

Улучшение качества эпитопного картирования методами глубокого обучения

Цириков С.А., группа 18.Б11-мм, мат-мех СПбГУ

Научный руководитель: д.ф.-м.н., проф. Граничин О.Н.

Консультант: исследователь JetBrains Research, Зенкова Н.В.

Рецензент: к.ф.-м.н., Ерофеева В.А.

Предметная область

Исследование комплексов антиген-антитело важно для понимания механизмов работы иммунной системы живых организмов и при разработке лекарственных средств.

Компьютерное моделирование позволяет снизить стоимость и повысить скорость таких исследований.

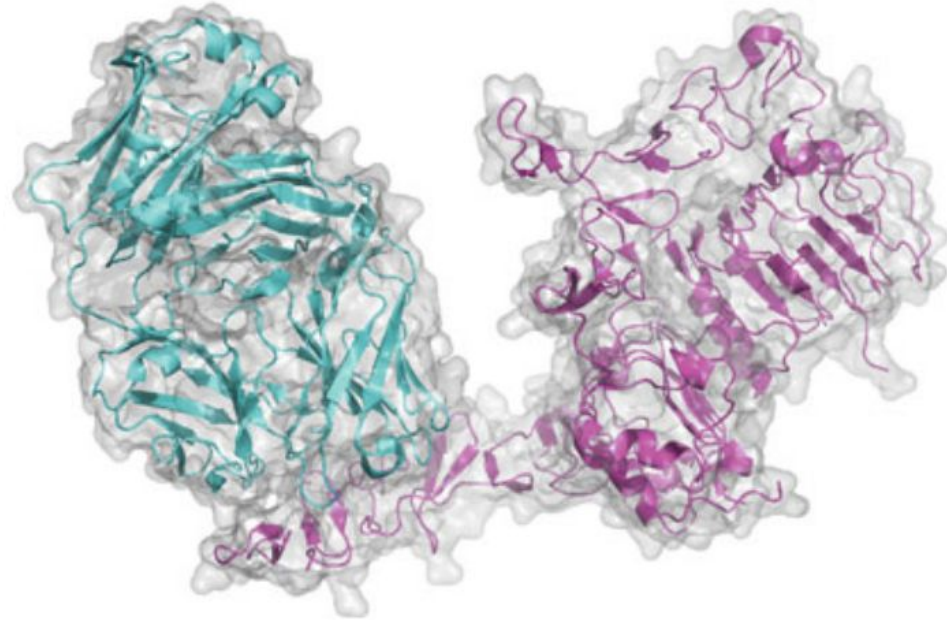


Рис. из J.Nilvebrant, J.Rockberg, “An Introduction to Epitope Mapping”, 2018, комплекс антитело-антиген

Мотивация

Существующие модели стараются максимизировать recall, в то время как для прикладных задач возникает требование к высокому показателю метрики precision.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

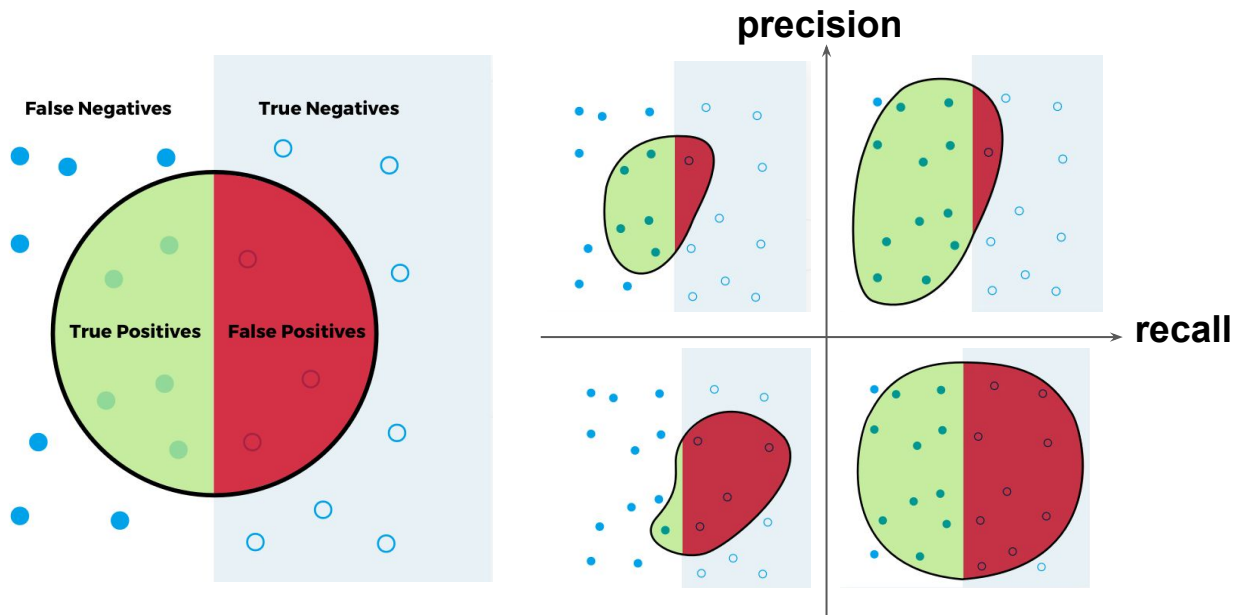


Рис. с портала stradigi.ai,
соотношение метрик precision и recall

Цель и задачи

Целью данной работы является создание модели машинного обучения для решения задачи эпитопного картирования с метрикой качества precision, превосходящей аналоги.

