CRF-LSTM Text Mining Method Unveiling the Pharmacological Mechanism of Off-target Side Effect of Anti-Multiple Myeloma Drugs

Kaiyin Zhou [1] **Sheng Zhang** [2] **Xiangyu Meng** [3] **Qi Luo** [2] **Yuxing Wang** [1]
Ke Ding [2] **Yukun Feng** [2] **Mo Chen** [1] **Kevin B Cohen** [4] **Jingbo Xia** [1*]
1. College of Informatics, Huazhong Agricultural University, China
2. College of Science, Huazhong Agricultural Univerisity, China
3. Center for Evidence-based and Translational Medicine, Zhongnan Hospital of Wuhan University, China
4. School of Medicine, University of Colorado, U.S.
*. Correspondence author: xiajingbo.math@gmail.com

Abstract

Off-target effects played a vital role in the pharmacological understanding of drug efficacy and this research aimed to use text mining strategy to curate molecular level information and unveil the mechanism of off-target effect caused by the usage of anti-multiple myeloma (MM) drugs. After training a hybrid CNN-CRF-LSTM neural network upon the training data from TAC 2017 benchmark database, we extracted all of the side effects of 16 anti-MM drugs from drug labels, and combined the results with existed database. Afterwards, gene targets of anti-MM drugs were obtained by using structure similarity, and their related phenotypes were retrieved from Human Phenotype Ontology. Furthermore, linked phenotypes to candidate genes and adverse reaction of known drugs formed a knowledge graph. Through regulation analysis upon intersected phenotypes of drugs and target genes, an off-target effect caused by SLC7A7 was found, which with high possibility unveiled the pharmacological mechanism of side effect after using combination of anti-MM drugs.

1 Introduction

Drug genetics aimed to discern the association between drugs and adverse reaction, and allowed to personalized medication (Stephen, 2011). Associating off-target effects with adverse reaction of drugs to discover the new pharmacological effect of them is a daunting task when using experimental method alone.(Eugen et al., 2012).

As a pioneer work, Lountkine et al.,(Eugen et al., 2012) explored a computational method to predict novel off-target effects of 656 marketed drugs. By using the chemoinformatics information, like ligand affinity, Similarity Ensemble Approach (SEA) (Keiser et al., 2009) was used to calculate structural similarity of drugs and targets, and the relations between drugs and targets was rebuild. In the meantime, the adverse reaction (ADR) of drug targets were curated from authoritative database including DrugBank, GeneGo Metabase, and Thompson Reuters Integrity. Thus, a large scale drug-target-ADR network was built, and the coincidental overlap of ADR among target gene and drugs potential gene gave illuminative explanation for the mechanism of off-targets effect.

Generally, the mechanism of off-target candidate filtering requires the prerequisite of target-drug pair indications. So far, this pair information has been widely predicted by inferring the similarity both in chemical structure and relevance info. Andreas et al., (Andreas et al., 2007) used chemical structure information to infer the drug-target pair, while Monica et al., (Monica et al., 2008) used phenotypic side effect similarities to make the inference. As a large-scale bioinformatics attempt, Mohan et al., (Mohan et al., 2008) exploited a huge training set of 10 million compounds with known in-vitro activities, predicted both primary and secondary pharmacology for 1279 molecules, and over 30 thousands possible interactions were predicted for these drugs.

Multiple myeloma (MM) is one of the most common hematological malignancies, the incidence of which ranks second just next to non-Hodgkin lymphoma. Although recent advances in MM treatment has largely improved the patients clinical outcome, it remains an incurable disease due to drug-resistance and relapse which are almost inevitable (Terpos, 2017). Common adverse drug reactions (ADRs) related to anti-MM treatment include hematologic toxic effects (eg. anemia, neutropenia and thrombocytopenia), thrombosis, impaired immune function, pe-

Proceedings of the BioNLP 2018 workshop, pages 166–171
Melbourne, Australia, July 19, 2018.

ripheral neuropathy, and gastrointestinal toxic effects (eg. mucositis, diarrhea), among many others. These ADRs bring harm to patients health and quality of life, and may result in premature discontinuation of treatment due to intolerance to side effects. Since the underlying mechanisms are largely unclear, currently they are mostly managed with symptomatic and/or supportive care, along with dosage reduction or treatment discontinuation (McCullough et al., 2018). A better understanding on the mechanisms will help us find ways to effectively cope with the above mentioned safety concerns in treating MM.

