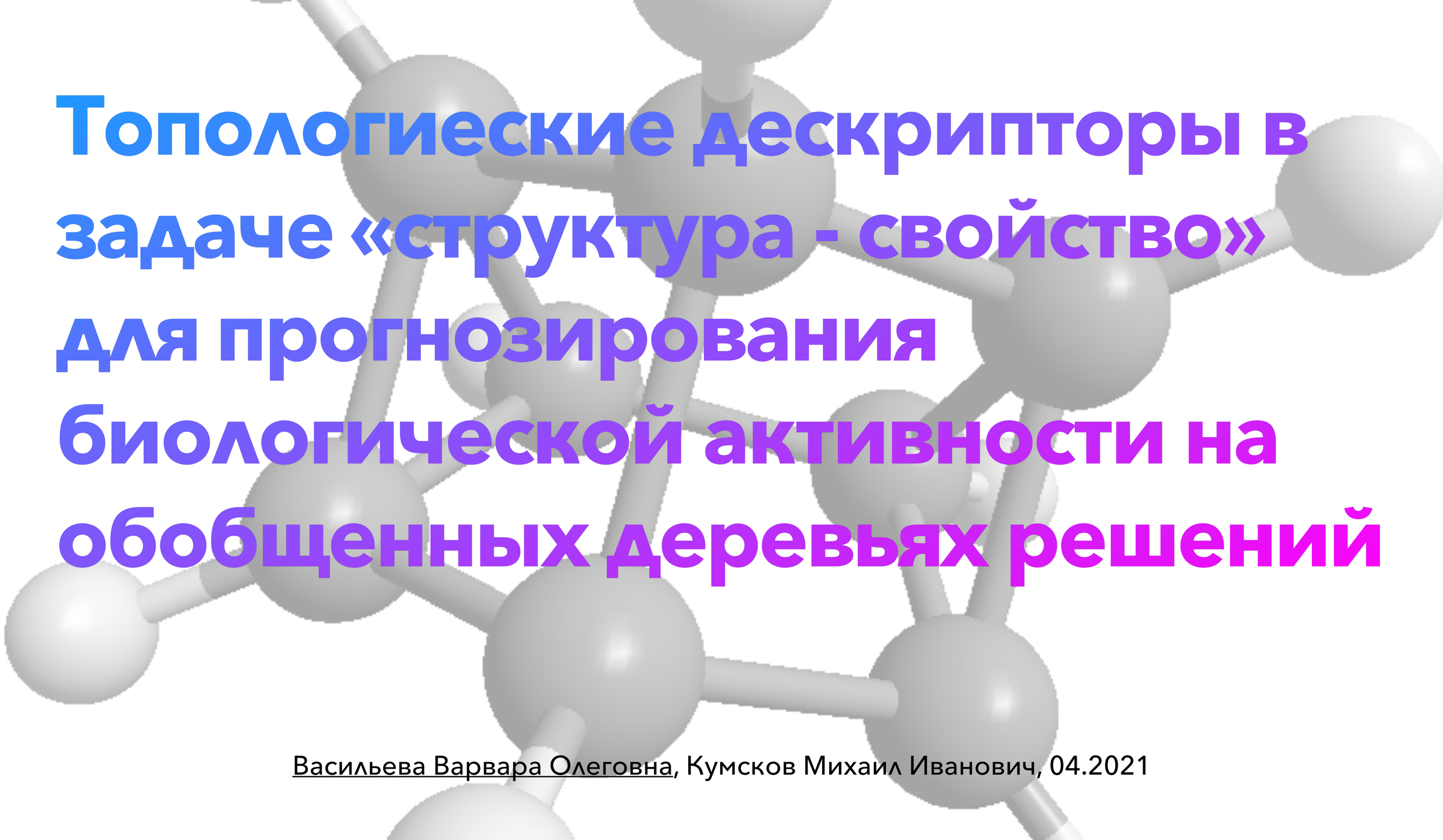
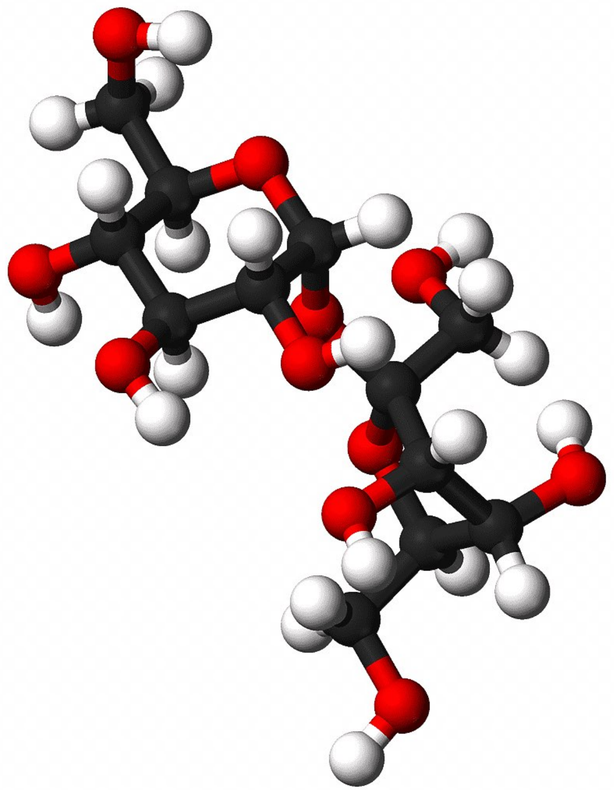


Топологиеские дескрипторы в задаче «структура - свойство» для прогнозирования биологической активности на обобщенных деревьях решений

