Применение сетей Колмогорова-Арнольда для анализа и дизайна PAM-специфичности CRISPR-Cas систем

Комкова Ирина Владимировна

руководитель: Шмаков Сергей Анатольевич

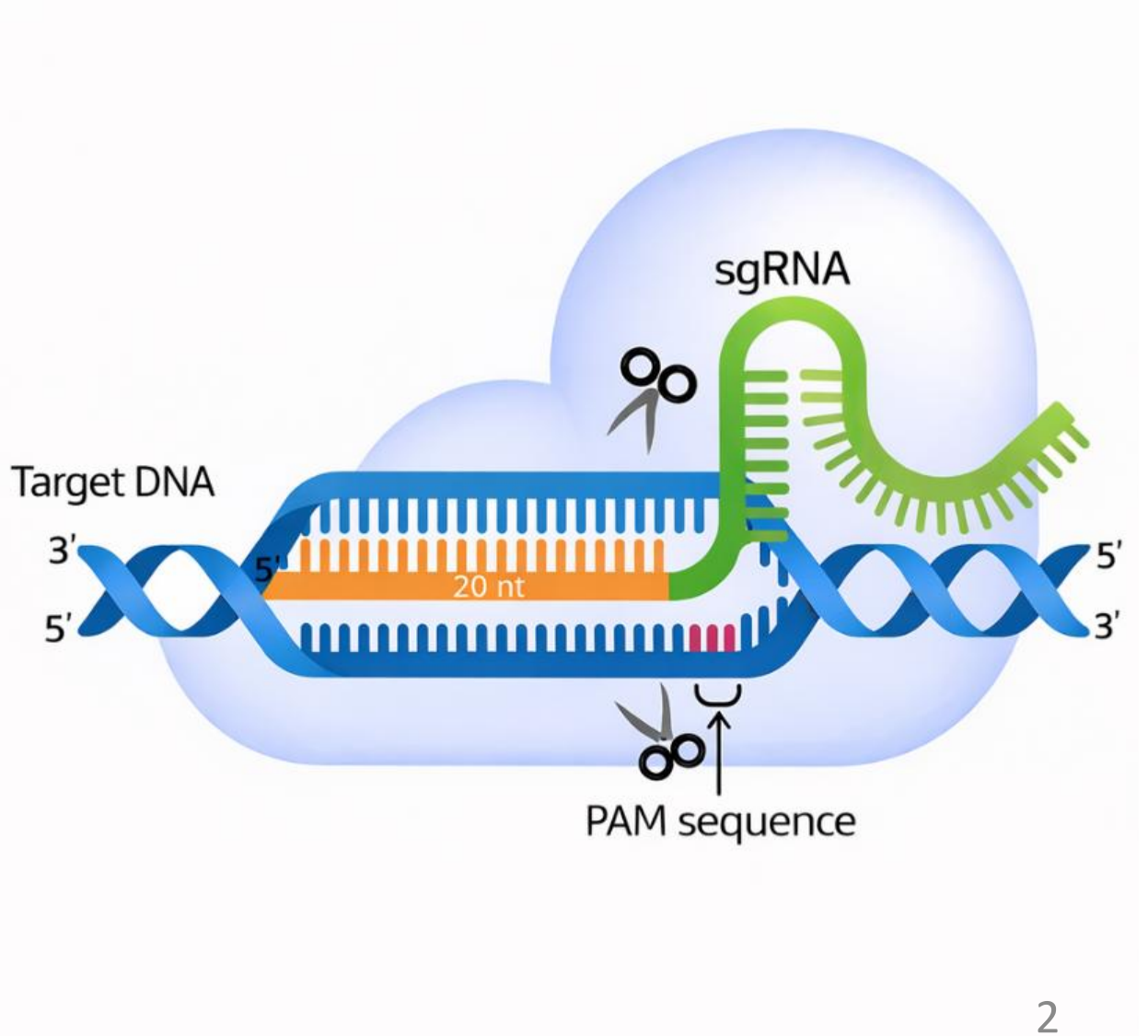
Научно-учебная группа эволюционной геномики,

