ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

Лекция №21

Хемоинформатика
и
виртуальный скрининг

Новоселецкий Валерий Николаевич
к.ф.-м.н., доц. каф. биоинженерии
valery.novoseletsky@yandex.ru

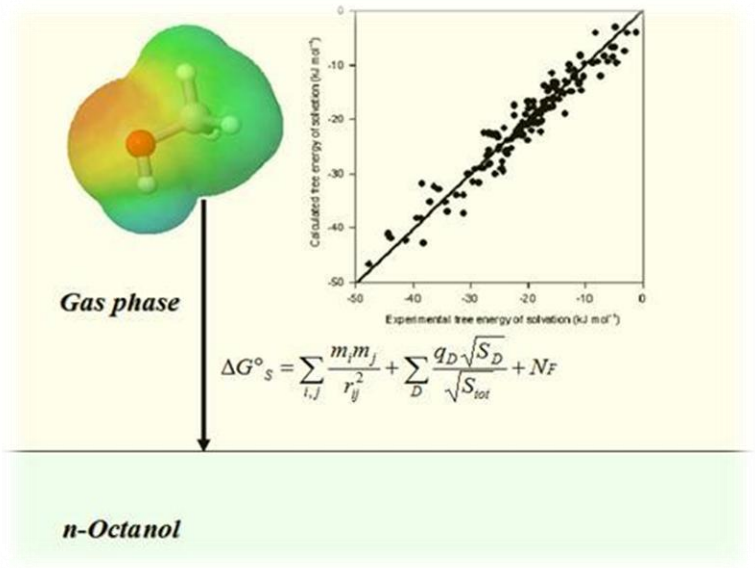
Сайт курса http://intbio.org/bioinf2019-2020

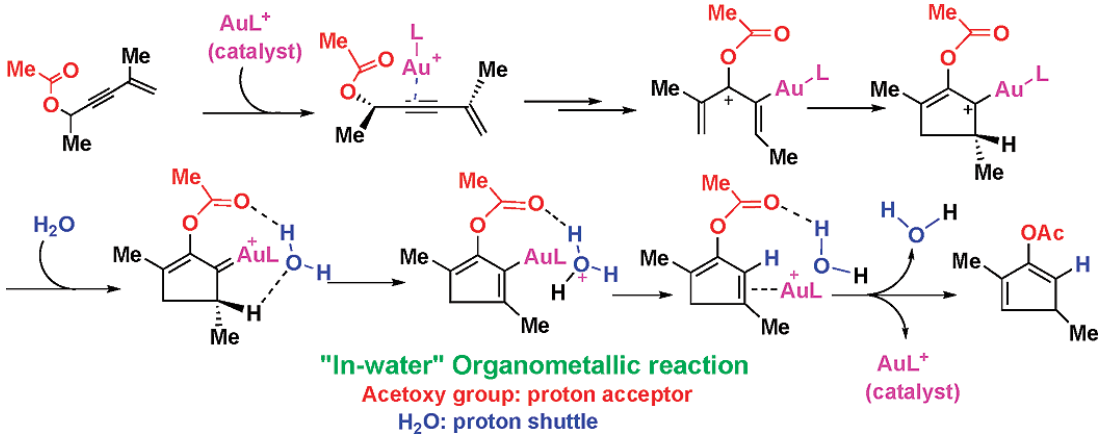
Хемоинформатика

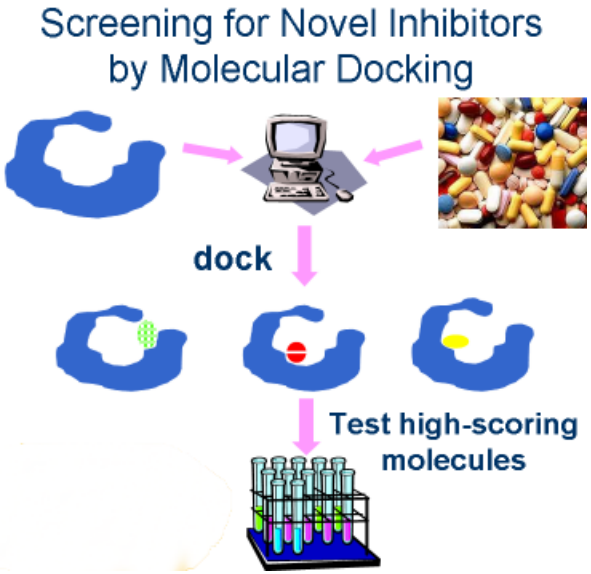
– это применение методов информатики для решения химических задач.

Область применения:

- Предсказание свойств химических соединений (QSPR)
- Поиск по химическому подобию, фармакофорный поиск, виртуальный скрининг
- Компьютерный синтез

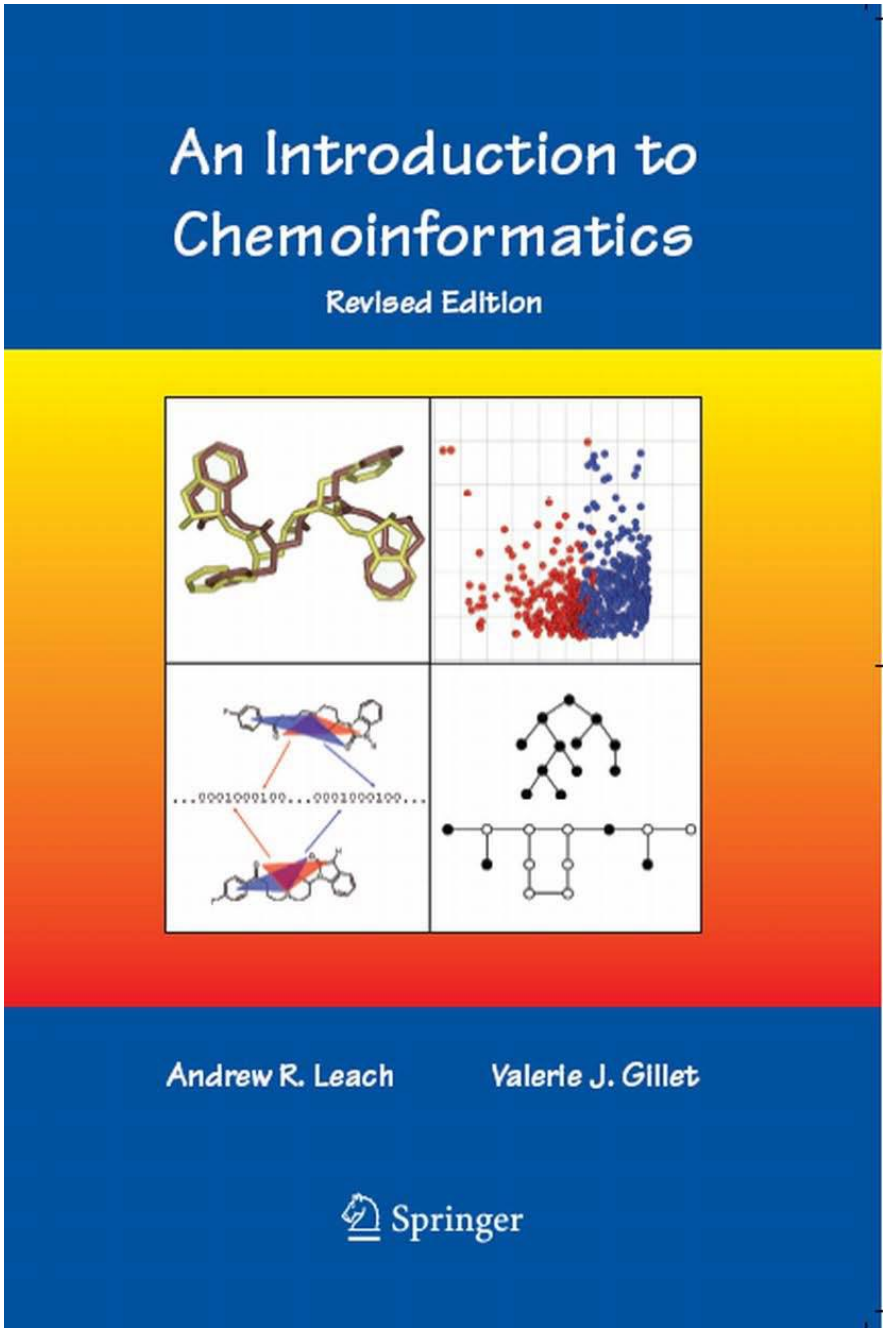




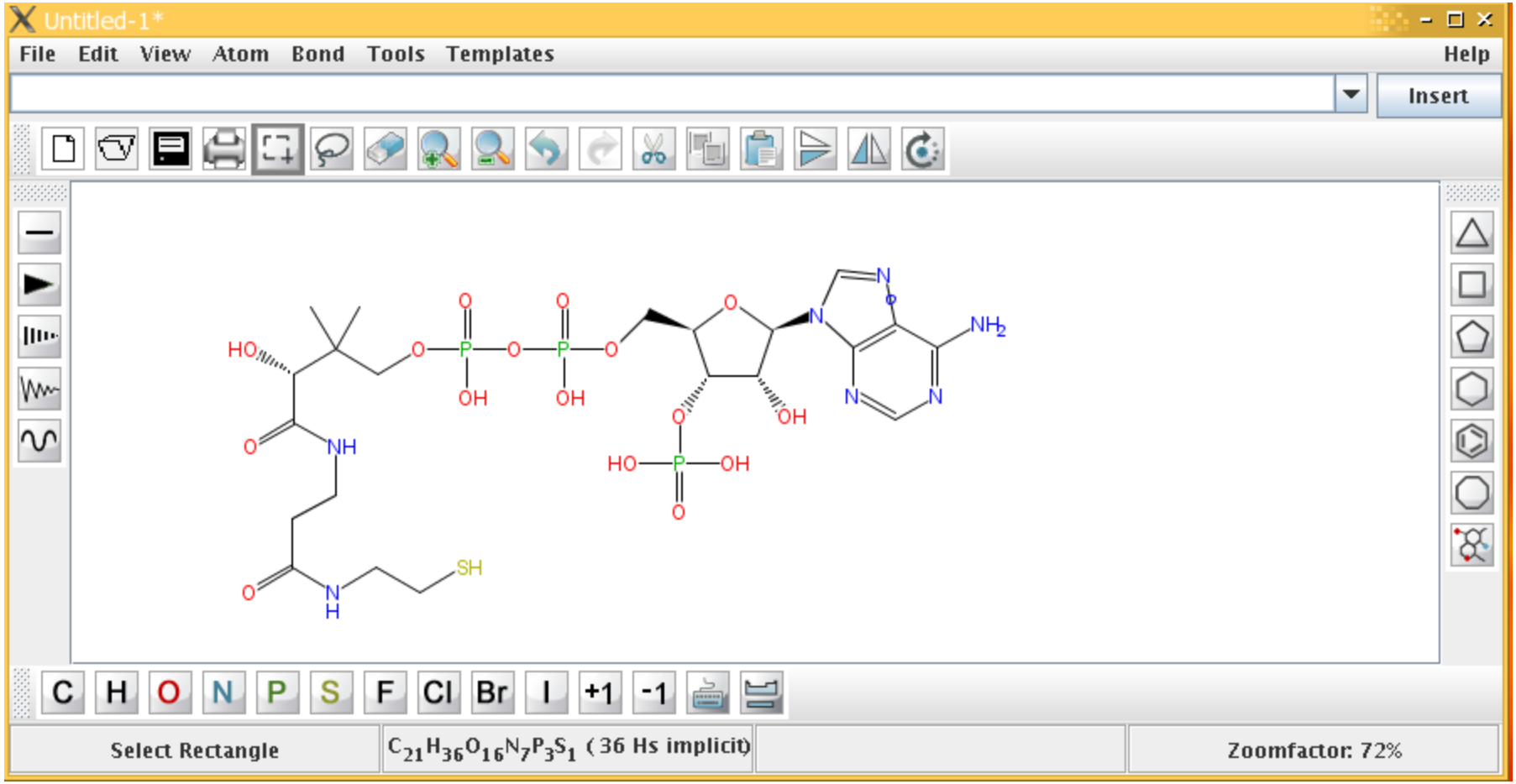


Хемоинформатика и все-все-все...

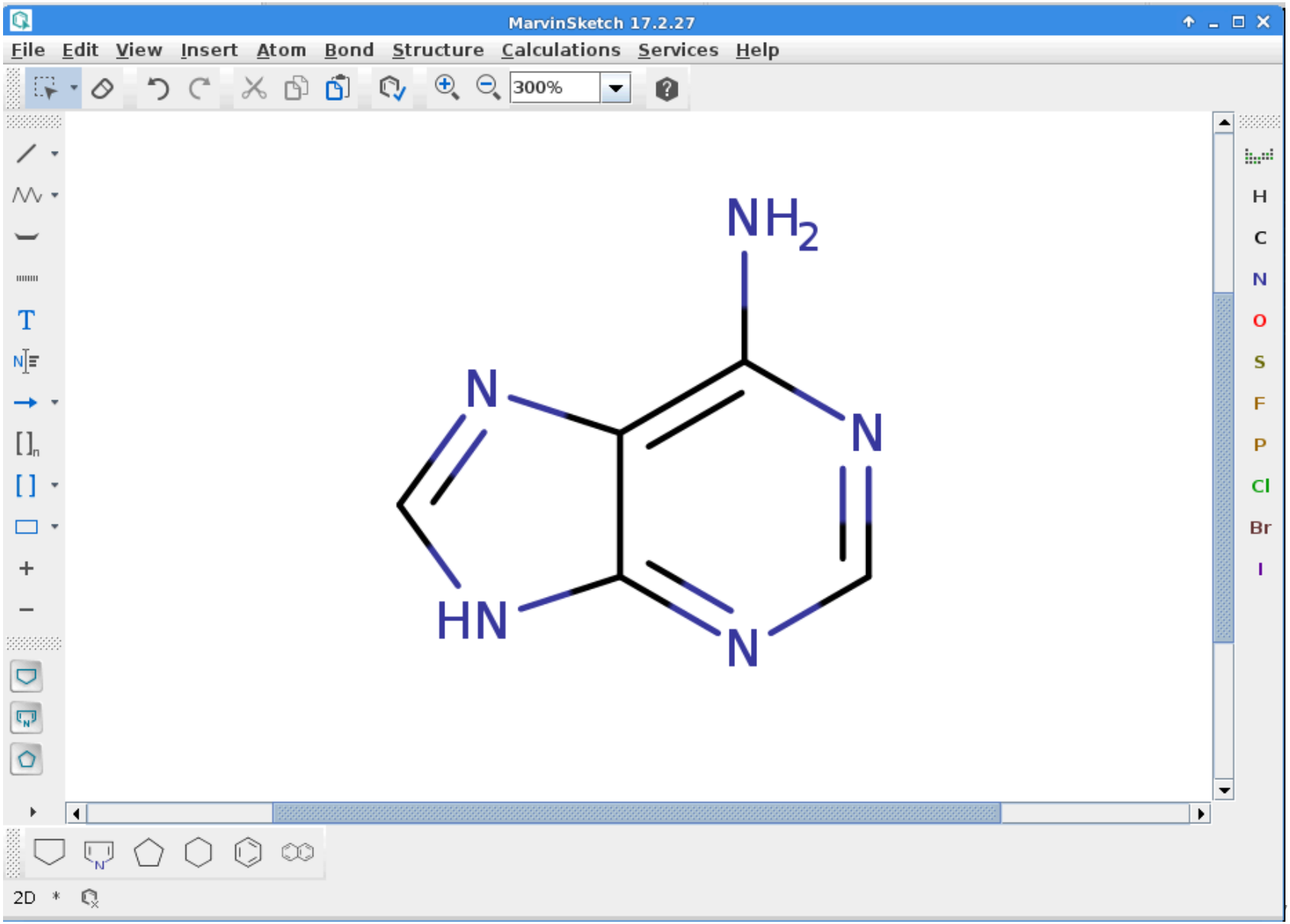
Что почитать?