Васильева Варвара Олеговна, Кумсков Михаил Иванович, 04.2021

Задача QSAR (QSPR)



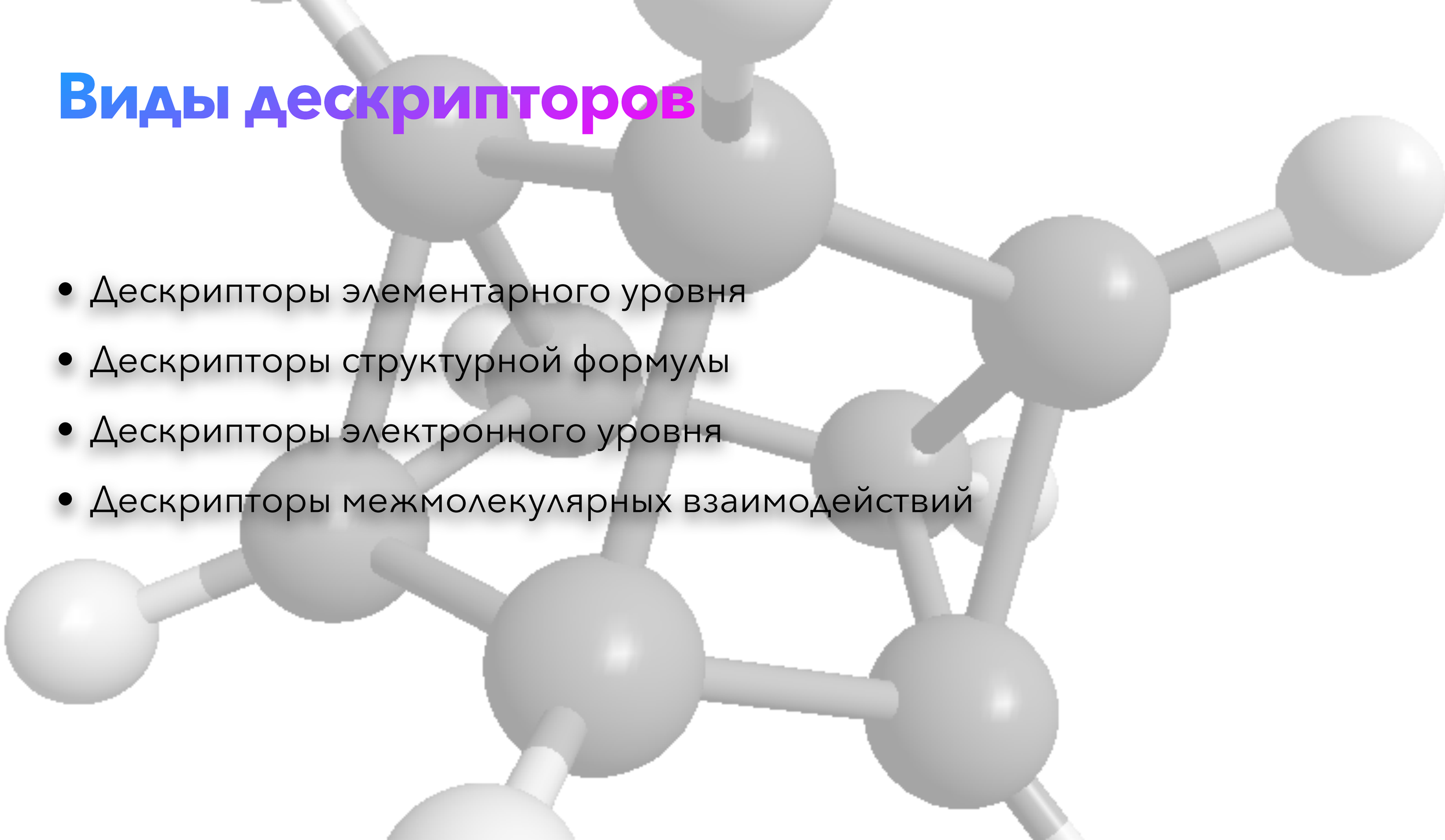
- активность (+1, -1)
- свойство (R)

