ИБМХ
Федеральное государственное бюджетное научное учреждение
«Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича»

Филимонов Д.А., Лагунин А.А., Погодин П.В., Поройков В.В.

PASS AFFINITIES ИЛИ ДЕЙСТВИТЕЛЬНО ЛИ НАИВЕН МЕТОД ПОЛУЧЕНИЯ (Q)SAR ОЦЕНОК «NAIVE BAYES»?

XXIII Российский национальный конгресс “Человек и лекарство”
XXII Симпозиум “Биоинформатика и компьютерное конструирование лекарств”

IBMC
Institute of Biomedical Chemistry
Moscow, Russia

Программа

- Почему Байес «наивный»?
- (Q)SAR оценки «Naïve Bayes»
- Молекулярная биофизика
- Не наивный «наивный» Байес
- Биофизика и «теория катастроф»
- «Размерность активности»
- PASS Affinities
- Биофизика, теория катастроф и QSAR
- Выводы

Формула Байеса

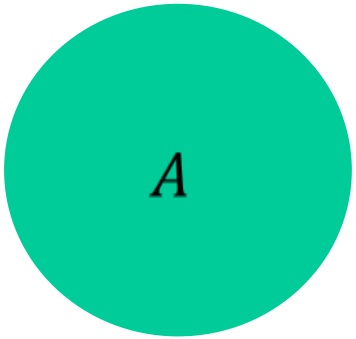
Вероятность $P(A|C)$ того, что органическое соединение C с дескрипторами $D_1, D_2, \ldots, D_m$ имеет активность A:

$$P(A|C) = P(A|D_1, D_2, \ldots, D_m) = \frac{P(D_1, D_2, \ldots, D_m|A) \cdot P(A)}{P(C)}$$

$P(D_1, D_2, \ldots, D_m|A)$ – условная вероятность того, что соединение с активностью A имеет дескрипторы $D_1, D_2, \ldots, D_m$;

$P(A)$ – априорная вероятность активности A;

$P(C) = P(D_1, D_2, \ldots, D_m)$ – априорная вероятность соединения C.

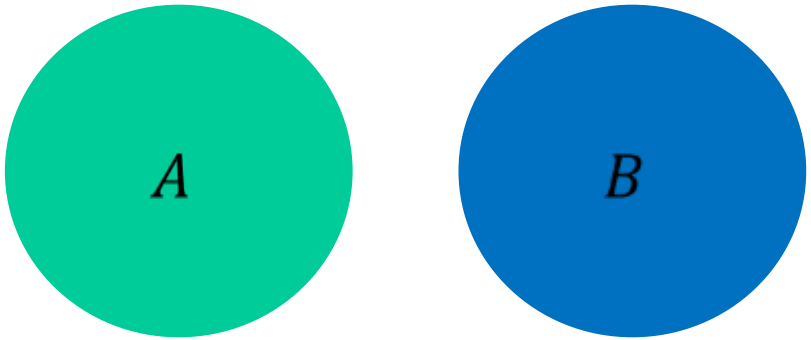


Формула Байеса

$$P(A|D_1, D_2, \ldots, D_m) = \frac{P(D_1, D_2, \ldots, D_m|A) \cdot P(A)}{P(C)}$$

$$P(B|D_1, D_2, \ldots, D_m) = \frac{P(D_1, D_2, \ldots, D_m|B) \cdot P(B)}{P(C)}$$

$$\frac{P(A|D_1, D_2, \ldots, D_m)}{P(B|D_1, D_2, \ldots, D_m)} = \frac{P(D_1, D_2, \ldots, D_m|A) \cdot P(A)}{P(D_1, D_2, \ldots, D_m|B) \cdot P(B)}$$

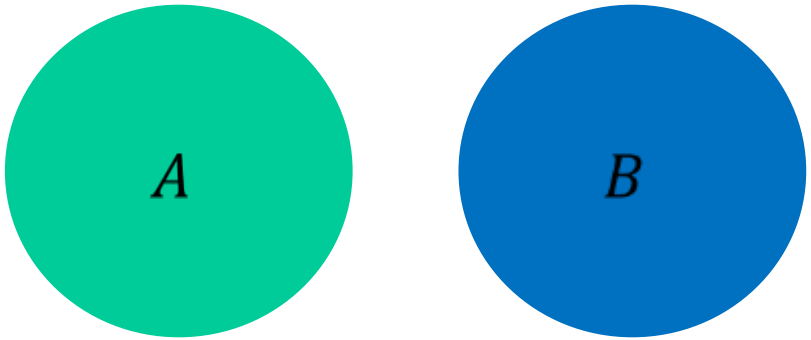


Байес «наивный» - Naïve Bayes

$$P(D_1, \ldots, D_m|A) \cong \prod_{i=1}^{m} P(D_i|A)$$

$$P(D_1, \ldots, D_m|B) \cong \prod_{i=1}^{m} P(D_i|B)$$

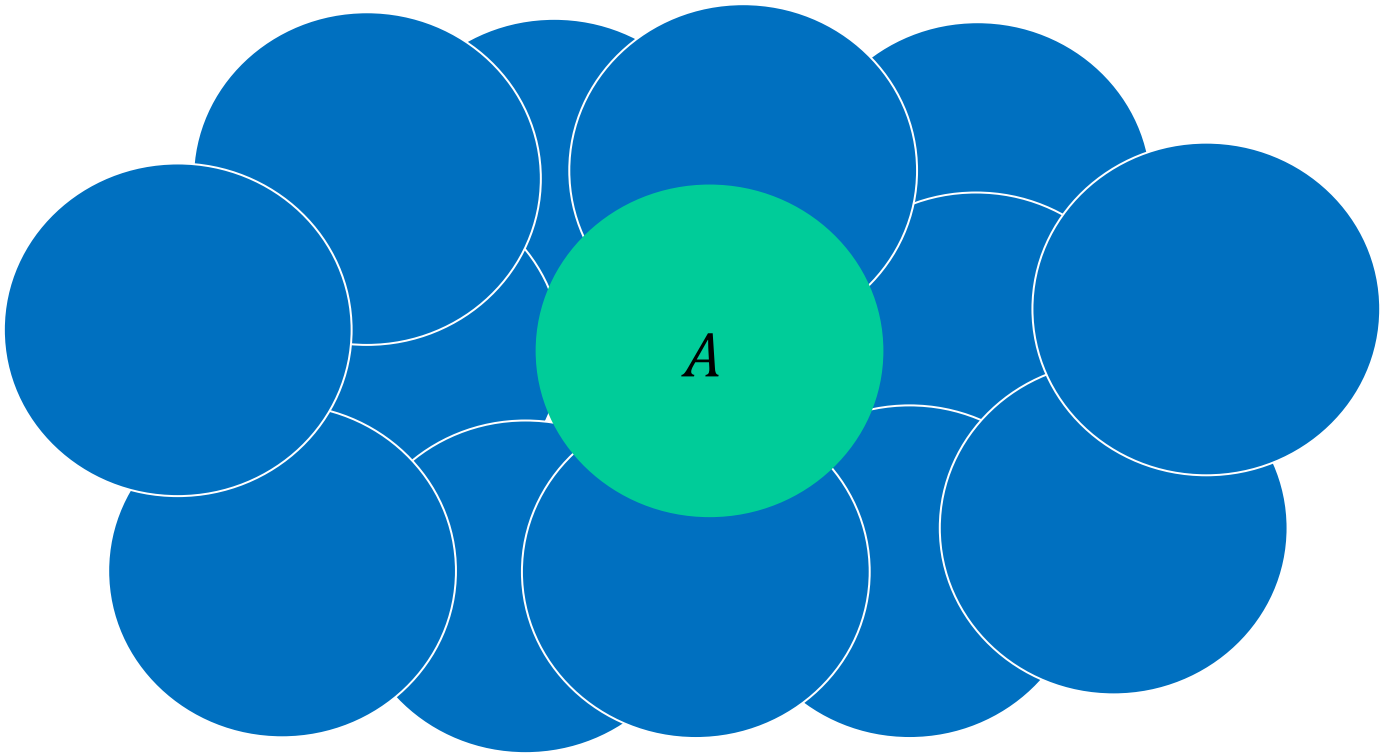
$$ln\left[\frac{P(A|D_1, D_2, \ldots, D_m)}{P(B|D_1, D_2, \ldots, D_m)}\right] \cong ln\left[\frac{P(A)}{P(B)}\right] + \sum_i ln\left[\frac{P(D_i|A)}{P(D_i|B)}\right]$$



(Q)SAR оценки «Naïve Bayes»

$$ln\left[\frac{P(A|C)}{1-P(A|C)}\right] \cong ln\left[\frac{P(A)}{1-P(A)}\right] + \sum_{i=1}^{m}\left\{ln\left[\frac{P(A|D_i)}{1-P(A|D_i)}\right] - ln\left[\frac{P(A)}{1-P(A)}\right]\right\}$$

$$= a_0 + \sum_i a_i d_i(C)$$



Молекулярная биофизика - аффинность

Результат молекулярного узнавания:

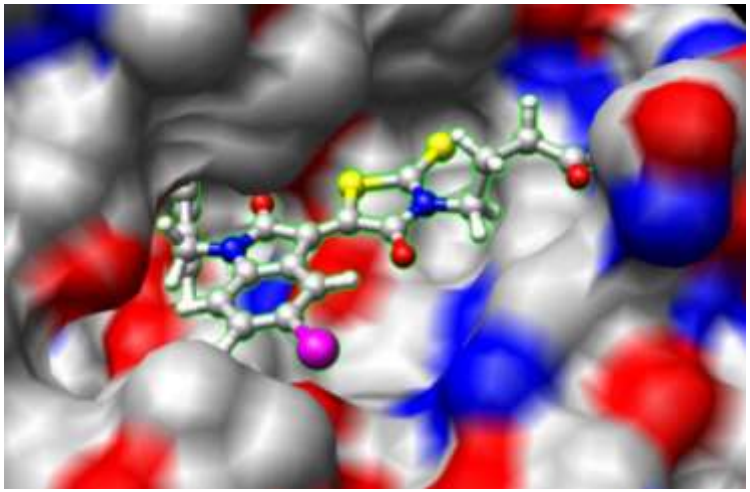
$$L + P = LP + \Delta G$$

Константа диссоциации – иное представление аффинности:

$$K_d = \frac{[L][P]}{[LP]}$$

Прямое соответствие между ΔG и K_d:

$$\Delta G = -RTln(K_d)$$



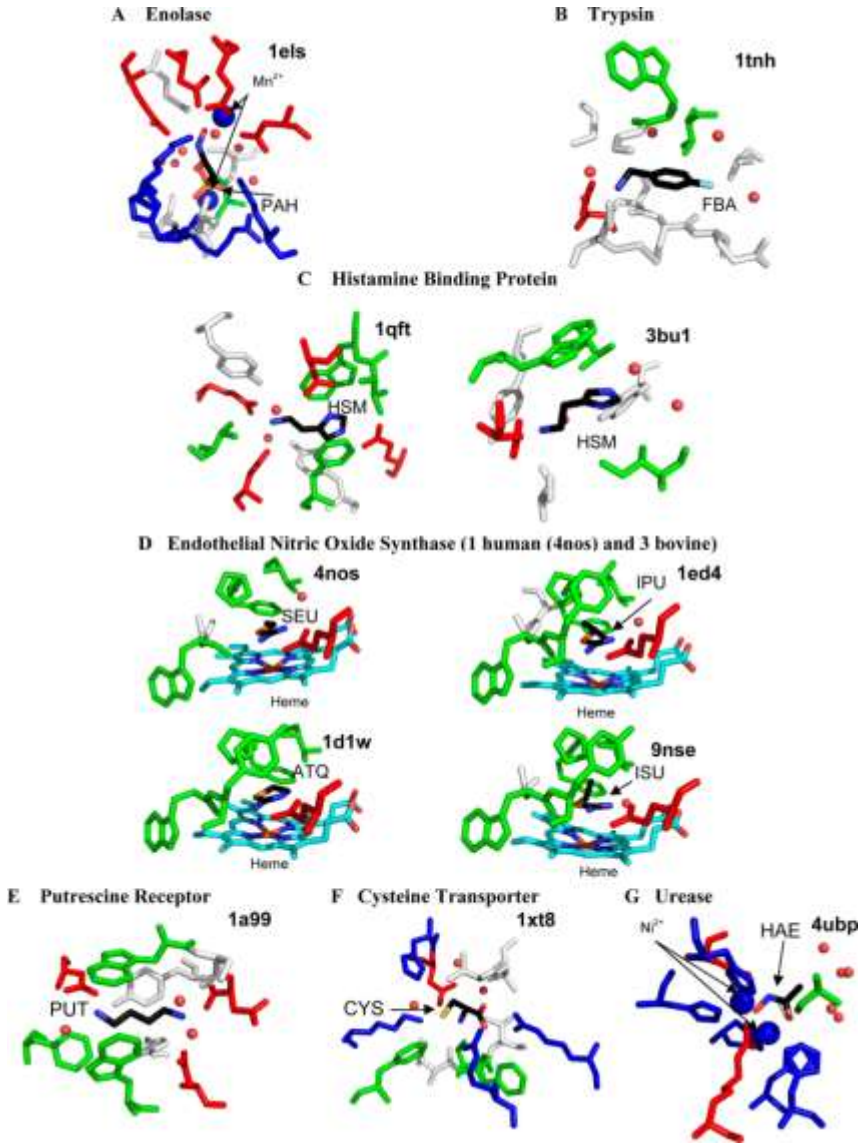
Молекулярная биофизика - аффинность

$$P(A) = Pr\{\Delta G > \Delta G_*\}$$

$$\Delta G = \sum_i \Delta G_i$$

$$P(A|D_i) = Pr\{\Delta G + \Delta G_i > \Delta G_*\}$$
$$= Pr\{\Delta G > \Delta G_* - \Delta G_i\}$$

$$P(A) = 1 - F_A(\Delta G_*)$$
$$P(A|D_i) = 1 - F_A(\Delta G_* - \Delta G_i)$$



R. D. Smith, A. L. Engdahl, J. B. Dunbar, Jr., and H. A. Carlson, *J. Chem. Inf. Model.*, 2012, 52, 2098−2106.

Не наивный «наивный» Байес

$$\Delta G_* = F_A^{-1}\big(1 - P(A)\big)$$

$$\Delta G_* - \Delta G_i = F_A^{-1}\big(1 - P(A|D_i)\big)$$

$$\Delta G = \sum_i \Delta G_i = \sum_i \Big[F_A^{-1}\big(1 - P(A)\big) - F_A^{-1}\big(1 - P(A|D_i)\big)\Big]$$

Наиболее привычно считать, что $F_A(*)$ – нормальное.

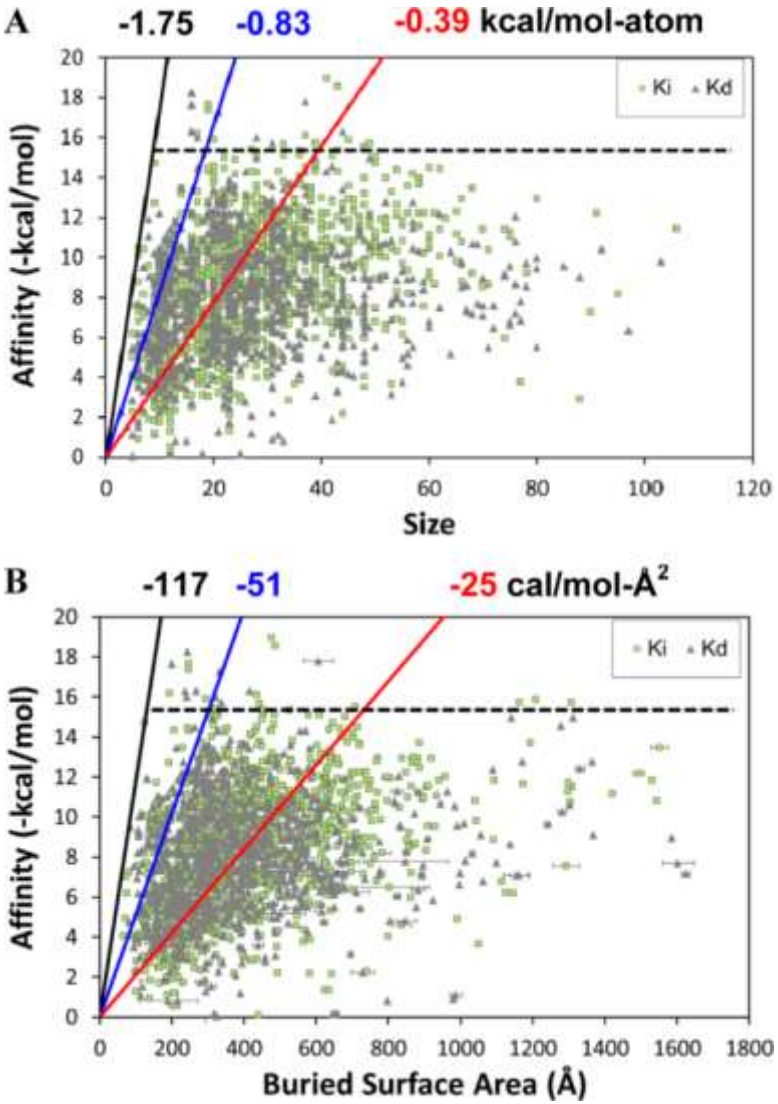
Аппроксимация – распределение Ферми оно же логистическое:

$$\Phi(x) = \frac{1}{1 + Exp(-1.6x)}$$

Немедленно получаем:

$$\Delta G_i \sim ln\left[\frac{P(A|D_i)}{1 - P(A|D_i)}\right] - ln\left[\frac{P(A)}{1 - P(A)}\right]$$

Биофизика и «теория катастроф»



Среднее
- 0.39 ккал/моль·атом
- 0.017 эВ/атом

Медиана
- 0.34 ккал/моль·атом
- 0.015 эВ/атом

Предел
- 1.75 ккал/моль·атом
- 0.076 эВ/атом

Максимум
- 15 ккал/моль
- 0.651 эВ
- K_d = 5 пМ

R. D. Smith, A. L. Engdahl, J. B. Dunbar, Jr., and H. A. Carlson, *J. Chem. Inf. Model.*, 2012, 52, 2098−2106.

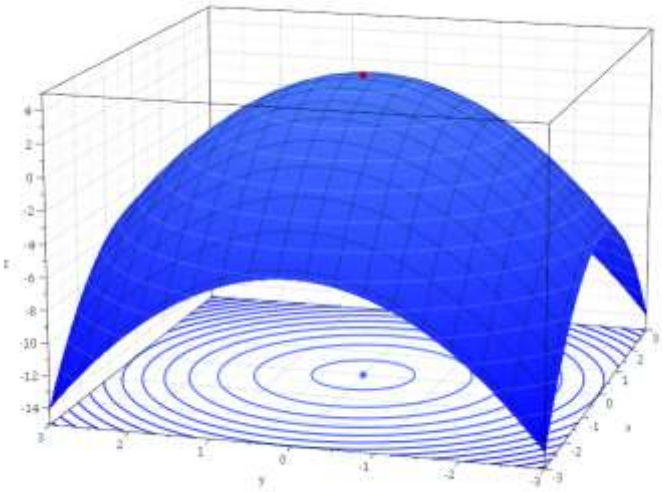
Биофизика и «теория катастроф»

