

В работе использовались четыре датасета, содержащих значений константы диссоциации и сорока четырёх дескрипторов для реакций с четырьмя таргетными белками.

Таблица 1. Базовая статистика значений константы ингибирования

Таргет	Число измерений	Среднее значение	Медиана	Стандартное отклонение
cb2	5948	2393.951	158.245	7549.403
b2ar	821	1260362	250	34899865
ckr5	215	325.4782	11	2428.429
drd1	1703	4266.803	360x	19781.7

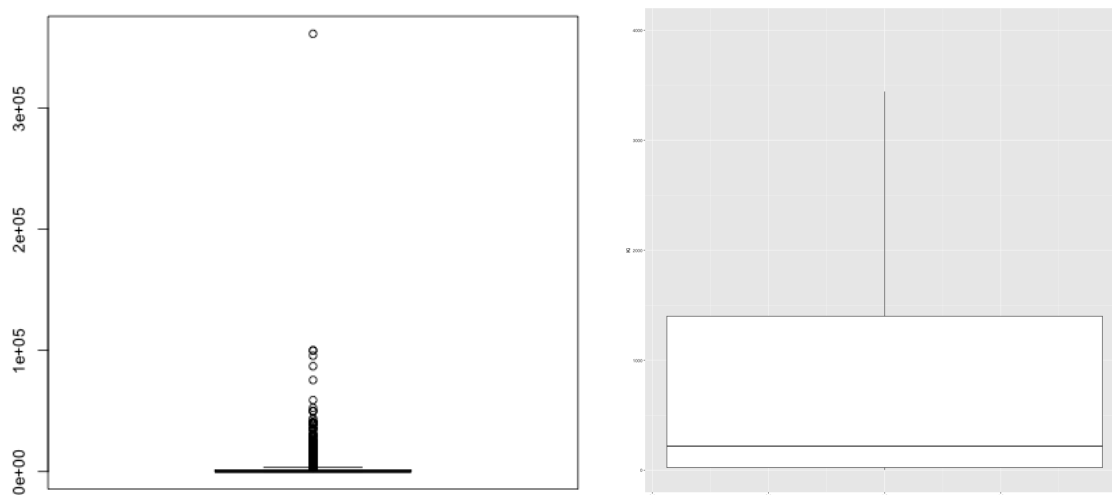


Рисунок 1. Box-plots распределения данных в датасете cb2 с учётом выпадающих значений (слева) и без них (справа).