Постановка задачи

Признаки

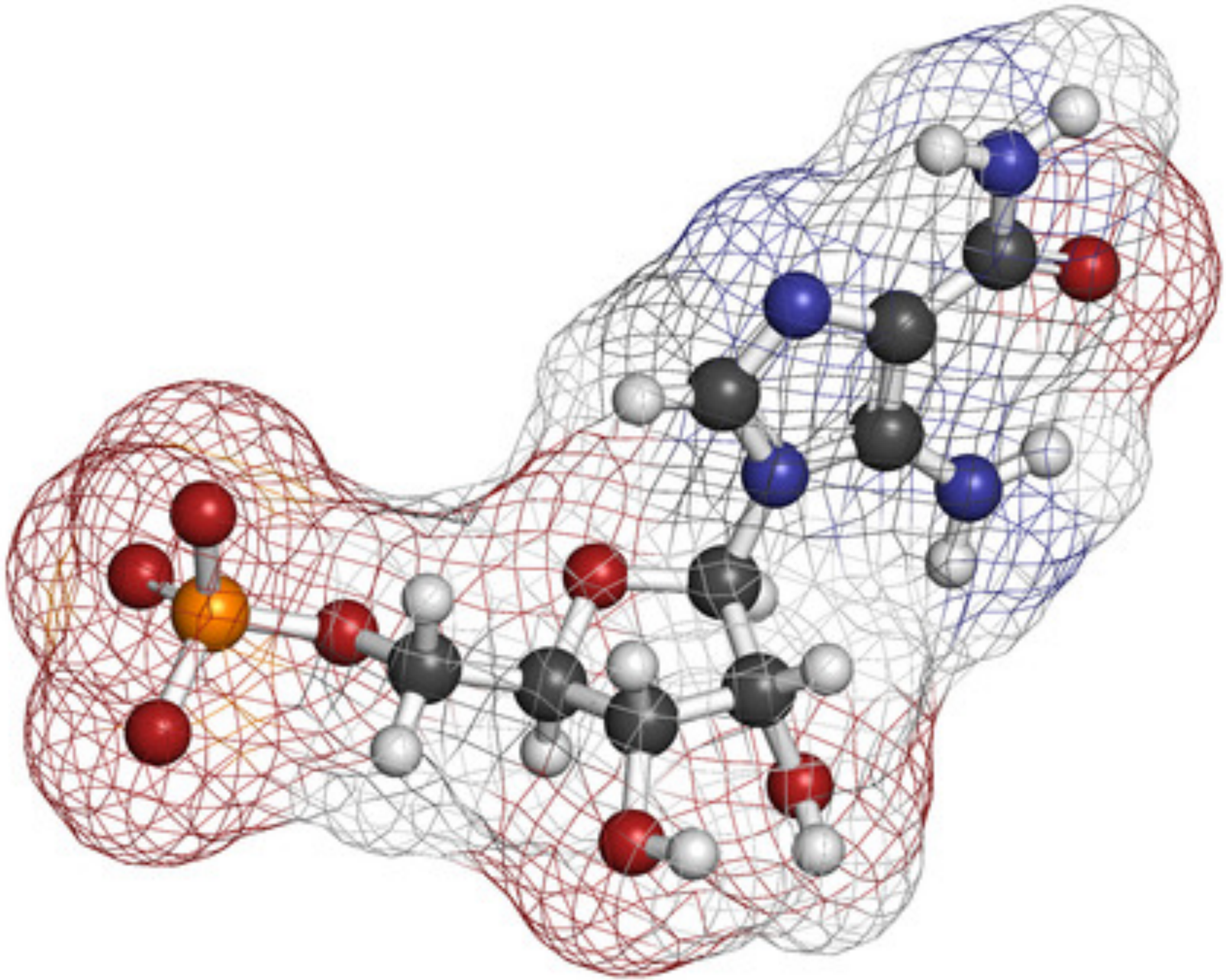
Зависимость

Виды дескрипторов



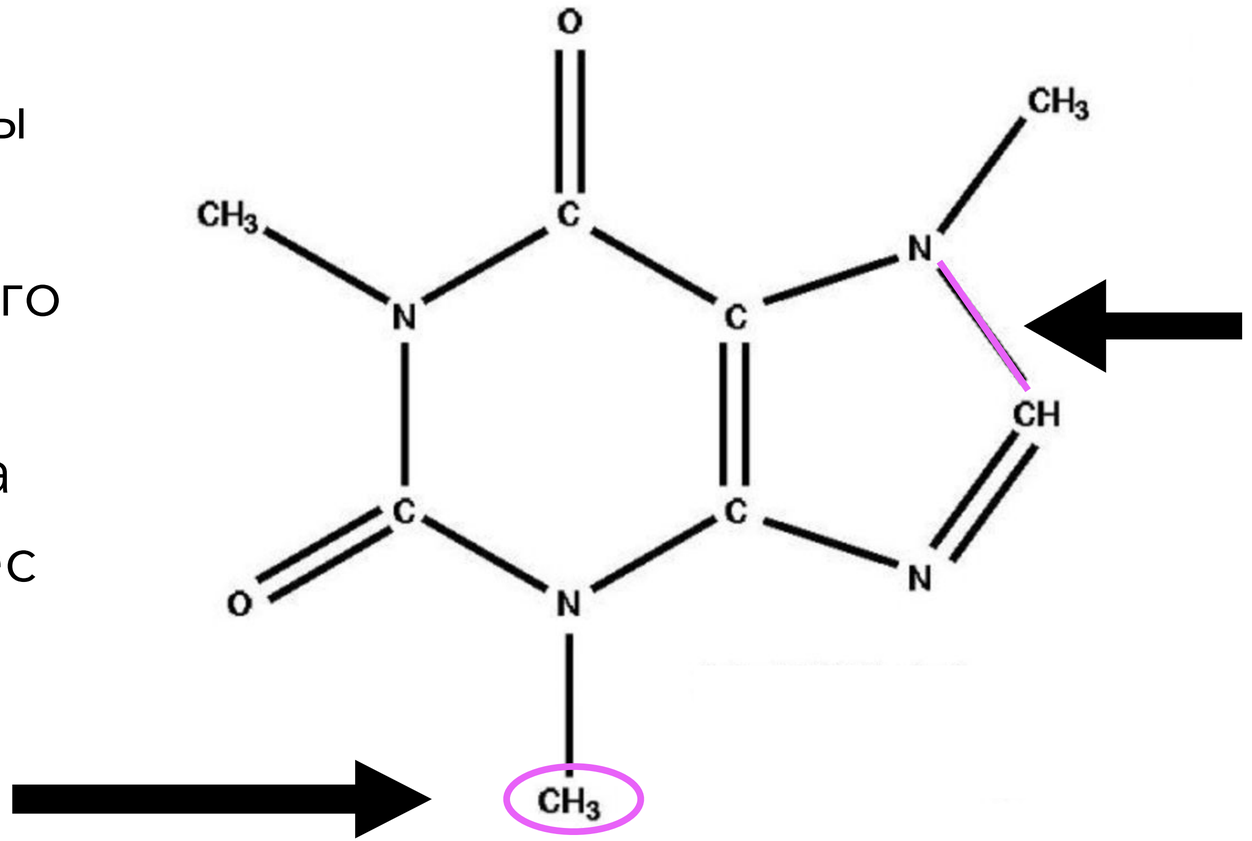
- Дескрипторы элементарного уровня
- Дескрипторы структурной формулы
- Дескрипторы электронного уровня
- Дескрипторы межмолекулярных взаимодействий

