Seq3seq Fingerprint: Towards End-to-end Semi-supervised Deep Drug Discovery

Xiaoyu Zhang
The University of Texas at Arlington
Arlington, Texas
xiaoyu.zhang2@mavs.uta.edu

Sheng Wang
The University of Texas at Arlington
Arlington, Texas
sheng.wang@mavs.uta.edu

Feiyun Zhu
The University of Texas at Arlington
Arlington, Texas
feiyun.zhu@uta.edu

Zheng Xu
The University of Texas at Arlington
Arlington, Texas
zheng.xu@mavs.uta.edu

Yuhong Wang
National Center for Advancing
Translating Sciences, NIH
Rockville, Maryland
yuhong.wang@nih.gov

Junzhou Huang*
The University of Texas at Arlington
Tencent AI Lab
jzhuang@uta.edu

ABSTRACT

Observing the recent progress in Deep Learning, the employment of AI is surging to accelerate drug discovery and cut R&D costs in the last few years. However, the success of deep learning is attributed to large-scale clean high-quality labeled data, which is generally unavailable in drug discovery practices.

In this paper, we address this issue by proposing an end-to-end deep learning framework in a semi-supervised learning fashion. That is said, the proposed deep learning approach can utilize both labeled and unlabeled data. While labeled data is of very limited availability, the amount of available unlabeled data is generally huge. The proposed framework, named as **seq3seq fingerprint**, automatically learns a strong representation of each molecule in an unsupervised way from a huge training data pool containing a mixture of both unlabeled and labeled molecules. In the meantime, the representation is also adjusted to further help predictive tasks, e.g., acidity, alkalinity or solubility classification. The entire framework is trained end-to-end and simultaneously learn the representation and inference results. Extensive experiments support the superiority of the proposed framework.

CCS CONCEPTS

• **Theory of computation → Unsupervised learning and clustering**; **Structured prediction**; • **Applied computing → Molecular sequence analysis**; **Sequencing and genotyping technologies**; *Bioinformatics*; Imaging;

KEYWORDS

Semi-Supervised Learning; Unsupervised Learning; Structured Prediction; Learning Representation; Sequence to Sequence Learning; Deep Learning; Drug Discovery; Virtual Screening; Molecular Representation; Imaging; Computational Biology

ACM Reference Format:
Xiaoyu Zhang, Sheng Wang, Feiyun Zhu, Zheng Xu, Yuhong Wang, and Junzhou Huang. 2018. Seq3seq Fingerprint: Towards End-to-end Semi-supervised Deep Drug Discovery. In *ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3233547.3233548

1 INTRODUCTION

In the past few years, the application of Artificial Intelligence (AI) technologies in drug discovery has become significant and increasingly popular. Observing the most recent rapid growth of a key technology in AI, namely **deep learning** (or **deep neural network**), the whole industry and academia are looking towards AI to speed up the drug discovery, cut R&D cost and decrease the failure rate in potential drug screening trials [6].

However, the previous success of deep learning in multiple applications, e.g., image understanding [8, 34], medical imaging [16, 23, 36, 40], video understanding [2, 46], bioinformatics [43–45], and machine translation [19], etc., has implied a reliance on large-scale high-quality labeled data-sets. The training procedure of those deep-learning-based state-of-the-art models generally involve millions of labeled samples. In the meantime, however, for the drug discovery tasks, the scale of labeled data-set stays around only thousands of examples due to the insanely high cost of obtaining the clean labeled data through the biological experiments. The available amount of the labeled training data is absolutely insufficient to secure the success of the application of deep learning in the drug discovery [27]. This huge gap between the requirement and availability of the labeled data in drug discovery has become a bottleneck of applying deep learning techniques into drug discovery.

Given the high cost of obtaining sufficient labeled data points, it seems impractical to increase the labeled data-set scale to a satisfactory level. To address this issue, we propose a semi-supervised deep learning modeling strategy. In simple terms, the proposed deep

*This work was partially supported by U.S. NSF IIS-1423056, CMMI-1434401, CNS-1405985, IIS-1718853, and NSF CAREER grant IIS-1553687.

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5794-4/18/08…$15.00
https://doi.org/10.1145/3233547.3233548

learning framework can learn from both labeled and unlabeled data, while the unlabeled data is almost infinitely available. For instance, the ZINC data-set [17] is publicly available and contains over 35 million unlabeled molecule data. With such scale of data being used, the deep learning model is expected to be trained with enough representation power to help the inference task.

In this paper, we propose a semi-supervised data-driven multi-task deep-learning-based drug discovery method, named as **seq3seq fingerprint**. The reasons behind this naming are two-fold: 1) this is the **next-generation seq2seq fingerprint** [43], whose major upgrade is that the original two-stage pipeline has been combined into an multi-task one-stage end-to-end pipeline to ensure much more decent inference performance; 2) the seq3seq fingerprint framework contains **three** ends with one input and two outputs while the seq2seq fingerprint contains **two** ends with one input and one output.

To briefly introduce the proposed seq3seq fingerprint framework, the seq3seq fingerprint network can be considered as a pipeline with one input and two outputs. The designed neural network can take the molecule inputs for training, **with or without labels**. The input is the raw sequence representation of a molecule, namely SMILE representation. Examples are referred in Figure 1. The two outputs will correspond to the two tasks inside this network. The first one is the **self-recovery**. The network is expected to be able to generate a vector representation which is able to be recovered back to original raw sequence representation. The second task is the **inference** whenever the label is available. For instance, it can be a task to predict the acidity, alkalinity or solubility of a single molecule. The two tasks are trained within the same network in an end-to-end fashion. As a result, in a specific inference task, the vector representation will be able to provide both good recovery performance and inference performance. Also, the network can be trained inside a mixture data pool with both labeled and unlabeled data, which is sufficient enough to ensure the fine training of the neural network.

The benefits of the seq3seq fingerprint are three folds: 1) the training phase of seq3seq fingerprint takes both labeled and unlabeled data into consideration, which is able to provide both strong vector representation and good inference performance. 2) it is data-driven, eliminating the reliance on expert's subjective knowledge. 3) since the unlabeled data is almost unlimited in practice, it will significantly complement the sole training with labeled data, ensuring a final good inference performance.