In this research, we proposed a novel pharmacological knowledge discovery strategy which integrated both Biomedical natural language processing (BioNLP) and medical informatics. The adverse reactions (ADRs) were trained by newly released ADR training data (Demner-Fushman et al., 2018), and were extracted on-line with large-scale of text mining upon 16 anti-MM drugs by using conditioned random field (CRF) and long short term memory (LSTM) neural networks. Subsequently, Human Phenotype Ontology (HPO) (Sebastian et al., 2017) and Ligand Similarity prediction were used to calculate the target phenotypes. Bioinformatics analysis hinted that an off-target gene, SLC7A7, played vital role in the side effect of a combination usage of anti-MM drugs.

2 Material and Method

2.1 Data Resource

Marketed drugs for MM were collected from drugs.com (Drugs). After searching anti-MM chemicals and removing drug synonyms, 16 drugs were extracted from the original pharmaceutical list, and drug targets were collected from SwissTargetPedction (David et al., 2014), as shown in supplementary table, Table S1 (Sixteen anti-MM drugs their possible targets). Meanwhile, drug labels were extracted from DailyMED database (National Library of Medicine and Services, 2005).

Human Phenotype Ontology (HPO) (Sebastian et al., 2017) provides standardized vocabulary of phenotypic abnormalities in human diseases. From HPO, matches of target genes and their corresponding phenotype terms were retrieved, as shown in table S2(Phenotype matching result for specific gene).

2.2 Sequence labeling by BioNLP Algorithm

2.2.1 Vector representation of tokens

Regarding the input form for a neural network, word embedding, controlled vocabulary - DISORDER, and part of speech (POS) are used for vector representation of tokens.

- Pre-trained Embeddings: Compared with randomly initialized word embeddings, pre-trained word embeddings generally yield better experimental results. 200 dimensional embeddings of GloVe ((Pennington et al., 2014)) was chosen, instead of word2vec word vectors, as GloVe is more preferable for named entity recognition tasks than word2vec (Ma and Hovy, 2016).

- DISO is a standardized dictionary from Metathesaurus of UMLS. The dictionary consists of the following 12 subtypes, i.e. acquired abnormality, anatomical abnormality, cell or molecular dysfunction, congenital abnormality, disease or syndrome, experimental model of disease, finding, injury or poisoning, mental or behavioral dysfunction, neoplastic process, pathologic function, and sign or symptom.

- The NLTK toolkit is taken into consideration to obtain the POS of each token. Randomly initialized feature weights was assigned to each POS type, and a lookup operation convert each sentence into a POS-embedding vector.

2.2.2 Integration of CRF and LSTM for sequence labeling

For sequence labeling task as ADR extraction, CRF is a popular mathematical method which defines the probability of the annotation of the label sequence $\mathbf{L} = (l_1, l_2, ..., l_l)$, given the observation sequence $\mathbf{O} = (o_1, o_2, ..., o_l)$: $exp(\sum_j \lambda_j t_j(l_{i-1}, l_i, \mathbf{O}, i)) + \sum_k \mu_k s_k(l_i, \mathbf{O}, i))$, where $t_j(l_{i-1}, l_i, \mathbf{O}, i))$ is a transition feature function that represents the transition distribution of label pair $\{l_{i-1}, l_i\}$ based on observation sequence $\mathbf{O}$, while $s_k(l_i, \mathbf{O}, i)$ refers to state feature function that quantify the state distribution of the label y_i given the observation sequence $\mathbf{O}$. The mechanism of CRF is to optimize the parameters λ_j and μ_k, and maximize the probability of $P(\mathbf{L}|\mathbf{O})$: $P(\mathbf{L}|\mathbf{O}, \lambda, \mu) = \frac{1}{Z(\mathbf{O})} exp(\sum_j \lambda_j t_j(l_{i-1}, l_i, \mathbf{O}, i)) + \sum_k \mu_k s_k(l_i, \mathbf{O}, i))$, where $Z(\mathbf{O})$ is for normalization (Lafferty et al., 2001).

