спецкурс
“СУЧАСНІ КОМП’ЮТЕРНІ МЕТОДИ В ХІМІЇ”

Регресійні моделі QSAR

В.В.Іванов

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

МЕТОДИ ПОБУДОВИ СТАТИСТИЧНИХ МОДЕЛЕЙ

регрессія

- *Метод наименьших квадратів (LS)*
- *Нелінійний метод наименьших квадратів (NLS)*
- *Неповний метод наименьших квадратів (PLS)*
- *Адитивні схеми (ADD)*
- *Порівнювальний аналіз молекулярних полей (CoMFA)*
- *Дискримінаційний аналіз (DA)*
- *Факторний аналіз (FACTOR)*
- *Штучні нейронні мережі (NET)*
- *Кластерний аналіз (CLSTR)*

Операционная система: DOS, WINDOWS, Linux.

Регресійні моделі біологічної активності (ВА)

1) Регресія лінійна (полілінійна)

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

2) Регресія нелінійна

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

3) Адитивні схеми

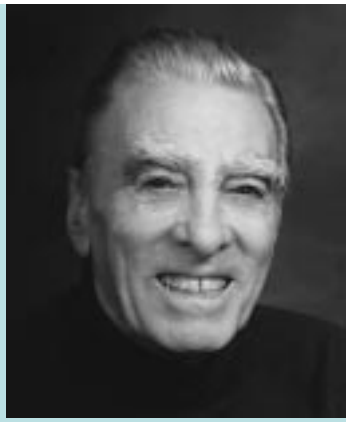
$$BA = \sum_{\text{fragments (i)}}^{N} n_i \chi_i + \sum_{\text{corrections (j)}}^{C} k_j F_j$$

χ_i Парціальні величини (інкременти)

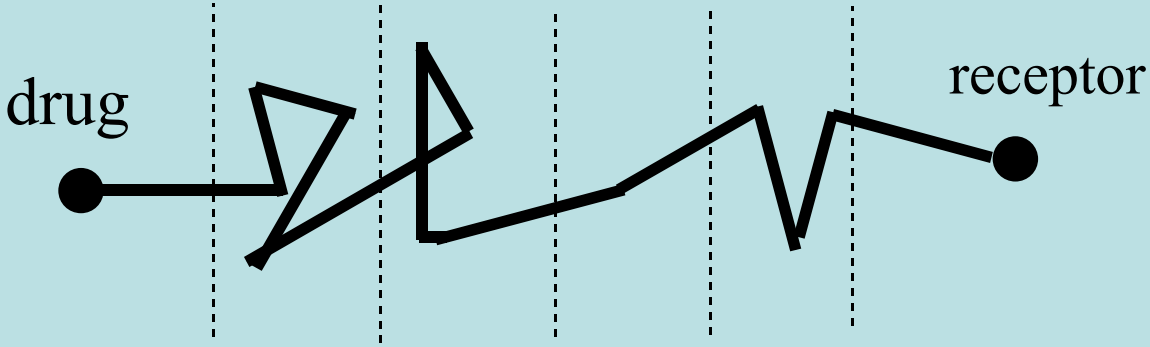
4) Оцінка прогностичної здатності : r, σ

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

Теорія Хэнча (Corwin Hansch, 1962)



