Improving Molecular Property Prediction on Limited Data with Deep Multi-Label Learning

1st Hehuan Ma *
hehuan.ma@mavs.uta.edu

2nd Chaochao Yan *
chaochao.yan@mavs.uta.edu

3rd Yuzhi Guo *
yuzhi.guo@mavs.uta.edu

4th Sheng Wang *
sheng.wang@mavs.uta.edu

5th Yuhong Wang †
yuhong.wang@nih.gov

6th Hongmao Sun †
sunh7@mail.nih.gov

Junzhou Huang *‡
jzhuang@uta.edu

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313553

***Abstract*—Acquiring labeled data has been widely recognized as a major challenge in molecular property prediction. Since it generally requires a series of specialized biochemical experiments which are time-consuming, costly, as well as labor-intensive. The deficiency of labeled property data makes it difficult to learn a good prediction model. Here, we propose an RNN-based multi-label molecular property prediction method to alleviate the data scarcity issue in two stages: 1) utilize the abundant unlabeled SMILES data to pre-train a seq2seq model whose encoder learns to generate molecular fingerprint based on the given SMILES; and 2) finetune the pre-trained model on the labeled molecular property data. Since labeled data is limited, we train those properties with limited sample size jointly with other properties which contain relatively sufficient samples. This approach brings in the idea of multi-label training, which is able to pre-train and fine-tune the encoder network, as well as train the prediction network with a data augmentation strategy. Extensive experiments on molecular property prediction demonstrate that our proposed method has achieved superior performance compared with the state-of-the-art approaches on properties with limited sample size.**

***Index Terms*—Molecule Property Prediction, SMILES sequence learning, Recurrent Neural Network, Semi-supervised Learning, Unlabeled Data, Bioinformatics**

I. Introduction

Molecular property prediction has been a significant task in drug discovery area. Recently, the amount of available compounds and biological activity data increased exponentially due to the experimental techniques such as High-throughput screening (HTS) and parallel synthesis [1], [2]. Effectively utilizing these large-scale chemical data would be a fruitful strategy to tackle property prediction problem. Deep learning has been widely known for its capability of taking advantage of massive amount of data [3], which leads to another surge in drug discovery domain. It would significantly save the R&D costs for drug research procedure while decreasing the failure rate in potential drug screening trials, as well as speed up the overall drug discovery process by exploiting the large-scale chemical data [4]. Lots of researches have been working on property prediction from various aspects, such as analyzing the structure of the molecule [5]–[7], or extracting the local features by looking into the chemical features like bounds relatives [8]. These methods generally perform prediction task by running the target property individually, and achieve good performance with sufficient labeled data. However, the prediction performance is usually unsatisfied when the labeled data is limited.

To address such problem, some previous work turns the attention to explore how to take advantage of the enormous unlabeled data. Certain attempts have successfully improved the prediction performance by adding unlabeled data as part of the training process, e.g., seq2seq and seq3seq [9]–[12]. Nevertheless, there is another way to address this issue from a different perspective, which is to explore the potential information of limited labeled data effectively. Specifically, we propose to jointly co-train multiple properties of data, by taking those properties into the pool filled with several other properties and train them together. This multi-label idea would perform as a data augmentation for the limit-sample properties.

In this paper, we propose a data-driven Multi-label Recurrent Neural Networks based molecular property prediction technique to maximize the utilization of available data, not only unlabeled data but also the labeled data. Specifically, it can be divided into two fundamental parts, unsupervised task and supervised task. First, we applied sequence-to-sequence (seq2seq) learning on the massive unlabeled data, which is an enormous collection of molecule SMILES. It is inspired by semi-supervised learning in Nature Language Processing (NLP) [13]–[15]. The SMILES sequence is input into the recurrent neural network, and converted to a vector representation called molecular fingerprint. The fingerprint then is reconstructed back to SMILES to update the network. By fully training the unlabeled data, we are able to get a pre-trained network for translating SMILES to vector with high efficiency. The intermediate fingerprint can be used for further investigation, in our case, which would be the property prediction task. The second part is to overcome the difficulty of acquiring sufficient labeled data by establishing a multi-label supervised model on a combined dataset with missing labels. As we mentioned before, training each property prediction task is ineffective and costly. In our proposed method, the input

* Hehuan Ma, Chaochao Yan, Yuzhi Guo, Sheng Wang, and Junzhou Huang are with the Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, Texas, United States.
† Yuhong Wang and Hongmao Sun are with National Center for Advancing Translating Sciences, NIH, Rockville, Maryland, United States.
‡ Corresponding author.

to prediction network is a data matrix with multiple property label information, which can be an original dataset collected from specialized experiments [1], [2], or formed manually by various single property. This data matrix is then fed into a prediction network to perform either classification tasks, regression task or both.