In the meantime, LSTM is a special Recurrent neural networks(RNNs) which could capture time dynamics via cycles in the graph, and especially, is capable of capturing long-distance dependencies with the employment of a special cell and three grates, i.e. input gate, forget gate, and output gate. Supposing that t represents a time point, x_t is the input vector at time t. i_t, f_t, c_t, o_t stand for different gates state at time t. W_i, W_f, W_c, W_o are the weight matrices for hidden state h_t. U_i, U_f, U_c, U_o denote the weight matrices of different gates for input x_t. b_i, b_f, b_c, b_o denote the

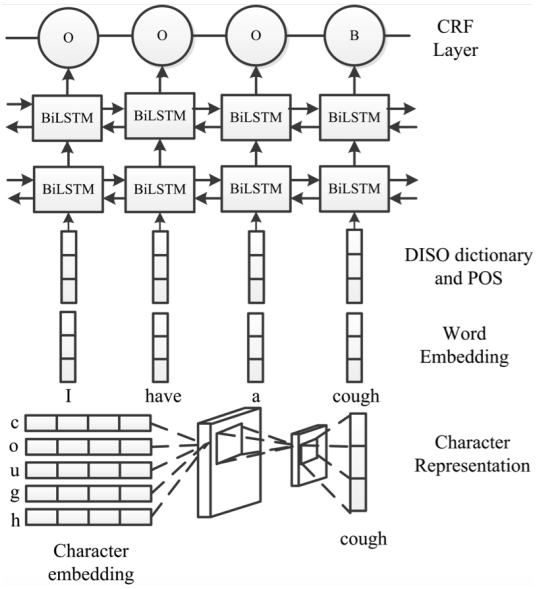

Figure 1: The idea of the CNN-LSTM-CRF sequence labeling method

bias vectors from different gates. And the formulas for LSTM unit at time t are:

$$\begin{aligned} i_t &= \sigma(W_i h_{t-1} + U_i x_t + b_i) \\ f_t &= \sigma(W_f h_{t-1} + U_f x_t + b_f) \\ c_t &= f_t * c_{t-1} + i_t * \tanh(W_c h_{t-1} + U_c x_t + b_c) \\ o_t &= \sigma(W_o h_{t-1} + U_o x_t + b_o) \\ h_t &= o_t * \tanh(c_t) \end{aligned} \quad , \quad (1)$$

where σ is the element-wise sigmoid function and $*$ is the element-wise product. And h_t is the hidden state, namely the finally output of LSTM unit at time t.

To achieve a better semantic understanding in biologic domain, a combined BLSTM-CNNs-CRF neural network was put forward by Ma et al.(Ma and Hovy, 2016), where CNNs are utilized to model character-level information, bi-directional LSTM (BLSTM) is used to capture past and future information respectively, and CRF is employed to decode the best label sequence. In order to further improve the labeling accuracy for this specific task, double-BLSTM layer is taken into consideration instead of single-BLSTM layer, namely BLSTM, mentioned in Ma et al.(2016).