В силу леммы Морса ΔG в окрестности максимума:

$$\Delta G = \Delta G_{max} - \sum_i Q_i^2 = \Delta G_{max} - R^2$$

Q_i – «расстояния» от точки максимума

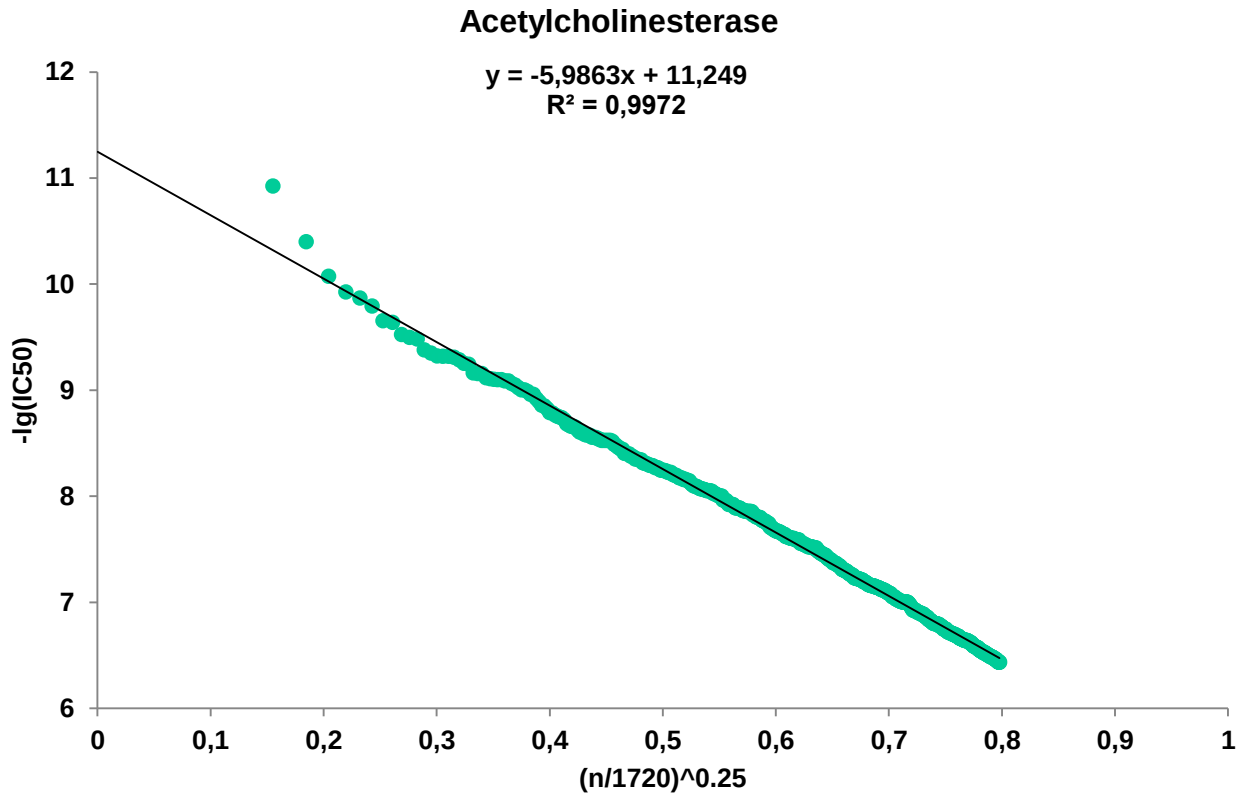
R – радиус сечения многомерного параболоида



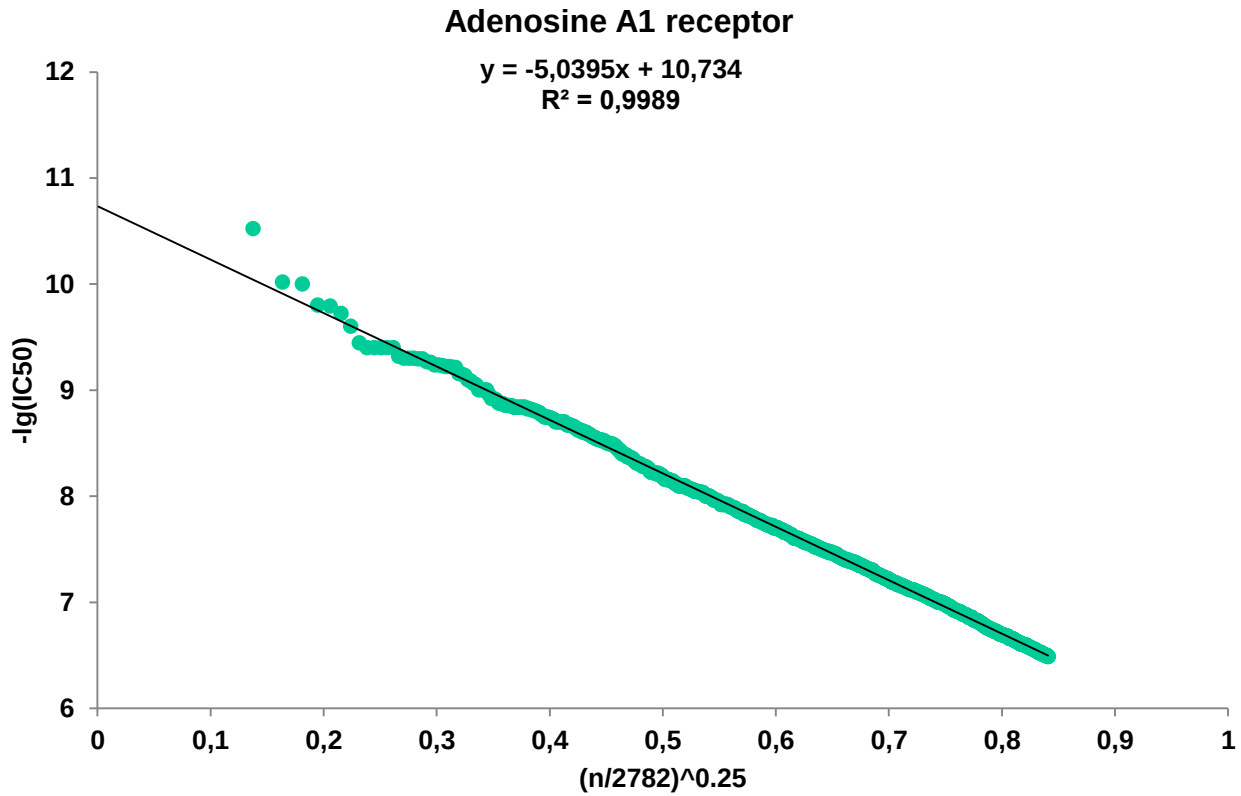
Получаем из леммы Морса:

$$Pr\{\Delta G > \Delta G_*\} \sim (\Delta G_{max} - \Delta G_*)^{\frac{m}{2}}$$

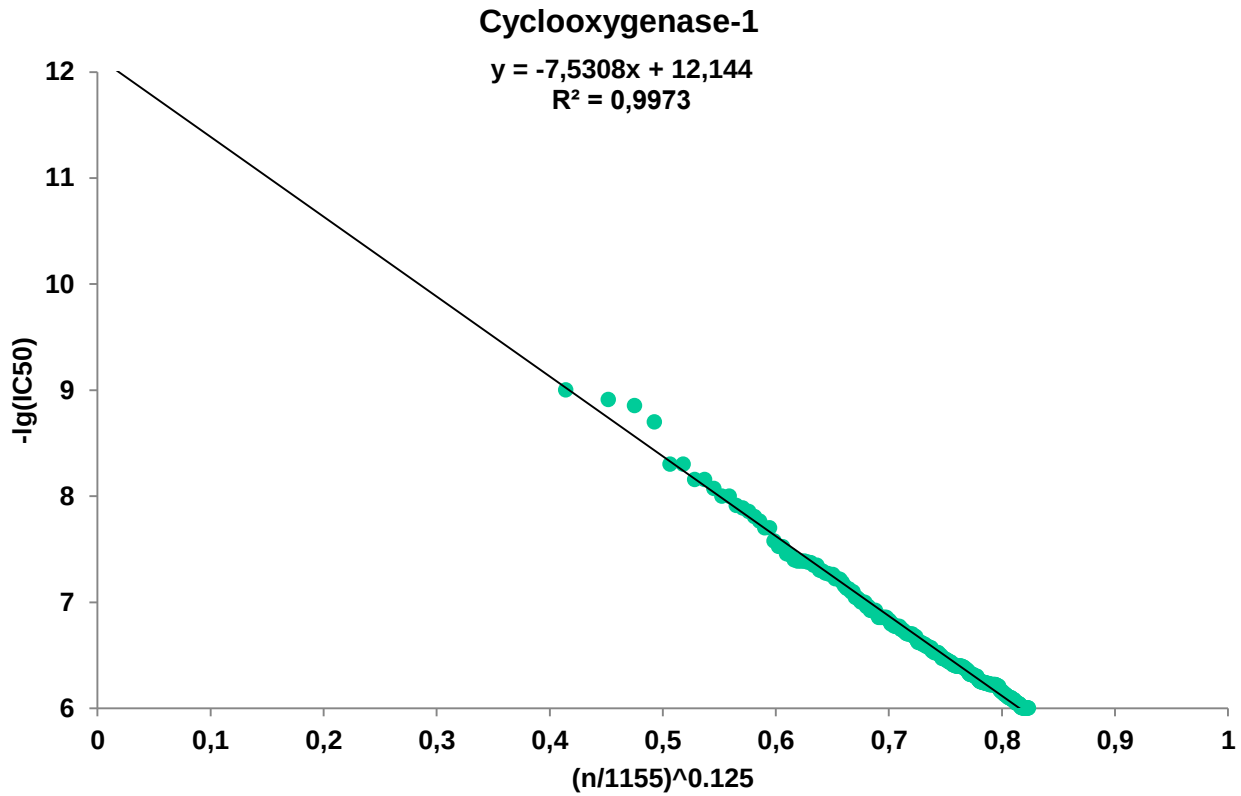
«Размерность активности»



