Chemical Language Modelling

اعداد طالبة الماجستير

لبنى ليث طه

Introduction

- Language models are autoregressive models for sequence generation that have shown impressive progress recently in natural language understanding using deep neural networks
- These advancements are driven by architecture improvements like the Transformer; a powerful neural network for sequential data that uses self attention. Transformers have found use in many significant scientific applications like protein structure prediction and design and various tasks in cheminformatics.

Introduction

- The foundation of chemical language modelling lies in the representation of chemical structures and reactions.
- This field is driven by the imperative to equip computers with the ability to interpret, generate, and innovate within the realm of chemical information.

Data Used in CLM

1- The use molecular graphs and make use of geometric deep learning to learn representations directly on atoms and bonds

2- The use SMILES (Simplified molecular input line entry specification) string representations that linearize molecular graphs into strings.

3- XYZ files, Crystallographic Information files (CIFs), or Protein Data Bank files (PDBs).

Models and Methodologies

Architectures of Generative Neural Networks:

- **(A)** Recurrent neural network: unfolds over time which is specialized for processing a sequence of input vectors From an initial hidden state vector.
- **(B)** Variational autoencoder: devised to learn a probabilistic generative model as well as its posterior, respectively known as decoder and encoder.
- **(C)** Generative adversarial network: It is an implicit generative model in the sense that it allows for inference of model parameters without requiring one to specify a likelihood
- **(D)** Adversarial autoencoder: is proposed as a standard AE regularized by an adversarial learning (AL) procedure

