спецкурс
“СУЧАСНІ КОМП’ЮТЕРНІ МЕТОДИ В ХІМІЇ”

Регрессионные модели QSAR

В.В.Иванов

Materials Chemistry Department
V. N. Karazin National University,
61077, Kharkiv, Ukraine
ivv34@yahoo.com

МЕТОДЫ ПОСТРОЕНИЯ СТАТИСТИЧЕСКИХ МОДЕЛЕЙ

регрессия

- *Метод наименьших квадратов (LS)*
- *Нелинейный метод наименьших квадратов (NLS)*
- *Неполный метод наименьших квадратов (PLS)*
- *Метод наименьших модулей (LM)*
- *Аддитивные схемы (ADD)*
- *Сравнительный анализ молекулярных полей (CoMFA)*
- *Дискриминантный анализ (DA)*
- *Факторный анализ (FACTOR)*
- *Искусственные нейронные сети (NET)*
- *Кластерный анализ (CLSTR)*

Операционная система: DOS, WINDOWS, Linux.

Регрессионные модели биологической активности (BA)

1) Регрессия линейная (полилинейная)

$$BA = a_0 + a_1 d_1 + a_2 d_2 + a_3 d_3 + ...$$

2) Регрессия нелинейная

$$BA = a_0 + a_1 \log(\alpha d_1 + 1) + a_2 d_2 + \exp(\beta d_3) + ...$$

3) Аддитивные схемы

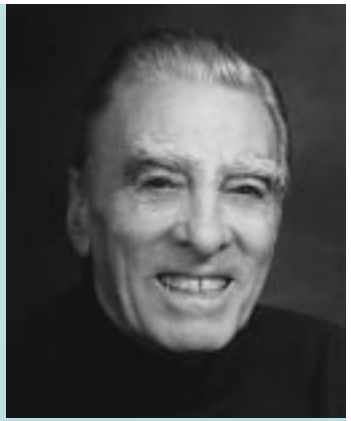
$$BA = \sum_{\text{fragments (i)}}^{N} n_i \chi_i + \sum_{\text{corrections (j)}}^{C} k_j F_j$$

χ_i Парциальные величины (инкременты)

4) Оценка прогностической способности: r, σ

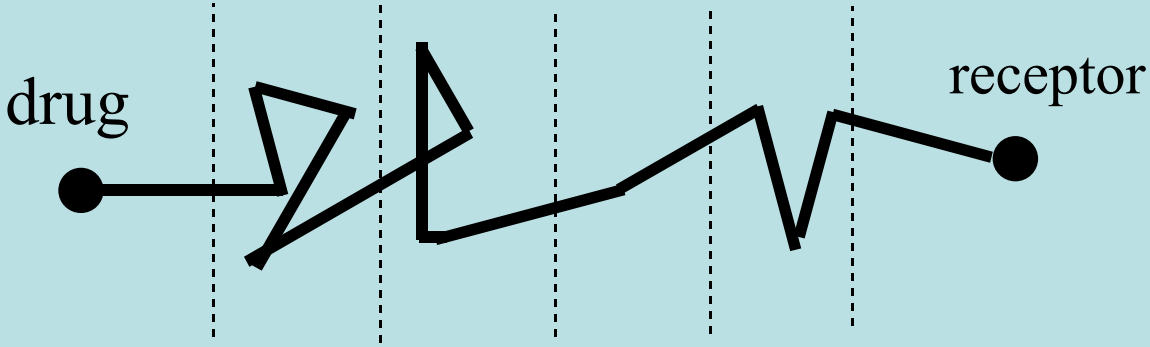
Live-one-out cross-validation (LOO) – процедура перекрестного оценивания

Теория Хэнча (Corwin Hansch, 1962)



Значение липофильности

$$\lg P = \lg \frac{C_{1-\text{octanole}}}{C_{\text{water}}} = \lg C_{1-\text{octanole}} - \lg C_{\text{water}}$$

