Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery

Zheng Xu
The University of Texas at Arlington
701 S. Nedderman Drive
Arlington, Texas 76019
zheng.xu@mavs.uta.edu

Sheng Wang
The University of Texas at Arlington
701 S. Nedderman Drive
Arlington, Texas 76019
sheng.wang@mavs.uta.edu

Feiyun Zhu
The University of Texas at Arlington
701 S. Nedderman Drive
Arlington, Texas 76019
feiyun.zhu@uta.edu

Junzhou Huang*
The University of Texas at Arlington
701 S. Nedderman Drive
Arlington, Texas 76019
jzhuang@uta.edu

ABSTRACT

Many of today's drug discoveries require expertise knowledge and insanely expensive biological experiments for identifying the chemical molecular properties. However, despite the growing interests of using supervised machine learning algorithms to automatically identify those chemical molecular properties, there is little advancement of the performance and accuracy due to the limited amount of training data.

In this paper, we propose a novel unsupervised molecular embedding method, providing a continuous feature vector for each molecule to perform further tasks, e.g., solubility classification. In the proposed method, a multi-layered Gated Recurrent Unit (GRU) network is used to map the input molecule into a continuous feature vector of fixed dimensionality, and then another deep GRU network is employed to decode the continuous vector back to the original molecule. As a result, the continuous encoding vector is expected to contain rigorous and enough information to recover the original molecule and predict its chemical properties. The proposed embedding method could utilize almost unlimited molecule data for the training phase. With sufficient information encoded in the vector, the proposed method is also robust and task-insensitive. The performance and robustness are confirmed and interpreted in our extensive experiments.

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

ACM-BCB '17, August 20-23, 2017, Boston, MA, USA

ACM ISBN 978-1-4503-4722-8/17/08…$15.00
https://doi.org/10.1145/3107411.3107424

CCS CONCEPTS

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

KEYWORDS

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

1 INTRODUCTION

In the most recent decade, every drug company with R&D department has carried out numerous initiatives for speeding up its drug discovery process [14]. Drug discovery is the process through which potential new medicines are identified. Modern drug discovery is usually implemented as drug compound selection, while, for every candidate chemical compound, the chemical drug properties, e.g., affinity, selectivity, metabolic stability, are biologically tested in the lab environment. Once all the properties pass the drug requirement tests, it will be selected as a new potential drug candidate. However, this process is excessively expensive and labor-intensive, and costs hundreds of million dollars each year.

Therefore, using machine learning methods to automatically predict the chemical properties has recently raised great interests in the drug discovery community [3, 5, 27, 35]. However, the majority of machine learning algorithms take fixed-length continuous feature vectors as inputs [55–57]. However, the nature of molecules makes it extremely hard to represent molecules with fixed-length vectors [11, 23, 44]. The readers might refer to Figure 1 to grab some intuition. As a simple example, we may consider H_2O (water) and O_2 (oxygen). They differ in atom types, numbers as well as bond types. One might find it is tricky to represent each molecule as a fixed-length vector. So a large class of research papers has been published to generate the fixed-length continuous vector representation for molecules. Overall, the choice of the representation of molecules is at the heart of the machine learning-based drug discovery [6, 8, 29, 34].

Table 1: Comparison among different types of fingerprint methods, in three different aspects: 1) if the design of the fingerprint requires biologists' expertise knowledge, 2) if the fingerprint has enough information to be reverted to original SMILE representation, and 3) if the fingerprint method requires many labeled data. DL is short for Deep learning while FP is short for FingerPrint.

Properties	Non-data driven Methods		Supervised DL FP [13, 26, 42]	Seq2seq FP (**ours**)
	Hash-based [17, 25, 28, 36]	Local feature[30, 37]		
Without biologist guide	✓	×	✓	✓
Revertible	×	×	×	✓
Less thirsty on label data	✓	✓	×	✓

Traditionally, the design of new fixed-length vector molecular representations, named **fingerprints**, is not data-driven and based on human expertise knowledge [7, 22, 32, 38, 39]. One type of those design is based on some hashing procedure, e.g., Extended Connectivity FingerPrint (ECFP) [36]. Those fingerprints are usually efficient in speed, but is much like a lossy compression in the imaging area [33, 50–52] and the operation is non-invertible. The other sort of non-data-driven fingerprint is based on local sub-structures of molecules. Biologists look for several highly related chemical molecular sub-structures for specific tasks and design the fingerprint feature vector accordingly. Representative works are [30, 37]. However, this kind of design obviously requires years of expertise experience and is highly task-sensitive. To sum up, the non-data-driven fingerprint is either limited in encoding enough information or highly lean to expensive and accurate human knowledge. Hence it has raised a great demand for the data-driven fingerprints, which does not require years of human guide and expensive biological experiments.

Observed the recent success of deep learning on imaging understanding [43, 47] or natural language processing [20, 31], there are a few attempts made in applying deep neural network to generate fingerprints. Among the most famous ones are the neural fingerprint [13] as well as [18, 26, 42, 46]. However, most supervised deep learning methods are data-hungry and usually completely fail when data scale is limited [53, 58], and unfortunately this is usually the case in the drug discovery due to the insane expensiveness of the lab experiment.

In this paper, we propose an unsupervised data-driven deep-learning-based molecular fingerprint method, named **seq2seq fingerprint**. To overcome the issues mentioned above, 1) the proposed method is data-driven, without any human expertise knowledge required. 2) the fingerprints generated by the proposed method are completely revertible to original molecular representations, ensuring the sufficiency of information encoded in the fingerprint vector. 3) the proposed method employs an unsupervised training on a **huge unlabeled** dataset, sufficiently releasing the horsepower of deep neural network. We illustrate a comparison among all mentioned fingerprint methods and our seq2seq fingerprint method.