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

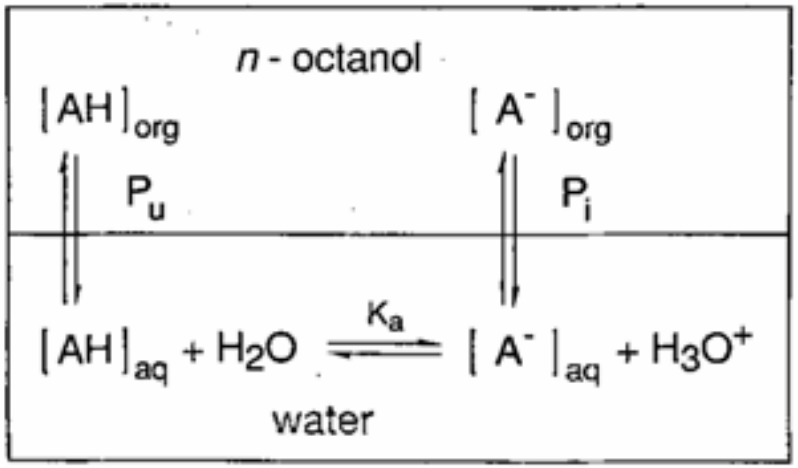


Параметри замісників

$$\text{aromatic}: \quad \pi_X = \log P_{C_6H_5X} - \log P_{C_6H_6}$$

$$\text{aliphatic}: \quad \pi_X = \log P_{RX} - \log P_{RH}$$

Програми для розрахунку ліпофільності: HyperChem 6.0 DRAGON ACD/logP

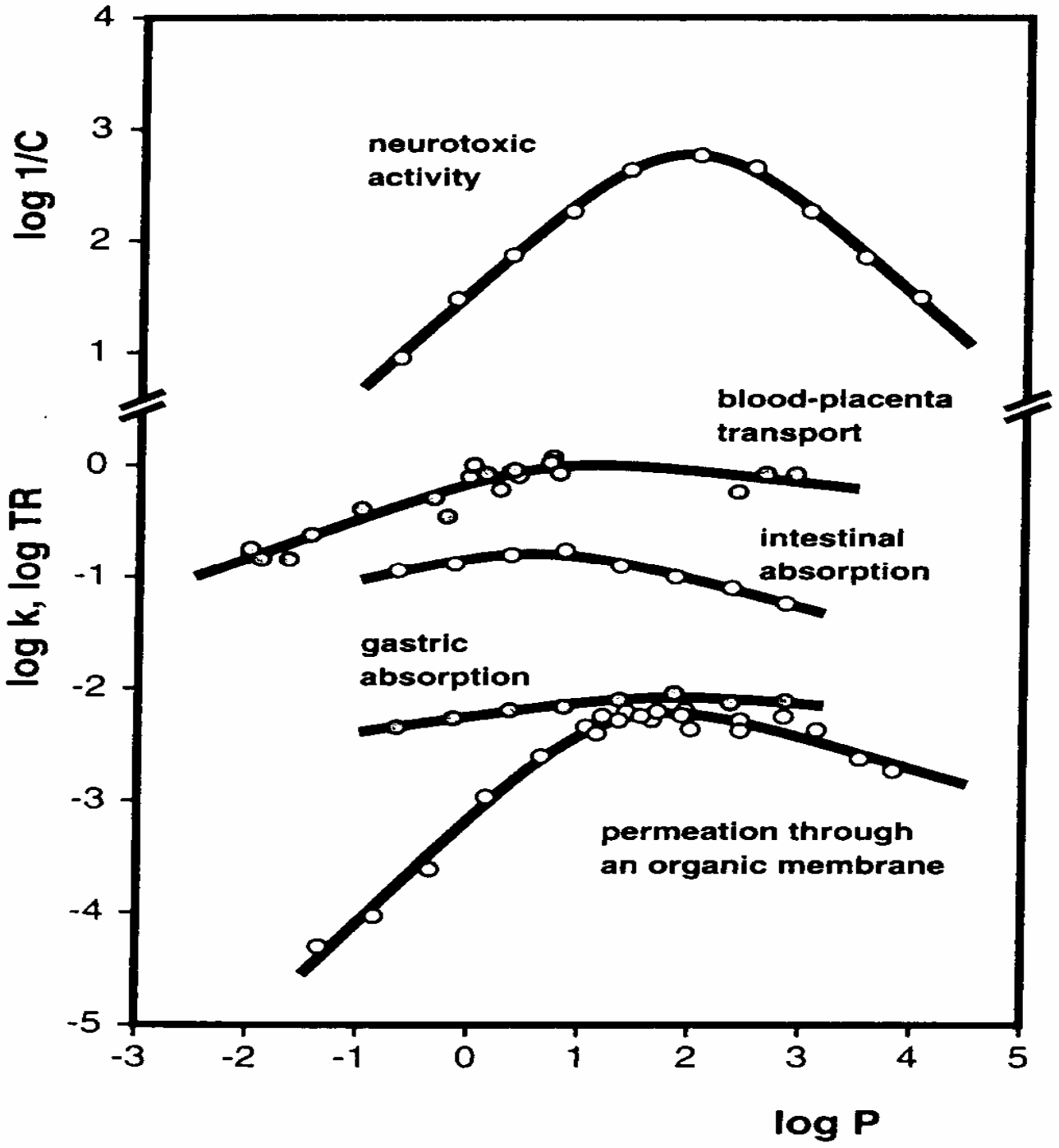


$$XH \rightleftharpoons X^- + H^+$$

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

