

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ  
КАФЕДРА БИОМЕДИЦИНСКОЙ ИНФОРМАТИКИ

**Исследование архитектур нейронных сетей для предсказания  
свойств лекарственно-подобных молекул**

Курсовая работа

Благодарного Артёма Андреевича  
ФПМИ, БМИ, 4 курс 3 группа

Научный руководитель:  
профессор, доктор  
физико-математических наук,  
Тузииков А.В.

**Цель работы:**

Изучить современные архитектуры нейронных сетей для предсказания свойств лекарственно-подобных молекул

**Задачи работы:**

1. Изучить современные архитектуры нейронных сетей.
2. Провести эксперимент.
3. Проанализировать результат.

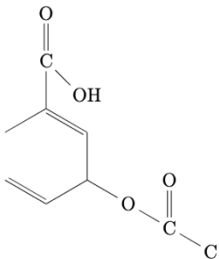
## Жизненный цикл лекарства: ADMEТ



- ▶ **Absorption** — попадание в кровоток
- ▶ **Distribution** — распределение по тканям
- ▶ **Metabolism** — биотрансформация
- ▶ **Excretion** — выведение из организма
- ▶ **Toxicity** — токсическое воздействие

## SMILES: Линейная нотация структур

### Acetylsalicylic acid



#### SMILES String:

```
CC(=O)Oc1ccccc1C(=O)O
```

- `c1ccccc1`: Ароматическое кольцо
- `(=O)`: Двойная связь с кислородом
- `()`: Ветвления функциональных групп

## Архитектура: Morgan + MLP

```
MorganMLP(  
  (net): Sequential(  
    (0): Linear(in_features=1024, out_features=1024, bias=True)  
    (1): ReLU()  
    (2): Dropout(p=0.1, inplace=False)  
    (3): Linear(in_features=1024, out_features=512, bias=True)  
    (4): ReLU()  
    (5): Dropout(p=0.1, inplace=False)  
    (6): Linear(in_features=512, out_features=128, bias=True)  
    (7): ReLU()  
    (8): Linear(in_features=128, out_features=1, bias=True)  
  )  
)
```

## Архитектура: SMILES + CNN

```
SmilesCNN(  
  (embedding): Embedding(64, 64, padding_idx=0)  
  (conv1): Conv1d(64, 32, kernel_size=(4,), stride=(1,))  
  (conv2): Conv1d(32, 64, kernel_size=(6,), stride=(1,))  
  (conv3): Conv1d(64, 96, kernel_size=(8,), stride=(1,))  
  (mlp): Sequential(  
    (0): Linear(in_features=96, out_features=32, bias=True)  
    (1): ReLU()  
    (2): Linear(in_features=32, out_features=32, bias=True)  
    (3): Linear(in_features=32, out_features=1, bias=True)  
  )  
)
```

## Архитектура: GCN

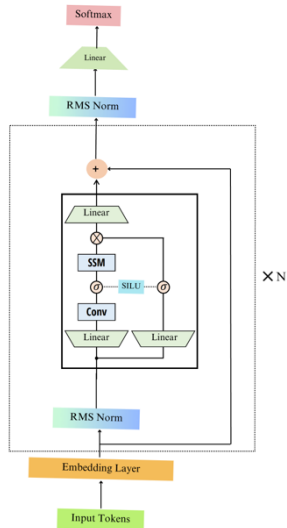
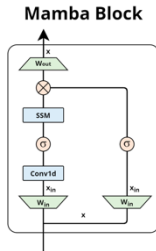
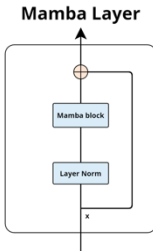
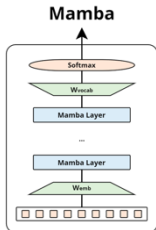
```
GCN(  
    (convs): ModuleList(  
      (0): GCNConv(9, 100)  
      (1-4): 4 x GCNConv(100, 100)  
    )  
    (head): Linear(in_features=100, out_features=1, bias=True)  
)
```