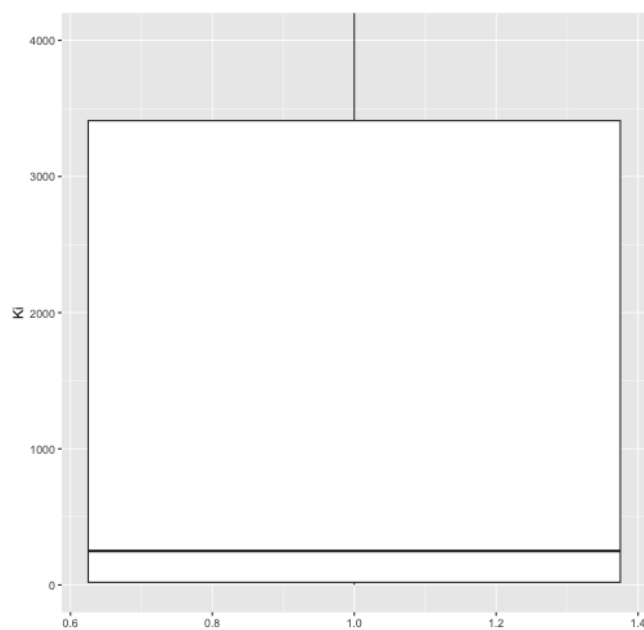


Рисунок 2. Box-plot распределения данных для b2ar без учета выпадающих значений

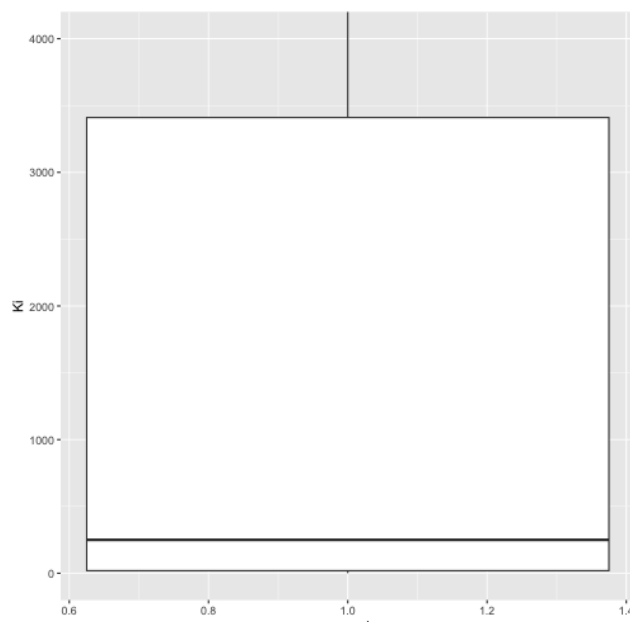


Рисунок 3. Box-plot распределения данных для skr5 без учета выпадающих значений

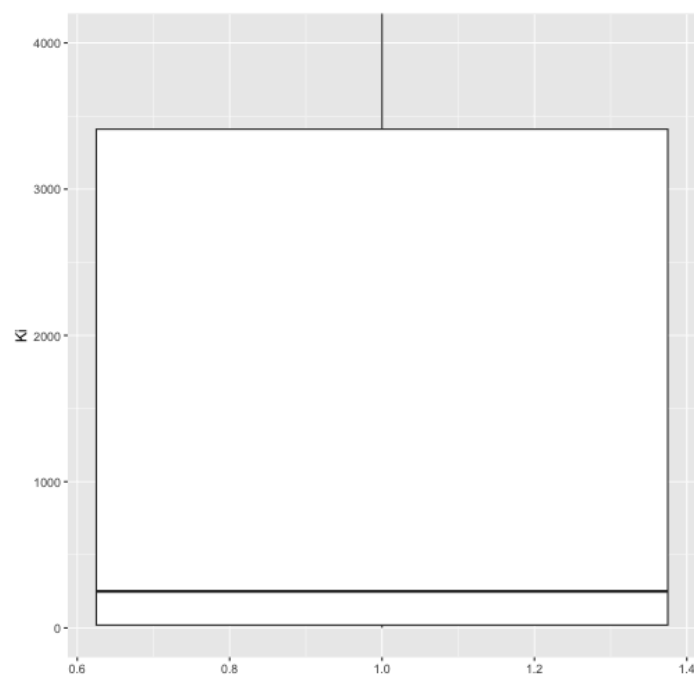


Рисунок 4. Box-plot распределения данных для drd1 без учета выпадающих значений

Тесты на нормальность (тест Колмогорова-Смирнова):

data: cb2\_cor\$Ki  
D = 0.37558, p-value < 2.2e-16

data: b2ar\_cor\$Ki  
D = 0.49323, p-value < 2.2e-16

data: ckr5\_cor\$Ki  
D = 0.44671, p-value < 2.2e-16

data: drd1\_cor\$Ki  
D = 0.41461, p-value < 2.2e-16

Можно сделать вывод, что среди всех целевых белков константа ингибирования не распространена нормально.

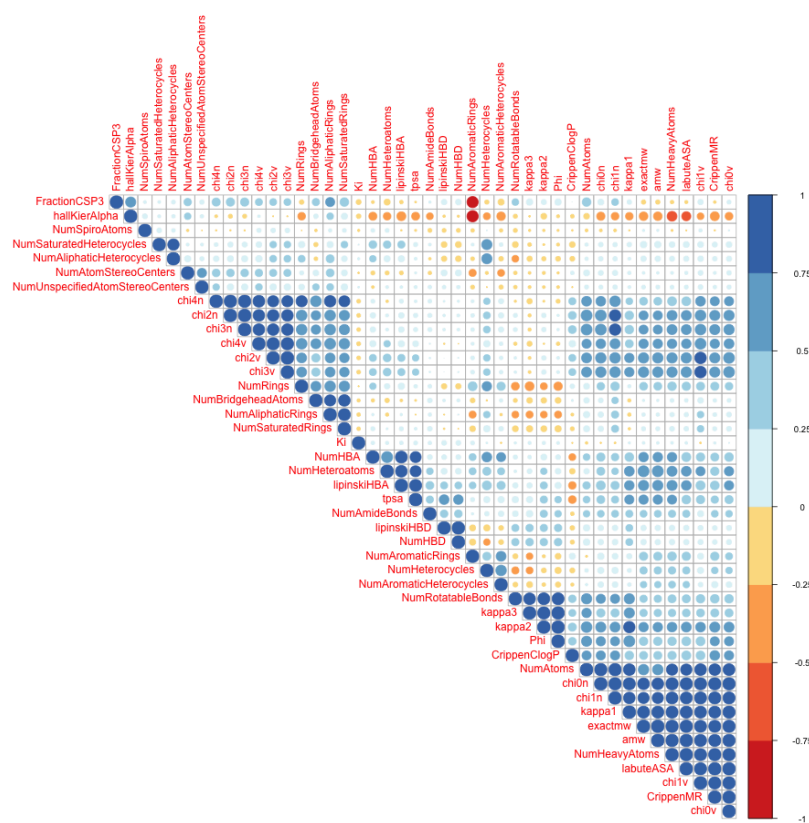


Рисунок 5. Коррелограмма для всех дескрипторов и  $K_i$  в датасете cb2.