3D-QSAR



Конкретная постановка задачи

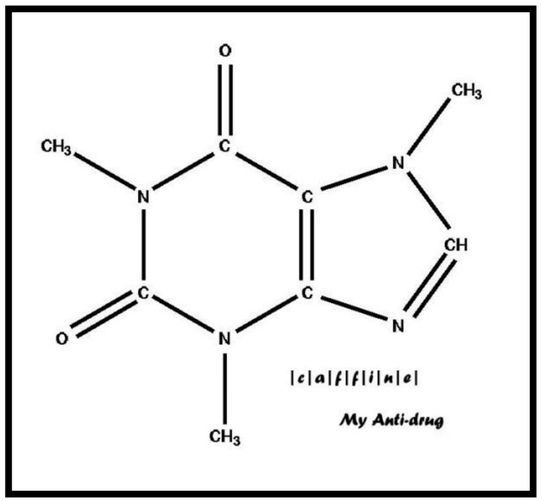
- координаты
- символ химимческого элемента
- заряд ядра
- атомный вес
- атомный радиус

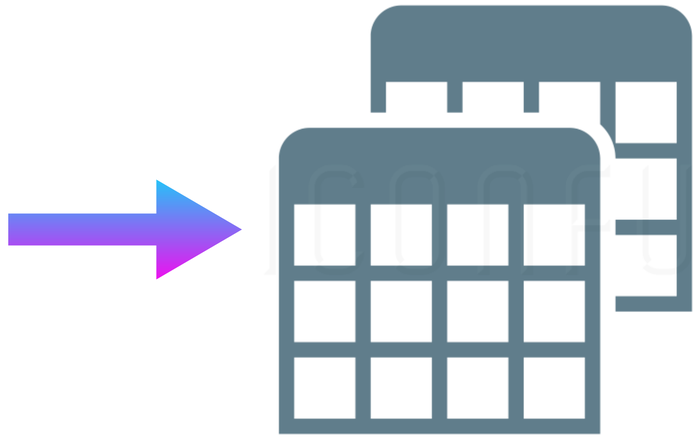


- кратность
- длина
- порядок связей

Конкретная постановка задачи

