Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Факультет математики и компьютерных наук имени профессора Н.И. Червякова
Кафедра математического моделирования

Утверждена распоряжением по факультету от 18.03.2025 № 127-р-25.00

Допущена к защите
«____» _____ 20____ г.

Зав. кафедрой математического моделирования, к.ф.-м.н., доцент Ляхов П.А.

_____

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧ ХИМИИ

Рецензент:
Самойленко Ирина Владимировна,
кандидат технических наук, доцент, доцент кафедры информационных систем СтГАУ

Выполнил:
Бобров Анатолий Анатольевич
студент 2 курса,
группы ПМИ-м-о-23-1
направления подготовки 01.04.02 Прикладная математика и информатика
очной формы обучения

_____

Нормоконтролер:
Ляхов Павел Алексеевич
кандидат физико-математических наук, заведующий кафедрой математического моделирования факультета математики и компьютерных наук имени профессора Н.И. Червякова

_____

Научный руководитель:
Ляхов Павел Алексеевич
кандидат физико-математических наук, заведующий кафедрой математического моделирования факультета математики и компьютерных наук имени профессора Н.И. Червякова

_____

Дата защиты
«____» _____ 2025 г.

Оценка _____

Ставрополь, 2025

Набор данных ESOL является широко используемым эталонным набором данных в хемоинформатике и машинном обучении для прогнозирования растворимости молекул в воде. Он был представлен статье 2004 года [15] под названием «ESOL: оценка растворимости в воде непосредственно по молекулярной структуре». Набор данных был разработан для обучения и оценки моделей, которые предсказывают растворимость в воде (логарифмическая растворимость в моль/л). Набор содержит 1128 соединений с экспериментально измеренными значениями растворимости.

Вторым набором данных будет «freesolv». Этот набор данных содержит информацию о свободной энергии гидратации химических соединений. **Свободная энергия сольватации** — это изменение свободной энергии Гиббса при переходе молекул вещества из вакуума в растворитель.

Для начала работы нам понадобится подключение следующих библиотек:

Библиотека «csv» — это встроенный модуль, который предоставляет функциональные возможности для чтения из файлов формата CSV (значения, разделенные запятыми) и записи в них. Это упрощает работу с табличными данными за счет автоматической обработки синтаксического анализа и форматирования.

«json» — это встроенный модуль, предоставляющий функции для работы с данными в формате JSON (JavaScript Object Notation). Она позволяет кодировать объекты Python в строки JSON и декодировать строку JSON обратно в объекты Python.

«time» — это встроенный модуль, который предоставляет функции для операций, связанных со временем, включая измерение временных интервалов, задержку выполнения и преобразование между форматами времени. Она обычно используется для сравнения производительности.

«itertools» — это встроенная библиотека, предоставляющая эффективные, экономящие память инструменты для работы с итераторами.

«RDKit» — это программное обеспечение для хемоинформатики и машинного обучения с открытым исходным кодом, предназначенное для работы с молекулярными данными. Оно предоставляет инструменты для манипулирования,

анализа и визуализации химических структур, что делает его незаменимым в разработке лекарств, материаловедении и вычислительной химии.

«PyTorch» — это платформа машинного обучения с открытым исходным кодом. Она предназначена для гибкого и быстрого создания и обучения моделей глубокого обучения, использующих ускорение графическим процессором.

«NumPy» — это базовая библиотека для научных вычислений на Python. Она обеспечивает поддержку больших многомерных массивов, а также обширную коллекцию высокоуровневых математических функций для работы с ними.

«Scikit-learn» — это библиотека машинного обучения с открытым исходным кодом для Python, предназначенная для предоставления простых и эффективных инструментов для интеллектуального анализа данных.