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

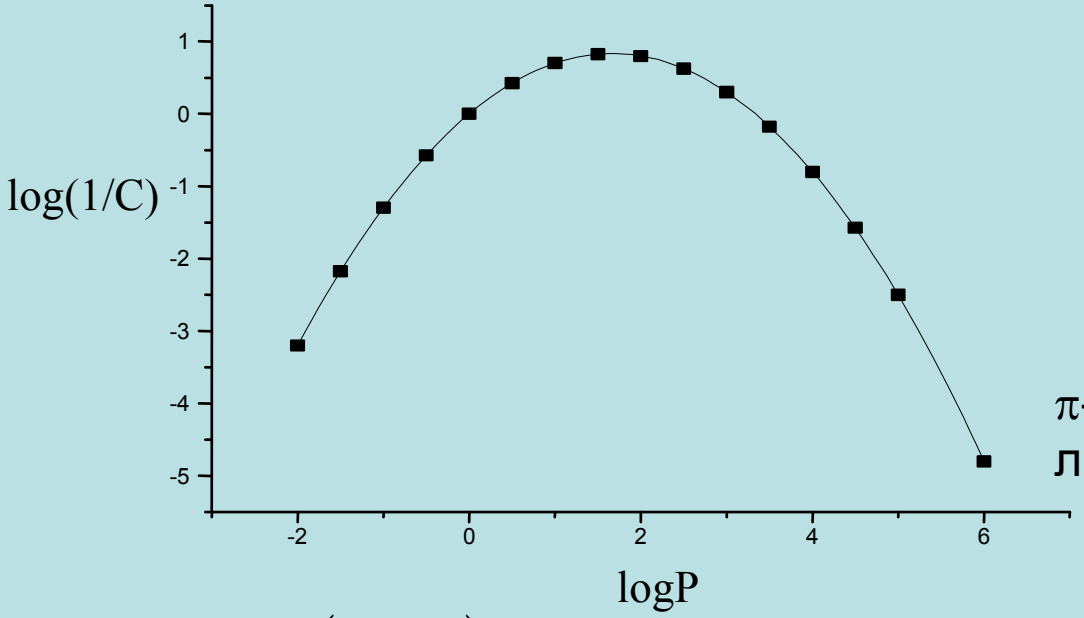
$$\lg D = \lg P - \lg(1 + 10^{pK-pH})$$


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

Регресійні моделі Біологічної активності

Рівняння Хэнча:

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

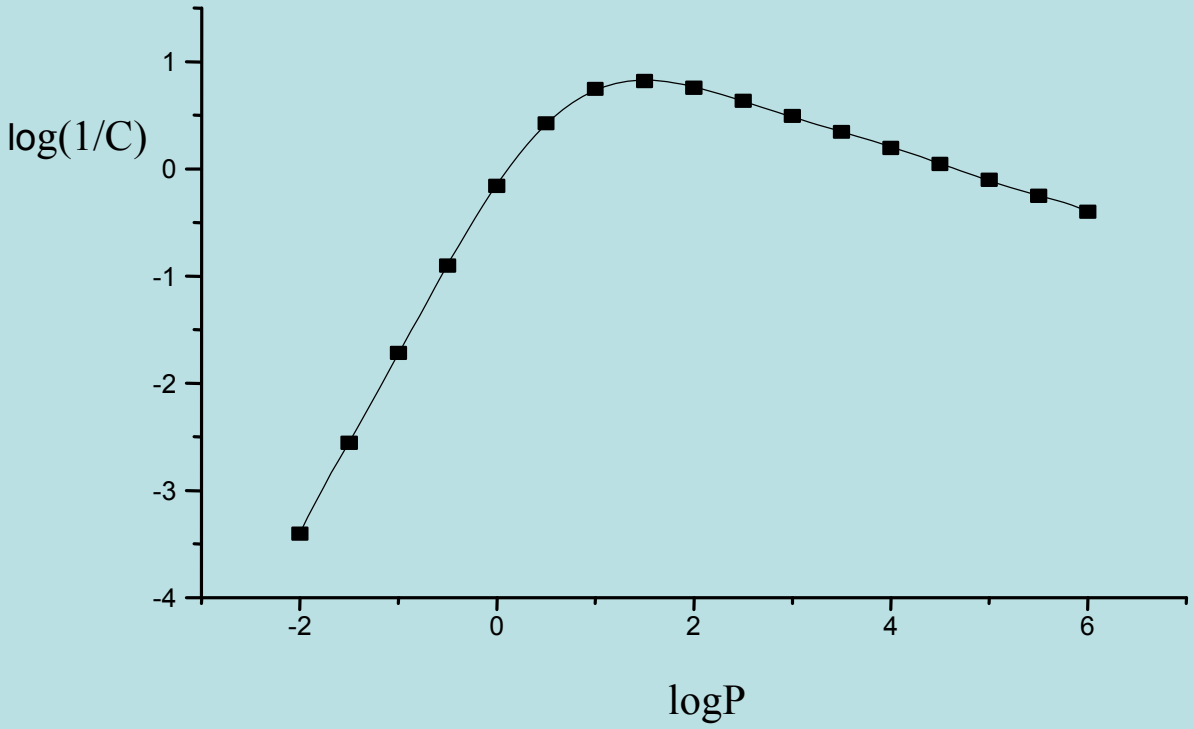


π- парціальна ліпофільність (замісника)

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

Білінійна модель (Kubiny)