N x



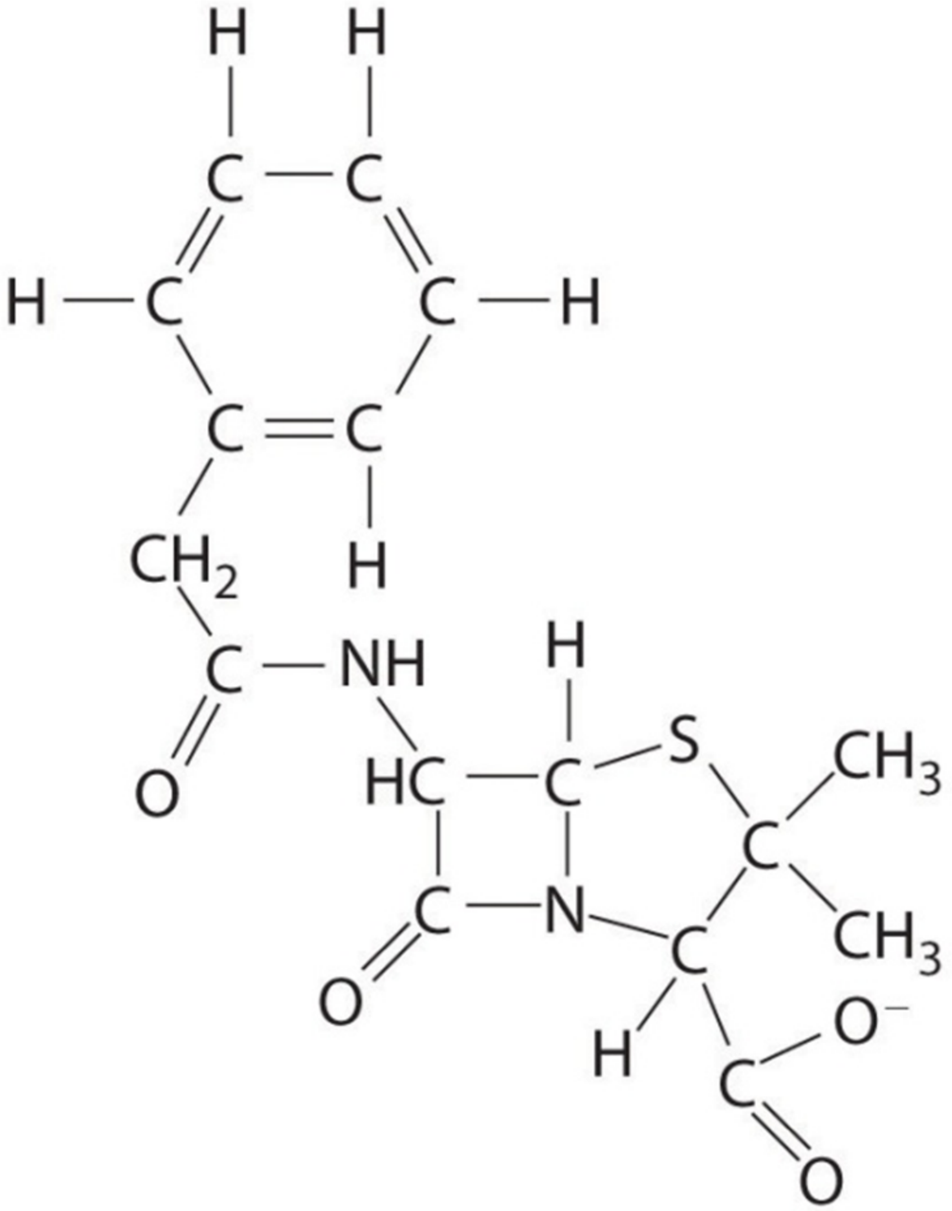


$F(x)$

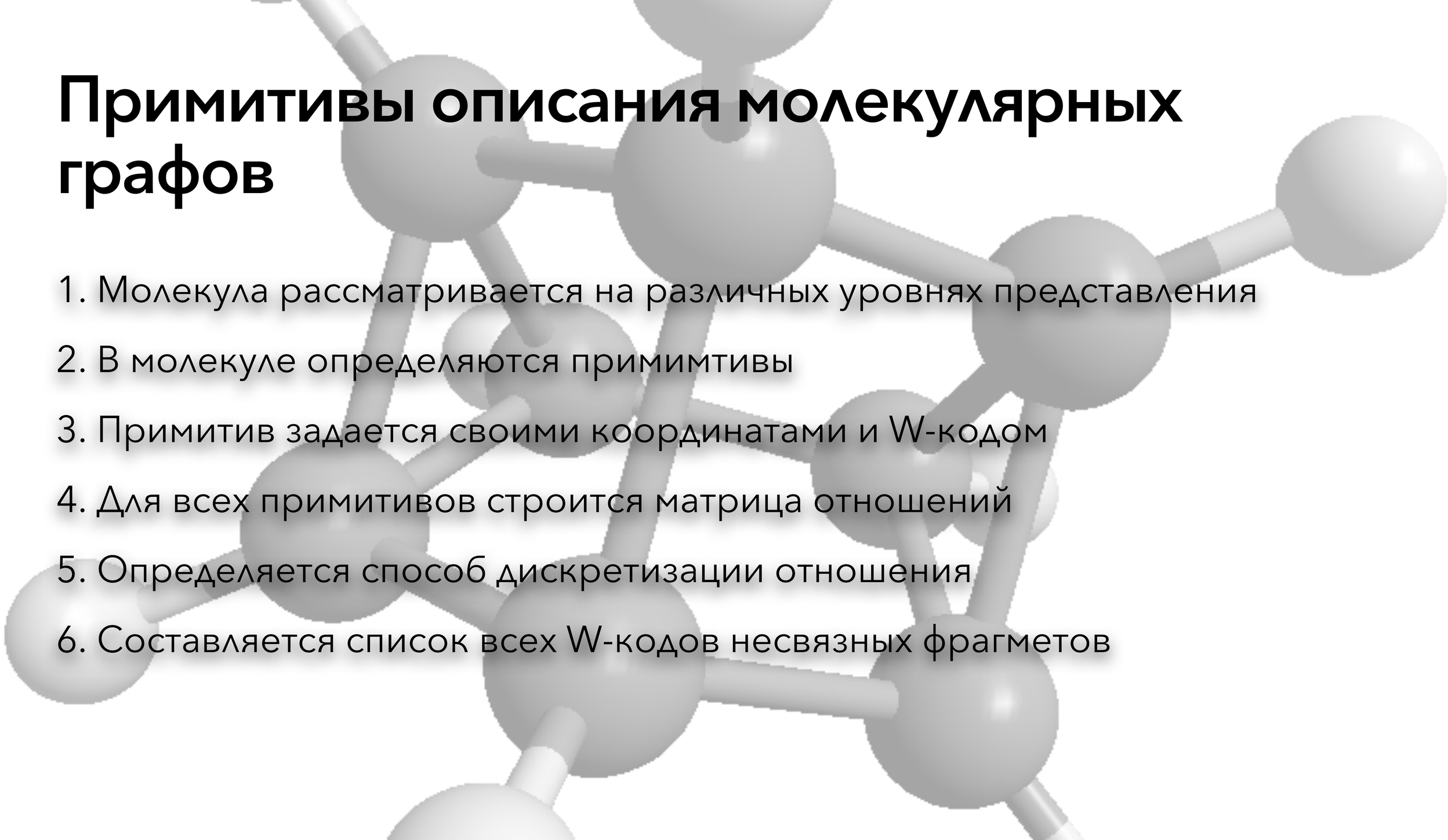
Способ построения матрицы Молекула - Дескриптор

NNpbr:

- NN - имя атом;
- p - степень атома
- b - тип связи
- r - маркер положения атома в кольцевой системе