Молекулярные редакторы. JChemPaint



Молекулярные редакторы. MarvinSketch



Представление структуры молекул

Молекулярный граф – связный неориентированный граф, находящийся во взаимно-однозначном соответствии со структурной формулой химического соединения таким образом, что вершинам графа соответствуют атомы молекулы, а рёбрам графа — химические связи между этими атомами.

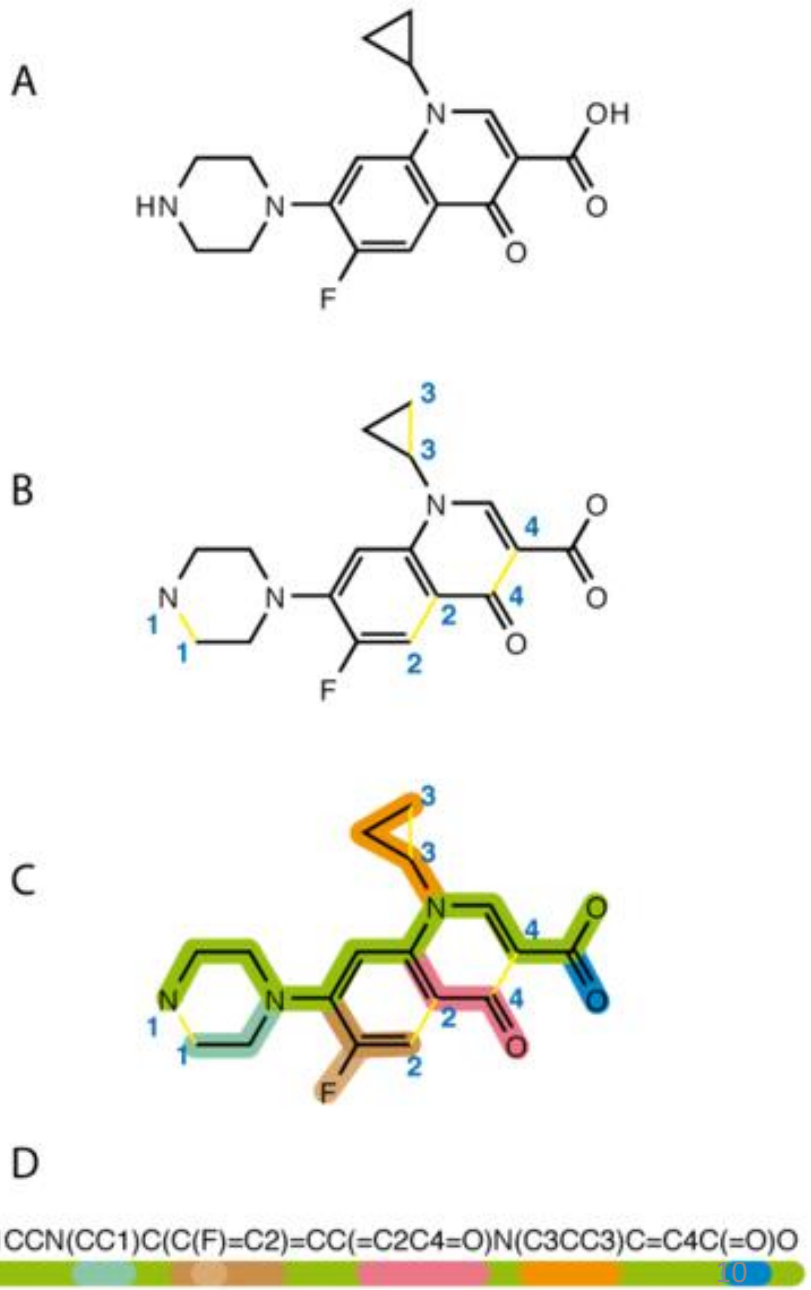
Способы записи:

- Линейные нотации (SMILES, SMARTS, SLN, InChI)
- Матрица смежности
- Структурные файлы (общие – MOL, SDF,…, специальные - MOL2, HIN,…)
- Chemical Markup Language

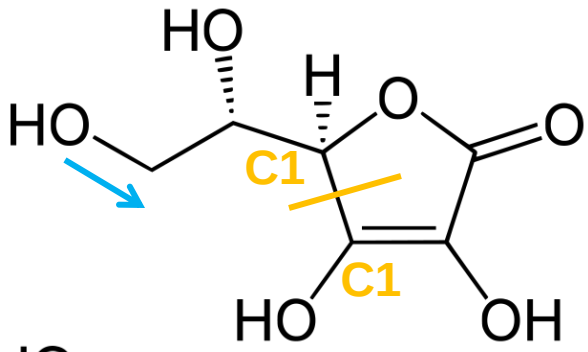
Линейные нотации. SMILES

SMILES (*Simplified Molecular Input Line Entry Specification, спецификация упрощенного представления молекул в строке ввода*) — система правил (спецификация) однозначного описания состава и структуры молекулы химического вещества с использованием строки символов **ASCII** (1980е - …).

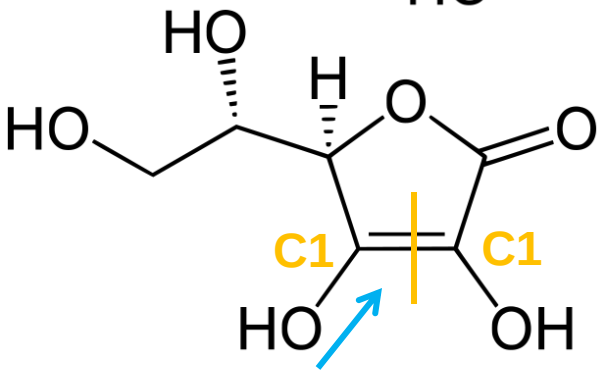
Вода	O
Этанол	CCO
Углекислый газ	O=C=O
Синильная кислота	C#N
Циклогексан	C1CCCCC1
Бензол	c1ccccc1



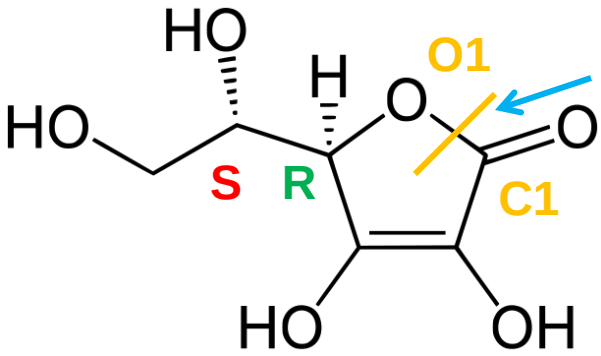
Линейные нотации. SMILES



OCC(O)C1OC(=O)C(O)=C1(O)



OC=1C(OC(=O)C=1O)[C@@H](O)CO



O=C1C(O)=C(O)[C@H](O1)[C@@H](O)CO

@ - перечисление заместителей по убыванию старшинства против часовой стрелки

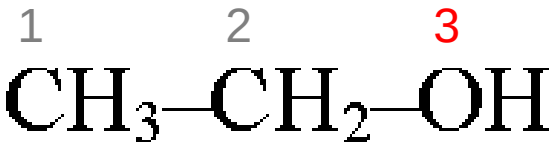
@@ - по часовой

https://jacobdurrant.com/2014/05/20/online-smiles-string-to-3d-visualizer/

Линейные нотации. InChI

InChI (*International Chemical Identifier*) — текстовый идентификатор химического соединения для стандартизации кодирования молекулярной информации и представления её в читаемом виде (2005 - …).

Этанол InChI=1/C2H6O/c1-2-3/h3H,2H2,1H3

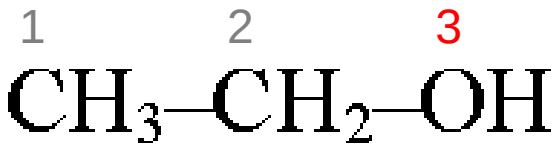


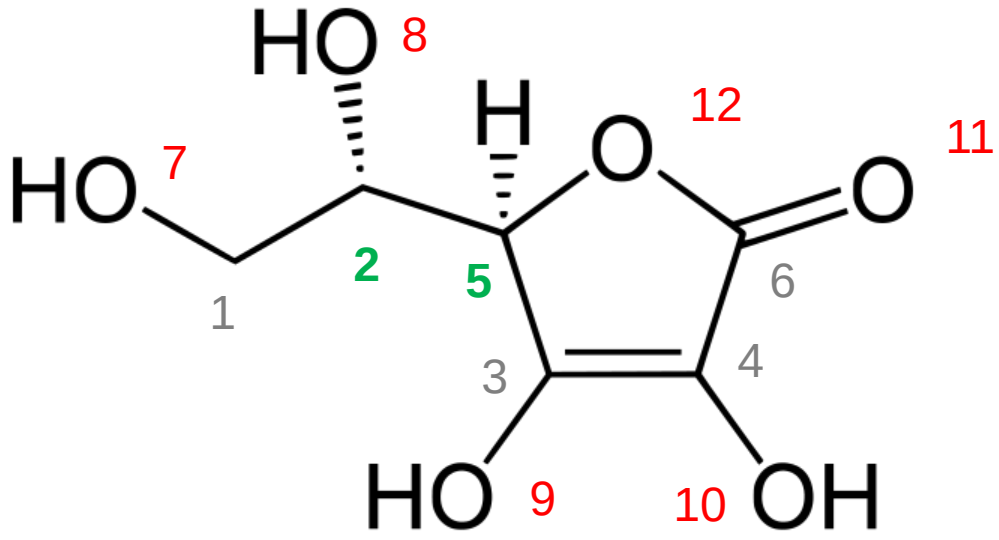
Линейные нотации. InChI

InChI (*International Chemical Identifier*) — текстовый идентификатор химического соединения для стандартизации кодирования молекулярной информации и представления её в читаемом виде (2005 - …).

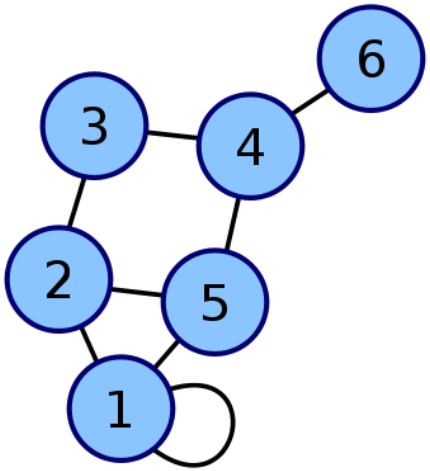
Этанол InChI=1/C2H6O/c1-2-3/h3H,2H2,1H3

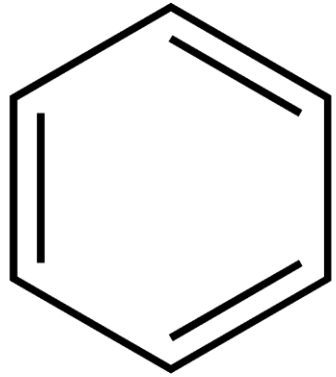
Аскорбиновая кислота InChI=1/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H,1H2/t2-,5+/m0/s1

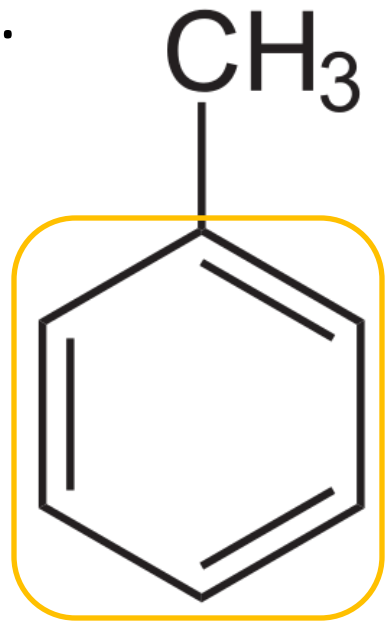




Представление структуры молекул. Матрица смежности







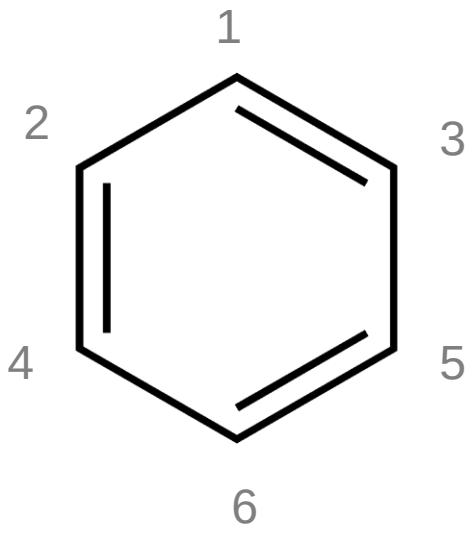
$$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