Our fingerprint is designed based on a recent breakthrough model, called *sequence-to-sequence learning* (seq2seq learning). The seq2seq learning method comes from a seemingly unrelated area, English-to-French translation. The seq2seq learning method takes an English sentence as the input, encodes it into a *meaning* vector and then translates it back to a French sentence as the output. The crux of our method is similar, but differs in the way that we set both the input and output of the seq2seq learning as the same SMILE string, a text representation of a molecule. We map the SMILE string to a fixed-sized vector and then *translates* it back to the original SMILE string. The intermediate fixed-sized vector is extracted as the **seq2seq fingerprint**. Once the model is well-trained, the intermediate feature vector is considered to encode all the information to recover the original molecular representation. Hence, the seq2seq fingerprint is expected to capture the rigorous information with which we can accurately predict the molecular properties.

The benefits of the seq2seq fingerprint are three folds: 1) the training phase of seq2seq fingerprint is completely **label-free**, avoiding the costly and labor-intensive label acquiring procedure. 2) it is data-driven, eliminating the reliance on expert's subjective knowledge. 3) since the unlabeled data is almost unlimited in practice, we can fully utilize the power of deep learning, without suffering from the short supply of labeled data.

The technical contributions of this paper are summarized as: 1) the seq2seq fingerprint method is clearly the first attempt to apply the seq2seq learning method to perform drug discovery tasks, coupling two seemingly unrelated areas. 2) several important adaptations are made into the original seq2seq learning to suit drug discovery applications:

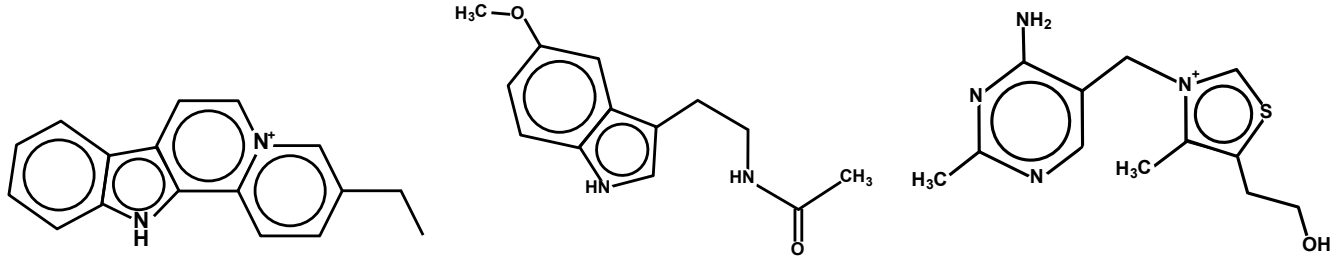
- GRU cell is used, instead of LSTM, to accelerate the training process,
- Attention Mechanism is employed to centralize the fingerprint space,
- Dropout layer is added to overcome the over-fitting issue during the training phase,
- An extra fingerprint extraction layer set is added to pull the fingerprint out.

3) extensive experiments confirm the superior performance on different tasks over the state-of-the-art methods.

The rest of the paper is organized as follows. We summarize several related work, in both drug discovery and sequence to sequence learning, 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 present several related work. First, we introduce the initial representation of molecules, i.e., how the molecular data is persisted in the data store. Second, we list several state-of-the-art



(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.

fingerprint methods, including the most recent ones using deep learning techniques. Last but not least, we briefly describe our cornerstone learning method, i.e., seq2seq learning, with several of its related work in language translation area.

2.1 SMILE Representations of Molecules

Initially, the molecules are represented through the Simplified Molecular-Input Line-Entry system (SMILE) [45], 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. 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. We show some more complex examples in Figure 1.

2.2 Fingerprint Methods

Hash-based Fingerprints. Many hash-based has been developed to generate unique molecular feature representation [17, 25, 28]. One of the most famous ones being Extended-Connectivity FingerPrint (ECFP) [36]. Circular fingerprints generate each layer's features by applying a fixed hash function to the concatenated features of the neighborhood in the previous layer. 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, [30, 37]. Biologist look for several important task-related sub-structures (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 generalize very poorly for other tasks.

Supervised Deep Learning-based Fingerprints. The growth of deep learning has provided the flexibility and performance to create the molecular fingerprint from data samples, without explicit human guide, [2, 13, 26, 40, 42]. The state-of-the-art work is the neural fingerprint [13]. 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 data-hungry deep neural network. To acquire enough labeled data, biologists need to perform a sufficiently large number of tests on chemical molecules, which is extremely expensive.

2.3 Encoder-Decoder Structured Neural Network

Variational Auto-Encoder. Variational Auto-Encoder (VAE) model [12] shares some similar structure as our method, which uses a *encoder* to encode the original representation to a vector or scalar then a *decoder* to decode the vector to original representation. The difference is that the VAE model puts the assumption that the embedded space follows some specific Gaussian distribution. In most recent months, the authors become aware of an unpublished VAE report in drug discovery [19]. However, there is no evidence and experimental results to support the Gaussian assumption on the embedded fingerprint space. Moreover, we still lack the evaluation on how the VAE will perform in the predictive tasks.

Generative Adversarial Network. Generative Adversarial Network (GAN) [21] has recently become popular in the machine learning area. A GAN is constructed by a *discriminator* and a *generator*. The discriminator acts as a cop to distinguish the training data samples from the samples generated from the generator. Hence, the learning process actually learns from both training data set and the generated fake data samples. It works well when the scale of data sample is limited. But such network is hard to train and we are not aware of any publicly available report that documents the attempt to adapt GAN into drug discovery.