The technical contributions of this paper are summarized as: 1) the seq3seq fingerprint method is obviously the first attempt to utilize both labeled data and unlabeled data for sequence-based end-to-end deep learning in drug discovery. 2) several important features are enabled in the seq3seq fingerprint to help inference:

- this is the first **end-to-end** framework coupling both the recovery and inference task.
- the proposed framework is general enough to suit **different prediction tasks**, e.g., classification, regression, etc.
- it is feasible to use **different inference network structures**, e.g., Convolutional Neural Networks (CNNs), Multi-Layer Perceptrons (MLPs), etc.

3) extensive experiments demonstrate the superior performance on different tasks over both supervised and unsupervised state-of-the-art fingerprint methods.

The rest of the paper is organized as follows. We summarize several related work in drug discovery, in Section 2. In Section 3, we describe our entire pipeline in details. We show our experiment results in Section 4, demonstrating the superior performance of our method. We conclude and discuss the future direction of our paper in Section 5.

2 RELATED WORK

In this section, we briefly introduce several related works. First, we present the raw representation of molecules, namely SMILE representation, i.e., the persistence form of the molecular data in the cold data storage. Second, we list a few state-of-the-art fingerprint methods, including the ones using human-designed and hash-based features.. Finally, we briefly describe some most recent deep learning based methods, e.g., neural fingerprint [5], seq2seq fingerprint [43].

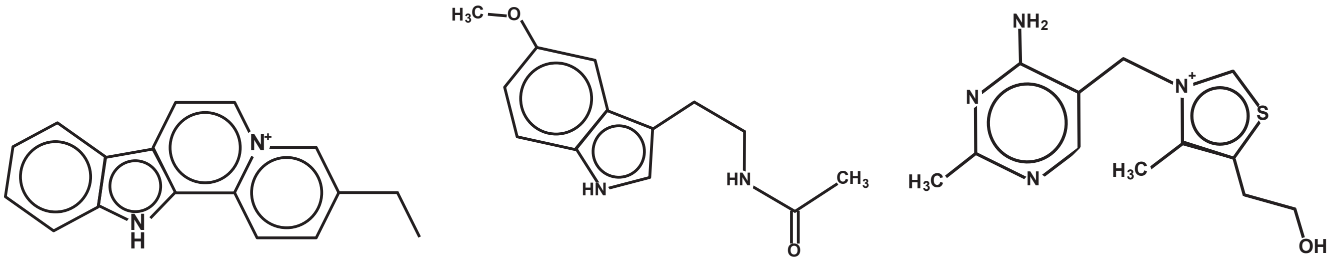
2.1 SMILE Representations of Molecules

Initially, the molecules are stored in the form of a sequence representation, namely the Simplified Molecular-Input Line-Entry system (SMILE) [37], which is a line notation for describing the structure of chemical species using text strings. The SMILE system represents the chemical structures in a graph-based definition, where the atoms, bonds and rings are encoded in a graph and represented in text sequences. Simple examples of SMILE representations are 1) dinitrogen with structure $N \equiv N$ (N#N), 2) methyl isocyanate with structure $CH_3 - N = C = O$ (CN=C=O), where corresponding SMILE representations are included in the brackets. Simply speaking, the letters, e.g., C, N, generally represent the atoms, while some symbols like −, =, # represent the bonds. We show some more complicated examples in Figure 1.

2.2 Fingerprint Methods

Hash-based Fingerprints. Many hash-based methods has been developed to generate unique molecular feature representation [12, 15, 24]. One important class is called **circular fingerprints**. Circular fingerprints generate each layer's features by applying a fixed hash function to the concatenated features of the neighborhood in the previous layer. One of the most famous ones is Extended-Connectivity FingerPrint (ECFP) [29]. However, due to the non-invertible nature of the hash function, the hash-bashed fingerprint methods usually do not encode enough information and hence result in lower performance in the further predictive tasks.

Biologist-guided Local-Feature Fingerprints. Another mainstream of traditional fingerprint methods is designed based on the biological experiments and the expertise knowledge and experience, e.g., [26, 30]. Biologists look for several important task-related substructures (fragments), e.g., $CC(OH)CC$ for pro-solubility prediction, and count those sub-structures as local features to produce fingerprints. This kind of fingerprint methods usually work well for specific tasks, but poorly generalize for other tasks.



(a) Flavopereirin: CCc(c1)ccc2[n+]1ccc3c2Nc4c3cccc4 **(b) Melatonin: CC(=O)NCCC1=CNc2c1cc(OC)cc2** **(c) Thiamine: OCCc1c(C)[n+](cs1)Cc2cnc(C)nc2N**

Figure 1: The examples of SMILE representations.

2.3 Deep-learning-based Models

The growth of deep learning [20, 39] has provided the great flexibility and performance to create the molecular fingerprint from data samples, without explicit human guide, [3, 9, 18, 31, 35, 43]. In this subsection, we discuss two major classes, namely supervised and unsupervised learning models.

Supervised Models. Many of deep learning-based fingerprint methods are still trained in a supervised-learning fashion [31, 38], which is using only labeled molecular data samples as inputs and adjusting model weights according to their labels [21]. However, as mentioned earlier, the performance of the deep supervised learning models are generally limited by the availability of the labeled data. The state-of-the-art work is the neural fingerprint [9]. The neural fingerprint mimics the process of generating circular fingerprint but instead the hash function is replaced by a non-linear activated densely connected layer. This method is based on the deep graph convolutional neural network [13, 21, 22, 25]. There are also few attempts that address the insufficient label issue by using few-shot learning strategies, e.g., [4]. To secure a satisfactory performance and acquire enough labeled data, biologists need to perform a sufficiently large number of tests on chemical molecules, which is prohibitively expensive.

Unsupervised Models. Recently, few unsupervised fingerprint methods, e.g., seq2seq fingerprint [43], are proposed to alleviate the issue brought by the insufficient labeled data. These models generally train deep neural networks to provide strong vector representations using a big pool of unlabeled data. The vector representation model is thereafter used for supervised training with other models, e.g., Adaboost [10], GradientBoost [11], and RandomForest [14], etc. Since the deep models are trained with a sufficiently large data-set, the representation is expected to contain enough information to provide good inference performance. However, this type of methods are not trained end-to-end, meaning that the representation only adjusts to the recovery task of the original raw representation. It is robust to the specific labeled task, but might not provide optimal inference performance for each task.

