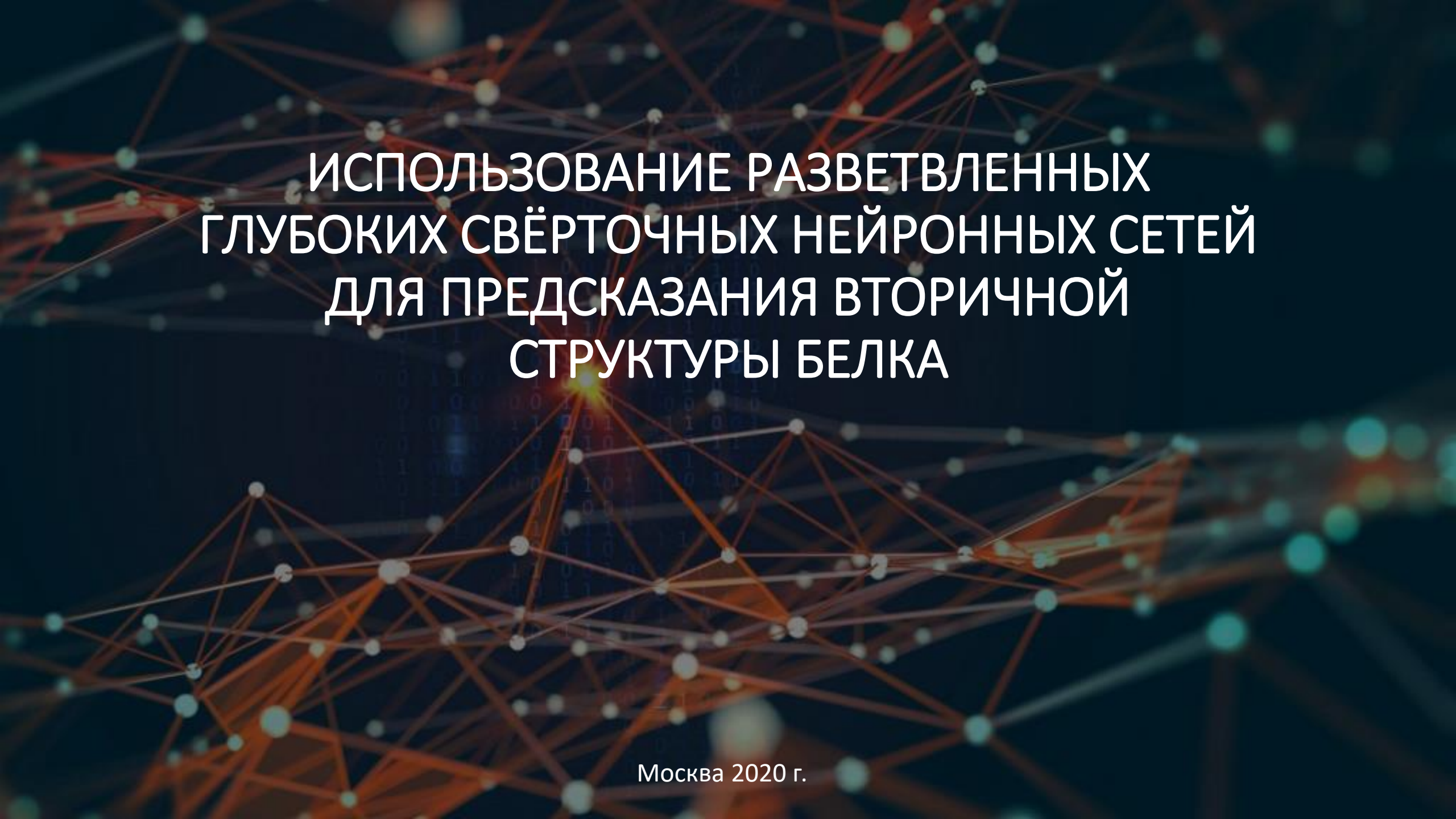


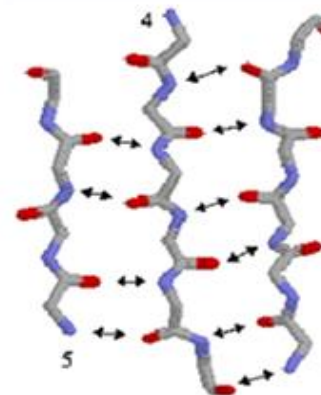
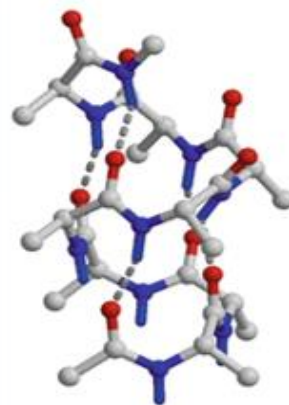
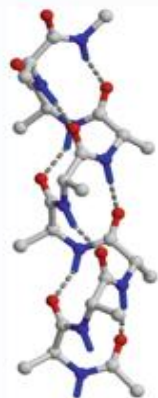
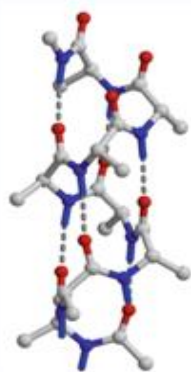