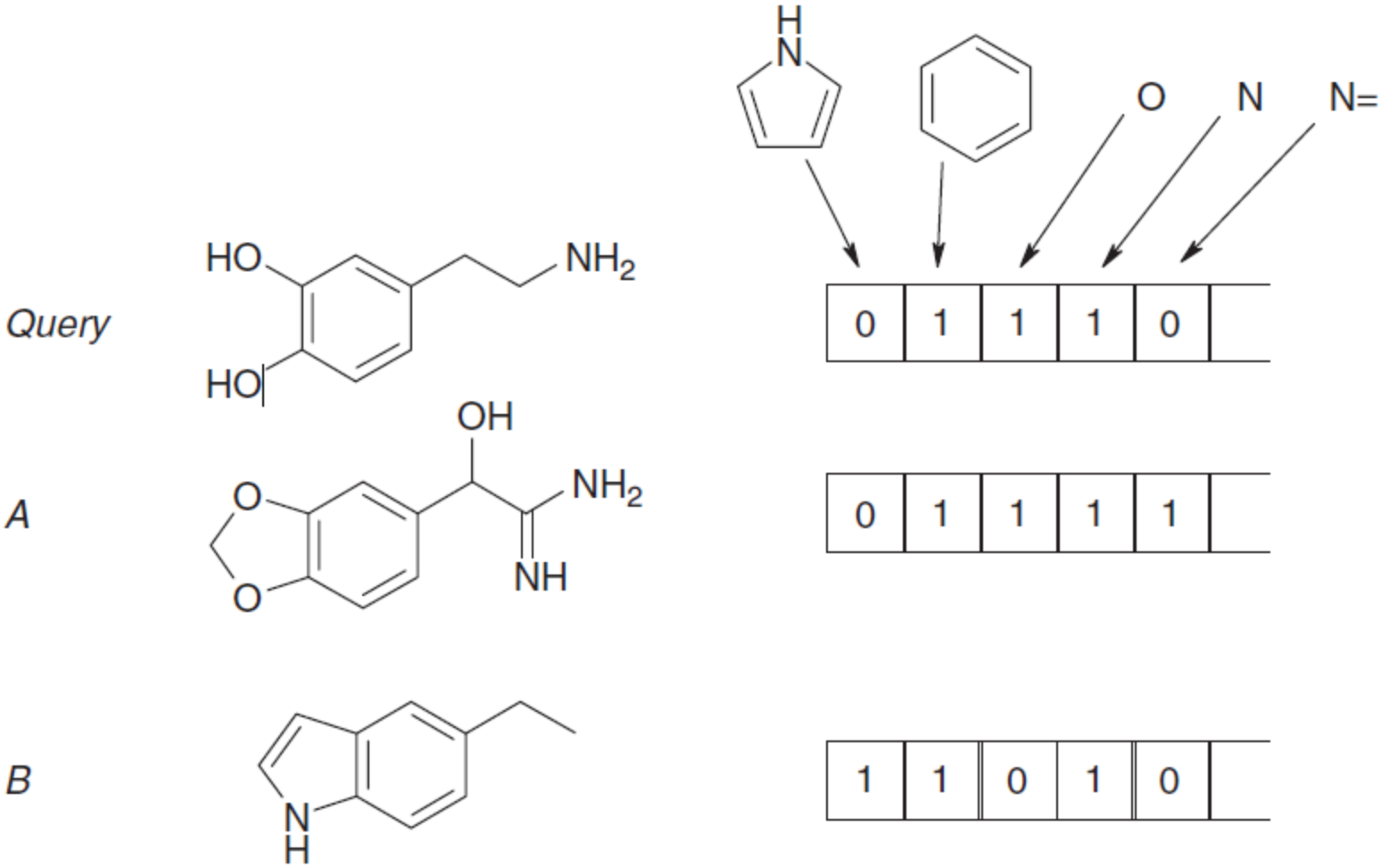
$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Представление структуры молекул. Структурные файлы. MOL

```
benzene
ACD/Labs0812062058
6  6  0  0  0  0  0  0  0  0  1 V2000
    1.9050   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.9050   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7531   -0.1282    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7531   -2.7882    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3987   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3987   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0  0  0  0
  3  1  2  0  0  0  0
  4  2  2  0  0  0  0
  5  3  1  0  0  0  0
  6  4  1  0  0  0  0
  6  5  2  0  0  0  0
 M  END
```



Представление структуры молекул. Битовые строки



Предсказание свойств соединений

Поиск количественных соотношений структура-свойство — процедура построения моделей, позволяющих по структурам химических соединений предсказывать их разнообразные свойства.

Основная гипотеза – сходные соединения имеют сходные свойства.

QSPR - *Quantitative Structure-Property Relationship* – физические и физико-химические свойства:

Температуры плавления и кипения

Вязкость

Давление насыщенных паров

Плотность

Химические сдвиги в спектрах ^{1}H ЯМР

Растворимость

...

QSAR - *Quantitative Structure-Activity Relationship* – биологические свойства

Предсказание свойств соединений

Молекулярные дескрипторы:

"*The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment.*" (Todeschini and Consonni, 2000)

- **Теоретические** (число кратных связей, наличие молекулярных фрагментов, число доноров и акцепторов Н-связей, ...)

- **Экспериментальные** (гидрофобность, поляризуемость, показатель преломления, ...)

Дескрипторы инвариантны,

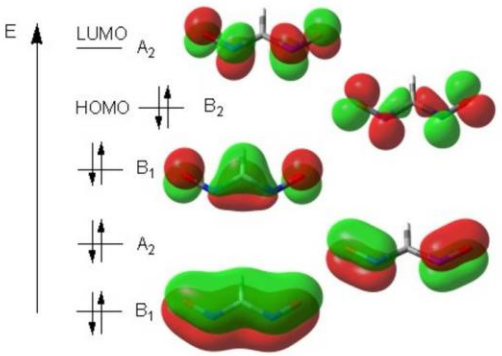
т.е. не зависят от положения молекулы в пространстве.

Молекулярные дескрипторы

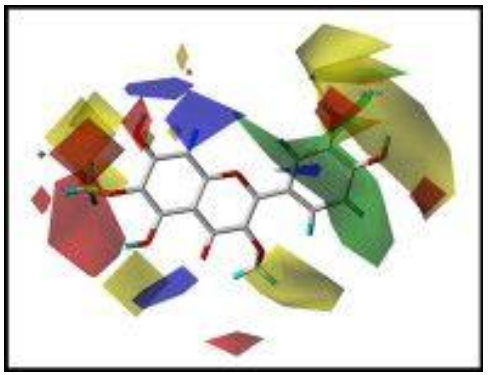
Фрагментные дескрипторы – отражают факт наличия фрагмента в молекулярном графе (*бинарные*) или число вхождений фрагмента (*целочисленные*)

Физико-химические дескрипторы – соответствуют измеряемым физ-хим величинам (липофильность (LogP), молярная рефракция (MR), молекулярный вес (MW), молекулярные объемы и площади поверхностей,...)

Квантово-химические дескрипторы — величины, получаемые в результате квантово-химических расчетов (энергии граничных орбиталей, частичные заряды на атомах, порядки связей,...)



Дескрипторы молекулярных полей — величины, аппроксимирующие значения молекулярных полей путем вычисления энергии взаимодействия пробного атома, помещенного в узел решетки, с рассматриваемой молекулой



Другие…

Предсказание свойств соединений

В самой общей форме:

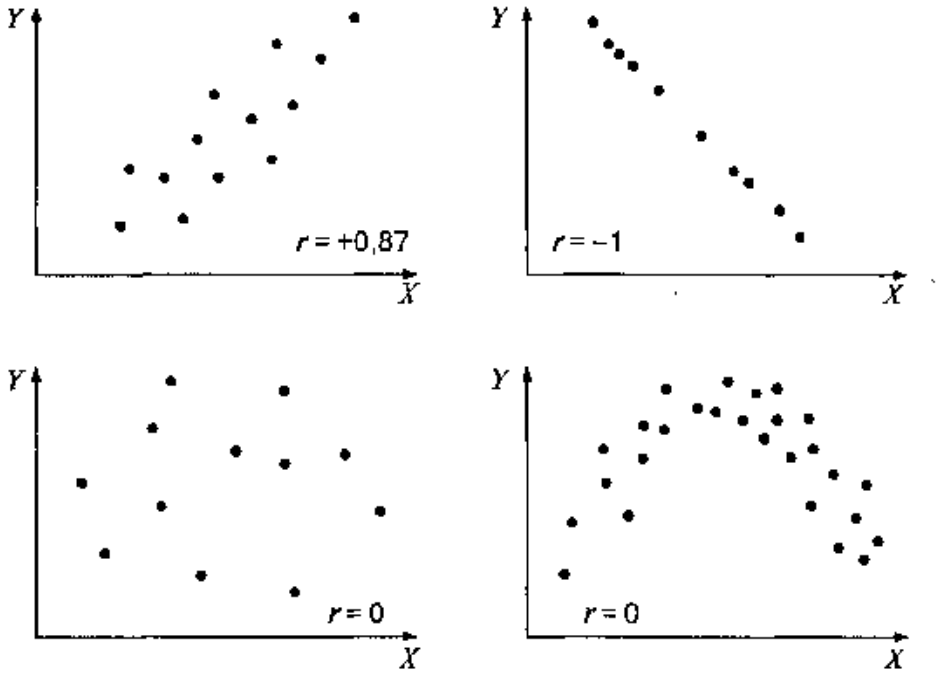
значение свойства – это некая функция от некого набора дескрипторов.

Цель: найти оптимальную функцию и оптимальный набор.

Выявленная связь должна быть проверена =>

- **Сравнение модели с экспериментом (коэффициент корреляции)**
- Разбиение данных на обучающую и тестовую выборки:
- - Перекрестная проверка (cross-validation) (для маленьких выборок)
- - Рандомизация (для больших выборок)

Сравнение модели с экспериментом. Коэффициент корреляции



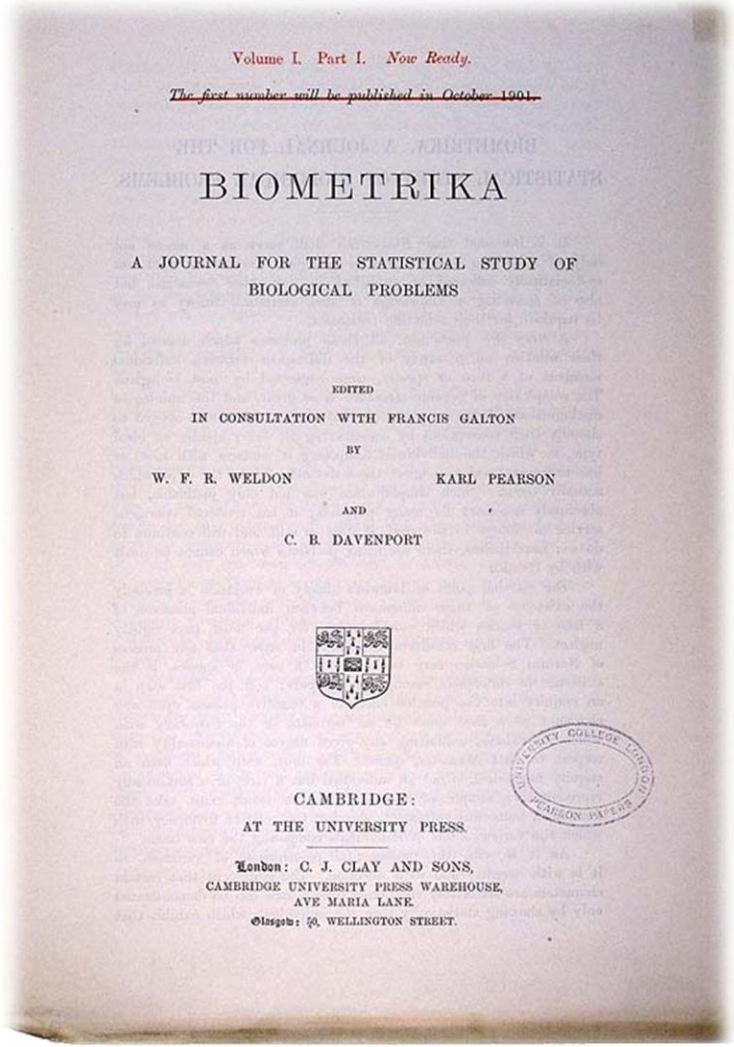
Карл Пирсон
(1857 – 1936)

$$r_{XY} = \frac{\mathrm{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}.$$

Отсутствие корреляции означает неадекватность выбранной модели

Наличие корреляции не означает причинно-следственной связи

Сравнение модели с экспериментом. Коэффициент корреляции



Volume I. Part I. *Now Ready.*

~~*The first number will be published in October 1901.*~~

BIOMETRIKA

A JOURNAL FOR THE STATISTICAL STUDY OF BIOLOGICAL PROBLEMS

EDITED
IN CONSULTATION WITH FRANCIS GALTON
BY
W. F. R. WELDON KARL PEARSON
AND
C. B. DAVENPORT

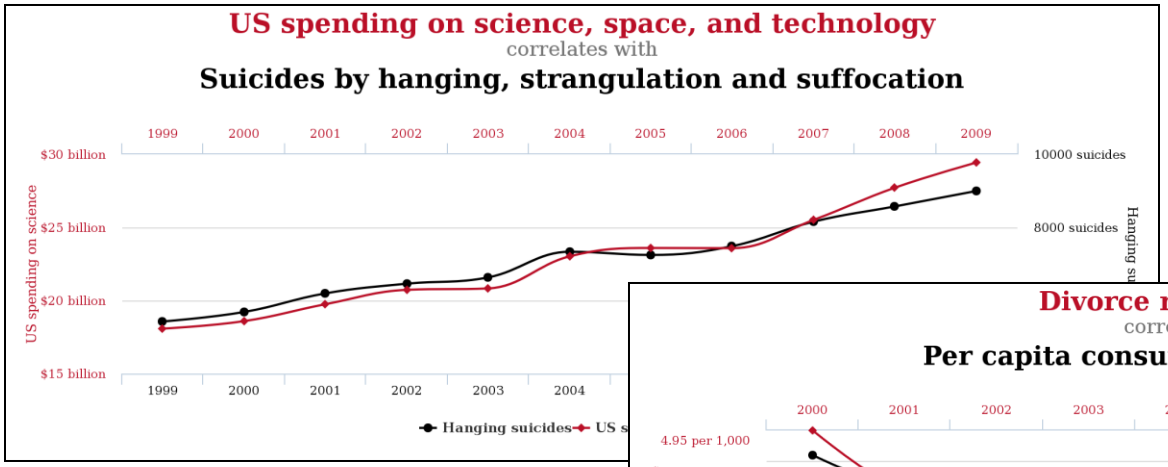
CAMBRIDGE:
AT THE UNIVERSITY PRESS.

