapies targeted at the primary disease processes (Ashbaugh et al., 1967). Accordingly, there is a need for methods of early detection (Janz and Ware, 2013). There has been recent recognition of the clinical and biological heterogeneity within ARDS (Dowdy et al., 2006; Sweeney et al., 2018; Yehya et al., 2019) that reflects our incomplete understanding of the biology of ARDS.

Acute Respiratory Distress Syndrome (ARDS) is an illness that typically develops in people who are significantly ill or have serious injuries. Within a few hours, patients with ARDS will develop severe shortness of breath, low blood pressure, and unusually rapid breathing (Mayo Clinic, 2020). ARDS is characterized by fluid build-up that occurs in the alveoli of the lungs. The buildup of fluid prevents the lungs from filling up with air which results in less oxygen reaching the bloodstream (Katzenstein et al., 1976; Matute-Bello et al., 2008; Johns Hopkins Medicine,). The lack of sufficient oxygen explains why patients with ARDS are placed on supplemental oxygen for milder symptoms while severe cases are placed in a mechanical ventilation system. ARDS is also a systemic inflammatory disease which suggests that while it is typically found to affect the respiratory system, it tends to affect other organ systems as well. The risk of death from ARDS increases with age and severity of illness while those that survive may experience lasting damage to their lungs (Ashbaugh et al., 1967; Force et al., 2012). The most common cause of ARDS is sepsis. Sepsis is characterized by a serious and widespread infection of the bloodstream. Another common cause of ARDS is severe pneumonia. A more recent cause of ARDS are patients that develop a severe case of COVID-19. These types of cases where patients develop ARDS can often be fatal, and those that do survive and recover from ARDS may have lasting pulmonary scarring (Ware and Matthay, 2000; Rubenfeld et al., 2005; Johns Hopkins Medicine,).

Additional contributions to the knowledge about inheritance of ARDS and/or pathogenesis will be of great benefit in moving forward with successful clinical translation of new diagnostic,

OPEN ACCESS

Edited by:
Haiquan Li,
University of Arizona, United States

Reviewed by:
Tong Zhou,
University of Nevada, Reno, United States
Arun Upadhyay,
Feinberg School of Medicine, Northwestern University, United States

***Correspondence:**
Inimary T. Toby
itoby@udallas.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *30 July 2021*
Accepted: *17 November 2021*
Published: *09 December 2021*

Citation:
Quintanilla E, Diwa K, Nguyen A, Vu L and Toby IT (2021) A Data Report on the Curation and Development of a Database of Genes for Acute Respiratory Distress Syndrome. Front. Genet. 12:750568. doi: 10.3389/fgene.2021.750568

preventive, and therapeutic strategies (Vincent et al., 2006; Constantin et al., 2010; Tejera et al., 2012; Chiumello and Marino, 2017).

The NIH-NHLBI ARDS Network was a research network formed to study treatment of Acute Respiratory Distress Syndrome in 1994. The goal of the Network was to efficiently test promising agents, devices, or management strategies to improve the care of patients with ARDS. During its 20 years of service, 5,527 patients were enrolled in 10 randomized controlled trials and one observational study. Additional trials informed best practices by suggesting no role for routine use of corticosteroids, beta agonists, pulmonary artery catheterization, or early full calorie enteral nutrition. The ARDS Network also developed new outcome measures (ventilator free days) and promoted innovative and efficient techniques (factorial designs and coenrollment) to speed the discovery of new treatment approaches for patients with ARDS (ARDS Network,). This network provided a robust amount of specimen for research experiments and has enabled the research community access to request these samples for secondary analysis.

NCBI GEO contains ~222 ARDS patient samples from high throughput sequencing experiments, some of which utilize specimen derived as part of the ARDS network project. NCBI GEO serves as a resource to support the deposition of datasets from multiple sequencing platform options and accommodates a variety of sample groupings and associated metadata (National Center for Biotechnology Information,). Another NCBI resource, dbGAP, as of the time for this report contained 2 published datasets from sequencing studies done in ARDS. Both GEO and dbGAP do not provide a direct output file containing primary level curated genes, gene functions, chromosomal tags, reference paper id and associated variants from published ARDS studies. The process of extraction of these types of gene lists from external resources and the data parsing required for secondary analysis and follow up computational work is often cumbersome and requires sophisticated Bioinformatics approaches. Here we present, ARDS DB, a comprehensive database for genes and variants specifically related to ARDS. The ARDS DB framework provides gene and variant information and associated metadata derived from primary level curation of experimentally verified studies. The caveat of a dedicated gene database for deeper analysis of ARDS is that it provides the user with a centralized location to retrieve pertinent information. ARDS DB is freely available via an open-source repository and represents a major step towards filling a gap in computational resources for bench biologists and clinicians.