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

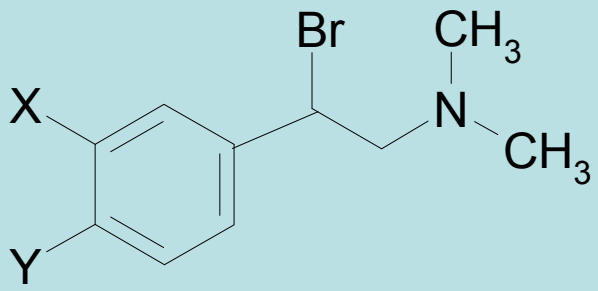


Валідація регресійних рівняннь процедура Leave-One-Out (LOO)

$$Y = a_0 + a_1X_1 + a_2X_2 + ... + a_MX_M$$

№	X_1	X_2	…	X_M	Y(exp)	Y(calc)	Y(pred)
1							
2							
3							
…							
N							

Фунгістатична активність



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

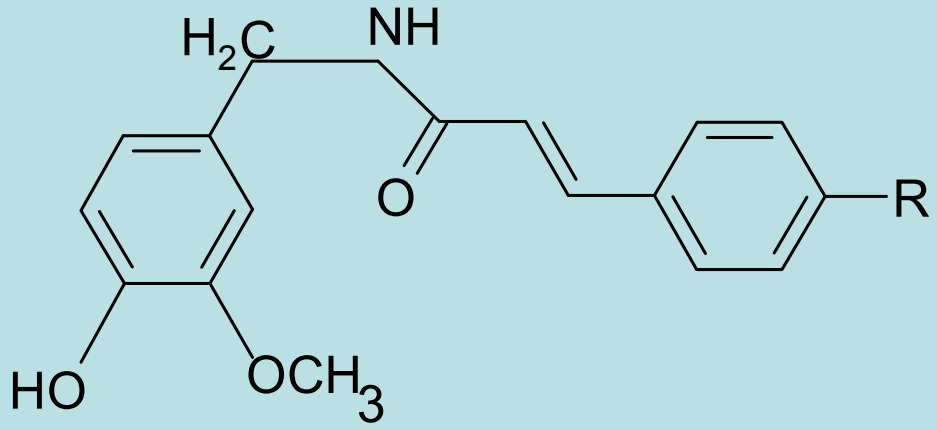
$$\log 1/C = 1.15\pi - 1.46\sigma^{+} + 7.82$$

токсичність набору ароматичних сполук по відношенню до *Tetrahymena pyriformis*: (війчасті)

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

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

Анальгетична дія Похідних капсаіцину



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

$$-\log EC_{50} = 1.137 - 0.068 \cdot MR, \quad R = 0.726, \; Q = 0.554$$

$$-\log EC_{50} = 0.770 - 0.815 \cdot \pi \quad R = 0.943, \; Q = 0.877$$

pKa органічних кислот

15 одноосновных насичених 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

заряд на кисні карбонільної групи (x1)
заряд на кисні оксі-групи (x2),
заряд на водні оксі-групи (x3),

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

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

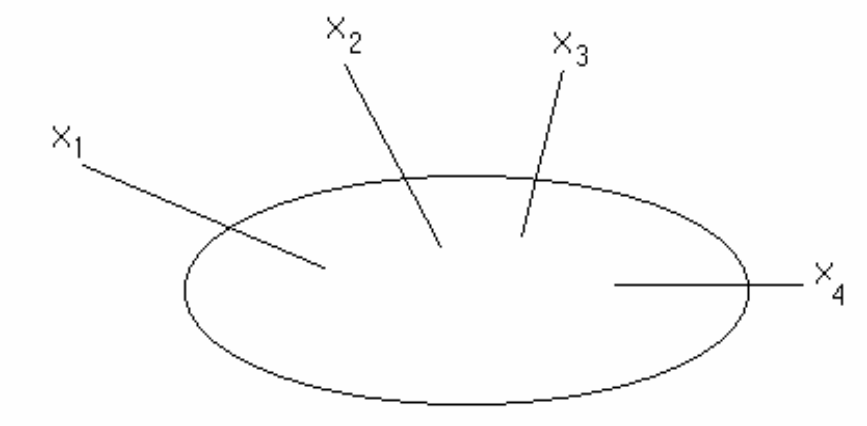
Адитивні схеми

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

Використовується для обчислення ліпофільностей (lgP), молекулярних рефракцій и т.ін.