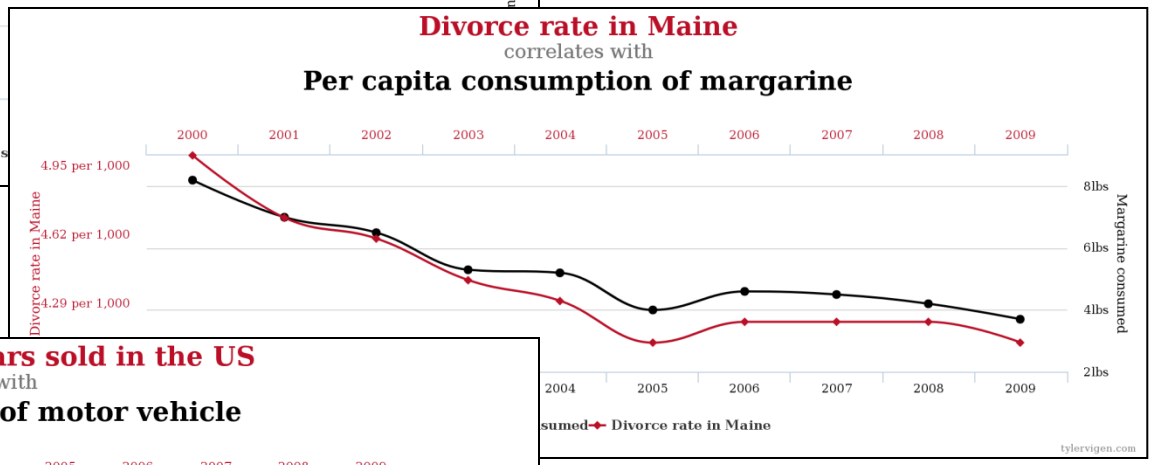
London: C. J. CLAY AND SONS,
CAMBRIDGE UNIVERSITY PRESS WAREHOUSE,
AVE MARIA LANE.
Glasgow: 50, WELLINGTON STREET.

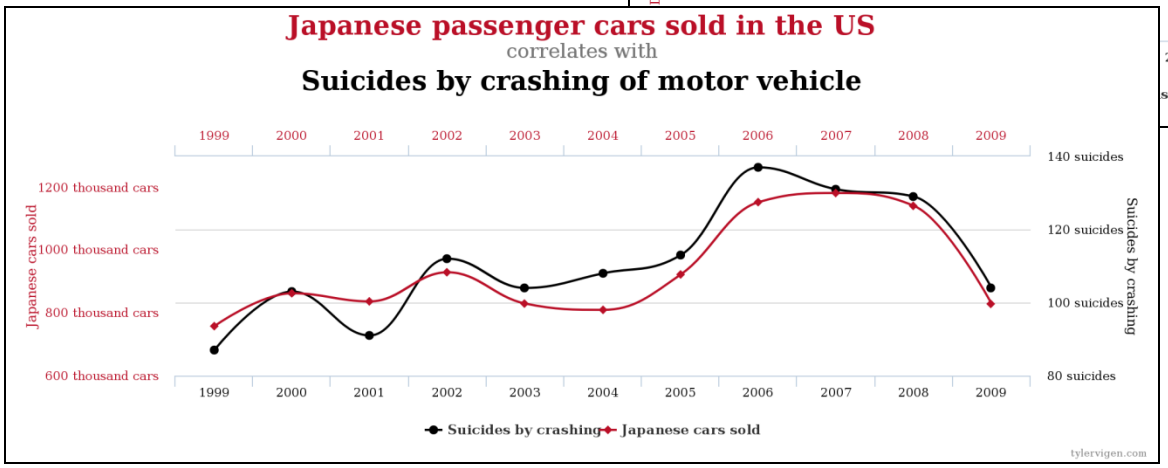
Карл Пирсон
(1857 – 1936)

Неожиданные корреляции

http://tylervigen.com/spurious-correlations





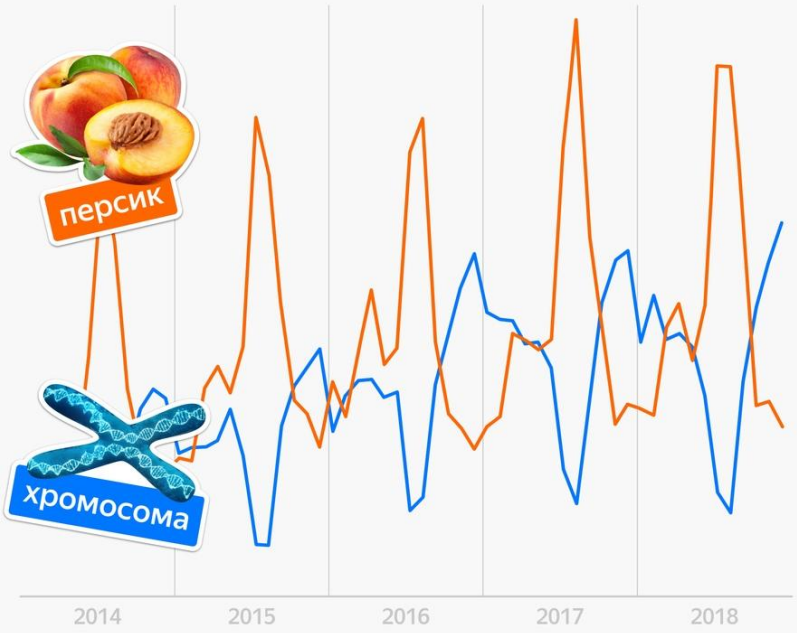


Наличие корреляции не всегда что-то означает. Хотя… ☺

Неожиданные корреляции

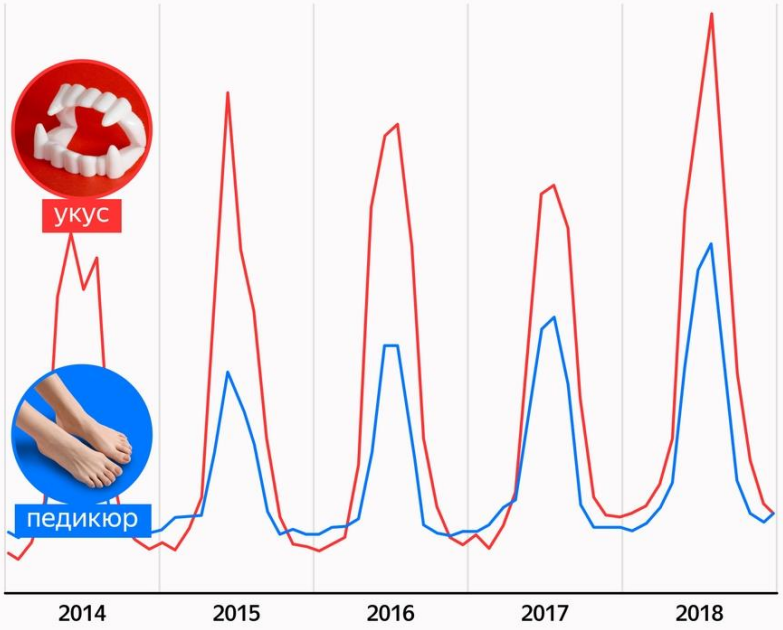
Парадоксы в Поиске — Яндекс

Когда в Поиске снижается доля запросов со словом хромосома, становится больше запросов со словом персик



Парадоксы в Поиске — Яндекс

Когда в Поиске повышается интерес к педикюру, взлетает доля запросов со словом укус



Наличие корреляции не всегда что-то означает. Хотя... ☺

Предсказание свойств соединений

В самой общей форме:

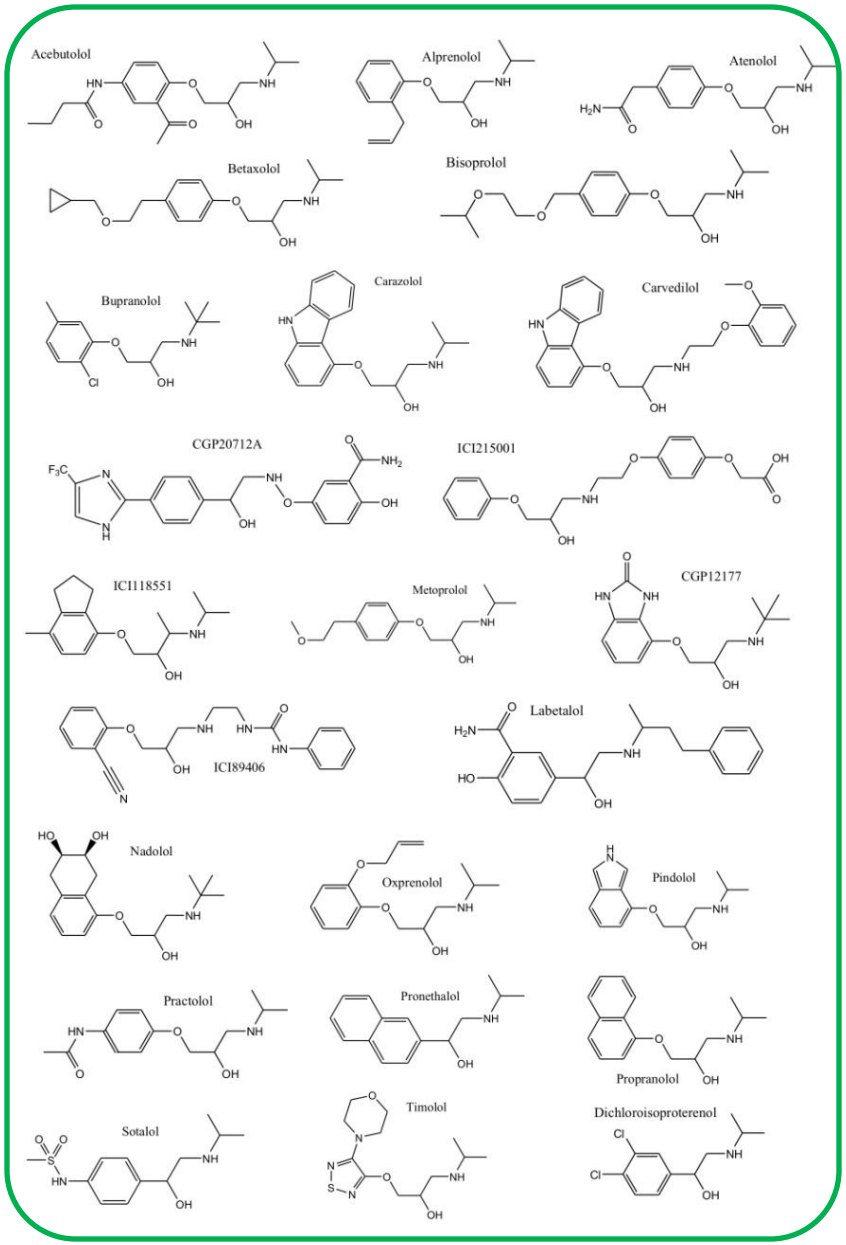
значение свойства – это некая функция от некого набора дескрипторов.

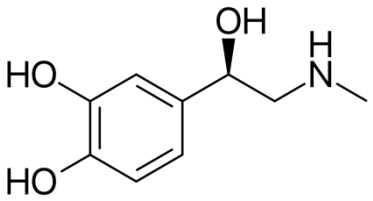
Цель: найти оптимальную функцию и оптимальный набор.

Выявленная связь должна быть проверена =>

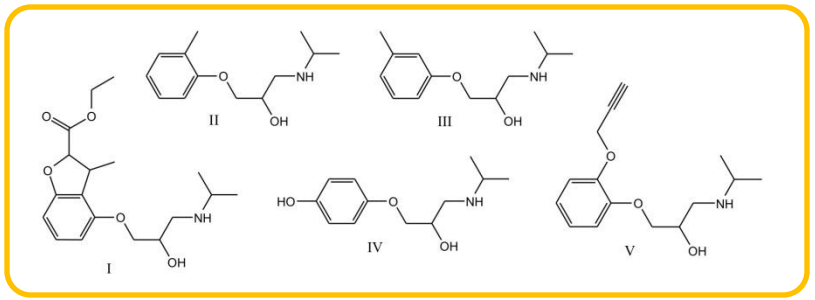
- **Сравнение модели с экспериментом (коэффициент корреляции)**
- **Разбиение данных на обучающую и тестовую выборки:**
- - Перекрестная проверка (cross-validation) (для маленьких выборок)
- - Рандомизация (для больших выборок)

Обучающая и тестовая выборки





Лиганды бета2-адренэргического рецептора

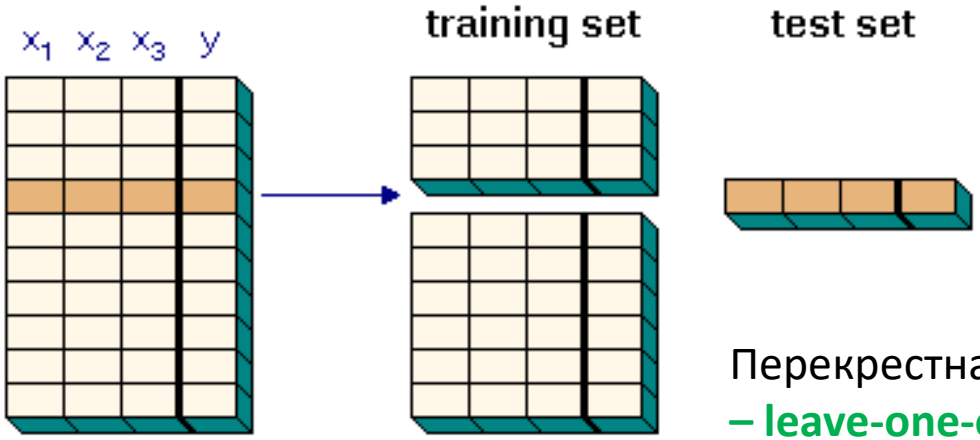


Выборки должны быть репрезентативными, т.е. отражать генеральную совокупность со всей возможной полнотой

r – коэффициент корреляции по обучающей выборке (хорошо, когда > 0,9)

q – коэффициент корреляции по тестовой выборке (приемлемо, когда > 0,6)

Перекрестная проверка и рандомизация



Перекрестная проверка с одним исключенным
– leave-one-out cross-validation

Для изучения свойств деревьев необязательно рассматривать каждое дерево в лесу

Но! Важно выбрать именно типичные деревья!