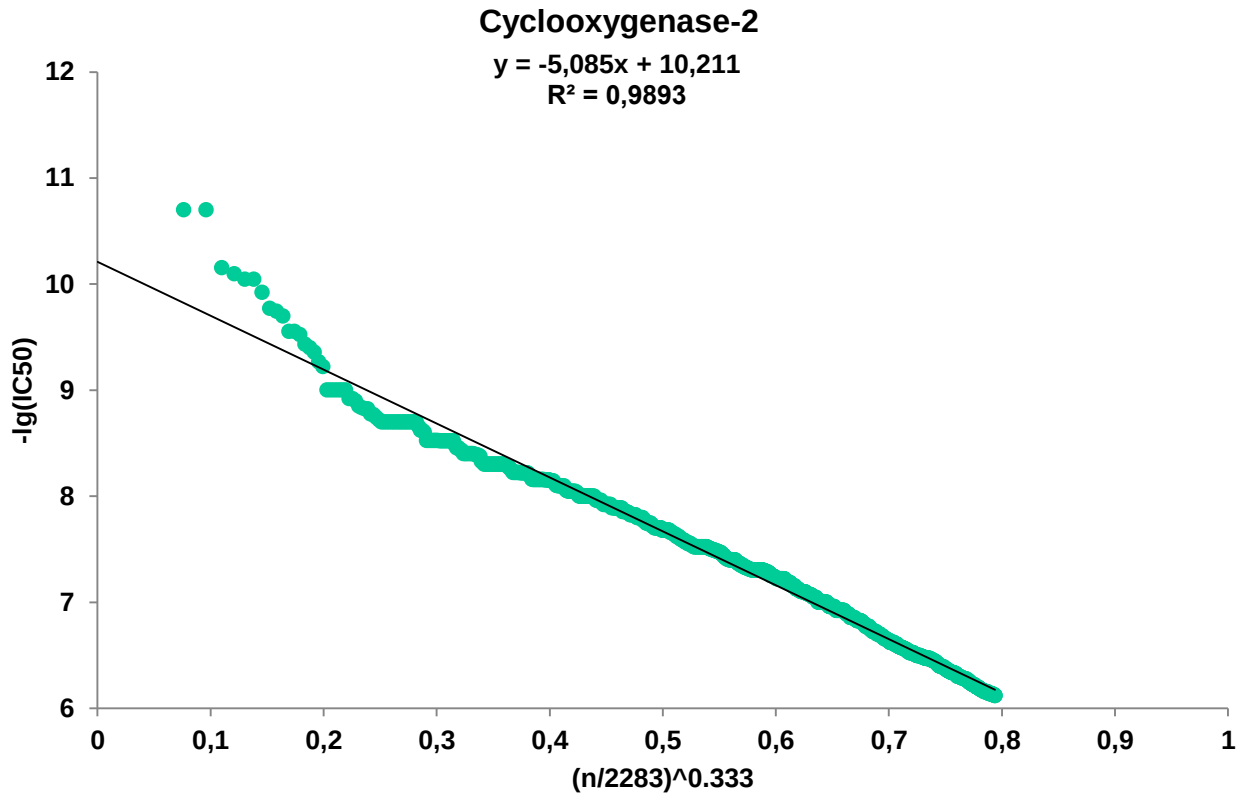
«Размерность активности»



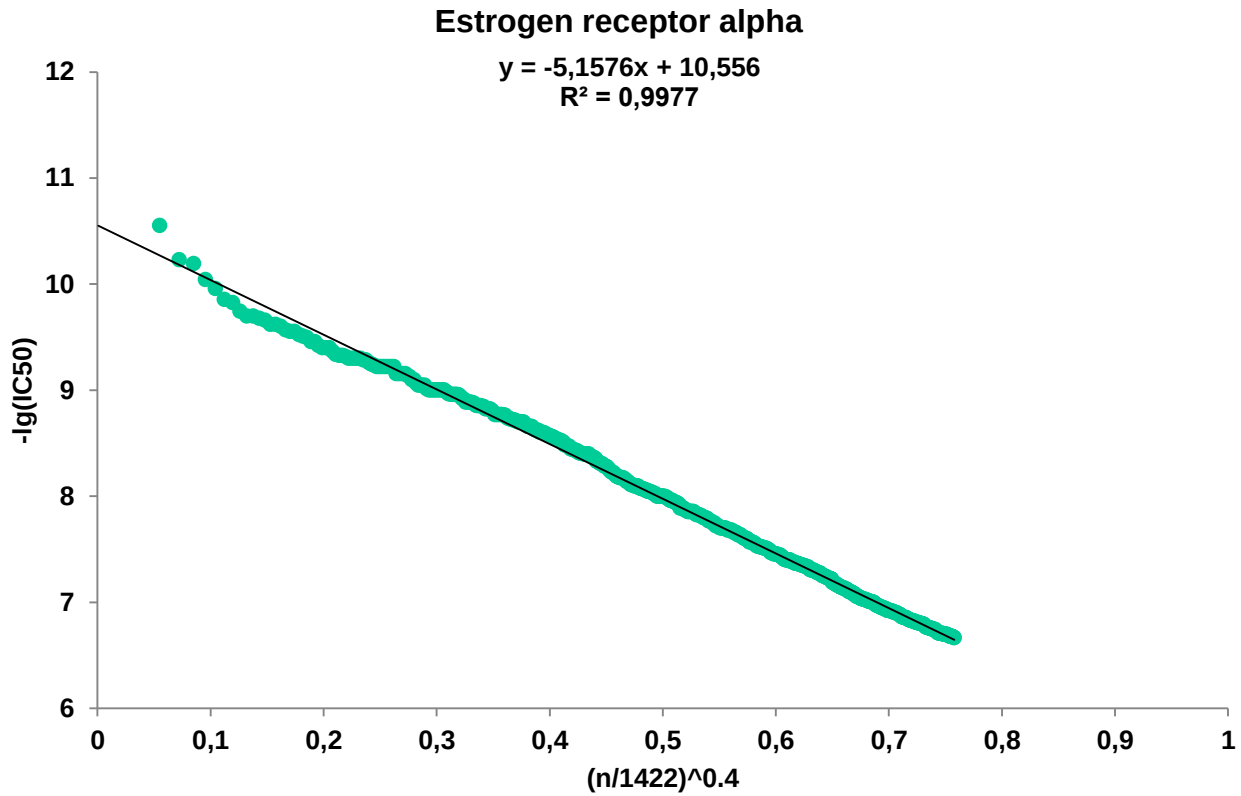
«Размерность активности»