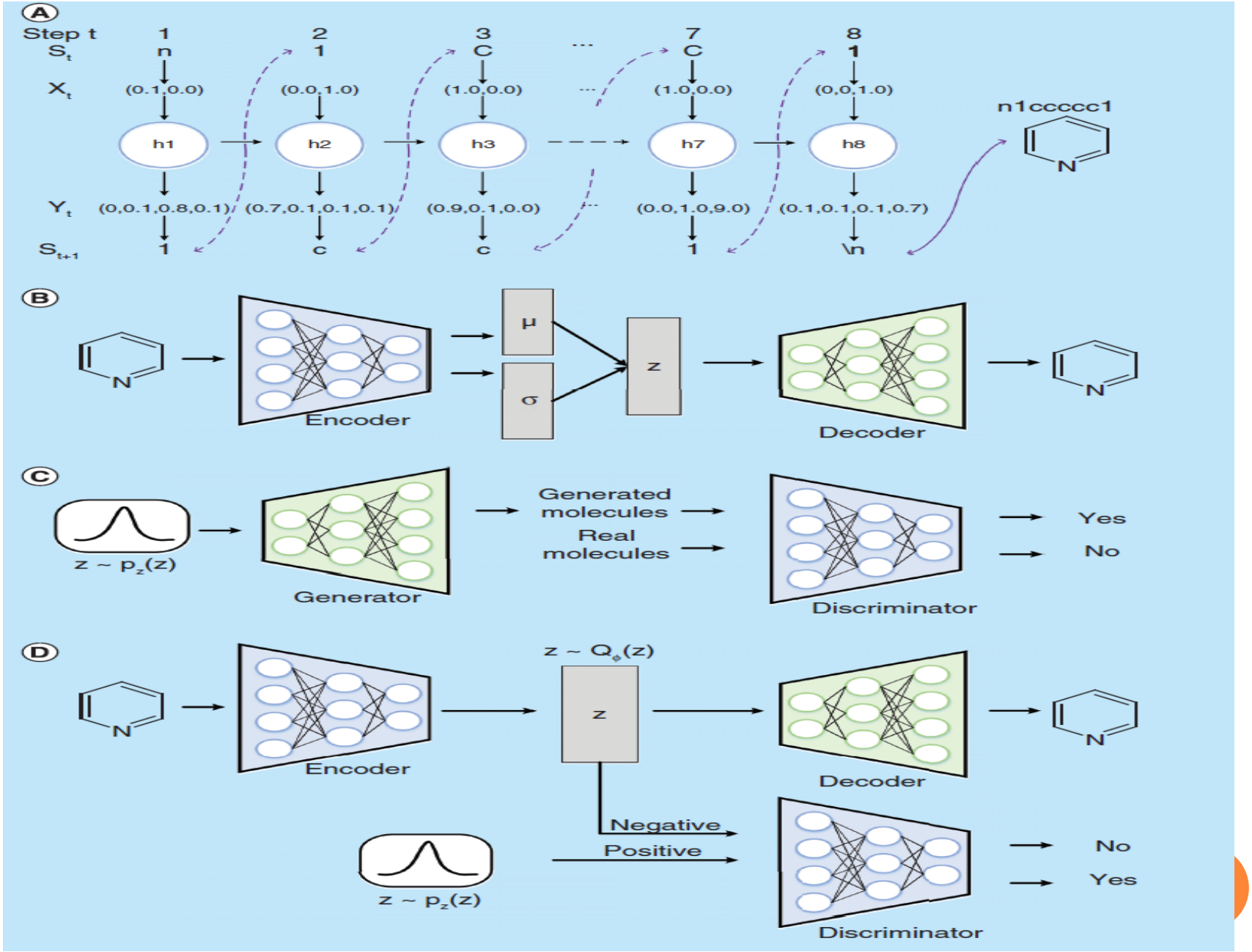

Figure 1: architectures of generative neural networks

Models and Methodologies

Optimization of Generated Molecules using:

A. **Transfer learning** is a simple and popular optimization strategy depending on fine-tuning a general model using a small dataset with certain specific property. It is often applied to recurrent neural network-only-based models.
B. **Bayesian optimization** is a sequential optimization strategy based on a smooth latent space only constructed by auto encoder models to find the best molecules with certain optimal property.

Models and Methodologies

C. **Reinforcement learning** is a preferable optimization strategy via jointly training, not only preserving the generative ability, but maximizing the reward objective about certain property. Generative adversarial network and recurrent neural network-based generative models often use this method for biased drug design.

D. **Conditional generation** is a spring-up optimization strategy. During the optimization process, one or more objectives are converted into one or more conditional codes, which are concatenated to the original input vector for guiding biased molecular generation. It can be applied to multi objective optimization, and may be suitable for the four generative models with ingenious computational design.

Models and Methodologies

- Deep learning models form the backbone of chemical language modeling.
- Recurrent Neural Networks (RNNs) and Transformer models, known for their ability to capture sequential dependencies and long-range dependencies, play a crucial role.
- These models are trained on extensive datasets containing diverse chemical information, ranging from molecular structures to reaction pathways.

Models and Methodologies

- The training process involves exposing the models to a rich corpus of chemical data, allowing them to learn the intricacies of chemical language. This includes understanding the relationships between atoms, predicting molecular properties, and deciphering the nuances of chemical reactions. The result is a model capable of not only comprehending chemical information but also generating novel molecular structures or predicting outcomes of chemical transformations.

The Process of CLM

1- Data Collection: Gather a diverse dataset of chemical structures, reactions, or properties to train the language model.

2- Data Preprocessing: Clean and preprocess the chemical data, which may involve tokenization, normalization, and encoding for effective input into the model.

3- Model Architecture: Choose or design a suitable neural network architecture for chemical language modeling. Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), or Transformer architectures are commonly used.

4- Training: Train the language model using the preprocessed chemical data. This involves feeding input sequences and adjusting the model parameters based on the predicted outputs to minimize a defined loss function.

The Process of CLM

5- Evaluation: Assess the performance of the trained model using separate validation datasets to ensure it generalizes well to new, unseen data.

6- Fine-Tuning: Adjust the model parameters or architecture based on the evaluation results to improve performance.

7- Deployment: Once satisfied with the model's performance, deploy it for chemical language-related tasks, such as predicting molecular properties, generating novel compounds, or assisting in drug discovery.

Bata Banks Used in CLM

Some of the data banks used are:

- PDB
- ZINC
- ChEMBL

Programs Used in CLM

- like RDKit, Open Babel, and Cheminformatics Toolkit
- TensorFlow, PyTorch, and Keras.
- Python