Аналогично в клинических исследованиях

The Literary Digest

NEW YORK — OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the Union, is now finished, and in the table below we record the figures received up to the hour of going to press.

These figures are exactly as received from more than one in every five voters polled in our country—they are neither weighted, adjusted nor interpreted.

Never before in an experience covering more than a quarter of a century in taking polls have we received so many different varieties of criticism—praise from many; condemnation from many others—and yet it has been just of the same type that has come to us every time a Poll has been taken in all these years.

A telegram from a new
nia asks: "Is it true
has purchased The Lite
telephone message only
these lines were written:

lican National Committee purchased The Literary Digest?" And all types and varieties, including: "Have the Jews purchased The Literary Digest?" "Is the Pope of Rome a stockholder of The Literary Digest?" And so it goes—all equally absurd and amusing. We could add more to this list, and yet all of these questions in recent days are but repetitions of what we have been experiencing all down the years from the very first Poll.

Problem—Now, are the figures in this Poll correct? In answer to this question we will simply refer to a telegram we sent to a young man in Massachusetts the other day in answer to his challenge to us to wager $100,000 on the accuracy of our Poll. We

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens to-day, the Hon. James A. Farley, Chairman of the Democratic National Committee. This is what Mr. Farley said October 14, 1932:

" 'Any sane person can not escape the implication of such a gigantic sampling of popular opinion as is embraced in The Literary Digest straw vote. I consider this conclusive evidence as to the desire of the people of this country for a change in the National Government. The Literary Digest poll is an achievement of no little magnitude. It is a Poll fairly and correctly conducted.' "

table of the voters from

10 млн анкет, 2,3 млн ответов…

И поражение Лэндона со счётом 8 : 523

Какие свойства можно предсказывать?

Физические свойства индивидуальных низкомолекулярных соединений

- Температура кипения (BP)
- Вязкость
- Плотность
- Показатель преломления
- Температура плавления (MP)
- Константы ионизации (кислотности или основности)
- ...

Спектроскопические свойства

- Положение длинноволновой полосы поглощения симметричных цианиновых красителей
- Химические сдвиги в спектрах ^{1}H ЯМР
- ...

Физические свойства, обусловленные межмолекулярными взаимодействиями

- Растворимость в воде (LogSw)
- Коэффициент распределения *n*-октанол/вода (LogP)
- ...

Физические и физико-химические свойства полимеров

- Показатель преломления полимеров
- Коэффициент проницаемости через полиэтилен низкой плотности
- ...

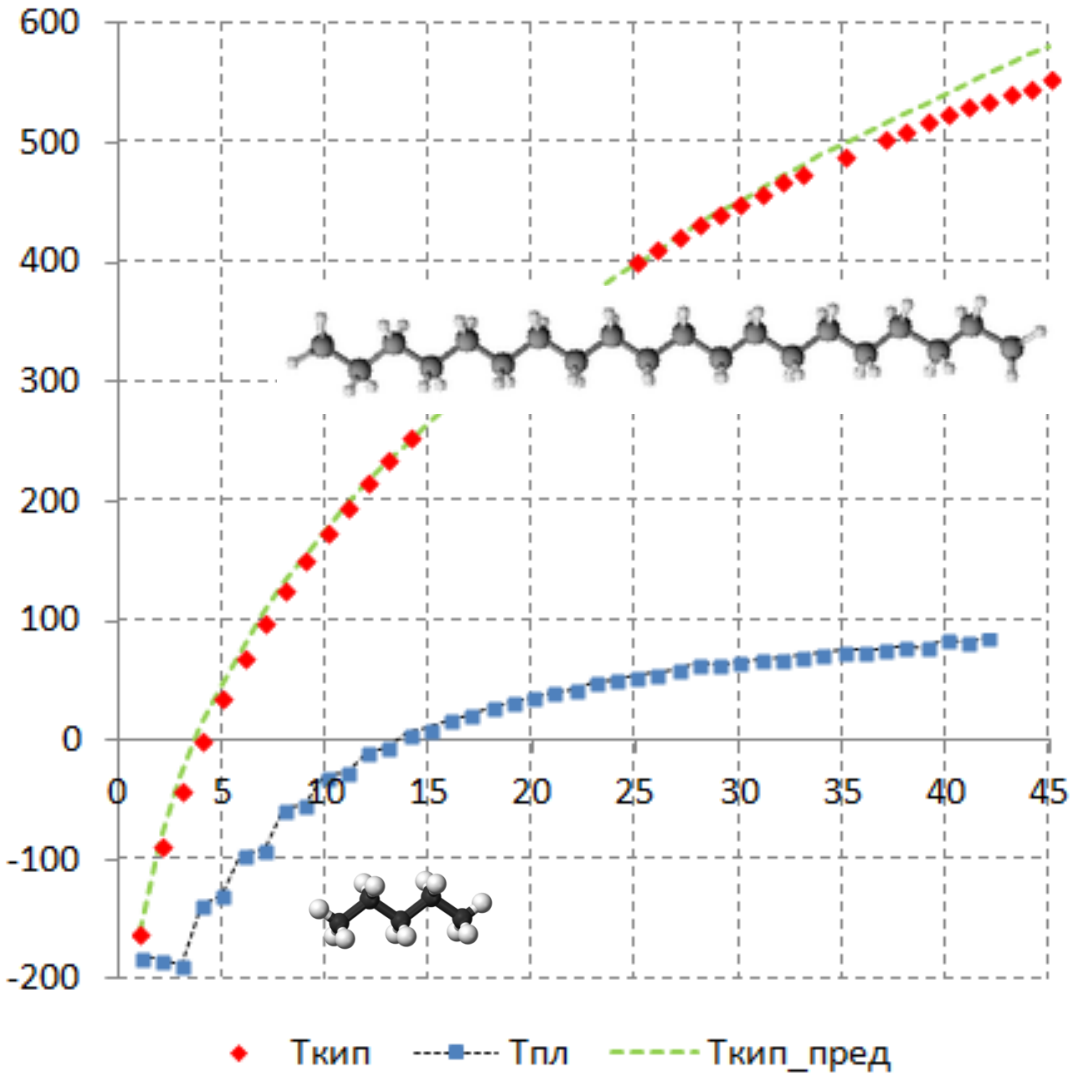
Предсказание температур кипения

Предсказание температуры кипения линейных алканов

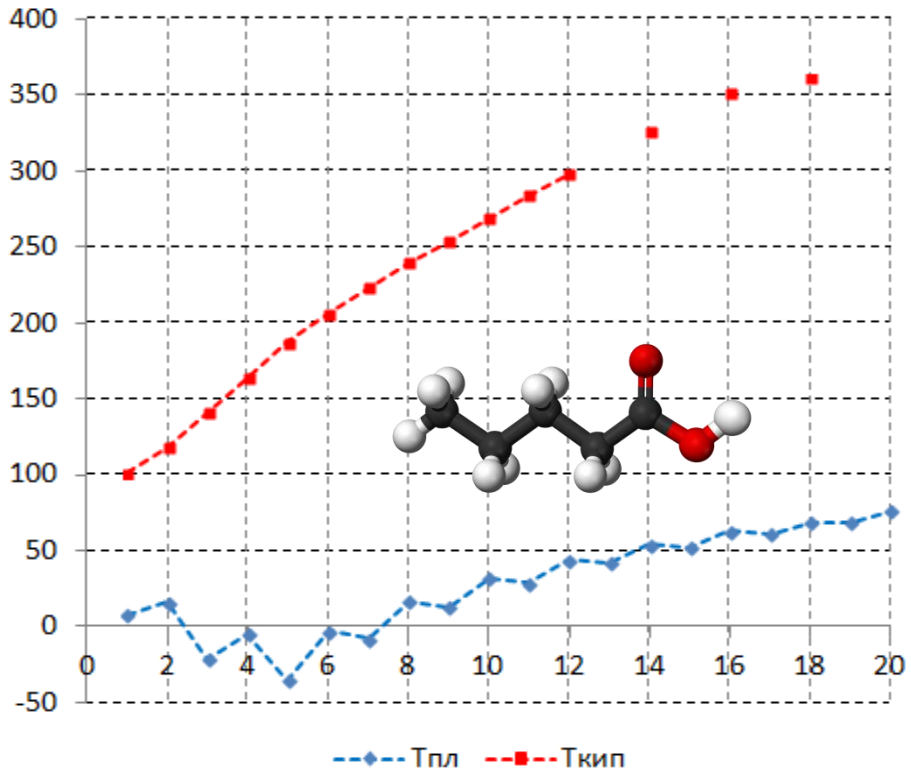
$CH_3 — (CH_2)_{n-2} — CH_3$

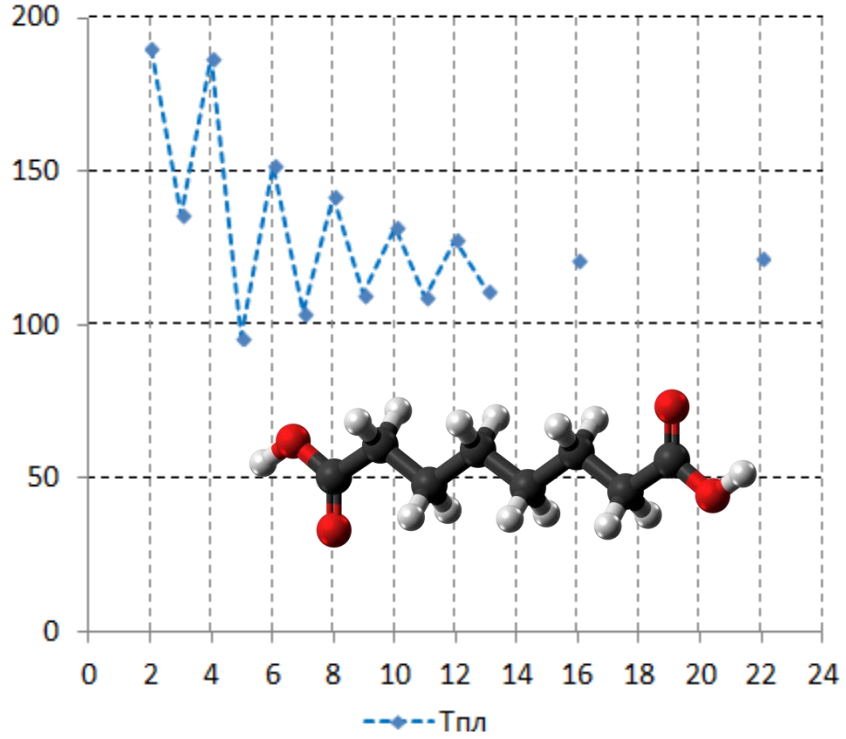
$$T(n) = 295 \cdot n^{0.33} - 455$$

$$R^2 = 0.999645$$



Температуры фазовых переходов карбоновых и дикарбоновых кислот





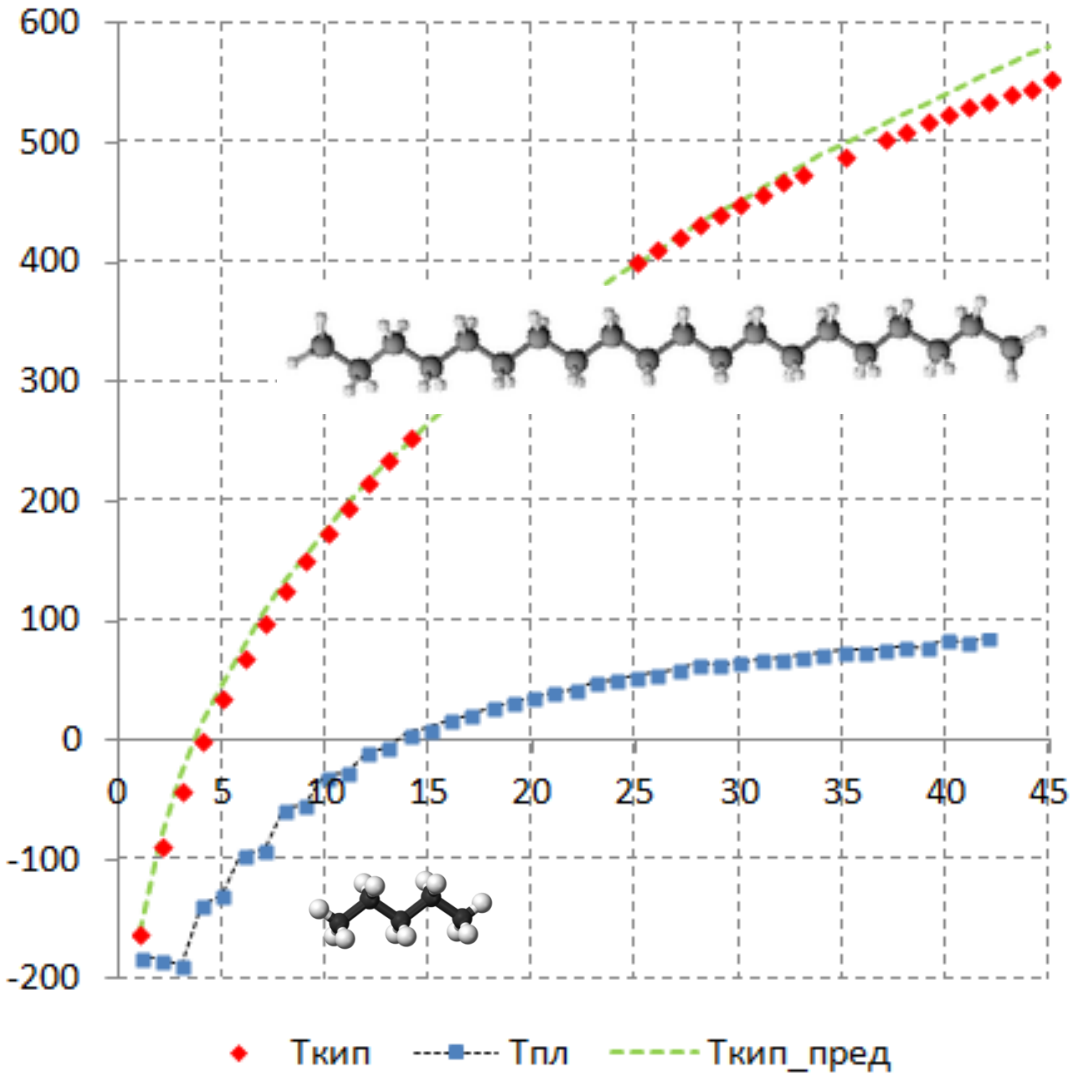
Предсказание температур кипения

Предсказание температуры кипения линейных алканов

$CH_3 — (CH_2)_{n-2} — CH_3$

$$T(n) = 295 \cdot n^{0.33} - 455$$

$$R^2 = 0.999645$$

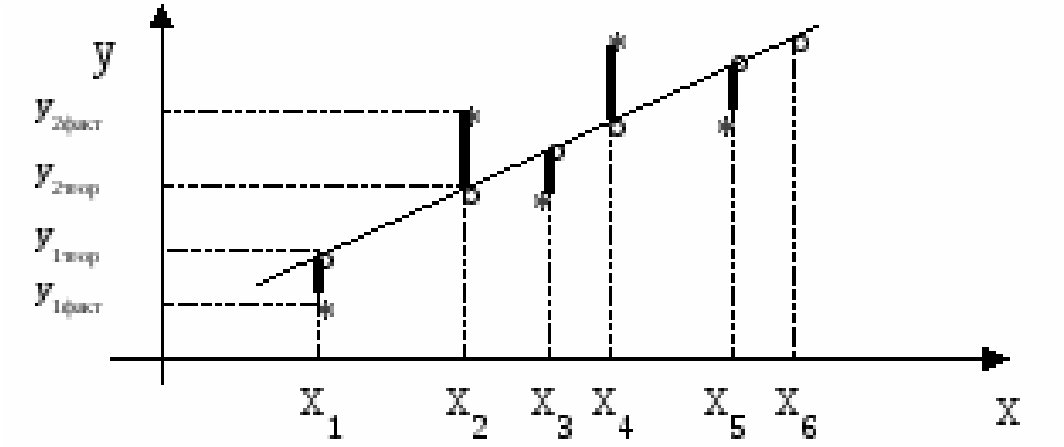


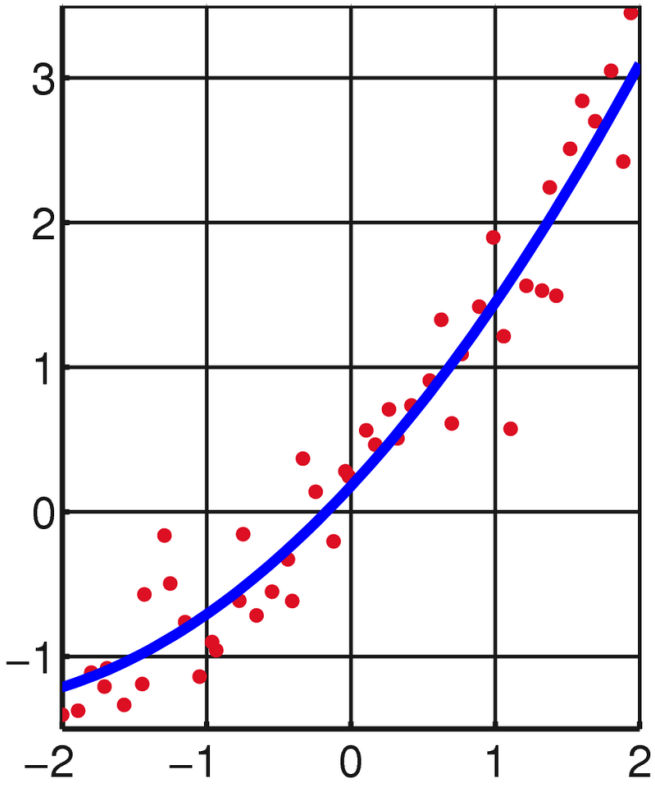
$$T_i(a,b,c) = \mathrm{a} \cdot \mathrm{n}_\mathrm{i}^\mathrm{b} + c$$

Переопределённая система уравнений

Регрессионный анализ

Метод наименьших квадратов (Гаусс, 1795; Лежандр, 1805)



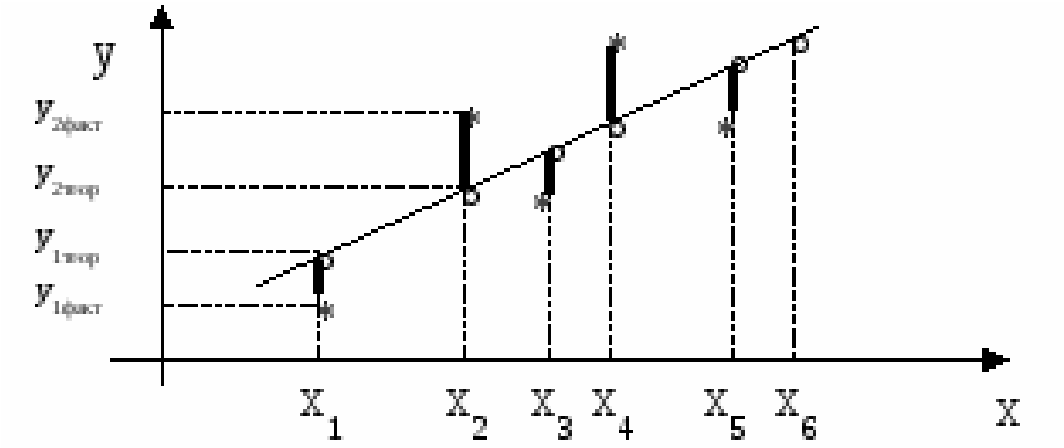


$$x = \{x_i\},\ i = 1, 2, 3...$$

$$\sum_i (y_i - f(x_i))^2 \to \min_x$$

Регрессионный анализ

Метод наименьших квадратов (Гаусс, 1795; Лежандр, 1805)



$$x = \{x_i\},\ i = 1, 2, 3\ldots$$

$$\sum_i (y_i - f(x_i))^2 \to \min_x$$

Простейший случай – линейная регрессия

Функция должна быть линейной по коэффициентам:

$$y_i = a + bx_i + cz_i$$

$$y_i = a + bx_i + cz_i + dx_i z_i$$

$$y_i = a + bx_i + cz_i + bcx_i z_i$$

Предсказание коэффициента гидрофобности

$$\log P_{ow} = \sum_i n_i \alpha_i$$

n_i – *число атомов типа i*

α_i – *вклад атома типа i*

Table I. Classification of atoms, and their contributions to octanol-water partition coefficient which is a measure of hydrophobicity.

Type	Description[a]	Hydrophobic[b] Contribution	No. of Compounds	Frequency of Use	T-test	Molar Refraction[c]
	C in:					
1	:CH_3R, CH_4	−0.6037	360	548	100.00	2.3000
2	:CH_2R_2	−0.4295	216	454	100.00	2.3071
3	:CHR_3	−0.3426	45	50	100.00	2.4926
4	:CR_4	−0.1155	24	24	74.32	2.3000
5	:CH_3X	−1.0578	157	224	100.00	3.4006
6	:CH_2RX	−0.8188	257	402	100.00	3.2624
7	:CH_2X_2	−0.1540	5	5	51.00	3.6770
8	:CHR_2X	−0.5995	73	118	100.00	3.0137
9	:$CHRX_2$	0.0095	27	27	7.85	3.225
10	:CHX_3	0.5134	4	4	96.02	3.2401
11	:CR_3X	−0.4807	14	14	99.97	2.6140
12	:CR_2X_2	0.2853	2	2	58.14	3.1488
13	:CRX_3	0.5335	34	36	100.00	2.3010
14	:CX_4	1.1114	6	6	100.00	3.3559
15	:—CH	−0.1654	25	31	97.01	3.5071

http://www.vcclab.org/lab/alogps/

9000+ соединений, 115 типов атомов
(Ghose et al., 1986, 1998)

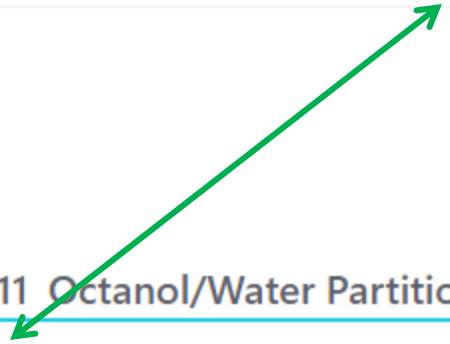
Предсказание коэффициента гидрофобности

PubChem Pentane (Compound)

3 Chemical and Physical Properties

3.1 Computed Properties

Property Name	Property Value	Reference
Molecular Weight	72.15 g/mol	Computed by PubChem 2.1 (PubChem release 2019.06.18)
XLogP3	3.4	Computed by XLogP3 3.0 (PubChem release 2019.06.18)

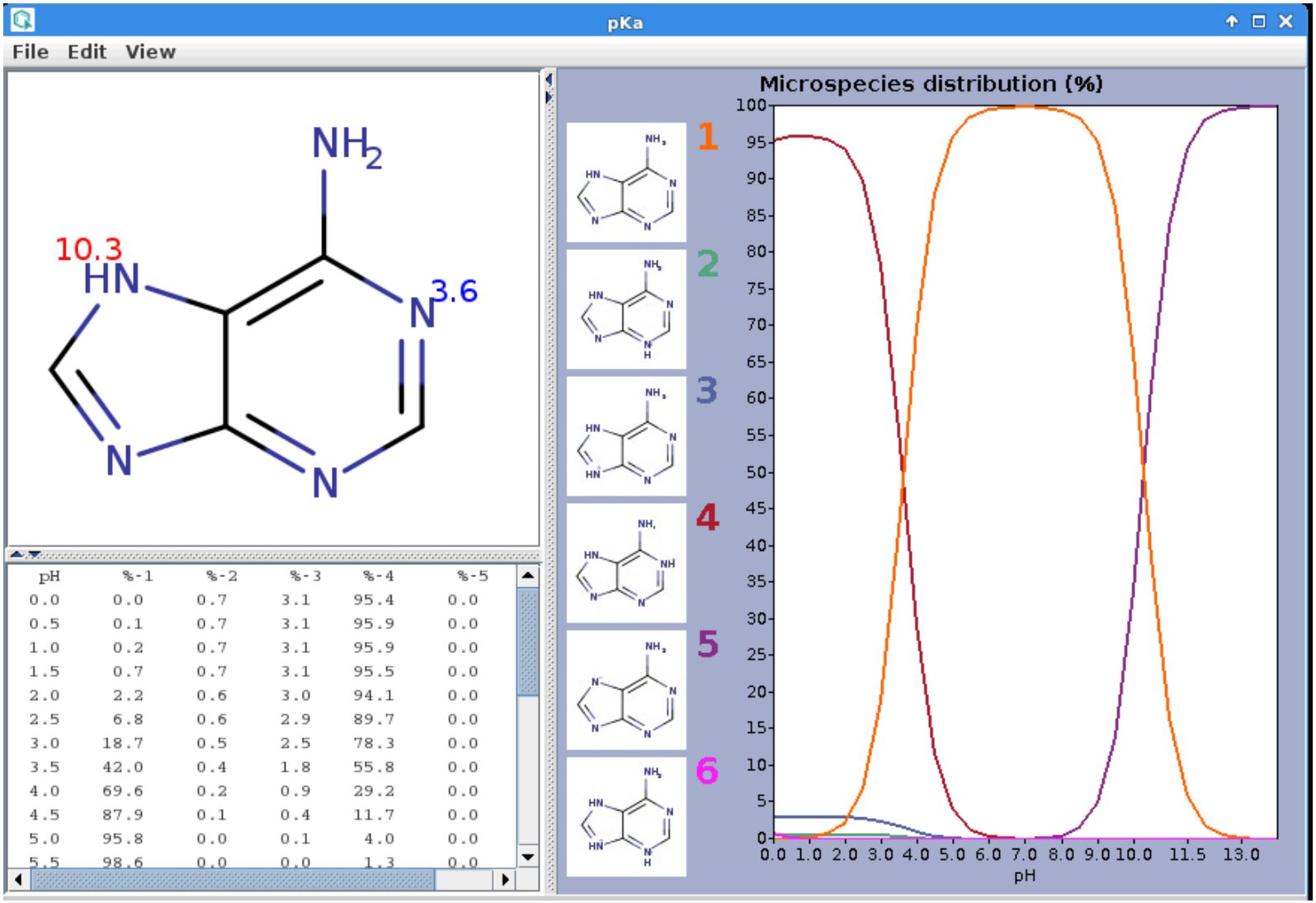


3.2.11 Octanol/Water Partition Coefficient

3.39 (LogP)

HANSCH,C ET AL. (1995)

Предсказание pKa



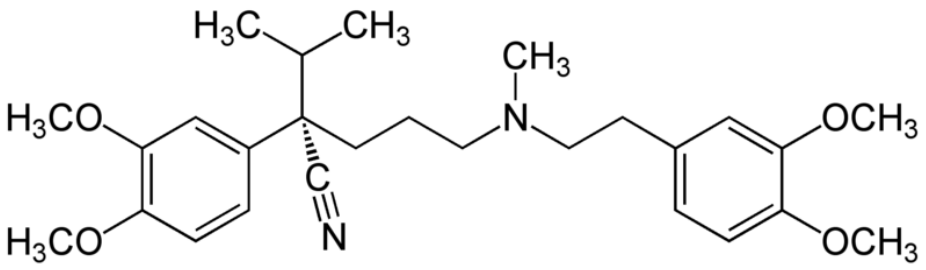
pH	%-1	%-2	%-3	%-4	%-5
0.0	0.0	0.7	3.1	95.4	0.0
0.5	0.1	0.7	3.1	95.9	0.0
1.0	0.2	0.7	3.1	95.9	0.0
1.5	0.7	0.7	3.1	95.5	0.0
2.0	2.2	0.6	3.0	94.1	0.0
2.5	6.8	0.6	2.9	89.7	0.0
3.0	18.7	0.5	2.5	78.3	0.0
3.5	42.0	0.4	1.8	55.8	0.0
4.0	69.6	0.2	0.9	29.2	0.0
4.5	87.9	0.1	0.4	11.7	0.0
5.0	95.8	0.0	0.1	4.0	0.0
5.5	98.6	0.0	0.0	1.3	0.0

Exp. Data (1959)**: Acidity (pKa)** 4.15 (secondary), 9.80 (primary)

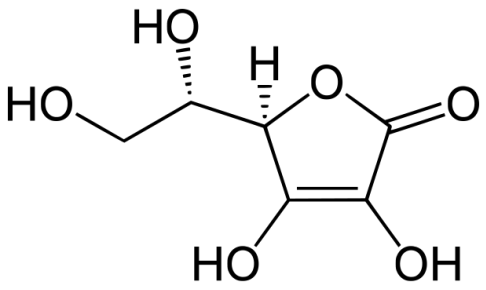
Предсказание биологической активности

«Правило пяти» (Lipinski, 1997) (Rule of thumb):

- Не более 5 доноров водородных связей
- Не более 10 акцепторов водородных связей
- Относительная молярная масса не более 500
- LogP не более 5



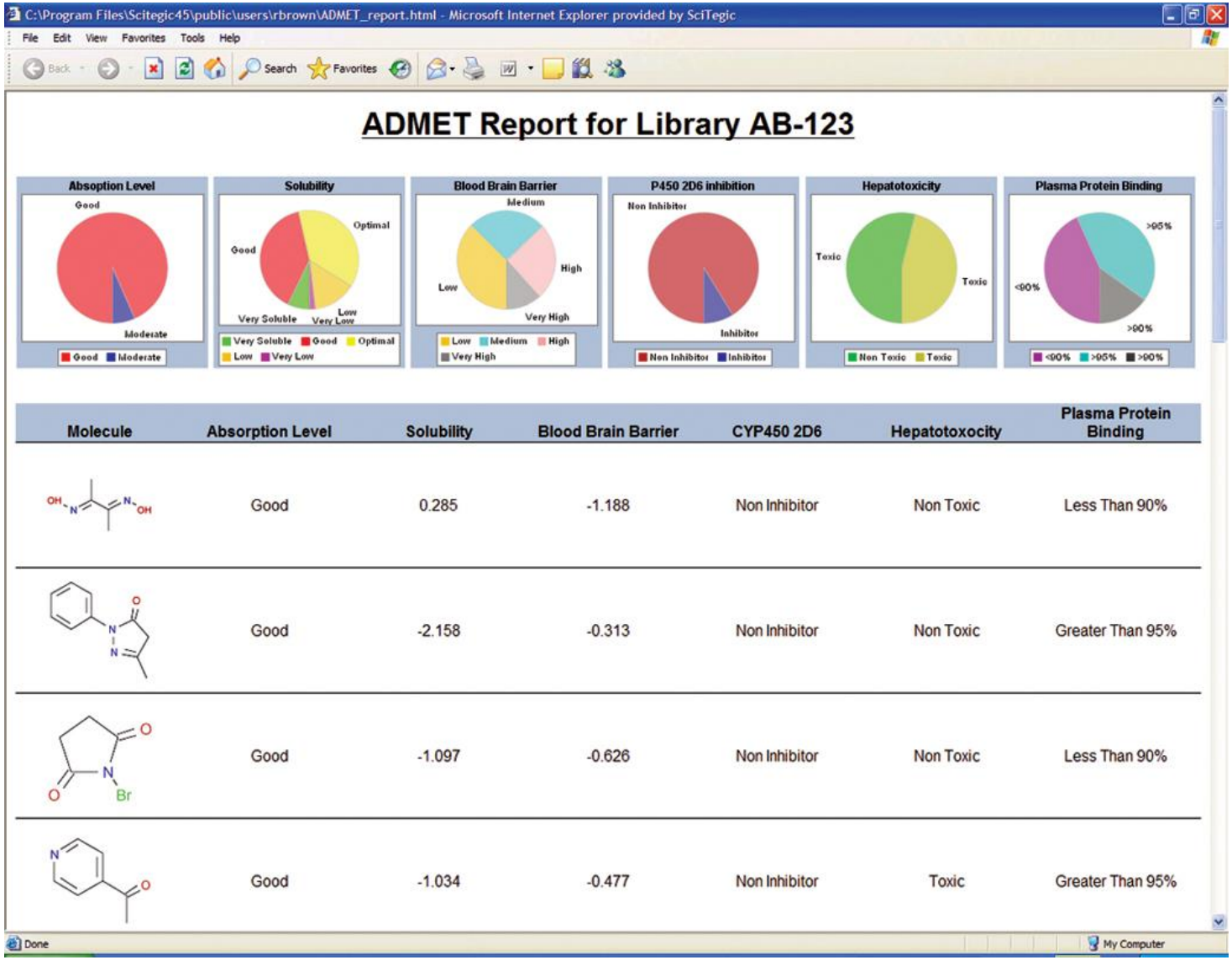
Верапамил
Mw=454 LogP=3,79



Аскорбиновая кислота
Mw=176 LogP=-1,9

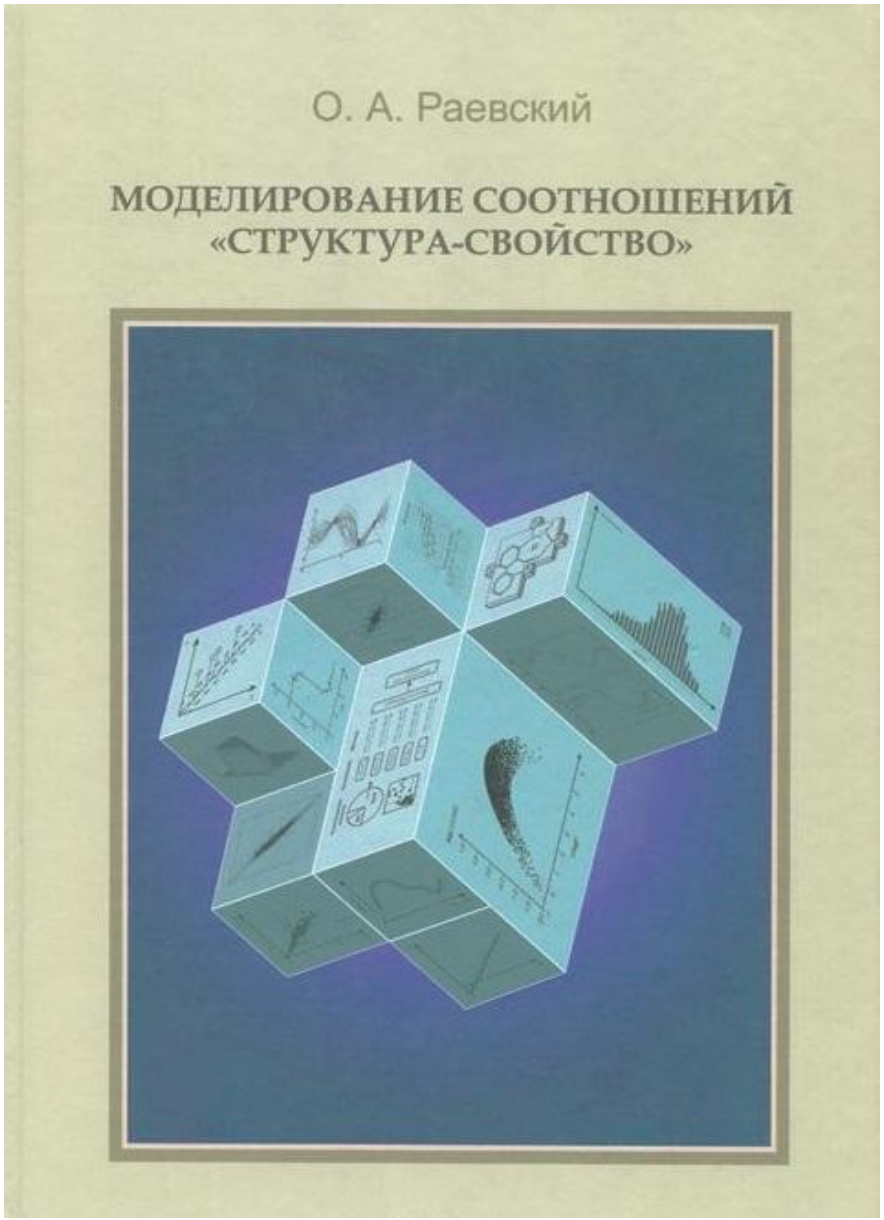
Предсказание биологической активности

ADMET – Absorption, Distribution, Metabolism, Excretion, and Toxicity



Предсказание свойств соединений

Успех QSAR-модели зависит от точности исходных данных, выбора подходящих дескрипторов и статистических методов и полноценной проверки модели.

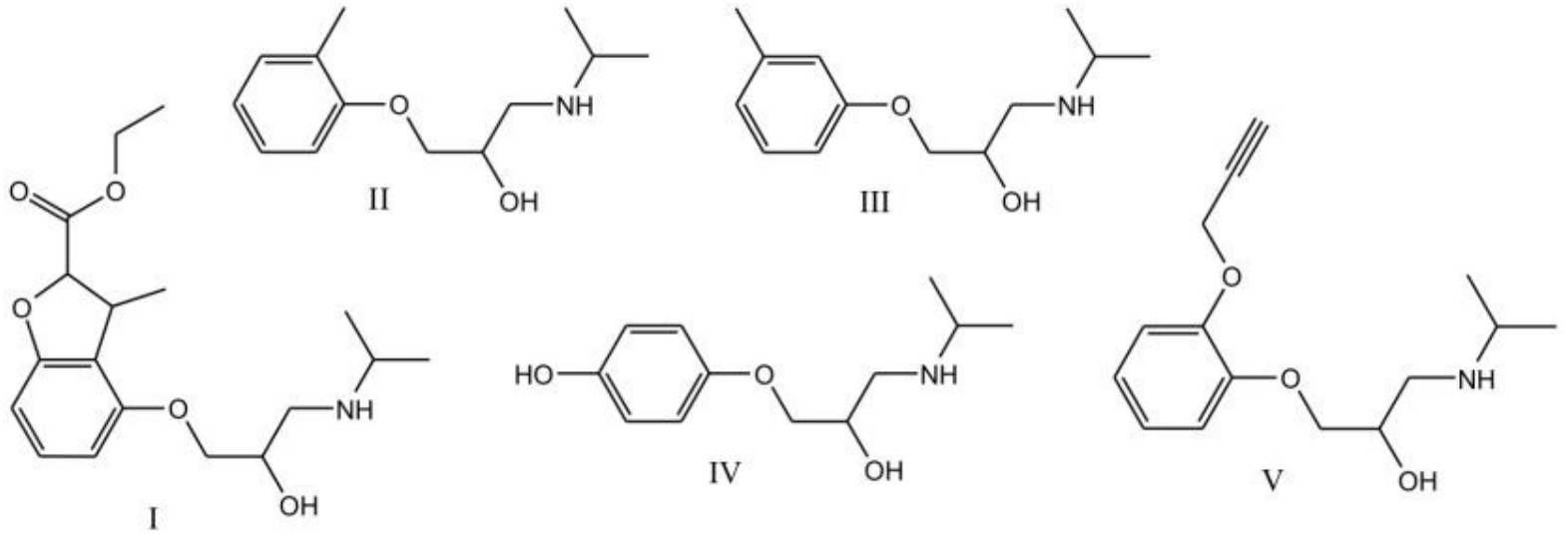


Предсказание свойств соединений

В самой общей форме:

значение свойства – это некая функция от некого набора дескрипторов.

Схожие молекулы должны обладать схожими свойствами



Насколько похожи эти молекулы?

О мерах сходства

Коэффициент Танимото (для битовых строк Xi и Yi) (1960):

$$S_T = \frac{\sum_i (X_i \wedge Y_i)}{\sum_i (X_i \vee Y_i)}$$

Λ – логическое И
V – логическое ИЛИ

A	1	0	1	1	0	1	$\lvert A\rvert = 4$
B	1	1	0	1	0	0	$\lvert B\rvert = 3$
$A \wedge B$	1	0	0	1	0	0	$\lvert A \wedge B\rvert = 2$
$A \vee B$	1	1	1	1	0	1	$\lvert A \vee B\rvert = 5$

$$S_T(A, B) = \tfrac{2}{5}$$

О мерах сходства

Коэффициент Жаккара («коэффициент флористической общности») (P. Jaccard, 1901)

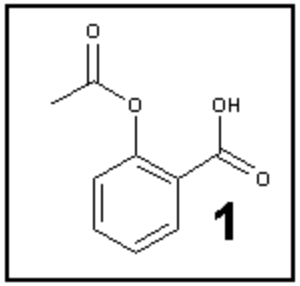
$$K_J = \frac{c}{a+b-c}$$

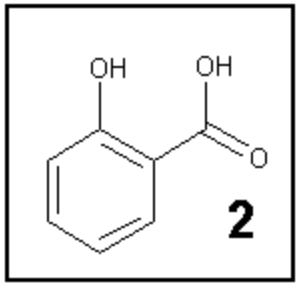
Первый предложенный коэффициент сходства

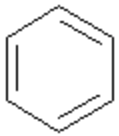
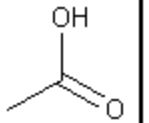
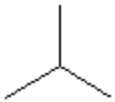
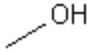
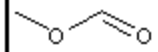
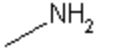
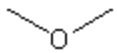
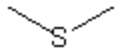
a — количество видов на первой пробной площадке,
b — количество видов на второй пробной площадке,
c — количество видов, общих для 1-ой и 2-ой площадок.

Молекулярное сходство

Similarity Searching





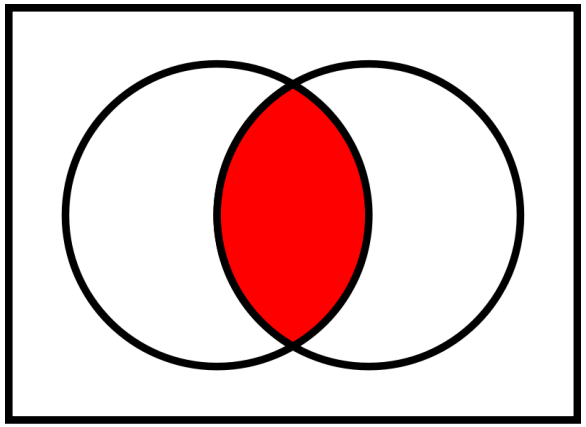
1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0
								

A = Number of bits set in both = 3
B = Number of bits set in (1), but not in (2) = 2
C = Number of bits set in (2), but not in (1) = 0

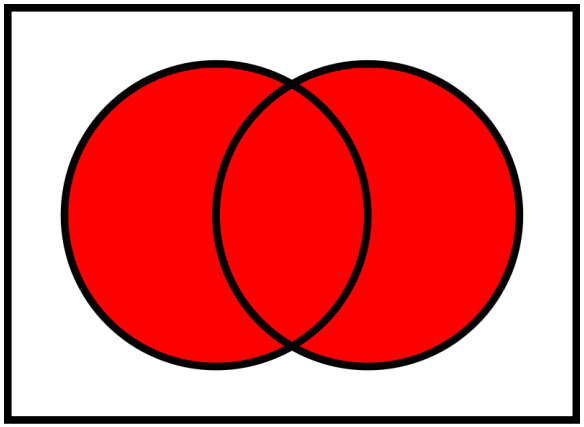
TANIMOTO COEFFICIENT = A / (A + B + C)
= 3 / (3 + 2 + 0) = 0.6 or 60%

О мерах сходства

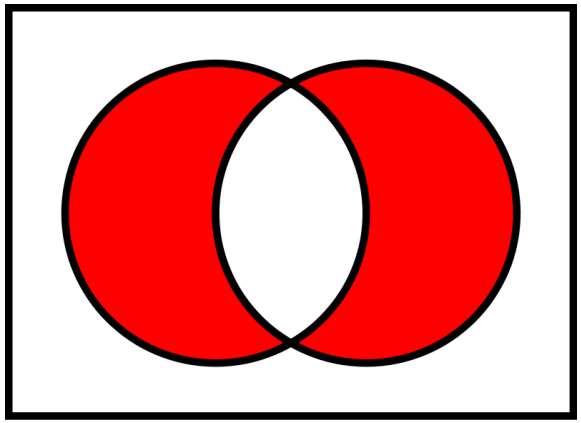
Диаграммы Венна («Eulerian Circles»,1880):



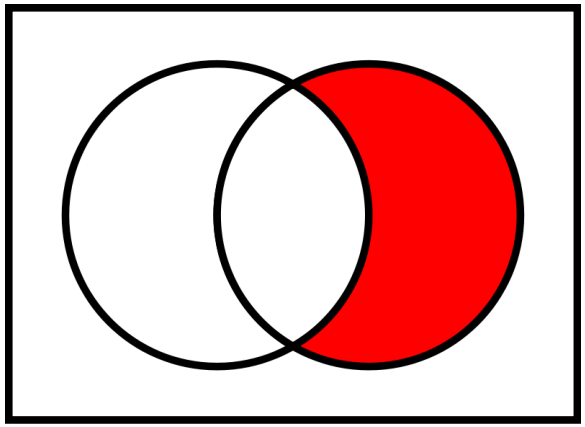
Пересечение



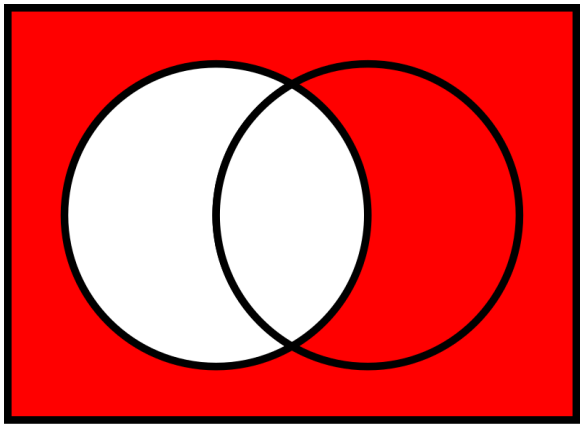
Объединение



Симметричная разность



Относительное дополнение



Абсолютное дополнение

О мерах сходства

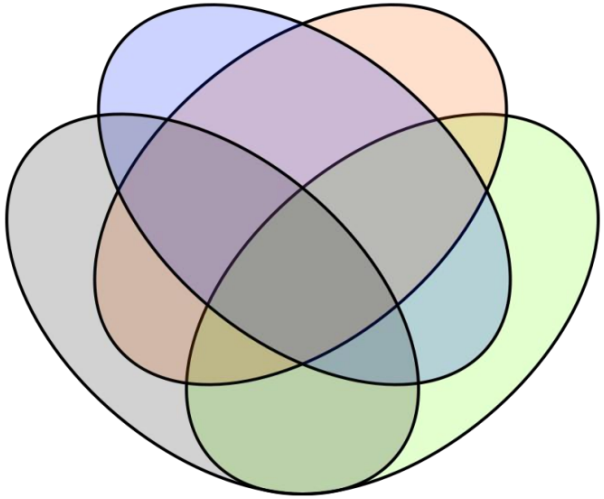


Диаграмма Венна для 4х множеств: показаны все возможные пересечения множеств

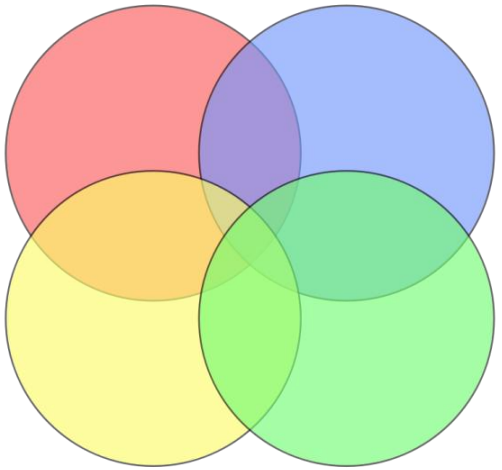
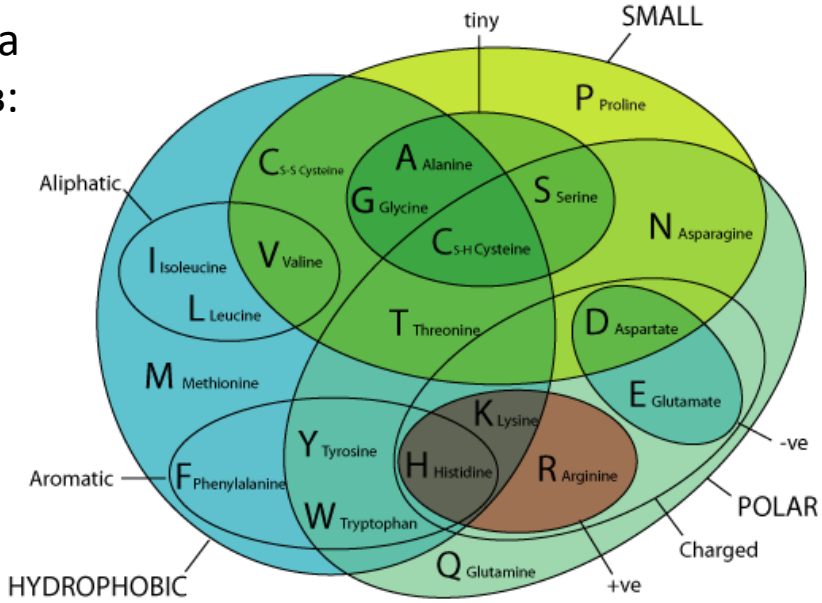
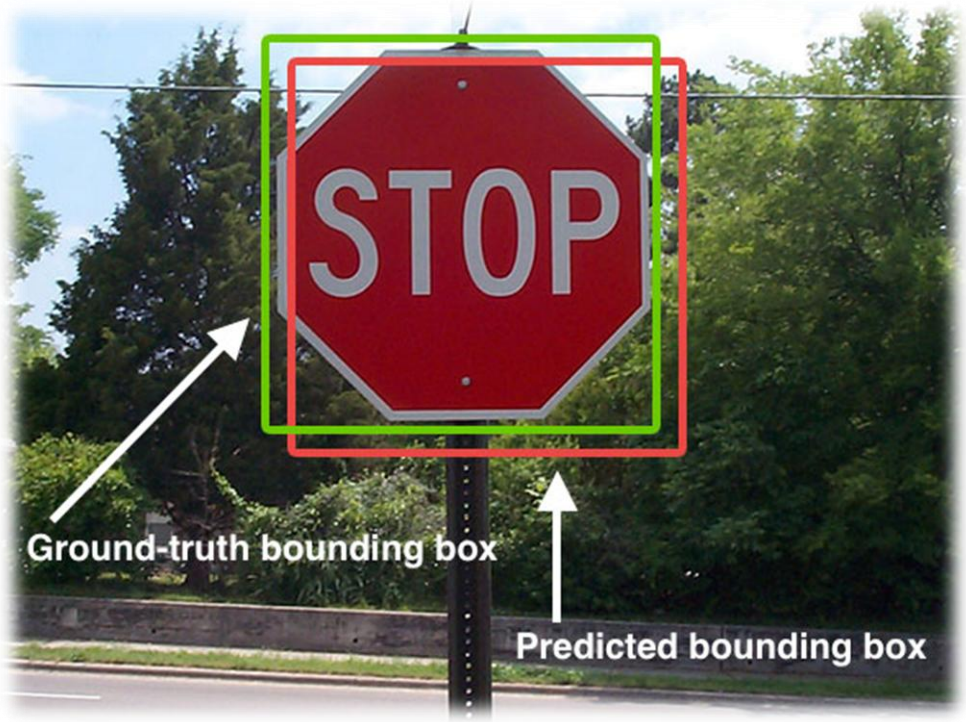


Диаграмма Эйлера для 4х множеств: показана лишь часть парных пересечений



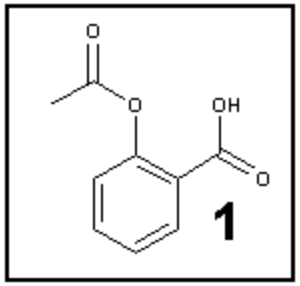
О мерах сходства

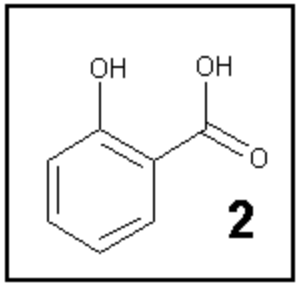


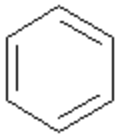
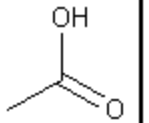
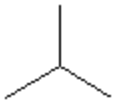
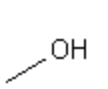
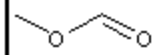
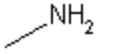
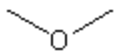
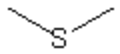
$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Молекулярное сходство

Similarity Searching





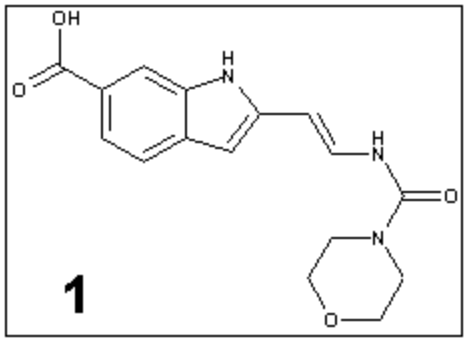
1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0
								

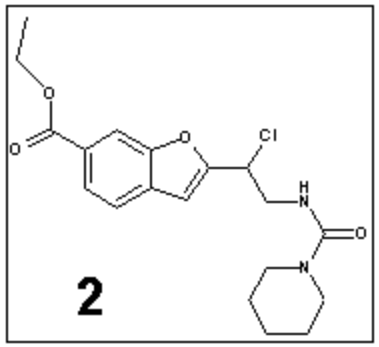
A = Number of bits set in both = 3
B = Number of bits set in (1), but not in (2) = 2
C = Number of bits set in (2), but not in (1) = 0

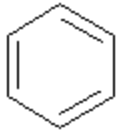
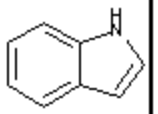
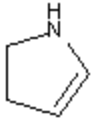
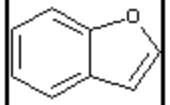
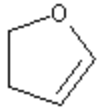
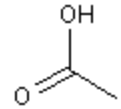
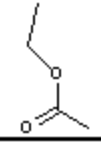
TANIMOTO COEFFICIENT = A / (A + B + C)
= 3 / (3 + 2 + 0) = 0.6 or 60%

Молекулярное сходство

Similarity Searching: Problem 1





1								
2								
								
1								
2