Sequence to Sequence Model. The sequence to sequence model [41] has been recently used in English-to-French translation and demonstrated as a breakthrough success. The basic strategy of sequence to sequence learning is to map the input sequence, e.g., an English sentence, to a fixed-sized vector using one deep Long Short-Term Memory (LSTM) network, and then map the vector to the target sequence, e.g., the translated French sentence, with another deep LSTM network. The fixed-sized vector is considered as an intermediate representation and contains the *"meaning"* of sentences.

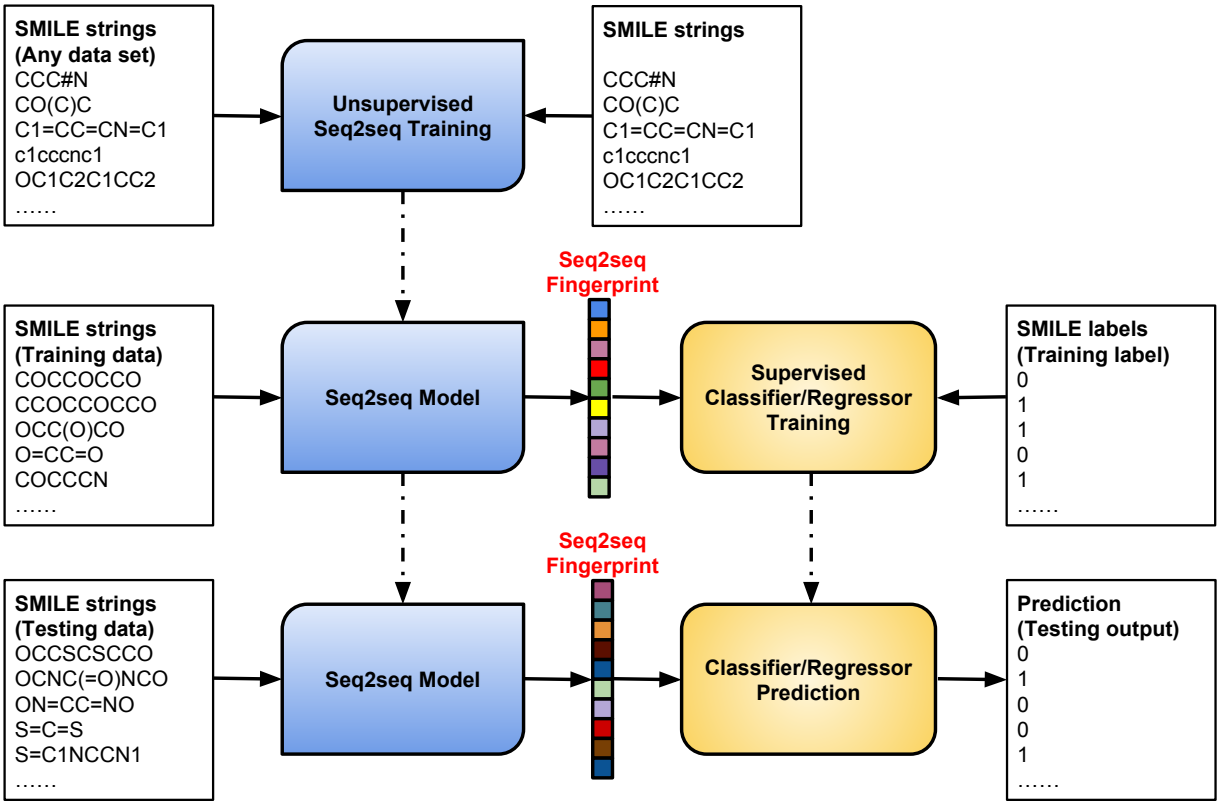


Figure 2: This figure shows how the entire pipeline works. 1) The seq2seq fingerprint model is trained on a large pool of unlabeled SMILE data. 2) The trained model is fed with SMILE strings to generate the seq2seq fingerprint. 3) Coupling the fingerprint and label, the pairs are fed into supervised classifiers/regressors to train a predictive model.

3 METHODOLOGY

In the sequel, we present the entire pipeline of the proposed method. First, we overview the entire pipeline with an introduction of the crux of our fingerprint method. Second, we detail each step of our method and our improvements and adaptations upon the original seq2seq learning method. Last, we discuss our methods to end this section.

3.1 Overview

The entire pipeline of our method consists of three steps: 1) we first train the seq2seq fingerprint model on a **huge** pool of unlabeled training data. 2) Then the trained model is used to generate the seq2seq fingerprint for the labeled data set. 3) The resulting fingerprints and their labels are fed to some supervised learning method to train a predictive model, e.g., Gradient Boosting, Multi-Layer Perceptron (MLP). An illustration of the pipeline is shown in Figure 2. As a result, the entire pipeline is able to **transfer** knowledge from a large number of unlabeled data samples to the supervised training on a relatively small labeled dataset and thus improve the final predictive performance.

The crux of our unsupervised seq2seq fingerprint method, **seq2seq fingerprint**, is to set both input and output sequences to the same SMILE string for each molecule in sequence-to-sequence learning for unsupervised training, or simply **translate the SMILE string to itself**. Since the intermediate vector is considered to maintain the "meaning" of the sequence, we thus extract the intermediate vector as the fingerprint. While the sequence to sequence learning [41] could in principle directly work with our idea, however, there are still many drawbacks yet to limit the application in the molecular predictive tasks.