3 METHODOLOGY

In this section, we describe the details of our semi-supervised seq3seq fingerprint model. First, an overview of the proposed seq3seq fingerprint model is given. The proposed semi-supervised model is trained in an end-to-end fashion by completing two tasks, a self-recovery task for molecule (without any label) and an inference task (with specific classification/regression label). After that, we describe the recovery task and the inference task in detail, their loss functions and how the two tasks are trained. Then the semi-supervised loss is described. In the end, we offer a multi-task scaffolding view from frame-semantic parsing [33] in natural language processing area to explain the proposed model.

3.1 Overview

Different from traditional models [5, 43], the proposed seq3seq fingerprint model works in a semi-supervised fashion. It means that our training data comes from two sources, the labeled data, for classification/regression, as well as the unlabeled data. The labeled data contains the SMILE strings for molecule data and their labels, such as acidity or other molecular activities. The unlabeled data contains just molecular SMILE strings and the unlabeled data is almost infinitely available. The proposed seq3seq fingerprint model takes the mixture of the labeled data and unlabeled data together as training inputs to the network. The work flow is depicted in Figure 2. The semi-supervised training is done by two tasks: the self-recovery task and the inference task. The whole pipeline is illustrated in Figure 3.

3.2 The Duo Tasks in Seq3seq Fingerprint Model

The Self-recovery Task The self-recovery task is to learn a vector representation (usually noted as **fingerprint** in the drug discovery literature) for each input molecular SMILE string. This task also requires the SMILE string of the molecule can be recovered from its fingerprint vector. It is an unsupervised learning problem since no label information is required in training. As shown in Figure 3, this task contains a perceiver network and an interpreter network. This structure is motivated by the seq2seq model [32, 43]. The original seq2seq model is used in machine translation [32]. It is to learn a vector representation from a sentence in a given language, e.g., English, then translate the learned representation into another language such as French. Seq2seq fingerprint [43] combines the idea from seq2seq learning and the idea of auto-encoder to learn the vector representation for molecule.

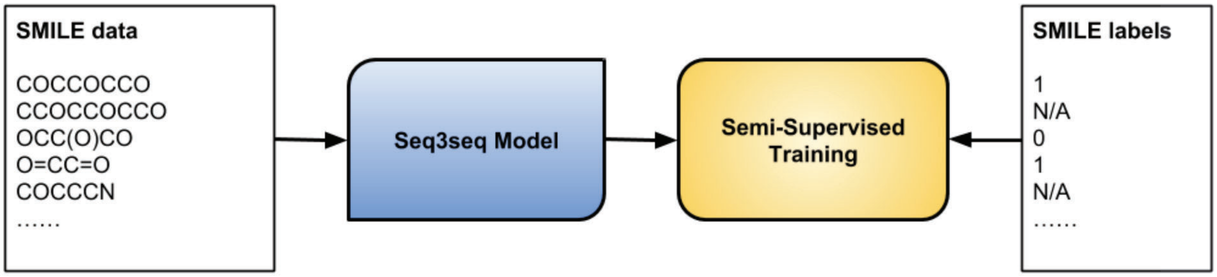


Figure 2: This figures shows how semi-supervised training is used for our proposed model. We mix the unlabeled data and labeled data together to train our proposed model. The SMILEs with label 0/1 come from labeled dataset and the SMILEs without labels (N/A in the figure) come from unlabeled dataset.

We generalize the idea of seq2seq [5, 43] in two views. First, the perceiver network and the interpreter network in the proposed seq3seq fingerprint model can be any recurrent deep neural networks such as LSTM, GRU neural networks. The only limitation is that the perceiver network could map the string tokens into a vector representation and the interpreter could map the vector back into string tokens. Second, we introduce unlabeled molecule data into our training process to learn better representations. Instead of using the SMILE strings of only the labeled molecule data, we take advantage of the **almost infinite** unlabeled data and use both unlabeled and labeled data for the self-recovery task to learn a more accurate vector presentation than those models which only use labeled data or unlabeled data separately. The loss function in our proposed model follows the one in [43]. It is the sum of multiple cross-entropy loss and we denote it as $\mathcal{L}_{unsup}$.

The Inference Task The inference task in the proposed seq3seq fingerprint model is to predict the activity of molecules. In the proposed model, the inference task includes the perceiver network and the inference network. The perceiver network is shared in both self-recovery and inference tasks. It is trained by both labeled and unlabeled data in an end-to-end fashion. The inference network maps the seq3seq fingerprint to a final inference result on a certain prediction task. The structure of the inference network can be any trainable network which maps the vector into a inference value. It allows huge flexibility for the choice of the inference network. For instance, it could be a Convolutional Neural Network (CNN), a Multi-Layer Perceptron (MLP) or even a single fully-connected layer. Depending on whether the inference task is classification or regression, the loss for the inference task $\mathcal{L}_{sup}$ could be either classification loss (usually a cross entropy loss) or regression loss (usually a ℓ_1 smooth/ℓ_2 distance loss). Since computing the $\mathcal{L}_{sup}$ needs labels, the inference task is only trained on labeled data.

3.3 End-to-end Semi-supervised Learning

As shown in Figure 3, the semi-supervised loss $\mathcal{L}_{semi}$ combines the unsupervised loss $\mathcal{L}_{unsup}$ and the supervised loss $\mathcal{L}_{sup}$ together as

$$\mathcal{L}_{semi} = \mathcal{L}_{unsup} + \lambda \mathcal{L}_{sup}. \quad (1)$$

where λ is a hyper-parameter of the proposed model to balance the two tasks. The proposed model is trained with both supervised data and unsupervised data. When the data is unlabeled, the supervised loss $\mathcal{L}_{sup}$ will be zero. Thus, in this case, only the part of the model in self-recovery task will be trained. While the data is labeled, both the part of the model in self-recovery and inference will be trained. The end-to-end training avoids the multi-stage training, i.e., pre-trained model training or separated classifier training [43]. As a result, the proposed end-to-end model is expected to provide an optimal inference performance as well as shorter training time for specific task than that in a multi-stage model from [43].