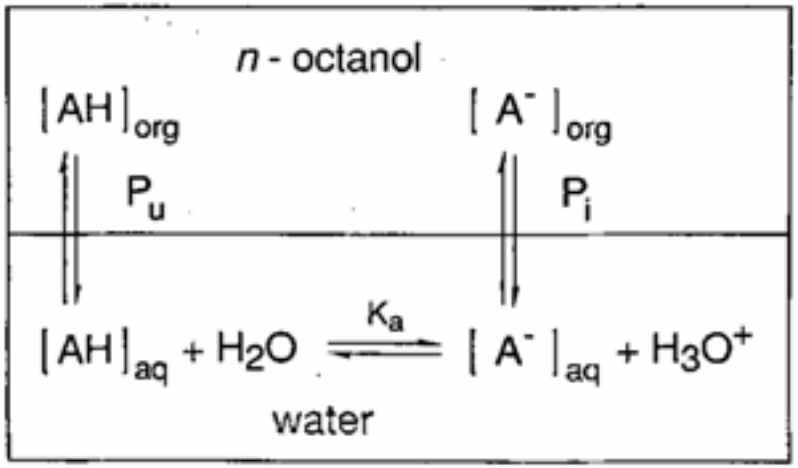


Параметры заместителя

$$\pi_X = \lg P_{C_6H_5X} - \lg P_{C_6H_6}$$

$$\pi_X = \lg P_{RX} - \lg P_{RH}$$

Программы для расчета липофильности: HyperChem 6.0 DRAGON ACD/logP

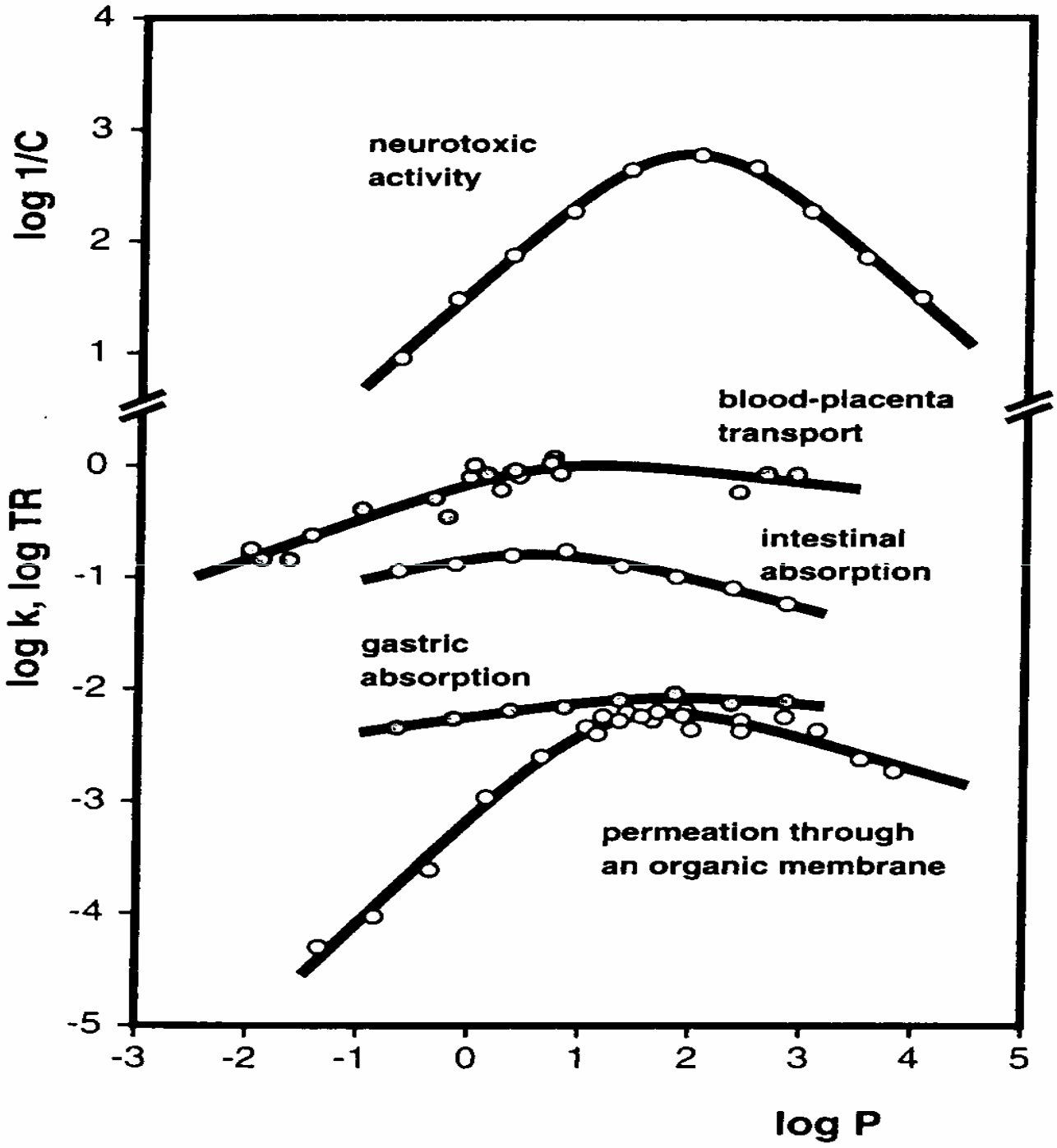


$$XH \rightleftharpoons X^- + H^+$$

$$\lg D = \lg P - \lg(1 + 10^{-pK+pH})$$

$$X + H^+ \rightleftharpoons XH^+$$

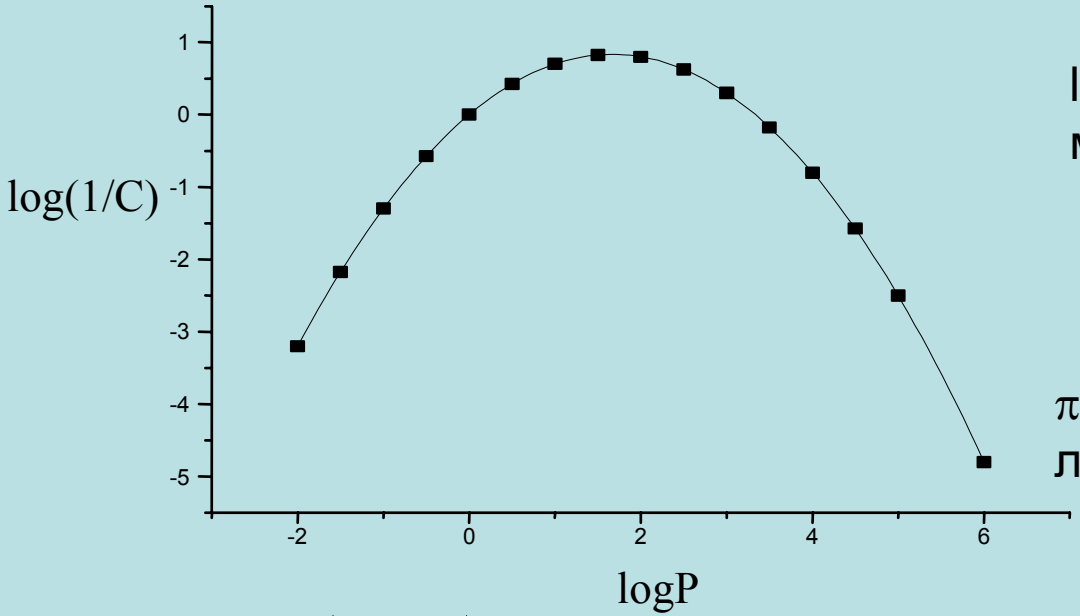
$$\lg D = \lg P - \lg(1 + 10^{pK-pH})$$


log 1/C
log k, log TR
log P
neurotoxic activity
blood-placenta transport
intestinal absorption
gastric absorption
permeation through an organic membrane
4
3
2
1
0
-1
-2
-3
-4
-5
-3
-2
-1
0
1
2
3
4
5