Примитивы описания молекулярных графов



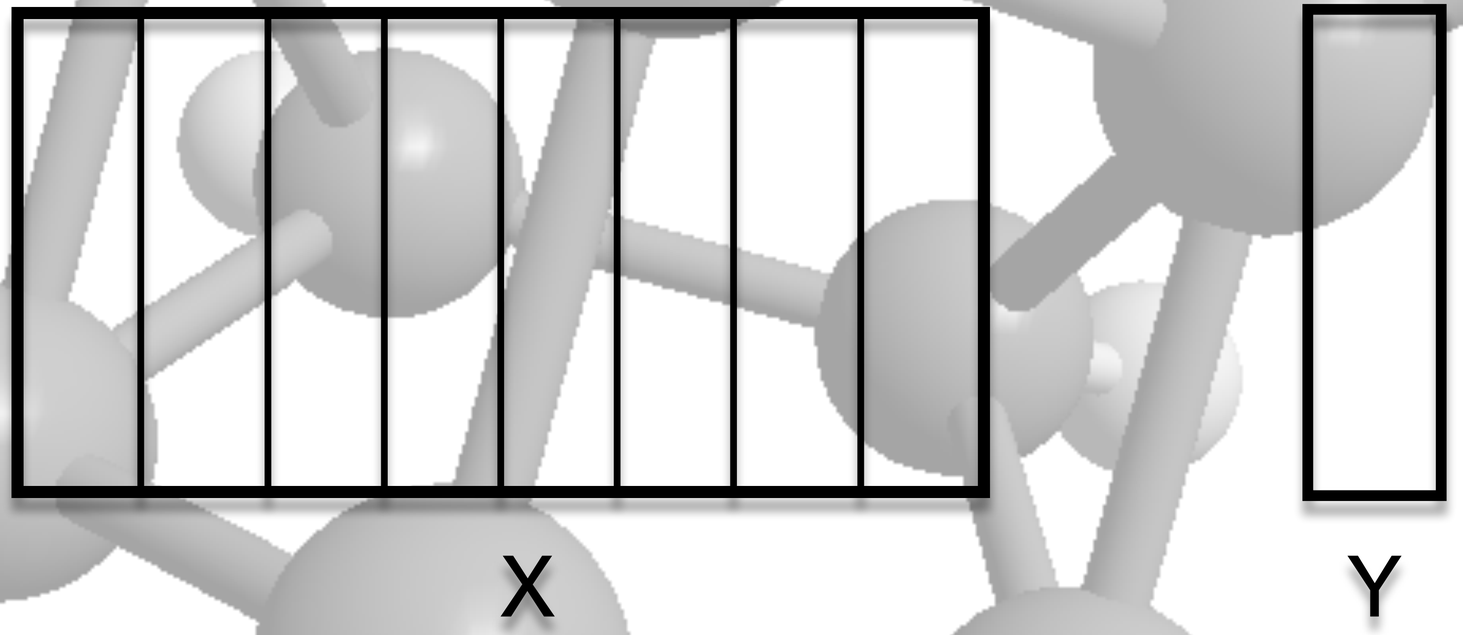
1. Молекула рассматривается на различных уровнях представления
2. В молекуле определяются прими́мтивы
3. Примитив задается своими координатами и W-кодом
4. Для всех примитивов строится матрица отношений
5. Определяется способ дискретизации отношения
6. Составляется список всех W-кодов несвязных фрагметов

Поиск функциональной зависимости

$$Ax = b$$

	0	1	2	3	4	5	6	7	8	9	...	167	168	169	170	171	172	173	174	175	176
0	0.0	0.0	3.0	0.0	2.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	3.0	0.0	3.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	4.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Метод Группового Учета Аргументов



Метод Группового Учета Аргументов

$$Cor(a,b) = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{(\sum a_i^2)(\sum b_i^2)}}$$

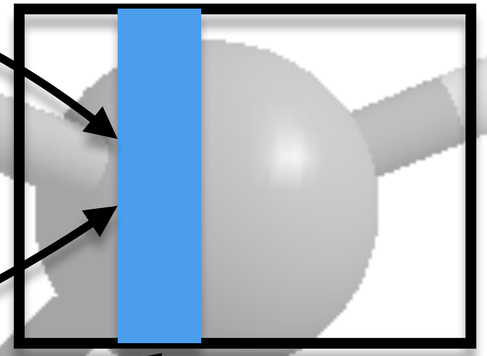
$$R^2(Y,\hat{Y}) = 1 - \frac{\sum_{i=1}^{N} \varepsilon_i^2}{\sum (Y_i - Y_{cp})^2}$$

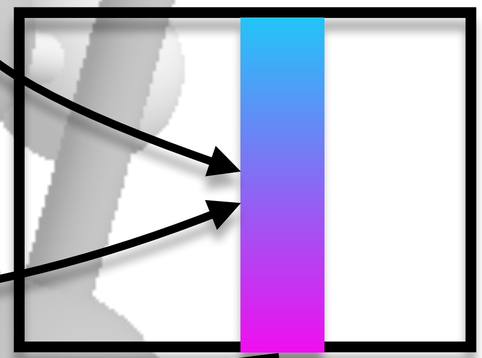
Параметры:

- Q - размер буфера
- C - порог корреляции
- I - количество итераций

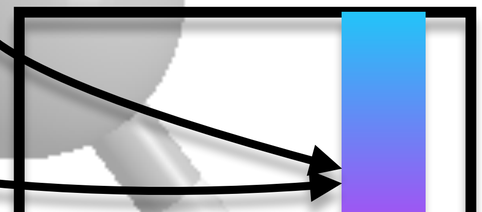
Метод Группового Учета Аргументов

X

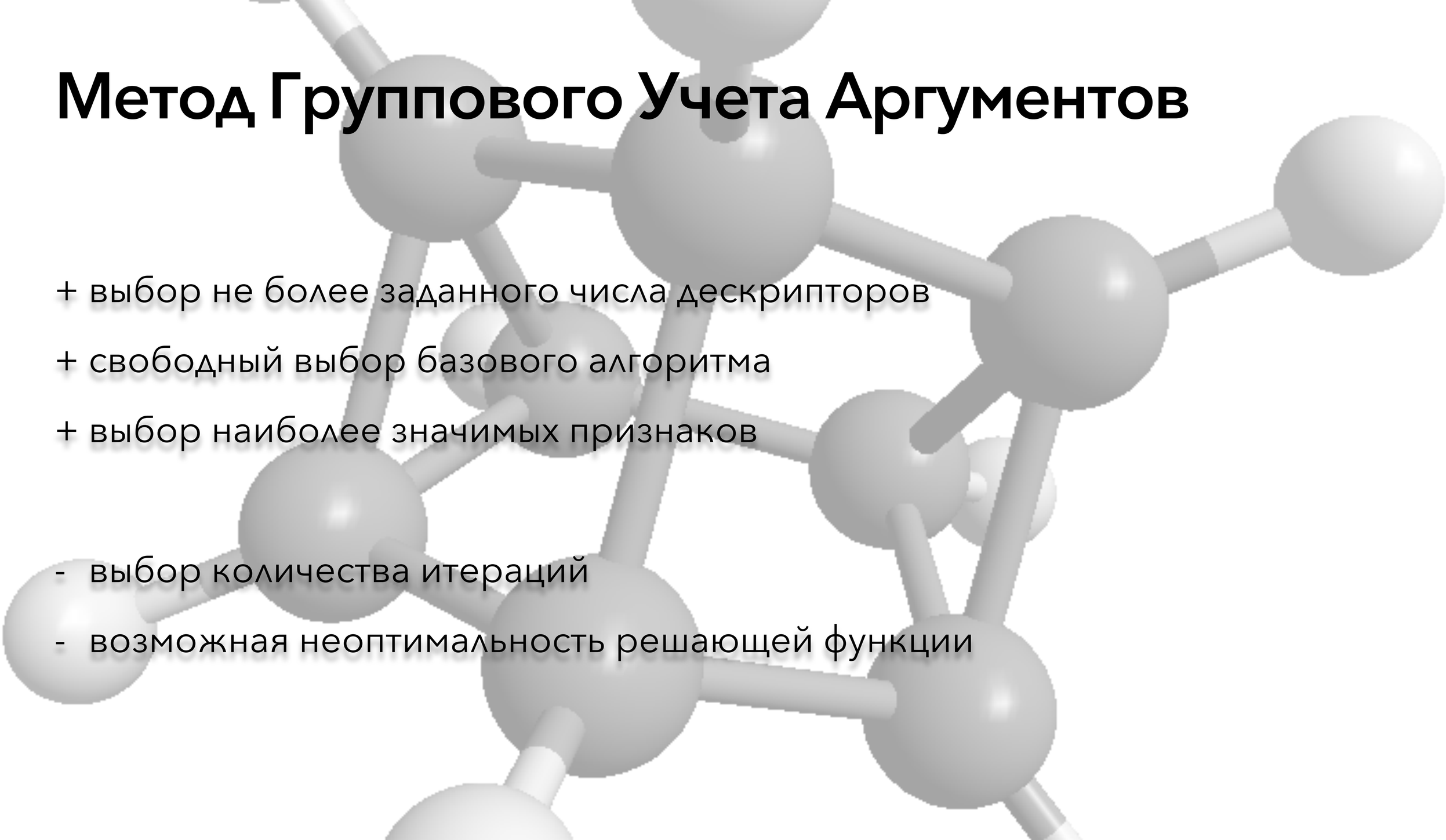




Буфер



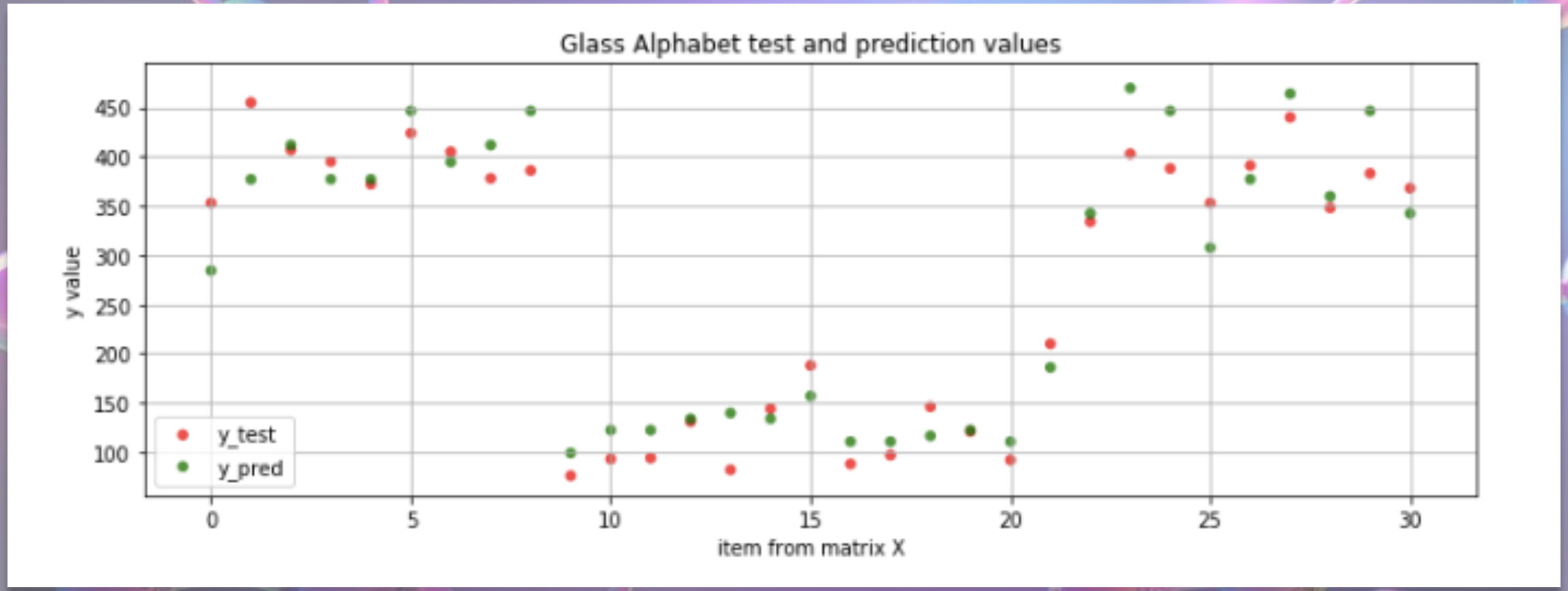
Метод Группового Учета Аргументов



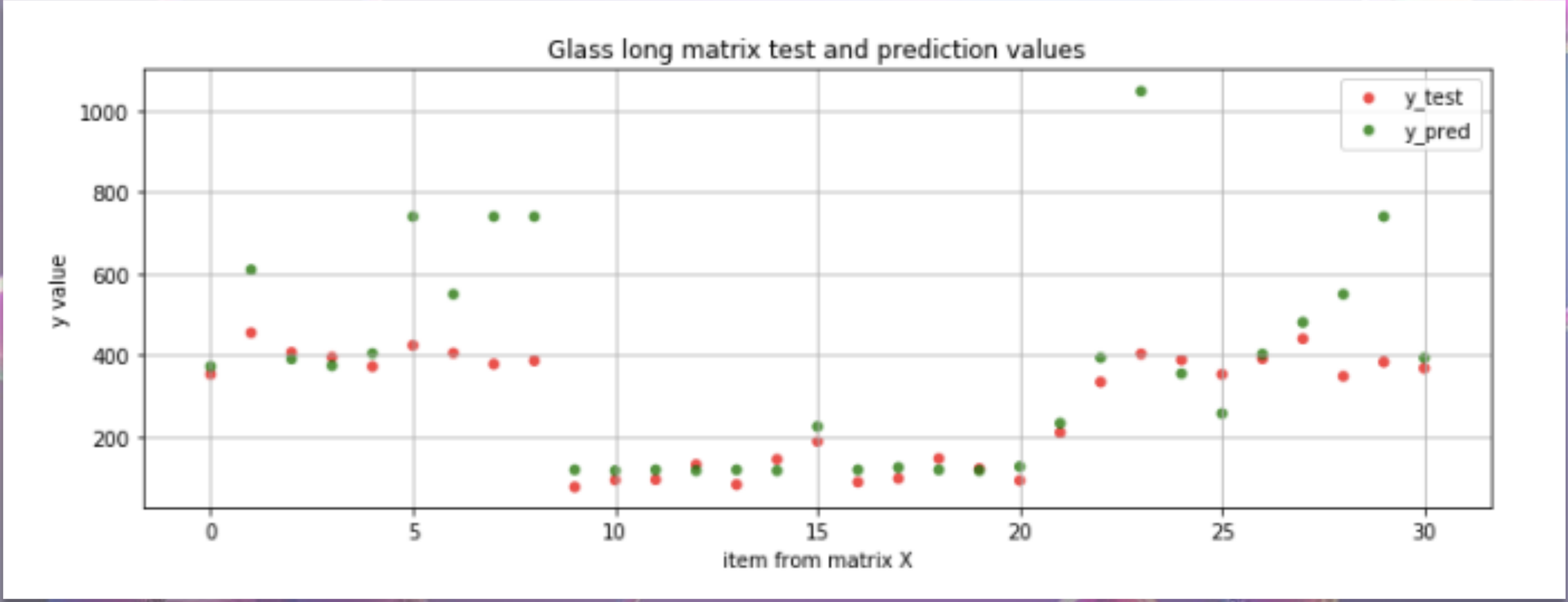
+ выбор не более заданного числа дескрипторов
+ свободный выбор базового алгоритма
+ выбор наиболее значимых признаков

- выбор количества итераций
- возможная неоптимальность решающей функции

Результаты. Задача регрессии



Результаты. Задача регрессии



Результаты. Классификация с использованием кластеризации

	алгоритм кластер-анализа			precision	recall	f1-score
bzr NNdb*	к-средних	CLUSTER # 0	-1	0.91	1.00	0.95
			1	0.00	0.00	0.00
			accuracy	0.91		
		CLUSTER # 1	-1	0.71	0.92	0.80
			1	0.50	0.18	0.27
			accuracy	0.69		
		CLUSTER # 2	-1	0.78	0.60	0.68
			1	0.71	0.86	0.78
			accuracy	0.74		
	DBSCAN	CLUSTER # -1 (отказ)	-1	0.89	0.89	0.89
			1	0.86	0.86	0.86
			accuracy	0.88		
		CLUSTER # 0	-1	0.66	0.79	0.72
			1	0.65	0.48	0.55
			accuracy	0.65		
		CLUSTER # 1	-1	0.95	1.00	0.98
			1	0.00	0.00	0.00
			accuracy	0.95		
		CLUSTER # 2	-1	0.62	0.71	0.67
			1	0.75	0.67	0.71
			accuracy	0.69		

Выводы

- Рассмотрена задача QSAR
- Показаны способы описания и маркировки дескрипторов и формирования матрицы Молекула-Признак
- Описан метод группового учета аргументов и показаны результаты его работы
- Исследован метод группового учета аргументов в задаче классификации для сбалансированных выборок (при прогнозировании на всех молекулах) и несбалансированных (в кластерах)

Спасибо за внимание!

Vasilyeva Varvara, 2021