RatsPub: a webservice aided by deep learning to mine PubMed for addiction-related genes

Mustafa Hakan Gunturkun[1], Efraim Flashner[2], Tengfei Wang[1], Megan K. Mulligan[2], Robert W. Williams[2], Pjotr Prins[2], Hao Chen[1,*]

1. Department of Pharmacology, Addiction Science and Toxicology, University of Tennessee Health Science, Memphis, TN 38103 USA
2. Department of Genetics, Genomics and Informatics, University of Tennessee Health Science, Memphis, TN 38103 USA

* Corresponding Author

Abstract

Interpreting and integrating results from omics studies on addiction phenotypes require a comprehensive survey of the extant literature. Most often this is conducted by *ad hoc* queries of the PubMed database. Here, we introduce RatsPub, a literature mining web service that searches user-provided gene symbols in conjunction with a set of systematically curated keywords related to addiction, as well as results from human genome-wide association studies (GWAS). We have organized over 300 keywords into six categories forming an ontology. The literature search is conducted by querying the NIH PubMed server using a programmatic interface. Abstracts are retrieved from a local copy of the PubMed archive. The main results presented to the user are sentences containing the gene symbol and at least one keyword. These sentences are presented in the browser through an interactive graphical interface or using tables. Results are linked to the source PubMed records. GWAS results are displayed using a similar method. We wrote a natural language processing module that uses deep convolutional neural networks to distinguish sentences describing systemic stress vs cellular stress. The automated and comprehensive search strategy provided by RatsPub facilitates the integration of new discoveries from addiction omic studies with existing literature and improves analysis and modeling in the field of addiction biology. RatsPub is free and open source software. The source code of RatsPub and the link to a running instance is available at https://githhub.com/chen42/ratspub.

Keywords: addiction, literature mining, PubMed, web service

1. Introduction

We describe a web service and application—*Relationships to Addiction Through Searches of PubMed* (RatsPub) (http://rats.pub) (RRID: SCR_018905**)**—that automatically extracts relevant information from PubMed and NHGRI-EBI GWAS catalog on the relationship of any gene with several known biological processes related to addiction. To this end, we created an ontology by identifying six key categories of concepts related to addiction. We created a list of keywords for each concept. User supplied genes are paired with these keywords to search NIH PubMed. Relevant abstracts are then retrieved from a local PubMed database. We extract sentences that contain both the genes and keywords and present them to the user in a graphical or tabular format. In addition, we also query the GWAS catalog to retrieve addiction related associations.

Omic studies are becoming the main driving force for discovering molecular mechanisms of substance abuse. For example, genome-wide association studies (GWAS) have become the main platform of discovery on genetic variants responsible for phenotypes related to substance abuse disorders. One recent human GWAS identified over 500 variants associated with smoking and alcohol usage related traits [1]. A GWAS on alcohol use disorder identified 18 genome-wide significant loci [2]. GWAS on opioid [3] or cocaine use disorders [4] have also been conducted or are ongoing.

Specialized databases, such as the GWAS catalog [5], are available for directly searching the genetic variant – phenotype associations.

Model organisms, ranging from mammals to flies to worms, have all been used to investigate mechanisms associated with drug abuse related phenotypes. Genetic mapping studies have been conducted on cocaine, opioids, nicotine, alcohol, etc. using these species [6–9]. Because of the relatively small sample size and limited genetic diversity, a trade off of these studies is that the quantitative trait locus (QTL) often contains multiple genes.

Even more results have been reported on the changes at the transcriptome or epigenome level. Typically, these studies identify genes affected by either acute or chronic exposure to abused substances.

In these omics studies, understanding the role of genes in addiction is a challenging task that requires thorough integration of existing knowledge. Statistics driven gene ontology, or pathway analysis, are often employed for this purpose. However, extensive review of the primary literature is ultimately needed to provide a comprehensive and nuanced mechanism. For many scientists, this starts as simple searches of PubMed based on their domain knowledge. Unfortunately, these ad hoc searches often miss important information because of the inherent complexity of the biology of addiction and the amount of time required for conducting these searches, tracking the results, filtering the abstracts, and reading them. The task of literature searches is especially daunting when more and more genes are beginning to be identified in a single study.

We rely mostly on keyword matching to select relevant sentences. However, sometimes the same keyword can have multiple meanings. For example, stress promotes initial drug use, escalates continued drug use, precipitates relapse and is a major factor contributing to drug addiction [10]. Stress in this context refers to the body's response to internal and external challenges and is mediated by activating the hypothalamic–pituitary–adrenal axis. In addition, stress can also refer to the responses of cells to perturbations of their environment, such as extreme temperature, mechanical damage, or accumulation of metabolites, etc. These responses often involve the activation of specific molecular pathways. Both systemic and cellular stress have a large collection of literature. Cellular stress is much less relevant to addiction compared to systemic stress. Displaying stress-related sentences as two separate groups will improve the user experience. We therefore developed a deep learning model, specifically, a convolutional neural network, to separate sentences describing cellular stress from those that describe system stress. By providing both a web-interface and the source code, we hope this application will be both easy to use, expand, and adaptable for other similar applications.

2. Methods

2.1 System Overview

RatsPub is a free and open source web application (Fig. 1). The source code and URL of a running instance is available at http://github.com/chen42/ratspub. The main user interface contains a search box that accepts up to 30 gene symbols from the user. Each gene symbol is then paired with each one of the six mini ontologies to query PubMed. Record identifiers (i.e. PMIDs) are then retrieved. The title and abstract of these records are then obtained from a mirrored copy of PubMed on the local server. Sentences containing at least one gene symbol and one keyword are retained. A local copy of NHGRI-EBI GWAS catalog is also searched for associations between the queried genes and addiction or psychiatric disease related phenotypes. The PubMed and GWAS catalog query results are then combined and presented to the user as an interactive graph. Alternatively, these results are also available as a table. Both formats provide a summary of the gene-keyword relationships and for the users to review the original sentences, which are linked to PubMed.