The detailed algorithm steps are shown in Figure 1. For each word in training text, the character-level representation vector computed by CNN, the DISO and POS feature got by lookup random initialization weights, concatenated with word embedding vector are designed as the input of the double-BLSTM network. And the output vectors of double-BLSTM are fed to the CRF layers to jointly decode the best label sequence. The flowchart of a specific labeling employment example is presented in the following.

For instance, "I have a cough." where "cough" is the target word. After the sentence being separated into words, the words are broken into letters, which can be embedded into a one-hot vector to compute the character representation vector by CNN. The character-level representation vectors of each words, their DISO and POS representation vector and word embeddings, computed by glove, are combined as the inputs of double-BLSTM, which has double-layer of two processes, i.e. the past(left) and the future(right). The past process takes information only from 'I' to 'cough' while the future process takes information only from 'cough' to 'I'. These two pieces of information was concatenated as the final outputs of double-BLSTM and, simultaneously, the inputs of CRF. With the utilization of CRF, the labels of sentence are tagged as 'O O O B'.

2.3 Phenotype matching algorithm

To decide whether two phenotype words match or not, two criteria were applied. First, both phenotypes are available in the database with the same $is\_a \cdot ID$; second, the word embedding distance of two terms are small sufficiently. The algorithm is shown in the following.

- If both phenotypes are available in the database with the same $is\_a \cdot ID$, the output will be $True$.
- If not, each target phenotype is converted into a word embedding, and if the distance of the two vectors is less than a threshold value t, the two phenotype terms are matched. Otherwise, the two terms are not matched.

Algorithm 1 Phenotype matching algorithm

Input: Term A, term B, threshold value t
Output: $True/False$

```
if (A ∈ HPO) ⋀ (B ∈ HPO) ⋀ (A·is_a·ID = B·is_a·ID) then
    return True
else if Cosine Distance(A, B) < t then
    return True
else
    return False
end if
```

2.4 Flowchart of the proposed strategy for off-target side effect prediction

The purpose of this research is to find the co-occurrence of phenotype from both drug and the related protein, so as to illuminate the pharmacological mechanism of the drug side effect.

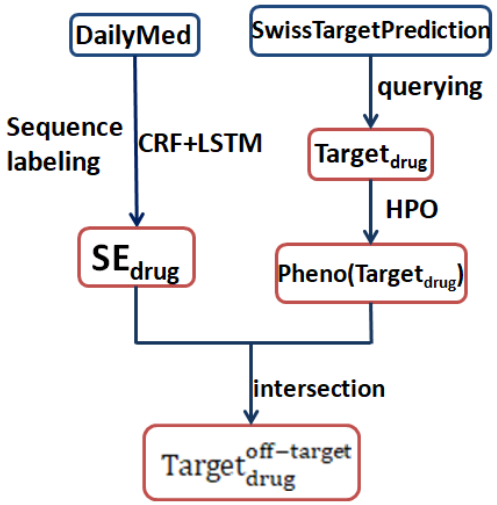

Figure 2: Flowchart of the proposed off-target mechanism discovery

By using an integration of the CRF and LSTM text mining algorithms, sequence labeling was carried on to extract side effects, SE_{drug}, of anti-MM drugs from DailyMed drug labels. Potential drug targets, $Target_{drug}$, were filtered by querying SwissTargetPredcition Database. Meanwhile, related phenotype of $Target_{drug}$, i.e., $Pheno(Target_{drug})$, was obtained by using Human Phenotyping Ontology (HPO). Subsequently, off-target gene, $Target_{drug}^{off-target}$, of corresponding drugs were filtered out by intersection analysis of SE_{drug} and $Pheno(Target_{drug})$.