Регрессионные модели Биологической активности

Уравнение Хэнча:

$$\log 1/C = a \cdot \lg P - b \cdot \lg^2 P + \sum_{subst.} c_i \sigma_i + \ldots$$



lgP – липофильность молекулы

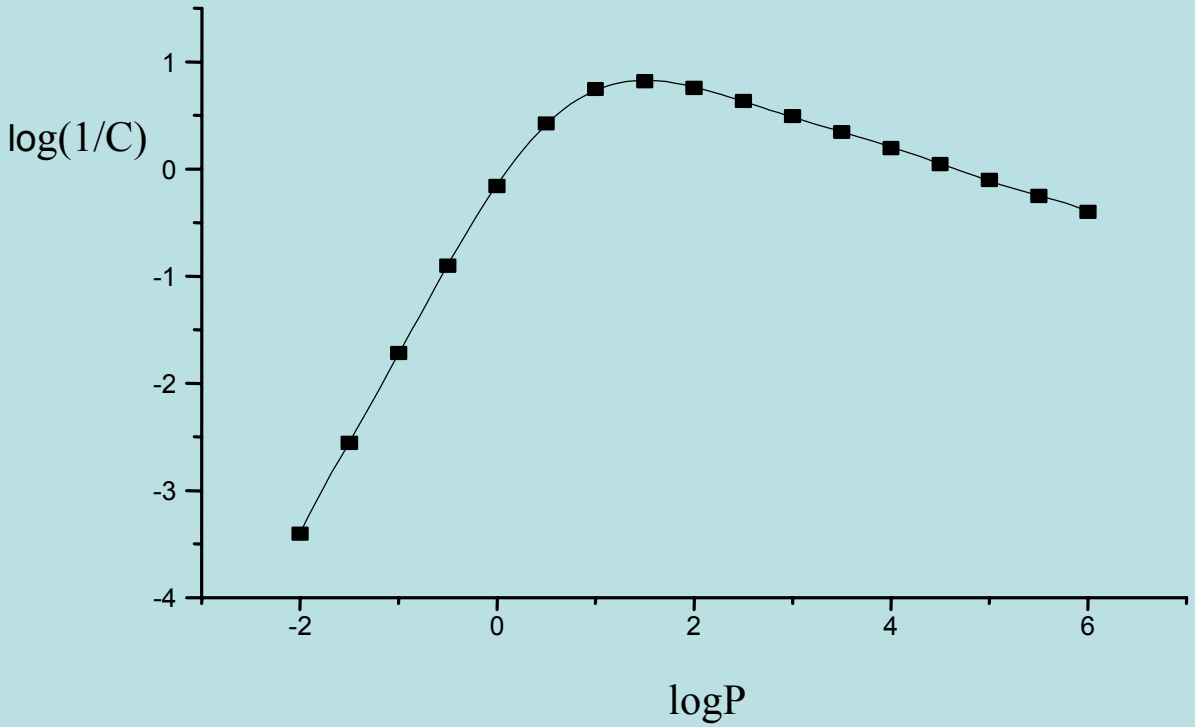
π- парциальная липофильность (заместителя)

$$\log 1/C_0 = a\left(\sum \pi_i\right)^2 + b\sum \pi_i + c\sum \sigma_i + d\sum E_{si} + \ldots + const$$

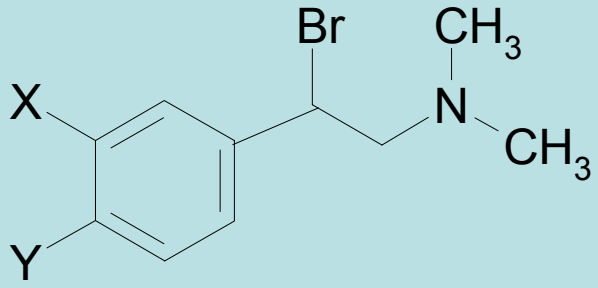
Билинейная модель (Kubiny)

$$\log 1/C_0 = a \log P + b \log(\beta P + 1) + ...$$

P – константа распределения в-ва в системе октанол-вода



Фунгистатическая активность



N,N-dimethyl-α–bromophenethylamines
(X,Y=H,F,Cl,Br,I,CH_3)

$$\log 1/C = 1.15\pi - 1.46\sigma^{+} + 7.82$$

токсичность набора ароматических соединений по отношению к *Tetrahymena pyriformis*:

$$\lg \frac{1}{IGC_{50}} = 0.603 \cdot \lg P - 0.33 \cdot E_{LUMO} - 1$$

$$(N = 239,\ r^2 = 0.8,\ q^2 = 0.796,\ \sigma = 0.335,\ F = 476)$$

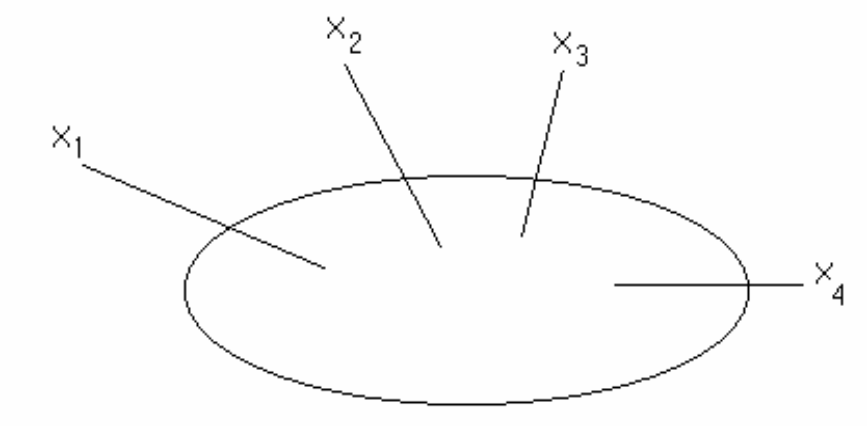
Аддитивные схемы

$$\log(1/C) = \sum_{\text{fragments (i)}}^{N} n_i \chi_i + \sum_{\text{corrections (j)}}^{C} k_j F_j$$

Используется для вычисления липофильностей (lgP), молекулярных рефракций и т.д.

χ_i - Парциальные величины (инкременты)