First, at least in theory, our model can train on a pool of infinite molecular data. While the LSTM is famous for its slow training, the time invested in the training process on a large amount of data is absolutely unbearable. Second, the original sequence to sequence learning does not explicitly output the embedding vector, and therefore lacks an extra layer in order to output the fingerprint vector. Third, as argued in [4], when the length of the input sequence

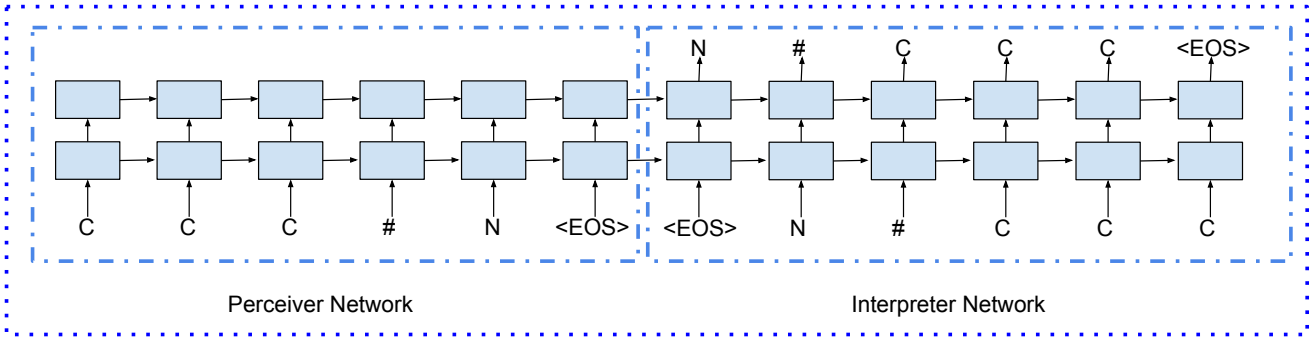


Figure 3: An example on how the unsupervised training works. The perceiver network understands the molecule SMILE representation, e.g., CCC#N, and encodes it into a fixed-length vector, i.e., seq2seq fingerprint. The interpreter network will then translate the fingerprint back to the sequence, e.g. N#CCC.

grows, the performance of the neural network decreases rapidly. However, SMILE representation usually contains several tens of characters (up to 250 characters), which is too long to be handled by the original sequence to sequence model. Finally, due to the large training data scale, the number of model parameters tends to be relatively smaller than demanded, yielding the over-fitting issue.

Here, we propose the seq2seq fingerprint, with various of improvement upon the original sequence-to-sequence learning [41] used in English-to-French translation to generate an effective fingerprint for drug discovery tasks. We detail each step in the following sub-sections.

3.2 Unsupervised Seq2seq Training

To train a fingerprint generator on a huge unlabeled dataset, we first employ a deep Gated Recurrent Unit (GRU) network, named *perceiver network*, to map the original molecular SMILE string to a fixed-sized vector, i.e., the seq2seq fingerprint. Then another deep GRU neural network, called *interpreter network*, is used to generate the original SMILE string back from the seq2seq fingerprint. A workflow illustration of our method is shown in Figure 3. In the following, we show, in the descent order of importance from the authors' perspective, the details we altered to adapt the drug discovery application.

GRU Units The Gated Recurrent Unit (GRU) is used in our experiment instead of LSTM. GRU is famous for its LSTM-like performance but faster training process. A GRU network computes a sequence of outputs $(s_1, \ldots, s_T)$ from the input sequences $(x_1, \ldots, x_T)$ by iterating

$$\begin{aligned} z_t &= \sigma_g(W_z x_t + U_z s_{t-1} + b_z) \\ r_t &= \sigma_r(W_r x_t + U_r s_{t-1} + b_r) \\ h_t &= \tanh(U_h x_t + W_h(s_{t-1} \circ r_t)) \\ s_t &= (1 - z_t) \circ h_{t-1} + z_t \circ s_{t-1}. \end{aligned} \tag{1}$$

A GRU cell has two gates: the update gate z and the reset gate r. Each gate has the trainable parameter W, U, b. The activation function σ for each gate is usually the sigmoid function. GRU also has the *"hidden memory"* h, which holds another set of trainable parameters U, W. In contrast with LSTM which has three gates, it has similar performance but faster training speed [10].

Attention Mechanism So far, the only connection between the perceiver and interpreter networks is the sharing hidden memory. When the sequence becomes longer, it becomes extremely challenging to pass the information from the perceiver to the interpreter network through the hidden memory. To address this issue, the attention mechanism is employed to establish a stronger connection and provide soft-alignment between the perceiver and interpreter networks. More details are referred to [4].

Dropout Layer One of the most favorable features in our model is the capability to use nearly unlimited molecular training data. However, the over-fitting issue will come to play if we grow our data unrestrictively. To enhance the generalizability of our model, we add dropout layer to each input, output gate and yet we do not add the dropout for the hidden memory transferring gate, following the practices in [54].

What do we inherit? While we improve the original sequence-to-sequence model from several aspects, we keep using the reverse technique introduced in [41], where the source sequence is mapped to the reverse sequence of the target. For example, instead of mapping a, b, c to α, β, γ, we map a, b, c to γ, β, α. This trick is observed to make it easier for the Stochastic Gradient Descent (SGD) algorithm to *"establish communication"* between the source and target sequences. Another important technique we keep is the *bucket training*, where all the training sequences are distributed into several buckets, and all the sequences in the same bucket are padded to the same length. This technique can parallel the training process on GPUs for acceleration.