Для контроля генераторов случайных величин нужно использовать следующий код:

```
numpy.random.seed(0)
torch.manual_seed(0)
```

После настройки генераторов случайных величин можно приступать к чтению данных. Файл «freesolv.csv» со свойствами молекул содержит следующие столбцы:

- «iupac» - название вещества по IUPAC;
- «smiles» - символьное обозначение SMILES;
- «expt» - экспериментальное значение свободной энергии сольватации (кДж/моль);
- «calc» - вычисленное значение свободной энергии сольватации (кДж/моль);

Файл «esol.csv» со свойствами молекул содержит множество столбцов, однако нас интересуют следующие:

- «smiles»;
- «measured log solubility in mols per litre» - экспериментальное значение логарифма растворимости;

- «ESOL predicted log solubility in mols per litre» - вычисленное значение логарифма растворимости;

Для представления SMILES в виде молекулярного графа, необходимо использовать функцию rdkit.Chem.MolFromSmiles.

Для решения задачи регрессии можно использовать графовые нейронные сети. Используем простую модель на основе сверточной графовой нейронной сети (GCNConv). Функционал для данного типа нейронных сетей предоставляет модуль «torch_geometric». В качестве наборов данных мы использовали «freesolv» и «esol».

Свойства атомов в молекуле закодированы методом «One-Hot Encoding»:

- Номер атома представлен в виде 118-мерного вектора, поскольку в периодической таблице Д.И. Менделеева насчитывается 118 элементов;
- Формальный заряд представлен одним целым числом;
- Одним числом представлена принадлежность атома ароматическому циклу (1, если принадлежит, в противном случае 0);
- Хиральность атома представляется 9-мерным вектором;
- Гибридизация атома тоже обозначается 9-мерным вектором;

В итоге получим 138-мерный вектор свойств атома. Тип связи между атомами обозначается числом, равным кратности связи между атомами.

Подбор гиперпараметров нейронной сети производился следующим образом:

1) Размер скрытого слоя: 20n (n = 1 – 10);
2) Скорость обучения: 0.001, 0.0025, 0.005, 0.01, 0.025, 0.05.

Итого испытано 60 комбинаций гиперпараметров. После обучения лучшая модель имеет ошибку составила 1.1934. Для «freesolv» ошибка составила 0.9344.

Результаты обучения собраны в таблицу 3.1 ниже для наглядности.

Таблица 3.1. Значения функции потерь для моделей.

Датасет / Модель	ESOL	Freesolv
Исследования	0.8861	2.6467
https://github.com/Anatoliibobrov/ChemFPNN перцептрон на основе Morgan Finger Print	1.6459	2.3562
GCN	1.1934	0.9344

Обе модели показали лучшую точность, чем исследование на наборе данных «freesolv». Как можем заметить, графовая нейронная сеть лучше справилась с задачей, чем обычный перцептрон. Сравнительные графики функция потерь-эпоха приведены на рисунках 3.1 и 3.2.

