DeepGeneMD: A Joint Deep Learning Model for Extracting Gene Mutation-Disease Knowledge from PubMed Literature

Feifan Liu*† , Xiaoyu Zheng†, Bo Wang, Catarina Kiefe
University of Massachusetts Medical School, Worcester, MA 01605

Abstract

Understanding the pathogenesis of genetic diseases through different gene activities and their relations to relevant diseases is important for new drug discovery and drug repositioning. In this paper, we present a joint deep learning model in a multi-task learning paradigm for gene mutation-disease knowledge extraction, DeepGeneMD, which adapts the state-of-the-art hierarchical multi-task learning framework for joint inference on named entity recognition (NER) and relation extraction (RE) in the context of the AGAC (Active Gene Annotation Corpus) track at 2019 BioNLP Open Shared Tasks (BioNLP-OST). It simultaneously extracts gene mutation related activities, diseases, and their relations from the published scientific literature. In DeepGeneMD, we explore the task decomposition to create auxiliary subtasks so that more interactions between different learning subtasks can be leveraged in model training. Our model achieves the average F1 score of 0.45 on recognizing gene activities and disease entities, ranking 2nd in the AGAC NER task; and the average F1 score of 0.35 on extracting relations, ranking 1st in the AGAC RE task.

1 Introduction

Drug repositioning has been regarded as a highly promising strategy for translational medicine (Wang and Zhang, 2013). One pharmacological hypothesis is that if a disease is caused by a mutated gene with gain of function (GOF) or loss of function (LOF), an antagonist/agonist chemical targeting the GOF/LOF mutated gene is a drug candidate for this disease (Wang and Zhang, 2013). Therefore, identifying and understanding the pathogenesis of genetic diseases as well as drug actions becomes an essential task. Among ways to test the above drug discovery hypothesis, computational methods through data mining (i.e. in silico) attract increasing attention over experimental methods (i.e. in vivo or in vitro) as the former ones are more cost-effective and time-efficient (Gachloo et al., 2019).

PubMed contains over 28 million biomedical article abstracts (Fiorini et al., 2018) and continues to grow rapidly, providing a valuable data resource to mine and extract this type of knowledge in a large scale. The 2019 AGAC shared tasks (Wang et al., 2018) are organized to facilitate efforts of extracting gene mutation-disease knowledge. In this study, we will focus on task 1 and task 2. Task 1 is a NER task where 12 concept entities representing different gene activities (e.g. variation, interaction, cell physiological activity, gene, protein, etc.), diseases, and regulatory actions (e.g. regulation, positive_regulation, negative_regulation, etc.) will be identified from free-text PubMed abstracts, while Task 2 is a RE task where "ThemeOf" and "CauseOf" relations will be extracted among entities recognized in Task 1. For instance, in the sentence "The [mutation]$_{\text{Variation}}$ resulted in a severe [loss]$_{\text{Negative\_Regulation}}$ of [DAX1]$_{\text{Gene}}$ [repressor activity]$_{\text{Molecular\_Physiological\_Activity}}$.", there are three relations among 4 entities: (1) CauseOf: "mutation" → "loss"; (2) ThemeOf: "repressor activity" → "loss"; (3): ThemeOf: "DAX1" → "repressor activity". Detailed definitions of each entity and relation may be found in (Wang et al., 2018).

* Correspondence: feifan.liu@umassmed.edu
† Two authors contribute equally.

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 77–83

Recently, text mining approaches have been developed to assist in the discovery of novel associations between existing drugs and new indications for hypothesis generation in connection with drug repurposing (Andronis et al., 2011). The emergence of deep learning approaches in natural language processing (NLP) propelled text-mining based drug knowledge discovery research, especially on the NER task (Gachloo et al., 2019). Effectively training deep neural networks, however, typically requires a large number of labeled samples, which are often prohibitively expensive to obtain in real-life applications (Zhang and Yang, 2018). As a popular solution to this data insufficient problem, Multi-Task Learning (MTL) (Caruana, 1997) has been widely applied and has led to successes across all applications of machine learning, including speech recognition (Deng et al., 2013), NLP (Collobert and Weston, 2008), computer vision (Ren et al., 2015) and drug discovery (Ramsundar et al., 2015).

In this paper, we proposed DeepGeneMD, a joint deep learning approach in a multi-task learning setting for mining gene mutation-disease knowledge from the biomedical literature. Inspired by the state-of-the-art hierarchical multi-task learning (HMTL) approaches (Sanh et al., 2018), we further explore how to create additional subtasks interacting with each other in a hierarchical manner. To this end, we take into account the task's inherent compositionality and decompose the NER task into three subtasks. Compared with HMTL, this creates additional levels of learning hierarchy between NER decomposed subtasks and original NER. The hypothesis is that through task decomposition, we can enrich the interactions among the semantic representations learned at each level of the hierarchy, which enables DeepGeneMD to incorporate diverse signals from related tasks to learn more effective representations for each task with optimal generalizability. The contributions of this study are:

(1) Propose DeepGeneMD to extend hierarchical multi-task learning through task decomposition and enriched inter-task inter-actions.

(2) Apply advanced word representations to initialize semantic representations of input sentences.

(3) Demonstrate the effectiveness of the proposed approach given limited annotated data.

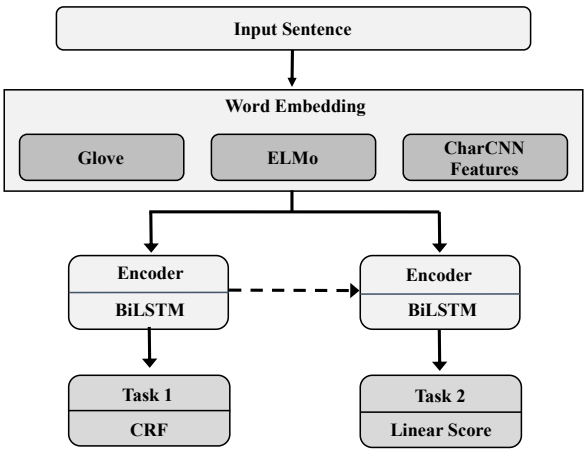

Figure 1: The HMTL (Sanh et al.) architecture for AGAC tasks

2 Hierarchical Multi-Task Learning

The hierarchical model trained in the multi-task setup (Hierarchical Multi-Task Learning, HMTL) introduces a hierarchical inductive bias between different tasks by supervising low-level tasks at the bottom layers of the model architecture and supervising higher-level tasks at higher layers (Hashimoto et al., 2017; Sanh et al., 2018). The assumption is that lower-level tasks require less linguistic understanding than higher-level complex tasks while learning different levels of linguistic properties in the hierarchical end-to-end fashion enables the higher-level tasks to leverage the shared representation of the low-level tasks.

We formulated the 2019 AGAC task 1 and 2 into a hierarchical multi-task learning problem, which can be addressed using the HMTL architecture similar to (Sanh et al., 2018). As shown in Figure 1, the task 1 (NER, recognize gene activity concepts and disease entities) is considered as a lower-level task while task 2 (RE, extract relationship among concept/entity pairs) as a higher-level task, and the dashed lines indicate interactions among tasks. For a given input sentence, the embedding layer concatenates the Glove word-level embedding (Pennington et al., 2014), contextual ELMo (Peters et al., 2018) word embeddings and convolutional neural network (CNN) based Character-level word embeddings (Chiu and Nichols, 2016) as each word's expanded embeddings (e_W). The encoder of Task 1 takes the word embedding through multilayer BiLSTM (Lample et al., 2016) and outputs an encoded sequence (e_{NER}) into the final Conditional Random Field (CRF) layer for inferring the NER output. The encoder of Task 2 takes as the input the

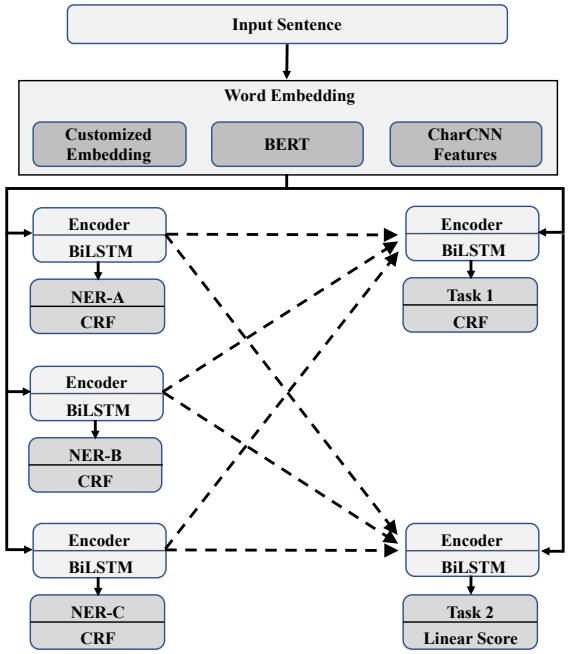

Figure 2: The architecture of the proposed DeepGeneMD model

concatenated word embedding, i.e. e_W, with the learned vector representation, i.e. e_{NER}, from the encoder of Task 1 into a linear scorer (Sanh et al., 2018) for RE inferences. Note that the two tasks don't depend on each other's output explicitly, but RE does use the intermediate encoder representation from NER to make better decisions.

3 The DeepGeneMD System

Most existing efforts in HMTL approaches are limited to existing tasks of interest, however, auxiliary tasks have been shown helpful in multi-task learning (Liebel and Körner, 2018; Niu et al., 2019). Motivated by this idea, we introduced the DeepGeneMD model to create auxiliary subtasks into the HMTL structure to further explore the potential of HMTL approaches. Compared with previous work, the following summarizes the differences in our model:

- Upgrade the word representations using state-of-the-art counterparts as well as customized ones trained on domain data.
- Integrate task decomposition to enable more interactions in the HMTL learning structure.
- Design the hierarchical linking structure to accommodate decomposed subtasks, as shown in Figure 2.

3.1 Word Embeddings

Although Glove is trained on a very large corpus, it may still lack domain coverage when processing medical texts. To overcome this challenge, we utilized in our model a customized word embedding (Jagannatha and Yu, 2016) trained through skip-gram setting using all PubMed open access articles, 99,700 EHR notes, and English Wikipedia articles in 2015. This embedding contains 3 billion tokens and the embedding dimension is 200.

BERT (Bidirectional Encoder Representations from Transformers) builds upon recent work in pre-training contextual representations, and have demonstrated new state-of-the-art performance when applied on various NLP tasks (Devlin et al., 2018), compared with previous models, e.g. ELMo (Peters et al., 2018). Therefore, we exploited the BERT representations in the DeepGeneMD model to provide contextual representations of each word in the input sentence. Following (Sanh et al., 2018), we also used character CNN word embeddings to accommodate the out of vocabulary (OOV) problems. As shown in Figure 2, the input of our model will be mapped to a concatenated vector of customized embedding, BERT, and character CNN embeddings.

3.2 Task Decomposition

The rationale of task decomposition is two-folds. First, it could create auxiliary subtasks to be engaged in the HMTL structure, and the supervision on those auxiliary tasks is expected to provide additional information through sharing their learned language representations. Second, decomposed subtasks reduce the complexity compared with the original task, holding the potential of learning from a unique perspective. In this study, we applied the task decomposition on the AGAC NER task in which there are 12 types of entities to be identified, such that each subtask recognizes a subset of entity types. We empirically set the number of subtasks as 3 based on the hypothesis that too many subtasks may introduce noise during model training.

To determine which entity goes to which subtask, we calculated a statistical measure, roleRatio, for each entity as in equation (1) which is expected to capture statistical characteristics regarding the role each entity plays when relating to other entities.

$$roleRatio = Freq_{rel\_head}/Freq_{rel\_tail} \quad (1)$$

Here "$Freq_{rel\_head}$" and "$Freq_{rel\_tail}$" indicate respectively how many times the entity serves as the head and tail of a participating relationship in the training data. Each relation starts from the head entity and points to the tail entity. Based on the value of roleRatio, we split all the entities into 3 subgroups, each containing 4 entity types:

Subgroup	Entities
A	PosReg (positive regulation), NegReg (negative regulation), Reg (regulation), Interaction
B	Gene, Pathway, Protein, Disease
C	Enzyme, Var (variation), CPA (cell physiological activity), MPA (Molecular physiological activity)

Table 1: Subgroups of 12 Entities for Task Decomposition

In subgroup A, the roleRatio values of all the entities are all less than 1 indicating they are more likely to be the tail entity of a relation. For entities in subgroup B and C, we split them in a stratified way, each of them containing both high and low roleRatio entities, e.g. Gene from subgroup B and Enzyme from subgroup C have the largest roleRatio of 27 and 14.5 respectively.

The corresponding subtasks to identify those subgroups are denoted as NER-A, NER-B, and NER-C respectively, and the original NER for 12 entities as NER.

3.3 Interaction Linking Structure

There are different ways to link different subtasks in the HMTL structure. In our model, we designed the structure as shown in Figure 2. The dashed lines indicate interaction connections between tasks. The task pointed by the arrow is on the higher-level of HMTL layer, which has access to the learned language representations from all the other tasks pointing to it. For instance, the outputs of BiLSTM encoders for NER-A, NER-B, and NER-C are concatenated as the part of the input of another two higher-level tasks: (1) NER for 12 entities (Task 1) (2) RE for two relations (Task 2). In addition, as NER-A, NER-B and NER-C can also produce outputs for Task 1, we can combine their prediction result in a simple ensemble manner, which may lead to better performance.

4 Experiments

4.1 Preprocessing

We randomly selected 25 (10%) documents from the training data as the validation set. The model is trained on the remaining 225 documents and the performance evaluated on the validation set is used for model tuning. All the entities are labeled through BIOUL (Begin, Inside, Outside, Unit, Last) labeling schema.

4.2 Hyperparameters and Implementation Details

We applied the same hyperparameter setting used in (Sanh et al., 2018) except the following adjustment based on validation performance: (1) we increased the dropout rate from 0.2 to 0.25 for NER related tasks; (2) We increased the dropout rate from 0.2 to 0.3 for the RE task.

We used various batch sizes (4, 8, 16, 32 and 64) for the RE task when training the DeepGeneMD system. The resulting five settings are denoted as DeepGeneMD-4, DeepGeneMD-8, DeepGeneMD-16, DeepGeneMD-32, and DeepGeneMD-64. We also trained an HMTL Model using the structure in Figure 1 but with our new word representations, denoted as HMTL-New.

We adopted the same training method called proportional-sampling as in (Sanh et al., 2018): after each parameter update, a task is randomly selected and a batch of the dataset attached to this task is also randomly sampled. The probability of sampling a task is proportional to the relative size of each dataset compared to the size of all the datasets.

4.3 Results

As mentioned earlier, NER results can be taken from different subtask module, and RE results can be taken from different training settings with different batch size. We tried different merging strategies when submitting results to the organization committee. In total, we submitted three runs:

- **Run1**: DeepGeneMD-4 for task 1; HMTL-New for task 2.
- **Run2**: Merged results from original NER task in DeepGeneMD-4 and three subtasks (NER-A, NER-B, NER-C) in DeepGeneMD-16 for task 1; DeepGeneMD-8 for task 2.
- **Run3**: Merged results from original NER task in DeepGeneMD-4, NER-A subtask in DeepGeneMD-16, NER-B subtask in DeepGeneMD-32 and NER-C subtask in DeepGeneMD-64 for task 1; DeepGeneMD-8 for task 2.

