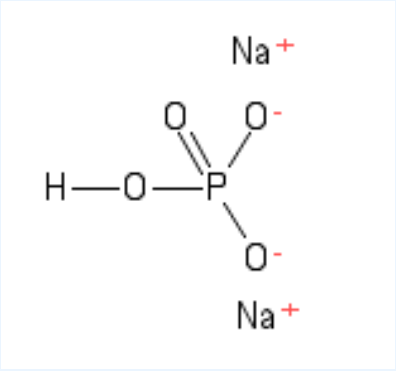
Одномерные формы отображения химического вещества в базах данных

Na_2HPO_4

Это:

гидроортофосфат натрия,
натрий-гидрофосфат,
кислый фосфат натрия двузамещенный,
динатрийгидротетраоксофосфат(V),
disodium orthophosphate,
hidrogenoortofosfato de disodio,

HNa_2O_4P ;
$Na_2O \cdot 0,5H_2O \cdot 0,5P_2O_5$;
$2Na^+ HPO_4^{2-}$
и т. д.



Вещество в справочнике, каталоге

Блок идентификаторов

IUPAC name	[hide] Ethane
Identifiers	
CAS number	74-84-0 ✓
PubChem	6324
EC number	200-814-8
UN number	1035
RTECS number	KH3800000
SMILES	[hide] CC
InChI	[hide] 1/C2H6/c1-2/h1-2H3
ChemSpider ID	6084

Wikipedia

14420 **Ethane**, 99+%
C_2H_6 *FW* 30·07 [74-84-0] *mp* -172° *bp* -88° *fp* -135° *RTECS* KH3800000
EINECS 200-814-8 *TSCA* *Merck* **12**,3767 *BRN* 1730716 *UN* 1035
EXTREMELY FLAMMABLE / KEEP COLD
R:12
S:9-16-33

$CH_3–CH_3$ **110g** **172.00**

Каталог реактивов Lancaster

CAS Registry Number

CASRN, CAS RN, CAS Number, CAS#

— номер, под которым химическое вещество (или смесь веществ) зарегистрировано в *Chemical Abstracts Service*.

Присваивается в хронологическом порядке, химический смысл не закладывается.

Формат: **7558-79-4**.

CAS Registry Number: особенности регистрации

Свой *CASRN* присваивается каждому химическому **объекту**, например:

Цис-1,2-дихлорэтен,
транс-1,2-дихлорэтен,

1,2-дихлорэтен (без указания изомерии)

C_2HDCl_2 (без указания изомерии)
цис-$C_2D_2Cl_2$

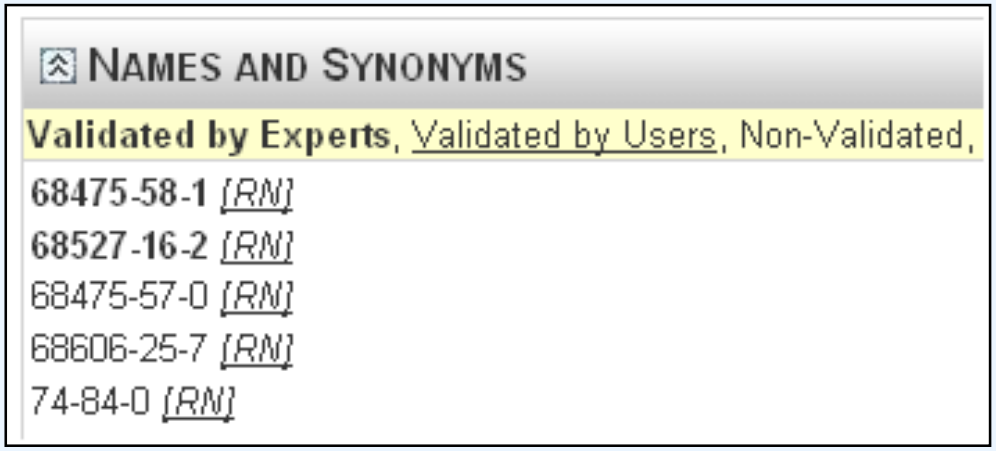
и т. д.

(в том числе, какая-нибудь особенно важная смесь, содержащая это вещество)

CAS Registry Number
проблемы использования

а) Доступ к полному списку *CASRN* – платный.

б) Плодятся неточности в бесплатном WWW.