3.3 Fingerprint Extraction

During the fingerprint extraction stage, we only feed-forward the *perceiver* network, leaving the interpreter network behind to save computational resources. Moreover, the original sequence to sequence model does not explicitly output the embedding vector, which brings us challenges to extract fingerprints we need. A fixed unit fully connected layer together with a GRU cell state concatenation layer is injected between the perceiver network and interpreter

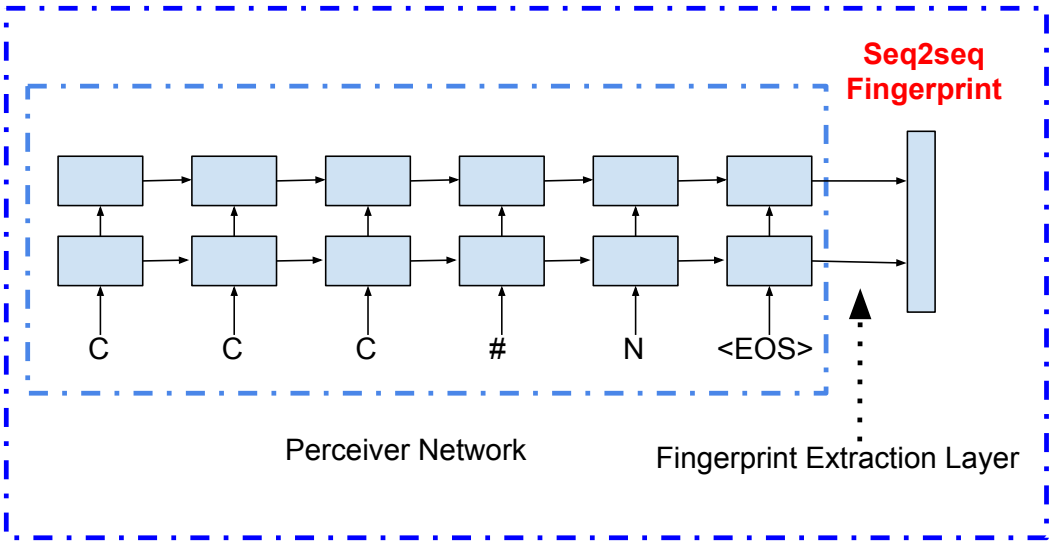


Figure 4: The illustration of how to extract the seq2seq fingerprint. Only the perceiver network is feed-forwarded with an extra fingerprint extraction layer to extract the resulting seq2seq fingerprint.

network to extract the seq2seq fingerprint from the network. The illustration of this process is in Figure 4.

3.4 Supervised Training on Labeled Data

Since our method embedded the molecular graph into a vector space with fixed dimension, the resulting fingerprint can be almost trained with almost all popular regressors or classifiers. Those methods include but not limited to linear Support Vector Machine (SVM), ν-support vector machine, and ensemble methods, e.g., AdaBoost, Extra Trees, etc. In our experiments, we investigate our fingerprints with three ensemble methods: AdaBoost, GradientBoost and Random Forest.

3.5 Discussion

Our method can indeed transfer knowledge from unlabeled data to the labeled data training. However, it is not technically semi-supervised, since the unlabeled data is not directly used in the supervised training. So we still name our fingerprint method as unsupervised.

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 recovery performance of the seq2seq fingerprint method, i.e., how the SMILE self-translation performance is. Finally, we show the superior performance on two predictive tasks for our seq2seq fingerprint method.

4.1 Experiment Setup

Unsupervised Train Dataset Our training data was collected from a combination of two large datasets: LogP and PM2-full datasets. Those datasets were obtained from National Center for Advancing Translational Sciences (NCATS) at National Institutes of Health (NIH). The training dataset contains 334,092 valid molecular SMILE representations.

Labeled Datasets We performed the classification on two smaller datasets:

- **LogP**: LogP dataset contains a total of 10,850 samples. Each sample contains a pair of a SMILE string and a water-octanol partition coefficient (LogP) value. A certain threshold of 1.88 is suggested by an NCATS expert. Samples with LogP value smaller than 1.88 will be classified as the negative samples, while the opposites are considered the positive samples.
- **PM2-10k**: PM2-10k dataset contains 10,000 pairs of SMILE strings and binary promiscuous class labels.

Comparison Methods We compared our seq2seq fingerprint method with two state-of-the-art methods: the ECFP [36] (circular fingerprint) and the neural fingerprint method [13]. The circular fingerprint is a hand-crafted hash-based feature. The neural fingerprint is constructed on a supervised deep graph convolutional neural network. The circular fingerprint was generated through RDKit [1] and we use Multi-Layer Perceptron for the future predictive task as suggested in [13]. We obtained the neural fingerprint from https://github.com/HIPS/neural-fingerprint and we carefully followed the authors' instructions to apply our datasets.

Infrastructure and Software The seq2seq fingerprint method was implemented through tensorflow package [1], and the trained models used in our experiments were trained on a workstation with Intel i7 6700K @ 4.00 GHz CPU, 16 Gigabytes RAM and a Nvidia GTX 1080 GPU. We performed the hyper-parameter grid search and the training process of the classifiers on the TACC

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

Table 2: The comparison of classification accuracy on the LogP data.

			Adaboost (Ours)			GradientBoost (Ours)			RandomForest (Ours)		
	Circular	Neural	512	768	1024	512	768	1024	512	768	1024
Mean	0.3674	0.6080	0.7044	0.6837	0.7342	0.7350	0.7149	**0.7664**	0.6895	0.6664	0.6845
StDev	0.0074	0.0135	0.0042	0.0097	0.0042	0.0060	0.0058	0.0043	0.0061	0.0100	0.0032

Table 3: The comparison of classification accuracy on the PM2-10k data.

			Adaboost (Ours)			GradientBoost (Ours)			RandomForest (Ours)		
	Circular	Neural	512	768	1024	512	768	1024	512	768	1024
Mean	0.3938	0.5227	0.5535	0.5561	0.6036	0.5741	0.5713	**0.6206**	0.5316	0.5282	0.5481
StDev	0.0114	0.0112	0.0132	0.0070	0.0147	0.0086	0.0151	0.0198	0.0110	0.0081	0.0088

Lonestar 5 cluster [2]. In addition to the traditional MPI package for distributed grid-search, we used a more flexible master-worker task distribution package for Python called *dgsearch*. The code for training the seq2seq fingerprint will become publicly available after the acceptance of this paper.

