Predicting Chemical-Disease Associations with Ontological Embeddings and Artificial Intelligence

Ian Coleman

Student Number: 9106160090

March 2019

Thesis
MSc Bioinformatics

Dr. Robert Hoehndorf
King Abdullah University of Science and Technology
&
Professor Bas Zwaan
Wageningen University

Abstract

Background

Human activity has introduced a vast array of chemicals into our environment. For the most part, the impact of these chemicals on human health is poorly understood. Legislation is moving towards proactively requiring proof of chemical benignity, rather than the traditional use-first-and-research-later approach. This places pressure on the toxicological assessment process, which currently depends on resource-intensive *in vitro* and epidemiological methods. An effective computational model could significantly reduce the burden of this process.

Results

We present a methodology to process biomedical ontological information through an artificial neural network in the prediction of positive chemical-disease associations. A training database is created with a wide array of chemicals, primarily but not exclusively environmental chemicals and drugs. Features are created for each chemical-disease pair using established gene and phenotype associations and leveraging an array of ontologies. On cross-validation, this model displays high specificity and sensitivity (AUC = 0.92) in predicting positive chemical-disease associations. Specific predictions are provided and an evaluation process described.

Conclusion

Computational models are required to radically improve the efficacy of toxicological assessment processes. One such model, a combination of ontological semantic mining and a neural network is proposed here and shown to have the ability to provide accurate chemical-disease association predictions.

Table of Contents

Introduction

Human activity has radically altered our environment and lifestyle over recent centuries. As a result, the quantity and variety of chemicals to which we are exposed has changed significantly. For the most part, the impact of these chemicals on our complex biology is poorly understood. The expanded use of naturally occurring and synthetic materials has been integral to human advancement in effectively every industry, however, the approach of treating these substances as innocent until proven guilty has also resulted in widespread damage. Some harmful examples have become commonplace in our environment before their danger was recognised, such as with chlorofluorocarbons and flame retardants (Zha et al. 2019), (Parson 2003). Water sources across the world are found to feature pollutants inadvertently acquired from industrial processes, potentially having a subversive negative health impact on large populations of people (Bondy and Campbell 2017). The same is true for our air, with increasing levels of evidence suggesting a major morbidity burden resulting from poor air quality both in poor countries and in rich (B.-Y. Yang et al. 2018; Sweileh et al. 2018; Kilian and Kitazawa 2018).

A major barrier to effective regulation of environmental substances is the difficulty of proving them benign or harmful - the introduction of one entity into a complex biological system can have unpredictable and insidious effects that may take many years to detect and verify. We have, historically, taken a use-first-research-later approach to chemical tools, but in an ideal world each would be investigated in an effective way before implementation. It seems that legislation is moving in this direction, with a 2016 act 'The Frank R. Lautenberg Chemical Safety for the 21st Century Act' giving the United States Environmental Protection Agency ("EPA") the onus to test a chemical's toxicity before its commercial use (Epa et al. 2016). With conventional methods, proactive assessment is wildly impractical in time and resource consumption, but computational models may offer a solution. Bioinformatic methods are opening up the potential to carry out a wide breadth of biological research in a resource-efficient way, enabling us to test hypotheses at a rate that would previously have been unfeasible. Computational models of this nature have transformed drug discovery and drug repurposing among other fields (Hoehndorf et al. 2014; Karaman and Sippl 2018; Schuler et al. 2017). We propose developing such a model to predict chemical toxicity, specifically, predicting whether a given chemical-disease pair will be positively associated. Such a tool could pose a major advance in the efficacy of toxicological assessment and enable targeted, informed and effective interventions against harmful environmental chemicals.

Existing chemical toxicity prediction methods tend to be based on one of three concepts, or some combination; (i) chemical similarity based prediction, (ii) common biological intermediates based prediction (see Figure 1), or (iii) else chemical structure based prediction (W. Zhang et al. 2018). In the chemical similarity approach, a dataset is created with established chemical-disease associations, largely gathered from published animal studies. The chemical to be assessed is then compared with each of the chemicals in this dataset, based on a set of features that may include chemical structure, chemical class, drug target, and it is inferred that the chemical will have the same toxicological profile as its closest homologue. This method necessitates the existence of and established toxicological evidence for a chemical homologue, which is frequently unavailable and expensive to acquire.

The method of using common biological intermediates effectively matches up chemicals and diseases that are associated with the same gene, protein, and/or metabolic pathway, and inferring that the chemical and the disease will in turn be associated with each other. This can only succeed when a given chemical-disease pair have at least one known common intermediate, which is very often not the case.

We put forward a method that aims to capture the strengths of these existing methods while minimising their pitfalls. We create a deep set of features for each chemical and each disease, capturing a myriad of information including but not limited to semantic information, associated biological entities, and interaction networks. Importantly, each feature represents not only a trait of the given disease/chemical for which it is created, but also captures information for multiple layers of associated entities. For example, when we draw on gene associations we record not only that chemical 1 is associated with gene B, but also that gene B is associated with gene C and that gene C is associated with gene D. So if disease 1 is associated with gene D we can identify a possible association between chemical 1 and disease 1. These extended network representations capture a richer understanding of an entity and give a significantly reduced dependence on established common biological intermediates or toxicologically assessed chemical homologues. Semantic embeddings of this nature have shown value in various biological settings (Cohen and Widdows 2017). Figure 2 depicts a conceptual version of our model.

Biological network information is provided by ontologies. Biological ontologies are a powerful tool for capturing the complexity of biological systems. An ontology is a specification of a conceptualisation of a domain (Gruber 1993). An ontology breaks a domain into defined classes and maps the relationship of these classes to each other, information which can then be used to annotate biological entities and provide their relationships to each other for a given biological domain. Utilising this context, particularly for multiple biological domains, allows us to more accurately capture the nature of a biological variable, treating it as an integral part of a complex system rather than as an isolated entity. The representation of biological data in the formalised structure and standardised language of ontologies allows data science to leverage information that would otherwise be too ambiguous and qualitative to process.