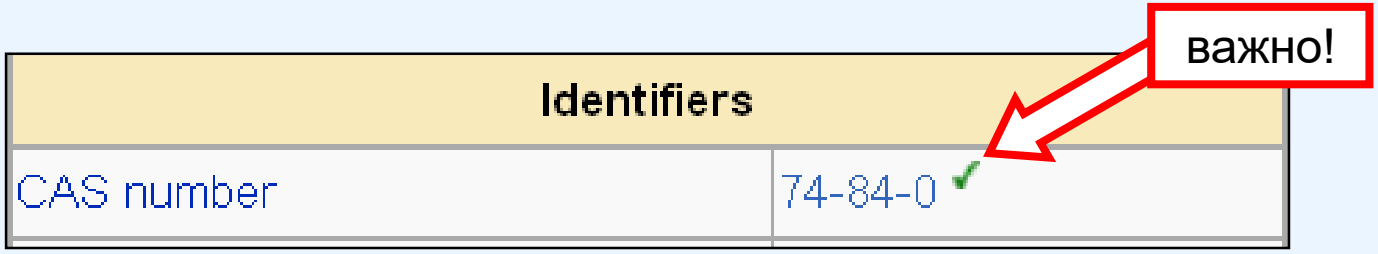


в) *CASRN* ⟺ химический объект *и*
CASRN ⟺ химическое вещество.

CAS Registry Number
где найти правильный код?

а) В редактируемых научных базах данных.
б) (В каталогах реактивов) (?).
в) Официальный краткий список:
Common Chemistry (http://www.commonchemistry.org/)

г) В Wikipedia (в статье о конкретном веществе).



Дополнительный материал:
CAS Registry Number и справочник Common Chemistry
http://www.abc.chemistry.bsu.by/bulchinf/2009_1_6-8.pdf

Двумерная графическая формула

(2D-структура)

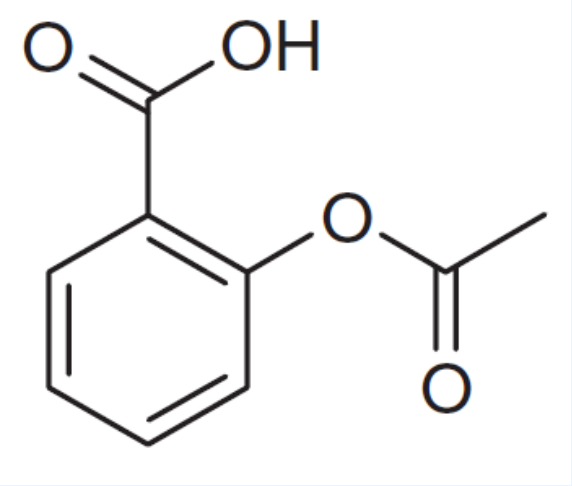
Топология структуры:

вид атомов и порядок их соединения друг с другом.

Топография структуры:

расположение атомов в пространстве.

Двумерная графическая формула отображает топологию структуры.



ацетилсалициловая кислота (аспирин)

Линейная нотация (линейная запись) –

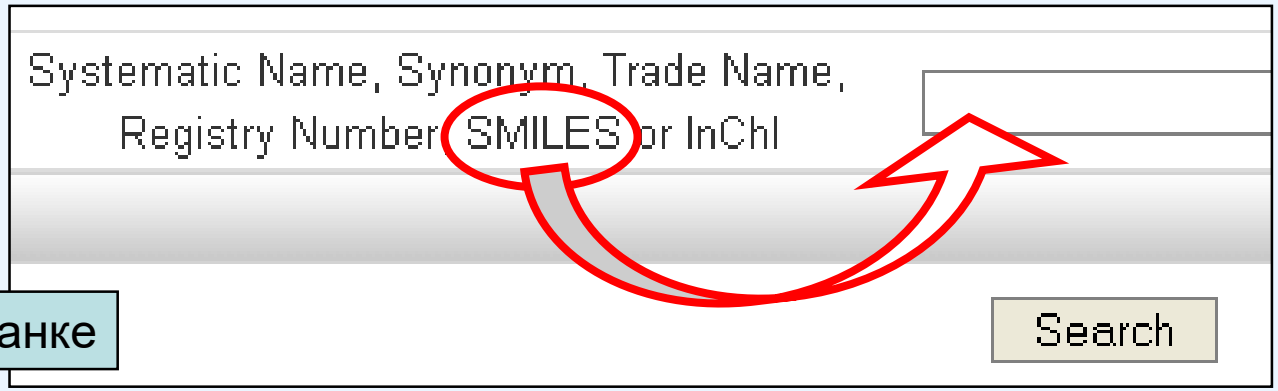
одномерная форма отображения химического объекта в виде строки буквенно-цифровых символов