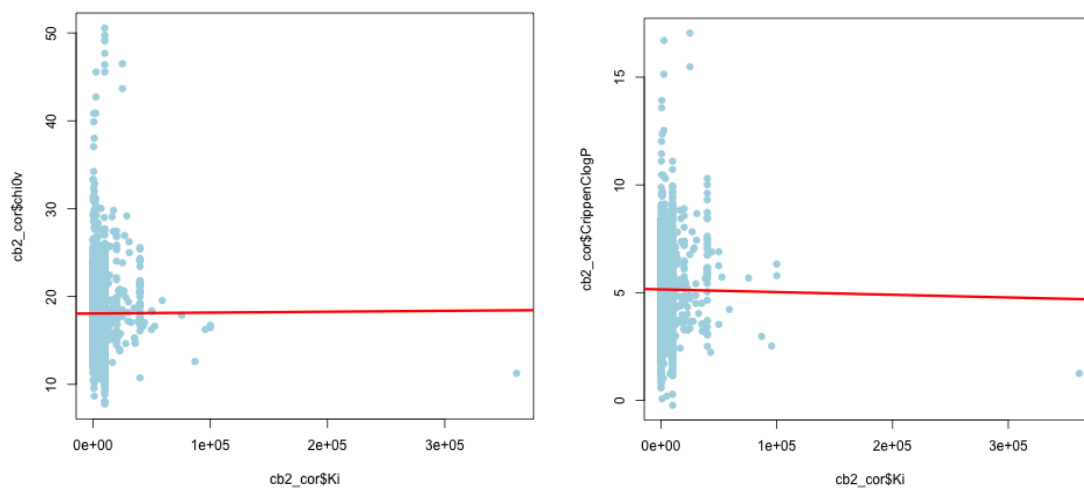


Рисунок 6. Примеры графиков корреляции  $K_i$  с некоторыми из дескрипторов (cb2)