3 Result

3.1 Database querying result

In total, 48 types of anti-MM drugs are collected by searching drug.com. And with the 48 drug names as searching condition, 16 different drugs and 16 corresponding labels are extracted from 27 drug labels, acquired by DailyMED. Among the 16 drugs, 2 are protein drugs, and the left 14 non-protein drugs are taken to predict their potential targets with the utilization of SwissTargetprediction, where 15 potential targets can be obtained from each drug. Searching the 15 potential targets in HPO, targets, not only our predicted targets but also targets existing in HPO, are achieved. With the application of HPO, drugs, potential target genes, and corresponding phenotype are related with each other. Meanwhile, corresponding ADRs form acquired drug labels can be collected with the strategy of sequence labeling, and improvement of the relationship between drugs and ADRs can be achieved through drugs.com, where related drugs' ADRs are collected. Eventually, drugs, potential targets, and data of overlapping ADRs are acquired via artificial recognition. And it is revealed in the result that under the circumstance of a certain drug, its potential targets have a tight relation with its ADRs.

3.2 Phenotype matching and phenotype coincidence

For the trained samples in table S3, F-Score and Matthews Correlation Coefficient (MCC) were calculated, and a best threshold $t = 0.57$ was obtained. Here $F - score = 2\frac{Precision \times Recall}{Precision + Recall}$, and $MCC = \frac{TP \times TN - FP \times FN}{\sqrt[2]{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$. The selection of t is shown in figure 3. The best F-score and MCC are 0.733, 0.622 separately.

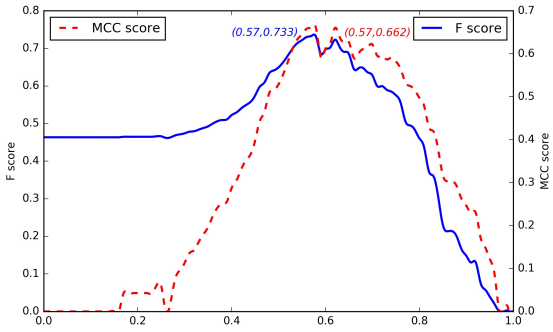

Figure 3: threshold selection for phenotype matching

By using algorithm 1, the gene whose phenotypes in HPO are highly consistent with drug ADRs were retrieved, and the coincidence were evaluated by Jaccard similarity coefficient. Among all the intersection of phenotype terms, the most prominent output pair is melphalan-SLC7A7 for Jaccard value being 0.280 and melphalan-CA2 for Jaccard value being 0.198. As shown in table 1, phenotype coincidence for melphalan and SLC7A7/CA2 is clear, that hinted that the two genes possibly play roles in the side effects of the drug.

3.3 Knowledge discovery of off-target side effect

An illuminative evidence comes from Melphalan, a common anti-MM drug. Through intersection analysis of $SE_{Melphalan}$ and $Pheno(Target_{Melphalan})$, anemia, thrombocytopenia and diarrhea were found to be the same phenotypes of the drug Melphalan and the possible target SLC7A . Observing its target genes are NR3C1, NR0B1, ANXA1, NOS2, NR1L2, and its possible target gene is SLC7A, we found that, after taking another anti-MM drug Prednisone, mRNA level of target genes goes down and that of SLC7A

Gene: SLC7A	
Known ADRs	**Off-target effect**
Sparse hair	Alopecia
Thrombocytopenia	**Thrombocytopenia**
Leukopenia	Leukopenia
Diarrhea	**Diarrhea**
Nausea	Nausea
Anemia	**Anemia**
Vomiting	**Vomiting**
Muscle weakness	Muscular parlysis
Respiratory insufficiency	Dyspnea

Table 1: Consistency of ADRs of melphalan alkeran evomela in clinical records and off-target curations