Один из способов отображения топологии структуры в форме линейной нотации:

SMILES

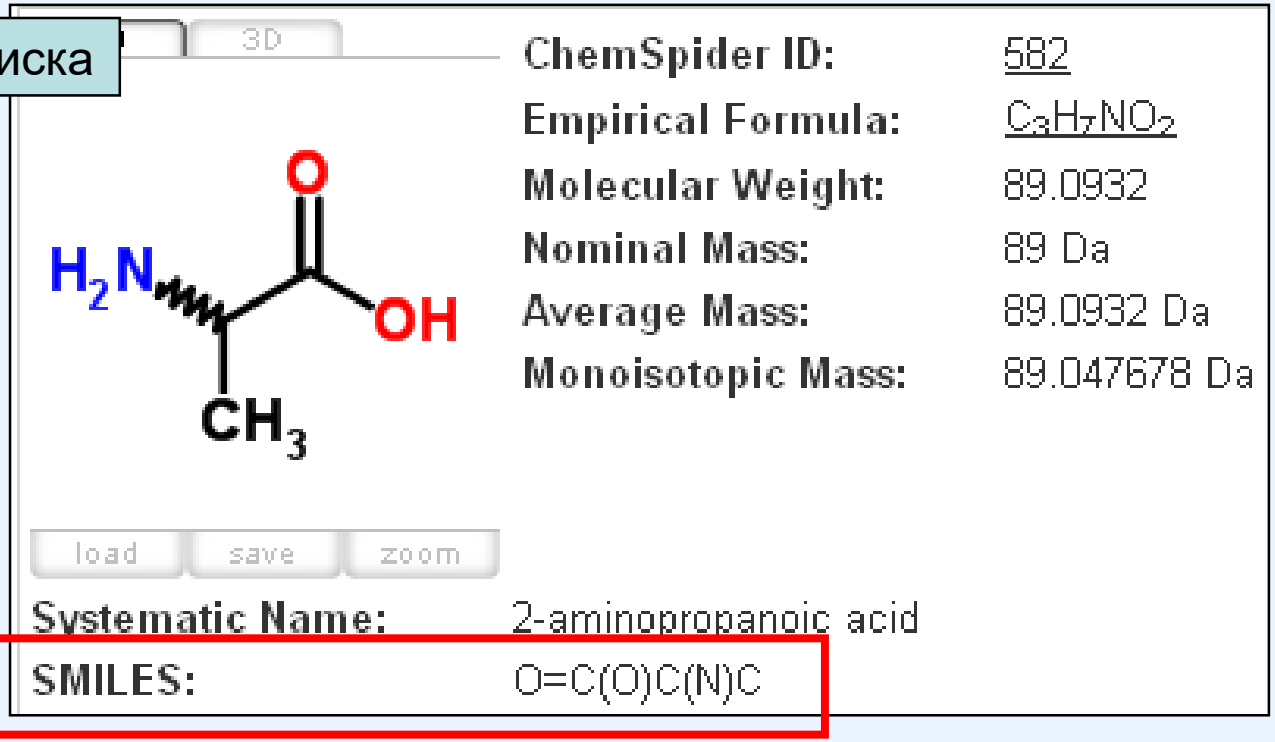
SMILES

[СМАЙЛЗ]



на поисковом бланке

в результатах поиска



Пример:

O=C(O)C(N)C

Синтаксис SMILES. Атомы.

В общем случае,
атомы отображаются символами химических элементов и записываются в квадратных скобках,

например: [As].

Без квадратных скобок
можно отображать атомы "органических" элементов
в "низших нормальных" валентных состояниях:

B(III) C(IV) N(III, V*) O(II) F(I)
P(III, V), S(II, IV, VI) Cl(I)
Br(I)
I(I)

* На самом деле – N(IV)

Водород при этих атомах, насыщающий свободные валентности, можно не указывать.