χ_i - Парціальні величини (інкременты)

Один із варіантів адитивного методу - Метод Фри-Вільсона



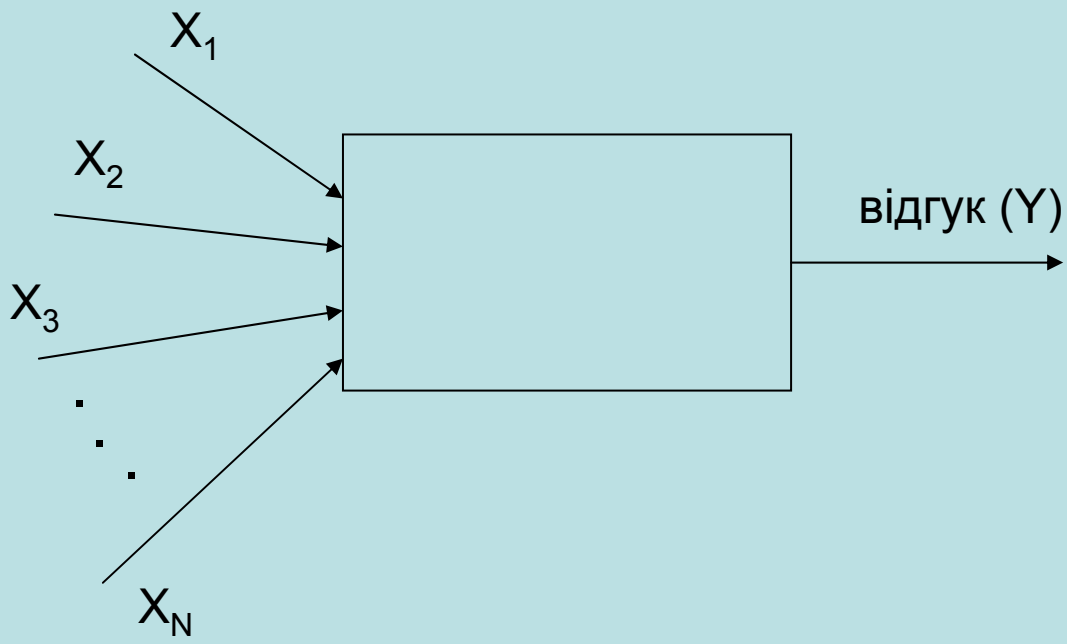
$$-\log C = y_0 + \sum_{i,j} n_{ij} \chi_{ij}$$

n_{ij} -Кількість замісників типу і в положенні j

Оцінка прогностичної способности: r, σ

Live-one-out cross-validation (LOO) – процедура перехресного оцінювання

Факторний аналіз. (аналіз головних компонент)



Фактор:

$$\varphi = c_y Y + c_1 X_1 + c_2 X_2 + ... + c_N X_N$$

Латентні змінні

Неповний метод найменших квадратів

Partial Least-Squares / Projection on Latent Structures (PLS)

число об'єктів (молекул) << число змінних (дескрипторів)

зазвичай для МНК $M \geq (3 \div 5)N$

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

1) PCR - Регресія головних компонентів (факторів)

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

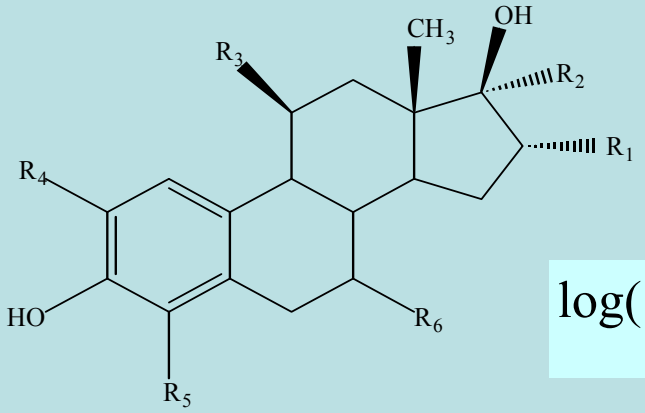
2) PLS - Регресія факторів з більшим внеском відгуку (y)