3.4 A Multi-task Scaffolding View of Seq3seq Fingerprint

In [43], the authors viewed seq2seq fingerprint as a machine translation problem in the Natural Language Processing (NLP) area, with both source and target language set to be the SMILE representation. Interestingly, the proposed seq3seq fingerprint model can be viewed, to some extent, as **a multi-task scaffolding framework** [33] in the NLP area as well. In [33], the authors focus on solving the frame-semantic parsing problem, which is basically finding the *action* (frame) with its associated objects from a sentence. For example, in sentence "Alice loves Bob.", the frame is "loves" with its associated objects being "Alice" and "Bob". However, a single sequence-to-frame network model generally performs poorly in this task. In [33], they proposed to use a multi-task framework to refine the predictions. Besides the frame parsing task, they also introduce the syntactic parsing task. The second task is basically predicting the word categories, e.g., nouns, adverbs, adjectives, etc. For the previous "Alice loves Bob." sentence, the result will be that "Alice" being noun, "loves" being verb and "Bob" being another noun. In [33], it is demonstrated that the second task significantly helps the success of the main (frame parsing) task. To sum up, the multi-task scaffolding frame parsing framework utilizes a second *syntactic parsing* task to reinforce the main task which is the *frame parsing*. Our seq3seq fingerprint can be viewed in a very similar fashion: the **self-recovery task** serves as the auxiliary task to augment the main **prediction task**. This modification is also further demonstrated superior in our experiments described in Section 4.

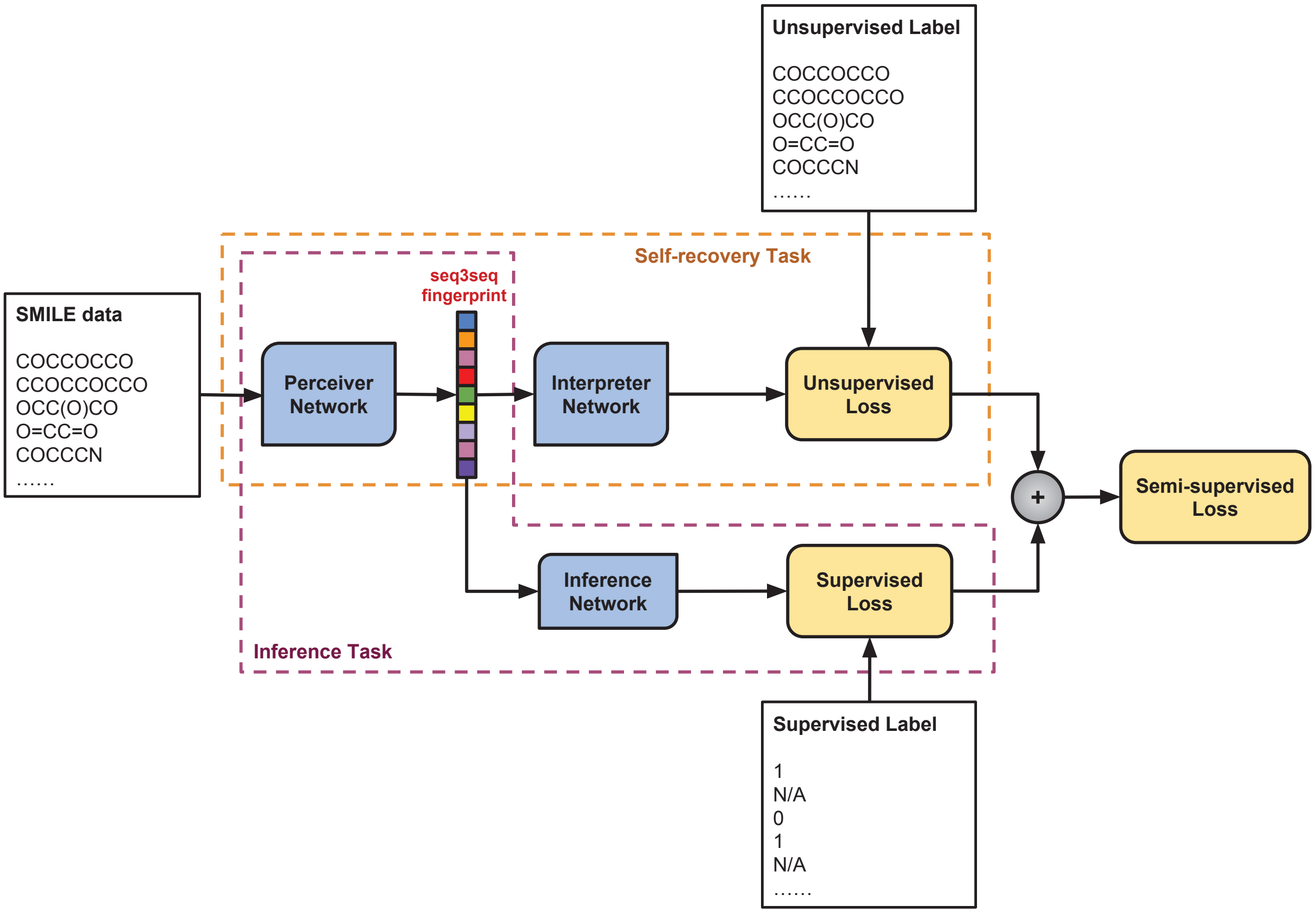


Figure 3: This figure shows the proposed seq3seq fingerprint model. The proposed model is trained through two tasks: a self-recovery task and an inference task. The self-recovery task contains a perceiver network and an interpreter network; the inference task shares the perceiver with self-recover task and has an inference network. The semi-supervised loss is the sum of supervised loss and unsupervised loss.

4 EXPERIMENTS

In this section, we first detail the experimental setup, e.g., the data set description, hardware and software settings, etc. Then we report the benchmark performance of the seq3seq fingerprint methods among state-of-the-art methods. Furthermore, to show the flexibility of our methods and complete our experiments, we offer ablation studies for the sensitivity of the hyper-parameters of our seq3seq fingerprint models, e.g., the multi-task balance weight λ, the Recurrent Neural Network (RNN) layer hidden size and layer number, etc.