ИСПОЛЬЗОВАНИЕ РАЗВЕТВЛЕННЫХ ГЛУБОКИХ СВЁРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ПРЕДСКАЗАНИЯ ВТОРИЧНОЙ СТРУКТУРЫ БЕЛКА

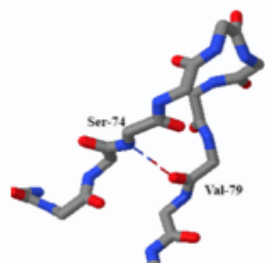
Москва 2020 г.

Восемь типов вторичной структуры белка по DSSP:

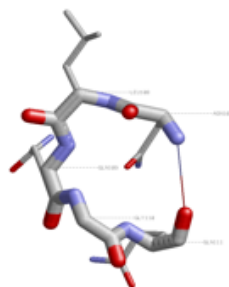
H = альфа-спираль G = спираль 3/10 I = пи-спираль E = бета-тяж



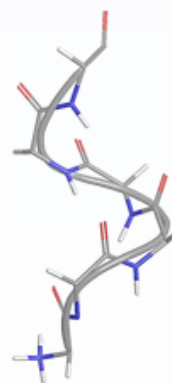
B = остаток в изолированном
бета-мостике



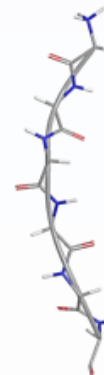
T = поворот с
водородной связью



S = изгиб



C = нерегулярные петли



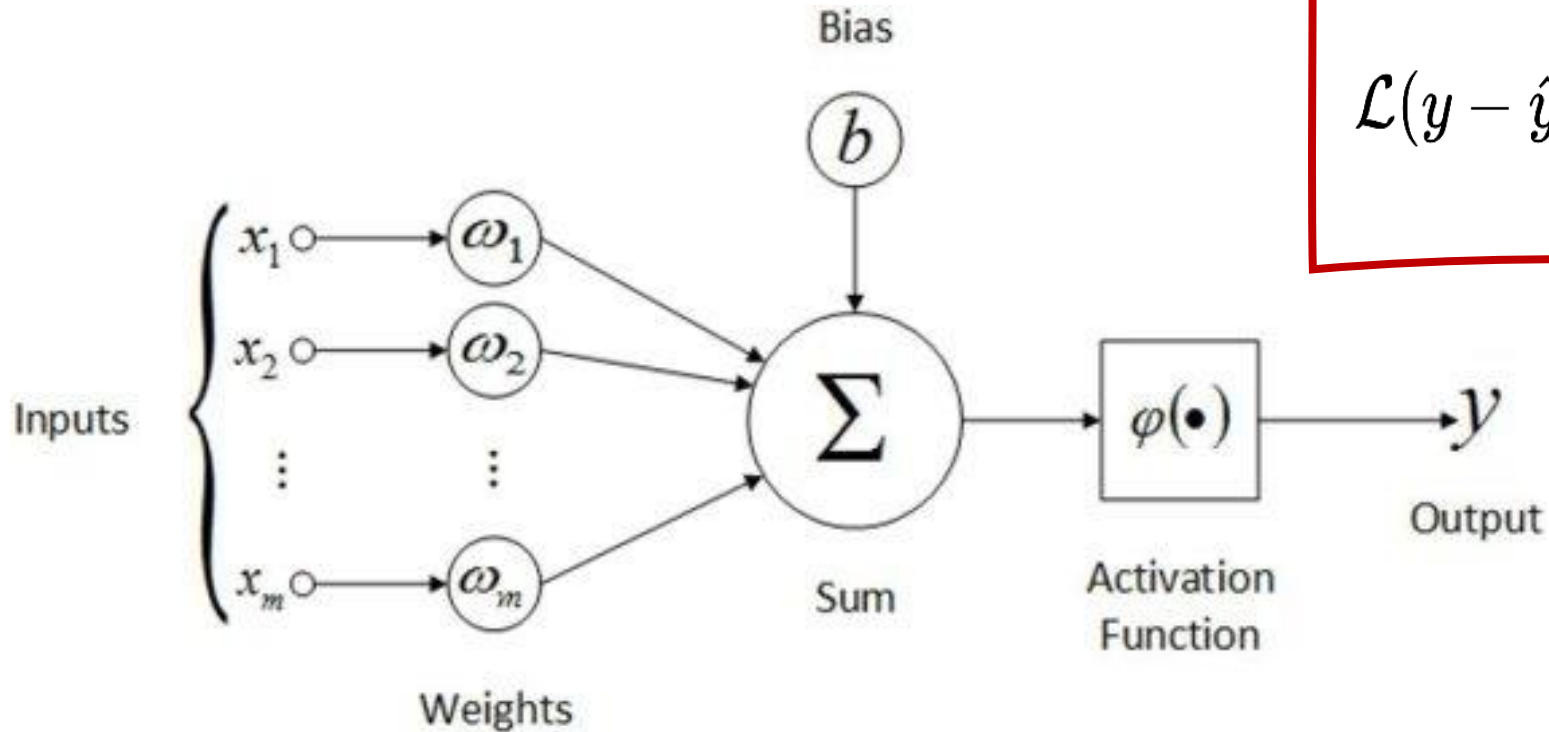
Цель работы:

- Сконструировать модель, предсказывающую с высокой точностью тип вторичной структуры белка (Q8) по аминокислотной последовательности.

Задачи:

- Подбор и обработка обучающего набора данных;
- Конструирование оптимальной сверточной нейронной сети (CNN);
- Обучение и оптимизация гиперпараметров полученной модели;
- Предсказание вторичной структуры белка для тестового набора данных.

Базовая архитектура перцептрона.

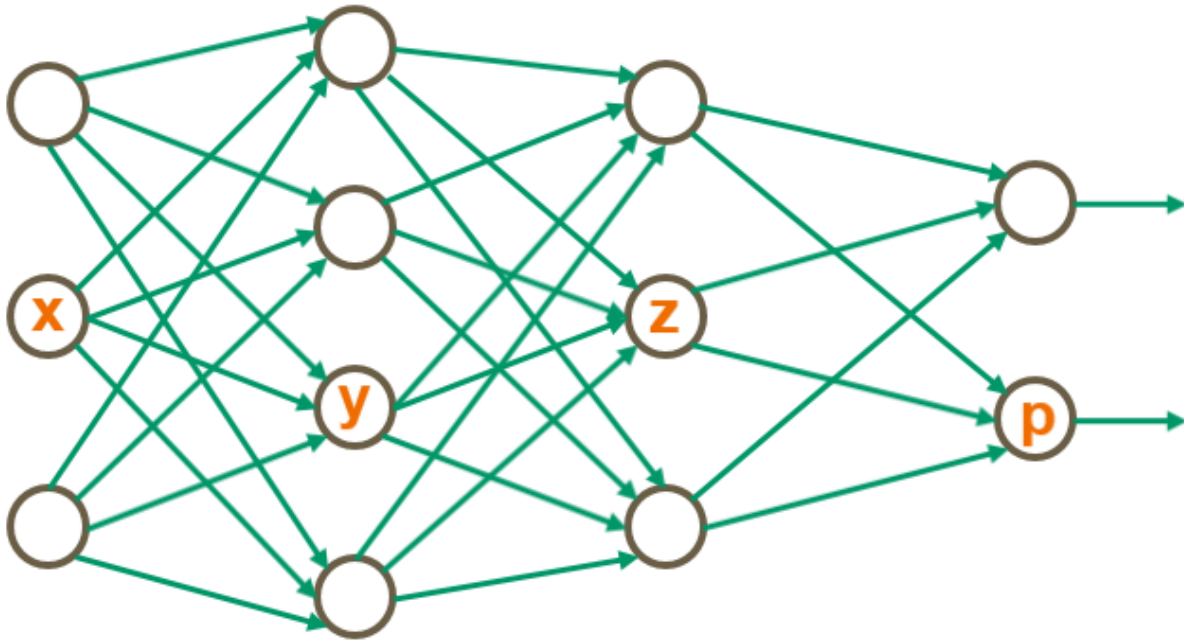


Функция потерь:

$$\mathcal{L}(y - \hat{y}) = - \sum_{i=1}^n y_i \log \hat{y}_i$$

[X1, X2, X3,..., Xm] – входные данные; Y – предсказанное значение.

Архитектура искусственных нейронных сетей



Каждый уровень выполняет умножение матриц: эти операции эффективно обрабатывают графические процессоры (GPU)