Для достижения указанной цели были поставлены следующие задачи:

- Сделать обзор существующих моделей
- Выбрать набор данных для обучения
- Спроектировать и реализовать модель глубокого обучения
- Провести эксперименты, сравнить результаты с другими моделями
- Разработать прототип системы для использования предобученной модели

Обзор

Модель	Представление	Вход	Подход
PECAN	Графовое	Комплекс	CAN
PInet	Геометрическое	Комплекс	GAN
Epitope3D	Геометрическое	Антиген	AdaBoost

Таб., сравнение актуальных моделей в задаче эпитопного картирования

Данные

База данных	Количество молекул, шт.	Тип
PRISM	6001	Белки
PDB	10714	Белки
DBD5	409	Антигены
Anbase	570	Антигены

Таб., сравнение баз данных, используемых в задачах взаимодействия антитела и антигена

Модель

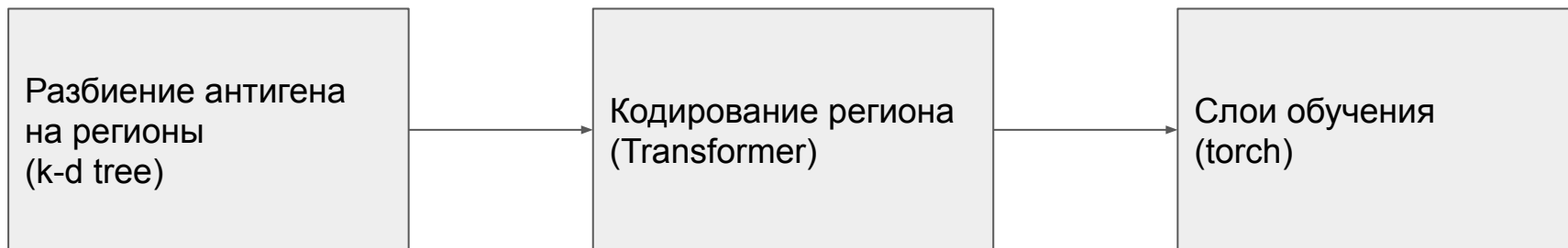


Рис., архитектура модели

Сравнение результатов

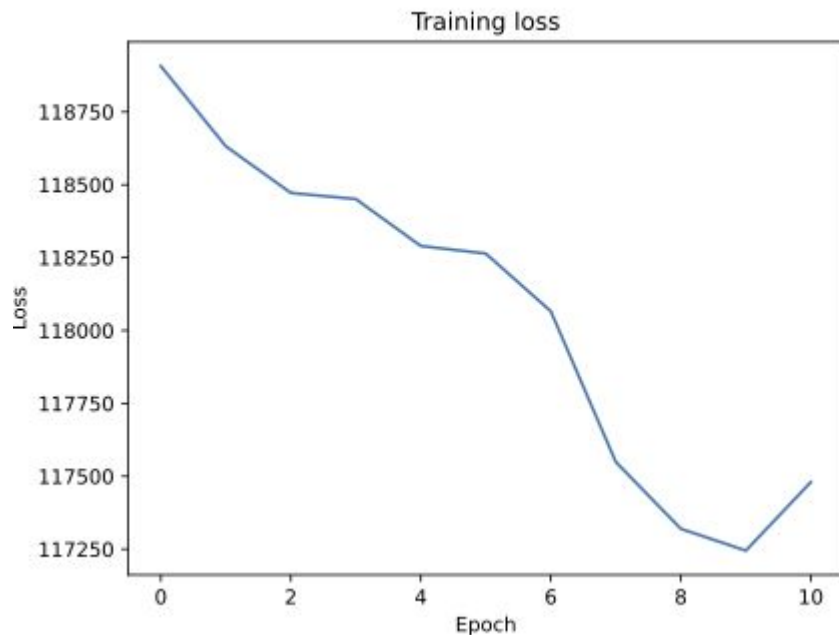


Рис., график функции потерь по эпохам

Модель	precision	recall
Epitope3D	0.084	0.191
Модель	0.153	0.090

Таб., сравнение метрик качества

Результаты

- Проведён обзор существующих моделей: PECAN, PInet, Epitope3D, Molformer
- Выбрана для обучения база данных Anbase
- Спроектирована и реализована модель глубокого обучения по архитектуре Transformer
- Проведены эксперименты, сделано сравнение метрик precision и recall с моделью Epitope3D, достигнуто увеличение precision на 80%
- Для использования предобученной модели разработан прототип системы на языке программирования python