Гидриды

Объект	*Строка SMILES*
метан CH_4	C
аммиак NH_3	N
вода H_2O	O
сероводород H_2S	S
хлороводород HCl	Cl
арсин AsH_3	[AsH3]

Водород - в неявной форме

Водород - в явной форме

Ковалентная химическая связь

Соседние атомы записывают рядом.

Ковалентная связь отображается так:

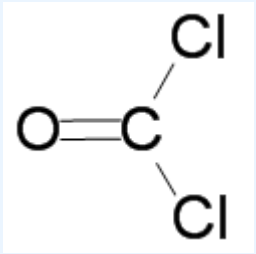
одинарная	никак (иногда -)
двойная	=
тройная	#

Например:

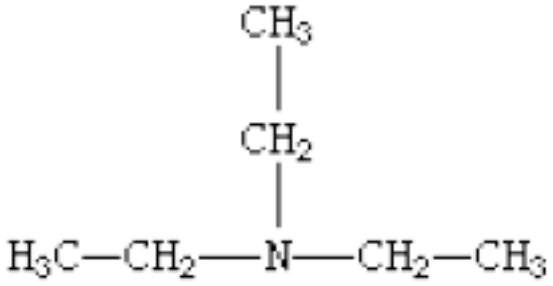
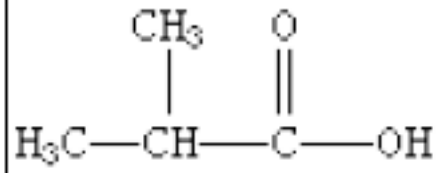
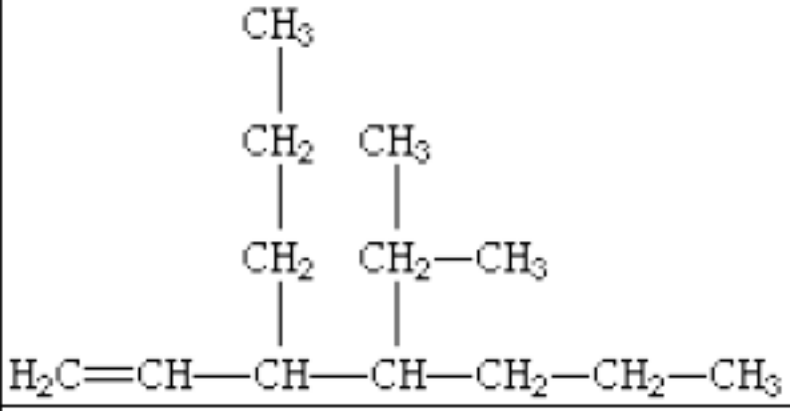
Объект	*Строка SMILES*
этан CH_3CH_3	CC
пропан $CH_3CH_2CH_3$	CCC
углекислый газ	O=C=O
синильная кислота	C#N

Боковые цепи (заместители)

Боковую цепь указывают в круглых скобках после символа того атома, к которому она присоединена.



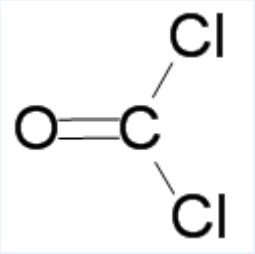
O=C(Cl)Cl

		
CCN(CC)CC	CC(C)C(=O)O	C=CC(CCC)C(C(C)C)CCC

Daylight Theory Manual v. 4.9, 2008.

Стандартная (каноническая) запись

Возможны многочисленные варианты записи:



O=C(Cl)Cl , ClC(Cl)=O
ClC(=O)Cl , C(Cl)(Cl)=O
и т. д.

Все они считаются правильными и каждый может использоваться как поисковый термин.

Один из вариантов является стандартным (каноническим). Стандартный генерируют по правилам, которые мы изучать не будем.

В базах данных информация хранится на основе канонических форм.

Компьютер сам преобразовывает запись пользователя в каноническую форму.

Ионы и ионные соединения

Заряд иона указывают внутри квадратных скобок.