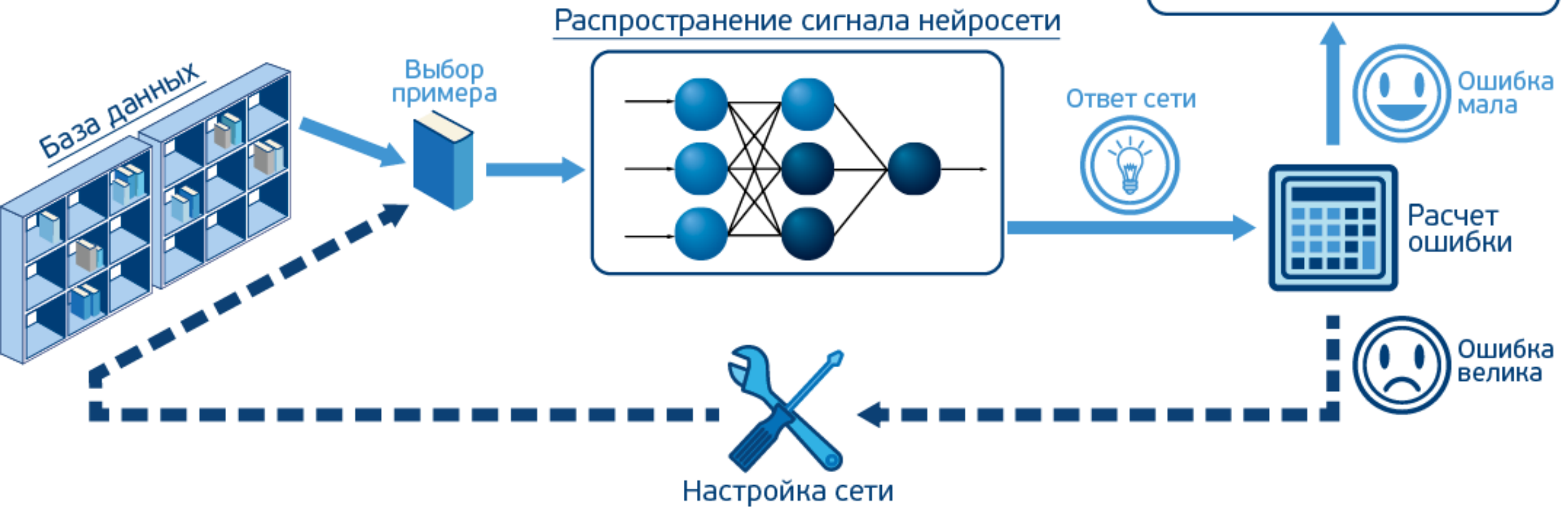
$$y_j = W_{ji} \cdot x_i + b_j$$

The matrix representation of the equation $y_j = W_{ji} \cdot x_i + b_j$ shows a 4x1 column vector (output) equal to a 4x3 matrix (weights) multiplied by a 3x1 column vector (input), plus a 4x1 column vector (bias).

$$z_k = W'_{kj} \cdot y_j + b'_k$$

The matrix representation of the equation $z_k = W'_{kj} \cdot y_j + b'_k$ shows a 3x1 column vector (output) equal to a 3x4 matrix (weights) multiplied by a 4x1 column vector (input), plus a 3x1 column vector (bias).