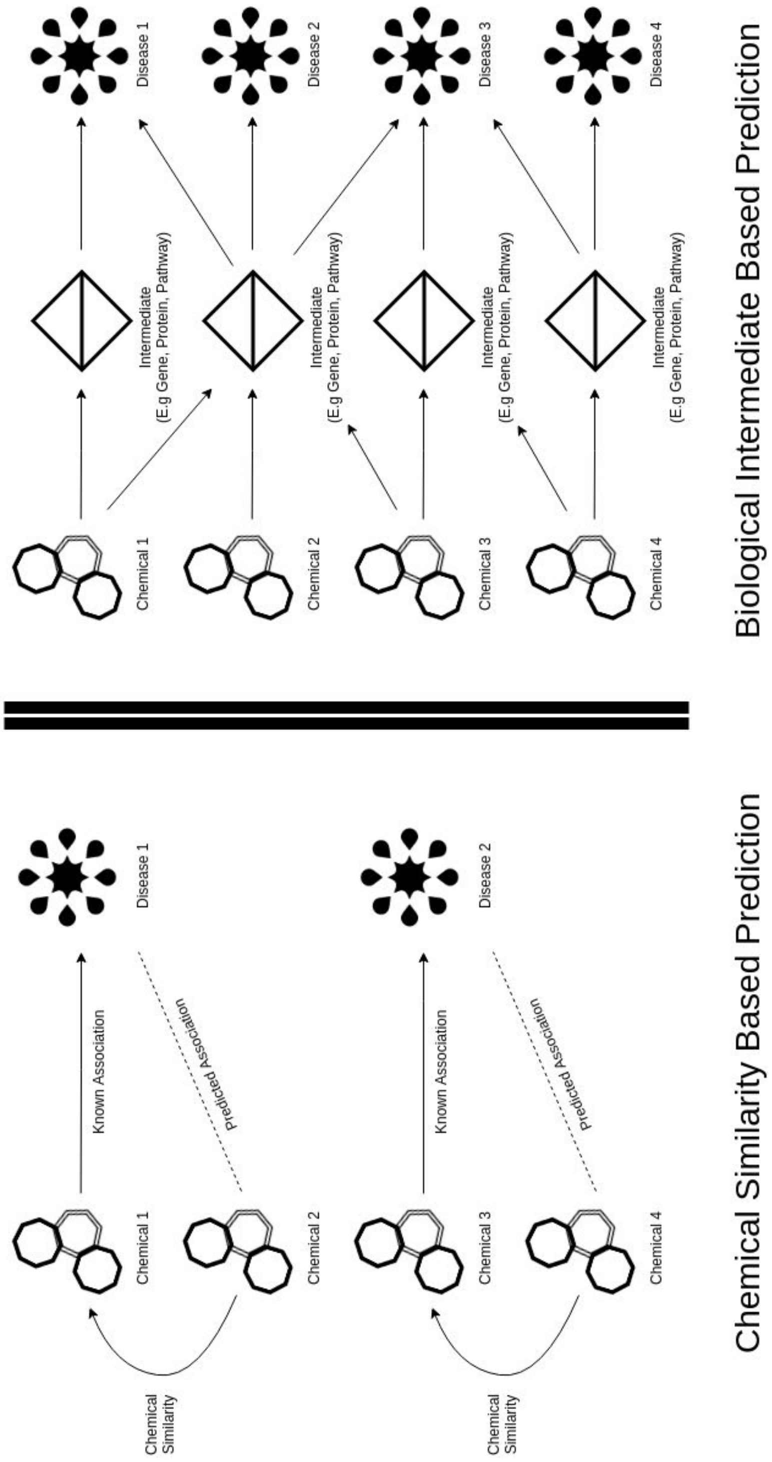

Figure 1. Prominent Existing Chemical Toxicity Prediction Methodologies

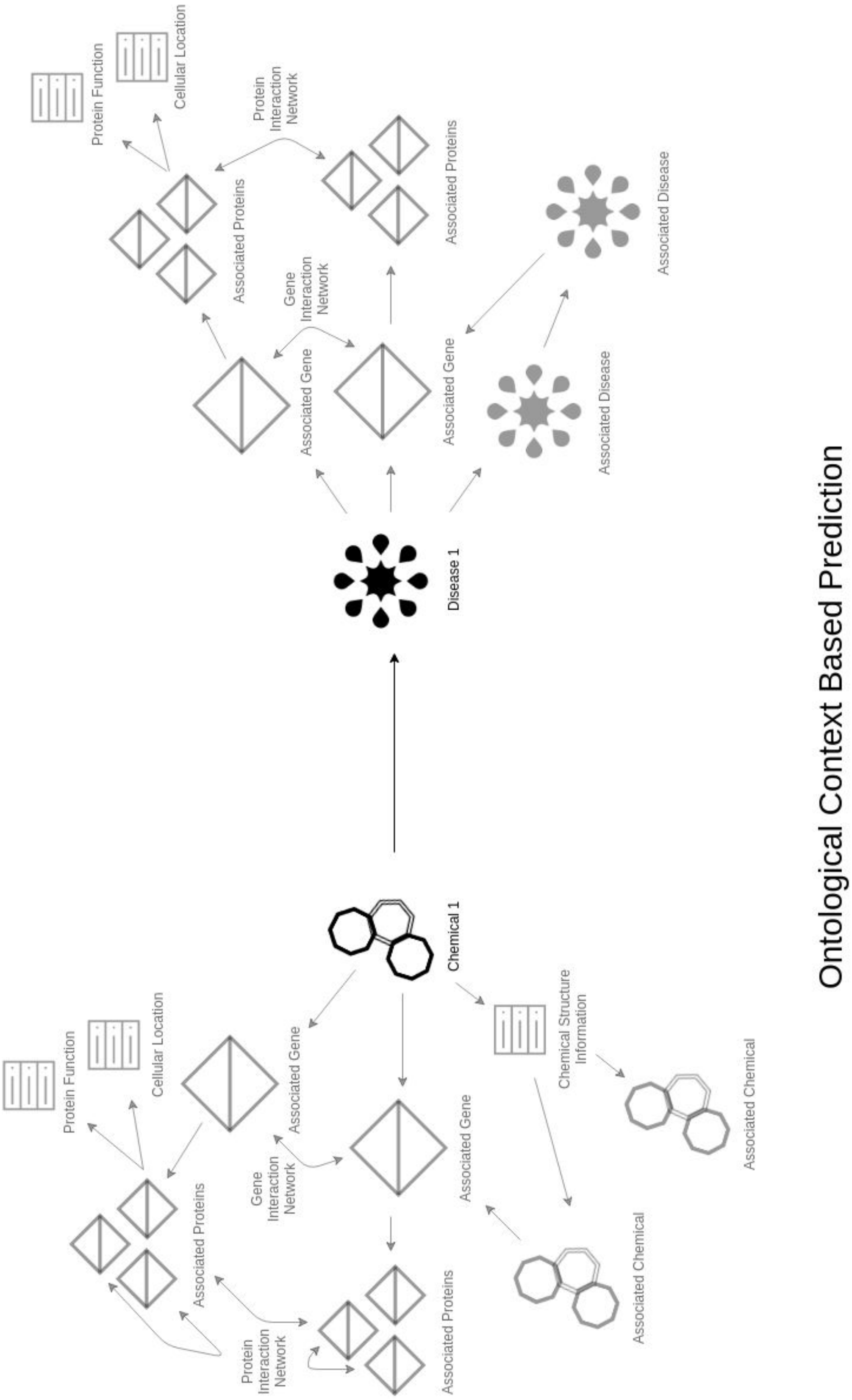

Figure 2. Ontological Context Based Chemical Toxicity Prediction

It is our aim, as outlined above, to represent each chemical and each disease with number vectors derived from a composite of networks, each network being a subset of an ontology centered around the chemical/disease or its proxy. To learn usable feature vectors from the semantic structure of our ontologies we will use Machine Learning ("**ML**") methods popularised in Natural Language Processing ("**NLP**"), where a word is represented by its context words - words that co-occur alongside that word in a corpus of text as weighted by their frequency, using vector arithmetic (Pennington, Socher, and Manning 2014; Henry, Cuffy, and McInnes 2018). In our case, the words are biological entities and the corpus of text is a biological ontology. NLP techniques such as Continuous Bag of Words ("**CBOW**") and Skip-gram models have been successfully implemented in multiple biological use cases (Alshahrani and Hoehndorf 2018; Liu-Wei, Kafkas, and Hoehndorf 2018; Pakhomov et al. 2016).

Upon adding learned feature vectors for each chemical and disease to our dataset, supervised learning methods are then utilised to build a predictive model for the existence of positive chemical-disease associations. For this, we could utilise a straightforward distance measure such as cosine similarity or else advanced kernel methods, Neural Networks ("**NN's"**) or a variety of other methods. The inherent complexity of the features and their rendering as number vectors leaves them sufficiently unintuitive that domain knowledge cannot be easily applied to inform the direction of the feature space, as such a NN with its greater flexibility to define its own feature space likely poses a better choice than options such as kernel methods. NN's perform well in a wide range of use cases, scale well to handle large datasets, are relatively easily implemented and well established in the scientific literature and as such are the method of choice (Z. Zhang 2016; Molaie et al. 2014; Dawson 2018).