The innovation of our proposed method mainly contribute to four points: 1) It effectively utilizes the enormous unlabeled molecule data information, as well as the limit labeled molecular property data. 2) By feeding and training multiple properties data jointly into a neural network, the prediction results on the properties with limited samples are significantly improved by learning from those properties with relatively more data points. 3) The overall results on all properties perform better than training each property individually. 4) Regression and classification can be employed at the same time during our supervised training process, which assures the variety of different input labels.

II. Related Work

1) Hand-crafted Biologist-guided Hash-based Fingerprints: One conventional used traditional feature extraction method is to design the molecule fingerprint manually by experts based on biological experiments and chemical knowledge, e.g., [8], [16], [17]. This type of fingerprint methods generally work well for particular tasks but lacks universality. Hash-based methods have then been developed to address the issue of biologist-guided local-feature fingerprints. It aims to generate unique fingerprint based on different molecular features [8], [17], [18]. One critical approach is called circular fingerprint [19]. Nonetheless, it has a very notable problem, since the characteristic of the hash function is non-invertible, it might not be able to catch enough information when converting.

2) Sequence-based Models: SMILES sequence is a breakthrough for studying molecular property prediction by deep learning methodologies, e.g., seq2seq fingerprint [9], and seq3seq fingerprint [10]. These models spot at the potentially useful information of almost infinite molecule SMILES sequence data by adequately training them to obtain strong vector representation of the molecule. These vectors then go through other supervised models to perform property prediction, e.g., GradientBoost [20], RandomForest [21], SVM [22]. The Seq3seq fingerprint is an end-to-end semi-supervised learning method, which combines the training unlabeled data part and further prediction part together in one framework. It directly takes the generated fingerprint from the unsupervised learning network to predict the property labels.

III. Methodologies

Our proposed deep multi-label learning prediction contains two principal parts, unsupervised task and supervised task. The unsupervised task exploits the enormous unlabeled data, and the supervised task overcomes the dilemma of limited labeled data.

A. Network structure

The framework of our proposed method is shown in Fig. 1. The upper part is the unsupervised task, which is responsible for training a proper pre-trained model. SMILES data goes through the encoder network to generate a fingerprint, then enters the decoder network to recover back to SMILES sequence. The unsupervised loss is calculated and used to update the network. The encoder and decoder network share similar fundamental parts, which can be various recurrent neural network models, here in our experiment, LSTM network is implemented [23]. The pre-trained model then is employed in the following supervised prediction task, along with both SMILES and the labels as the input. A mask function is performed in the prediction network to eliminate the influence of missing labels.

B. LSTM Unit

The Long Short-Term Memory (LSTM) [23] is the most widely used recurrent neural network. LSTM has three gates: input gate, forgot gate, and output gate. A LSTM network computes a sequence of network outputs $(y'_1, \ldots, y'_T)$ from the input sequence $(x_1, \ldots, x_T)$ by iterating

$$f_t = \sigma_g(W_f X_t + U_f h_{t-1} + b_f) \tag{1}$$

$$i_t = \sigma_g(W_i x_t + U_i y'_{t-1} + b_i) \tag{2}$$

$$o_t = \sigma_g(W_o x_t + U_o y'_{t-1} + b_o) \tag{3}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c + U_c h_{t-1} + b_c) \tag{4}$$

$$h_t = o_t \circ \sigma_h(c_t). \tag{5}$$

The LSTM cell has a "forgot" gate f_t which is to block some of the previous states to pass through the entire sequence. i_t and o_t are the input and output gates for the LSTM cell at time step t. c_t and h_t are the LSTM cell state and hidden state. σ represents the activation function. In our experiments, σ_g is the sigmoid function, and σ_c and σ_h are the hyperbolic tangent functions.

C. Unsupervised Pre-train Model Training

The unsupervised encoder network takes the SMILES sequence representations as the input, then outputs the corresponding vector fingerprints. The generated fingerprints then go through a decoder network to transfer the vector back to a sequence. The output sequence is compared with the input SMILES to calculate the loss, which is back-propagated to update the encoder network weights.