Один из вариантов аддитивного метода - Метод Фри-Вильсона

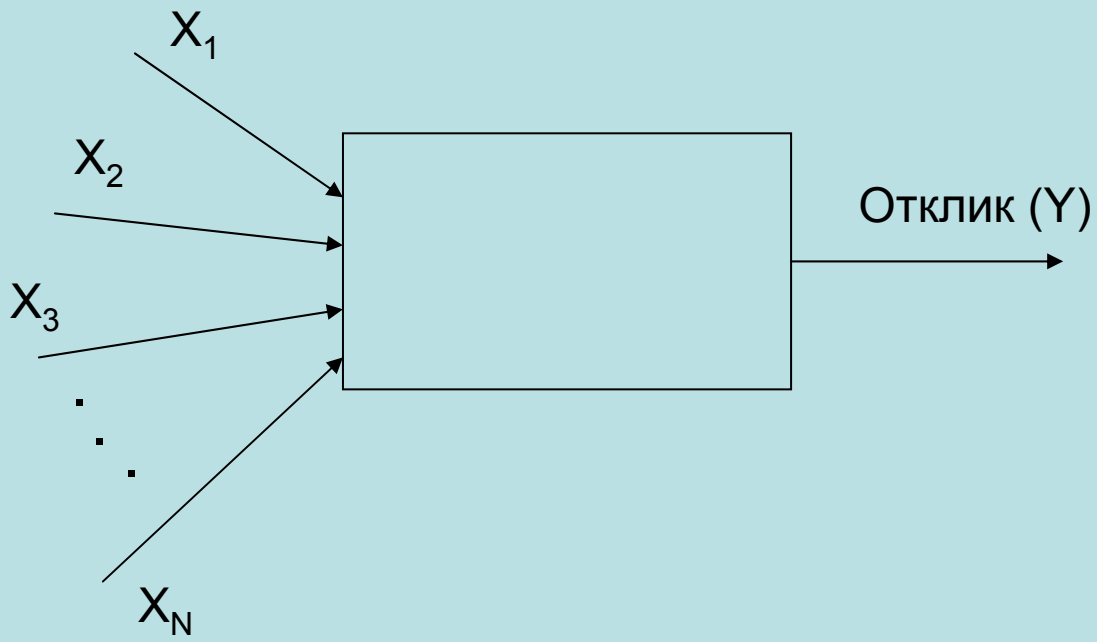


$$-\log C = y_0 + \sum_{i,j} n_{ij} \chi_{ij}$$

n_{ij} -Кол-во заместителей типа i в положении j

Оценка прогностической способности: r, σ
Live-one-out cross-validation (LOO) –
процедура перекрестного оценивания

Факторный анализ (анализ главных компонент)



Фактор:

$$\varphi = c_y Y + \boxed{c_1 X_1 + c_2 X_2 + \ldots\ldots} + c_N X_N$$

Латентные переменные

НЕПОЛНЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Partial Least-Squares / Projection on Latent Structures (PLS)
число объектов (молекул) << числа переменных (дескрипторов)

Обычно для МНК $M \geq (3 \div 5)N$

$$y = a_0 + \sum_{i=1}^{N>>M} a_i x_i$$

1) PCR - Регрессия главных компонент (факторов)

$$y = a_1 PC_1 + a_2 PC_2 + \ldots + a_p PC_p$$

2) PLS - Регрессия факторов большим вкладом отклика (y)

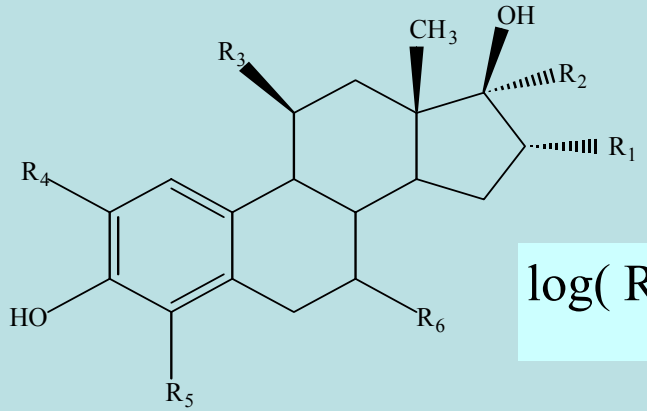
$$y = a_1 PC_1(y) + a_2 PC_2(y) + \ldots + a_p PC_p(y)$$

PC_i – факторы. Находятся с учетом факторной структуры y

P << число объектов (молекул)