4.1 Experiment Setup

Datasets As we mentioned in the introduction, the seq3seq fingerprint can be trained from a mixture of both unlabeled and labeled data. In practices, we usually use an unlabeled data set of a much larger size than that of a labeled dataset.

Unlabeled Dataset For (large) unlabeled dataset, we use ZINC drug-like datasets [17]. ZINC is a free database of commercially-available compounds for virtual screening. The drug-like dataset from ZINC contains 18,691,354 molecular SMILE representations.
Labeled Dataset Two additional datasets, LogP and PM2-10k, were used for semi-supervised training and test. They are obtained from National Center for Advancing Translational Sciences (NCATS) at National Institutes of Health (NIH). Each of them contains around 10,000 molecular SMILE representations with multiple scores, each score quantifies some chemical property. Classification was conducted on LogP and PM2-10k.

- **LogP**: Totally 10,850 samples were used from LogP, Each sample contains a pair of a SMILE string and a water-octanol partition coefficient (LogP) value. A threshold of 1.88 is used to label the data. For those samples with LogP value smaller than 1.88 were classified as negative samples, the rest were labeled as positive samples.

Table 1: The comparison of classification accuracy on the LogP data. We report the average classification accuracy (Mean) and the corresponding Standard Deviation (StDev) of 5-fold cross-validation result.

	Circular [29]	Neural [9]	seq2seq [43]	seq3seq (Ours)
Mean	36.74%	60.80%	76.64%	**89.72%**
StDev	0.74%	1.35%	0.43%	0.41%

Table 2: The comparison of classification accuracy on the PM2-10k data. We report the average classification accuracy (Mean) and the corresponding Standard Deviation (StDev) of 5-fold cross-validation result.

	Circular [29]	Neural [9]	seq2seq [43]	seq3seq (Ours)
Mean	39.38%	52.27%	62.06%	**68.45%**
StDev	1.14%	1.12%	1.98%	0.80%

- **PM2-10k**: PM2-10k dataset contains 10,000 samples of SMILE strings and binary promiscuous class labels. Similarly, a threshold of 0.024896 was used to classify each SMILE. Samples with value larger than the threshold were considered as positive 1; otherwise, labeled as 0.

We mix the ZINC drug-like dataset with the labeled dataset and train the recovery and inference task simultaneously on the mixed dataset.

Neural Network Structures As we mentioned earlier, the proposed seq3seq fingerprint framework is super flexible in the choice of the network structure. Theoretically, both perceiver and interpreter network can use any stacked Recurrent Neural Network (RNN) with different layers and layer hidden sizes. Also the RNN cell can be formed in different types, e.g., LSTM, GRU, etc. Due to the page limit of this paper, we hereby assume the perceiver and interpreter network always use the same type of RNN cells with the same number of layers and hidden sizes. In this section, we only discuss Gated Recurrent Unit (GRU) [7] as the RNN cell. Also, we limit the discussion of the inference network to a single densely connected layer with the output number equaling the number of the classification class number. For simplicity, we use $GRU - L - H$ to represent the network structure, where GRU is the RNN cell type, $L \in N^+$ is the stacked RNN layer number and $H \in N^+$ is the RNN cell hidden size. For instance, $GRU - 2 - 256$ represents a seq3seq model where both perceiver and interpreter network use 2-layer GRU cell with 256 hidden units.

Learning Hyper-parameters For optimization, we use the Stochastic Gradient Descent (SGD) with a heuristic learning rate decaying schedule. The initial learning rate is 0.5 for any training models. The learning rate will be decayed by a factor of 0.99 if the test loss does not decrease after 600 training steps. The training will automatically halt if the learning rate is smaller than $1e - 7$. Under the above hyper-parameter sets, the training of each model in the semi-supervised setting can generally finish within a few hours.

Evaluation Metrics Given that we have two tasks of our semi-supervised learning framework, i.e., recovery and inference task, we report two evaluation metrics for each model we trained. For recovery task, we use an Exact Match Accuracy (EMA) for evaluation. This metric measure the portion of the exactly recovered sequence within the entire set of sequences. Furthermore, we report the classification accuracy (hereafter SSLA for Semi-Supervised Learning Accuracy) for our classification task.

Comparison Methods We compare our semi-supervised method with the unsupervised seq2seq fingerprint method [43] as well as several other state-of-the-art methods: the ECFP [29] (circular fingerprint) and the neural fingerprint method [9]. We download the official implementation of the seq2seq fingerprint [1] and carefully follow the experimental setting of the authors. The circular fingerprint is a hand-crafted hash-based feature that was generated through RDKit [2]. The neural fingerprint implementation is obtained from https://github.com/HIPS/neural-fingerprint, which we slightly modify to adapt our dataset file format.

Infrastructure and Software The seq3seq fingerprint method was implemented through Tensorflow package [1], and our semi-supervised model was trained in a self-hosted 16-GPU cluster platform with Intel i7 6700K @ 4.00 GHz CPU, 64 Gigabytes RAM and four Nvidia GTX 1080Ti GPUs on each workstation. The code will be released upon the acceptance of this paper.

4.2 Comparison with State-of-the-art Methods

In Table 1 and 2, we report the 5-fold cross validation average classification accuracy on LogP and PM2-10k datasets. The proposed methods are compared with ECFP (circular) fingerprint [12], neural fingerprint [5] and seq2seq fingerprint [43]. For seq2seq fingerprint, according to their paper, the seq2seq fingerprint with length 1024 + Gradient Boosting always provides best performance, so we only report those results on our paper.

It is shown that on both datasets, the seq3seq fingerprint always provides best inference performance. On LogP dataset, our seq3seq model performs significantly superior than the other state-of-the-art methods, up to 13% in terms of classification accuracy (SSLA in the tables). Compared with circular fingerprint, the seq3seq fingerprint is data-driven and contains enough information to be recovered. The performance of neural fingerprint is generally limited by the availability of the labeled data. Seq2seq fingerprint is the closest

[1] https://github.com/XericZephyr/seq2seq-fingerprint

[2] http://www.rdkit.org

work in terms of accuracy for now since it can be also trained on the huge pool of unlabeled data, extracting a good representation and train/infer with a sophisticated classification model. However, seq2seq fingerprint is, unfortunately, not an end-to-end framework, which means the recovery and inference training of seq2seq fingerprint are separate. The unsupervised recovery training can bring in considerable amount of noise in the representation which limits further improvements of the inference performance. The seq3seq fingerprint, which uses the inference task to correct the recovery task during training, can constantly provide the best performance among all of the comparison methods.