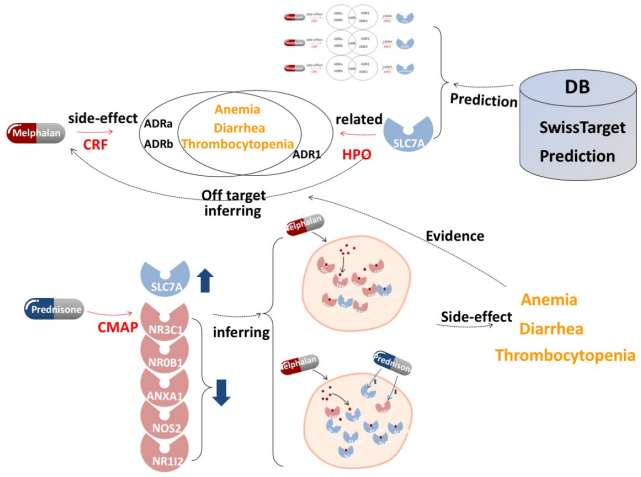

Figure 4: Mechanism of off-target side effects via functioning of SLC7A7 after MVP drug usage

goes up. That made it high chance for off-target event of SLC7A to manifest its off-target side effects: $Pheno(Target_{Melphalan})$. Thus SLC7A is with high chance the factor of the off-target effect.

4 Discussion

Mechanism of off-target effect is illustrated in this section. First, literature evidences are shown to address the side effect after anti-MM drug usage, and then the up/down regulatory mRNA-level tendency of on/off targets are shown.

4.1 Literature evidence

It was reported that a combined usage of melphalan, prednisone, and bortezomib (MPV) is regarded as common treatment for the high-risk MM patient, while neutropenia, thrombocytopenia, anemia, and gastrointestinal symptoms were common after MPV treatment (Kyle and Rajkumar., 2009).

Meanwhile, SLC7A is a heterotrimeric amino acid transporter (HAT) y+LAT-1 gene located on chromosome 14q11.2. It was reported that mutation in SLC7A caused Lysinuric Protein Intolerance. Then, delayed physical development, intestinal malabsorption, vomiting, and failure to thrive are the prominent clinical manifestations (Lawson and Loyd, 2013).

4.2 Up/Down regulation of target/off-target gene

Drug usage of Prednisone is treated as exposure in comparison analysis, and the connectivity map (CMAP) is used to unveil the up/down regulation by analyzing the before/after mRNA level of patience. We input the target genes as down regulated genes and the off-target genes as up regulated, and the output off-target gene is Prednisone, with significant P value, 0.01029.

As shown in figure 4, after taking Prednisone, as it mentioned above, the expression levels of target genes are down regulated while the off-target genes are up, the steady state is broken. In this condition, more off-target proteins lead to more combination with Melphalan than usual, which contribute to more significant side effects.

Here, we infer that the usage of Prednisone lead to an up regulation of SLC7A, and it arises competition between SLC7A and the drug targets, i.e., NR3C1, NR0B1, ANXA1, NOS2, NR1L2. The binding of SLC7A to Melphalan brings the off-target effect. Therefore, thrombocytopenia, anemia, and gastrointestinal symptoms can be easily observed after combined usage of Melphalan and Prednisone.

5 Conclusion

Sequence labeling of biomedical entities, e.g., side effects or phenotypes, was a long-term task in BioNLP and MedNLP communities. Thanks to effects made among these communities, adverse reaction NER has developed dramatically in recent years (Demner-Fushman et al., 2018). As an illuminative application, to achieve knowledge discovery via the combination of the text mining result and bioinformatics idea shed lights on the pharmacological mechanism research.

Acknowledgments

This work is funded by the Fundamental Research Funds for the Central Universities of China (Project No. 2662018PY096). We expressed our gratitude to Pierre Zweigenbaum for discussion of WAPITI and CRF, and to Köhler Sebastian et. al. for offering help in the HPO resource. We also thank anonymous reviewers for their kind suggestions.

References

Bender Andreas, Josef Scheiber, Meir Glick, John W. Davies, Kamal Azzaoui, Jacques Hamon, Laszlo Urban, Steven Whitebread, and Jeremy L. Jenkins. 2007. Analysis of pharmacology data and the prediction of adverse drug reactions and offtarget effects from chemical structure. *ChemMedChem*, 2(6):861–873.