MATERIALS AND METHODS

The data extraction process for the development of version 1 ARDS DB began in June of 2020 and consisted of 2 phases (**Figure 1**). The first phase consisted of retrieval of the related information obtained from 222 samples deposited at NCBI GEO with their associated papers published in Pubmed resource. Next, during the second phase a relevance text mining algorithm was employed via PubMed. The algorithm was based on the standard PubMed Best Match sort using a weighted term frequency algorithm. This approach calculates the frequency with which terms, in this case, Acute Respiratory Distress Syndrome, appear in PubMed records. Those frequencies are then applied in a weighted fashion to return a ranked list of PubMed citations that match the query terms.

An updated feature of the algorithm includes machine learning to re-rank the top articles returned. This algorithm combines over 150 signals that are helpful for finding best matching results. Most of these signals are computed from the number of matches between the search terms and the PubMed record, while others are either specific to a record (e.g., publication type; publication year) or specific to a search (e.g., search length). The new ranking model was built on relevance data obtained from anonymous PubMed search logs that were aggregated over an extended period of time (pubmed reference). The data was filtered by search term and species to include only those results pertaining to humans. Of the 202 articles identified, a manual curation process was employed to extract gene and/or variant lists. A comprehensive literature review was built into the process to compile gene lists. This search strategy was repeated across all matching articles. The relevant files for each gene sets or variants lists were extracted and further parsed using statistical analysis.

Statistical assessments were performed for each of these extracted lists using the R Bioconductor package (R Core Team, 2017). The criteria applied for the statistical evaluation was ≥1.5 fold change and $p \leq 0.05$. Genes and variants found to be statistically significant were included in the final criteria. For each gene that was listed in the database, the corresponding PMID of the research article was included. www.genecards.org was used to obtain the official gene name as well as any other corresponding alias name for the gene (GeneCards,). Both the gene name from the published study and all other alias names for that gene were included in the database. Using the official gene name, the chromosome and position of the gene were also included as part of the metadata assembled. The gene names were then verified by direct import into "NCBI's genome data viewer" and conducting a search for the gene. A summary of each gene's function, as well as its chromosomal location and the start and stop site on the chromosome were documented. DAVID analysis was further employed to extract detailed gene description (DAVID,). For analysis within DAVID, the "Entrez_gene_ID" option was selected as the identifier name. The gene lists submitted were converted into official gene symbol as part of the curation tasks.

The gene's relation to ARDS patient outcomes is indicated in the database as provided in the published study. The two categories for patient outcomes included in ARDS DB are increased susceptibility or mortality which was found to be associated with the differentially expressed genes. Some genes were reported to cause both increased susceptibility and mortality, which are indicated within the ARDS DB. In addition, the gene function in ARDS, pathogenesis related information was included if pertinent to the primary research article. Permanent digital object identifiers (DOIs) for the original research articles are provided for each entry. The database was

designed using a structured query language (SQL) architecture and is publicly available via the Zenodo open source ecosystem. https://zenodo.org/record/4033491#.YQN_cY5Khyw.

DATASET

ARDS-DB contains a total of 238 genes that were found to be differentially expressed in ARDS patients. It contains the following types of metadata: official gene symbol, as well as any alias names that the gene could be associated with; NCBI gene ID, Chromosomal location, start and stop sites, variant id where relevant, as well as its corresponding location and type, which is listed in the database. The reference and primary publication where each gene was found are listed in the database, as well as a summary of the gene function. Lastly, the association between the gene and patient outcomes is provided where pertinent as well as a summary of the relatedness of the gene to ARDS patients. The corresponding reference containing this information is provided.

A major strength of ARDS DB is that it has been created through a pipeline consisting of intensive manual curation efforts, combined with machine learning algorithms. The synergy of these 2 approaches will ensure ease of continuous update as new data is deposited. Another strength is that the resource conveys specificity for ARDS and will help researchers looking for a centralized location to search for genes and variants. The detailed information on chromosomal location allows for ease of searching against any novel variants being assessed as comparison. The gene function information provided enables the user to learn quick facts about the gene and its role in signaling processes. The inclusion of patient outcomes provides clinicians quick reference information that will be informative to place the gene or variant in context for further consideration.

Currently the database is provided in a downloadable SQL format, which requires the user to download and compile it locally using a SQL-based interface. To address this, our future plan is to migrate ARDS DB into a stand-alone web-based resource. We would like to provide a web interface with easy access for bench biologists and clinicians that will offer advanced search features as well as data analysis and visualization all within the same ecosystem. With the availability of ARDS DB, users will be able to categorize and further understand the gene relationships involved in ARDS and the associated variants from published studies. The availability of variant locations will facilitate the direct comparison with novel variants or unique cases of familial ARDS such as that reported recently (Toby et al., 2020). An additional use for the database is to identify genes for training set to help build machine learning (ML) models to elucidate variations in ARDS patient outcomes. ML based assessments (such as Clustering algorithms, Random forest algorithms) and methods to include specialized sequence data such as from RNA seq and specialized sequencing technologies will be of particular focus. Potential associations of whole genome data to more specific patient cohorts for clinicians to better understand cases of familial ARDS will be of importance in future work.

