ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

РАЗРАБОТКА ГЕНЕРАТИВНОЙ НЕЙРОННОЙ СЕТИ ДЛЯ ГЕНЕРАЦИИ ПОТЕНЦИАЛЬНЫХ МУЛЬТИКИНАЗНЫХ ИНГИБИТОРОВ С ЗАДАННЫМИ СВОЙСТВАМИ

Т. Д. Войтко

Белорусский государственный университет, г. Минск;
timvaitko@gmail.com;
науч. рук. – А.Д. Карпенко, ст. преп.

Работа посвящена разработке генеративной нейронной сети для генерации мультикиназных ингибиторов. Мишенями являются нативная и мутантные киназы, в качестве основной архитектуры нейронной сети был выбран гетероэнкодер. При помощи полученной архитектуры были сгенерированы 1083 уникальные молекулы, некоторые из них по результатам докинга сравнимы с уже существующими ингибиторами по разным мишеням.

Ключевые слова: машинное обучение; нейронные сети; гетероэнкодеры; LSTM сети; полносвязные сети; ингибиторы нативной киназы; ингибиторы мутантной киназы.

Одной из нововведенных архитектур в докладах исследователей по машинному обучению и дизайну лекарств выступает модель гетероэнкодера, которая является усовершенствованием автоэнкодера. На данном этапе, основной задачей является построение самой архитектуры гетероэнкодера. Основное отличие гетероэнкодера от автоэнкодера есть способность обрабатывать сразу несколько представлений молекулы. В исследованиях гетероэнкодеров [3] рассматривают различные архитектуры автоэнкодеров, объединенные в гетероэнкодеры, экспериментируя с различными типами данных.

Также в задачу входит генерация молекул с заданными свойствами при помощи полученной генеративной модели: требуется получить ингибитор мультикиназа, который будет устойчив к множеству мутаций заболеваний лейкемии, будет иметь достаточно маленькую энергию связывания, будет содержать в своей структуре 2-ариламинопиримидин.

Принципиальное отличие нашей работы от уже существующих [1, 3] состоит в архитектуре гетероэнкодера, а также в выбранной функции потерь для обучения этого гетероэнкодера.

Концепт нашей работы заключается в следующем: для входных данных мы хотим получить их представление в латентном пространстве, то есть получить эмбеддинги, а далее, основываясь на этих эмбеддингах

и задаваемой энергии связывания, мы хотим восстановить наши данные, а точнее восстановить строку формата SMILES и строку формата канонического SMILES.

Передавать энергию связывания будем декодирующему слою для достижения следующей цели: как известно, чем меньше энергия связывания, тем крепче связь ингибитор – белок, поэтому требуется разработать архитектуру, которой можно будет передавать желаемую энергию связывания для генерации. Для нас это значения от -14 до -10.

Для реализации данной задачи была разработана архитектура гетероэнкодера с тремя энкодерами и двумя декодерами, на рисунке 1 обзорно представлена архитектура модели. Энкодерами для строковых форматов была выбрана архитектура, состоящая из двух слоев LSTM, состояния из которых становятся эмбеддингами и попадают на объединяющий слой.

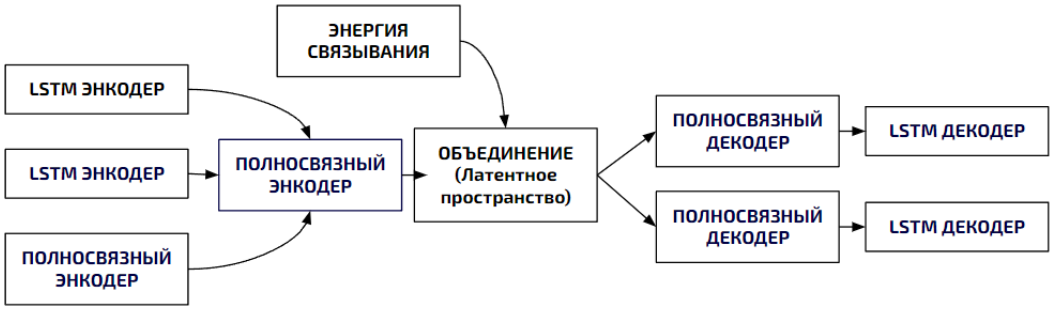

Рис. 1. Архитектура разработанного гетероэнкодера

Числовые характеристики обрабатываются полносвязной нейронной сетью прямого распространения, состоящей из двух полносвязных слоев, слоя батч-нормализации и еще одного полносвязного слоя, результаты работы последнего являются эмбеддингами для числовых признаков. Все полученные выше эмбеддинги попадают на конкатенирующий слой, где образуют один вектор. После этот вектор нормализуется на слое батч-нормализации и передается на полносвязный слой, после чего к ним присоединяется желаемое значение энергии. Результаты работы этого слоя, то есть обработанные эмбеддинги и энергия, образуют латентное пространство.