Gfeller David, Aurlien Grosdidier, Matthias Wirth, Antoine Daina, Olivier Michielin, and Vincent Zoete. 2014. Swisstargetprediction: a web server for target prediction of bioactive small molecules. *Nucleic acids research*, 42(W1):W32–W38.

Dina Demner-Fushman, Sonya E Shooshan, Laritza Rodriguez, Alan R Aronson, Francois Lang, Willie Rogers, Kirk Roberts, and Joseph Tonning. 2018. A dataset of 200 structured product labels annotated for adverse drug reactions. *Scientific data*, 5:180001.

Drugs. Drugs.com. https://www.drugs.ca.

Lounkine Eugen, Michael J. Keiser, Steven Whitebread, Dmitri Mikhailov, Jacques Hamon, Jeremy L. Jenkins, Paul Lavan, and et al. 2012. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, 486(7403):361–367.

Michael J. Keiser, Vincent Setola, John J. Irwin, Christian Laggner, Atheir I. Abbas, Sandra J. Hufeisen, Niels H. Jensen, and et al. 2009. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181.

Robert A. Kyle and S. Vincent Rajkumar. 2009. Treatment of multiple myeloma: a comprehensive review. *Clinical Lymphoma and Myeloma*, 9(4):278–288.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

William E Lawson and James E Loyd. 2013. Interstitial and restrictive pulmonary disorders. In *Emery and Rimoin's Principles and Practice of Medical Genetics*, pages 1–22. Elsevier.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Kristen B McCullough, Miriam A Hobbs, Jithma P Abeykoon, and Prashant Kapoor. 2018. Common adverse effects of novel therapies for multiple myeloma (mm) and their management strategies. *Current hematologic malignancy reports*, pages 1–11.

Health National Library of Medicine, National Institutes of Health and Human Services. 2005. Dailymed.com. https://dailymed.nlm.nih.gov/dailymed/index.cfm.

Prerna Mewawalla and Abhishek Chilkulwar. 2017. Maintenance therapy in multiple myeloma. *Therapeutic advances in hematology*, 8(2):71–79.

Rao Mohan, Michael Liguori, Srinivasa Mantena, Scott Mittelstadt, Eric Blomme, and Terry Van Vleet. 2008. Computational prediction of off-target pharmacology for discontinued drugs. *The FASEB Journal*.

Campillos Monica, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. 2008. Drug target identification using side-effect similarity. *Science*, 31(5886):263–266.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Köhler Sebastian, Nicole A. Vasilevsky, Mark Engelstad, Erin Foster, Julie McMurry, Sgolne Aym, Gareth Baynam, and et al. 2017. The human phenotype ontology in 2017. *Nucleic acids research*, 45(D1):D865–D876.

Neidle Stephen. 2011. *Cancer drug design and discovery*. Academic Press.

Evangelos Terpos. 2017. Multiple myeloma: Clinical updates from the american society of hematology annual meeting 2016. *Clinical Lymphoma, Myeloma and Leukemia*, 17(6):329–339.

Zi-Hang Zeng, Jia-Feng Chen, Yi-Xuan Li, Ran Zhang, Ling-Fei Xiao, and Xiang-Yu Meng. 2017. Induction regimens for transplant-eligible patients with newly diagnosed multiple myeloma: a network meta-analysis of randomized controlled trials. *Cancer management and research*, 9:287.

A Supplemental Material

Attached are the supplementary tables.

Table S1. Sixteen anti-MM drugs and their possible targets, (https://github.com/kyzhouhzau/crf-lstm-text/blob/master/Table%20S1.xlsx).

Table S2. Phenotype matching result for specific gene, (https://github.com/kyzhouhzau/crf-lstm-text/blob/master/Table%20S2.xlsx).

Table S3. Positive and negative samples and their distance, (https://github.com/kyzhouhzau/crf-lstm-text/blob/master/Table%20S3.xlsx)