D. Supervised Multi-label Prediction

The pre-trained model is applied in the following supervised prediction task. Our proposed method merges multiple property datasets into a matrix with the property labels as the columns, and all observed SMILES as row index. Since different property datasets contain different number of samples, each SMILES in the matrix might only have one or some of the properties information. During the training process, when the label of the SMILES is missing, we set the associated loss as 0. By training multiple properties together,

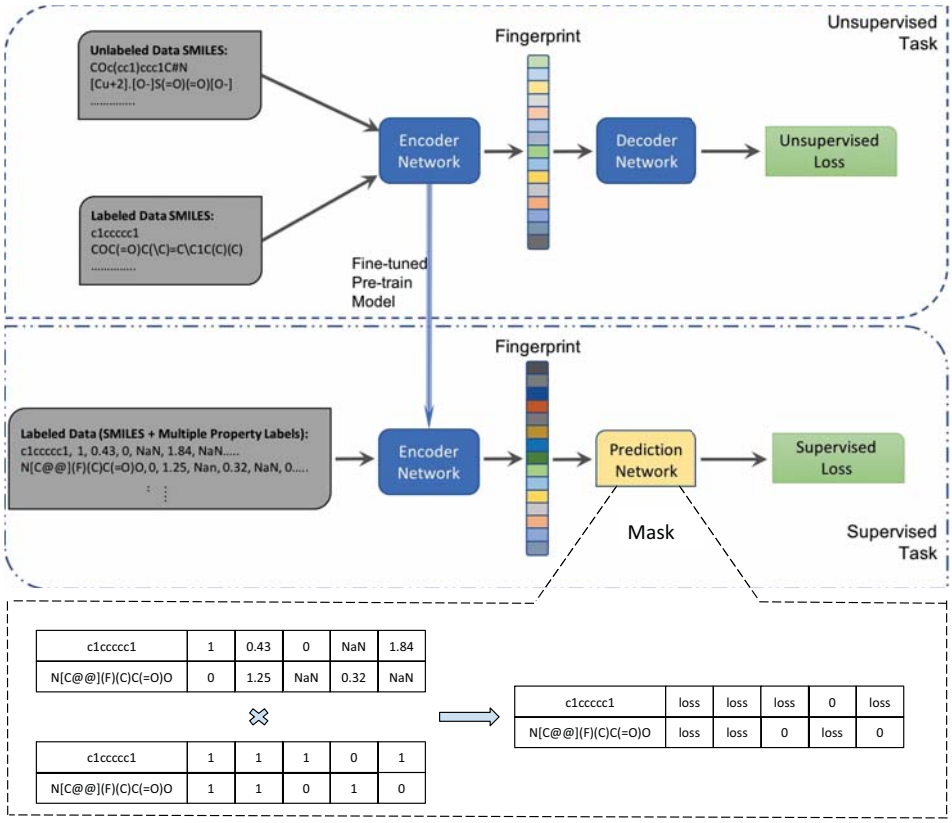

Fig. 1: Network structure. The upper part is the unsupervised network part, and a fine-tuned encoder network from this stage is used in the lower part (the prediction network) to perform the supervised tasks.

the performance is improved significantly on the properties with limited data. In addition, the prediction network is able to perform regression and classification at the same time by specifying the classification property index. The detail results are presented and discussed in the next section.

E. Loss Function

1) Unsupervised task: We apply the cross-entropy loss for the unsupervised task. The token vocabulary $\{v_1, v_2, \ldots, v_N\}$ of SMILES sequence is unique and limited. Set $z_t \in \mathbb{R}^N$ as the output token distribution from the LSTM cell outputs, and $l_t \in \mathbb{R}^N$ as the one-hot vector of the given original SMILES sequence token at time step t. Thus the unsupervised loss $\mathcal{L}_{unsup}$ is given by:

$$\mathcal{L}_{unsup} = \sum_{t=1}^{T} l_t^T \log(z_t). \tag{6}$$

2) Supervised task: For the supervised task, we use the **softmax** loss. It calculates the probabilities of each target class over all possible target classes. The calculated probabilities are then used to determine the target class for the given inputs.