Table 3: The performance variations with λ and GRU model parameters for LogP data. Layer: the stacked layer number of RNN cells. LD: Latent Dimension (hidden size) of RNN cells. EMA: Exact Match Accuracy for self-recovery task. SSLA: classification accuracy for inference task.

Layer	LD	λ	EMA	SSLA
2	128	1	86.31%	89.46%
		0.1	91.80%	89.62%
		0.01	90.23%	81.05%
		0.001	91.42%	64.95%
2	256	1	93.59%	90.18%
		0.1	94.52%	89.35%
		0.01	95.77%	84.65%
		0.001	95.48%	69.16%

Table 4: The performance variations with λ and GRU model parameters for PM2-10k data. Layer: the stacked layer number of RNN cells. LD: Latent Dimension (hidden size) of RNN cells. EMA: Exact Match Accuracy for self-recovery task. SSLA: classification accuracy for inference task.

Layer	LD	λ	EMA	SSLA
2	256	1	87.48%	65.28%
		0.1	89.84%	64.85%
		0.01	91.73%	62.37%
		0.001	91.31%	50.66%
3	256	1	82.40%	64.90%
		0.1	87.61%	67.92%
		0.01	89.33%	68.24%
		0.001	90.25%	50.07%

4.3 Sensitivity Analysis of Multi-task Weight Balance Parameters

In multi-task machine learning practice, the weight balancing hyper-parameters among different tasks (in our case, λ in the loss function) are sometimes critical and sensitive to data. This might not be an intriguing feature in practices. However, our method is quite robust and tolerant with λ variations. In this subsection, we report our sensitivity studies of λ. We choose different scale of λ to see how the final model performance responds to the variance of λ , showing the robustness of our method with regard to different weight balancing hyper-parameters.

In Table 3, 4 as well as Figure 5, we vary λ in the logarithm scale with a base of 10. We tried $10^0, 10^{-1}, 10^{-2}, 10^{-3}$. On both datasets, it looks that within a quite wide range of λ, i.e., $10^{-2} - 10^0$, the performance is quite robust to the change of λ. The reason behind this robustness might be the huge unlabeled data pool used in the training process. Given the model has been trained with a sufficiently large (up to dozens of millions) molecular data pool, the resulting model will automatically adjust to a small task weight perturbation.

4.4 The Ablation Study of Neural Network Structures

In this section, we provide a comprehensive study of the impacts of different layers and layer hidden sizes of our seq3seq fingerprint models. We report the 5-fold cross validation Exact Match Accuracy (EMA) and the classification accuracy (SSLA) in Table 5 and 6 for each of the two datasets, respectively. Figure 4 (a) and (b) also illustrates the trends when varying the layer numbers and layer hidden sizes.

Inference Task It is super exciting to reveal the **robustness of classification accuracy to the change of network structures** on both datasets. In Figure 4, the classification accuracy (blue bars) almost stays at the same height when varying the layer numbers and layer hidden sizes. This implies the importance of the representation learning inside the seq3seq fingerprint. This further support the positive effects of the large-scale (up to dozens of millions) unlabeled data utilization.

When the inference is super robust to the network changes, for self-recovery task (in terms of EMA), we observe a decreasing trend when increasing the layer depth (numbers). Meanwhile, the increasing number of hidden units inside each layer generally yields better EMA. This suggests that the improvement of self-recovery task has higher reliance on the layer hidden sizes. Deeper network might not always be an elixir for a simple auxiliary task like self-recovery. This observation might help future network design. To simultaneously ensure high inference performance and reduce training time (deeper network generally takes longer to train.), it might be a good idea to use reasonably deep and wide RNN networks.

5 CONCLUSIONS

In this paper, we discuss a new semi-supervised deep learning based molecular prediction system, called **seq3seq fingerprint**. Our model is the first attempt in sequence-based deep learning method utilizing both unlabeled and labeled data for drug discovery. The reinforcement from the unlabeled data is demonstrated to significantly improve the inference performance by enhancing the representation power of the perceiver network. As a result, the superior inference performance over multiple state-of-the-art methods is revealed in our extensive experiments.

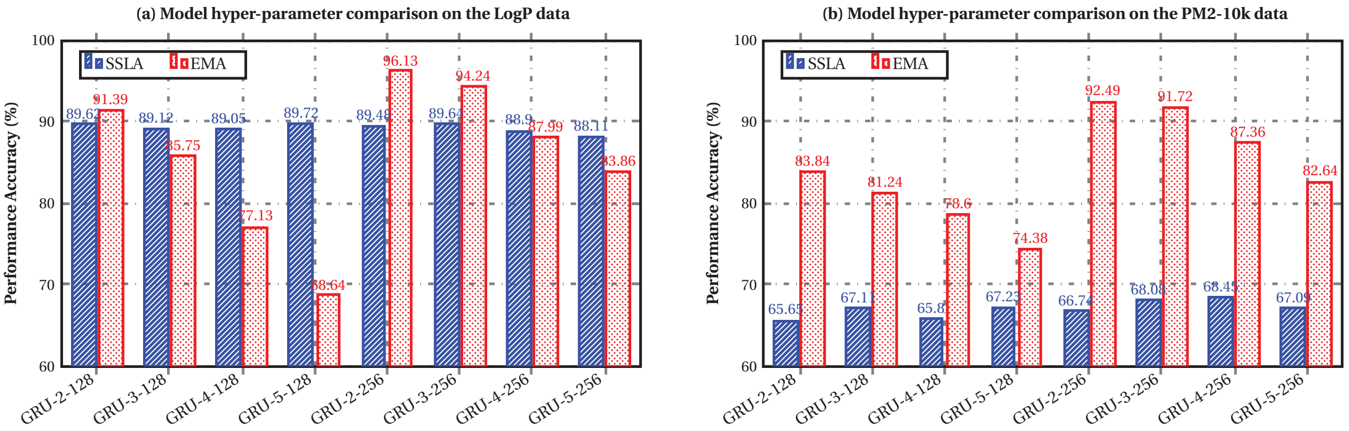
In the future, a potential direction might be improving the training algorithm [28, 41, 42]. Furthermore, our seq3seq fingerprint method still share some common aspects with Natural Language Processing (NLP) area as the seq2seq fingerprint does [43]. As described in Section 3, it looks that we have found a new direction

Table 5: The comparison of 5-fold cross validation classification accuracy among different seq3seq GRU models on the LogP data. Both average (Mean) and Standard Deviation (StDev) are reported for the 5-fold splits. FP Length: FingerPrint Length. SSLA: classification accuracy for inference task. EMA: Exact Match Accuracy for self-recovery task.