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

PC_i – фактори. Знаходяться з урахуванням факторної будови y

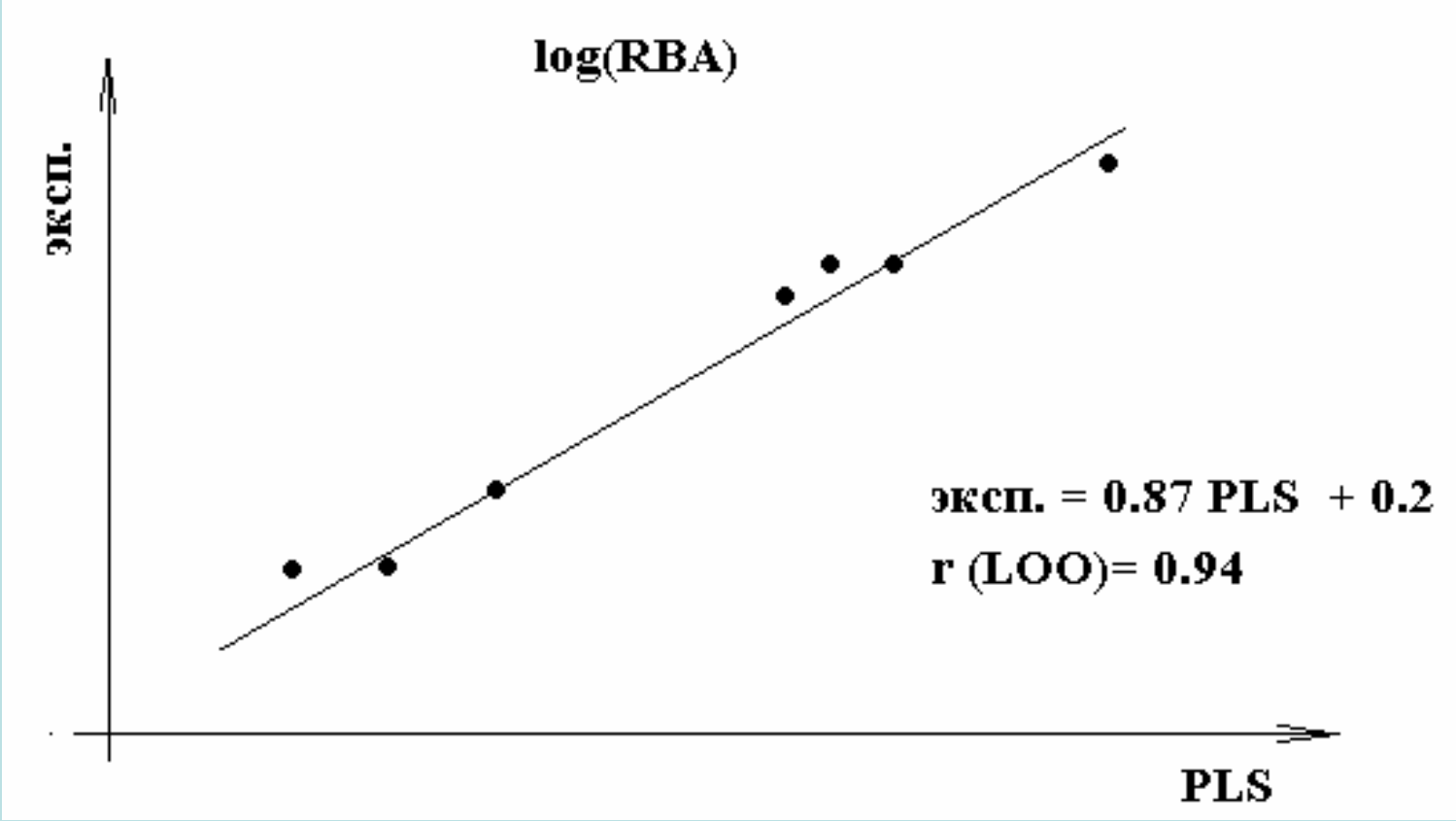
P << число об'єктів (молекул)

Активність естрадіолів (PLS)