4.2 Seq2seq Fingerprint Recovery Performance

Throughout the entire experiment sections, we discuss three variants of our seq2seq fingerprint, varying in feature vector lengths as 512, 768, and 1024. Each model was trained for 24 hours. These three models differ only in the number of GRU layers and yet with the same Latent Dimension (LD). We report the training details and the recovery power of each fingerprint model in Table 4. The recovery performance is evaluated through the perplexity and Exact Match (EM) accuracy. The perplexity is calculated by the entropy of the probability distribution over the training set. The EM accuracy is the ratio between the exactly recovered SMILE strings and the total number of SMILEs in the test sets.

Table 4: The reconstruction performance with different number of GRU layers.

Model	Layer	LD	Perplexity	EM Accuracy
seq2seq-512	2	256	1.00897	94.24%
seq2seq-768	3	256	1.00949	92.92%
seq2seq-1024	4	256	1.01472	90.26%

Table 4 reveals a decreasing trend of recovery performance when we increase the layer number of stacked GRU cells. One might expect a deeper GRU network to have a better EM accuracy, which contrasts with the observation. The reason might be complex. First, the training of longer seq2seq fingerprint might take longer time to have better performance. Also, increasing the length of fingerprint actually expands the representation space of molecules, leaving more null spaces in the fingerprints. However, this observation does not indicate a longer fingerprint will decrease the performance in other tasks, as shown in the next subsection.

[2]https://www.tacc.utexas.edu/systems/lonestar

4.3 LogP Solubility and PM2 Promiscuous Classification

In this section, we report the classification performance of all three seq2seq models with fingerprint lengths 512, 768, and 1024, compared with the circular fingerprint [36] and neural fingerprint [13]. We use three ensemble classifiers for our seq2seq fingerprints: Adaboost [15], GradientBoost [16], and RandomForest [24]. We report the accuracy means and standard deviations of 5-fold classification cross validation on both LogP and PM2-10k data, in Table 2 and 3 respectively. All results are the 100-run averages to reduce the randomness. We also show the impact of seq2seq fingerprint length on the accuracy in Figure 5.

From the Table 2 and 3, we observe, on both data sets, our methods significantly outperform the circular and neural fingerprints, regardless of classifiers and fingerprint lengths. The circular fingerprint is hashing-based and abandons a large portion of information and is not invertible to original molecule, while our fingerprint is completely invertible and encodes rigorous information. One might argue if the ensemble classifiers will improve the performance of circular fingerprint. According to [13] and our preliminary experimental observation, the results will be worse if we switch the MLP classifier to ensemble classifiers, e.g., GradientBoost, due to the limited length and information of circular fingerprint. The neural fingerprint is a supervised deep learning-based algorithm, and it could be highly limited by the amount of labeled training data. While our method transfers knowledge from a fairly large amount of unlabeled data, our method could outperform the neural fingerprint method in classification tasks. Overall, our seq2seq fingerprints encode rigorous information for molecules and could train on a huge amount of data to achieve task-insensitive performance.

In Figure 5, however, despite the lower recovery performance of seq2seq-1024 fingerprint, it does always provide the best classification performance, while, surprisingly, the seq2seq-768 seems to always have lower classification performance. The longer fingerprints might have more information for ensemble classification methods, but might also bring in noise. While the noise takes the major effects, the performance might decrease. But when the information is encoded enough, the performance will boost.

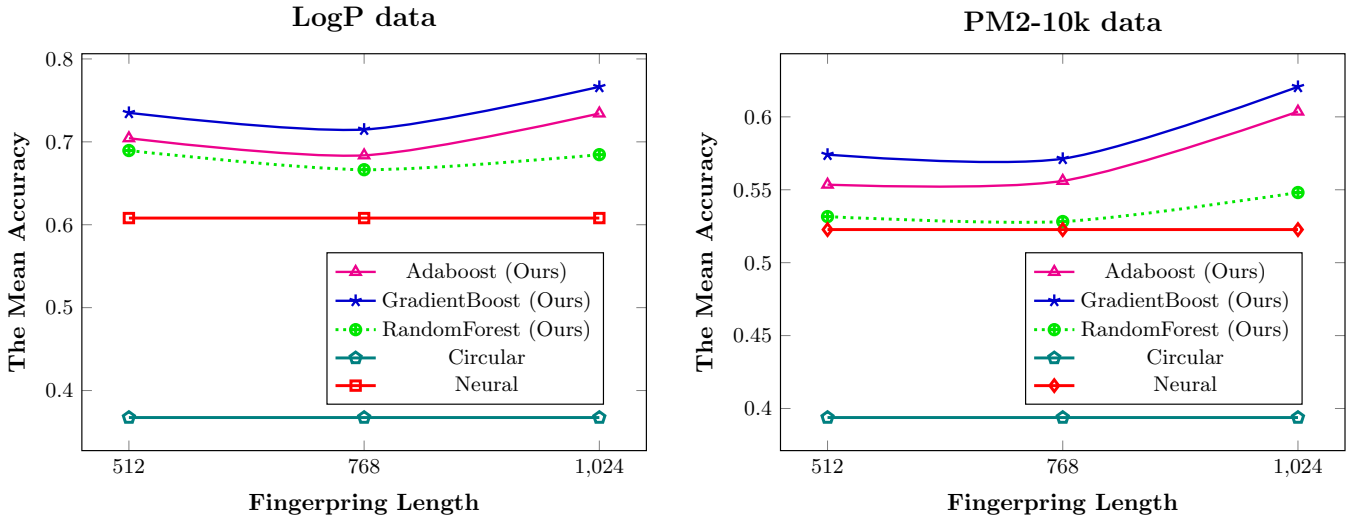


Figure 5: The mean accuracy of five methods. AdaBoost, GradientBoosting and RandomForest our methods. The circular and neural fingerprint are the state-of-the-art methods.