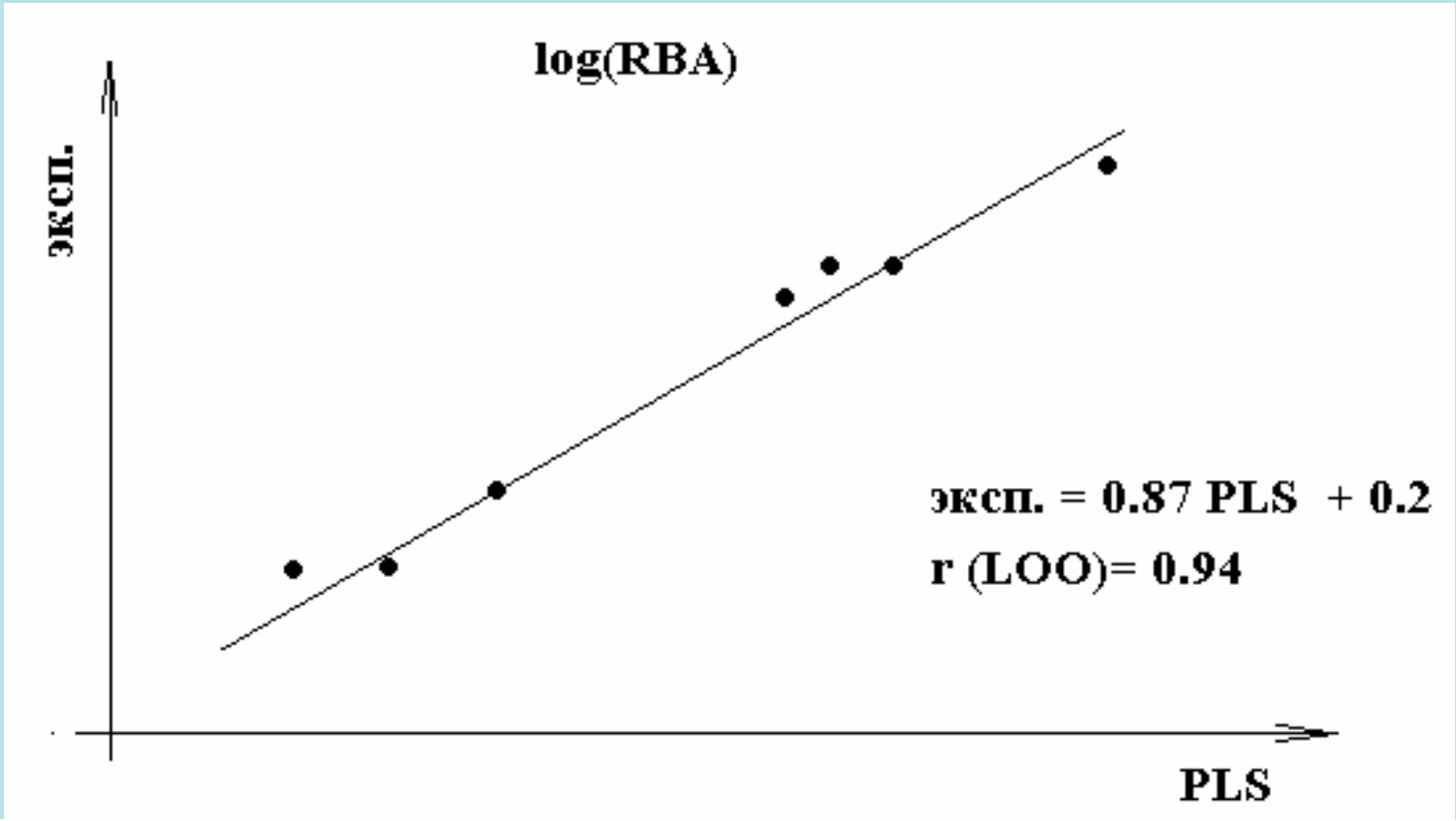
Активность эстрадиолов

PLS



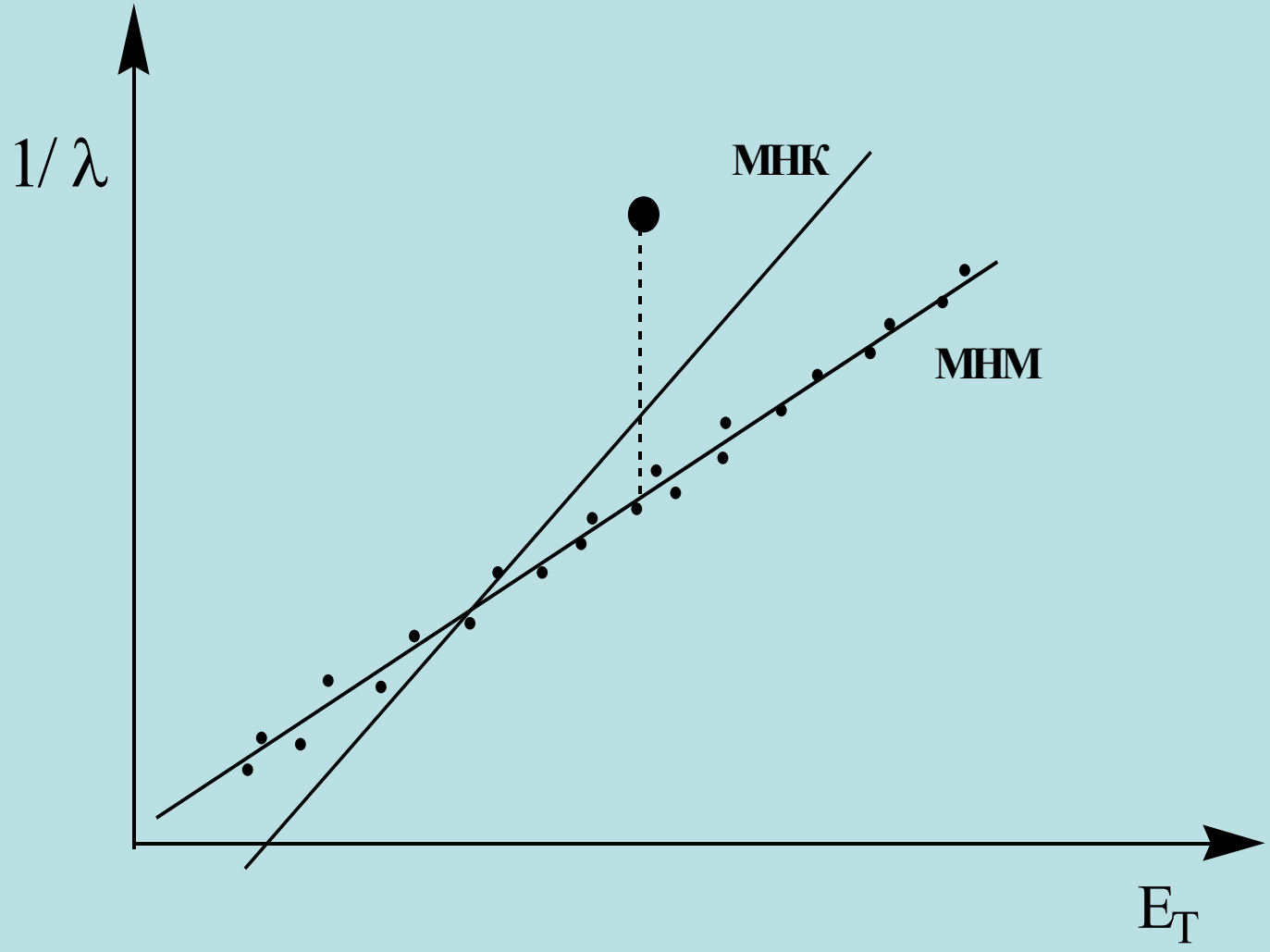
Активность: Log(RBA)

$$\log(\mathrm{RBA}) = \sum_i a_i T_i + \sum_i b_i ET_i + \sum_i c_i G_i + \ldots$$



Least Absolute Deviation (LAD)

метод наименьших модулей



Least Absolute Deviation (LAD)

метод наименьших модулей (мнм)

$$F = F(a_0, a_1, a_2, \ldots a_k) = \sum_i (y_i - \hat{y}_i)^2$$ МНК

$$F = F(a_0, a_1, a_2, \ldots a_k) = \sum_i p_i (y_i - \hat{y}_i)^2$$ Взвешенный МНК

$$p_i = \frac{1}{\sigma_i^2}$$

$$p_i = \frac{1}{|y_i - \hat{y}_i|}$$

МНМ

$$W = W(a_0, a_1, a_2, \ldots a_k) = \sum_i \frac{1}{|y_i - \hat{y}_i|} (y_i - \hat{y}_i)^2 = \sum_i |y_i - \hat{y}_i|$$

pKa органических кислот.