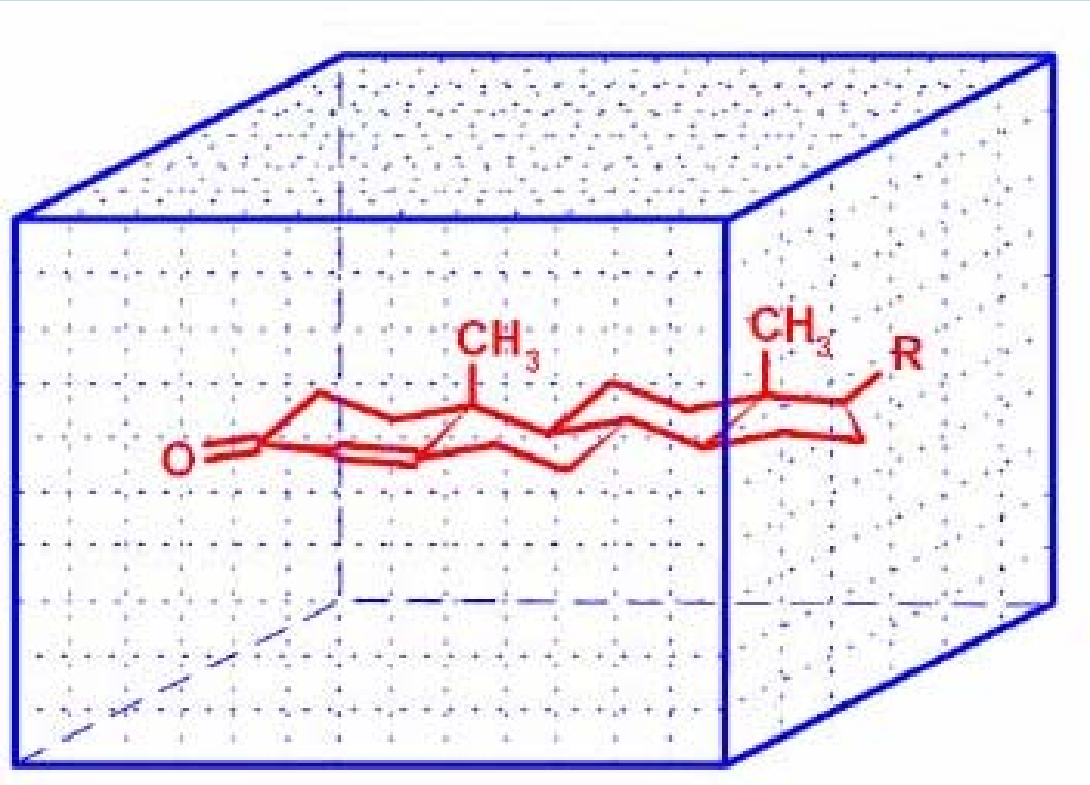


Активність: Log(RBA)

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



Comparative molecular field analysis (CoMFA) порівняльний аналіз молекулярних полів



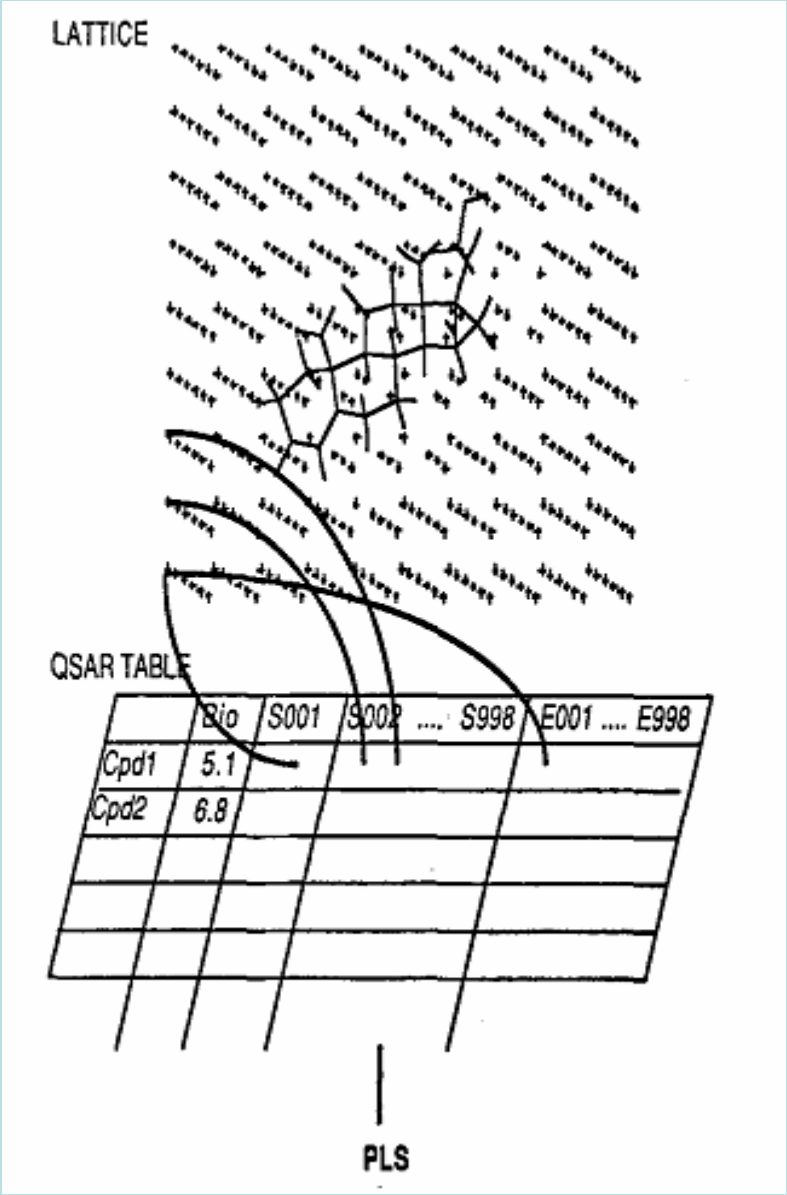
U_i - Електростатичний потенціал в точці

$$U_i = \sum_{i,M} \frac{q_M}{|r_i - r_M|}$$

S_i - стерический потенциал в точке

$$S_i = 4\sum_M \varepsilon_M \left(\left(\frac{\sigma_M}{r_i - r_M} \right)^{12} - \left(\frac{\sigma_M}{r_i - r_M} \right)^{6} \right)$$