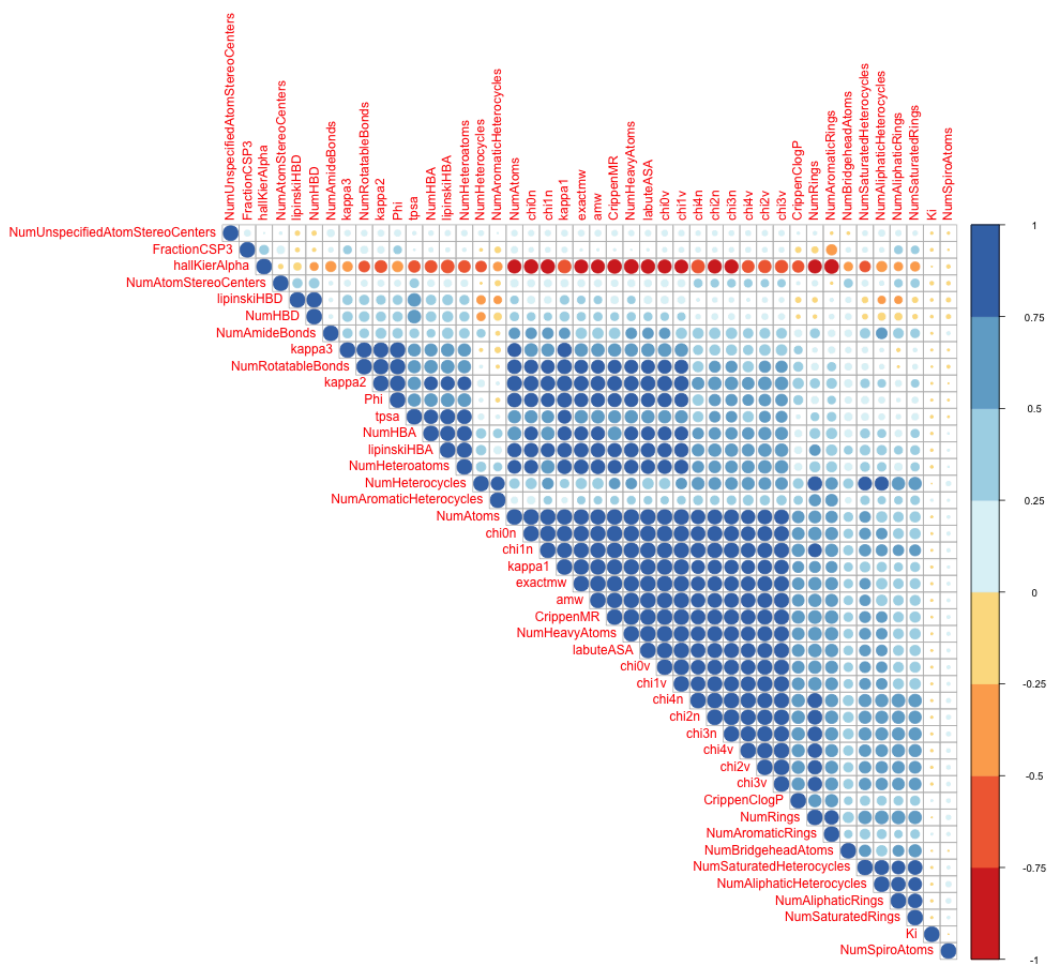


Рисунок 7. Коррелограмма для всех дескрипторов и  $K_i$  в датсете b2ar.

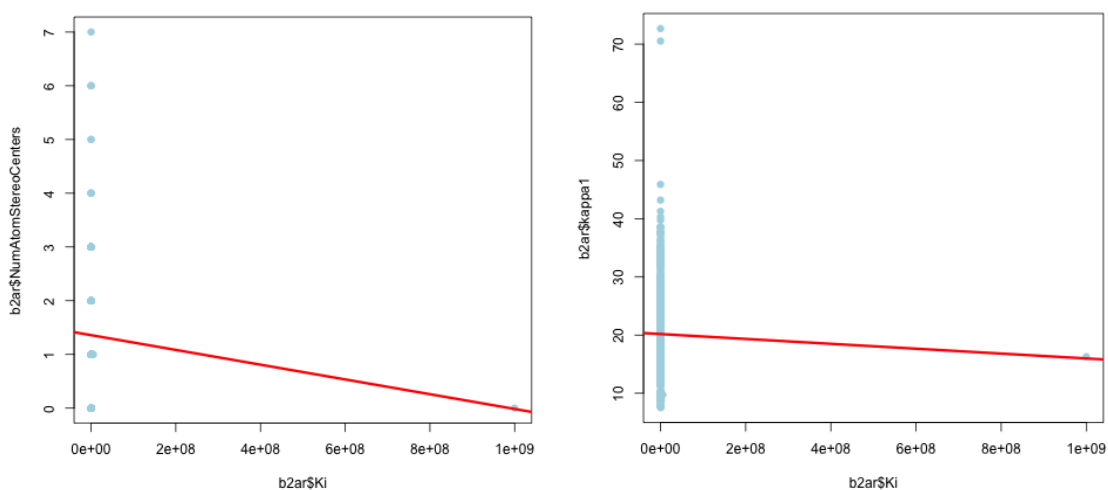


Рисунок 8. Примеры графиков корреляции  $K_i$  с некоторыми из дескрипторов (b2ar)

Отсутствие сильной корреляции каких-либо дескрипторов с целевой переменной говорит о целесообразности подбора других дескрипторов для решения задачи.