Marrying machine learning to structured biological data such as biomedical ontologies has proven to be an effective tool elsewhere, exposing new insights in a variety of use cases (Liu-Wei, Kafkas, and Hoehndorf 2018), (Alshahrani and Hoehndorf 2018). While NN's are 'black boxes' that give us predictions without revealing the patterns they have identified, their raw predictive power makes them worthwhile for use cases where prediction accuracy is the priority.

The United States' EPA has published a strong case for the need for a computational model to empower the toxicological assessment of chemical compounds, and the advantage of using ontologies in doing so (R.-L. Wang, Edwards, and Ives 2019). We set forward such a model in this report. We then demonstrate that this model can be used to predict positive chemical-disease associations with a level of accuracy that renders it of practical use and we discuss it in the context of other models and approaches in the field.

Methodology

Data

We first create feature vectors for each chemical and disease, then feed these into an artificial NN. Our workflow is depicted in figure 3.

An initial training dataset was constructed from The Comparative ToxicoGenomic Database (“**CTD**”) (Davis et al. 2018), each row representing a chemical-disease pair, with a binary variable indicating whether or not the chemical and the disease are positively correlated. Molecular reagents, environmental chemicals and clinical drugs comprise our chemical set, the disease set consisting of any disease with at least one chemical association as per CTD. This initial set contained only positively correlated chemical-disease pairs, so a control set of observations was made by randomising chemical-disease combinations and removing any of these pairs that are known by CTD to be correlated, the remaining randomised pairs were then assumed to be uncorrelated. This left us with a balanced dataset of a layout depicted in Table 1.

Chemical	Disease	Correlation
Chemical A	Disease 1	1
Chemical B	Disease 1	0
Chemical B	Disease 2	1

Table 1. The Layout of the Initial Dataset

Gene-disease associations, gene-chemical associations, disease-phenotype associations, and chemical-disease associations were downloaded from CTD on September 25th 2018. Curated associations were selected, inferred associations removed, resulting in 17,739 observations; 8,920 correlated plus 8,819 uncorrelated chemical-disease pairs, covering 568 chemicals and 2,501 diseases.