Объект	*Строка SMILES*
Fe^{2+}	[Fe+2]
H_3O^+	[OH3+]
NH_4^+	[NH4+]

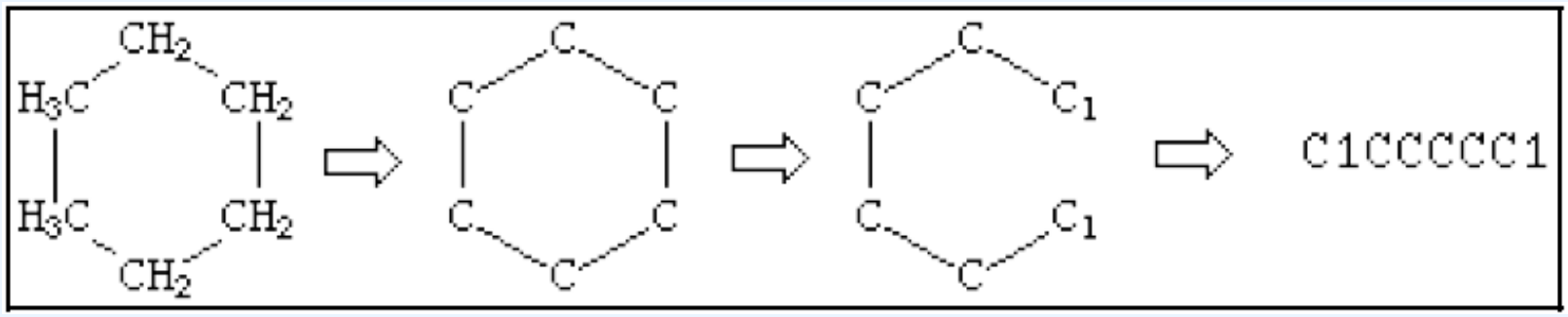
Обратим внимание на порядок записи цифр и знака "плюс"

Точкой отделяют автономные частицы.
Например, катион и анион.

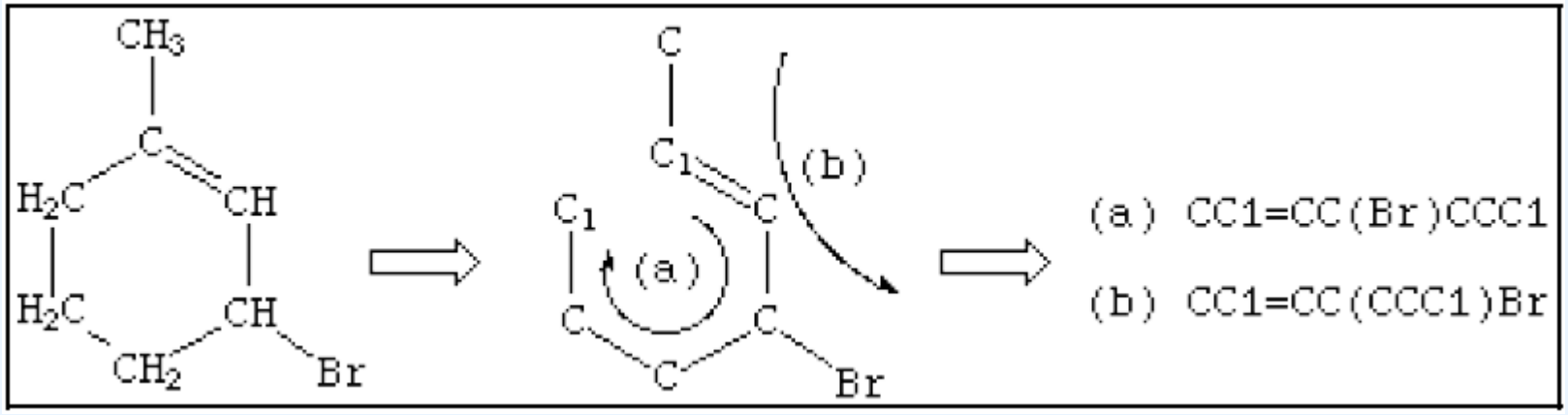
Объект	*Строка SMILES*
NaOH	[Na+].[OH-]

Циклы

Два атома цикла нумеруют одним и тем же числом и связь между этими атомами условно разрывают:



Допускаются варианты выбора основной цепи:



Daylight Theory Manual v. 4.9, 2008.