Convolutional neural networks are one of the deep learning methods which was initially designed for two dimensional image processing [11]. They use a linear operation called convolution besides the regular neural network components, and explore the important patterns in a data by identifying both local and global features of the data. Here, we implement a one dimensional convolutional neural network to classify sentences describing stress to either cellular stress or system stress.

2.2. Sources of data: PubMed and GWAS catalog

We created a copy of the entire PubMed abstract on our server following instructions provided by the NCBI [12]. This allows us to rapidly retrieve the abstracts and bypass the limits imposed by NCBI on automated retrievals to prevent system overload. This local copy is updated automatically every week on our server.

We also store a local copy of the GWAS catalog database (i.e. a tab separated text file). This file is updated manually upon every new release of the catalog. This allows us to perform customized and rapid queries of addiction and psychiatric phenotypes.

2.3. Mini-ontology for addiction related concepts

We created a mini-ontology for addiction related concepts (Table S1). This ontology has three levels. The top level has the following six categories: addiction, drugs, brain, stress, psychiatric diseases and molecular function. The second level is composed of relevant keywords and the third level includes subconcepts of the keywords or commonly used spelling or acronyms for the keywords. We include all the third level words in our automated search operations and present the user aggregated summary at the second level in the graphic or tabular result sections. The matching keywords at the third level are highlighted using bold font when the sentences are displayed. Users have the option to skip any category to speed up the query.

2.4. Query processing and user interface

We wrote the web-service in the Python programming language and used the Flask library as the web application framework [13]. Users of the web service have the option of creating an account for the purpose of saving search results for later reviews. Query terms provided by the user are first paired with all the keywords. Keywords belonging to the same second level ontology terms are combined using the boolean OR operator before joining with the gene symbol using the AND operator. The E-utilities provided by the NCBI Entrez system [12] are used to send the query to the PubMed database (Esearch) and to retrieve PMIDs (Efetch). Corresponding records for each PMID is obtained from the local copy of PubMed and the xtract tool is used to parse the titles and abstracts. The Python NLTK library [14] is then used to tokenize the abstracts into sentences. Python regular expressions are used to find sentences that contain at least one instance of a query gene and one instance of a keyword. The number of *abstracts* containing such sentences are then counted. The gene is also searched in the GWAS catalog for associations with drug abuse and psychiatric disease. The number of associations are also counted. A network graph is constructed using the Cytoscape Javascript library [15], where all genes, keywords, and GWAS terms are used as nodes, and a connection is made between nodes describing a gene and a keyword. The number of abstracts are used as the weight of the edge. This interactive graph allows a user to click on the edge to review the corresponding sentences. All sentences are linked to their original PubMed abstract. The user can also click on a keyword to see the synonyms used and launch searches using these synonyms, or click on a gene to query its relationship with the top 100 addiction related genes (described below in the results section). A table view of the same content is also available.

Queries can also be initiated by placing the terms in the URL. For example, to start a search for chrna5 and BDNF genes against the keyword categories drug, stress, addiction, and GWAS, the following hyperlink can be used:

http://rats.pub/progress?type=drug&type=stress&type=GWAS&type=addiction&query=CHRNA5+RGMA

This allows RatsPub to be embedded directly into other webservices.

The RatsPub source code is distributed as free and open source software and can therefore easily be installed on other systems. The whole service with dependencies is described as a byte reproducible GNU Guix software package [16]. As a convenience, through this package description, we are providing a Docker container to run software locally which can be found through our website.

2.5. Convolutional neural network to classify sentences describing stress

To differentiate sentences describing systemic stress vs cellular stress, we developed an artificial neural network to conduct a binary classification (Fig. 2). To create a training corpus, we used a word2vec embeddings library based on PubMed and PubMedCentral data [17] to retrieve words that are similar to examples of systemic stress and cellular stress (e.g, restraint, corticosterone, CRH, and oxidative stress respectively). We then manually crafted two PubMed queries to retrieve abstracts related to systemic or cellular stress:

A. (CRF OR AVP OR urocortin OR vasopressin OR CRH OR restraint OR stressor OR tail-shock OR (social AND defeat) OR (foot AND shock) OR immobilization OR (predator AND odor) OR intruder OR unescapable OR inescapable OR CORT OR corticosterone OR cortisol or ACTH OR prolactin OR PRL OR adrenocorticotropin OR adrenocorticotrophin) AND stress NOT (ROS OR oxidative OR redox-regulation OR nitrosative OR nitrative OR hyperglycemia OR carbonyl OR lipoxidative OR Nrf2-driven OR thiol-oxidative)

B. (ROS OR oxidative OR redox-regulation OR nitrosative OR nitrative OR hyperglycemia OR carbonyl OR lipoxidative OR Nrf2-driven OR thiol-oxidative) AND stress

We downloaded all the PubMed abstracts returned from these two queries. Manually examining some of the abstracts confirmed the relevance of the results. We then extracted all sentences containing the word *stress* from each set and kept 9,974 sentences from the "systemic stress" class and 9,652 sentences from time "cellular stress" class as our stress training/validation corpus. We maintained another set of 10,000 sentences as the testing corpus, 5,000 sentences for each class.