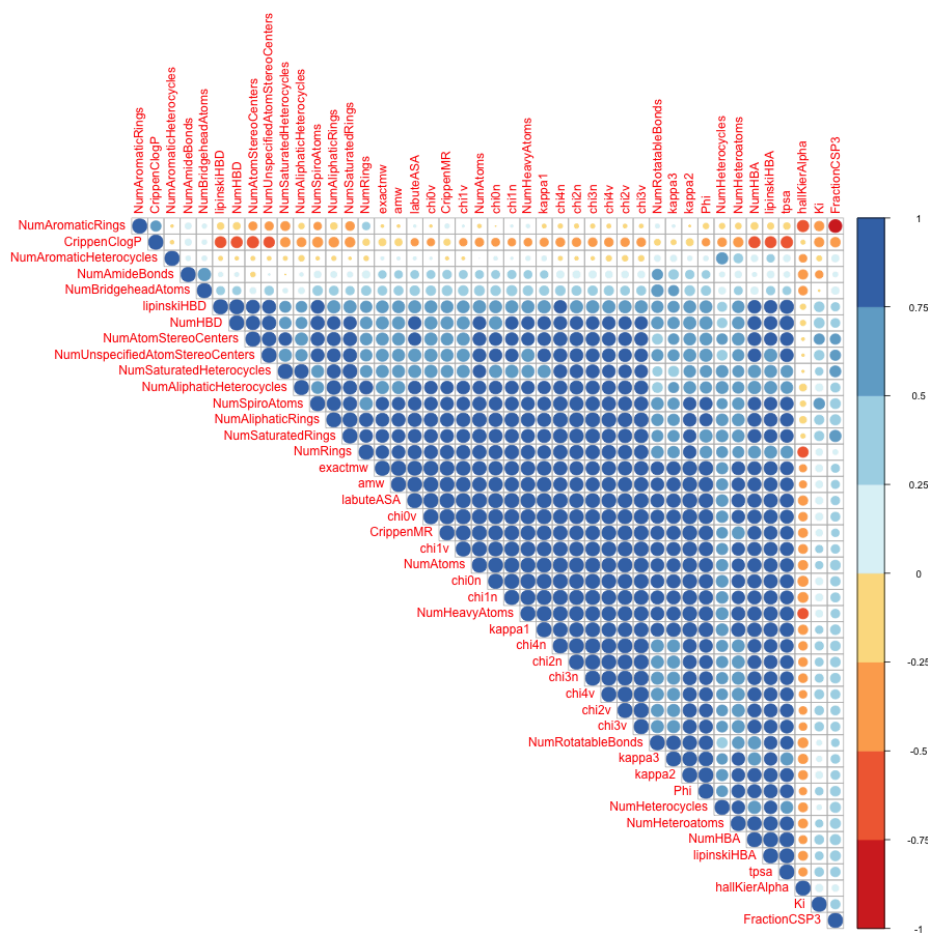


Рисунок 9. Коррелограмма для всех дескрипторов и Ki в датасете skr5.

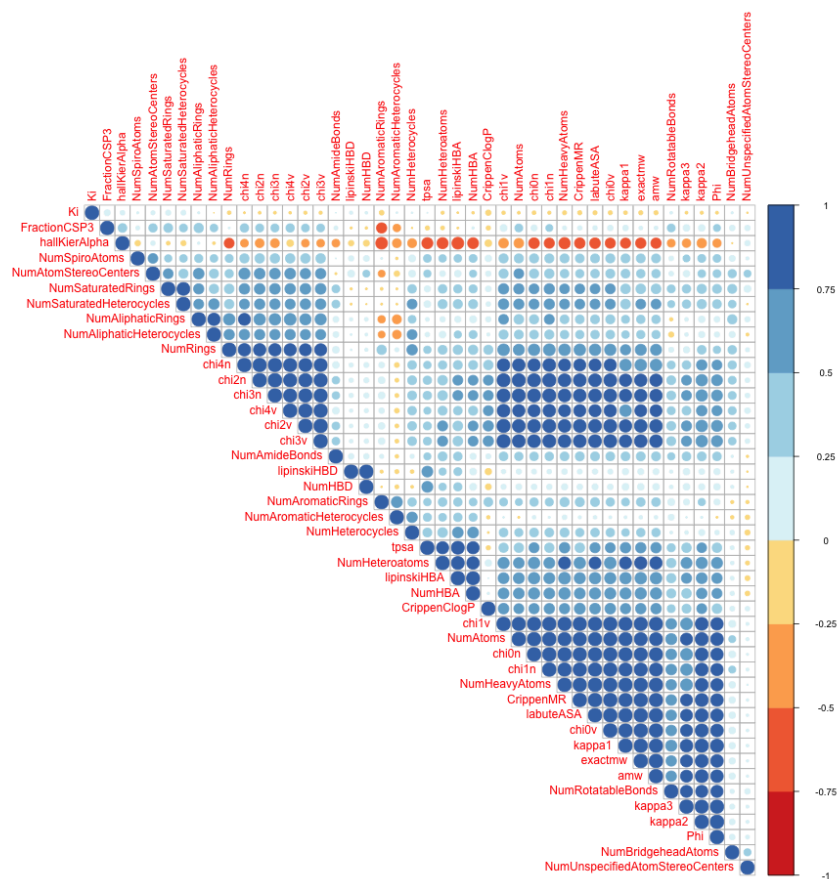


Рисунок 7. Коррелограмма для всех дескрипторов и Ki в датасете skr5.