5 CONCLUSIONS

In this paper, we discuss a new unsupervised molecular representation system, called **seq2seq fingerprint**, based on the idea from the recent breakthrough on the English-to-French language translation, named sequence to sequence learning model. Our model translates the molecular SMILE string to the SMILE itself, while at the same time generates a fixed length fingerprint vector. The experiments on classification task demonstrate its superior performance. Also, the nature of our data-driven label-free model brings us even more benefits. 1) Our fingerprint system is completely unsupervised, meaning it will never be limited by the expensive label collection process. In fact, it could utilize each of every valid molecule, theoretically reaching the amount of infinite. 2) Contrast to the supervised learning models trained with very limited data samples, the seq2seq fingerprint is trained from a sufficiently large pool of samples, and therefore it is more robust to the specific task.

This seq2seq fingerprint is definitely not the end. It widely opens tons of new possibilities. Also due to the long training time, we might introduce efficient distributed training strategy [1, 48]. There are still many hyperparameters in our training algorithms, in the future, we might want to pick an optimal method for hyperparameter tuning [49]. Another quick future work would lie on how to embed some label information [9] to the fingerprint training to enhance its performance on the future machine learning task. Those type of semi-supervised learning could be a trend in the future drug discover tasks.

Acknowledgment The authors would like to thank NVIDIA for GPU donation and the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial. The authors would also like to thank Zhongxing Peng and Texas Advanced Computing Center (TACC) for generously providing computational resource for the preliminary experiments. 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. (2015). http://download.tensorflow.org/paper/whitepaper2015.pdf
[2] 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).
[3] Mark Ashton, John Barnard, Florence Casset, Michael Charlton, Geoffrey Downs, Dominique Gorse, John Holliday, Roger Lahana, and Peter Willett. 2002. Identification of Diverse Database Subsets using Property-Based and Fragment-Based Molecular Descriptions. *Molecular Informatics* 21, 6 (2002), 598–604.
[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
[5] Guy W Bemis and Mark A Murcko. 1996. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry* 39, 15 (1996), 2887–2893.
[6] Jeffrey M Blaney and J Scott Dixon. 2007. Distance geometry in molecular modeling. *Reviews in Computational Chemistry, Volume 5* (2007), 299–335.
[7] Frank R Burden. 1989. Molecular identification number for substructure searches. *Journal of Chemical Information and Computer Sciences* 29, 3 (1989), 225–227.
[8] Raymond E Carhart, Dennis H Smith, and R Venkataraghavan. 1985. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences* 25, 2 (1985), 64–73.
[9] Guangliang Cheng, Feiyun Zhu, Shiming Xiang, Ying Wang, and Chunhong Pan. 2016. Semisupervised hyperspectral image classification via discriminant analysis and robust regression. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9, 2 (2016), 595–608.
[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
[11] Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. 2008. On the Art of Compiling and Using'Drug-Like'Chemical Fragment Spaces. *ChemMedChem* 3, 10 (2008), 1503–1507.
[12] Carl Doersch. 2016. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908* (2016).
[13] 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.
[14] David Edwards. 2004. Accelerating Drug Development with Precision Dosing Techniques. In *PMPS*.
[15] 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.
[16] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[17] 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.

[18] 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).

[19] Rafael Gómez-Bombarelli, David Duvenaud, José Miguel Hernández-Lobato, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. 2016. Automatic chemical design using a data-driven continuous representation of molecules. *arXiv preprint arXiv:1610.02415* (2016).

[20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[22] Osman F Güner. 2000. *Pharmacophore perception, development, and use in drug design*. Vol. 2. Internat'l University Line.

[23] Thomas A Halgren. 1996. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *Journal of Computational Chemistry* 17, 5-6 (1996), 553–586.

[24] 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.

[25] 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.

[26] 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.

[27] Xiao Qing Lewell, Duncan B Judd, Stephen P Watson, and Michael M Hann. 1998. Recap retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of chemical information and computer sciences* 38, 3 (1998), 511–522.

[28] 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.

[29] Ramaswamy Nilakantan, Norman Bauman, J Scott Dixon, and R Venkataraghavan. 1987. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *Journal of Chemical Information and Computer Sciences* 27, 2 (1987), 82–85.

[30] 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.

[31] 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.

[32] Robert S Pearlman and KM Smith. 2002. Novel software tools for chemical diversity. In *3D QSAR in drug design*. Springer, 339–353.

[33] 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.

[34] Anthony K Rappé, Carla J Casewit, KS Colwell, WA Goddard Iii, and WM Skiff. 1992. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American chemical society* 114, 25 (1992), 10024–10035.

[35] Sereina Riniker and Gregory A Landrum. 2013. Similarity maps-a visualization strategy for molecular fingerprints and machine-learning methods. *Journal of cheminformatics* 5, 1 (2013), 43.

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

[37] 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.

[38] Gisbert Schneider, Odile Clément-Chomienne, Laurence Hilfiger, Petra Schneider, Stefan Kirsch, Hans-Joachim Böhm, and Werner Neidhart. 2000. Virtual screening for bioactive molecules by evolutionary de novo design. *Angewandte Chemie International Edition* 39, 22 (2000), 4130–4133.

[39] Gisbert Schneider, Man-Ling Lee, Martin Stahl, and Petra Schneider. 2000. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *Journal of computer-aided molecular design* 14, 5 (2000), 487–494.

[40] 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.

[41] 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.

[42] 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).

[43] 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.

[44] Lutz Weber, Sabine Wallbaum, Clemens Broger, and Klaus Gubernator. 1995. Optimization of the biological activity of combinatorial compound libraries by a genetic algorithm. *Angewandte Chemie International Edition in English* 34, 20 (1995), 2280–2282.

[45] 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.