To clean the data and make it ready for deep learning, we split 19,626 sentences into words, removed punctuation marks, filtered the stop words and stemmed the words [18]. The words formed a vocabulary of size 23,153 and were tokenized by the Tokenizer library of Keras API. Then the tokenized sentences were split randomly into training and validation sets at 80% and 20%, respectively. We built a 1D convolutional neural network (CNN) in Keras on top of the Tensorflow framework [19]. The model includes an embedding layer that projects each word to a 32 dimensional space hence this layer produces a weight matrix with 23,153 x 32 dimensions. Sentences are padded to 64 words, resulting in 64x32 sized matrices in the model. After that, a one dimensional convolutional layer with 16 filters and a kernel size of 4 is implemented and activated by the rectified linear unit (ReLU). This layer produces a 4 x 32 x 16 weight matrix. Downsampling is performed by max pooling with a window size of 2. Then a flattened layer with 480 neurons is connected to two fully connected layers, one of which has 10 neurons activated with ReLU and the latter one is the final layer activated with a sigmoid function. We validate the model using 3,924 sentences, 1,997 of them belong to the "systemic stress" class, 1,927 sentences belong to the "cellular stress" class. These were selected randomly before training. To minimize the value of the loss function and update the parameters, Adamax optimization algorithm [20] was used with the parameters of learning rate=0.002, beta1=0.9, beta2=0.999. We measured the performance of the model by binary cross-entropy loss.

3. Results

We have written a command line and a graphical interface for searching the role genes play in addiction. The command line interface is more suitable for searching a large number of genes and requires the user to install the software and maintain a local mirror of PubMed. We used the command line interface to search all human genes. We first retrieved all (61,636) human genes from the NCBI gene website [21]. We then parsed the gene symbols and aliases and counted the total number of abstracts for each gene using E-Utils. Relevant sentences for all genes with more than 50 abstracts are then retrieved. We manually examined these results and removed 988 words from the list of gene symbols and aliases. The top 10 genes with the greatest number of addiction related abstracts are FOS, BDNF, TH, OPRM1, CNR1, DRD2, CREB1, SLC6A4, TNF and CYP2B6. Many of these genes are involved in the activation of neurons or neurotransmission. Because intracellular signaling is one of the categories, the list of top 100 genes also included some genes that primarily are known for their role in immune system function, such as TNF and IL6. The top genes are provided in Table S2. These genes and their associated sentences are available at the http://rats.pub website.

On the other hand, the graphical interface is more user friendly and can be used through our website. As a demonstration of the utility of the web interface, we entered the nine genes that reached suggestive significance in a recent genome wide association study of opioid cessation [22]. The graph view of the search results are shown in Figure 3. Genes and keywords are all shown as circles and lines connecting them show the number of abstracts containing the two circles they connect. Clicking on these lines brings up a new page that displays all sentences containing the words that the lines connect. An alternative tabular view of the same results is also available, where genes, addiction keywords, and number of abstracts are shown as separate columns. In addition, clicking gene names will launch a new search for sentences containing the target gene and 100 addiction-related genes.

Our results identified roles played by PTPRD, SNAP25 and MYOM2 in addiction, which were all discussed in the original publication [22]. In addition, our results found reports indicated the potential involvement of RIT2 and SYT4 in addiction. For example, RIT2 is associated with smoking initiation [1] and autism [23]. Recent publications indicated that RIT2 is involved in dopamine transporter trafficking [24] and plays a sex-specific role in acute cocaine response [25]. SYT4 is expressed in the hippocampus and entorhinal cortex [26] and regulates synaptic growth [27,28]. Further, SUCLA2P2 has been implicated in age of smoking initiation [29] and Schizophrenia [30]. This example demonstrated the utility of RatsPub in rapidly finding information that links a gene to addiction.

We designed a one-dimensional convolutional neural network with 4 hidden layers (Fig. 2) to differentiate two classes of sentences related to stress, namely systemic and cellular stress. We used hand crafted boolean queries of PubMed to obtain our training and validation corpus (approximately 8,000 sentences for training and 2,000 for validation for each class). These selections of keywords used in the queries were informed by an existing word2vec model [17] as described in the methods section.

The gradient based optimization algorithm Adamax is used to optimize the loss function with a learning rate of 0.002. During training, model accuracy (Fig. S1.A) increased rapidly during the first five epochs to approximately 0.995, while validation accuracy peaked at 0.991 at epoch five. On the other hand, model loss curve (Fig. S1.B) on the training dataset continued to decline after the initial drop and approached zero after 15 epochs. However, the loss on the validation data set started to increase after epoch five, indication model overfitting. Therefore, we used the weights that maximized the validation performance before overfitting (i.e. epoch five). By using these weights and parameters, our model has an AUC of 99.2% on the validation dataset.

We tested the model on a new dataset consisting of 5,000 system stress sentences and 5,000 cellular stress sentences. In order to evaluate the performance of the classification and summarize the results, we computed the confusion table for the test dataset (Table 1). The number of true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN) and the performance scores derived from these measures are computed in the table. The sensitivity of the model that is the proportion of predicted systemic class sentences to all sentences observed in this class is 97%. The similar measure for cellular class sentences, i.e., specificity is 94%. The prediction accuracy of the model that is the ability to distinguish two classes on the test dataset is 95.4% and the AUC is 98.9%.

We also checked the distribution of the predicted probabilities (Fig. S2) of the test dataset. The model predicts a probability of the class membership for each sentence. If the predicted probability of a sentence is more than 0.5, it is labelled as a system stress sentence. Otherwise the sentence is predicted to be a member of the cellular stress class. Among the system stress sentences in the test dataset, 88% of the sentences had predicted probabilities greater than 0.9. This shows the model's confidence of its prediction on stress sentences. Likewise, 88% of the cellular stress sentences had predicted probabilities less than 0.1. Therefore the model is 90% confident about the classification of the 88% of the cellular stress sentences.