$$\mathcal{L}_{sup} = \frac{e^{z_j}}{\sum_i e^{z_i}}. \tag{7}$$

The total supervised loss is the sum of each property. Property weight λ can be assigned accordingly to each property before training.

$$\mathcal{L}_{sup\_all} = \sum_{n} \mathcal{L}_{sup}. \tag{8}$$

F. Missing Labels Handling

In our experiment, a mask function is applied to eliminate the effects of the missing labels. After calculating the loss for each data point, the loss values are then multiplied by a matrix with the same size of the input data matrix. The mask matrix is formed by 0s and 1s, as the corresponding positions with missing labels are recorded as 0 since $w_{ij} * 0 = 0$, others are marked as 1. Thus, any weights connecting with the missing labels would have no influence on the further computation.

IV. Experiments

A. Experiment Setup

1) Dataset Description: As we mentioned before, two types of datasets are used in our experiment, the unlabeled and labeled dataset are explored in the pre-training, while the labeled dataset is used in the supervised property prediction tasks. One large unlabeled dataset is used to train the encoder network, which is the ZINC Dataset [24]. It is an open-source chemical dataset which is released to the public in late 2015. ZINC

contains over 35 million commercially available molecular compounds with multiple biologically relevant information, such as structure and properties. Here we pick the drug-like dataset, which contains 18,691,354 molecular in SMILES representation.

Two labeled datasets are used to perform property prediction tasks, NIH-17p and NCI. Within them, NIH-17p is used for both classification tasks and regression tasks, while NCI is only used in the regression tasks based on the label characteristic.

- NIH-17p is provided by the National Center for Advancing Translational Sciences (NCATS) at National Institutes of Health (NIH). It consists of 17 individual properties associated with molecular SMILES representations. The sample size of each property is different (shown under the "Sample-Numbers" in Table I and Table II); some of them have very limited data. For instance, property **pgb** has only **186** samples, **heptox** has **440** objects, and **vd** has **668** data points. On the other hand, some properties have up to almost 100 times data compared with them, e.g., the sample size of 2d6 is 15428, and 1a2 is 14226. NIH-17p contains both continuous labels (3/17) and binary labels (14/17). The classification tasks are established on the 14 properties, and the regression tasks are conducted on the remaining three properties. We randomly split the dataset into training, validation, and testing by the ratio of 80%, 10%, and 10%.
- NCI dataset is download from Deepchem [25]. It contains 11,738 unique molecule SMILES with eight cancer-related property labels, which are CCRE, HL-60, K-526, RPMI, A549, COLO, HCC, MALME. Since no missing labels are observed in NCI dataset, we manually generate a new dataset from the original NCI dataset with missing labels. The detailed approach is selecting two properties, says CCRE and HL-60, and randomly set 95% and 90% of the labels as missing, other properties remain unchanged. Next, the multi-label prediction is applied on this dataset to evaluate our proposed assumption, which is properties with more data would help improve the performance of the properties with limited data. The dataset is split by 80% training, 10% validating, and 10% testing as well. We repeat this approach 4 times (each time sets two different properties as the properties with "limited data") to eliminate the contingency.

B. Hyper-parameters Settings

For the unsupervised task, the encoder network is a 3 layers LSTM network with the input embedding dimension as 128, and the hidden size as 256. The decoder network is assigned with the same hyper-parameters with the encoder network, with the output dimension as 128. The optimizer used is Adam, and the dropout rate is 0.5. The supervised model is one layer fully connected neural network.

C. Evaluation Metric

Exact match accuracy and the cross-entropy loss are used to back-propagate and update the unsupervised neural network. The exact match accuracy is a measure of the portion of accurately recovered sequence within the entire samples. For the followed supervised tasks, the root mean squared error (RMSE) is used to measure the performance of the regression task, and the accuracy is used for validating the classification results.

TABLE I: The regression results of NIH-17p (RMSE, lower is better).

Property	Sample-Numbers	Circular-FP	Neural-FP	Seq2seq-FP	Seq3seq-FP	**Ours**
Logp	10851	1.3437	0.7085	1.4065	0.5864	**0.4534**
pampar	4071	0.7609	0.6517	0.7343	0.6959	**0.5578**
vd	668	0.5870	0.5789	0.6321	0.6183	**0.5015**

TABLE II: The classification results on NIH-17p data (accuracy, higher is better).