	GRU-2-128	GRU-3-128	GRU-4-128	GRU-5-128	GRU-2-256	GRU-3-256	GRU-4-256	GRU-5-256
FP Length	256	384	512	640	512	768	1024	1280
SSLA Mean	89.62%	89.12%	89.05%	**89.72**%	89.48%	89.64%	88.90%	88.11%
SSLA StDev	0.62%	0.22%	0.10%	0.41%	0.44%	0.42%	0.31%	0.40%
EMA Mean	91.39%	85.75%	77.13%	68.64%	96.13%	94.24%	87.99%	83.86%
EMA StDev	0.46%	0.53%	0.56%	0.80%	0.21%	0.31%	0.45%	0.41%

Table 6: The comparison of 5-fold cross validation classification accuracy among different seq3seq GRU models on the PM2-10k data. Both average (Mean) and Standard Deviation (StDev) are reported for the 5-fold splits. FP Length: FingerPrint Length. SSLA: classification accuracy for inference task. EMA: Exact Match Accuracy for self-recovery task.

	GRU-2-128	GRU-3-128	GRU-4-128	GRU-5-128	GRU-2-256	GRU-3-256	GRU-4-256	GRU-5-256
FP Length	256	384	512	640	512	768	1024	1280
SSLA Mean	65.65%	67.11%	65.80%	67.23%	66.74%	68.08%	**68.45**%	67.09%
SSLA StDev	0.19%	0.85%	0.61%	0.52%	0.57%	0.35%	0.80%	0.67%
EMA Mean	83.84%	81.24%	78.60%	74.38%	92.49%	91.72%	87.36%	82.64%
EMA StDev	0.45%	0.67%	0.88%	0.88%	0.37%	0.25%	0.29%	0.76%


(a) Model hyper-parameter comparison on the LogP data

(b) Model hyper-parameter comparison on the PM2-10k data

Figure 4: Impacts of the network structures on different metrics on both LogP and PM2-10k dataset. 1) The robustness of inference performance (SSLA, blue bars) is revealed. 2) The positive and negative correlations with regard to the self-recovery performance (EMA, red bars) are observed for RNN network depths and widths, respectively.

to invent new drug discovery methods. In the future, it might be interesting to further investigate bonds between drug discovery and NLP area, which might bring in many novel methods to further accelerate drug discovery research.

Acknowledgment The authors would like to thank NVIDIA for GPU donation and the NIH and UCSF for sharing the drug discovery datasets. The statements contained herein are solely of the authors and do not represent or imply concurrence or endorsement by NCI.

REFERENCES

[1] MartÃŋn Abadi and et.al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. http://download.tensorflow.org/paper/whitepaper2015.pdf

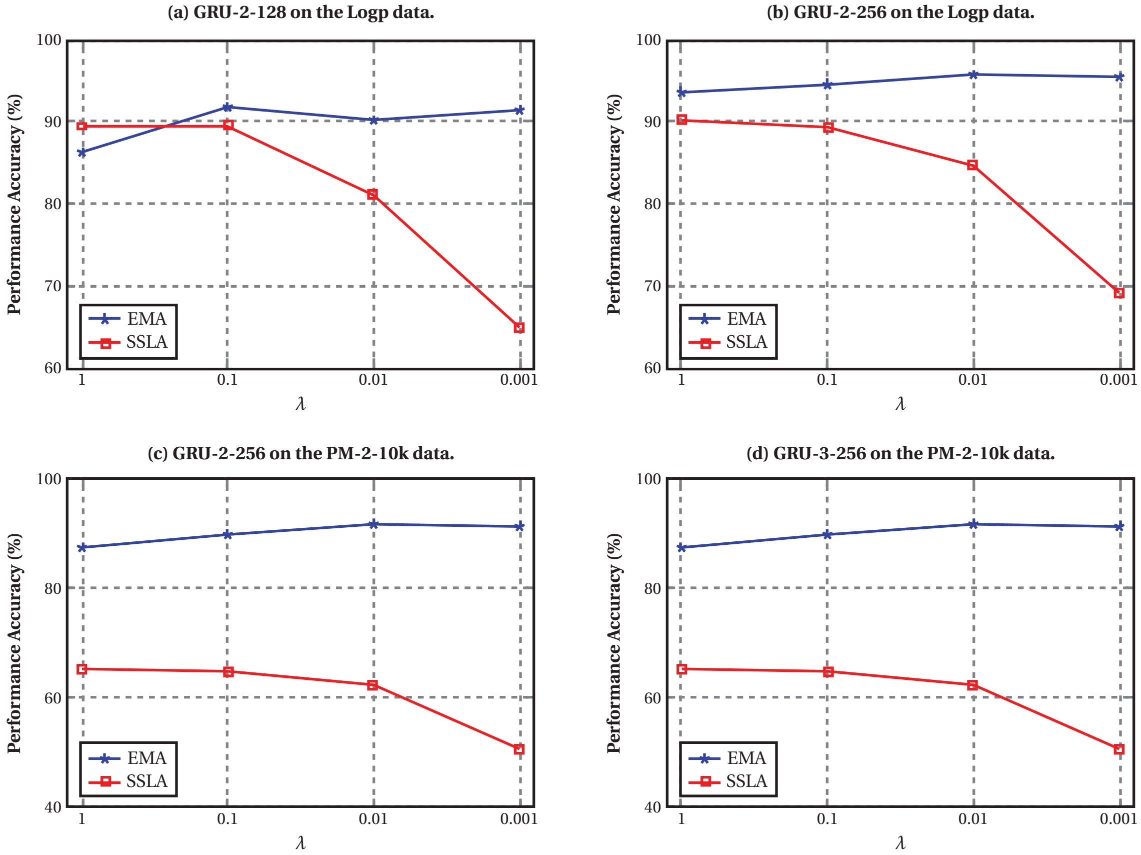
[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[3] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. 2016. Low Data Drug Discovery with One-shot Learning. *arXiv preprint arXiv:1611.03199* (2016).