The weights of the trained model are saved on the server and are used to make predictions for each retrieved sentence when the user clicks on the stress category (Fig. 4). As an example of run time performance, it took approximately 12 seconds to classify 3,908 sentences on CRF and stress.

4. Discussion

We present here a literature mining web application, RatsPub, that extracts sentences from a locally mirrored copy of PubMed abstracts containing user provided gene symbols and approximately 300 predefined addiction-related keywords organized into six categories. Associations between the gene symbol and psychiatric diseases, including addiction, from human GWAS results are also provided. The users can include up to 30 gene symbols in each search. The results are presented in a graphical or a tabular format, both provide links to review individual sentences that contain the gene and at least one keyword. Gene synonyms are also presented and can be included in additional searches. Stress related sentences are automatically classified into system vs cellular stress.

Scientists using omics methods to study addiction face a particularly challenging task when they conduct literature searches. Not only is the number of publications coming out every month too large to track for research, also the large number of genes they work with and the breadth of addiction science forms a great mountain of information that can not be easily mined. Typically, scientists manually conduct detailed searches in areas where they have expertise and the queries are much less thorough in other areas. These ad hoc queries are difficult to replicate when multiple genes are involved. RatsPub provides an interface that allows comprehensive queries of the role of any gene using a set of about 300 keywords. These keywords provide a comprehensive coverage of key concepts related to addiction. Although most of the functions provided by RatsPub can be carried out manually, it will require several orders of magnitude more time and effort. Even then, the results will have inevitable misses and will be difficult to review. In contrast, results provided by RatsPub are automatically organized by the mini ontology. All the genes and keywords can be seen in one graph or table, with informative sentences and abstracts readily available. RatsPub is an applied machine learning solution that helps mine the relevant literature and can act as an example for similar research areas. In contrast to manual searches, RatsPub offers a systematically structured search ability at a much improved speed.

RatsPub presents to the user sentences containing genes and keywords of interest to the user. Compared to phrases or abstracts, sentences are the most succinct semantic unit to convey a fact. Ding et al [31] compared different text processing units for text mining system design and found that the highest precision of information retrieval is achieved when phrases are used as the text unit whereas using sentences are more effective than both phrases and abstracts. Therefore, similar to our previous text mining tool [32], we continue to use sentences as the information unit. Unlike the commonly used gene ontology enrichment [33] or gene set enrichment [34] analysis, the literature analysis provided by RatsPub does not evaluate any statistical significance. Instead, these key sentences provide easy access to relevant prior research, where the nuanced detail can be easily obtained by following the link from the sentence to the abstract and then to the full text article.

Stress plays key roles in addiction. Using a convolutional network, we trained a model that achieved 97% sensitivity and 94% specificity in classifying sentences containing the word stress to either systemic stress or cellular stress. Training such a model requires large amounts of labeled data. Manually labeling these data is very labor intensive. Using an approach that is similar to some recent advances in automated data labeling [35], we carefully crafted two PubMed queries to obtain over 30,000 sentences that are mostly belonging to the correct category. This large corpus of text allowed us to achieve peak classification performance with less than 5 epochs of training (Fig. S1).

Gene synonyms represent a large challenge to any text mining approach. Not including synonyms will result in the loss of information. However, many synonyms, especially those that are short, have multiple meanings. For example, CNR is a synonym for the CNR1 gene. However, CNR is also an acronym for contrast noise ratio, frequently used in imaging analysis literature. We manually edited the list of aliases for the top 100 addiction related genes, which are shown in Table S2. For user supplied gene symbols, we do not include synonyms in the initial search to prevent the noise from “drawing out” the signal. However, we do provide users an option to either search individual synonyms or to conduct a combined search of all synonyms as a secondary step. We think this middle-of-the-road approach is the most efficient method to achieve a balance between computation and performance. Future work can potentially use deep learning to classify all PubMed abstracts for their relevance to addiction and thus exclude many abstracts containing short words that are not relevant to addiction from being confounded with gene synonyms.

Other future improvements for RatsPub are possible. For example, RatsPub uses PubMed abstracts as the source of data, rather than PubMed Central, which contains full-text articles. Lin [36] compared the effectiveness of information retrieval from abstract vs full text search and found that full text search, when indexed using paragraphs as the unit, is more effective than the abstract-only search. Several groups have reported either using full text search for curation [37,38] or using full text for analysis [39–41]. NCBI also provides an API for PubMed Central. However, the majority of the articles in PubMed Central are subject to traditional copyright restriction [42] and it is not feasible to establish a local mirror of the full-text collection. Interactively retrieving text via NCBI API is not feasible on the scale we need (e.g, several thousand articles at a time). Further, we anticipate full text may cause duplications of information and increase the noise in results.

Lastly, the mini ontology we use depends on the expertise of the authors and can be further improved by user input. We will also incorporate terms that are incorporated in addiction related ontologies, such as those that are available from the Open Biological and Biomedical Ontology Foundry (www.obofoundry.org).

Acknowledgements

Funding is provided by the following NIH/NIDA grants: U01DA047638 (HC, RWW), P30DA044223 (RWW, PP), and also by NIH/NIGMS grant R01GM123489 (RWW, PP).

Authors contribution

MHG conducted the research and drafted the manuscript. HC conceived of the project and conducted the initial research. EF, TW, MKM, RWW and PP contributed to the research. All authors revised and approved the manuscript.