When merging results from different task outputs, conflicts are empirically handled by

prioritizing outputs from three subtasks (NER-A, NER-B, NER-C) based on the assumption that they are tailored specifically to a subset of entities.

The overall performance of our three submitted runs is shown in Table 2. It is observed that Run 2 achieved the best F1 score of 0.35 for RE and Run 1 yielded the best F1 score of 0.45 for NER. It suggests that DeepGeneMD-8 benefits from task decomposition and more inter-task interactions for RE tasks. More experiments are needed to analyze each component's contribution to the whole learning structure.

Submission		Precision	Recall	F1
NER	Run1	0.36	0.59	**0.45**
	Run2	0.33	0.64	0.44
	Run3	0.34	0.62	0.44
RE	Run1	0.47	0.25	0.33
	Run2	0.4	0.31	**0.35**
	Run3	0.4	0.3	0.34

Table 2: Official Submission Results in AGAC

Entity Name	Precision	Recall	F1
Var	0.38	0.77	0.5
Pathway	-	0	0
MPA	0.19	0.48	0.27
CPA	0.12	0.14	0.13
Reg	0.63	0.46	0.53
PosReg	0.35	0.65	0.46
NegReg	0.41	0.66	0.5
Disease	0.45	0.57	0.5
Gene	0.33	0.7	0.45
Protein	0.42	0.08	0.14
Enzyme	-	0	0
Interaction	-	0	0
Overall	0.36	0.59	0.45

Table 3: Entity-level NER Performance of Run1

The entity-level performance for our best-performing NER run (Run 1) is presented in Table 3. The performance on each entity type varies, and most of them achieve higher recall (e.g. 0.77 for Var and 0.7 for Gene) except for Protein (recall of 0.08). There are three types of entities which the system fails to recognize: Pathway, Enzyme, Interaction. It may be due to the lack of training instances for those entities, which is demonstrated in Table 4. Those three entities have less than 30 examples (less than 1%) in training, compared with more than 200 examples in most entity types. It also explains the low recall for protein as it has less than 100 (2.77%) training instances.

Entity Name	Count	Percentage
Var	733	22.07%
Gene	526	15.84%
MPA	417	12.56%
NegReg	370	11.14%
Disease	334	10.06%
PosReg	327	9.85%
CPA	227	6.84%
Reg	215	6.47%
Protein	92	2.77%
Enzyme	29	0.87%
Interaction	27	0.81%
Pathway	24	0.72%
Overall	3321	100%

Table 4: Entity Statistics of Training Data

Relation	Precision	Recall	F1
CauseOf	0.54	0.32	0.4
ThemeOf	0.35	0.31	0.33
Overall	0.4	0.31	0.35

Table 5: Relation-level RE Performance of Run2

Table 5 shows the detailed performance of the best-performing run of our system on the relation extraction task. The system achieved similar recall value (~0.31-0.32) on both relations, but the much higher precision score for the "CauseOf" relation (0.54) than "ThemeOf" (0.35).

5 Error Analysis

We conducted some error analysis on the validation dataset and some examples are shown below.

- False Negatives

  [Loss of function]$_{Var}$ in [ROBO1]$_{Gene}$ is [associated]$_{Reg}$ with [tetralogy of Fallot]$_{Disease}$ and septal defects.

  In this sentence, our system only recognized "ROBO1" as Gene but failed on other entities. It could be due to the limited training data restricting the learning capacity of the model.

- False Positives

  In 2006, mutations in progranulin gene (GRN) that cause haploinsufficiency were found in familial cases of frontotemporal dementia (FTD).

In this case, our model incorrectly recognized "haploinsufficiency" as Var which is not annotated in the ground-truth. Here the contextual language (e.g. GRN, cause) confuses the system.

- Potential Annotation Error

  Gain-of-function mutations in PDR1, …

  For this example, the system identified "mutations" as Var, and "PDR1" as Gene which seems reasonable, but those are not annotated in the ground-truth.

6 Conclusion and Discussion

We developed the DeepGeneMD system in the hierarchical multi-task learning setup and applied it to extract gene mutation-disease knowledge from PubMed biomedical literature. By exploring task decomposition and new word embeddings, the resulting model demonstrated promising results, ranking 2nd in the NER Task and 1st in the RE Task among all participant teams. The idea of task decomposition and creating additional interactions among different subtasks can also apply to other applications in the hierarchical multi-task learning setting.

There are several limitations to this study. First, we applied a heuristic approach based on roleRatio value for the task decomposition, which is relatively ad-hoc and may not be optimal. Second, there are different structure candidates to engage different subtasks in an HMTL setting, and we simply made an empirical design for the current DeepGeneMD system, which may have limited the potential of mutual benefits of multiple learning tasks. Third, when merging results from different components, we assume that decomposed subtasks may have learned better knowledge regarding the corresponding subset of entities, but that assumption may not hold.

For future work, we plan to tune the hyper-parameters extensively and investigate whether applying different interaction linking structures among subtasks and leveraging various ways of task decomposition can further improve the system's performance. In addition, we will apply our framework on various datasets from different domains to evaluate its generalizability and robustness.

Acknowledgments

References

Christos Andronis, Anuj Sharma, Vassilis Virvilis, Spyros Deftereos, and Aris Persidis. 2011. Literature mining, ontologies and information visualization for drug repurposing. *Briefings in Bioinformatics*, 12(4):357–368, July.

Rich Caruana. 1997. Multitask Learning. *Machine Learning*, 28(1):41–75, July.

Jason Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

L. Deng, G. Hinton, and B. Kingsbury. 2013. New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8599–8603. May.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, October. arXiv: 1810.04805.

Nicolas Fiorini, Robert Leaman, David J. Lipman, and Zhiyong Lu. 2018. How user intelligence is improving PubMed. *Nature Biotechnology*, 36(10):937–945, October.

Mina Gachloo, Yuxing Wang, and Jingbo Xia. 2019. A review of drug knowledge discovery using BioNLP and tensor or matrix decomposition. *Genomics & Informatics*, 17(2), June.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933, Copenhagen, Denmark. Association for Computational Linguistics.

Abhyuday N Jagannatha and Hong Yu. 2016. Bidirectional RNN for Medical Event Detection in Electronic Health Records. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482, San Diego, California, June. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. , March.

Lukas Liebel and Marco Körner. 2018. Auxiliary Tasks in Multi-task Learning. *arXiv:1805.06334 [cs]*, May. arXiv: 1805.06334.

Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. 2019. Multi-task Character-Level Attentional Networks for Medical Concept Normalization. *Neural Processing Letters*, 49(3):1239–1256, June.

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, Doha, Qatar, October. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. , February.

Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. 2015. Massively Multitask Networks for Drug Discovery. *arXiv:1502.02072 [cs, stat]*, February. arXiv: 1502.02072.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv:1506.01497 [cs]*, June. arXiv: 1506.01497.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2018. A Hierarchical Multi-task Approach for Learning Embeddings from Semantic Tasks. *arXiv:1811.06031 [cs]*, November. arXiv: 1811.06031.

Yuxing Wang, Xinzhi Yao, Kaiyin Zhou, Xuan Qin, Jin-Dong Kim, Kevin Bretonnel Cohen, and Jingbo Xia. 2018. Guideline design of an active gene annotation corpus for the purpose of drug repurposing. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5. IEEE.

Zhong-Yi Wang and Hong-Yu Zhang. 2013. Rational drug repositioning by medical genetics. *Nature Biotechnology*, 31:1080–1082, December.

Yu Zhang and Qiang Yang. 2018. An overview of multi-task learning. *National Science Review*, 5(1):30–43, January.