[46] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2017. MoleculeNet: A Benchmark for Molecular Machine Learning. *arXiv preprint arXiv:1703.00564* (2017).

[47] 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.

[48] 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.

[49] Zheng Xu and Junzhou Huang. 2017. A General Efficient Hyperparameter-Free Algorithm for Convolutional Sparse Learning.. In *AAAI*. 2803–2809.

[50] 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.

[51] Zheng Xu, Yeqing Li, and Junzhou Huang. 2016. Accelerated sparse optimization for missing data completion. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 1267–1272.

[52] Jiawen Yao, Zheng Xu, Xiaolei Huang, and Junzhou Huang. 2015. Accelerated dynamic MRI reconstruction with total variation and nuclear norm regularization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 635–642.

[53] Jiawen Yao, Xinliang Zhu, Feiyun Zhu, and Junzhou Huang. 2017. Deep Correlational Learning for Survival Prediction from Multi-modality Data. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

[54] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329* (2014).

[55] Feiyun Zhu, Bin Fan, Xinliang Zhu, Ying Wang, Shiming Xiang, and Chunhong Pan. 2015. 10,000+ Times Accelerated Robust Subset Selection. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[56] Feiyun Zhu, Ying Wang, Bin Fan, Shiming Xiang, Geofeng Meng, and Chunhong Pan. 2014. Spectral unmixing via data-guided sparsity. *IEEE Transactions on Image Processing* 23, 12 (2014), 5412–5427.

[57] Feiyun Zhu, Ying Wang, Shiming Xiang, Bin Fan, and Chunhong Pan. 2014. Structured sparse method for hyperspectral unmixing. *ISPRS Journal of Photogrammetry and Remote Sensing* 88 (2014), 101–118.

[58] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. 2017. WSISA: Making Survival Prediction from Whole Slide Pathology Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

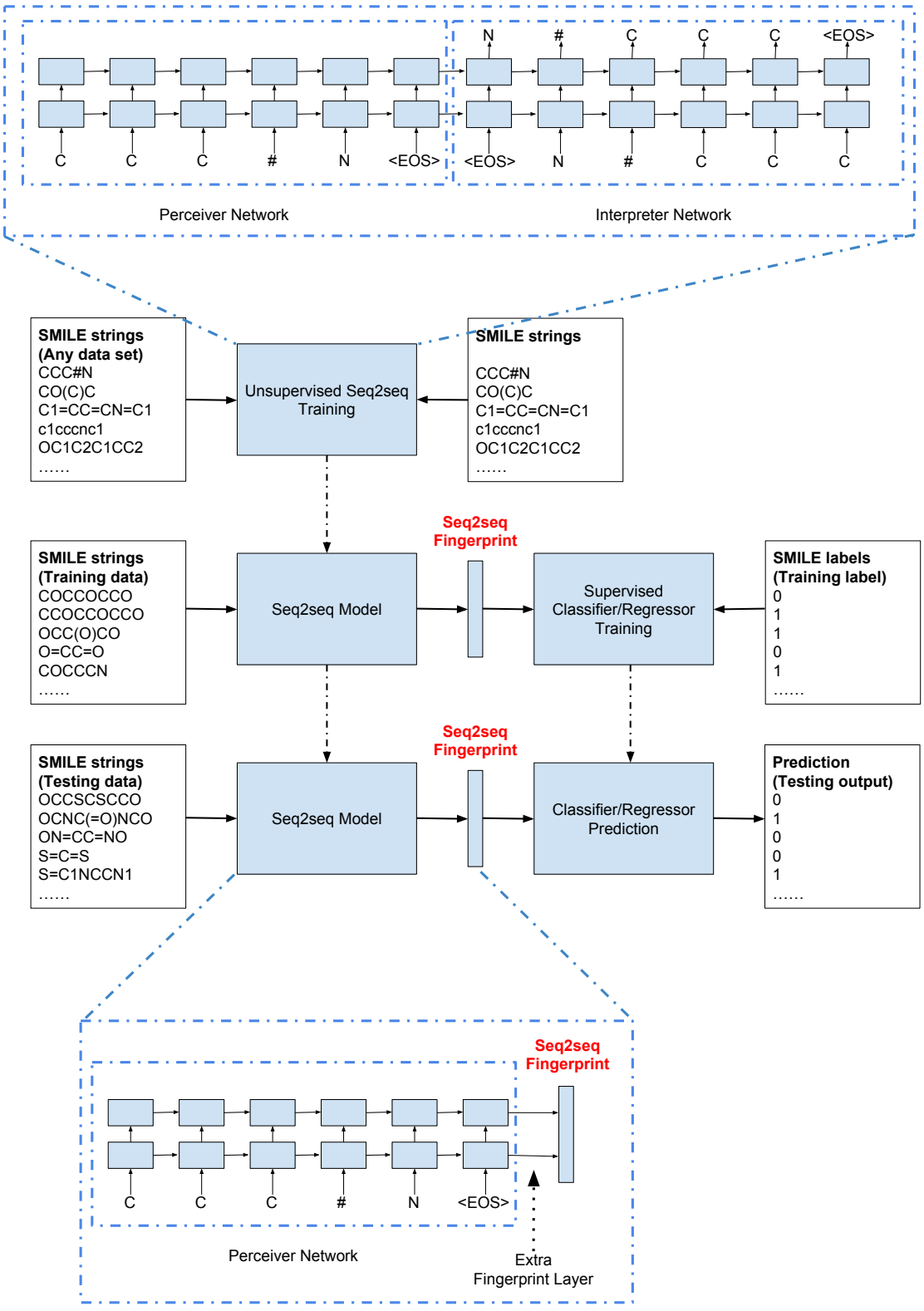


Figure 6: The illustration of overall pipeline for seq2seq fingerprint.