References

1. Liu M, Jiang Y, Wedow R, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet*. 2019;51(2):237-244.

2. Kranzler HR, Zhou H, Kember RL, et al. Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nat Commun*. 2019;10(1):1499.

3. Polimanti R, Walters RK, Johnson EC, et al. Leveraging genome-wide data to investigate differences between opioid use vs. opioid dependence in 41,176 individuals from the Psychiatric Genomics Consortium. *Mol Psychiatry*. Published online February 26, 2020. doi:10.1038/s41380-020-0677-9

4. Huggett SB, Stallings MC. Cocaine'omics: Genome-wide and transcriptome-wide analyses provide biological insight into cocaine use and dependence. *Addict Biol*. 2020;25(2):e12719.

5. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(D1):D1005-D1012.

6. Adkins AE, Hack LM, Bigdeli TB, et al. Genomewide Association Study of Alcohol Dependence Identifies Risk Loci Altering Ethanol-Response Behaviors in Model

Organisms. *Alcohol Clin Exp Res*. 2017;41(5):911-928.

7. Zhou Z, Blandino P, Yuan Q, et al. Exploratory locomotion, a predictor of addiction vulnerability, is oligogenic in rats selected for this phenotype. *Proc Natl Acad Sci U S A*. 2019;116(26):13107-13115.

8. Highfill CA, Baker BM, Stevens SD, Anholt RRH, Mackay TFC. Genetics of cocaine and methamphetamine consumption and preference in Drosophila melanogaster. *PLoS Genet*. 2019;15(5):e1007834.

9. Engleman EA, Katner SN, Neal-Beliveau BS. Caenorhabditis elegans as a Model to Study the Molecular and Genetic Mechanisms of Drug Addiction. *Prog Mol Biol Transl Sci*. 2016;137:229-252.

10. Koob GF, Schulkin J. Addiction and stress: An allostatic view. *Neurosci Biobehav Rev*. 2019;106:245-262.

11. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278-2324.

12. Kans J. *Entrez Direct: E-Utilities on the UNIX Command Line*. National Center for Biotechnology Information (US); 2020.

13. Flask. Accessed September 9, 2020. https://palletsprojects.com/p/flask/

14. Bird S, Klein E, Loper E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media Inc.; 2009.

15. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-2504.

16. Wurmus R, Uyar B, Osberg B, et al. PiGx: reproducible genomics analysis pipelines with GNU Guix. *Gigascience*. 2018;7(12). doi:10.1093/gigascience/giy123

17. Moen S, Ananiadou TSS. Distributional semantics resources for biomedical text processing. *Proceedings of LBM*. Published online 2013:39-44.

18. Brownlee J. *Deep Learning for Natural Language Processing: Develop Deep Learning Models for Your Natural Language Problems*. Machine Learning Mastery; 2017.

19. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv [csDC]*. Published online March 14, 2016. http://arxiv.org/abs/1603.04467

20. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint*

arXiv:14126980. Published online 2014. http://arxiv.org/abs/1412.6980

21. NCBI. NCBI Gene. Accessed June 14, 2020. https://www.ncbi.nlm.nih.gov/gene/

22. Cox JW, Sherva RM, Lunetta KL, et al. Genome-Wide Association Study of Opioid Cessation. *J Clin Med Res*. 2020;9(1). doi:10.3390/jcm9010180

23. Liu X, Shimada T, Otowa T, et al. Genome-wide Association Study of Autism Spectrum Disorder in the East Asian Populations. *Autism Res*. 2016;9(3):340-349.

24. Fagan RR, Kearney PJ, Sweeney CG, et al. Dopamine transporter trafficking and Rit2 GTPase: Mechanism of action and in vivo impact. *J Biol Chem*. 2020;295(16):5229-5244.

25. Sweeney CG, Kearney PJ, Fagan RR, et al. Conditional, inducible gene silencing in dopamine neurons reveals a sex-specific role for Rit2 GTPase in acute cocaine response and striatal function. *Neuropsychopharmacology*. 2020;45(2):384-393.

26. Crispino M, Stone DJ, Wei M, et al. Variations of synaptotagmin I, synaptotagmin IV, and synaptophysin mRNA levels in rat hippocampus during the estrous cycle. *Exp Neurol*. 1999;159(2):574-583.

27. Harris KP, Zhang YV, Piccioli ZD, Perrimon N, Littleton JT. The postsynaptic t-SNARE Syntaxin 4 controls traffic of Neuroligin 1 and Synaptotagmin 4 to regulate retrograde signaling. *Elife*. 2016;5. doi:10.7554/eLife.13881

28. Ó'Léime CS, Hoban AE, Hueston CM, et al. The orphan nuclear receptor TLX regulates hippocampal transcriptome changes induced by IL-1β. *Brain Behav Immun*. 2018;70:268-279.

29. Argos M, Tong L, Pierce BL, et al. Genome-wide association study of smoking behaviours among Bangladeshi adults. *J Med Genet*. 2014;51(5):327-333.

30. Ikeda M, Takahashi A, Kamatani Y, et al. Genome-Wide Association Study Detected Novel Susceptibility Genes for Schizophrenia and Shared Trans-Populations/Diseases Genetic Effect. *Schizophr Bull*. 2019;45(4):824-834.

31. Ding J, Berleant D, Nettleton D, Wurtele E. Mining MEDLINE: abstracts, sentences, or phrases? *Pac Symp Biocomput*. Published online 2002:326-337.

32. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*. 2004;5:147.

33. Osborne JD, Zhu LJ, Lin SM, Kibbe WA. Interpreting microarray results with gene ontology and MeSH. *Methods Mol Biol*. 2007;377:223-242.

34. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a

knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550.

35. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. *VLDB J*. 2020;29(2):709-730.

36. Lin J. Is searching full text more effective than searching abstracts? *BMC Bioinformatics*. 2009;10:46.

37. Müller H-M, Van Auken KM, Li Y, Sternberg PW. Textpresso Central: a customizable platform for searching, text mining, viewing, and curating biomedical literature. *BMC Bioinformatics*. 2018;19(1):94.

38. Van Auken K, Schaeffer ML, McQuilton P, et al. BC4GO: a full-text corpus for the BioCreative IV GO task. *Database* . 2014;2014. doi:10.1093/database/bau074

39. Wei Q, Collier N. Towards classifying species in systems biology papers using text mining. *BMC Res Notes*. 2011;4:32.

40. Verspoor K, Cohen KB, Lanfranchi A, et al. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*. 2012;13:207.

41. Islamaj Dogan R, Kim S, Chatr-Aryamontri A, et al. The BioC-BioGRID corpus: full text articles annotated for curation of protein-protein and genetic interactions. *Database* . 2017;2017. doi:10.1093/database/baw147

42. PMC Open Access. PubMed Central. Published 2020. Accessed September 6, 2020. https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

Table 1:Confusion Matrix of CNN on Test Data

		Predicted Class		
		Systemic Stress	Cellular Stress	
Actual Class	Systemic Stress	4,853 (TP)	147 (FN)	Sensitivity: 97%
	Cellular Stress	310 (FP)	4,690 (TN)	Specificity: 94%
		Negative Predictive Value: 97%	Negative Predictive Value: 94%	Accuracy: 95%

Figures:

Figure 1.

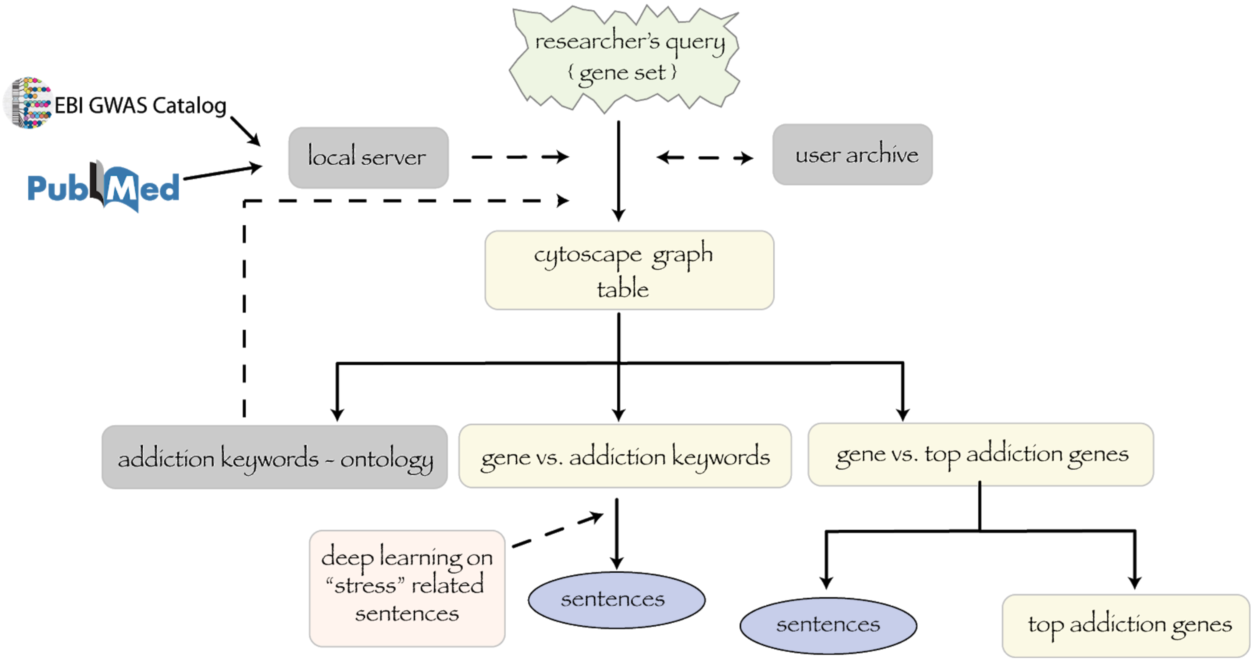

Figure 1: RatsPub allows researchers to query gene sets against many addiction related keywords and human GWAS. These keywords are predetermined by forming a small ontology, whereas the human GWAS data are extracted from the NHGRI-EBI GWAS catalog. Users have an option to choose keyword categories during the search. Searches are conducted using EUtils against the PubMed database but abstracts are retrieved from a locally mirrored copy of PubMed. The results are displayed as a cytoscape graph (Fig. 3) and a table. Results are archived on the server if the user chooses to log in. The graph and the table have many interactive elements, including displaying sentences that include the gene symbols and the addiction keywords. The number of unique abstracts and related sentences are shown separately. Sentences containing the keyword *stress* are classified using a convolutional neural network into one of two classes: systemic stress or cellular stress (Figs. 2 and 4). Sentences about the target gene and top addiction genes can be retrieved. Top addiction genes are ranked by the number of PubMed abstracts that contain the name of the gene and one or more addiction related keywords. Lastly, users can conduct secondary searches that contain synonyms of the target genes.

Figure 2.