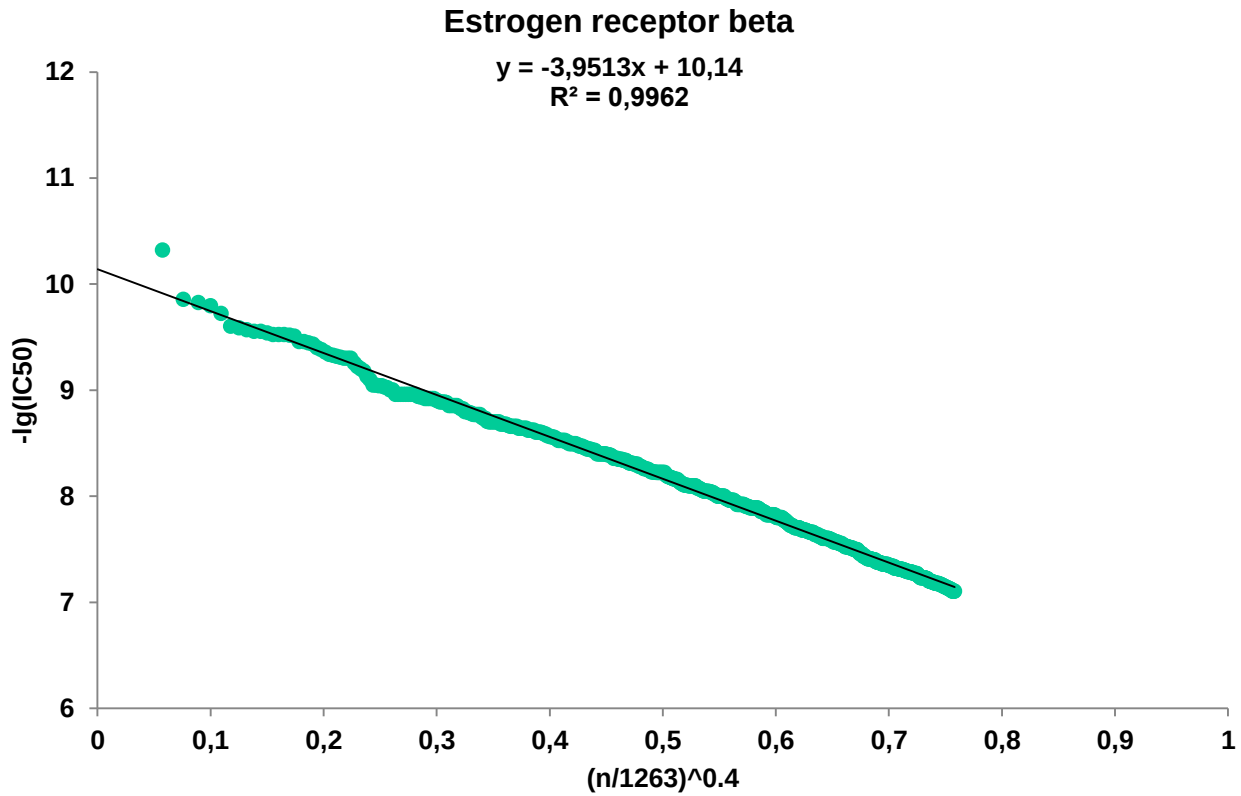
«Размерность активности»



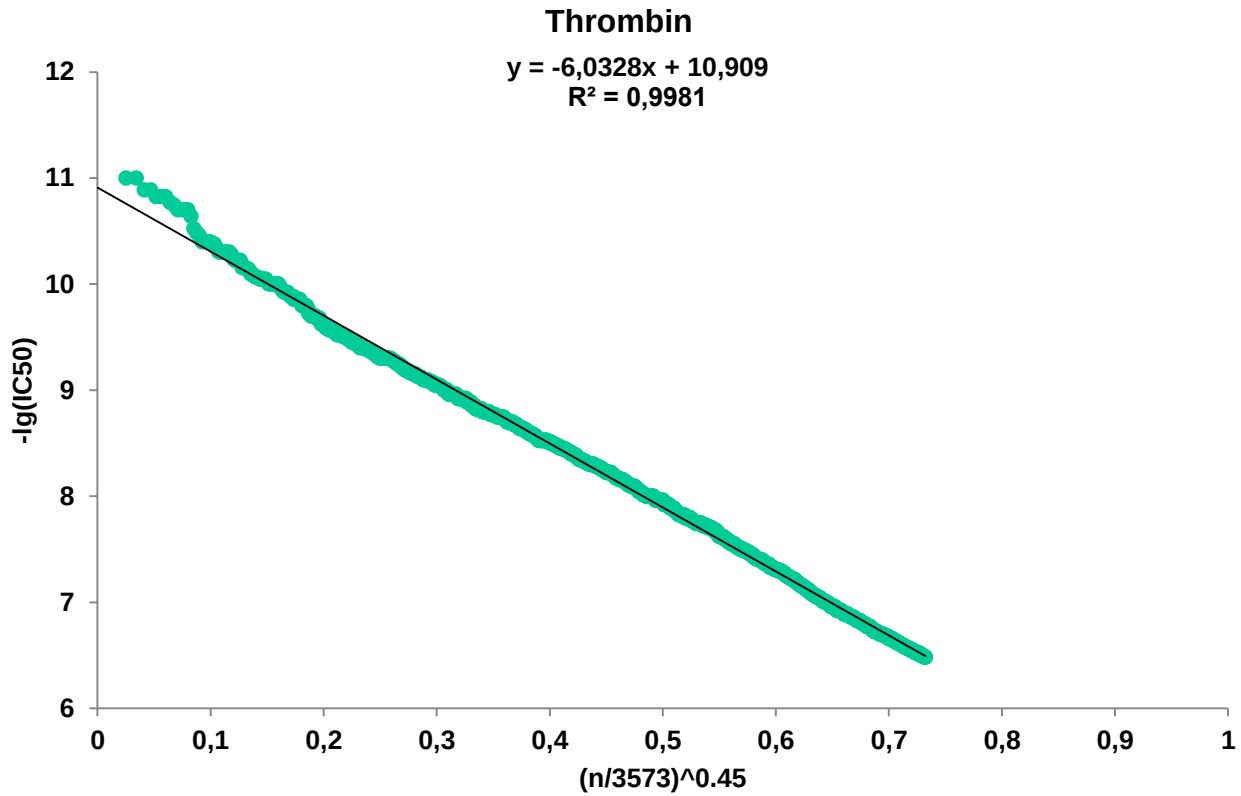
«Размерность активности»



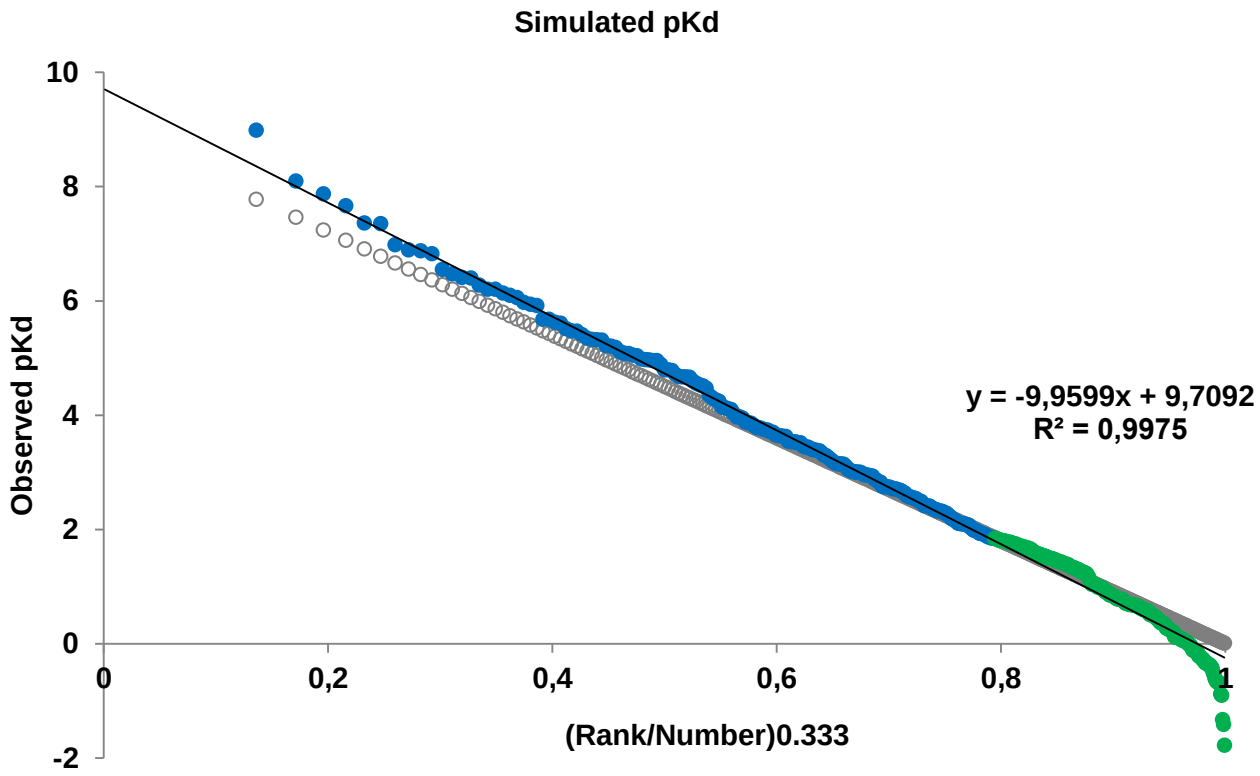
«Размерность активности»



«Размерность активности»



«Размерность активности» - симуляция



PASS Affinities

Из молекулярной биофизики и теории катастроф получаем:

$$P(A) = \left(\frac{\Delta G_{max} - \Delta G_*}{\Delta G_{max} - \Delta G_{min}}\right)^{\frac{m}{2}} = \frac{N_A}{N}$$

$$P(A|D_i) = \left(\frac{\Delta G_{max} - \Delta G_* + \Delta G_i}{\Delta G_{max} - \Delta G_{min}}\right)^{\frac{m}{2}} = \frac{N_{Ai}}{N_i}$$