Международная лаборатория статистической и вычислительной геномики,

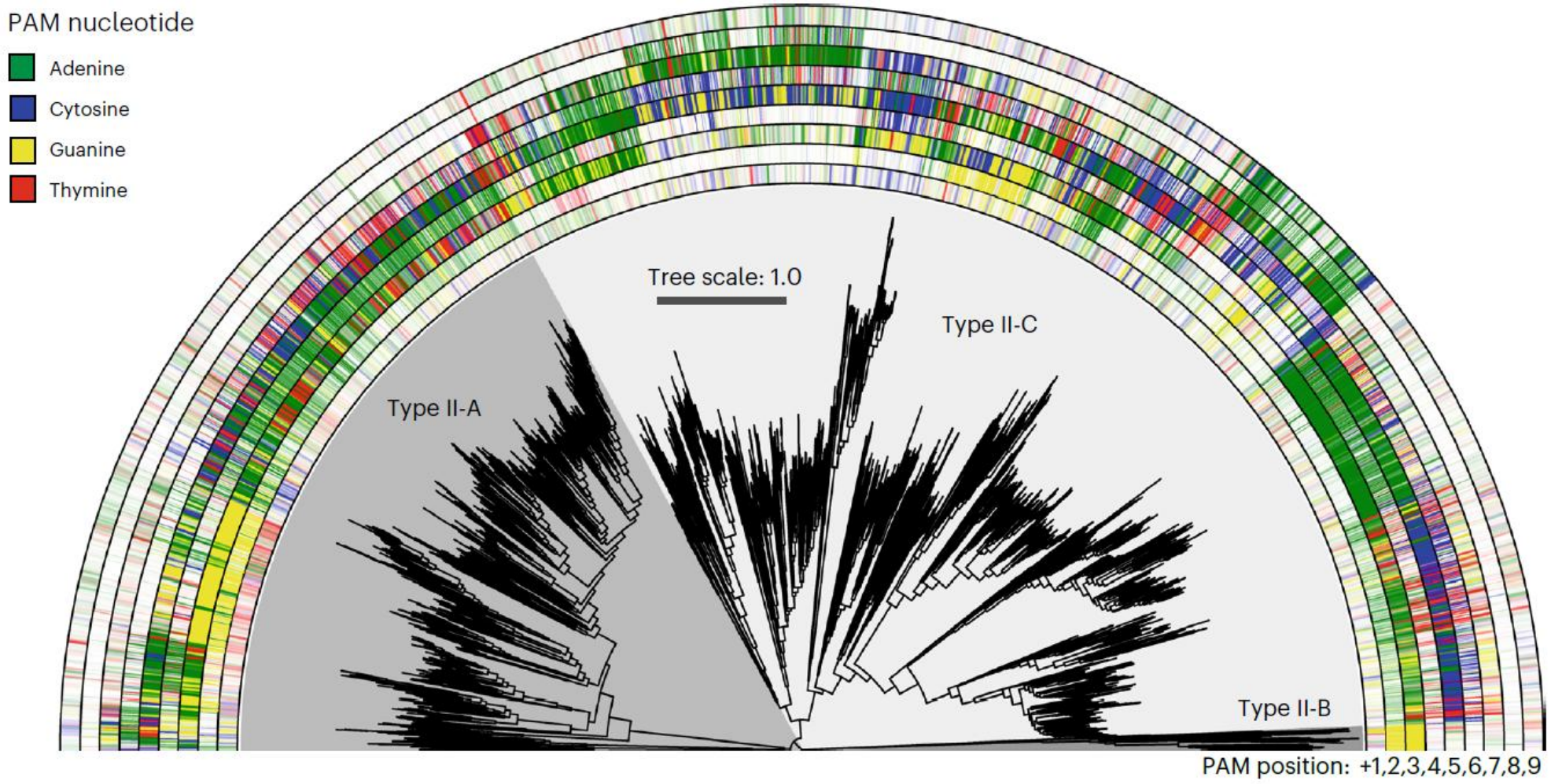
НИУ ВШЭ

CRISPR-Cas и PAM-специфичность

- PAM - короткий мотив рядом с мишенью (обычно 2–6 нт)
- CRISPR-Cas: sgRNA задаёт мишень (протоспейсер) – комплиментарна ей, Cas делает разрез ДНК
- PAM-специфичность - какие именно PAM этот Cas распознаёт



Разнообразие PAM



Постановка задачи

Зачем:

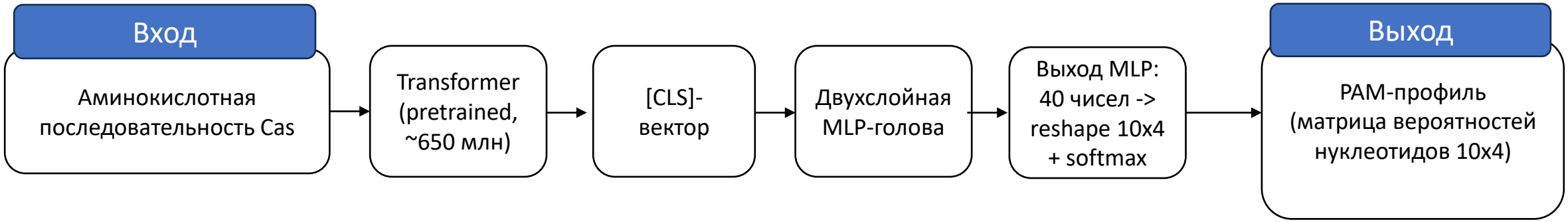
- Предсказать PAM для новых/малоизученных Cas
- Предложить дизайн Cas, который специфичен данному PAM

Нужно построить модель:

- вход: аминокислотная последовательность Cas-белка
- выход: матрица вероятностей A/C/G/T на каждой позиции PAM размером 10×4

Protein2PAM (Stephen Nayfach et al., 2025)

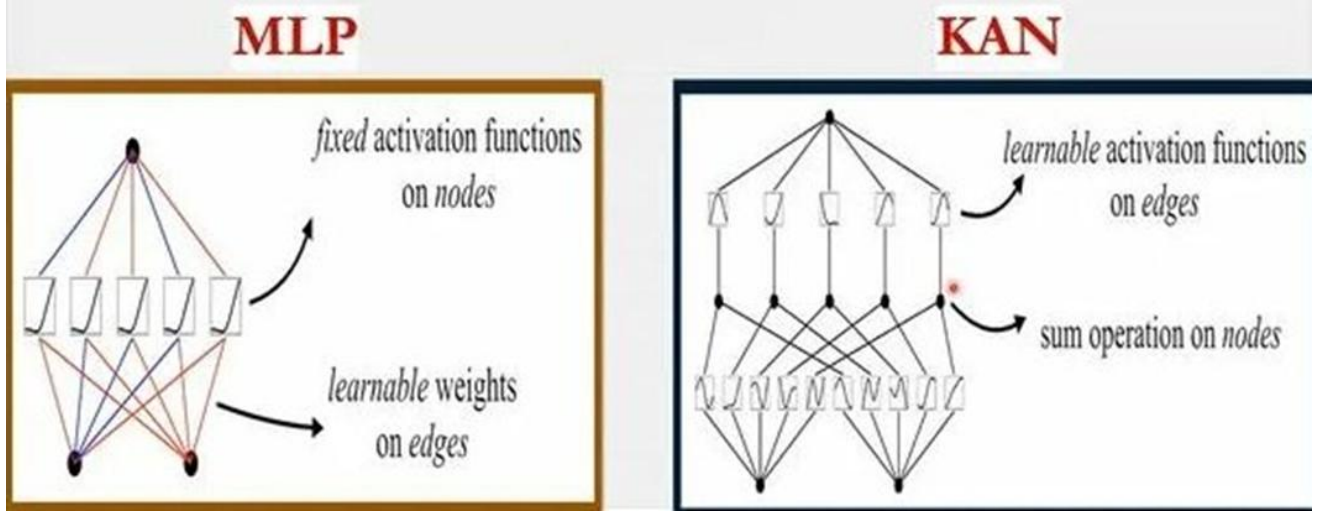
- Protein2PAM - protein language model
- Обучение на 45 тыс. пар Cas/PAM