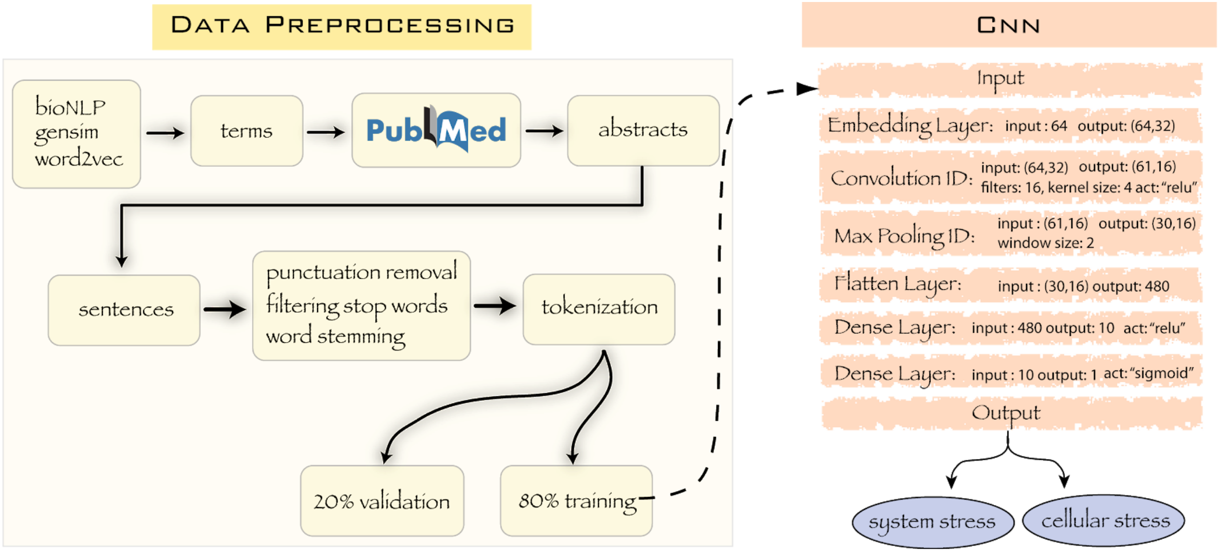
DATA PREPROCESSING
bioNLP
gensim
word2vec
terms
PubMed
abstracts
sentences
punctuation removal
filtering stop words
word stemming
tokenization
20% validation
80% training
CNN
Input
Embedding Layer: input : 64 output: (64,32)
Convolution 1D: input: (64,32) output: (61,16)
filters: 16, kernel size: 4 act: "relu"
Max Pooling 1D: input : (61,16) output: (30,16)
window size: 2
Flatten Layer: input : (30,16) output: 480
Dense Layer: input : 480 output: 10 act: "relu"
Dense Layer: input : 10 output: 1 act: "sigmoid"
Output
system stress
cellular stress

Figure 2: Pipeline for training the convolutional neural network that classifies sentences containing the word "stress". We used biomedical natural language processing tool [17] and the word2vec embeddings derived from PubMed and PMC text. The relevant terms with "system stress" and "cellular stress" were searched by using the cosine similarity tool in Python's Gensim library and the abstracts including these terms were fetched from PubMed. Abstracts then were parsed into sentences, punctuations were then removed, stop words were filtered, and all words were reduced to their stems. These words were then "tokenized" and were splitted into training (80%) and validation (20%) sets. Input layer of the model passed the training data to the embedding layer, which produced a 32 dimensional embedding vector for each word. After a 1D convolutional layer with 16 filters and a kernel size of 4, downsampling is implemented by a maximum pooling layer with window size of 2. Output of this is flattened to a 480 node layer and connected to two fully connected layers. We use the rectifier unit function to activate the neurons in the convolution layer and the dense layer. Last dense layer is activated by the sigmoid function. The final weights of the model were used to classify input sentences into either system stress or cellular stress.

Figure 3.