DATASET DESCRIPTION

The database is freely available in Zenodo and can be accessed through the following link: https://zenodo.org/record/4033491#.YPnI5BNKhQI or by searching within Zenodo for the following title: Acute Respiratory Distress Syndrome-Database of Genes (ARDS-DB). ARDS-DB is

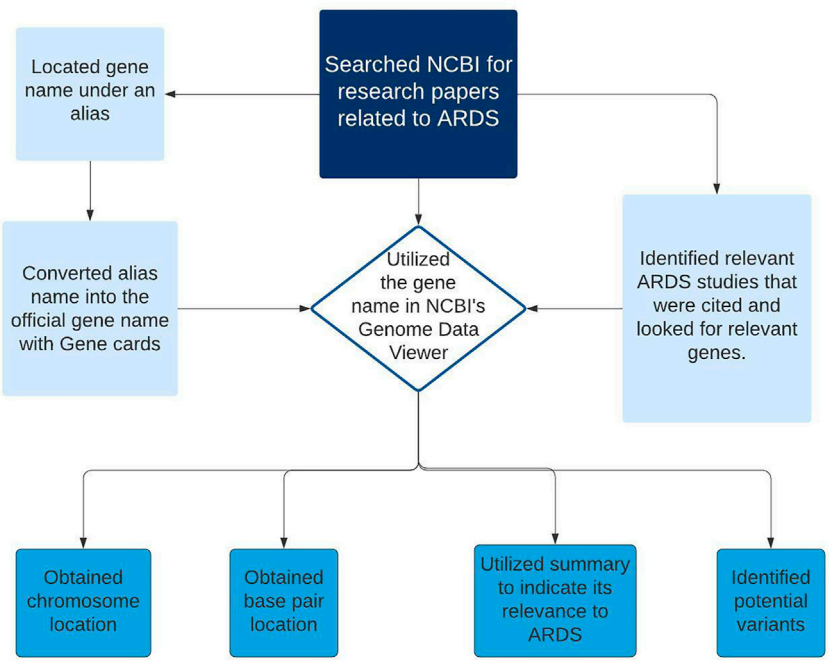


FIGURE 1 | Diagram showing the workflow steps (data collection and primary curation) for the development of ARDS DB.

accessible via user download and can be viewed using a SQL-based interface such as MySQL (https://www.mysql.com/downloads/) or SQL-lite browser (https://sqlitebrowser.org/).

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ Supplementary Material.

ETHICS STATEMENT

The study did not involve human participants or animals. The study consisted of secondary analysis work. An IRB protocol was obtained for doing this work.

AUTHOR CONTRIBUTIONS

EQ- contributed to the primary curation and data collection, development of the database, article preparation and lead role on content organization. KD- contributed to the primary curation and data collection and ML experiments. AN- contributed to the primary curation and data collection and ML experiments. LV- contributed to the primary curation and data collection. IT- contributed to the project design, data search strategies, data description, article preparation and overall supervision of the project.

FUNDING

This work is funded by a grant from The Nancy Cain and Jeffrey A. Marcus Science Endowment in Honor of President Donald A. Cowan (University of Dallas, PI: Inimary Toby)

REFERENCES

ARDS Network. The NHLBI ARDS Network. Available at: http://www.ardsnet.org/ Accessed: July 1, 2020.

Ashbaugh, D., Boyd Bigelow, D., Petty, T., and Levine, B. (1967). Acute Respiratory Distress in Adults. *Lancet* 290 (7511), 319–323. doi:10.1016/s0140-6736(67)90168-7

Chiumello, D., and Marino, A. (2017). ARDS Onset Time and Prognosis: Is it a Turtle and Rabbit Race? *J. Thorac. Dis.* 9 (4), 973–975. doi:10.21037/jtd.2017.03.147

Constantin, J.-M., Grasso, S., Chanques, G., Aufort, S., Futier, E., Sebbane, M., et al. (2010). Lung Morphology Predicts Response to Recruitment Maneuver in Patients with Acute Respiratory Distress Syndrome. *Crit. Care Med.* 38 (4), 1108–1117. doi:10.1097/ccm.0b013e3181d451ec

DAVID. Functional Annotation Tools. Available at: david.ncifcrf.gov/tools.jsp

Dowdy, D. W., Eid, M. P., Dennison, C. R., Mendez-Tellez, P. A., Herridge, M. S., Guallar, E., et al. (2006). Quality of Life after Acute Respiratory Distress Syndrome: a Meta-Analysis. *Intensive Care Med.* 32 (8), 1115–1124. doi:10.1007/s00134-006-0217-3

Force, A. D. T., Ranieri, V. M., Rubenfeld, G. D., Thompson, B. T., Ferguson, N. D., Caldwell, E., et al. (2012). Acute Respiratory Distress Syndrome: the Berlin Definition. *JAMA* 307 (23), 2526–2533. doi:10.1001/jama.2012.5669

GeneCards. GeneCards®: The Human Gene Database. Available at: www.genecards.org/ Accessed: July 1, 2020.

