Chemical and Textual Embeddings for Drug Repurposing

Galia Nordon
Technion- Israel
Institute of Technology
Haifa, Israel

Levi Gottlieb*
Rafael
Haifa, Israel

Kira Radinsky
Technion- Israel
Institute of Technology
Haifa, Israel

Abstract

Drug approval is a long and expensive process, that can take 10-15 years and more than 2 billion dollars. Therefore alternative techniques, such as drug repositioning, to identify new uses for approved drugs, has been gaining increasing attention. We examine the employment of different drug embeddings to predict successful drug repositioning. We study the employment of drug molecular structure and show that using larger chemical construct, such as large functional chemical groups, is much more effective than small sub-structures. We then study embeddings that are based on textual medical publications and compare them with the chemical-structure-based embeddings. We eventually present a novel embedding technique to combine the merit of the textual and chemical-based approaches. We provide empirical results on a repositioning benchmark set. Additionally, we present an application of such embedding as part of an ongoing repositioning research conducted with a major health care supplier, and identify a novel drug and indication. The pair has been verified on a corpus of 1.5 million patient EHR data.

1 Introduction

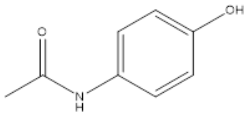
Drug development is a process that can last between 10 and 15 years, while only 0.1% of the drugs that enter pre-clinical testing progress to human testing, and only 20% of these are approved by the FDA (Suresh and Basu 2008). Given the low probability of a drug to succeed and the long development cycle of a new drug, repurposing of existing drugs to treat diseases became ever more attractive. Drug repurposing, also know as repositioning, is the process of discovering new indication for known drugs. Historically, the identification of such drugs was not systematic, but rather more fortuitous. A famous example is the repurpsoing of Sildenafil, an antihypertensive drug, to treat erectile dysfunction. In recent years, several computational approaches were developed for the task, including: (1) Electronic health records (EHRs) based approaches (Nordon et al. 2019); (2) Identifying genes associated with a disease; (3) Predicting binding site between a ligand (e.g., a drug) and a target protein using an optimization process of the chemical structure of the drug to best fit the protein structure; (4) Signature matching of a drug and a disease; In this work we focus on signature matching for the task of drug repurposing.

Signature matching requires representing a drug and a disease. In recent years, we identified two lines of research: (1) Chemical representations: computational approaches focusing on the drugs' chemical structures and their relationship to the biological activity (Jaeger, Fulle, and Turk 2018); (2) Natural language representations: literature-based-discovery approaches that aim to represent the drugs from medical publications (Yang et al. 2017; Weeber 2007); However, neither has been translated to a significant applicative success. In this paper, we study the merit of each representation, and present a novel approach for combining the two. We present a joint embedding of text and chemical structure in a joint representation space and show empirical gain for the task of drug repurposing.

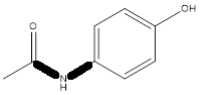
A chemical drug embedding is created relying on the drug's chemical structure. This structure is responsible for its functionality and effect on the body. There are several methods to represent a chemical compound: (1) A structural formula: a graph-like representation of a molecule, which is descriptive, but computationally expensive and is difficult to process. (2) Chemical nomenclature: SMILES (Weininger 1988) and IUPAC names (Favre and Powell 2013) are systematic naming conventions for chemical compounds. They are much easier to process by computational systems, although lack a good representation of the 3D structure of the molecule. In contrast to text-based embedding which is limited to words that exist in training vocabulary, chemical-based embedding can be used for novel molecules. This property is useful for drug discovery research. We study such representation for the task of drug repurposing and show that choosing the correct chemical representation is important. We show that using larger construct, such as large functional groups (IUPAC), is much more effective than small sub-structures (Jaeger, Fulle, and Turk 2018).

In recent years, several attempts were made to present textual drug embedding. Most methods (Zhang et al. 2019; Ngo et al. 2016) create the representation based on the tex-

*Dr. Gottlieb is an organic chemist who contributed to the chemical aspects of this paper.



(a) Structural formula, Name, IUPAC name, and SMILES notation



(b) Graph representation

Figure 1: Paracetamol Molecule

tual context of the drug as mentioned in medical literature. Algorithms such as word2vec (Mikolov et al. 2013b; De Vine et al. 2014), FastText (Bojanowski et al. 2017; Zhang et al. 2019) and Elmo (Peters et al. 2018) were shown to create useful representations in other domains. However, they are missing important factors. For example, it is not likely that the number of carbon atoms in a drug's molecule will be a differentiating factor between embeddings created based on text from general biomedical papers since these papers do not usually discuss such factors. Other chemical properties such as weight, solubility, toxicity may be more prominent but again, this is highly dependant both on the corpus being used for training (general medicine, pharmacology, organic chemistry, etc.) and also on the size and extent of the corpus. Another limitation of such embeddings is that they are limited to entities which appear in the training corpus and it is hard to use them for novel drugs. FastText mediates this problem somewhat by encoding sub-strings of each word in the corpus, hence capturing similar words that may have a very similar meaning. This is however a partial solution to the case of drugs and other chemical structures as they may have very different starting representations. For example, Acetaminophen, Parachetamol, and N-(4-hydroxyphenyl)ethanamide all refer to the same drug but are very different textually.

We study the trade-offs between the textual and chemical representations. We show that both are useful in predicting drugs to be tested for repurposing (Section 6.1). This might entail that there is a similar bias for selecting drug candidates for repurposing by researchers based on literature they read and chemical structures they study. We then present a novel embedding methodology that captures both the text context of the drugs and their chemical structure (Section 3.3). We leverage a corpus of 10562 drugs and indications tested for repositioning, and show that for the task of successfully reposition prediction, this hybrid approach yields the highest performance (Section 6.2). We then present a novel drug and indication our system identifies and verify it on a corpus of 1.5 million patient EHR data (Section 7).

2 Related Work

The chemical structure of drugs has been the subject of biomedical related research. Harel and Radinsky use generative models to discover novel drugs. The embeddings they use rely on chemical structure of known drugs (prototypes). Similarly, MolGAN (De Cao and Kipf 2018) is a generative model for molecule generation. It can be used with reinforcement learning to generate molecules with desired properties. Coley et al. used a graph convolutional neural network to predict products of organic reactions. We assess the common molecular-level drug representations, and show they perform poorly in the task of drug repositioning.

Literature based discovery (Swanson 1986) has been applied successfully to the fields of medicine and biomedicine. Both using textual repositories (Swanson and Smalheiser 1999; Spangler et al. 2014; Sybrandt, Shtutman, and Safro 2017; Lally et al. 2017) and in combination with other data (Choi, Chiu, and Sontag 2016; Nordon et al. 2019). The task of biomedical word embedding has also been the subject of diverse research, Choi, Chiu, and Sontag, show that biomedical embeddings based on different sources can differ in their representation of relations between concepts.

In this work, we apply both text and chemical representations for the task of predicting which drugs will be attempted by researchers for repositioning and for the task of predicting the repositioning result. We present a methodology for combining both text and chemical representations and show its merit for the latter task.

3 Embedding Algorithms

We review the different embedding methodologies, which we test for the task of drug repositioning (Section 6).

3.1 Text-based Drug Embeddings

Text-based embedding algorithms create representation of words based on the text and context in which they appear. The Word2Vec algorithm (Mikolov et al. 2013a) creates embeddings for words in a given corpus using a model trained to predict a word given the words around it (CBOW) or to predict the words around a given word (skip-gram). The result is a representation that utilizes the word's context and has been shown to be very expressive. Generally speaking, in this embedding, both drugs and indication are represented by the words describing them in the text. We used a corpus of sentences extracted from SemMedDB(Kilicoglu et al. 2012), containing 5,814,504 sentences from PubMed titles and abstracts that have been preprocessed with SemRep(Rindflesch and Fiszman 2003). The preprocessing includes a named-entity recognition step in which UMLS entities in the sentence are identified and mapped to their UMLS unique identifier (CUI). We chose to use this corpus due to this important named entity recognition step which mitigates much of the potential obscurity in the text.

3.2 Chemical-based Drug Embeddings

Chemical materials, including drugs can be described by their chemical structure. In these embeddings, we describe drugs based only on their molecular structure. We explore two types of common chemical embeddings:

Mol2Vec Representation. Extended Connectivity Fingerprints (ECFP) are topological sub-structures of molecules

that are used for measuring molecular similarity, characterising molecules etc. Consider a graph representation of a molecule as presented in Figure 1. A finger print is created starting from one vertex in the graph and traversing its surroundings with radius r. Different radii and different starting points may be used to create a set of fingerprints for a molecule. In Figure 1b, a fingerprint of radius 1 from the NH vertex is marked in bold.

Mol2vec (Jaeger, Fulle, and Turk 2018) transforms the SMILES representation of a molecule into a 'sentence' constructed of Extended Connectivity Fingerprints (Rogers and Hahn 2010). In our example, the SMILES representation of paracetamol is Oc1ccc(NC(C)=)cc1. This will be transformed into: *fingerprint(O,r=0) fingerprint(O,r=1) fingerprint(c,r=0) fingerprint(c,r=1) fingerprint(c,r=0) fingerprint(c,r=1)*, etc. Each fingerprint is a "word" and they are ordered according to the atoms in the SMILES representation. A word2vec model is then trained on the resulting corpus to produce an embedding for each fingerprint. The molecule's embedding is an aggregate of the embeddings of its fingerprints. Mol2Vec uses radii of 0, which is the vertex itself, and 1. Our paracetamol molecule will therefore be translated into a sentence of 22 words since there are 11 atoms in the molecule (excluding hydrogen) and we construct two fingerprints for each (at radii 0 and 1).

IUPAC Representation. We suggest using an alternative chemical representation based on IUPAC names. IUPAC names are chemical nomenclature standardized by the International Union of Pure and Applied Chemistry (IUPAC). The molecule name is constructed by concatenating the names of the functional groups in the molecule, their respectful position, and spatial relations between atoms and groups. For example: the IUPAC name of Paracetamol is N-(4-hydroxyphenyl)ethanamide. It consists of three functional groups: hydroxyl, phenyl and ethanamide.

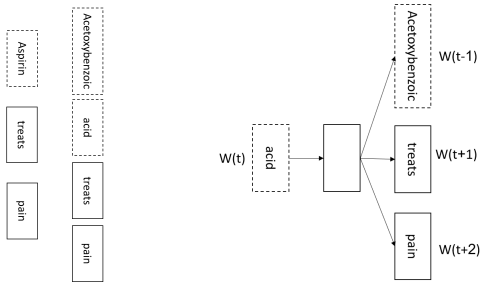
Functional groups are substructures of molecules which take part or affect the molecule's interactions. They are therefore more relevant to the material's effect. A chemical structure may be "broken" into an array of the simplest (primary) functional groups connected by linking atoms and chains. However, it is much more useful for our task not to use an assembly of the simplest functional groups but instead a higher hierarchy of building units (secondary functional groups). Using Paracetamol as an example, the hydroxyl and phenyl functions will be considered as a hydroxyphenyl unit. This representation uses larger constructs of the molecule which have a functional significance. We suggest a simple onehot encoding: each drug is represented by a vector v of length l, where l is the number of all IUPAC-derived functional groups in our data. We set $v[i] = 1$ if the drug contains the functional group and 0 otherwise.

3.3 Hybrid Model: Text and Chemical-based Drug Embeddings

The text representation carries some information about potential side effects which are not always conveyed in the chemical structure. On the other hand, the chemical structure holds deeper information about the molecule. We therefore present a hybrid model that combines the added information from the molecular structure with the expressive power of the drugs context. We refer to this method as IUPAC word2vec.

Table 1: Text Corpus Example

Processing	Sentence
Original sentence	Aspirin treats pain
word2vec corpus	C0004057 treats C0030193
IUPAC word2vec corpus	Acetoxybenzoic acid treats C0030193



(a) Drug translation (b) word2vec skip-gram model (Mikolov et al. 2013b)

Figure 2: IUPAC-based Word2vec Illustration

We enable this by inserting key structural components of the chemical representation into the text. We learn the representation of the chemical substructures given the textual context along with learning the representation of the words in the text given those substructures. The final chemical structure of the drug is then constructed by combining the representations of its chemical components (using mean aggregation). Specifically, we converted each drug CUI to the functional groups derived from its IUPAC name, therefore the IUPAC based vocabulary contains mainly functional groups and UMLS CUIs (for indications). Table 1 presents and example of the texts used for training the text-based representation (Section 3.1), and the transformed texts used for training the IUPAC word2vec representation.

Figure 2 illustrates this process (using the skip-gram method). The sentence "Aspirin treats pain" is translated into the sentence "Acetoxybenzoic acid treats pain", which is then used in the prediction model. The words "Acetoxybenzoic", "treats", and "pain" are predicted based on the input word "acid".

Formally, let $D = \{d \in C | d \textit{ is a drug}\}$ be a list of drugs, and $W = \{w \in C | w \notin D\}$ be words in a vocabulary, and $FG = \{fg \in IUPAC(d), \forall d \in D\}$ be the set of functional chemical groups. For a given corpus C we define: $A = W \cup FG$. The objective of the skip-gram model for our new corpus containing chemical structures is therefore given in Equation 1 by:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} \log p(a_{t+j}|a_t), \forall a \in A \qquad (1)$$

The softmax function is defined as:

$$p(a_O, a_I) = \frac{\exp\left({v'_{a_O}}^T v_{a_I}\right)}{\sum_{a=1}^{A} \exp\left({v'_a}^T v_{a_I}\right)} \quad (2)$$

where v_a and v'_a are the input and output vectors representing a and A is the number and functional groups derived from drug names in the vocabulary. As in (Mikolov et al. 2013b), negative sampling is used for complexity efficiency, and defined by the following objective:

$$\log \sigma(v'^T_{a_O} v_{a_I}) + \sum_{i=1}^{k} E_{a_i \sim P_n(a)}[\log \sigma(-v'^T_{a_i} v_{a_I})] \quad (3)$$

4 Dataset

RepoDB (Brown and Patel 2017) is a dataset developed as a benchmark for computational repurposing tasks. Based on ClinicalTrials.gov and DrugCentral[1], it contains pairs of drugs and indications. For each pair a status is provided: approved, terminated, withdrawn, or suspended. We differentiate between the status approved and the other three status which we collectively treat as unapproved. RepoDB contains 10562 such pairs, out of which 6677 are approved. Indications are identified using the UMLS CUIs and drugs using their DrugCentral id. We further map drug id to UMLS CUI and translate the drug to its IUPAC name. We retrieved the UMLS CUI for each drug in RepoDB using UMLS REST Api [2]. We retrieved the IUPAC name for each CUI using the NCI/CADD Chemical Identifier Resolver [3].

Date of status is not contained in the dataset. For our tasks, we add date of status by using the first date the two terms were mentioned together in a paper abstract on PubMed.

The data set contains 2074 unique indications with varying frequencies. The indications differ between status: there are 1229 unique indications with status approved, 1014 indications that are not approved, and only 169 unique indications that appear with both status approved or not approved in the data set. For drugs we see a similar dispersion. Out of 1572 unique drugs there are 1519 approved and 463 not approved. Only 410 unique drugs appear in the dataset with both status approved and not approved. In order to mitigate this bias, we limit the dataset to indications for which there are examples both in the approved and the unapproved groups. The resulting data set consists of 1290 pairs. We label each approved drug-indication pair as true and pairs with any other status (terminated, withdrawn or suspended) as false.

5 Empirical Evaluation

Our evaluation consists of two parts. First we create the four embedding models discussed in Section 3. Then, we use the RepoDB dataset to train simple models for two tasks: (1) Predict the results of a drug (re)purposing task; (2) Predict whether a drug (re)purposing task will be explored.

[1]http://drugcentral.org/

[2]https://documentation.uts.nlm.nih.gov/rest/home.html

[3]https://cactus.nci.nih.gov/chemical/structure

We use the same dataset to train and test our models. Minor differences between model datasets may accrue due to local missing information. For example: if a SMILES representation is missing or faulted, mol2vec embedding for that drug will be missing and the specific data point will be discarded.

5.1 Task 1: Predicting Drug Repurpose Study

Drug approval is a long and expensive process. The decision to initiate it is not taken lightly, considering various technical, scientific and financial aspects. Both the drugs being examined and the indications for which they are tested may be influenced by limitations of technology, recent discoveries, funding and facilities, and possibly even research trends. We therefore choose our first prediction task to be predicting whether a drug-indication pair will appear in RepoDB, regardless of the outcome. For this task, true pairs are pairs that appear in the RepoDB data set and false pairs are pairs which do not appear in it. We create false pairs by sampling from the drugs and indications in RepoDB, thus imitating their distribution in the real data. We experimented on several embedding sizes (on a external validation set), and as the dataset is small, the best performing word2vec embeddings was set to length 20. Similarly, for the IUPAC one hot embedding we set the dimensions of the onehot vector to 100 using PCA. For both, we train a simple neural network with one hidden layer. We train our networks on 80% of the data and test on 20%.

5.2 Task 2: Predicting Drug Repurpose Success

We train a simple neural network to classify between approved and unapproved pairs. As a training set, we used the data approved up to a cutoff year, and the test set is the data after that cutoff year. We tested three possible values for cutoff year: 2014,2000, and 1990. Naturally, the size of the training set increases with the years and the size of the test set decreases. The sizes of the training set ranges from 700 pairs in 2014 to 300 pairs in 1990 while the test set sizes ranges from 16 in 2014 to 350 in 1990. We use the same simple networks as described in Section 5.1

6 Results

6.1 Task 1: Predicting Drug Repurpose Study

Table 2 presents the average ROC AUC for 10 experiments. Overall, the onehot IUPAC and text word2vec embeddings give significantly better results than the two versions of IUPAC-based words2vec and mol2vec embeddings. We assume the significant difference (4σ) between the results in line 1 and 4 stems from the difference in drug embedding (as both have identical embeddings for indications), supporting our hypothesis that larger functional groups as expressed in IUPAC names are more effective than molecular fingerprints for our task. In fact, our results show that text word2vec embedding of drugs is superior to mol2vec (1.9σ). Onehot IUPAC is superior to both versions of IUPAC-based word2vec (3.12σ), suggesting that for this task, a hybrid representation for drugs is sub-optimal.

Table 2: Drug Study Prediction (AUC)

	Embedding	ROC AUC	std
1	onehot IUPAC	**0.75**	± 0.01
2	word2vec	0.70	± 0.03
3	IUPAC word2vec	0.65	± 0.03
4	mol2vec	0.62	± 0.03

Table 3: Drug Reurposing Prediction (AUC)

Cut year	IUPAC one hot	w2v	IUPAC w2v	m2v
2014	0.73 ± 0.04	0.7 ± 0.05	**0.84**± 0.03	0.76 ± 0.05
2000	0.70 ± 0.01	0.63 ± 0.01	**0.72**± 0.02	0.6 ± 0.01
1990	0.64 ± 0.01	0.52 ± 0.01	**0.7**± 0.02	0.6 ± 0.01

6.2 Task 2: Predicting Drug Repurpose Success)

Table 3 summarizes the results. As mentioned in Section 5.2, cutting the data at year 2014 produces the largest training set but a very small test set (16 to 29 examples). Cutting the data at year 1990 produces a training set and test set that are roughly the same size. We therefore focus on the experiment where the data is cut at year 2000, in this case the test set size is roughly 20% of the size of the training set.

We observe that IUPAC-based embeddings (onehot or word2vec) are superior. Mol2vec embeddings achieved comparable results to text-word2vec embeddings, although their drug embeddings are based on completely different datasets – text for word2vec and molecular fingerprints for mol2vec. A possible explanation for this result is that the classification model relies on the indication information for these cases. The hybrid IUPAC-based word2vec reaches the highest performance. We suggest that the superior performance is due to its more expressive representation of drugs. This allows the prediction model to use drug embeddings along side indication embeddings for its predictions. It is generally believed that text-based embedding of concepts is useful due to the fact that it utilizes the context of each word. In this example, we see that this might not be the best solution. The embedding based of pure chemical structure represented as IUPAC is superior.

7 Application: The Case of Doxazosine

One of the interesting applicative aspects of our finding is the ability to embed novel molecules using their IUPAC representation and use this embedding in machine learning task. We wish to test this and identify a new indication for a drug.

We received access electronic health records (EHR) collected for over 10 years from patients of a prominent health care provider in Israel. The data contains prescription purchases, diagnosis, demographic data, measurements and lab test results for more than 1.5M patients over these years. We select the Diabetes indication, as it has enough patient records and the disease improvement can be measured based on EHR data (such as blood tests). We apply our algorithms on pairs of drugs and the diabetes indication.

We wish to test whether patients taking the drugs ranked high by our system, indeed got better – i.e., their diabetes Hemoglobin A1C (HbA1c) values improved. In an attempt to eliminate as much patient variability among drug choices, we used propensity score matching (Rosenbaum and Rubin 1983) to examine whether a specific drug treatment achieved independently higher success rates. We used the following patient characteristics for the matching: DM-2 drug treatment, weight, age, BMI and smoking status. Re-sampling was allowed in the matching process. Patients with HgA1c levels higher than 6.5 were classified as successful treatment. In this section, we present a qualitative example for one such drug, which was explored by our partners in the medical center and verified.

Alpha 1 blockers are currently used for treatment of benign prostatic hyperplasia and were found to be correlated with improvement in type 2 diabetes ($p < 0.01$ in groups of over 200 patients). Doxazosine is an alpha 1 blocker. It does not appear in the repoDB dataset. Hence, for the scope of our experiment it is considered a novel drug. The IUPAC name of the drug is: *(RS)-2-[4-(2,3-Dihydro-1,4-benzodioxine-2-carbonyl)piperazin-1-yl]-6,7-dimethoxyquinazolin-4-amine*. We create a onehot IUPAdoxazosineC embedding for doxazosine using the functional groups from its name: *Dihydro, benzodioxine, carbonyl, piperazin, yl, dimethoxyquinazolin, amine*.

Comparing a sub-group of patients from the above that was additionally treated with alpha1 blockers (for benign prostate hyperplasia) with a matched group of patients that did not receive alpha blockers showed highly significant success rates for the treated group: 61% success rate for treated group and 53% success for untreated group. A chi-squared test with p-value 0.0004 and test statistic of 16.7. The groups were matched using propensity score matching. The treated group contained 1356 patients and the untreated group contains 1221 patients.

We look at the result of the classifier for predicting drug repurpose success for the pair doxazosine- non-insulin dependant diabetes and get a positive score (0.85). That is, our model predicts with high confidence that doxazosine will be successfully repurposed for treatment of non-insulin dependant diabetes. Interestingly, the result for predicting drug repurpose study for the same pair is rather low (0.27) indicating that the study of this drug for the treatment of non-insulin dependant diabetes is less likely to be conducted.

8 Conclusion

In this paper, we compared drug embeddings based on text, chemical structure and a combination of both in regards to two drug (re)positioning tasks. We show that embeddings based on chemical representations are as good as and sometimes better than text based embeddings. We further show that choosing the right chemical representation is crucial as using large functional group is superior to small, non-functional structures such as extended connectivity fingerprints. We show that the novel approach presented in this paper for combining textual and chemical representation is significantly better in the predicting repurpose success. Finally, we present a qualitative example of an immediate application of these results as the ability to embed novel molecules using their IUPAC representation and use this embedding in machine learning tasks. We plan to further combine this

embedding into the EHR based repurposing framework examining more repurposing candidates.

References

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.

Brown, A. S., and Patel, C. J. 2017. A standard database for drug repositioning. *Scientific data* 4:170029.

Choi, Y.; Chiu, C. Y.-I.; and Sontag, D. 2016. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings* 2016:41.

Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; and Jensen, K. F. 2019. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science* 10(2):370–377.

De Cao, N., and Kipf, T. 2018. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*.

De Vine, L.; Zuccon, G.; Koopman, B.; Sitbon, L.; and Bruza, P. 2014. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, 1819–1822. ACM.

Favre, H. A., and Powell, W. H. 2013. *Nomenclature of organic chemistry: IUPAC recommendations and preferred names 2013*. Royal Society of Chemistry.

Harel, S., and Radinsky, K. 2018. Prototype-based compound discovery using deep generative models. *Molecular pharmaceutics* 15(10):4406–4416.

Jaeger, S.; Fulle, S.; and Turk, S. 2018. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling* 58(1):27–35.

Kilicoglu, H.; Shin, D.; Fiszman, M.; Rosemblat, G.; and Rindflesch, T. C. 2012. Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics* 28(23):3158–3160.

Lally, A.; Bagchi, S.; Barborak, M. A.; Buchanan, D. W.; Chu-Carroll, J.; Ferrucci, D. A.; Glass, M. R.; Kalyanpur, A.; Mueller, E. T.; Murdock, J. W.; et al. 2017. Watsonpaths: scenario-based question answering and inference over unstructured information. *AI Magazine* 38(2):59–76.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Ngo, D. L.; Yamamoto, N.; Tran, V. A.; Nguyen, N. G.; Phan, D.; Lumbanraja, F. R.; Kubo, M.; and Satou, K. 2016. Application of word embedding to drug repositioning. *Journal of Biomedical Science and Engineering* 9(01):7.

Nordon, G.; Koren, G.; Shalev, V.; Horvitz, E.; and Radinsky, K. 2019. Separating wheat from chaff: Joining biomedical knowledge and patient data for repurposing medications. In *Proceedings of the 33rd AAAI conference on Artificial Intelligence*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Rindflesch, T. C., and Fiszman, M. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics* 36(6):462 – 477. Unified Medical Language System.

Rogers, D., and Hahn, M. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling* 50(5):742–754.

Rosenbaum, P. R., and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Spangler, S.; Wilkins, A. D.; Bachman, B. J.; Nagarajan, M.; Dayaram, T.; Haas, P.; Regenbogen, S.; Pickering, C. R.; Comer, A.; Myers, J. N.; et al. 2014. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1877–1886. ACM.

Suresh, P., and Basu, P. K. 2008. Improving pharmaceutical product development and manufacturing: impact on cost of drug development and cost of goods sold of pharmaceuticals. *Journal of Pharmaceutical Innovation* 3(3):175–187.

Swanson, D., and Smalheiser, N. 1999. Implicit text linkages between medline records: Using arrowsmith as an aid to scientific discovery. *Library Trends* 48(1):48–61.

Swanson, D. R. 1986. Fish oil, raynauds syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine* 30(1):718.

Sybrandt, J.; Shtutman, M.; and Safro, I. 2017. Moliere: Automatic biomedical hypothesis generation system. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, 1633–1642. New York, NY, USA: ACM.

Weeber, M. 2007. Drug discovery as an example of literature-based discovery. In *Computational Discovery of Scientific Knowledge*. Springer. 290–306.

Weininger, D. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* 28(1):31–36.

Yang, H.-T.; Ju, J.-H.; Wong, Y.-T.; Shmulevich, I.; and Chiang, J.-H. 2017. Literature-based discovery of new candidates for drug repurposing. *Briefings in bioinformatics* 18(3):488–497.

Zhang, Y.; Chen, Q.; Yang, Z.; Lin, H.; and Lu, Z. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data* 6(1):52.