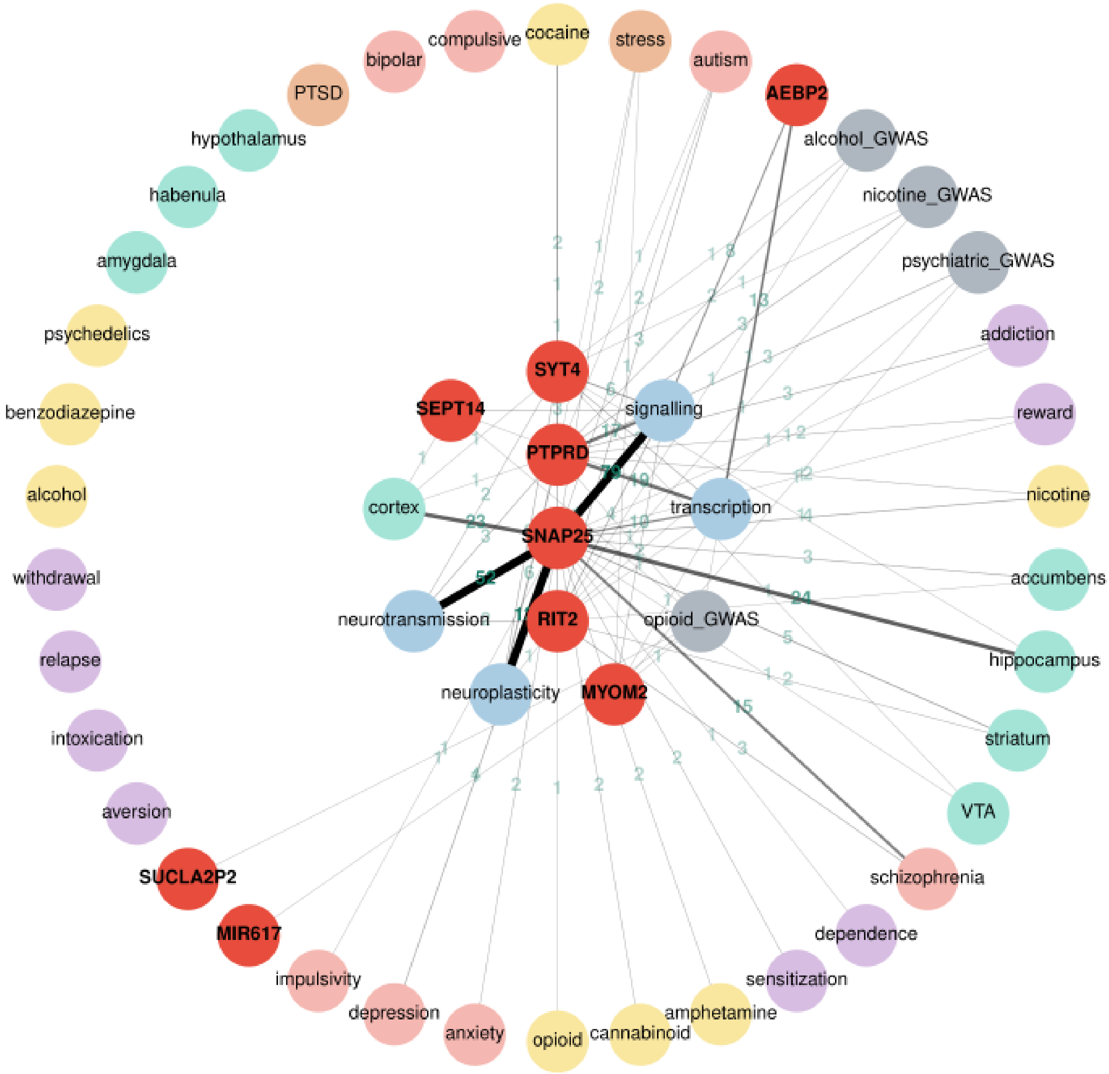

Figure 3: An interactive Cytoscape graph visualizing gene-keyword relationships. Nodes (circles) represent either search terms (in red) or keywords (colored according to the mini ontology; GWAS results are in grey). Clicking the keyword nodes displays the individual terms that are included in the search. Clicking the gene symbols displays their synonyms. The edges represent relationships between nodes. The number of PubMed abstracts where the gene symbol and keyword co-occur in the same sentence are displayed on the edges. The width of edge is correlated with the number of abstracts. Clicking on the edges shows these sentences, which are linked back to PubMed abstracts. Nodes can be moved about for better visibility of relationships. These genes were taken from a recent genome wide association study of opioid cessation [22].

Figure 4.

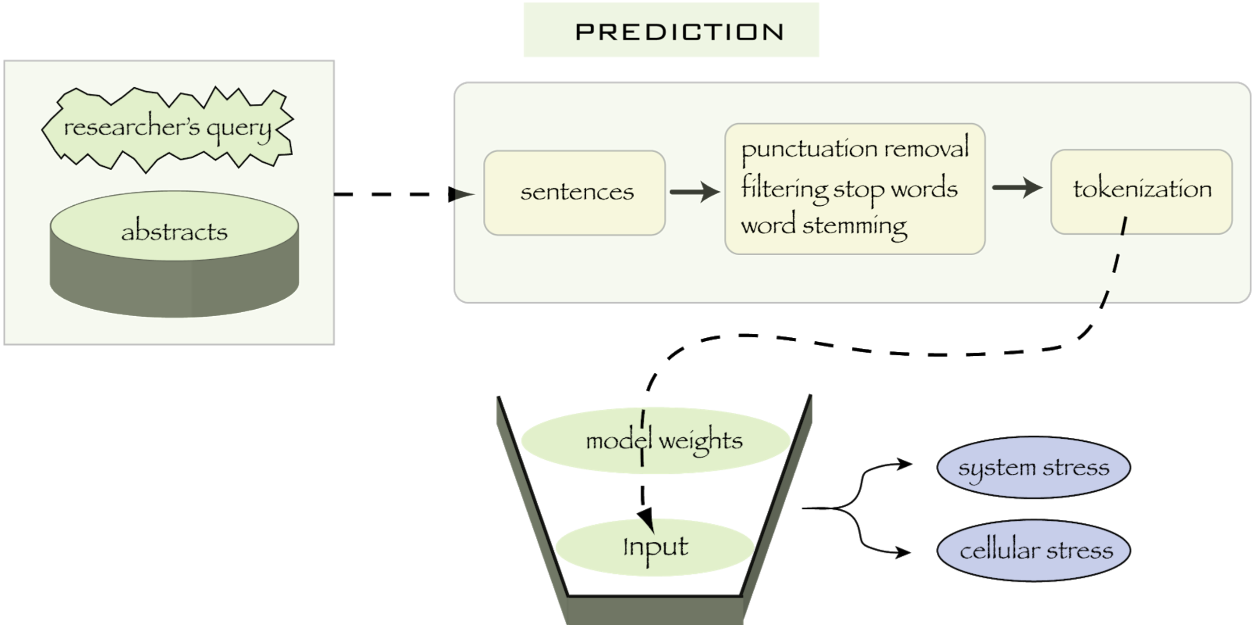

Figure 4. Steps for classifying sentences using a trained neural network. Abstracts are fetched from the locally mirrored copy of PubMed and are parsed into sentences. Punctuation marks and stop words are removed and the remaining words of the sentences are stemmed. The words are tokenized by using the Tokenizer library of the Keras API. The weight matrices of the trained model are multiplied by the sentence matrix to predict whether the input sentences are related to system stress or cellular stress.