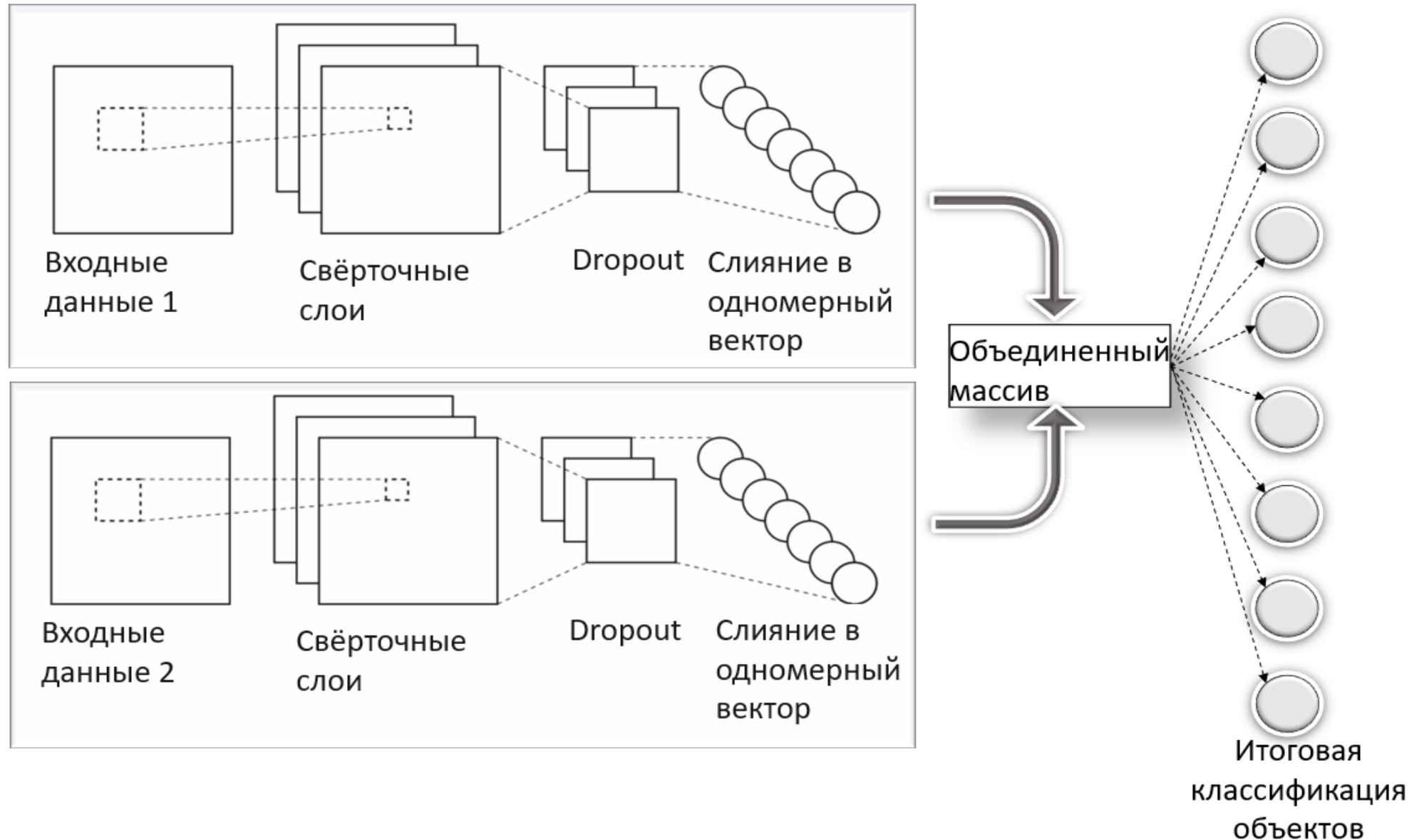
ПРОЦЕСС ОБУЧЕНИЯ НЕЙРОСЕТИ



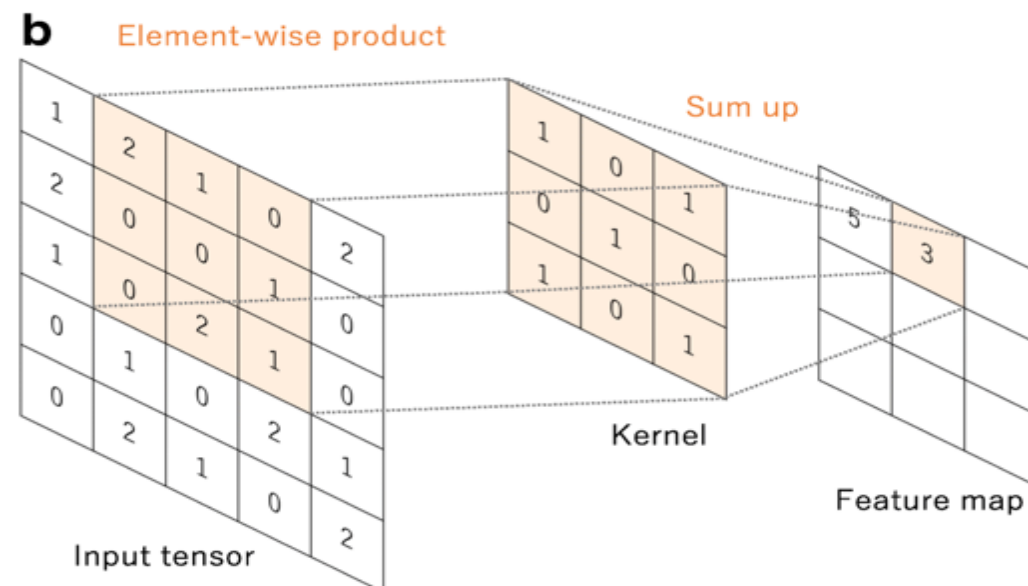
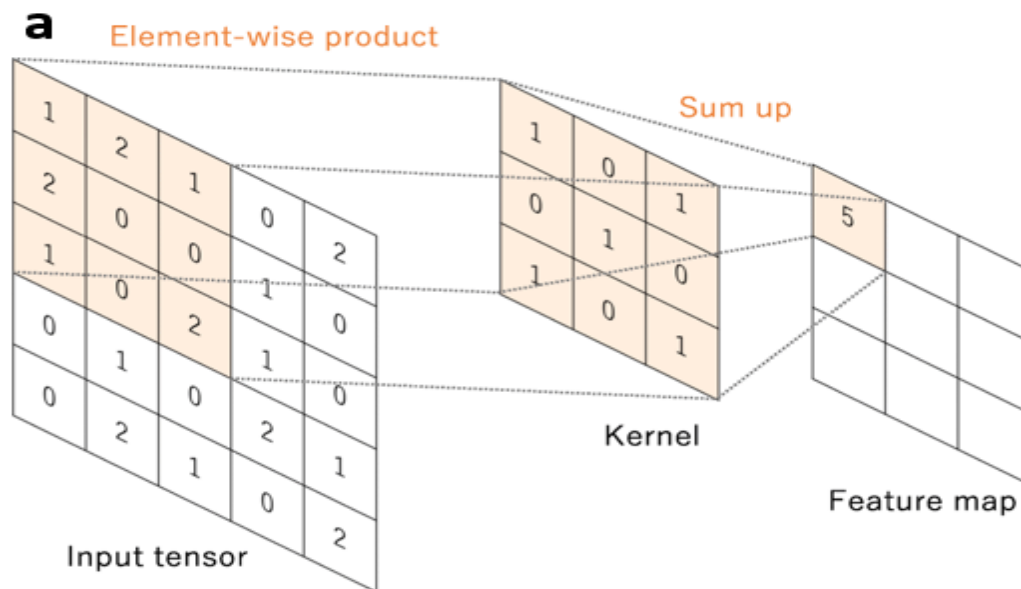
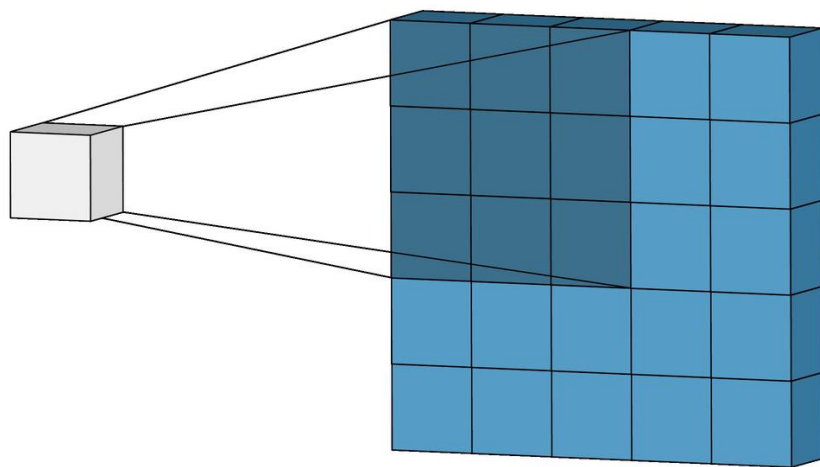
Общая схема алгоритма работы модели, классифицирующей аминокислоты по типам вторичной структуры.