[4] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. 2017. Low data drug discovery with one-shot learning. *ACS central science* 3, 4 (2017), 283–293.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[6] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz,



(a) GRU-2-128 on the Logp data.

(b) GRU-2-256 on the Logp data.

(c) GRU-2-256 on the PM-2-10k data.

(d) GRU-3-256 on the PM-2-10k data.

Figure 5: Impacts of the multi-task balance weights on different scales on both LogP and PM2-10k dataset. Within a very wide range (usually $10^{-2} - 10^{0}$), both self-recovery (EMA) and inference (SSLA) performance are quite robust to the change of λ.

Michael M Hoffman, et al. 2018. Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv* (2018), 142760.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[9] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*. 2224–2232.

[10] Yoav Freund and Robert E Schapire. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*. Springer, 23–37.

[11] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[12] Robert C Glen, Andreas Bender, Catrin H Arnby, Lars Carlsson, Scott Boyer, and James Smith. 2006. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* 9, 3 (2006), 199.

[13] Joseph Gomes, Bharath Ramsundar, Evan N Feinberg, and Vijay S Pande. 2017. Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity. *arXiv preprint arXiv:1703.10603* (2017).

[14] Tin Kam Ho. 1995. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, Vol. 1. IEEE, 278–282.

[15] Ye Hu, Eugen Lounkine, and Jürgen Bajorath. 2009. Improving the Search Performance of Extended Connectivity Fingerprints through Activity-Oriented Feature Filtering and Application of a Bit-Density-Dependent Similarity Function. *ChemMedChem* 4, 4 (2009), 540–548.

[16] Junzhou Huang and Zheng Xu. 2017. Cell Detection with Deep Learning Accelerated by Sparse Kernel. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing*. Springer, 137–157.

[17] John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. 2012. ZINC: a free tool to discover chemistry for biology. *Journal of chemical information and modeling* 52, 7 (2012), 1757–1768.

[18] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. 2016. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design* 30, 8 (2016), 595–608.

[19] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, Vol. 5. 79–86.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[21] Ruoyu Li and Junzhou Huang. 2017. Learning Graph While Training: An Evolving Graph Convolutional Neural Network. *arXiv preprint arXiv:1708.04675* (2017).

[22] Ruoyu Li, Sheng Wang, Feiyun Zhu, and Junzhou Huang. 2018. Adaptive Graph Convolutional Neural Networks. *arXiv preprint arXiv:1801.03226* (2018).

[23] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42 (2017), 60–88.

[24] HL Morgan. 1965. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chemical Documentation* 5 (1965), 107–113.

[25] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. 2016. Learning convolutional neural networks for graphs. In *International conference on machine learning*. 2014–2023.

[26] Noel M O'Boyle, Casey M Campbell, and Geoffrey R Hutchison. 2011. Computational design and selection of optimal organic photovoltaic materials. *The Journal of Physical Chemistry C* 115, 32 (2011), 16200–16210.

[27] Hao Pan, Zheng Xu, and Junzhou Huang. 2015. An effective approach for robust lung cancer cell detection. In *International Workshop on Patch-based Techniques in Medical Imaging*. Springer, 87–94.

[28] Zhongxing Peng, Zheng Xu, and Junzhou Huang. 2016. RSPIRIT: Robust self-consistent parallel imaging reconstruction based on generalized Lasso. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, 318–321.

[29] David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling* 50, 5 (2010), 742–754.

[30] Chetan Rupakheti, Aaron Virshup, Weitao Yang, and David N Beratan. 2015. Strategy to discover diverse optimal molecules in the small molecule universe. *Journal of chemical information and modeling* 55, 3 (2015), 529–537.

[31] Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. 2016. Computational Modeling of β-secretase 1 (BACE-1) Inhibitors using Ligand Based Approaches. *Journal of Chemical Information and Modeling* 56, 10 (2016), 1936–1949.

[32] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[33] Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A Smith. 2017. Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. *arXiv preprint arXiv:1706.09528* (2017).

[34] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning.. In *AAAI*, Vol. 4. 12.

[35] Izhar Wallach, Michael Dzamba, and Abraham Heifets. 2015. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855* (2015).

[36] Sheng Wang, Jiawen Yao, Zheng Xu, and Junzhou Huang. 2016. Subtype cell detection with an accelerated deep convolution neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 640–648.

[37] David Weininger. 1970. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. In *Proc. Edinburgh Math. SOC*, Vol. 17. 1–14.

[38] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* 9, 2 (2018), 513–530.

[39] Zheng Xu and Junzhou Huang. 2015. Efficient lung cancer cell detection with deep convolution neural network. In *International Workshop on Patch-based Techniques in Medical Imaging*. Springer, 79–86.

[40] Zheng Xu and Junzhou Huang. 2016. Detecting 10,000 Cells in One Second. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 676–684.

[41] Zheng Xu and Junzhou Huang. 2017. A general efficient hyperparameter-free algorithm for convolutional sparse learning.. In *AAAI*. 2803–2809.

[42] Zheng Xu, Yeqing Li, Leon Axel, and Junzhou Huang. 2015. Efficient preconditioning in joint total variation regularized parallel MRI reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 563–570.

[43] Zheng Xu, Sheng Wang, Feiyun Zhu, and Junzhou Huang. 2017. Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery. In *BCB*.

[44] Jiawen Yao, Sheng Wang, Xinliang Zhu, and Junzhou Huang. 2016. Imaging biomarker discovery for lung cancer survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 649–657.

[45] Feiyun Zhu, Jun Guo, Zheng Xu, Peng Liao, and Junzhou Huang. 2018. Group-driven Reinforcement Learning for Personalized mHealth Intervention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

[46] Andrew Zisserman, Joao Carreira, Karen Simonyan, Will Kay, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and Mustafa Suleyman. 2017. The Kinetics Human Action Video Dataset.