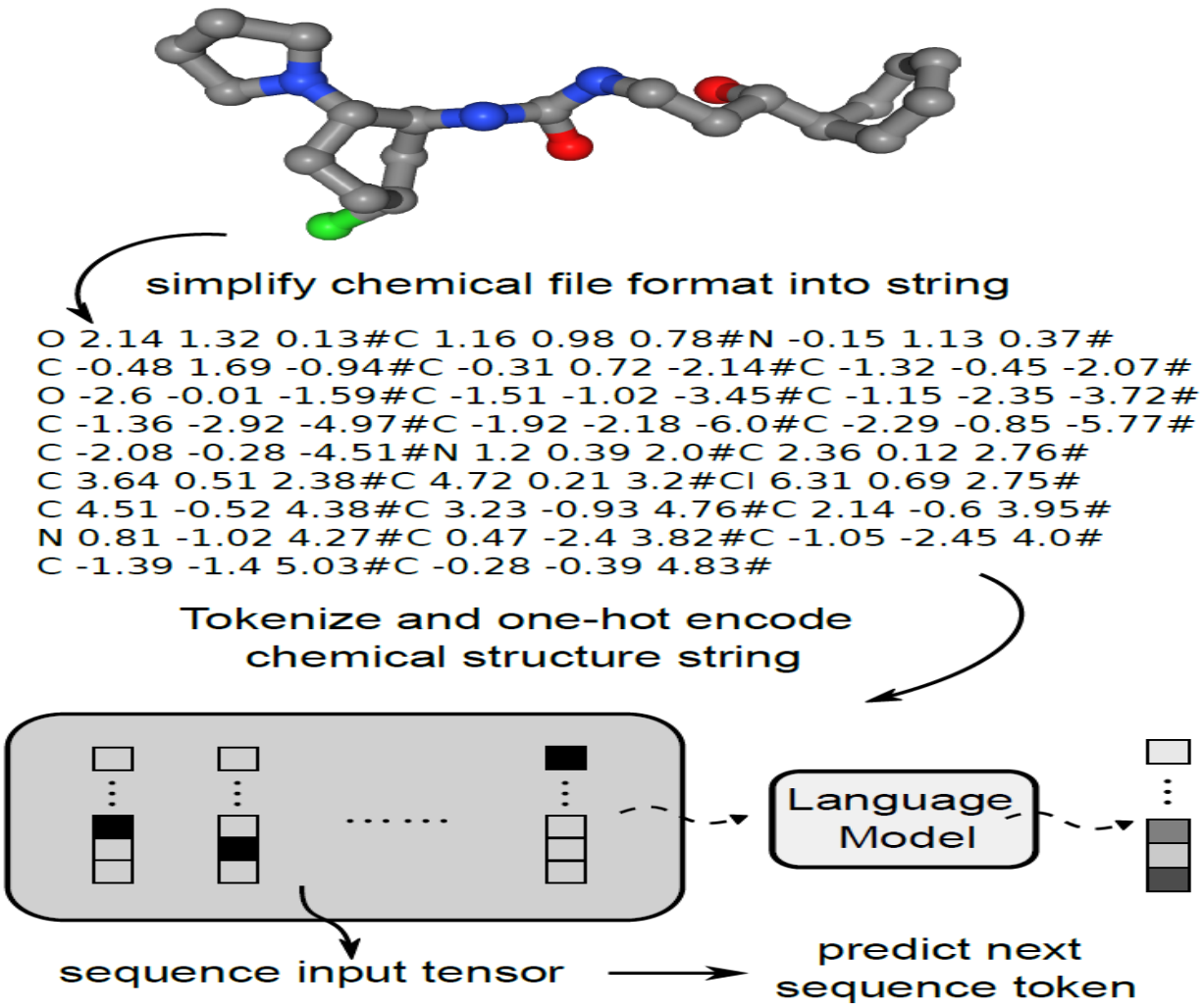

Figure 2: Overview of the training workflow

Applications and Future Prospects

- The applications of chemical language modeling are expansive.
- An important scientific objective is the exploration of chemical space, in order to discover new drugs and materials, these models assist in predicting molecular properties, identifying potential drug candidates, and optimizing chemical structures for enhanced bioactivity.
- In the automation of literature analysis, allowing researchers to sift through vast repositories of chemical information efficiently. This not only saves time but also enables researchers to focus on more complex tasks.

Example of CLM

- LSTM-based RNN model with different applications was developed. Based on 541,555 molecules from the ChEMBL dataset . They sampled 30,107 strings. Considering TL, three specific ligand subsets: 367 peroxisome proliferator-activated receptor γ (PPARγ) inhibitors; 1490 trypsin inhibitors, were used to fine-tune the pretrained model, respectively. For the PPARγ, among a set of 1000 generated samples, 96% were chemically valid. For the 1000 generated samples toward the trypsin inhibitors, 93% were validThey also proposed an expanded application of RNN-based generative models to fragment growing. Given a key fragment binding to thrombin, a library of molecules could be generated instantly without extensive similarity searching or external scoring.


THANK
YOU