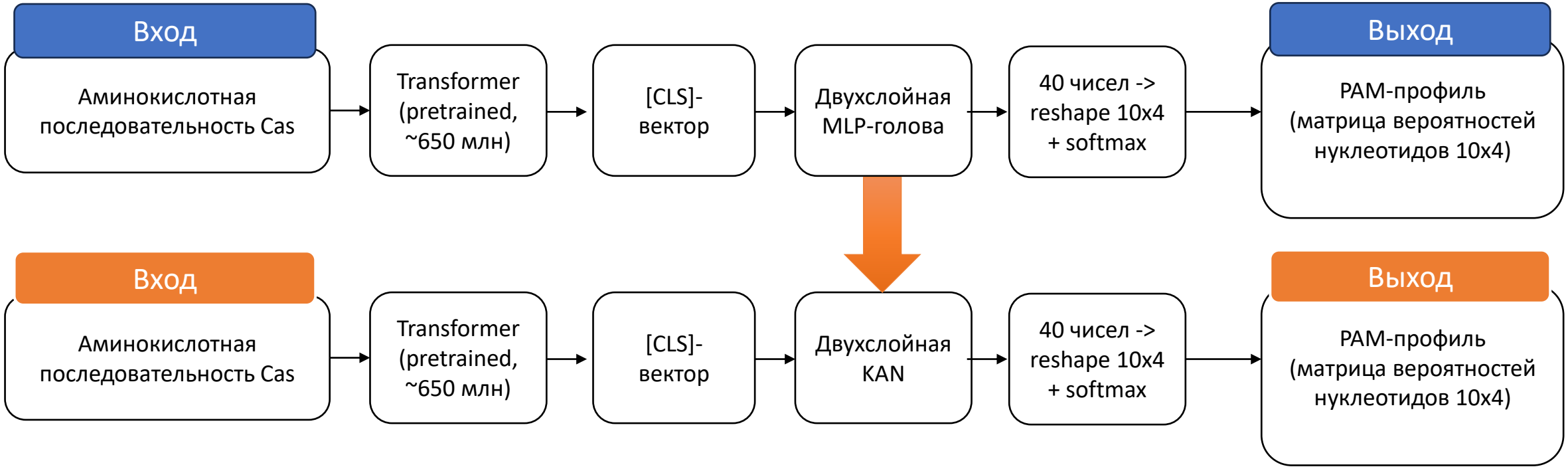
Почему KAN?

Kolmogorov-Arnold graph neural networks for molecular property prediction (Longlong Li et al., 2025)

Model	BACE	BBBP	ClinTox
Number of molecules	1,513	2,039	1,478
Number of average atoms	65	46	50.58
Number of tasks	1	1	2
D-MPNN[46]	0.809(0.006)	0.710(0.003)	0.906(0.007)
AttentiveFP[45]	0.784(0.022)	0.663(0.018)	0.847(0.003)
N-GramRF[47]	0.779(0.015)	0.697(0.006)	0.775(0.040)
N-GramXGB[47]	0.791(0.013)	0.691(0.008)	0.875(0.027)
PretrainGNN[48]	0.845(0.007)	0.687(0.013)	0.726(0.015)
GraphMVP[49]	0.812 (0.009)	0.724 (0.016)	0.791 (0.028)
MolCLR[50]	0.824 (0.009)	0.722 (0.021)	0.912 (0.035)
GEM[51]	0.856(0.011)	0.724(0.004)	0.901(0.013)
Mol-GDL[28]	0.863(0.019)	0.728(0.019)	0.966(0.002)
Uni-mol[52]	0.857 (0.002)	0.729 (0.006)	0.919 (0.018)
SMPT[53]	0.873 (0.015)	0.734 (0.003)	0.927 (0.002)
GNN-SKAN[40]	0.747(0.009)	0.676(0.014)	-
GraphKAN[39]	0.823(0.011)	0.731(0.017)	0.984(0.003)
KA-GNNs[30]	0.752(0.011)	0.721(0.003)	0.972(0.001)
KA-GCN	**0.890**(0.014)	**0.787**(0.014)	**0.992**(0.005)
KA-GAT	0.884(0.004)	0.785(0.021)	0.991(0.005)



Идея


Вход
Аминокислотная последовательность Cas
Transformer (pretrained, ~650 млн)
[CLS]-вектор
Двухслойная MLP-голова
40 чисел -> reshape 10x4 + softmax
Выход
PAM-профиль (матрица вероятностей нуклеотидов 10x4)
Вход
Аминокислотная последовательность Cas
Transformer (pretrained, ~650 млн)
[CLS]-вектор
Двухслойная KAN
40 чисел -> reshape 10x4 + softmax
Выход
PAM-профиль (матрица вероятностей нуклеотидов 10x4)

Дальнейшие шаги

- Обучить модель и сравнить с Protein2PAM
- Проверка качества: насколько хорошо модель предсказывает PAM-профиль для новых Cas
- Проверить предсказания модели в лаборатории

Благодарности

Мария Сергеевна Попцова