Property	Sample-Numbers	Circular-FP	Neural-FP	Graph-Representation	Seq2seq-FP	Seq3seq-FP	**Ours**
pgb	186	**72.22%**	66.67%	61.11%	**72.22%**	66.67%	**72.22%**
solub	6694	82.33%	80.17%	81.38%	62.64%	77.87%	**83.05%**
2c9	13064	78.05%	80.70%	**81.65%**	67.27%	79.38%	80.55%
2d6	15428	87.88%	88.08%	77.58%	86.38%	88.79%	**89.25%**
2c19	11833	76.15%	80.43%	**83.02%**	66.24%	77.44%	81.45%
rlm	10626	78.82%	73.03%	77.91%	62.58%	75.14%	**80.17%**
mmp	5970	88.40%	89.22%	86.79%	86.93%	87.91%	**90.36%**
ames	8224	**79.95%**	78.91%	78.59%	63.39%	74.97%	79.18%
pldc	4161	90.50%	**93.75%**	81.21%	91.75%	93.00%	92.75%
heptox	440	75.93%	62.96%	79.63%	62.96%	62.96%	**81.48%**
3a4	13433	78.57%	78.50%	78.29%	69.60%	75.30%	**79.64%**
pampac	4698	77.28%	76.88%	77.87%	68.56%	73.23%	**83.98%**
herg	3024	90.81%	91.17%	85.96%	85.16%	90.11%	**91.52%**
1a2	14226	83.38%	82.61%	83.88%	70.97%	84.78%	**85.06%**

D. Comparison Experiments

1) Regression task: NIH-17p.: The comparison experiments on regression task is conducted on both NIH-17p dataset and NCI dataset with four baseline methods, circular fingerprint (circular-FP) [19], neural fingerprint (neural-FP) [5], seq2seq fingerprint (seq2seq) [9], and seq3seq fingerprint (seq3seq) [10] . The circular-FP is generated by a hand-crafted hash-based algorithm to define the local features. The neural-FP is constructed by a supervised deep graph convolutional neural network. Seq2seq fingerprint and seq3seq fingerprint are RNN based models.

2) Regression task: NCI with missings.: Regression tasks are running on NCI dataset with missings too. We picked two properties in the proper order to conduct four experiments, each contains one property with 95% manually assigned missings, and one property with 90% missings, others remain still. The details about the data sample numbers can be found in Table III. Three comparison methods are applied on these datasets: circular-FP, neural-FP, and seq3seq fingerprint. Due to the limited position in the paper, we take out seq2seq since seq3seq fingerprint generally performs better.

3) Classification task.: One recently published method, molecular properties prediction utilizing graph-level representation (Graph-Representation) [26], has been added to perform classification task. This method proposed an idea of presenting molecular properties by learning graph-level features instead of node-level. Since it can only predict positive and negative labels, it is not included in the regression tasks. The dataset used for the classification experiments is formed by the 14 properties from NIH-17p dataset, which carry binary labels. Logp, pampar, and vd are excluded in the classification tasks, and used to conduct the regression tasks.

E. Experiment Results

The results of all the comparison methods and our proposed model are shown in the following tables. Table I is the RMSE results of running different models on NIH-17p dataset, Table II is the accuracy results of the 14 properties from NIH-17p dataset. Table III shows the results of the extensive experiments on NCI dataset with randomly assigning missings, the evaluation criterion is RMSE since all labels in the NCI dataset are continuous.

As observed, Table I and table II clearly show that properties with limited data have achieved up to 13.4% improvement, which are **vd** (668 samples), **heptox** (440 samples), and **pgb** (186 samples). For other properties, the overall results are also better than the baseline methods. Those properties with more data samples rarely improve since overall deep learning methods are able to obtain good results by training sufficient data adequately. We argue that the performance of pgb is same as the other two baselines probably due to the extremely small sample size (only 186). The sample size of the other two properties appear to be appropriate for multi-label training (approximately hundreds of samples).

The comparison experiments on the revised NCI dataset fully prove the effectiveness of our proposed method. By randomly assigning large proportion of missing labels to certain properties, which manually forms a dataset contains properties with limited data and properties with sufficient data, has confirmed that proposed multi-label learning is able to extract more information than training each property individually, especially boost the performance of the properties with limited data. The four runs with different "limited data" properties demonstrate the robustness of our method. Overall, training multiple properties together indeed improve the overall per-

TABLE III: The RMSE results of the regression tasks on the NCI dataset. Properties with limited data are marked as grey (lower is better).