Разветвленная глубокая свёрточная нейронная сеть

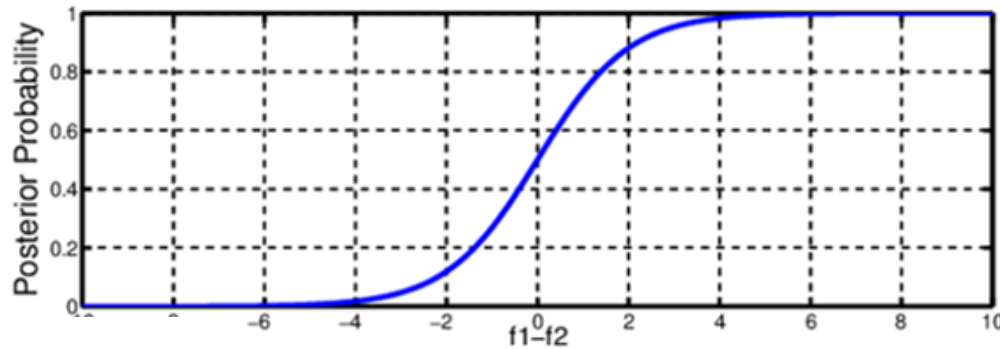


Алгоритм работы свёрточных слоёв



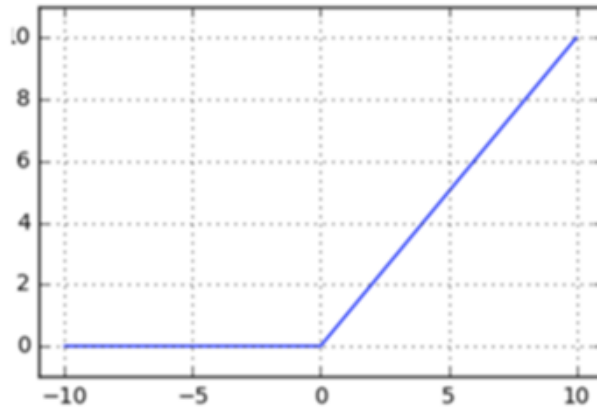
Применяемые в работе функции активации

Softmax



$$f_i(\vec{a}) = \frac{e^{a_i}}{\sum_k e^{a_k}}$$

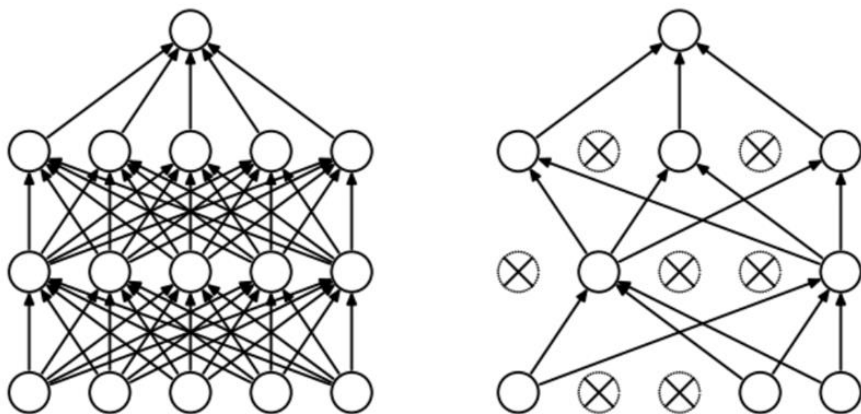
ReLU



$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$$

Методы борьбы с переобучением:

- Dropout



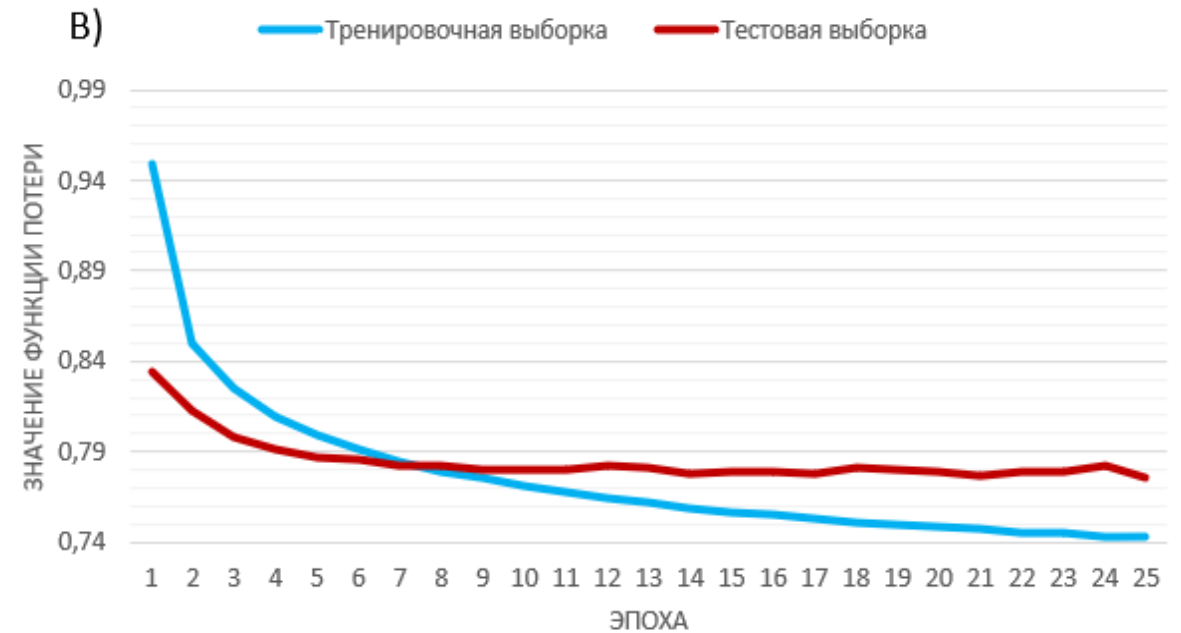
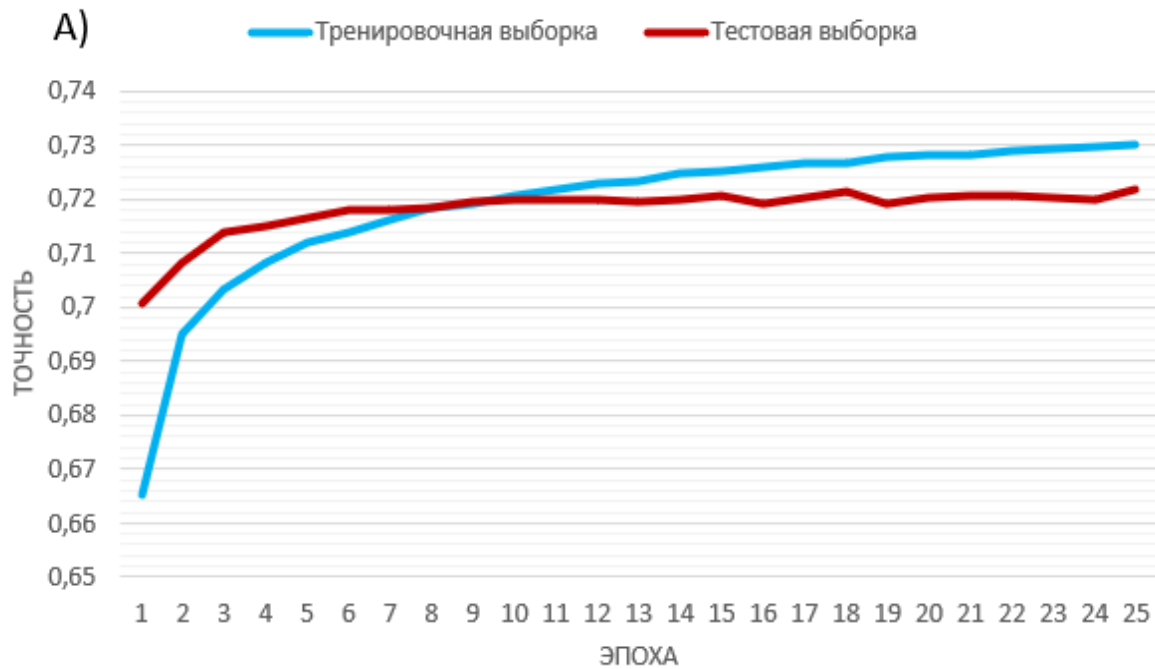
- Регуляризация Тихонова