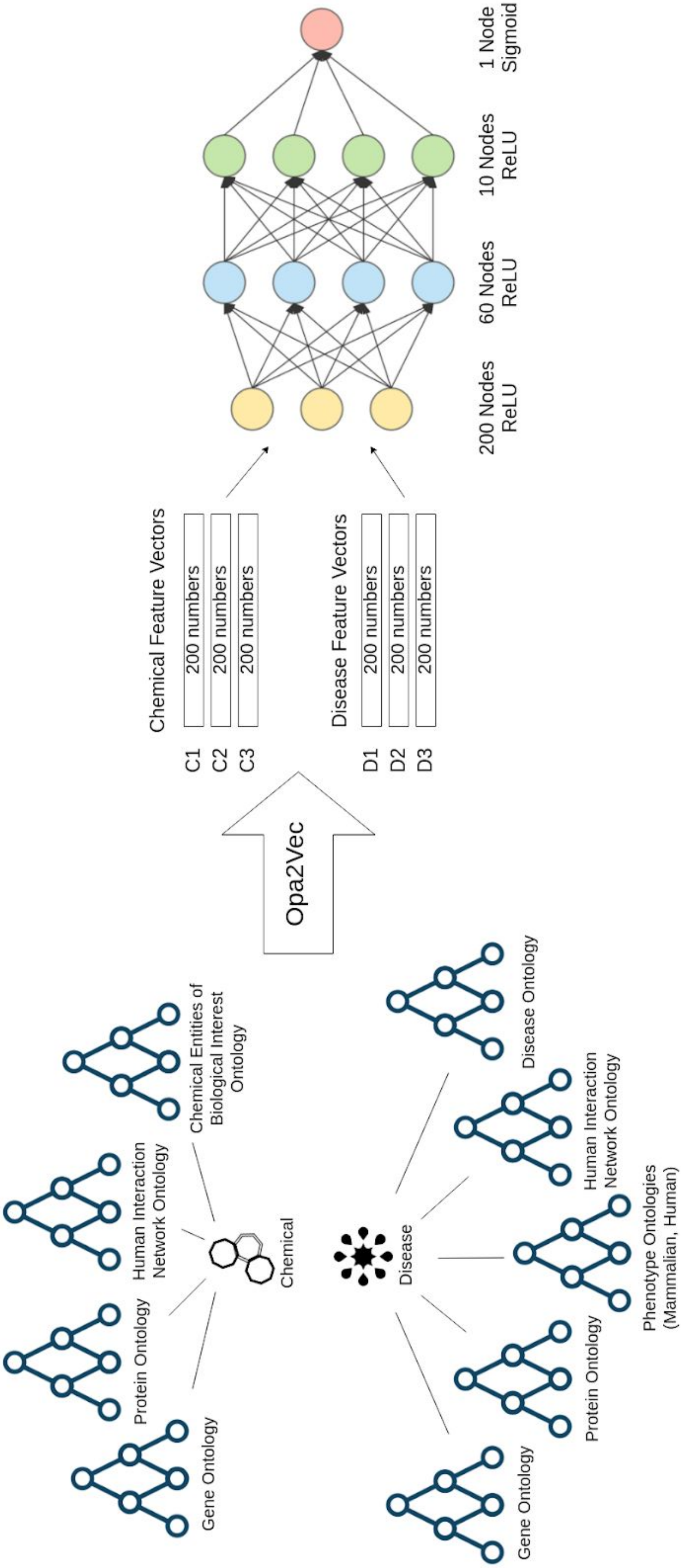

Figure 3. Workflow

Feature Creation

A number of tools have been validated for the leveraging of biological ontology data. Opa2vec was selected, a method that extracts not only the formal axioms of an ontology but also the often overlooked metadata such as entity descriptions and synonyms. Opa2Vec was a natural choice as it has relative ease of implementation, validation as a competitive method in the peer reviewed literature and, additionally, its creators and maintainers were at hand (Smaili, Gao, and Hoehndorf 2018).

Ontological Annotation

First of all, the chemicals and diseases in our dataset had to be annotated with ontological classes for our various ontologies. Gene Ontology ("**GO**") functions were downloaded on 17/9/18 and assigned to each chemical and disease via mutually associated genes (as per CTD) (The Gene Ontology Consortium 2016). As we selected for chemicals and genes that had established gene associations, every row had these features.

Disease MeSH IDs provided by CTD were mapped to Disease Ontology IDs (DOIDs) to enable mapping to phenotype ontologies. The Human Phenotype Ontology was downloaded on November 20th 2018 and diseases were annotated via their positively associated phenotypes (as per CTD) (Köhler et al. 2019). In total, 277,300 disease-phenotype associations were collected, 193,000 were applicable to our dataset resulting in 8,670 observations having these vectors. The Mammalian Phenotype Ontology was downloaded on November 20th 2018 and diseases were again annotated based on their positively associated phenotypes (as per CTD) (Smith and Eppig 2009). In total, 111,742 associations were collected, 88,633 were applicable to our dataset resulting in 8,670 observations having these vectors.

CTD chemical IDs were mapped to PubChem Chemical IDs (CIDs) - 15,148 were mappable, which were then further mapped to Chemical Entities of Biological Interest (ChEBI) IDs. The Chemical Entities of Biological Interest (ChEBI) ontology was downloaded on January 3rd 2019 and used to annotate the chemicals in our database (de Matos et al. 2010). ChEBI vectors were obtained for 10,825 rows, providing information on chemical structure and groupings.

The Disease Ontology was downloaded on January 16th 2019 and annotations created for the diseases in our database (Schriml et al. 2019). Every row had a Disease Ontology vector.

In order to capture information on protein and gene interactions, the Human Interaction Network Ontology (HINO) was downloaded on January 3rd 2019, chemicals and diseases were annotated to HINO classes via the proteins of their positively associated genes (as per CTD) (Özgür, Hur, and He 2016). 11,498 rows had a chemical HINO vector and 11,510 a disease HINO vector.

The Protein Ontology was downloaded on February 10th 2019, chemicals and diseases were annotated based on the proteins of their positively associated genes (as per CTD) (Natale et al. 2017). All rows had Protein Ontology vectors for both chemical and disease.

Feature Learning
With our entities annotated to ontologies, Opa2Vec was used to extract the ontological representation of these entities into vector-space features. A skipgram-model approach was used to capture information on neighbouring classes. The default Opa2Vec parameters were used as initial testing showed no clear improvement on varying these; a window size of 5, a minimum count parameter of 2 and an embedding size of 200.

Neural Net

An artificial NN was created using the TensorFlow framework (version 1.11.0) and the Keras library. The final NN is a multi-layered feed-forward model. The codebase used to construct the features and the neural net is available at https://github.com/colemai/msc-thesis.

NN Layers:

1. 200 nodes with a Rectified Linear Units (“**ReLU”**) activation function (Eckle and Schmidt-Hieber 2019).
2. 60 nodes with ReLU.
3. 10 nodes with ReLU.
4. 1 node with a sigmoid activation function.

The features were concatenated into a single vector to be fed into the NN. The outcome to be predicted was the existence of a positive correlation between a chemical and a disease - a binary outcome. CTD was used as the authority on the existence of a chemical-disease positive relationship, when such a relationship was in CTD’s curated database it was assumed to exist and otherwise was assumed not to. The fallibility of these assumptions is likely a mild source of error, though unavoidable. At one time the model was designed to find either positive or negative chemical-disease relationships but this was found to impede the predictive capacity for positive associations and so was abandoned in order to focus solely on positive associations i.e chemical toxicity. The logic of the model was assessed for circularity (training a model with information that it is intended to predict, and thereby artificially inflating its evaluation metrics) and the feature set adapted accordingly.