(a) Missing labels properties: CCRE & HL-60.

Prop (sample#)	Circular	Neural	Seq2seq	Seq3seq	Ours
CCRE (576)	0.8104	0.8591	0.8108	0.8218	**0.7914**
HL-60 (1134)	0.7464	0.6538	0.6704	0.6569	**0.6335**
K-526 (11738)	**1.0530**	1.0778	1.0694	1.0685	1.0587
RPMI (11738)	1.1251	1.0231	1.0371	1.0353	**1.0203**
A549 (11738)	**0.7543**	0.7571	0.7673	0.7563	0.7549
COLO (11738)	0.7554	0.7787	**0.7546**	0.7634	0.7605
HCC (11738)	0.7074	0.7115	0.7125	0.7048	**0.7035**
MALME (11738)	0.7673	0.7653	0.7683	0.7710	**0.7582**

(b) Missing labels properties: K-256 & RPMI.

Prop (sample#)	Circular	Neural	Seq2seq	Seq3seq	Ours
CCRE (11738)	0.9838	0.9896	0.9975	0.9921	**0.9828**
HL-60 (11738)	0.8068	0.8085	0.8077	0.8099	**0.8051**
K-526 (579)	0.9059	1.2606	0.9160	0.8759	**0.8548**
RPMI (1180)	1.1251	1.1329	1.0814	1.0778	**1.0654**
A549 (11738)	**0.7543**	0.7571	0.7673	0.7563	0.7549
COLO (11738)	0.7554	0.7787	**0.7546**	0.7634	0.7605
HCC (11738)	0.7074	0.7115	0.7125	0.7048	**0.7035**
MALME (11738)	0.7673	0.7653	0.7683	0.7710	**0.7582**

(c) Missing labels properties: A549 & COLO.

Prop (sample#)	Circular	Neural	Seq2seq	Seq3seq	Ours
CCRE (11738)	**0.9838**	0.9872	0.9975	0.9961	**0.9838**
HL-60 (11738)	0.8063	0.8085	0.8077	0.8123	**0.8043**
K-526 (11738)	**1.0530**	1.0860	1.0694	1.0689	1.0605
RPMI (11738)	1.1251	1.0276	1.0371	1.0292	**1.0217**
A549 (617)	0.8559	0.8840	0.8974	0.8531	**0.8475**
COLO (1191)	0.8244	0.8243	0.8130	0.8109	**0.8056**
HCC (11738)	0.7074	0.7115	0.7125	0.7083	**0.7024**
MALME (11738)	0.7673	0.7680	0.7683	0.7719	**0.7600**

(d) Missing labels properties: HCC & MALME.

Prop (sample#)	Circular	Neural	Seq2seq	Seq3seq	Ours
CCRE (11738)	0.9838	0.9896	0.9975	0.9903	**0.9813**
HL-60 (11738)	0.8063	**0.8025**	0.8077	0.8104	0.8073
K-526 (11738)	**1.0530**	1.0778	1.0694	1.0682	1.0602
RPMI (11738)	1.1251	1.0231	1.0371	1.0299	**1.0218**
A549 (11738)	0.7543	0.7571	0.7673	0.7636	**0.7524**
COLO (11738)	0.7554	0.7787	**0.7546**	0.7674	0.7611
HCC (603)	0.7769	0.7364	0.6841	0.6813	**0.6604**
MALME (1128)	0.9930	1.0233	0.9861	0.9940	**0.9707**

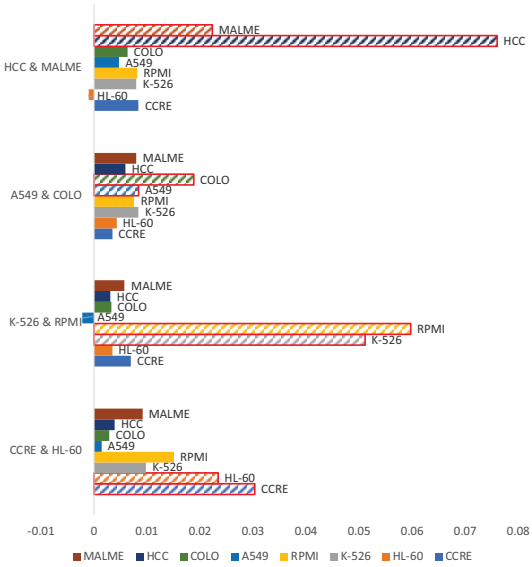

Fig. 2: Results Comparison of the NCI dataset. The figure shows the RMSE difference between our proposed method and the best baseline method in the four NCI experiments.

formance compared with training individual property.