$$J(w^{[1]}, b^{[1]}, \dots, w^{[L]}, b^{[L]}) = \frac{1}{m} \sum_{i=1}^m L(\hat{y}^{(i)}, y^{(i)}) + \frac{\lambda}{2m} \sum_{l=1}^L \|w^{[l]}\|_F^2$$

L – функция потерь, w – вес нейрона, λ – параметр регуляризации

- Гауссовский шум

Результаты обучения модели



Матрицы ошибок:

Истинные классы	L	11574	20	3182	237	0	912	571	1326	17822
	B	598	33	306	14	0	83	38	106	1178
	E	2217	18	14704	79	0	433	126	405	17982
	G	622	1	176	893	0	773	30	616	3111
	I	4	0	2	0	0	20	1	3	30
	H	898	2	312	316	0	23592	34	827	25981
	S	3668	3	910	160	0	623	1219	1695	8278
	T	1936	1	499	470	0	1596	247	5222	9971
	Σ	21517	78	20091	2169	0	28032	2266	10200	84353
			L	B	E	G	I	H	S	T
Предсказанные классы										

a)

Истинные классы	L	B	E	G	I	H	S	T
	65%	0%	18%	1%	0%	5%	3%	7%
	51%	3%	26%	1%	0%	7%	3%	9%
	12%	0%	82%	0%	0%	2%	1%	2%
	20%	0%	6%	29%	0%	25%	1%	20%
	13%	0%	7%	0%	0%	67%	3%	10%
	3%	0%	1%	1%	0%	91%	0%	3%
	44%	0%	11%	2%	0%	8%	15%	20%
	19%	0%	5%	5%	0%	16%	2%	52%
	L	B	E	G	I	H	S	T
Предсказанные классы								

b)

Истинные классы	L	B	E	G	I	H	S	T
	54%	26%	16%	11%	-	3%	25%	13%
	3%	42%	2%	1%	-	0%	2%	1%
	10%	23%	73%	4%	-	2%	6%	4%
	3%	1%	1%	41%	-	3%	1%	6%
	0%	0%	0%	0%	-	0%	0%	0%
	4%	3%	2%	15%	-	84%	2%	8%
	17%	4%	5%	7%	-	2%	54%	17%
	9%	1%	2%	22%	-	6%	11%	51%
	L	B	E	G	I	H	S	T
Предсказанные классы								

Процентное содержание ложноположительных результатов (a) и ложноотрицательных (b)

Матрица ошибок предсказания для тестового набора CB513

Результаты классификации аминокислотных остатков по типам вторичной структуры Q8 для CB513.

Классы структуры Q8	вторичной	Количество верных предсказаний	Количество ложноположительных результатов	Количество ложноотрицательных результатов
L(нерегулярные петли)		12522	9943	6248
B(остаток изолированном мостики)	в бета-	23	45	1145
E(бета-тяж)		14015	5387	3278
G(спираль 3/10)		716	1276	2218
I (пи-спираль)		0	0	30
H(альфа-спираль)		23534	4440	2389
S(изгиб)		1193	1047	7059
T(поворот водородной связью)	с	4822	4978	4749

Достигнутые результаты

- Были выбраны и проанализированы обучающий и тестовый наборы данных, также проведена их обработка.
- Была сконструирована модель машинного обучения на основе глубоких разветвленных сверточных нейронных сетей, решающая задачу классификации аминокислотных остатков на восемь классов вторичной структуры белка с точностью 0,722.
- Выявлено, что наилучшая точность предсказания достигается для таких классов вторичной структуры как α -спираль, β -лист, нерегулярные петли, изгибы, что может быть связано с их большим количеством в обучающем наборе и с выраженными особенностями остатков, относящихся к данным классам. Низкая точность предсказания наблюдается для остатков, относящихся к π -спирали, 3/10 спирали, остаткам в изолированном бета-мостике и поворот с водородной связью, что, как предполагается, связано с недостаточным количеством обучающих примеров для данных типов вторичных структур.