Feature Selection

The receiver operating characteristics (“**ROC**”) curve was used to judge each iteration of the model, specifically the area under the curve (“**AUC**”) metric. AUC is widely used in the scientific literature and is robust to imbalance between groups in a dataset (Bradley 1997). Features that contributed to the ROC AUC metric of the model were retained, as measured by an increase of at least an absolute 1 percent to an average AUC across three training iterations. Various combinations of the features were tested by forward/backward manual selection.

Parameter Optimisation

The data was split randomly by chemical into training set (60%, 342 chemicals), validation set, and test set (each 20%, 114 chemicals). The NN was trained with five-fold cross validation on the training set, using Adam Optimisation on a binary cross-entropy loss function (Ruder 2016). Model evaluation was carried out via accuracy and AUC metrics as tested on the holdout test set. A further evaluation was via a validation step on unseen databases.
Hyperparameters (the settings of the NN) were tuned on a trial and error basis, manually iterating through combinations and assessing by AUC.

Validation

In order to ultimately verify the potential of the model, further testing was carried out on unseen databases. Note that the model used in this validation step was trained on the CTD dataset and the unseen databases were simply fed into the established model to obtain predictions.

SIDER is an openly available database that documents side effects of consumer medications. 1,657 chemical-disease (drug-side effect) combinations were extracted, filtered down to this number by the presence of a disease rather than a phenotypic outcome variable as well as the availability of at least one positive gene association for each of the chemical and the disease in question. Our described features were then created for these chemical-disease pairs. The pretrained model was applied to this dataset to predict chemical-disease positive associations. Various combinations of the established features were tested by manual selection and AUC-based evaluation.

The Virtual Metabolic Human (VMH) holds data on metabolites and human disease, with an emphasis on the microbiome and nutrition. Fifty-two chemical-disease pairs were isolated via the API, this being the number of diseases with positive chemical associations where both the chemical and the disease had a gene listed as being correlated. The pretrained model was then imported and used to predict chemical-disease positive associations for this dataset. Various combinations of the established features were again tested by manual selection and AUC-based evaluation.

Results

Evaluation

The model was evaluated on a holdout set of our CTD dataset, showing strong predictive capacity.

Accuracy	0.84
AUC	0.92

A subjective assessment of false negative predictions and false positive predictions showed no obvious patterns by the class or nature of the diseases or chemicals in question, though it was found that entities with fewer positive gene associations, i.e less data, were more prone to erroneous predictions.

The pretrained model was then tested on our VMH and SIDER datasets:

Database	**VMH**	**SIDER**
Accuracy	0.55	0.54
AUC	0.61	0.61

Predictive scores here were predictably lower than on the CTD dataset, given that the model was both trained on and designed for the CTD use case. Of the feature set used in the initial CTD model, various combinations of features were trialed on the VMH and SIDER validation datasets. A model of GO and Protein Ontology features alone performed best on VMH. SIDER was found to answer best to GO and HINO vectors only. It is unclear why these reduced feature sets proved more efficacious than the full set. The chemicals and diseases from these validation sources came with fewer gene associations, which may have caused the less valuable features to have too few connections with which to provide any predictive capacity.

The confusion matrix of predictions on the holdout set of our CTD dataset:

	Predicted True	**Predicted False**	**Total**
Actually True	1,489	213	1,702
Actually False	293	1,524	1,817
Total	1,782	1,737	3,519

Predicted Chemical-Disease Associations

Ozone was selected as an environmental chemical for specific predictions. By the nature of the dataset, the NN was trained only with the information that ozone positively correlates with cachexia. The below tables show the highest and lowest ranking ozone disease predictions, as returned by the model. We additionally show in the 'Evidence' column that many of the highly ranked predictions were indeed established associations in CTD, while others were not found in CTD but are evidenced by peer-reviewed human or animal studies. The remainder are largely either co-morbidities from established associated diseases or else credible novel associations.

Chemical	Disease	Prediction Score	Evidence
Ozone	Inflammation	0.998	CTD
Ozone	Hypotension	0.998	Animal Study (Akcılar et al. 2015)
Ozone	Drug Eruptions	0.997	N/A
Ozone	Fever	0.997	Human Study Side Effect (Leaker, Barnes, and O'Connor 2013)
Ozone	Edema	0.997	Animal Study (Caudal et al. 2018)
Ozone	Wounds and Injuries	0.997	N/A
Ozone	Micronuclei, Chromosome-Defective	0.997	...Ozone is Epidemiologically Associated with Cytogenetic damage (Leaker, Barnes, and O'Connor 2013; Huen et al. 2006)
Ozone	Hypertrophy	0.996	N/A
Ozone	Chromosome Breakage	0.996	Epidemiology (Leaker, Barnes, and O'Connor 2013; Huen et al. 2006)
Ozone	Hyperplasia	0.996	N/A
Ozone	Nerve Degeneration	0.996	CTD: Association for Neurodegenerative Diseases
Ozone	Fibrosis	0.996	Expected in Lung Injury, Which is Associated as per CTD
Ozone	Lung Injury	0.996	CTD

Highest Ranked Disease Predictions for Ozone

The lowest rank predictions were indeed for disease states with no evidence linking them to ozone.

Chemical	Disease	Prediction Score	Evidence
Ozone	Yunis Varon syndrome	1.22E-06	N/A
Ozone	RETINITIS PIGMENTOSA 59	7.65E-06	N/A
Ozone	Pontocerebellar Hypoplasia Type 6	1.10E-05	N/A
Ozone	MANDIBULOFACIAL DYSOSTOSIS, GUION-ALMEIDA TYPE	1.40E-05	N/A
Ozone	Desbuquois syndrome	1.46E-05	N/A
Ozone	Smith-Magenis Syndrome	1.74E-05	N/A
Ozone	Orofaciodigital syndrome 6	1.88E-05	N/A
Ozone	PERRAULT SYNDROME 3	1.99E-05	N/A
Ozone	Holt-Oram syndrome	2.29E-05	N/A
Ozone	3C syndrome	2.30E-05	N/A
Ozone	COMBINED OXIDATIVE PHOSPHORYLATION DEFICIENCY 21	2.43E-05	N/A

Lowest Ranked Disease Predictions for Ozone

Discussion

We present an ontology-based method for predicting the toxicity of environmental chemicals using structural, functional, phenotypic and interaction network data of chemicals and diseases and their positively associated genes. Previous methods are built largely on established direct associations between biological entities while our methodology aims to gain more from existing knowledge by harnessing broad network representations of these associations and therefore draws also on indirect and even distant associations between entities. Additionally, previous methods for chemical-disease association prediction have, for the most part, been optimised specifically for drugs, whereas we took a more generalised approach in order to predict for environmental exposures.