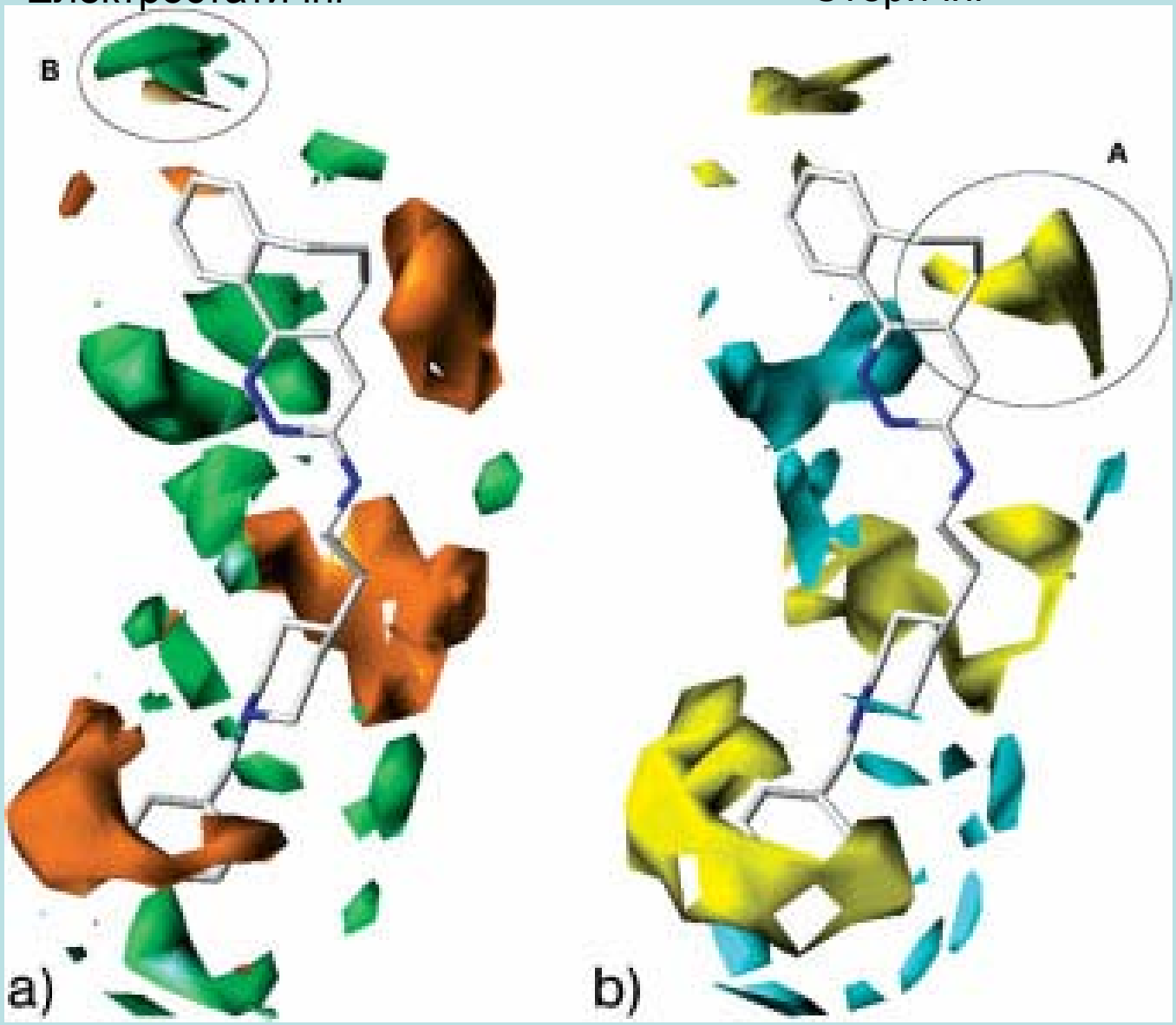
$$\log(1/C) = \sum_i a_i U_i + \sum_i b_i S_i$$


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

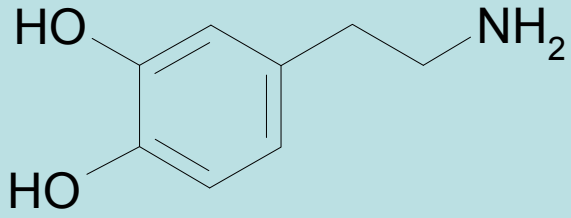
PLS коефіцієнти

Електростатичні

Стеричні



Дофамин (3,4-dihydroxyphenethylamine)- нейротрансміттер



Карта Електростатичного поля