15 одноосновных насыщенных кислот:

$HCOOH$, CH_3COOH, C_2H_5COOH, C_3H_7COOH,

$(CH_3)_2CHCOOH$, $CH_3(CH_2)_3COOH$, $(CH_3)_2CHCH_2COOH$,

$(CH_3)_3CCOOH$, CH_2FCOOH, $CH_2ClCOOH$, $CH_2BrCOOH$,

CH_2ICOOH, $CHCl_2COOH$, CCl_3COOH, CF_3COOH

OLS (мнк)

$$pK_a = -24.44 - 91.35x_2, \quad R^2 = 0.852, \quad Q^2 = 0.805$$

LAD (мнм)

$$pK_a = -20.53 - 79.06x_2, \quad R^2 = 0.839, \quad Q^2 = 0.798$$

X_2 - заряд на окси группе

pKa органических кислот

заряд на кислороде карбонильной группы (x1)
заряд на кислороде окси-группы (x2),
заряд на водороде окси-группы (x3),

OLS

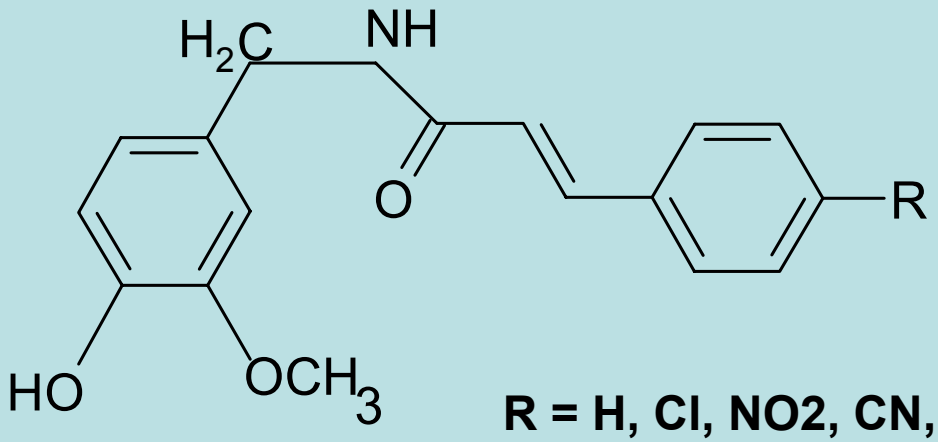
$$pK_a = -1.08 - 19.93x_1 - 46.80x_2 - 67.61x_3, \quad R^2 = 0.971, \quad Q^2 = 0.746$$

LAD

$$pK_a = -4.28 - 17.22x_1 - 55.12x_2 - 61.17x_3, \quad R^2 = 0.969, \quad Q^2 = 0.201$$

МЕТОД НАИМЕНЬШИХ МОДУЛЕЙ

АНАЛЬГЕТИЧЕСКОЕ ДЕЙСТВИЕ ПРОИЗВОДНЫХ КАПСАИЦИНА



R = H, Cl, NO2, CN, C6H5, N(CH3)2, I, NHCHO

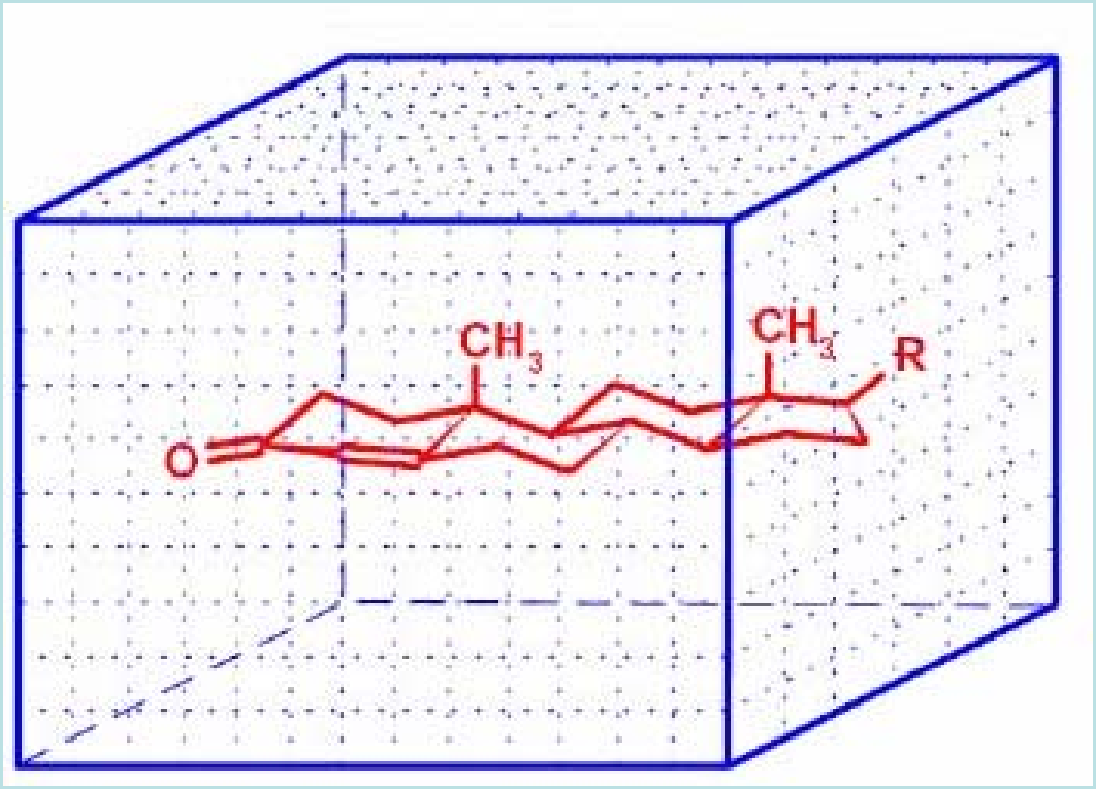
МНК: $-\lg EC_{50} = 1.137 - 0.068 \cdot MR, \quad r = 0.726, q = 0.554$