Most of the established chemical-disease association prediction methods use one, or some combination of chemical similarity based prediction, common biological intermediates based prediction or chemical structure based prediction. Luechtefeld *et al.* have brought a read-across method, with Jaccard-distance, to chemical structure based prediction for a broad range of chemicals (Luechtefeld et al. 2018). Mayr *et al.* have demonstrated an AUC of 84.6% on a CTD subset, using a deep NN on chemical structure data alone (Mayr et al. 2016). Wang et al. have trained a Support Vector Machine model to use a hybrid chemical structure and chemical similarity approach, specifically for drugs (Y. Wang et al. 2013). Yang et al. have implemented a Naive Bayes model to repurpose drugs based on their known side effects (L. Yang and Agarwal 2011). Common biological intermediates have been trialed in various ways including Yu et al.'s weighted tripartite network for common protein complexes (Yu et al. 2015; Wiegers et al. 2009).

The PREDICT tool is built on drug-drug similarity and disease-disease similarity, bringing a linear regression model to bear in predicting drug-disease features (Gottlieb et al. 2011). Zhang et al have shown great success with another drug-drug and disease-disease similarity tool, using a wide range of features including chemical substructures, targets, pathways, drug-drug interactions and disease semantic information, feeding this information into a similarity-constrained matrix factorization model (W. Zhang et al. 2018).

While our method similarly uses drug-drug and disease-disease similarity with a broad range of features, it can be differentiated from these other chemical-disease predictors by its vast harnessing of biological context. This wide web of information gives a reduced dependence on common biological intermediates or established toxicological assessments relative to other methods, as indirect and distant relationships between entities are now available to the model rather than only first degree associations. We harness, to some extent, each of chemical similarity based prediction, common biological intermediates based prediction and chemical structure based prediction, this last being provided by the ChEBI ontology. As such this model comprises a hybrid of the three most successful methodological classes in this field while adding the additional appendage of ontological semantic mining, an increasingly prevalent and accomplished methodology.

Similar ontological semantic-mining methodologies have shown success in detecting disease-gene candidates (Smaili, Gao, and Hoehndorf 2018), predicting host-pathogen interactions (Liu-Wei, Kafkas, and Hoehndorf 2018), predicting protein-protein interactions (S.-B. Zhang and Tang 2016) and their effect types (Yim et al. 2018). Kulmanov *et al.* have demonstrated an ontology based method to predict phenotypic outcomes from functional genetic information (Kulmanov et al. 2018). Althubaiti *et al.* have similarly developed a tool to accurately identify cancer driver genes (Althubaiti et al., n.d.).

Our model shows excellent predictive capabilities for the CTD dataset, with accuracy competing even with the reproducibility rate of animal studies - the gold standard source of toxicological knowledge. Animal studies show reproducibility of 78%–96% (sensitivity 50%–87%) in a range of mutagenicity, oral and dermal reaction measures (Luechtefeld et al. 2018), relative to our model's 84% accuracy and 91% AUC.

In the validation datasets, SIDER and VMH, our model displayed moderate predictive capacity, though at a lower rate than in the CTD setting. Given that SIDER is a particularly difficult use case, with many of the side effects being rare, and given that our model is not optimised for it, an AUC of 61% is not unsatisfactory and does indeed serve the purpose of validating our model's inherent predictive capacity. Similarly, our result for the VMH dataset shows genuine predictive abilities, while not of a field-leading AUC. The VMH dataset consists mostly of metabolites that are markers of a condition, while our CTD chemicals are most often causative agents of a disease, which is one of numerous possible reasons for the decreased relative AUC. Another is the relatively small size of the VMH dataset that we were able to isolate for our purposes. It is reasonable to assume that further work could rapidly adapt our method for these use cases.

This model poses a powerful screening tool to inform and/or complement conventional toxicity testing, such as *in vitro* and epidemiological approaches. However, gene-chemical and gene-disease associations are required in order for a chemical-disease pair to be effectively tested for association, with the model being unvalidated and unlikely to work for entities with no published gene associations and as such we do not fully sidestep the weaknesses of earlier approaches. This methodology is likely to become stronger over time as more gene associations are discovered for chemicals and diseases and as the biological ontologies develop and mature. A further strength is that it is an easily adaptable model, the current features can be rapidly updated and new features added.

While this model provides a valid, binary indication of chemical toxicity for a given disease or set of diseases, it remains an imperfect solution, not accounting for the quantity of the chemical involved nor modelling the cycle of that chemical into and out of the human body. Additionally, the model gives the classic Neural Net trade-off of providing powerful prediction while not giving any indication of what the patterns are that it uses to predict chemical-disease associations and therefore not leaving us wiser on that subject. Further research could incorporate the use of advanced techniques to decode insights into the patterns that this NN is finding (Maratea and Ferone 2019). Further research could additionally create an even testing ground to compare our method to other computational methods for superiority in various use-cases.

Our model appears to be viable for practical applications, which may include involvement in the process of environmental regulatory bodies in gauging the toxicity of chemicals, or by commercial entities in the safety assessment of novel or proposed products. The utilisation of this tool could pose a major saving in time and resources and could enable a seismic shift in the scope of toxicological assessment, making practical the attainment of a healthier environment.

Conclusion

A model is presented with a sufficient predictive capacity to be of practical use in informing the assessment of the toxicity of environmental chemicals. This is an easily implementable, powerful and generalised tool, comprising a hybrid of the best methodological classes in the field of chemical-disease association prediction with the addition of ontological semantic mining. Novel chemical-disease associations are additionally proposed.

References

Akcılar, Raziye, Sezer Akçer, Hasan Şimşek, Aydın Akcılar, Zeynep Bayat, and Osman Genç. 2015. “The Effect of Ozone on Blood Pressure in DOCA-Salt-Induced Hypertensive Rats.” *International Journal of Clinical and Experimental Medicine* 8 (8): 12783–91.

Alshahrani, Mona, and Robert Hoehndorf. 2018. “Semantic Disease Gene Embeddings (SmuDGE): Phenotype-Based Disease Gene Prioritization without Phenotypes.” *Bioinformatics* 34 (17): i901–7.

Althubaiti, Sara, Andreas Karwath, Ashraf Dallol, Adeeb Noor, Shadi Salem Alkhayyat, Rolina Alwassia, Katsuhiko Mineta, et al. n.d. “Ontology-Based Prediction of Cancer Driver Genes.” https://doi.org/10.1101/561480.

Bondy, Stephen, and Arezoo Campbell. 2017. “Water Quality and Brain Function.” *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph15010002.

Bradley, Andrew P. 1997. “The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms.” *Pattern Recognition* 30 (7): 1145–59.

Caudal, Victor, Justin Whitty, Elisabeth C. R. Snead, and Gregory S. Starrak. 2018. “Noncardiogenic Pulmonary Edema Associated with Ozone Exposure in Three Kittens.” *Journal of the American Veterinary Medical Association* 253 (10): 1328–33.

Cohen, Trevor, and Dominic Widdows. 2017. “Embedding of Semantic Predications.” *Journal of Biomedical Informatics* 68 (April): 150–66.

Davis, Allan Peter, Cynthia J. Grondin, Robin J. Johnson, Daniela Sciaky, Roy McMorran, Jolene Wiegers, Thomas C. Wiegers, and Carolyn J. Mattingly. 2018. “The Comparative Toxicogenomics Database: Update 2019.” *Nucleic Acids Research* 47 (D1): D948–54.

Dawson, Christian. 2018. *Applied Artificial Neural Networks*. MDPI.

Eckle, Konstantin, and Johannes Schmidt-Hieber. 2019. “A Comparison of Deep Networks with ReLU Activation Function and Linear Spline-Type Methods.” *Neural Networks: The Official Journal of the International Neural Network Society* 110 (February): 232–42.

Epa, U. S., OCSPP, OPPT, and CCD. 2016. “The Frank R. Lautenberg Chemical Safety for the 21st Century Act,” June. https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/frank-r-lautenberg-chemical-safety-21st-century-act.

Gottlieb, Assaf, Gideon Y. Stein, Eytan Ruppin, and Roded Sharan. 2011. “PREDICT: A Method for Inferring Novel Drug Indications with Application to Personalized Medicine.” *Molecular Systems Biology* 7 (June): 496.

Gruber, Thomas R. 1993. “A Translation Approach to Portable Ontology Specifications.” *Knowledge Acquisition*. https://doi.org/10.1006/knac.1993.1008.

Henry, Sam, Clint Cuffy, and Bridget T. McInnes. 2018. “Vector Representations of Multi-Word Terms for Semantic Relatedness.” *Journal of Biomedical Informatics* 77 (January): 111–19.

Hoehndorf, Robert, Tanya Hiebert, Nigel W. Hardy, Paul N. Schofield, Georgios V. Gkoutos, and Michel Dumontier. 2014. “Mouse Model Phenotypes Provide Information about Human Drug Targets.” *Bioinformatics* 30 (5): 719–25.

Huen, Karen, Laura Gunn, Paurene Duramad, Michael Jeng, Russell Scalf, and Nina Holland. 2006. “Application of a Geographic Information System to Explore Associations between Air Pollution and Micronucleus Frequencies in African American Children and Adults.”

Environmental and Molecular Mutagenesis 47 (4): 236–46.

Karaman, Berin, and Wolfgang Sippl. 2018. “Computational Drug Repurposing: Current Trends.” *Current Medicinal Chemistry*, May. https://doi.org/10.2174/0929867325666180530100332.

Kilian, Jason, and Masashi Kitazawa. 2018. “The Emerging Risk of Exposure to Air Pollution on Cognitive Decline and Alzheimer’s Disease – Evidence from Epidemiological and Animal Studies.” *Biomedical Journal*. https://doi.org/10.1016/j.bj.2018.06.001.

Köhler, Sebastian, Leigh Carmody, Nicole Vasilevsky, Julius O. B. Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, et al. 2019. “Expansion of the Human Phenotype Ontology (HPO) Knowledge Base and Resources.” *Nucleic Acids Research* 47 (D1): D1018–27.

Kulmanov, Maxat, Paul N. Schofield, Georgios V. Gkoutos, and Robert Hoehndorf. 2018. “Ontology-Based Validation and Identification of Regulatory Phenotypes.” *Bioinformatics* 34 (17): i857–65.

Leaker, Brian R., Peter J. Barnes, and Brian O’Connor. 2013. “Inhibition of LPS-Induced Airway Neutrophilic Inflammation in Healthy Volunteers with an Oral CXCR2 Antagonist.” *Respiratory Research* 14 (1): 137.

Liu-Wei, Wang, Şenay Kafkas, and Robert Hoehndorf. 2018. “Phenotypic, Functional and Taxonomic Features Predict Host-Pathogen Interactions.” https://doi.org/10.1101/508762.

Luechtefeld, Thomas, Dan Marsh, Craig Rowlands, and Thomas Hartung. 2018. “Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility.” *Toxicological Sciences: An Official Journal of the Society of Toxicology* 165 (1): 198–212.

Maratea, Antonio, and Alessio Ferone. 2019. “Deep Neural Networks and Explainable Machine Learning.” In *Advances in Biochemical Engineering/Biotechnology*, 253–56.

Matos, Paula de, Rafael Alcántara, Adriano Dekker, Marcus Ennis, Janna Hastings, Kenneth Haug, Inmaculada Spiteri, Steve Turner, and Christoph Steinbeck. 2010. “Chemical Entities of Biological Interest: An Update.” *Nucleic Acids Research* 38 (Database issue): D249–54.

Mayr, Andreas, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. 2016. “DeepTox: Toxicity Prediction Using Deep Learning.” *Frontiers of Environmental Science & Engineering in China* 3. https://doi.org/10.3389/fenvs.2015.00080.

Molaie, Malihe, Razieh Falahian, Shahriar Gharibzadeh, Sajad Jafari, and Julien C. Sprott. 2014. “Artificial Neural Networks: Powerful Tools for Modeling Chaotic Behavior in the Nervous System.” *Frontiers in Computational Neuroscience* 8 (April): 40.

Natale, Darren A., Cecilia N. Arighi, Judith A. Blake, Jonathan Bona, Chuming Chen, Sheng-Chih Chen, Karen R. Christie, et al. 2017. “Protein Ontology (PRO): Enhancing and Scaling up the Representation of Protein Entities.” *Nucleic Acids Research* 45 (D1): D339–46.

Özgür, Arzucan, Junguk Hur, and Yongqun He. 2016. “The Interaction Network Ontology-Supported Modeling and Mining of Complex Interactions Represented with Multiple Keywords in Biomedical Literature.” *BioData Mining* 9 (December): 41.

Pakhomov, Serguei V. S., Greg Finley, Reed McEwan, Yan Wang, and Genevieve B. Melton. 2016. “Corpus Domain Effects on Distributional Semantic Modeling of Medical Terms.” *Bioinformatics* 32 (23): 3635–44.

Parson, Edward A. 2003. “Eliminating Chlorofluorocarbons.” In *Protecting the Ozone Layer*, 147–72.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. “Glove: Global Vectors for Word Representation.” *Proceedings of the 2014 Conference on Empirical Methods in*

Natural Language Processing (EMNLP). https://doi.org/10.3115/v1/d14-1162.

Ruder, S. 2016. “An Overview of Gradient Descent Optimization Algorithms.” *arXiv Preprint*. https://doi.org/arXiv:160904747.

Schriml, Lynn M., Elvira Mitraka, James Munro, Becky Tauber, Mike Schor, Lance Nickle, Victor Felix, et al. 2019. “Human Disease Ontology 2018 Update: Classification, Content and Workflow Expansion.” *Nucleic Acids Research* 47 (D1): D955–62.

Schuler, James, Matthew L. Hudson, Diane Schwartz, and Ram Samudrala. 2017. “A Systematic Review of Computational Drug Discovery, Development, and Repurposing for Ebola Virus Disease Treatment.” *Molecules* 22 (10). https://doi.org/10.3390/molecules22101777.

Smaili, Fatima Zohra, Xin Gao, and Robert Hoehndorf. 2018. “OPA2Vec: Combining Formal and Informal Content of Biomedical Ontologies to Improve Similarity-Based Prediction.” *Bioinformatics* , November. https://doi.org/10.1093/bioinformatics/bty933.

Smith, Cynthia L., and Janan T. Eppig. 2009. “The Mammalian Phenotype Ontology: Enabling Robust Annotation and Comparative Analysis.” *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* 1 (3): 390–99.

Sweileh, Waleed M., Samah W. Al-Jabi, Sa’ed H. Zyoud, and Ansam F. Sawalha. 2018. “Outdoor Air Pollution and Respiratory Health: A Bibliometric Analysis of Publications in Peer-Reviewed Journals (1900 – 2017).” *Multidisciplinary Respiratory Medicine*. https://doi.org/10.1186/s40248-018-0128-5.

The Gene Ontology Consortium. 2016. “Expansion of the Gene Ontology Knowledgebase and Resources.” *Nucleic Acids Research* 45 (D1): D331–38.

Wang, Rong-Lin, Stephen Edwards, and Cataia Ives. 2019. “Ontology-Based Semantic Mapping of Chemical Toxicities.” *Toxicology* 412 (January): 89–100.

Wang, Yongcui, Shilong Chen, Naiyang Deng, and Yong Wang. 2013. “Drug Repositioning by Kernel-Based Integration of Molecular Structure, Molecular Activity, and Phenotype Data.” *PloS One* 8 (11): e78518.

Wiegers, Thomas C., Allan Peter Davis, K. Bretonnel Cohen, Lynette Hirschman, and Carolyn J. Mattingly. 2009. “Text Mining and Manual Curation of Chemical-Gene-Disease Networks for the Comparative Toxicogenomics Database (CTD).” *BMC Bioinformatics* 10 (October): 326.

Yang, Bo-Yi, Zhengmin Qian, Steven W. Howard, Michael G. Vaughn, Shu-Jun Fan, Kang-Kang Liu, and Guang-Hui Dong. 2018. “Global Association between Ambient Air Pollution and Blood Pressure: A Systematic Review and Meta-Analysis.” *Environmental Pollution* 235 (April): 576–88.

Yang, Lun, and Pankaj Agarwal. 2011. “Systematic Drug Repositioning Based on Clinical Side-Effects.” *PloS One* 6 (12): e28025.

Yim, Soorin, Hasun Yu, Dongjin Jang, and Doheon Lee. 2018. “Annotating Activation/inhibition Relationships to Protein-Protein Interactions Using Gene Ontology Relations.” *BMC Systems Biology* 12 (Suppl 1): 9.

Yu, Liang, Jianbin Huang, Zhixin Ma, Jing Zhang, Yapeng Zou, and Lin Gao. 2015. “Inferring Drug-Disease Associations Based on Known Protein Complexes.” *BMC Medical Genomics* 8 Suppl 2 (May): S2.

Zha, Ling, Yuri Kitamura, Tetsuhisa Kitamura, Rong Liu, Masayuki Shima, Norio Kurumatani, Tomoki Nakaya, Junko Goji, and Tomotaka Sobue. 2019. “Population-Based Cohort Study on Health Effects of Asbestos Exposure in Japan.” *Cancer Science*, January. https://doi.org/10.1111/cas.13930.

Zhang, Shu-Bo, and Qiang-Rong Tang. 2016. “Protein-Protein Interaction Inference Based on Semantic Similarity of Gene Ontology Terms.” *Journal of Theoretical Biology* 401 (July):

30–37.

Zhang, Wen, Xiang Yue, Weiran Lin, Wenjian Wu, Ruoqi Liu, Feng Huang, and Feng Liu. 2018. “Predicting Drug-Disease Associations by Using Similarity Constrained Matrix Factorization.” *BMC Bioinformatics* 19 (1): 233.

Zhang, Zhongheng. 2016. “A Gentle Introduction to Artificial Neural Networks.” *Annals of Translational Medicine* 4 (19): 370.