## Архитектура: NeuralFP (Frozen Encoder)

```
NeuralFP(  
  (encoder): GCN(  
    (convs): ModuleList(  
      (0): GCNConv(9, 100)  
      (1-4): 4 x GCNConv(100, 100)  
    )  
    (head): Identity()  
  )  
  (decoder): Sequential(  
    (0): Linear(in_features=100, out_features=200, bias=True)  
    (1): ReLU()  
    (2): Linear(in_features=200, out_features=100, bias=True)  
    (3): ReLU()  
    (4): Linear(in_features=100, out_features=50, bias=True)  
    (5): ReLU()  
    (6): Linear(in_features=50, out_features=1, bias=True)  
  )  
)
```



## Архитектура: Mamba



## Метрики оценки моделей



### Классификация

*Да/Нет (Токсичность, ГЭБ, Pgp)*



#### ROC-AUC

Стандарт для сбалансированных данных.



#### PR-AUC

Для несбалансированных данных (когда токсичных молекул мало).



### Регрессия

*Числа (Растворимость, Клиренс)*



#### MAE

Средняя абсолютная ошибка. Устойчива к выбросам.



#### Spearman $\rho$

Ранговая корреляция. Оценивает правильность ранжирования (порядок).

## Полученные результаты

Группа	Задача	Метрика	Morgan+MLP	SMILES+CNN	GCN	NeuralFP
PhysChem	AqSol	MAE ↓	1.1097	<b>0.9871</b>	1.3727	1.4733
	Lipo	MAE ↓	<b>0.6952</b>	0.7462	1.0099	0.9703
Absorption	Caco2	MAE ↓	<b>0.4822</b>	0.7722	1.0998	1.0300
	HIA	ROC-AUC ↑	<b>0.8045</b>	0.7985	0.7652	0.6182
	Pgp	ROC-AUC ↑	<b>0.8897</b>	0.8817	0.7401	0.8108
	Bioav	ROC-AUC ↑	0.5469	0.5612	<b>0.5833</b>	0.5182
Distribution	BBB	ROC-AUC ↑	0.8429	<b>0.9058</b>	0.4588	0.7388
	PPBR	MAE ↓	11.9905	<b>10.4078</b>	15.9136	16.9043
	VD	MAE ↓	2.4211	<b>2.3123</b>	2.5624	2.4784
Metabolism	CYP2D6-I	PR-AUC ↑	0.5425	<b>0.6008</b>	0.1764	0.2612
	CYP3A4-I	PR-AUC ↑	0.7866	<b>0.7958</b>	0.5398	0.5812
	CYP2C9-I	PR-AUC ↑	<b>0.7242</b>	0.6923	0.5775	0.5417
	CYP2D6-S	PR-AUC ↑	<b>0.7373</b>	0.5490	0.3902	0.4028
	CYP3A4-S	ROC-AUC ↑	<b>0.7239</b>	0.6355	0.5769	0.5545
	CYP2C9-S	PR-AUC ↑	<b>0.4223</b>	0.3883	0.3713	0.3680
Excretion	Half-Life	Spearman ↑	<b>0.6276</b>	0.0472	0.1115	-0.1117
	CL-Micro	Spearman ↑	<b>0.4243</b>	0.1770	0.0974	0.0004
	CL-Hepa	Spearman ↑	<b>0.2577</b>	0.2452	-0.0030	-0.0504
Toxicity	hERG	ROC-AUC ↑	0.7350	<b>0.8506</b>	0.6931	0.6456
	AMES	ROC-AUC ↑	0.7880	<b>0.8103</b>	0.6800	0.6212
	DILI	ROC-AUC ↑	0.8229	<b>0.8837</b>	0.7483	0.8264
	LD50	MAE ↓	<b>0.5645</b>	0.5680	0.7237	0.7192

## Продолжение работы:

1. Реализовать архитектуру Mamba.
2. Предобучить на строках Smiles.
3. Fine-tune под определённые задачи.
4. Сравнить полученные результаты с текущими моделями

Спасибо за внимание!