МНМ: $-\lg EC_{50} = 1.142 - 0.069 \cdot MR, \quad r = 0.726, q = 0.715$

МНК: $-\lg EC_{50} = 0.770 - 0.815 \cdot \pi \quad r = 0.943, \; q = 0.877$

МНМ: $-\lg EC_{50} = 0.767 - 0.708 \cdot \pi \quad r = 0.940, \; q = 0.846$

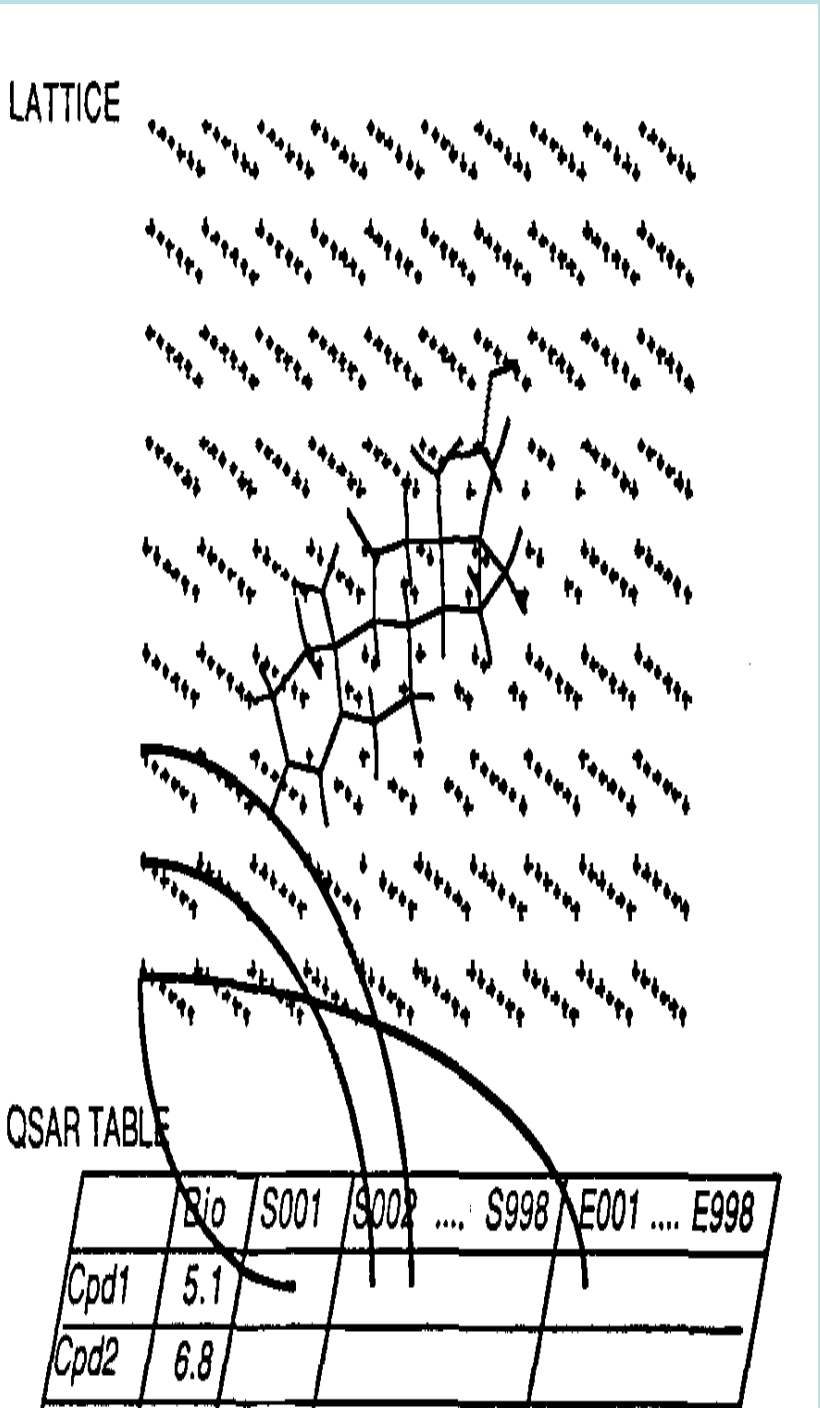
Comparative Molecular Field Analysis (CoMFA)
Сравнительный анализ молекулярных полей



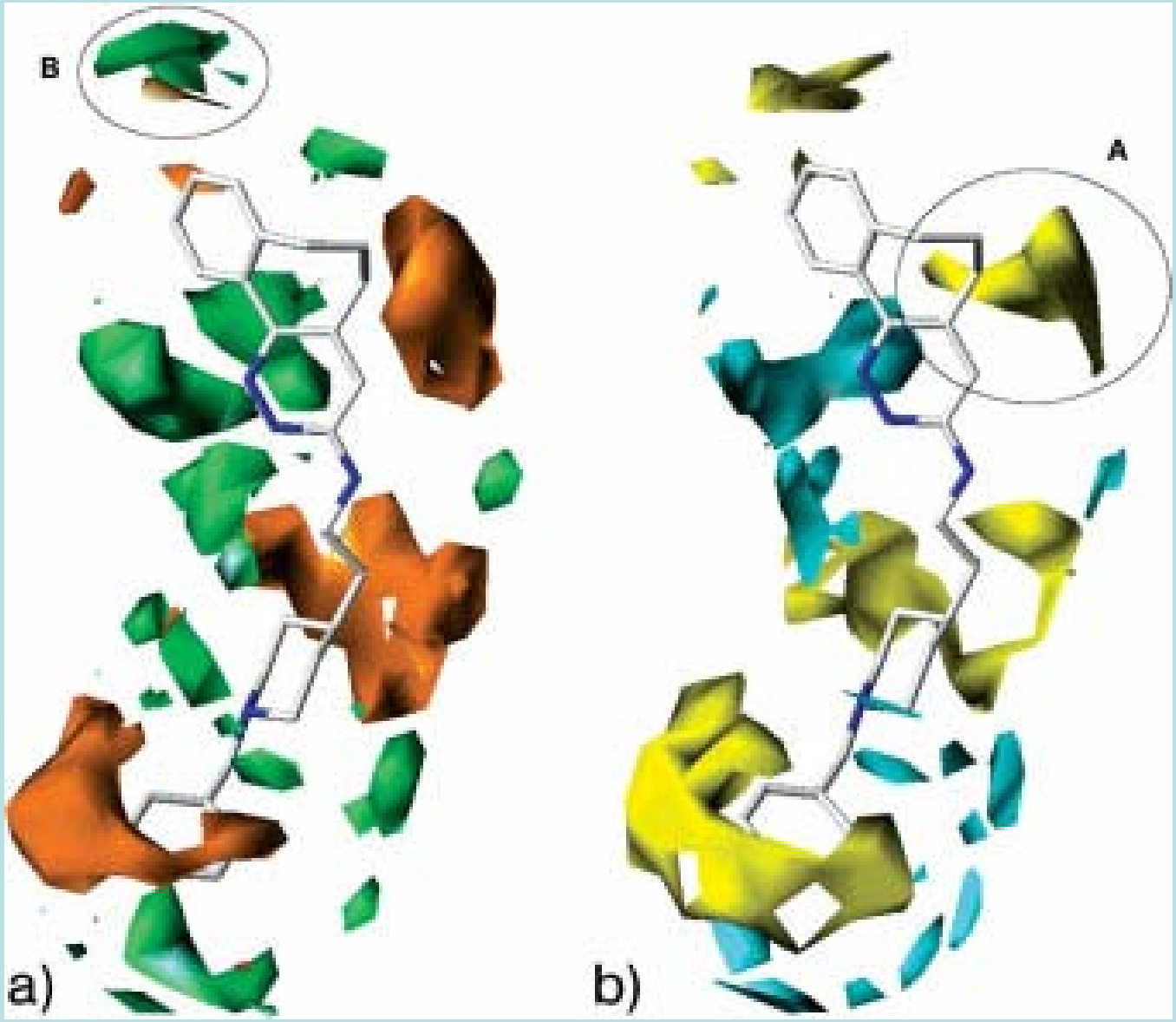
$$\log(1/C) = \sum_i a_i U_i + \sum_i b_i S_i$$