Далее, полученные элементы латентного пространства подаются на декодеры. Они попадают на два несвязных полносвязных слоя, выходы которых нормализуются и подаются в качестве состояний на слой LSTM. В качестве входных данных для LSTM слоя были представлены те же данные, что и ожидаются на выходе из нейронной сети. Эта идея была реализована с целью улучшить работоспособность декодера и

быстрее обучить модель. Выходы из LSTM слоев подаются на полносвязные слои с функцией активации softmax.

Программную реализацию данной архитектуры осуществим при помощи библиотеки Keras, на основе выкладок из [2].

Веса модели обучим методом минимизации функции потерь. Наша функция потерь будет состоять из двух частей. Первая часть есть категориальная кросс-энтропия (*CCE*). Также, дополнительно введем функцию, которая будет учитывать специфику наших данных. Назовем ее CustomChemLoss (*CCL*). Данная функция будет штрафовать за не содержание в декодированной молекуле 2-ариламинопиримидина и также будет штрафовать в случае если декодированная формула не является валидной. *CCL* равен 0, если молекула валидна и содержит 2-ариламинопиримидин, равен 1, если молекула валидна, но не содержит вышеупомянутое вещество и равен 5, если молекула невалидна.

Таким образом, формула ошибки для нашей модели:

$$Loss = CCE + \alpha\, CCL\ ,$$

где α – коэффициент значимости химической ошибки, является гиперпараметром. В нашей работе мы выбрали значение $\alpha = 0.1$.

Тренировочными данными для нейросети являются 108410 уникальных молекул, представленные в четырех типах. Валидационные данные представлены в аналогичном формате, и их 27102 уникальных молекул. В качестве оптимизатора был выбран ADAM.

После обучения из валидационной выборки были выбраны 1000 лучших по энергии связывания молекул, на основе их мы получили представление латентного пространства при помощи модели. Добавляя к этим векторам гауссовский шум, мы посимвольно сгенерировали новые соединения при помощи декодеров. Так получили 1083 новые молекулы, которые предположительно являются ингибиторами мультикиназа. Они были проверены на валидность, интерпретируемость и содержание 2-ариламинопиримидина при помощи библиотеки rdkit. По итогам молекулярного докинга получены энергии связывания по двум мишеням, гистограммы которых представлены на рисунке 2. Левая гистограмма – энергия связывания по нативной киназе, правая – по мутантной киназе.

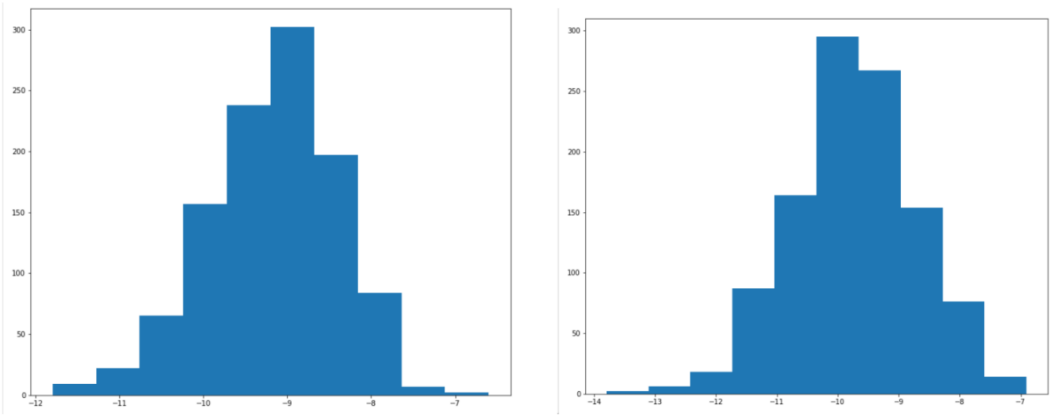

Рис. 2. Распределения энергий связывания

По данным распределениям можно говорить об предварительной успешности нашей работы: большое количество молекул находится в интересующем нас диапазоне, что позволяет перейти к дальнейшим исследованиям, таким как молекулярная динамика, сравнение с существующими ингибиторами по различным мишеням и т.д.

Библиографические ссылки

1. A De Novo Molecular Generation Method Using Latent Vector Based Generative Adversarial Network [Электрон.ресурс.] – https://www.researchgate.net/publication/333912936_A_De_Novo_Molecular_Generation_Method_Using_Latent_Vector_Based_Generative_Adversarial_Network
2. Rowel Atienza, Advanced Deep Learning with Keras/Rowel Atienza. – Published by Packt Publishing Ltd. Livery Place 35 Livery Street Birmingham B3 2PB, UK. – 2018. – 369p.
3. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders [Электрон.ресурс.] – https://www.researchgate.net/publication/328632724_Improving_Chemical_Autoencoder_Latent_Space_and_Molecular_De_Novo_Generation_Diversity_with_Heteroencoders.