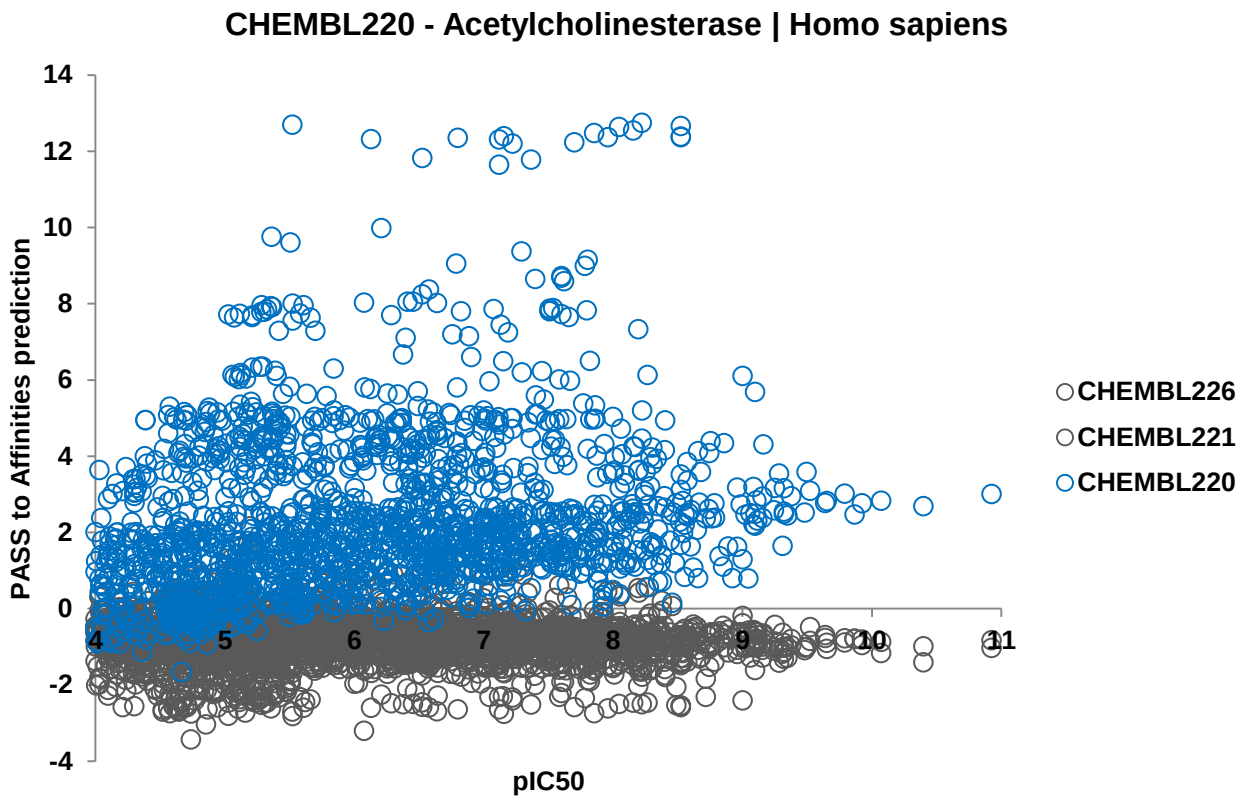
Откуда простейшими преобразованиями находим^

$$\Delta G = \sum_i \Delta G_i \sim A(ffinity) = \sum_i \left[\left(P(A|D_i)\right)^{1/\alpha} - \left(P(A)\right)^{1/\alpha}\right]$$

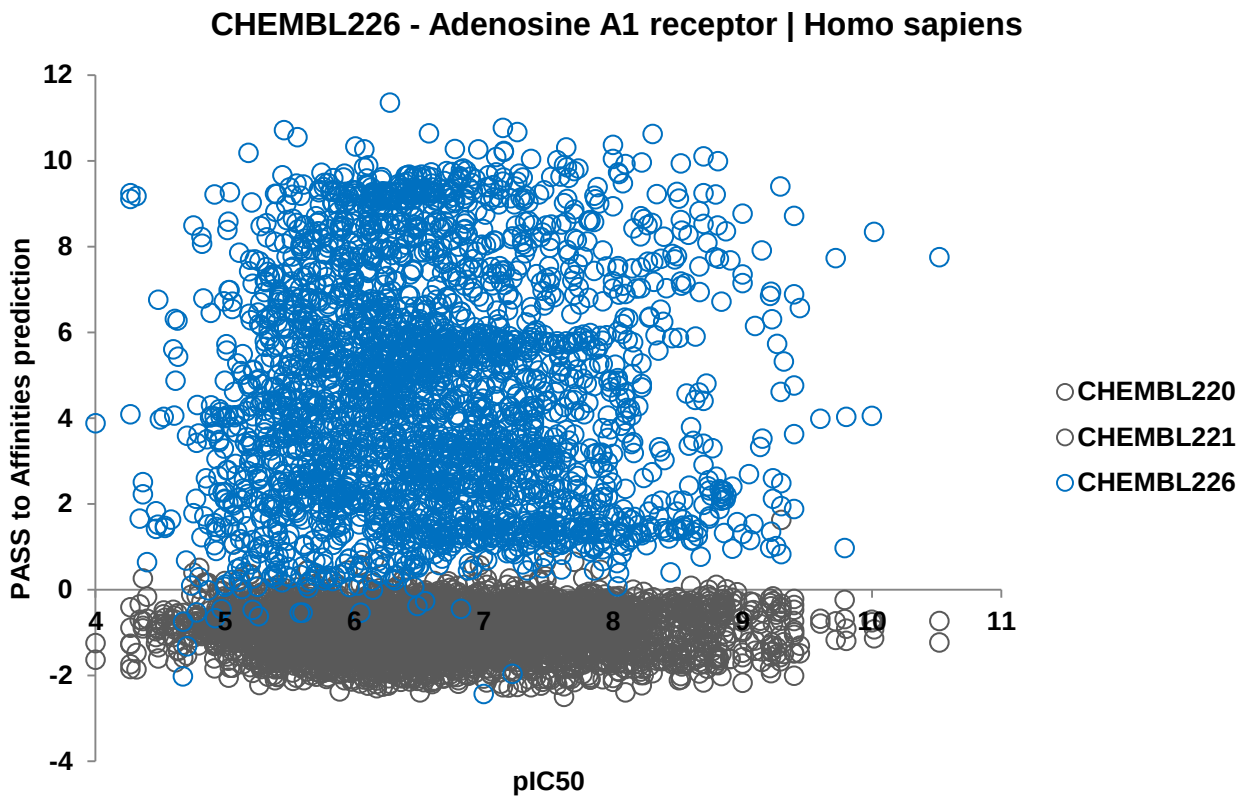
Новый алгоритм PASS Affinities основан на оценках:

$$A(ffinity) = \sum_i \left[\sqrt{P(A|D_i)} - \sqrt{P(A)}\right] = \sum_i \left[\sqrt{\frac{N_{Ai}}{N_i}} - \sqrt{\frac{N_A}{N}}\right]$$

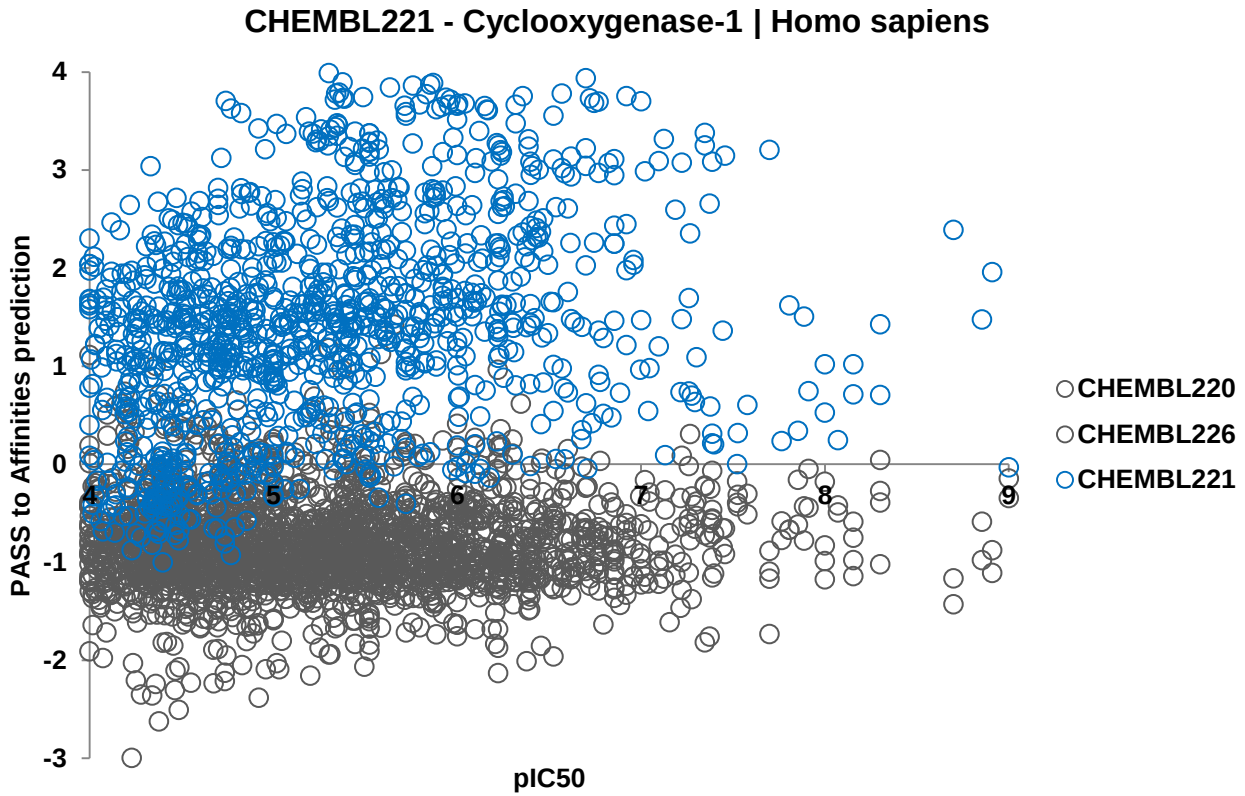
PASS Affinities
прогноз количественных данных по качественным