Ароматические соединения

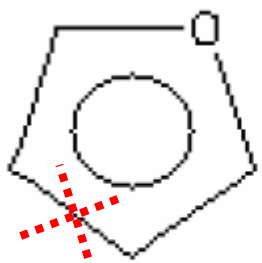
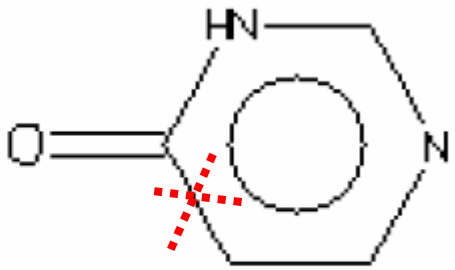
Химические символы атомов, образующих ароматические связи, записывают строчными буквами.

Пример 1. циклогексан C1CCCCC1

бензол c1ccccc1

фенол Oc1ccccc1

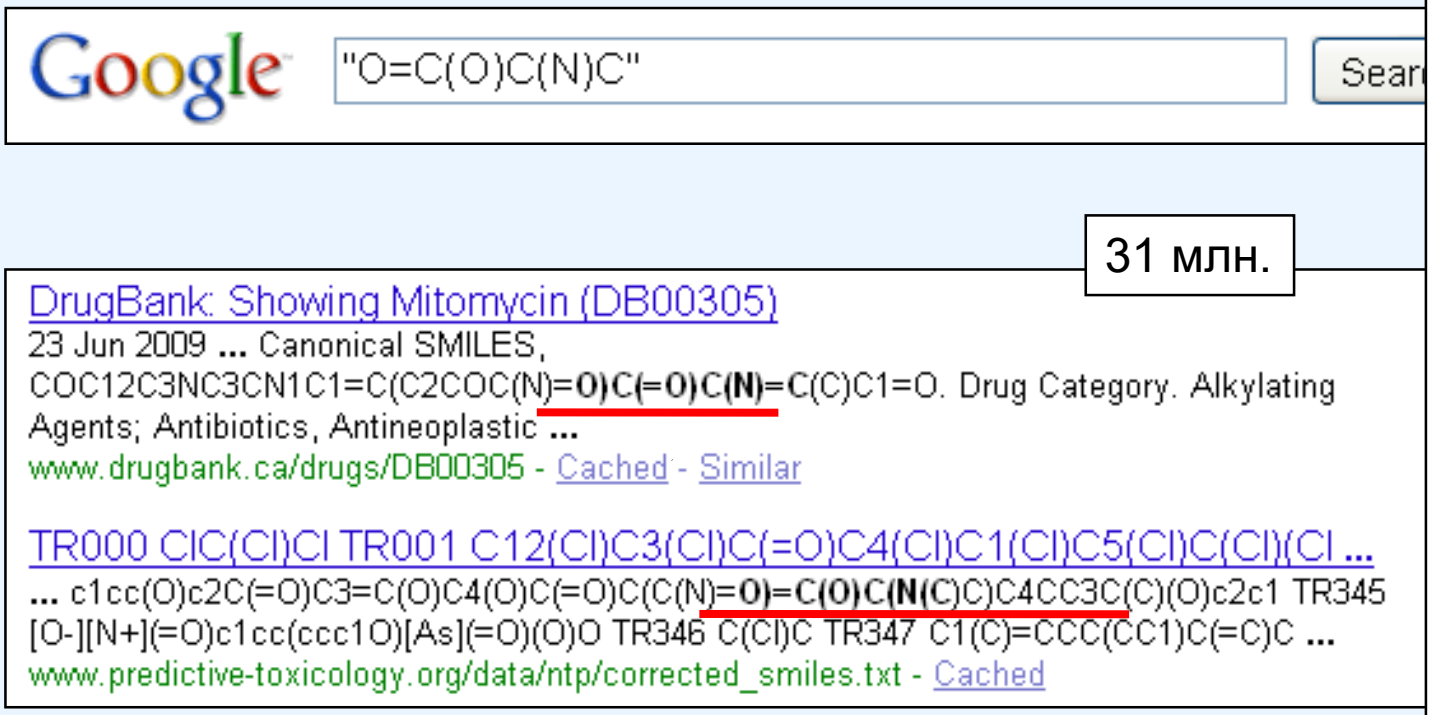
Пример 2.

	
c1cocc1	c1cnc[nH]c(=O)1

Daylight Theory Manual v. 4.9, 2008.

SMILES и Google

Код SMILES – буквенно-цифровой;
в принципе может быть компонентом запроса для универсальной поисковой системы.



Поиск работает, но есть проблемы:

а) запрос как фрагмент кода; как сумма терминов.

б) регистр букв тоже может иметь значение.