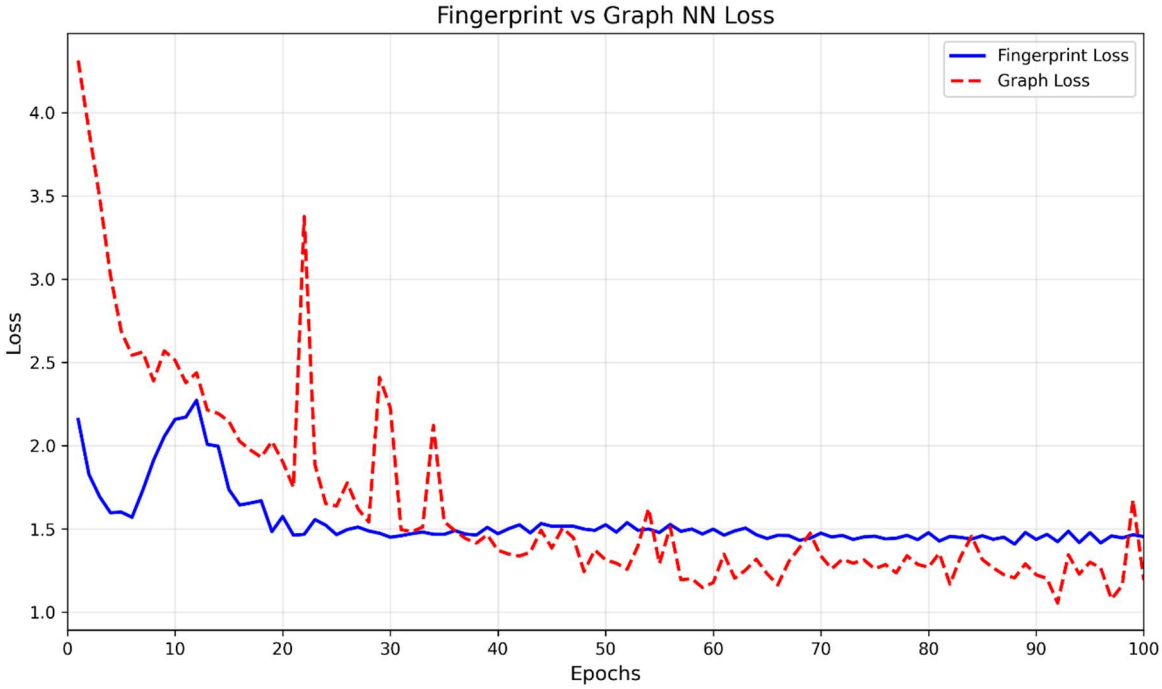

Рисунок 3.1 – сравнительный график обучения моделей на ESOL

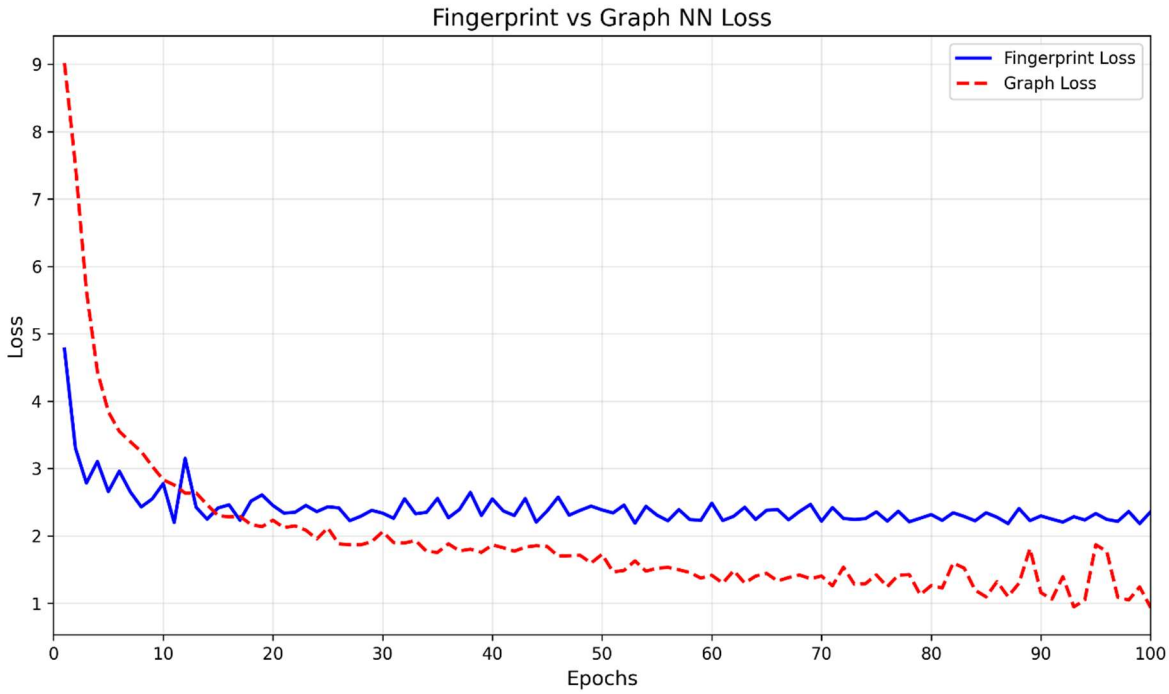

Рисунок 3.2 – сравнительный график обучения моделей на freesol