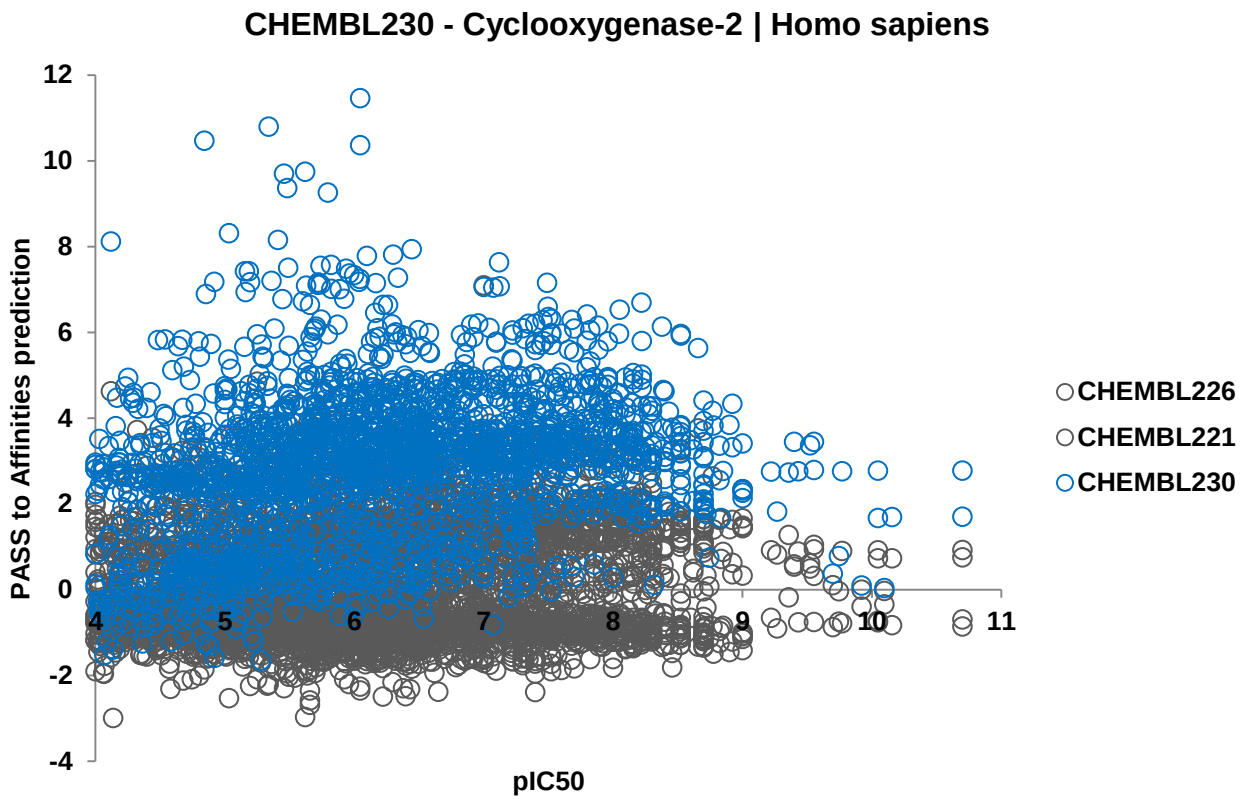
PASS Affinities
прогноз количественных данных по качественным



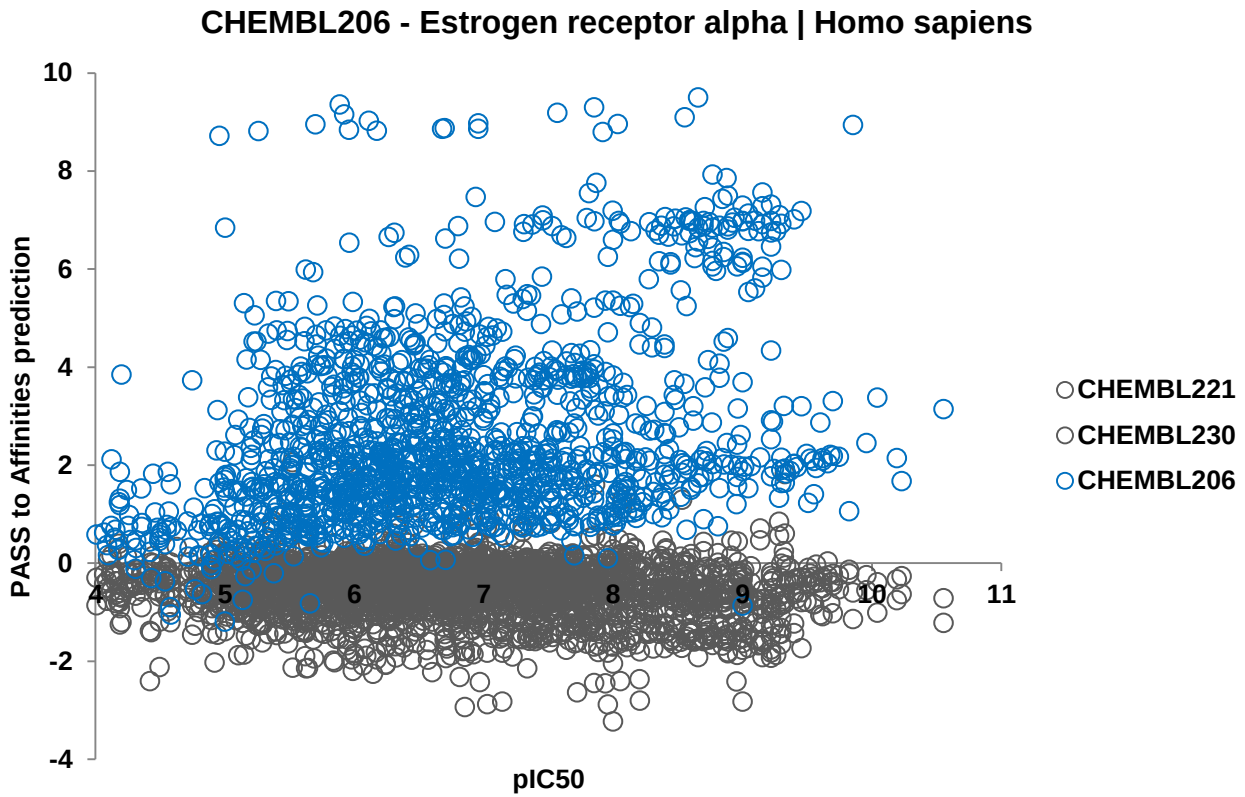
PASS Affinities
прогноз количественных данных по качественным



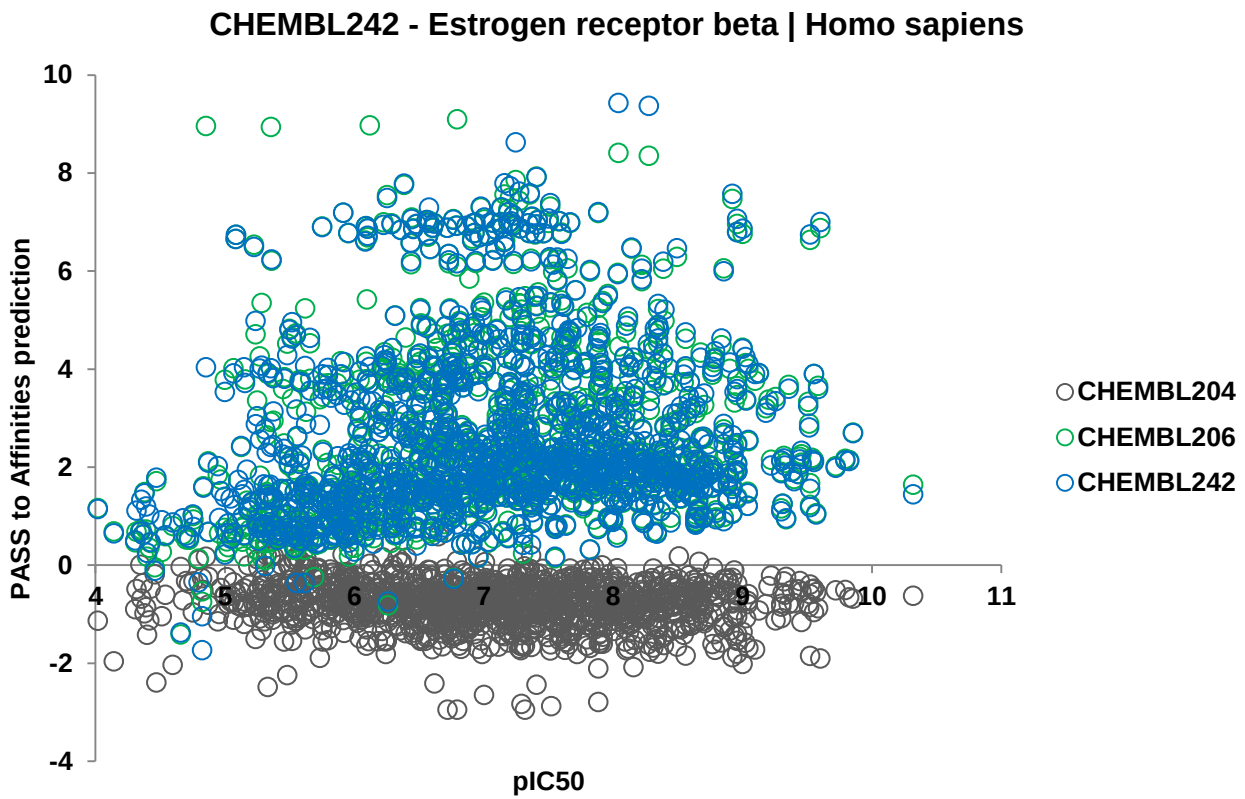
PASS Affinities
прогноз количественных данных по качественным