Janz, D. R., and Ware, L. B. (2013). The Needle in the Haystack: Searching for Biomarkers in Acute Respiratory Distress Syndrome. *Crit. Care* 17 (5), 192. doi:10.1186/cc13025

Johns Hopkins Medicine. COVID-19 Lung Damage. Available at: www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/what-coronavirus-does-to-the-lungs.

Katzenstein, A. L., Bloor, C. M., and Leibow, A. A. (1976). Diffuse Alveolar Damage-Tthe Role of Oxygen, Shock, and Related Factors. A Review. *Am. J. Pathol.* 85 (1), 209–228.

Matute-Bello, G., Frevert, C. W., and Martin, T. R. (2008). Animal Models of Acute Lung Injury. *Am. J. Physiol. Lung Cell Mol. Physiol.* 295 (3), L379–L399. doi:10.1152/ajplung.00010.2008

Mayo Clinic (2020). ARDS. Mayo Foundation for Medical Education and Research. Available at: www.mayoclinic.org/diseases-conditions/ards/symptoms-causes/syc-20355576 Accessed: June 13, 2020.

National Center for Biotechnology Information. National Center for Biotechnology Information. U.S. National Library of Medicine. Available at: www.ncbi.nlm.nih.gov/.

Phua, J., Badia, J. R., Adhikari, N. K. J., Friedrich, J. O., Fowler, R. A., Singh, J. M., et al. (2009). Has Mortality from Acute Respiratory Distress Syndrome Decreased over Time? *Am. J. Respir. Crit. Care Med.* 179 (3), 220–227. doi:10.1164/rccm.200805-722oc

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Rubenfeld, G. D., Caldwell, E., Peabody, E., Weaver, J., Martin, D. P., Neff, M., et al. (2005). Incidence and Outcomes of Acute Lung Injury. *N. Engl. J. Med.* 353 (16), 1685–1693. doi:10.1056/nejmoa050333

Sweeney, T. E., Thomas, N. J., Howrylak, J. A., Wong, H. R., Rogers, A. J., and Khatri, P. (2018). Multicohort Analysis of Whole-Blood Gene Expression Data Does Not Form a Robust Diagnostic for Acute Respiratory Distress Syndrome. *Crit. Care Med.* 46 (2), 244–251. doi:10.1097/ccm.0000000000002839

Tejera, P., Meyer, N. J., Chen, F., Feng, R., Zhao, Y., O'Mahony, D. S., et al. (2012). Distinct and Replicable Genetic Risk Factors for Acute Respiratory Distress Syndrome of Pulmonary or Extrapulmonary Origin. *J. Med. Genet.* 49 (11), 671–680. doi:10.1136/jmedgenet-2012-100972

Toby, I. T., Thomas, N. J., Thorenoor, N., Spear, D., DiAngelo, S., and Floros, J. (2020). Characterizing a Focused Landscape of Familial Acute Respiratory Distress Syndrome. *Biomark Applic.* 04, 141. doi:10.29011/2576-9588.100041

Vincent, J.-L., Sakr, Y., Sprung, C. L., Ranieri, V. M., Reinhart, K., Gerlach, H., et al. (2006). Sepsis in European Intensive Care Units: Results of the SOAP Study. *Crit. Care Med.* 34 (2), 344–353. doi:10.1097/01.ccm.0000194725.48928.3a

Ware, L. B., and Matthay, M. A. (2000). The Acute Respiratory Distress Syndrome. *N. Engl. J. Med.* 342 (18), 1334–1349. doi:10.1056/nejm200005043421806

Wellman, T. J., de Prost, N., Tucci, M., Winkler, T., Baron, R. M., Filipczak, P., et al. (2016). Lung Metabolic Activation as an Early Biomarker of Acute Respiratory Distress Syndrome and Local Gene Expression Heterogeneity. *Anesthesiology* 125 (5), 992–1004. doi:10.1097/aln.0000000000001334

Yehya, N., Thomas, N. J., and Wong, H. R. (2019). Evidence of Endotypes in Pediatric Acute Hypoxemic Respiratory Failure Caused by Sepsis. *Pediatr. Crit. Care Med.* 20 (2), 110–112. doi:10.1097/pcc.0000000000001808

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership