А. А. Рагойша

Информационные технологии в химии

2-й семестр

Избранные элементы хемоинформатики

Лекция 3

Хранение информации о трехмерной кристаллической структуре

Файлы в формате CIF
Crystallographic Information File

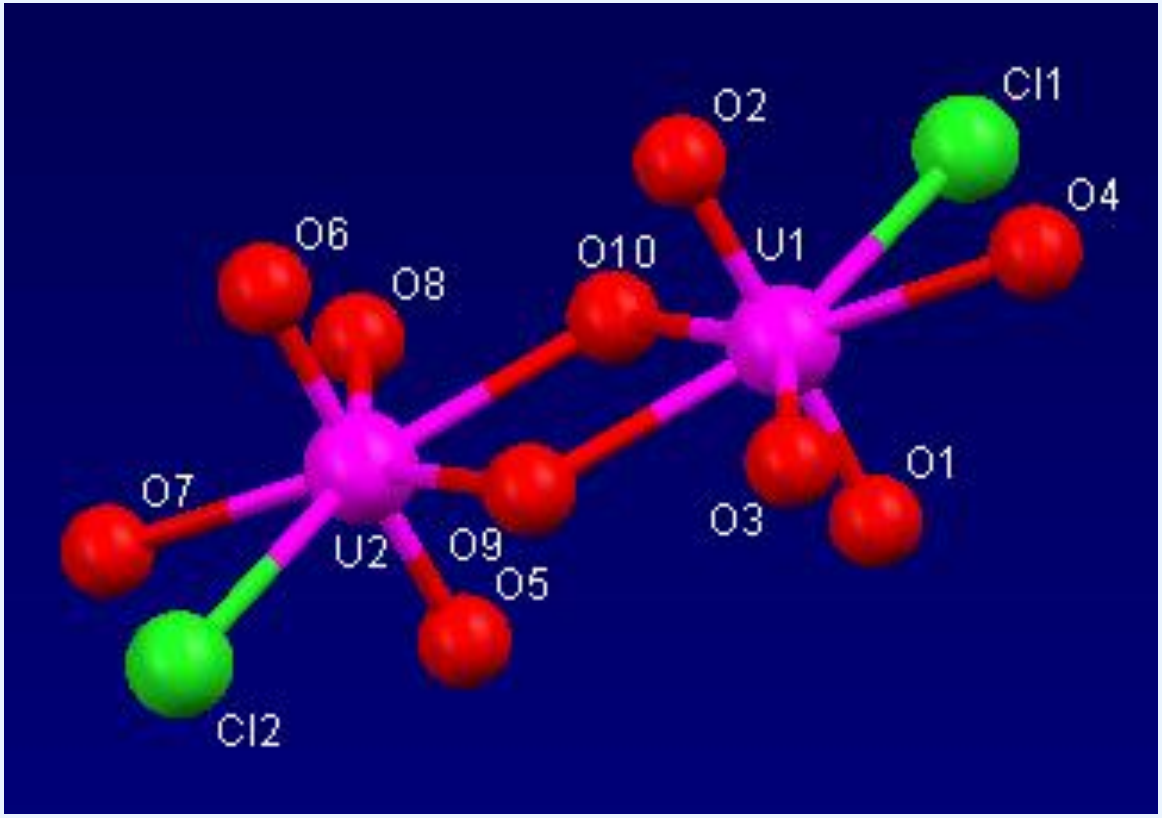
CIF – стандартный формат обмена кристаллографической информацией, разработанный Международным союзом кристаллографии.

Файлы в формате CIF могут содержать в себе:

- информацию о пространственном расположении атомов (т.е. координаты атомов и в явной форме межатомные расстояния, значения валентных углов) – причем это данные **экспериментальные**, а не рассчитанные из моделей;
- кристаллографические параметры;
- рентгенограммы;
- текстовый материал.

Фрагмент структуры, CIF-формат

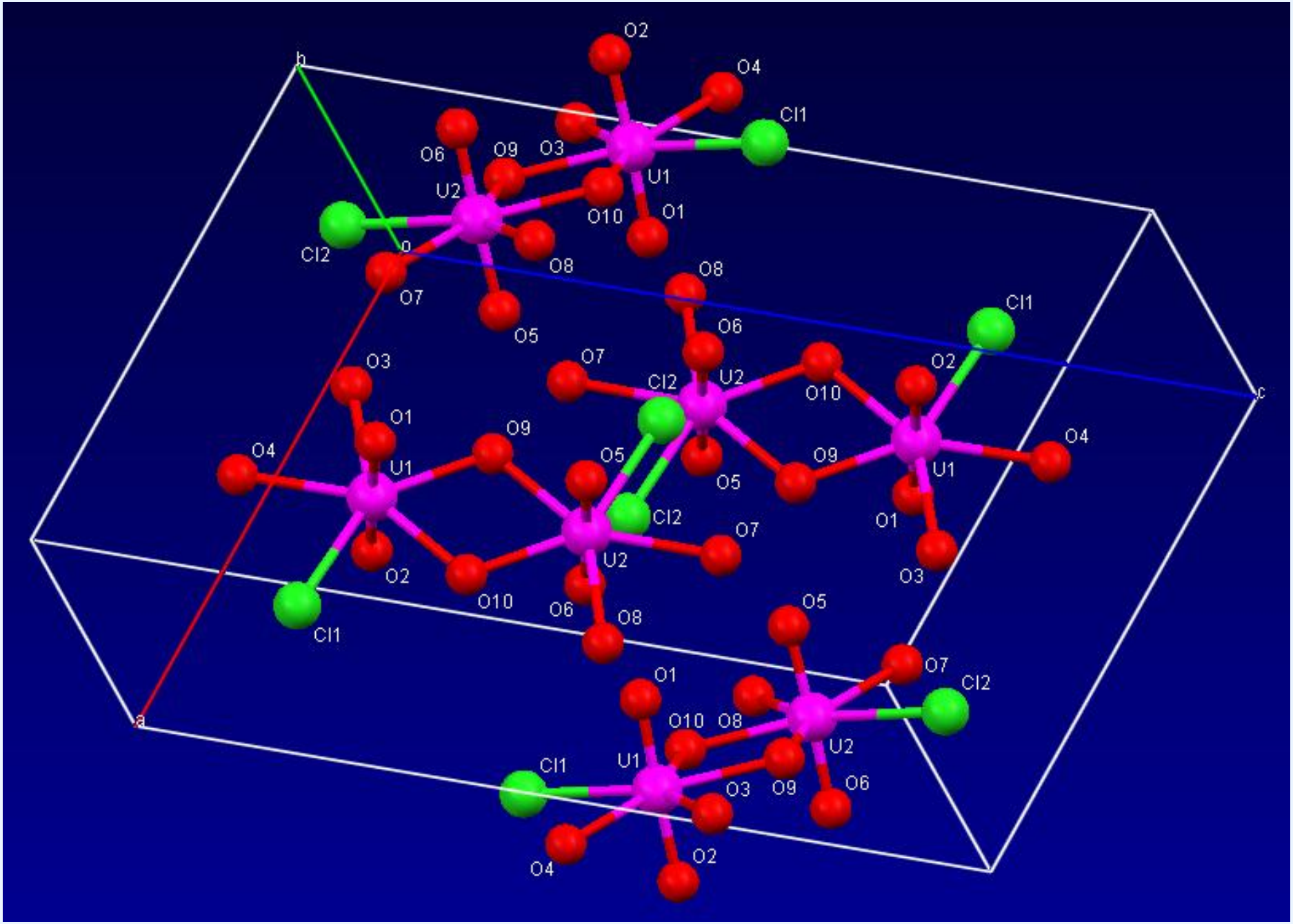
Визуализация фрагмента



Параметры структуры

	Atom1	Atom2	Type	Length
1	U1	Cl1	Unknown	2.751(3)
2	U1	O1	Unknown	1.746(9)
3	U1	O2	Unknown	1.790(9)
4	U1	O3	Unknown	2.396(9)
5	U1	O4	Unknown	2.49(1)
6	U1	O9	Unknown	2.382(9)

	Atom1	Atom2	Atom3	Angle
1	Cl1	U1	O1	91.0(3)
2	Cl1	U1	O2	91.1(3)
3	Cl1	U1	O3	141.7(2)
4	Cl1	U1	O4	72.9(2)
5	Cl1	U1	O9	143.1(2)


O2
O4
Cl1
O6
O9
O3
U1
U2
O10
O1
Cl2
O8
O7
O5
O8
O6
Cl1
O3
O7
Cl2
U2
O10
O2
O1
O9
O4
O5
U1
O9
O4
U1
O5
Cl2
O7
O1
U2
O2
O10
O6
O8
O3
Cl1
O5
O7
O1
Cl2
O10
O8
U2
Cl1
U1
O3
O9
O6
O4
O2
a
b
c

Структура CIF-файла

Используются только ASCII символы.

Разграничение: форма – содержание.

Каждый элемент информации
в формате:
Имя элемента (тэг) – значение.

Файл состоит из блоков.

Табличные данные в блоках loop_
Структура блока:
упорядоченный список имен и
упорядоченный список значений.

```
_chemical_formula_sum   'H10 C12 O10 U2'
_chemical_formula_weight     717.04
_symmetry_cell_setting     monoclinic
_symmetry_space_group_name_H-M  'P 21/n'
_symmetry_space_group_name_Hall '-P 2yn'
loop_
    _symmetry_equiv_pos_as_xyz
    'x, y, z'
    '-x+1/2, y+1/2, -z+1/2'
    '-x, -y, -z'
    'x-1/2, -y-1/2, z-1/2'
_cell_length_a      10.712(2)
_cell_length_b      6.1212(12)
_cell_length_c      17.662(4)
_cell_angle_alpha     90.00
_cell_angle_beta     95.47(3)
_cell_angle_gamma     90.00
_cell_volume     1152.8(4)
_cell_formula_units_Z     4
```

Таблица в CIF-файле

```
245  loop_
246      _geom_angle_atom_site_label_1
247      _geom_angle_atom_site_label_2
248      _geom_angle_atom_site_label_3
249      _geom_angle_site_symmetry_1
250      _geom_angle_site_symmetry_3
251      _geom_angle
252      _geom_angle_publ_flag
253      O1 U1 O2  .  .  177.7(4)  ?
254      O1 U1 O10 .  .  91.5(4)   ?
255      O2 U1 O10 .  .  87.9(4)   ?
256      O1 U1 O9  .  .  88.8(4)   ?
257      O2 U1 O9  .  .  88.9(4)   ?
```

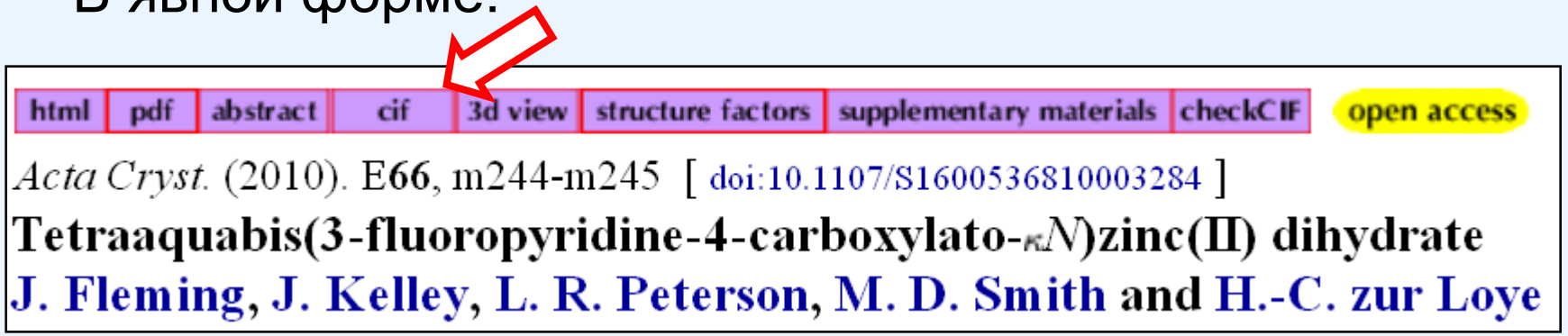
список имен

список значений

угол O2-U1-O9 равен $88,9^0$

Примеры CIF на сайтах журналов

- В явной форме:



- В неявной форме:

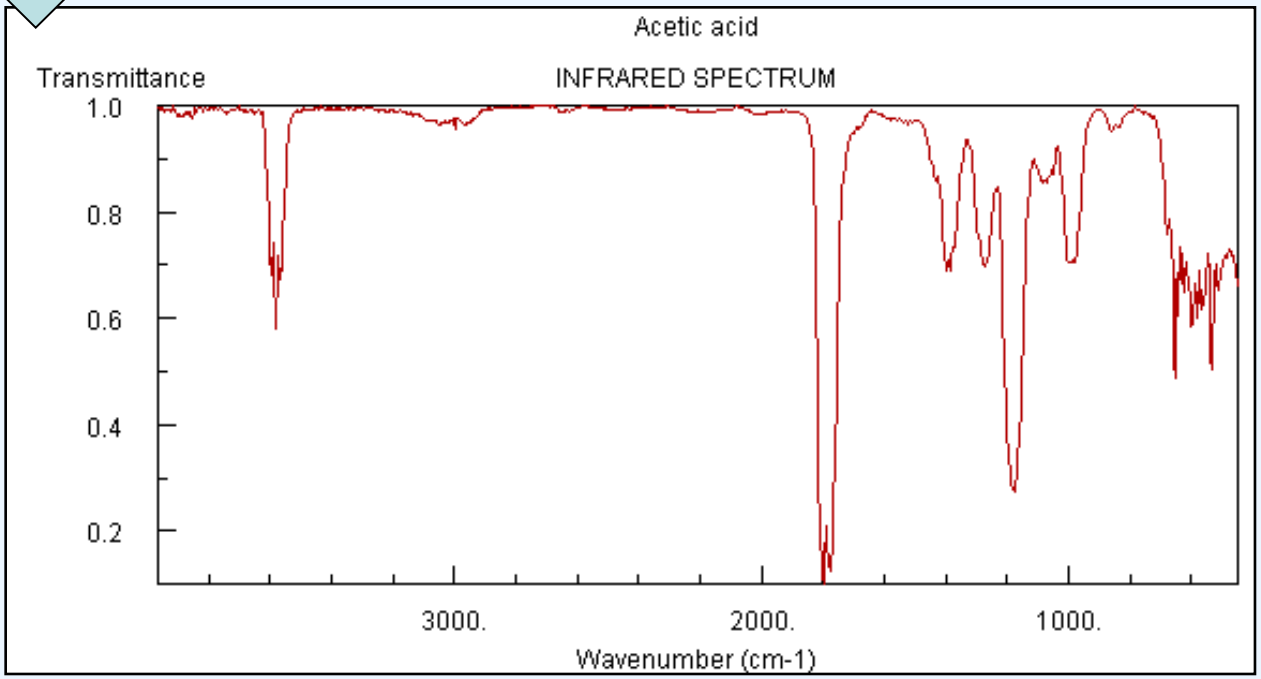
(Конечно же, CIF-файлы есть и в специализированных базах данных)

JCAMP-DX

```
##TITLE=Acetic acid
##JCAMP-DX=4.24
##DATA TYPE=INFRARED SPECTRUM
##OWNER=NIST Standard Reference
##CAS REGISTRY NO=64-19-7
##MOLFORM=C 2 H 4 O 2
##$NIST SOURCE=MSDC-IR
##STATE=gas
##XUNITS=1/CM
##YUNITS=ABSORBANCE
##XFACTOR=1.0
##YFACTOR=0.000139697
##DELTAX=4.0
##FIRSTX=450.0
##LASTX=3966.0
##FIRSTY=0.252992
##MAXX=3966
##MINX=450
##MAXY=1.39697
##MINY=0
##NPOINTS=880
##XYDATA=(X++(Y..Y))
450.0 1811 1597 1541 1481 1457
490.0 1449 1444 1542 1532 1646
530.0 1702 2994 2794 1627 1417
570.0 2084 1625 1849 2231 1931
610.0 1776 1582 1512 1679 1887
650.0 1788 3132 2876 1717 1253
```

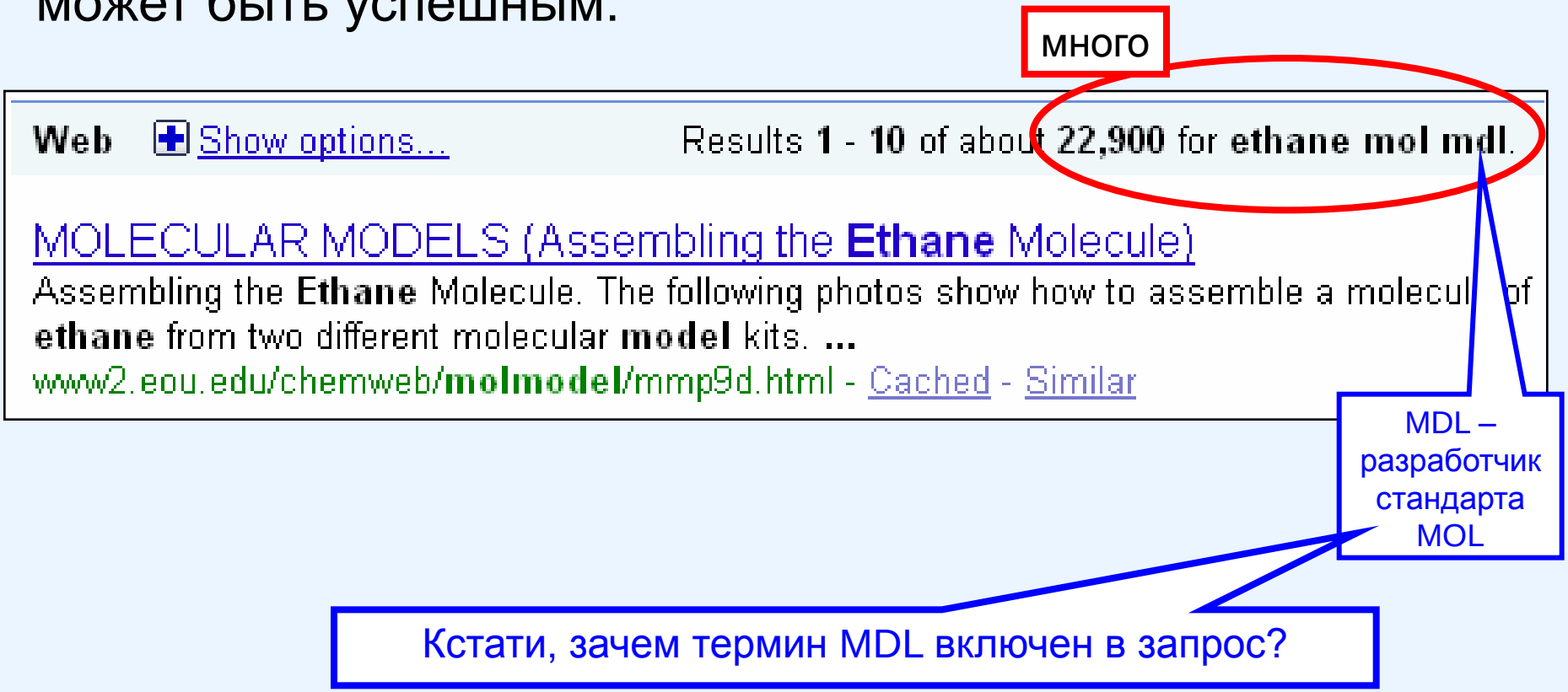
Стандарт ИЮПАК обмена спектральной информацией.

Файлы с расширениями **.dx**, **.jdx**.



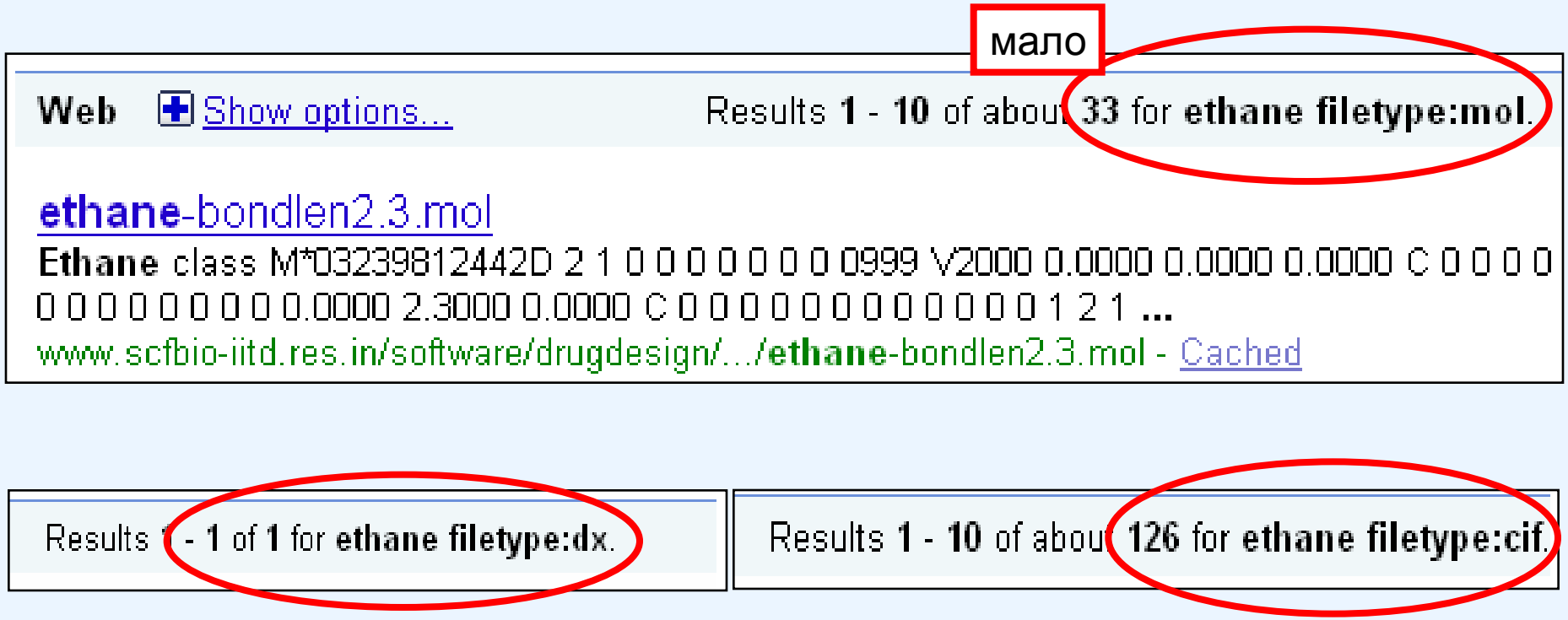
Google и химические файлы

Обнаружение химического файла
по *косвенным* признакам
может быть успешным:



Google и химические файлы

Поиск "в лоб" малоэффективен:
а) в базе данных мало таких файлов,
б) в названиях файлов – произвольные слова.



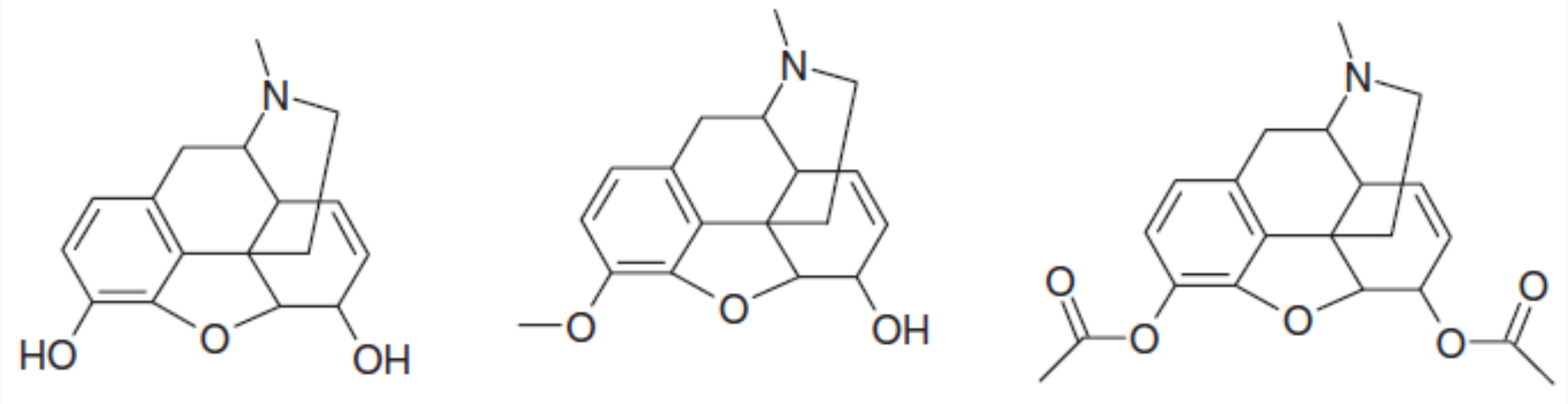
Краткий обзор некоторых иных понятий хемоинформатики

ОГРОМНЫЕ базы данных структурной информации о веществе - один из важнейших объектов хемоинформатики (особенно в области поиска биологически активных соединений)

1-й уровень рассуждений:
Одинаковые функциональные группы,
высокий уровень молекулярного подобия

близкая биологическая активность

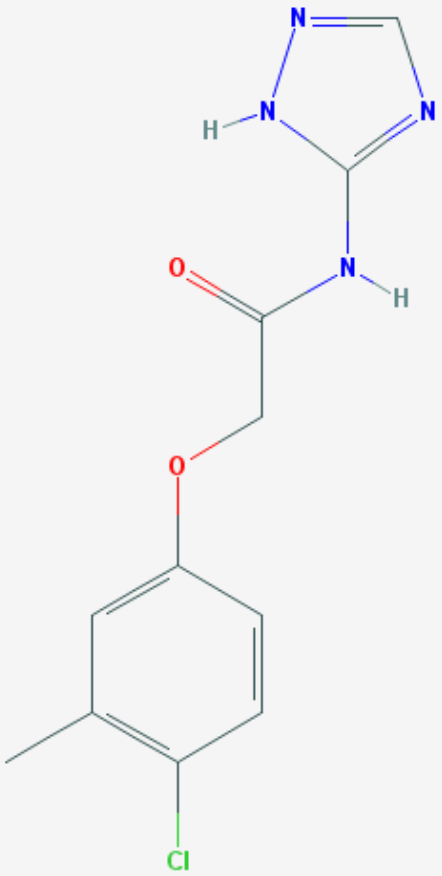
Пример: явно подобные структуры

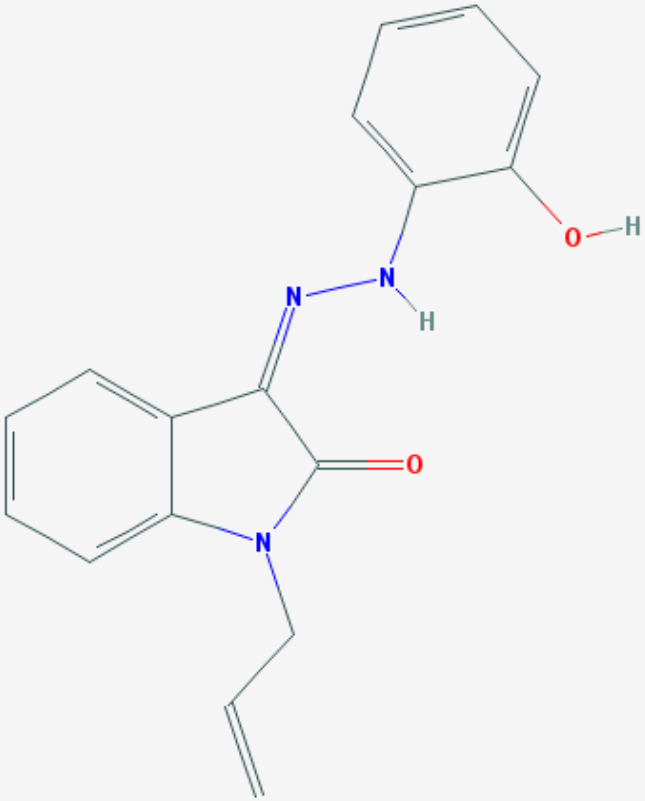


коэффициент Танимото ≈ 1

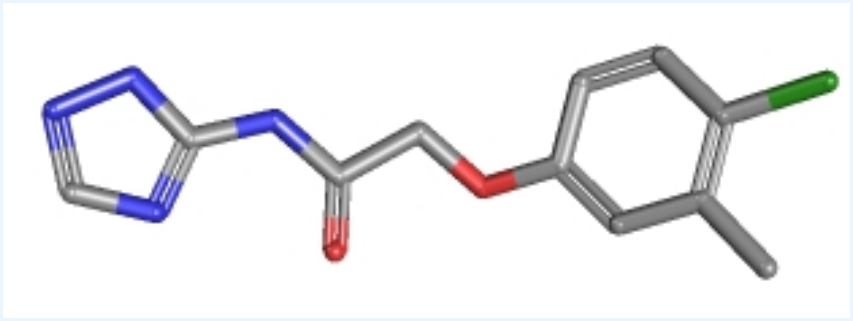
Все соединения действуют на опиоидные рецепторы

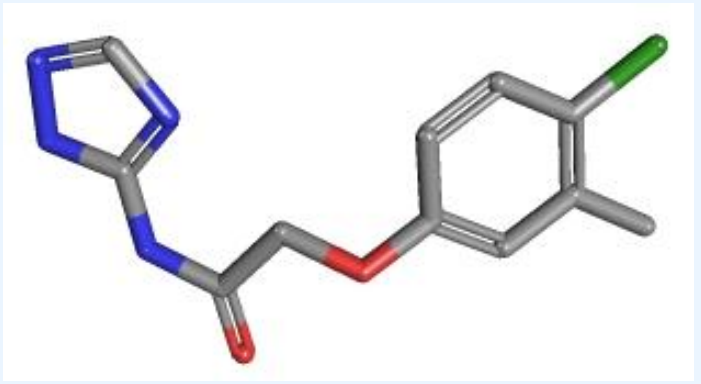
A.R.Leach, V.J.Gillet. An Introduction to Chemoinformatics. 2007.

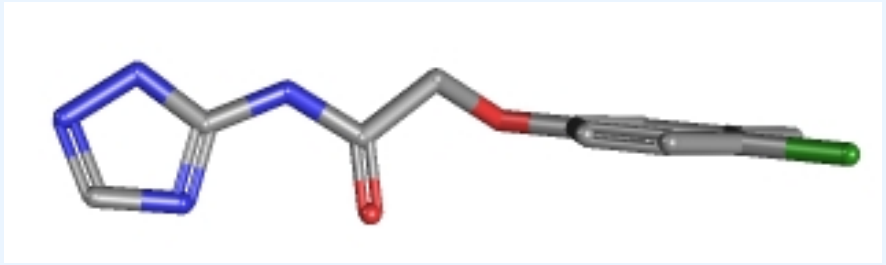


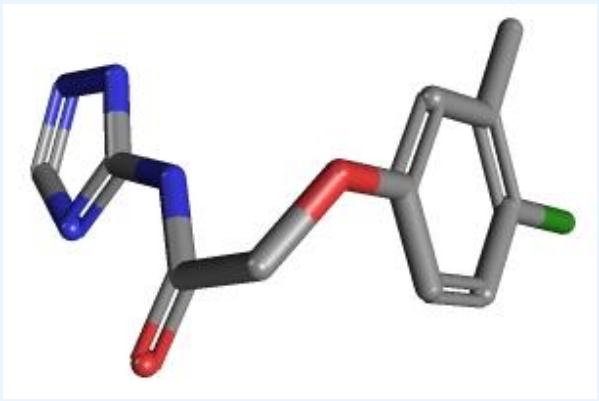


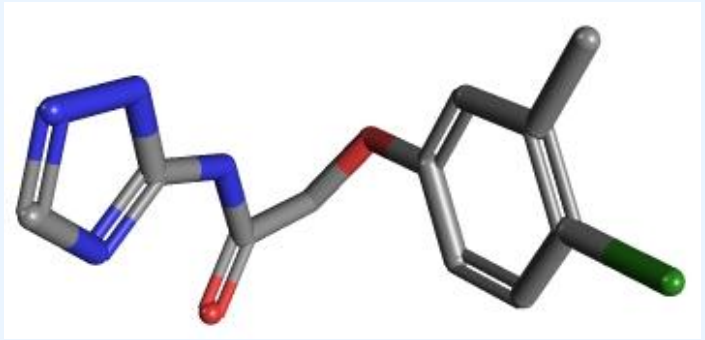
2D коэффициент Танимото = 0,43,
а
биологическое действие подобно

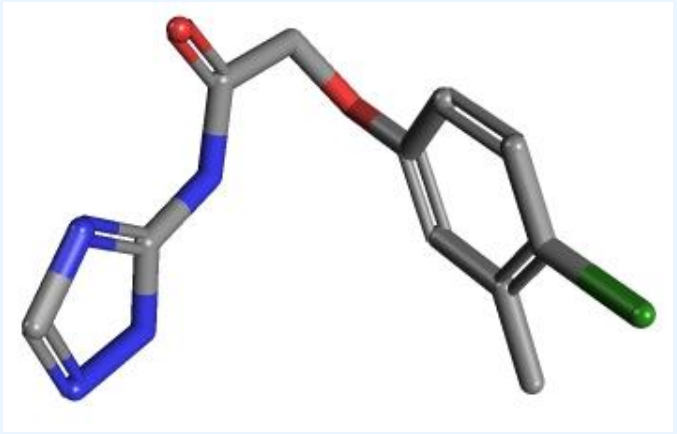




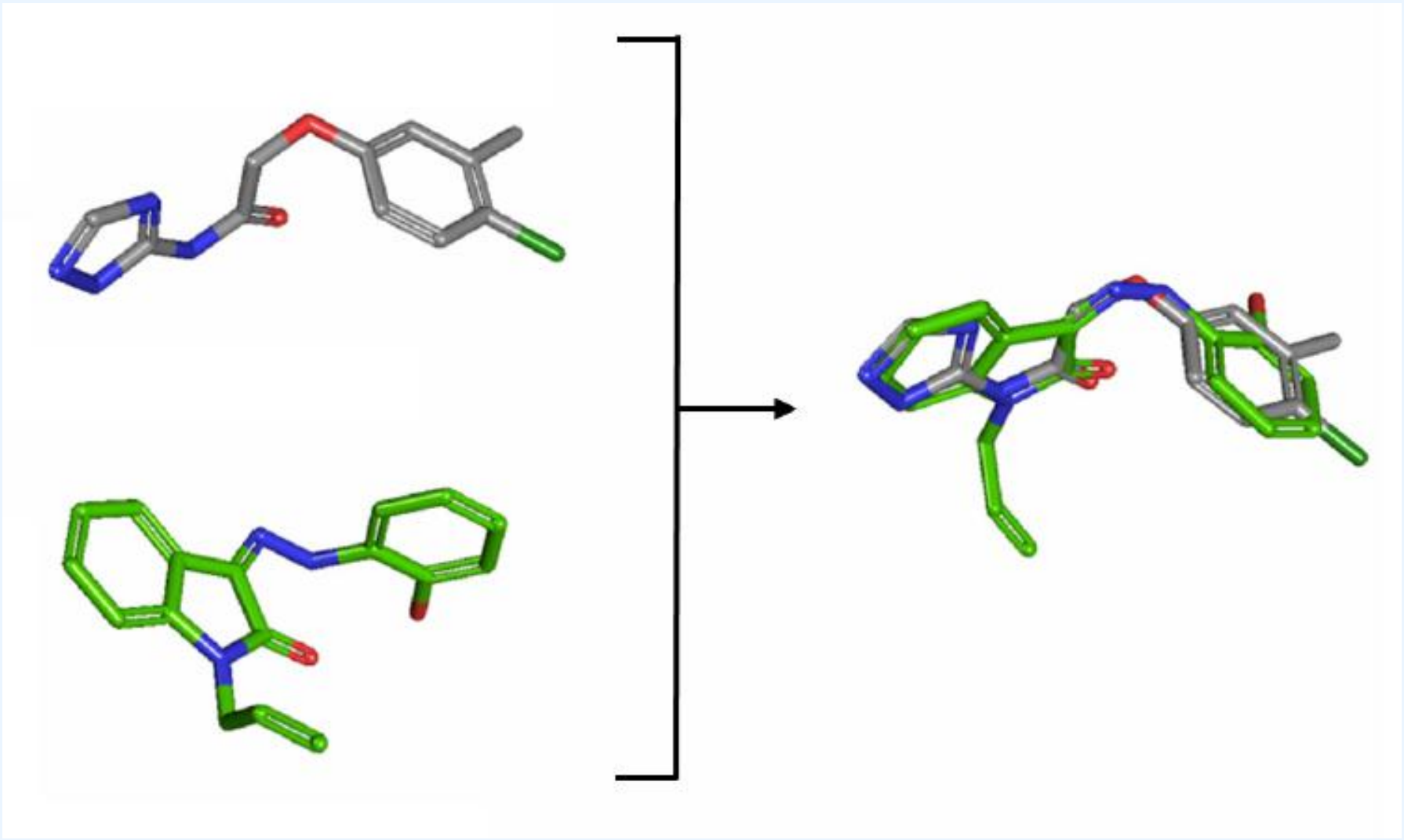








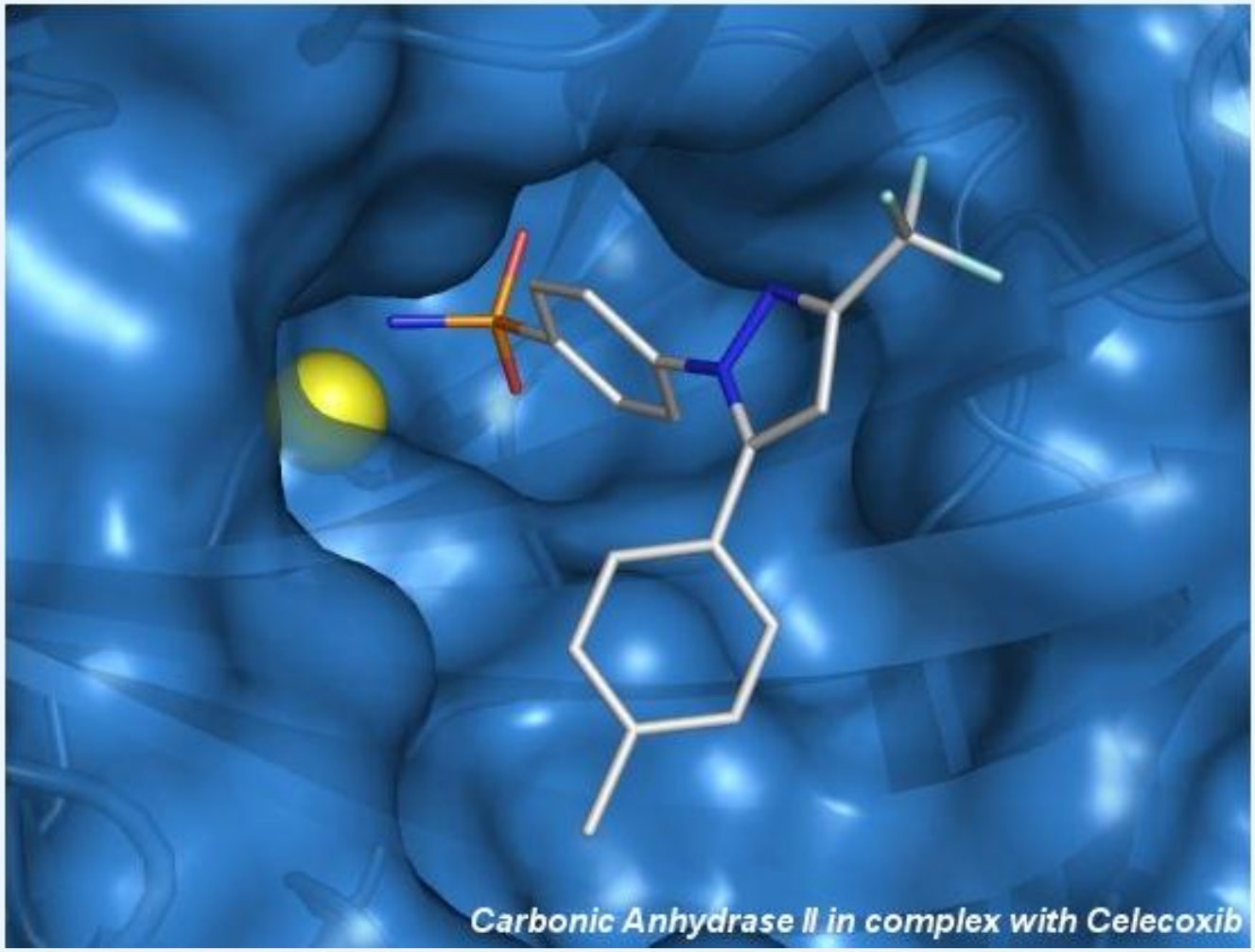
PubChem: 172 конформера



3D коэффициент Танимото (по форме молекулы) = 0,80

Комплекс "лиганд-белок"

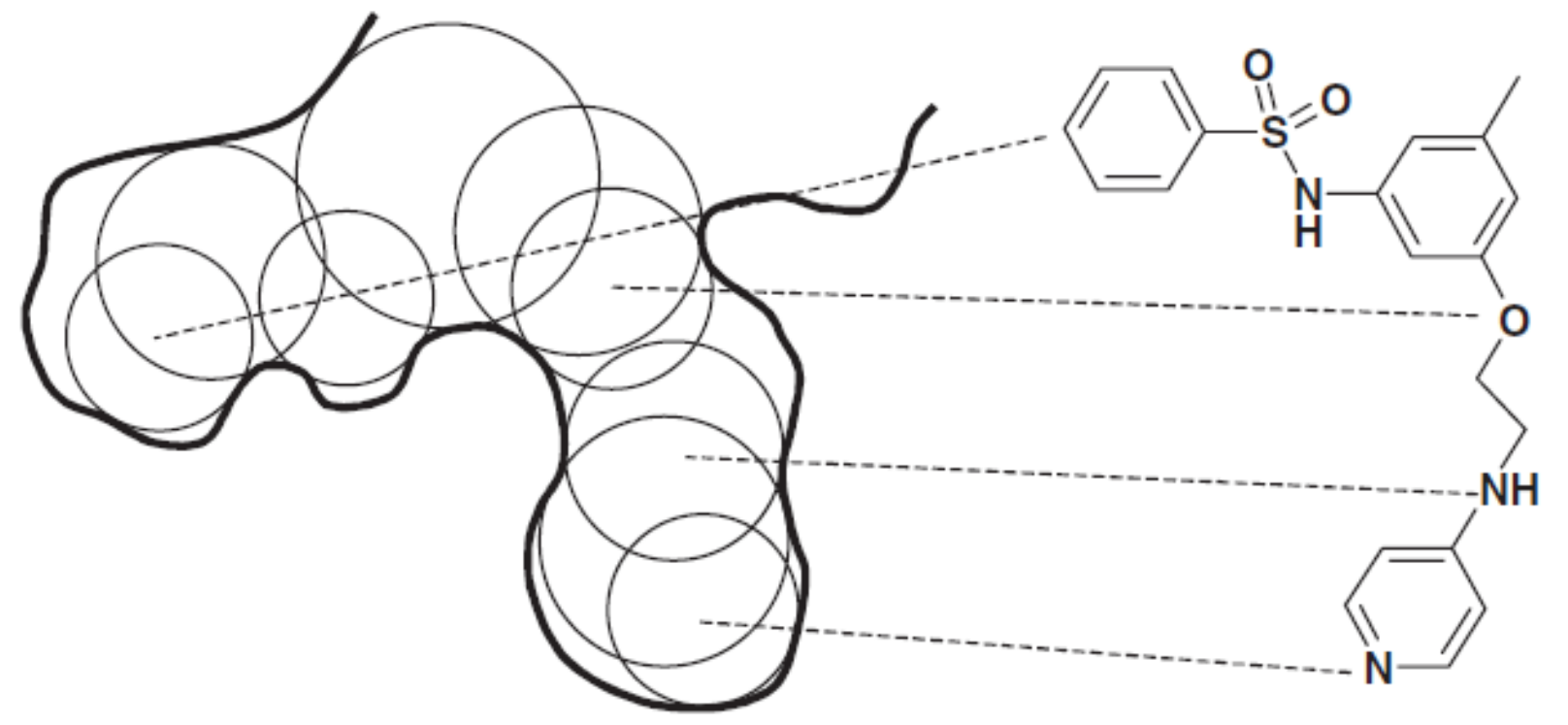
Рецепторы,
ферменты



AffinDB V1.0.2 - Affinity Database For Protein-Ligand Complexes
http://pc1664.pharmazie.uni-marburg.de/affinity/index.php.

Молекулярный докинг

Моделирование состава, конформации, взаимной ориентации, наиболее выгодных для образования устойчивого комплекса.



A.R.Leach, V.J.Gillet. An Introduction to Chemoinformatics. 2007.

Формальное описание молекулярной структуры с точки зрения заданной биологической активности

Фармакофор

Фармакофор
Pharmacophore

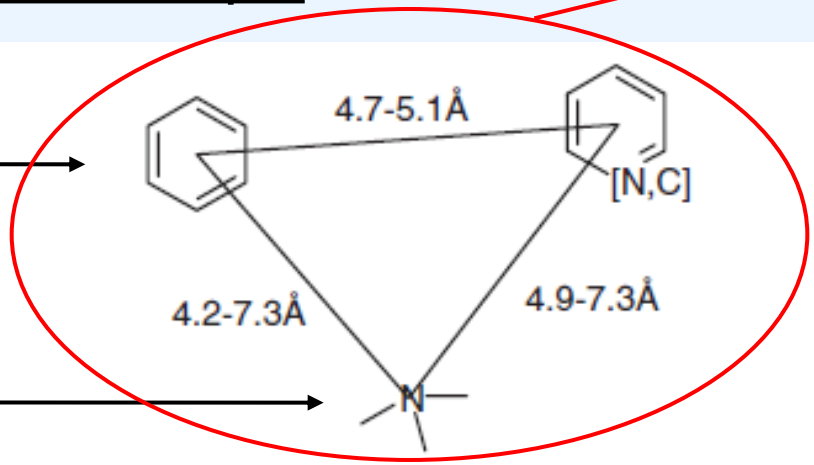
- Фармакофор — это набор пространственных и электронных признаков, необходимых для обеспечения оптимальных супрамолекулярных взаимодействий со специфической биологической мишенью, которые могут вызывать (или блокировать) ее биологический ответ.

Фармакофорные признаки: фармакофорные центры и интервалы расстояний между ними, необходимые для проявления данного типа биологической активности.

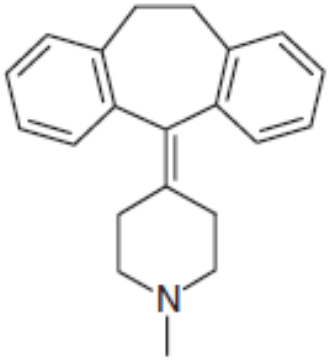
2D-фармакофор (пример)

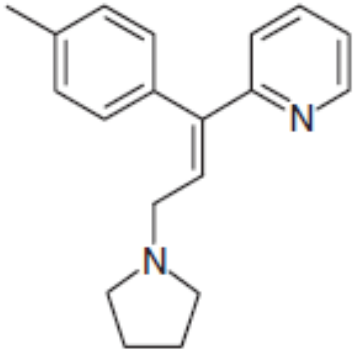
Фармакофорные центры

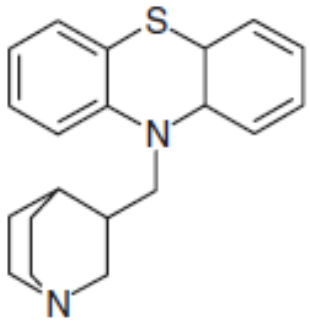
Фармакофор

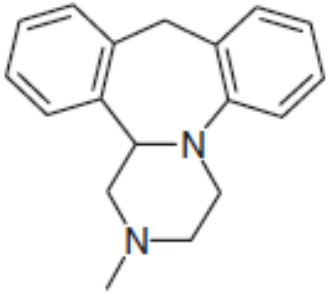


молекулы с таким фармакофором:





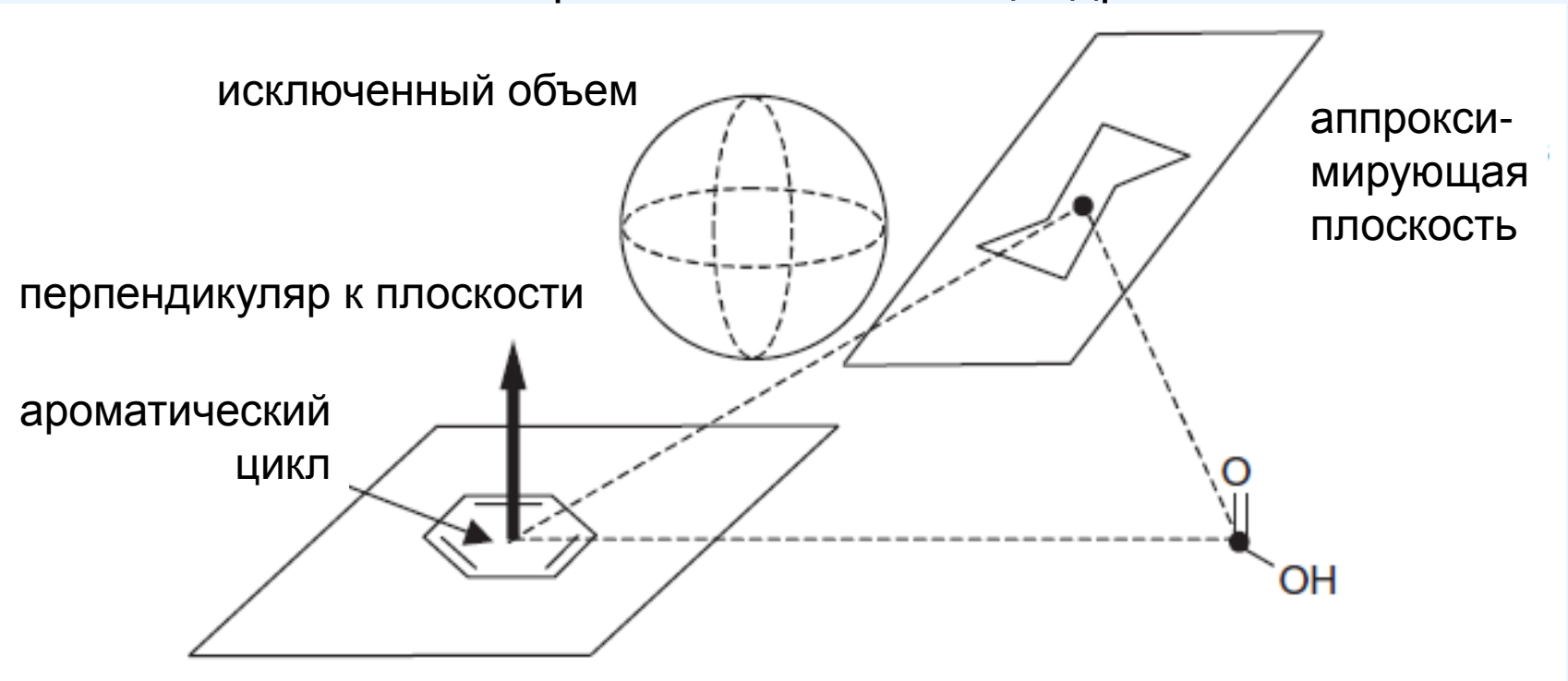




A.R.Leach, V.J.Gillet. An Introduction to Chemoinformatics. 2007.

3D-фармакофор (пример)

Фармакофорные центры: гидрофобные области, ароматические кольца, доноры и акцепторы водородной связи, анионные и катионные центры, гидрофобные и исключенные объемы, допустимые интервалы угловой ориентации векторов водородных связей и плоскостей ароматических колец и др.



A.R.Leach, V.J.Gillet. An Introduction to Chemoinformatics. 2007.

QSAR, QSPR

- QSAR [кью-сар]
  Quantitative Structure – Activity Relationship
  количественное соотношение структура – активность (биологическая)
- QSPR
  Quantitative Structure – Property Relationship
  количественное соотношение структура – свойство (физическое, физико-химическое)

Построение моделей, позволяющих по описанию структур химических соединений предсказывать свойства этих соединений (физические, химические, биологическую активность).

Молекулярные дескрипторы

Молекулярный дескриптор – численная величина, характеризующая молекулу (или химический объект).

Примеры:

молекулярная масса,
объем молекулы,
число ОН-групп.

Для чего нужен дескриптор?

постулируют:

Свойство вещества =

= ***f*** (дескриптор1, дескриптор2, дескриптор3, ...)

Фрагментные дескрипторы

Фрагментные дескрипторы:
оценивают вклад различных структурных частей молекулы в общее свойство:

- функциональная группа,
- донор, акцептор водородной связи,
- ароматический цикл,
- вращаемая связь и т. п.

Есть – нет,

Если есть – сколько,

В явной форме; в форме «отпечатка пальцев».

Пример: Предсказание способности вещества к биологическому разложению.

Table 2. Definitions of descriptors

Descriptor	
1) No. of Ar-X	2) No. of N
3) Molecular Weight	4) No. of Cl
5) No. of COOR	6) No. of $CH_2C(=X)X$

6 дескрипторов – успешность 77 %
50 дескрипторов – успешность 80 %
89 дескрипторов – успешность 82 %

7) No. of CH_2OH	
8) No. of 1,2,4-substituted benzene	
9) No. of Ar-X-Ar	10) No. of CH_2
11) No. of Ar-NH_2	12) No. of CH:CH

Journal of Computer Aided Chemistry , Vol.10, 1-9 (2009)

Физико-химические дескрипторы

Физико-химические дескрипторы – числовые характеристики, получаемые в результате моделирования физико-химических свойств химических соединений, либо величины, имеющие четкую физико-химическую интерпретацию.

Наиболее часто используются: липофильность (Log*P*), молярная рефракция (MR), молярная масса (MW), молекулярные объемы, площади поверхностей.

Другие дескрипторы: квантово-химические, дескрипторы молекулярных полей и т. д. и т. п.

Пример использования молекулярных дескрипторов при прогнозировании эффективности работы одного из сенсоров типа "химический нос".

Оказывается,

$$\text{Активность сенсора} = aE + bHB_D^2 + cMR^2$$

где E – энергия связи между сенсором и обнаруживаемой частицей,

HB_D – дескриптор, характеризующий донора водородной связи,

MR – молярная рефракция (характеризует размер молекулы и ее поляризуемость).

Липофильность log *P*

Липофильность (гидрофобность) характеризует способность растворяться в липидах (*и не только*).

Оценка способности вещества преодолевать клеточные мембраны.

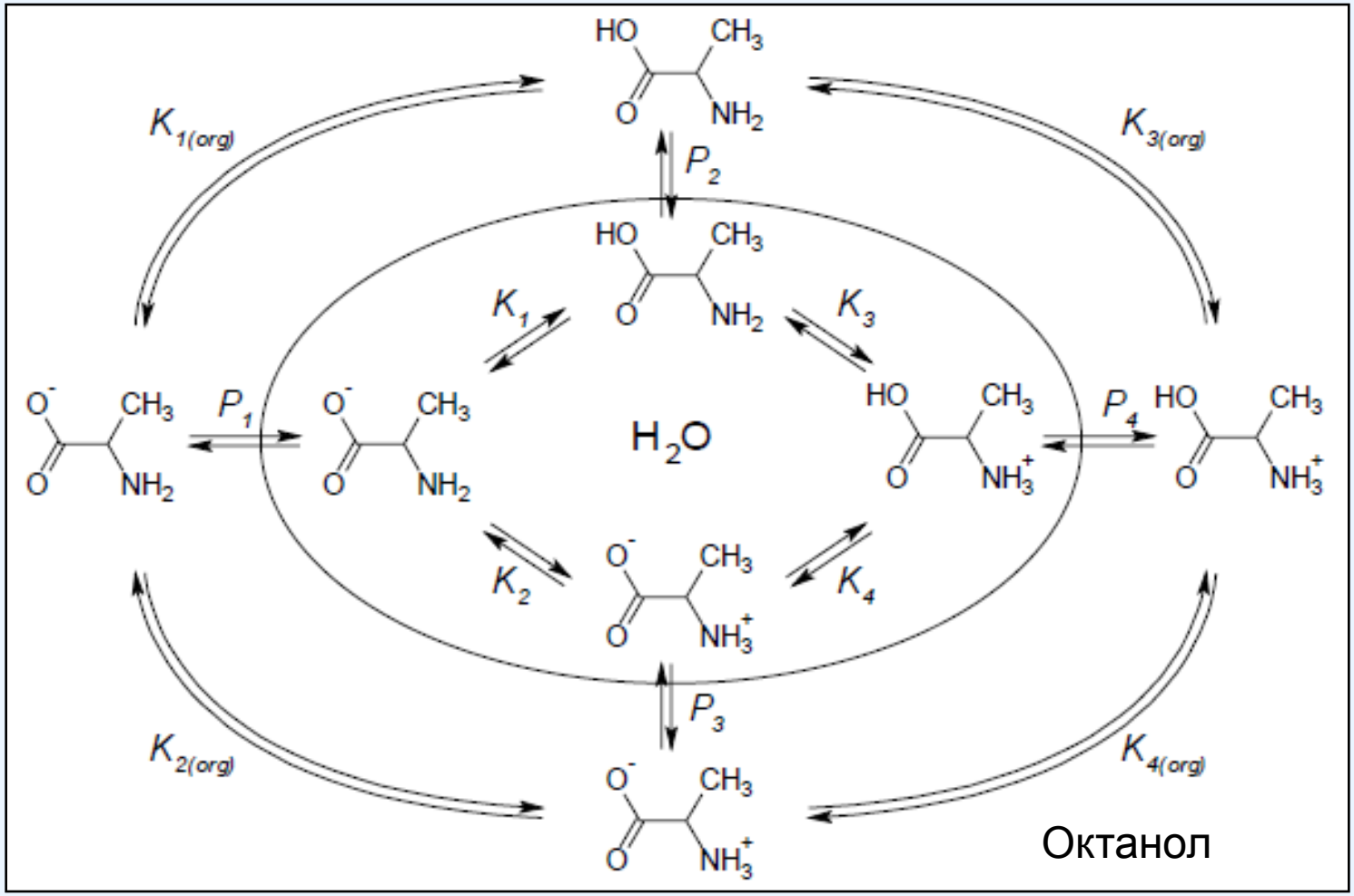
Моделируется: распределение вещества между октанолом и водой.

Коэффициент распределения :

$$P = \frac{c_{\text{октанол}}}{c_{\text{вода}}}$$

Липофильность (гидрофобность): log *P* = lg *P*

Если вещество образует ионы:



log *D*

$$\log D = \log\left(\frac{\sum a_i^{org}}{\sum a_i^{H_2O}}\right)$$

ACD/I-Lab. User's Guide. 2004.

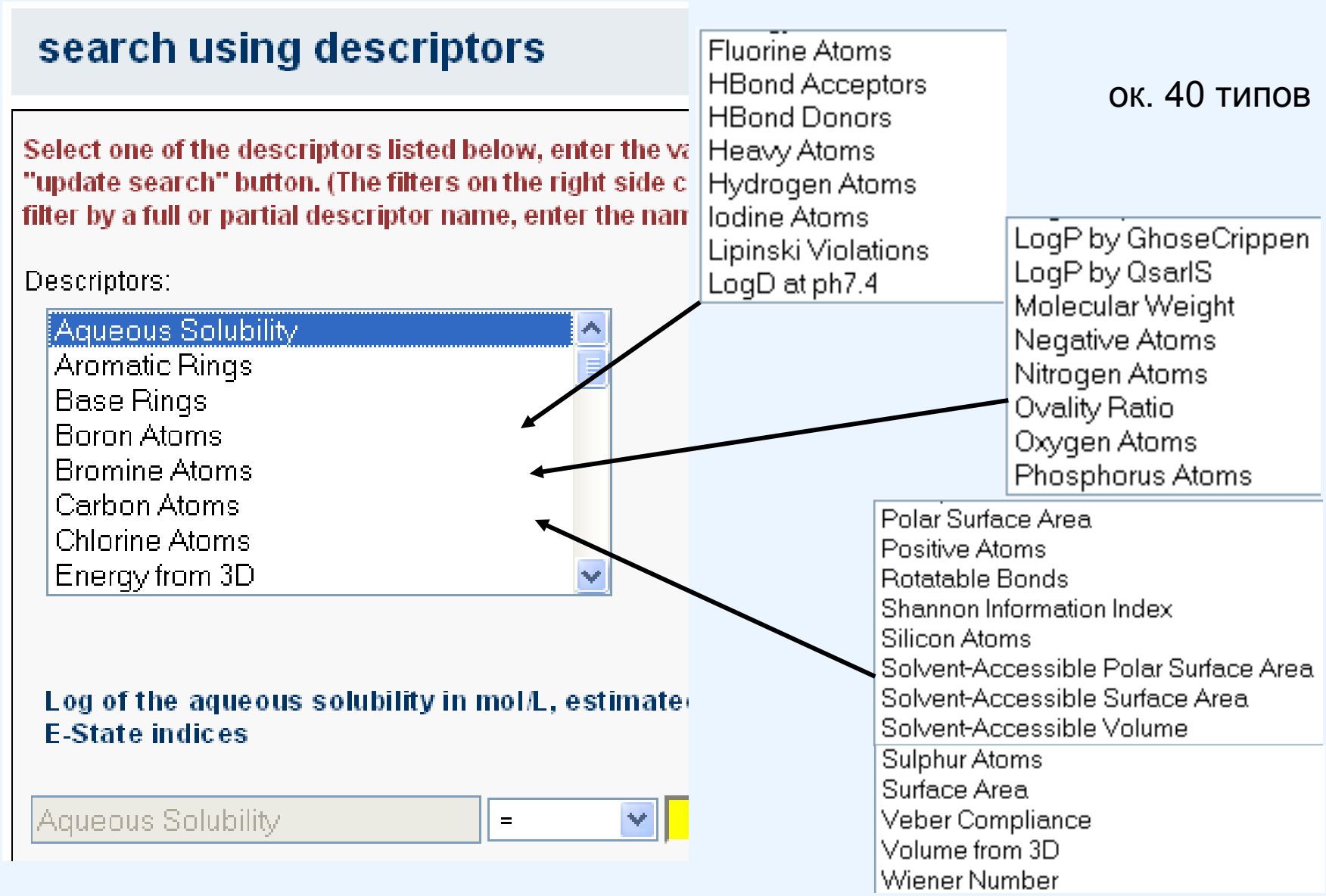
"Правило пяти" Липинского

Биологическая активность вещества (при приеме внутрь) более вероятна, если одновременно:

- Молярная масса ≤ 500.
- log*P* ≤ 5.
- Число доноров водорода водородной связи ≤ 5 (определяют по сумме OH- и NH-групп).
- Число акцепторов водорода водородной связи ≤ 10 (определяют по сумме атомов O и N).

У 70% веществ, имеющих признаки биологической активности: 0-2 донора водородной связи, 2-9 акцепторов водородной связи, 2-8 вращаемых связей, 1-4 цикла.

ZINC: Фрагмент поискового бланка с набором дескрипторов для запроса



ок. 40 типов

Обнаруживать корреляцию в тех ситуациях, когда переменных значительно больше, чем наблюдаемых фактов –
задача, обычная для химии

Data Mining

Интеллектуальный анализ данных — выявление скрытых закономерностей или взаимосвязей между переменными в больших массивах необработанных данных.

Английский термин «Data Mining» не имеет однозначного перевода на русский язык (*интеллектуальный анализ данных, добыча данных, вскрытие данных, информационная проходка, извлечение данных/информации*), поэтому в большинстве случаев используется в оригинале.

- Бритва Оккама:
  «Не следует множить сущее
  без необходимости»

«Матушка Природа
никогда не брилась
бритвой Оккама»

(*Garland R. Marshall, 2004*)