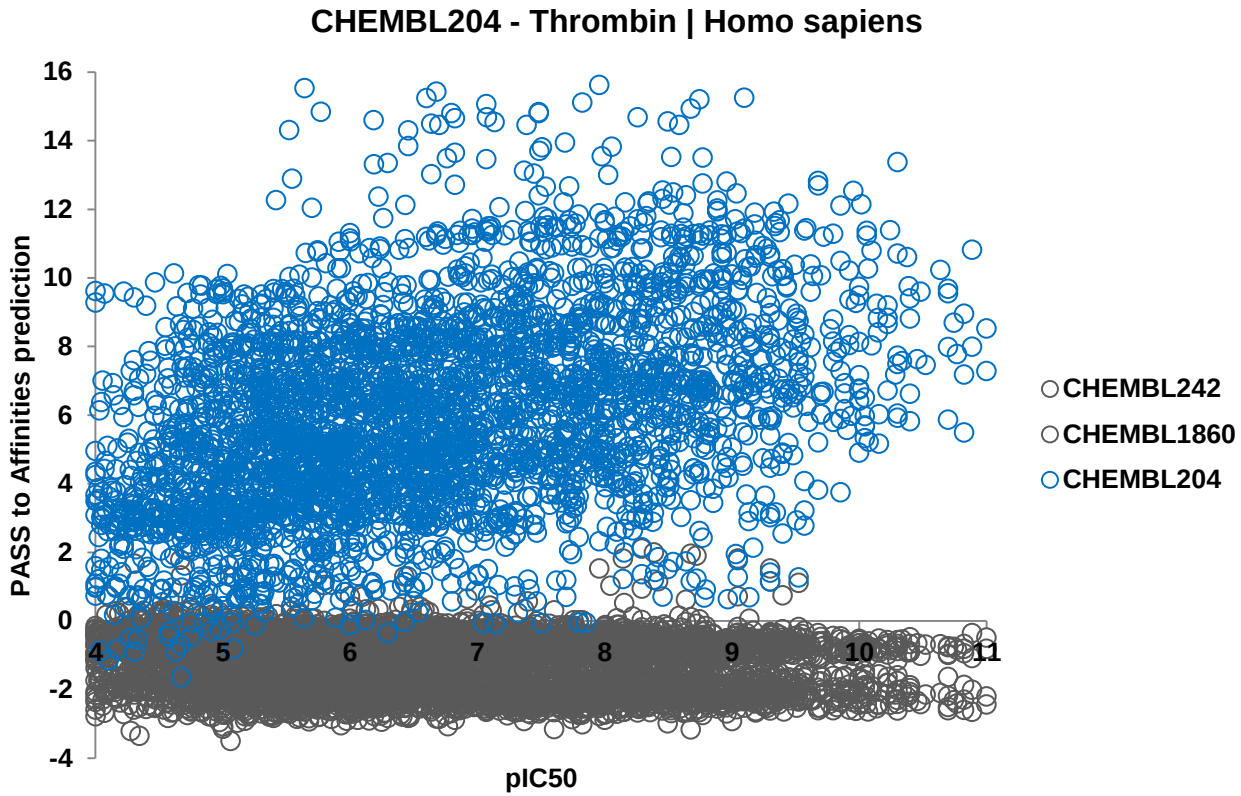
PASS Affinities
прогноз количественных данных по качественным



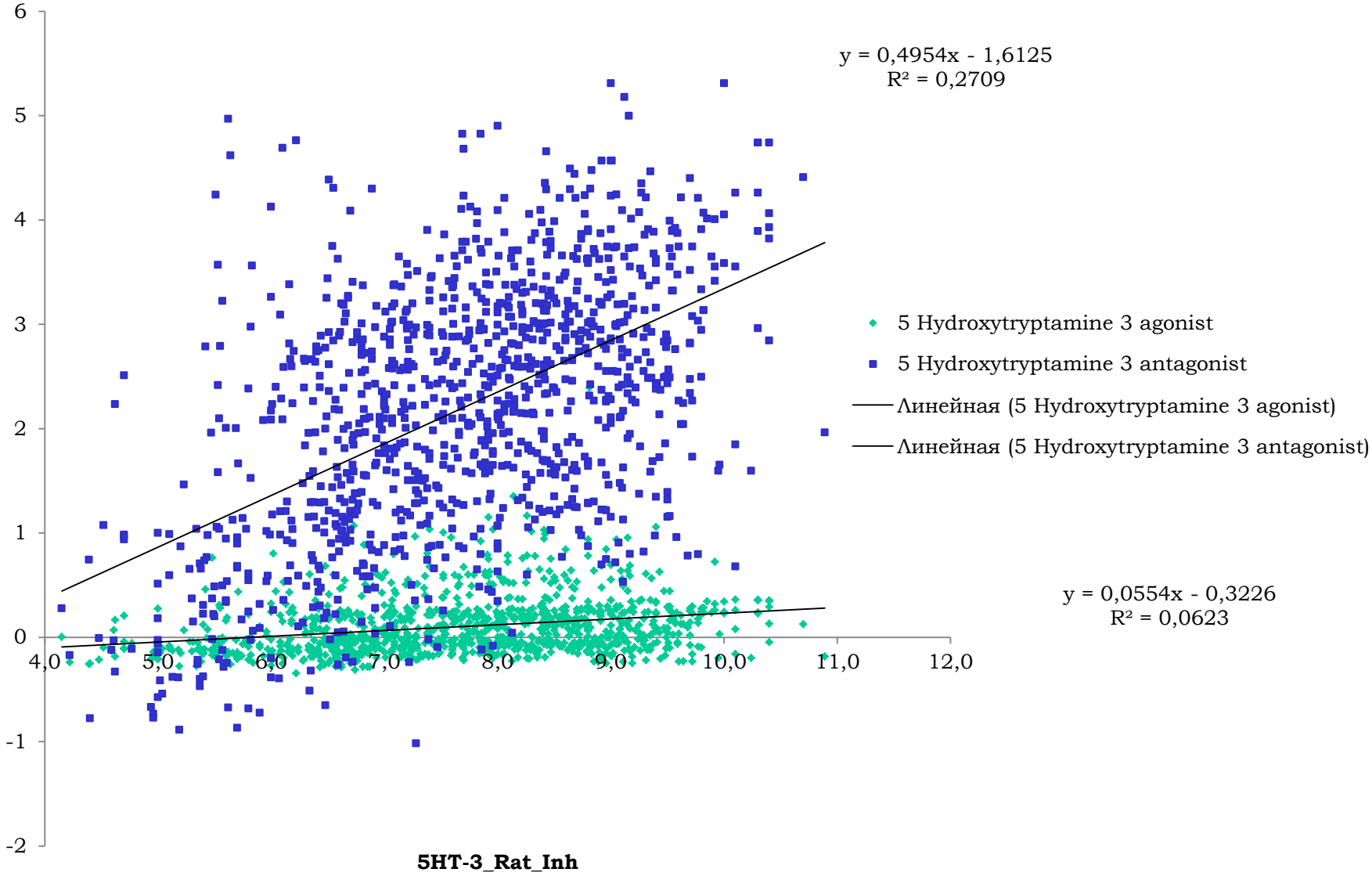
PASS Affinities
прогноз количественных данных по качественным



PASS Affinities
прогноз количественных данных по качественным



PASS Affinities
прогноз количественных данных по качественным



Биофизика, теория катастроф и QSAR

Активность $A(S)$ соединения S может быть представлена в виде:

$$A(S) \approx F(x_1, x_2, \dots, x_m) \equiv F(x)$$

Согласно лемме Морса в окрестности точки максимума $F(x)$ может быть представлена в виде:

$$F(x) = C - \sum_i u_i^2(x)$$

где $u_i(x)$ – гладкие преобразования координат $x_1, x_2, \dots, x_m$, которые можно записать в виде:

$$u_i(x) \approx c_i + \sum_k b_{ik} x_k$$

И, соответственно:

$$A(S) \approx C - \sum_i \left(c_i + \sum_k b_{ik} x_k \right)^2$$

Выводы

- Метод получения (Q)SAR оценок «Naive Bayes» известен как один из лучших. На основе молекулярной биофизики это находит свое естественное объяснение.

- Лемма Морса дает основу для степенного закона распределения аффинности лиганд-белковых комплексов.

- Сочетание знаний из разных областей науки позволило разработать новый компьютерный метод прогноза количественных оценок аффинности к макромолекулярным мишеням органических соединений по их структурным формулам на основе выборки с качественными данными о биологической активности органических соединений.

PASS Affinities или действительно ли наивен метод получения (Q)SAR оценок «Naive Bayes»?

Спасибо за внимание!