U_i Электростатический потенциал в точке i

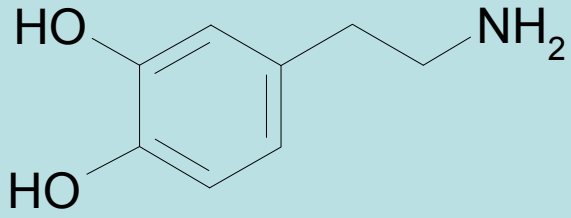
S_i Стерический потенциал в точке i


LATTICE
QSAR TABLE
Bio
S001
S002
....
S998
E001 .... E998
Cpd1
5.1
Cpd2
6.8

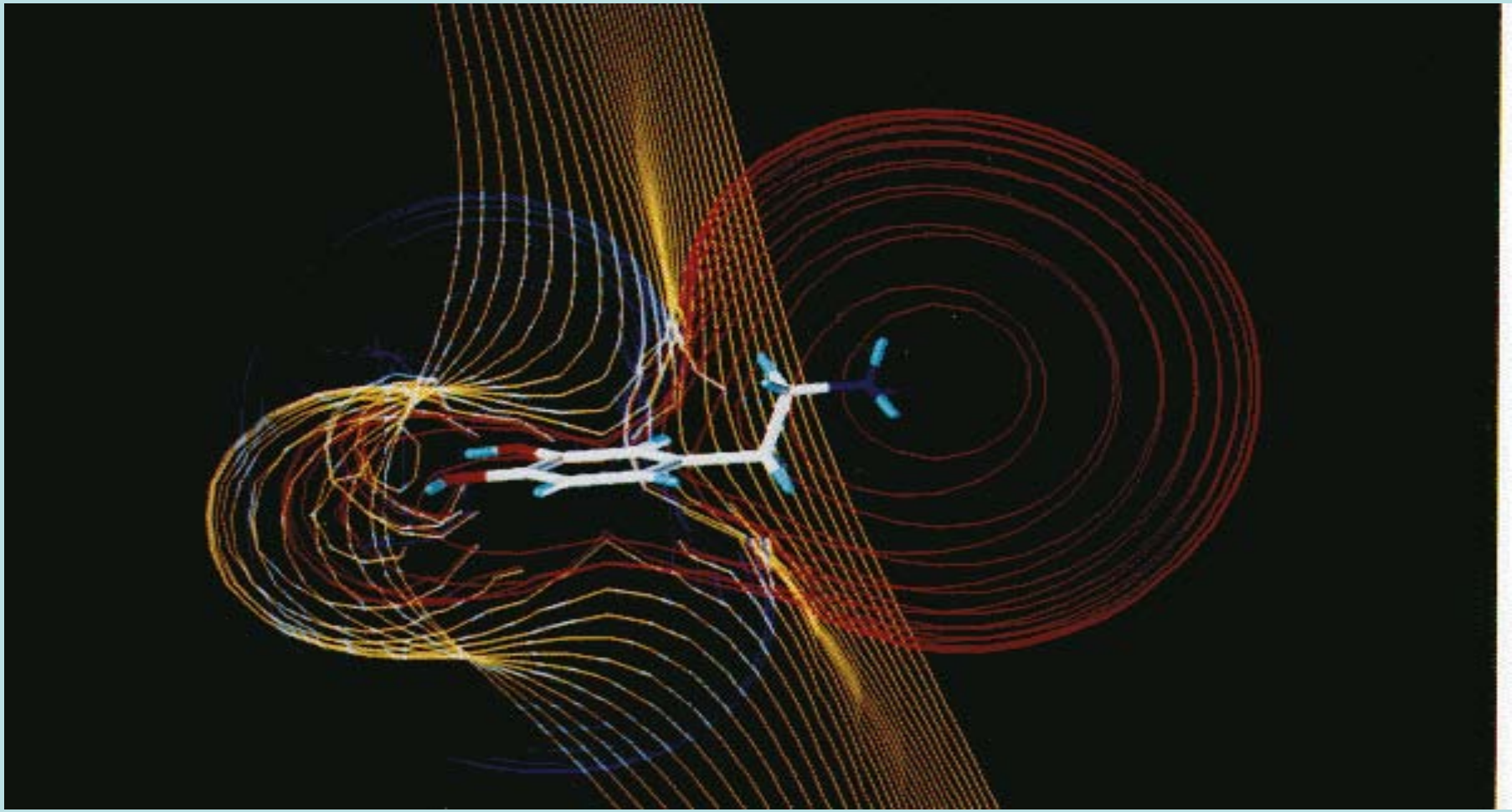
PLS коэффициенты

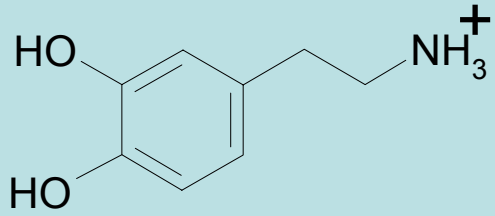


Дофамин (3,4-dihydroxyphenethylamine)- нейротрансмиттер



Карта Электростатического поля




HO
HO
NH3
+