V. Conclusion

In this paper, we propose an innovative idea of combining and training multiple labeled datasets together to perform deep multi-label learning for molecular property prediction. Our method takes advantages of both labeled data and unlabeled data in a multi-label learning fashion, which enables effectively improve the prediction performance from two perspectives, 1) we draw abundant information from vast unlabeled SMILES data to obtain a good encoder model for generating accurate fingerprints, and 2) we exploit labeled data by coordinately training multiple properties to promote the performance of small-sized properties. The results of extensive experiments demonstrate the effectiveness and robustness of our method, which can significantly improve the performance of properties with limited data.

References

[1] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, "The rise of deep learning in drug discovery," *Drug discovery today*, vol. 23, no. 6, pp. 1241–1250, 2018.

[2] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker *et al.*, "Pubchem substance and compound databases," *Nucleic acids research*, vol. 44, no. D1, pp. D1202–D1213, 2015.

[3] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2011.

[4] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, p. 20170387, 2018.

[5] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in neural information processing systems*, 2015, pp. 2224–2232.

[6] R. Li, S. Wang, F. Zhu, and J. Huang, "Adaptive graph convolutional neural networks," *arXiv preprint arXiv:1801.03226*, 2018.

[7] C. Yan, Q. Ding, P. Zhao, S. Zheng, J. Yang, Y. Yu, and J. Huang, "Retroxpert: Decompose retrosynthesis prediction like a chemist," 2020.

[8] R. C. Glen, A. Bender, C. H. Arnby, L. Carlsson, S. Boyer, and J. Smith, "Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to adme," *IDrugs*, vol. 9, no. 3, p. 199, 2006.

[9] Z. Xu, S. Wang, F. Zhu, and J. Huang, "Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery," in *BCB*, 2017.

[10] X. Zhang, S. Wang, F. Zhu, Z. Xu, Y. Wang, and J. Huang, "Seq3seq fingerprint: towards end-to-end semi-supervised deep drug discovery," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2018, pp. 404–413.

[11] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: A simple and general method for semi-supervised learning," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ser. ACL '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 384–394. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858681.1858721

[12] Z.-H. Zhou and J.-M. Xu, "On the relation between multi-instance learning and semi-supervised learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 1167–1174.

[13] P. Liang, "Semi-supervised learning for natural language," Ph.D. dissertation, Massachusetts Institute of Technology, 2005.

[14] A. Søgaard, "Semi-supervised learning and domain adaptation in natural language processing," *Synthesis Lectures on Human Language Technologies*, vol. 6, no. 2, pp. 1–103, 2013.

[15] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[16] N. M. O'Boyle, C. M. Campbell, and G. R. Hutchison, "Computational design and selection of optimal organic photovoltaic materials," *The Journal of Physical Chemistry C*, vol. 115, no. 32, pp. 16 200–16 210, 2011.

[17] H. Morgan, "The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service," *J. Chemical Documentation*, vol. 5, pp. 107–113, 1965.

[18] Y. Hu, E. Lounkine, and J. Bajorath, "Improving the search performance of extended connectivity fingerprints through activity-oriented feature filtering and application of a bit-density-dependent similarity function," *ChemMedChem*, vol. 4, no. 4, pp. 540–548, 2009.

[19] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[20] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[21] T. K. Ho, "Random decision forests," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 1. IEEE, 1995, pp. 278–282.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] S. Hochreiter and J. Schmidhuber, "Lstm can solve hard long time lag problems," in *Advances in neural information processing systems*, 1997, pp. 473–479.

[24] T. Sterling and J. J. Irwin, "Zinc 15–ligand discovery for everyone," *Journal of chemical information and modeling*, vol. 55, no. 11, pp. 2324–2337, 2015.

[25] B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing, and Z. Wu, *Deep Learning for the Life Sciences*. O'Reilly Media, 2019, https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

[26] R. Li and J. Huang, "Learning graph while training: An evolving graph convolutional neural network," *arXiv preprint arXiv:1708.04675*, 2017.