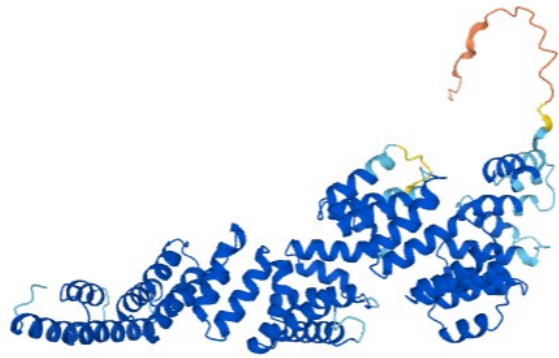
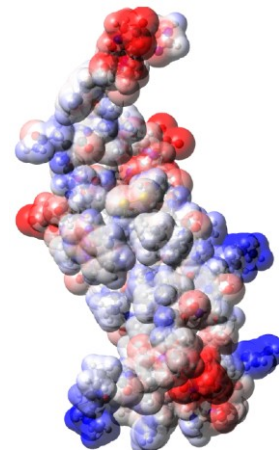
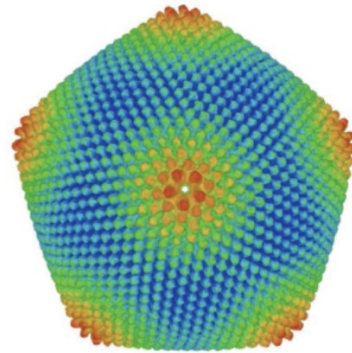
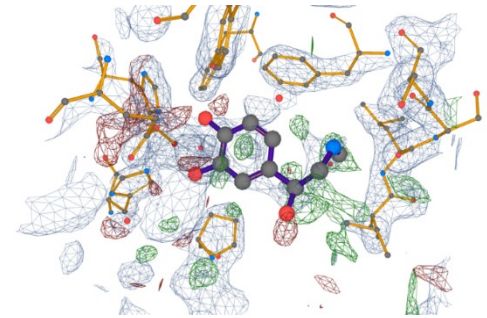


Computer model of a molecule

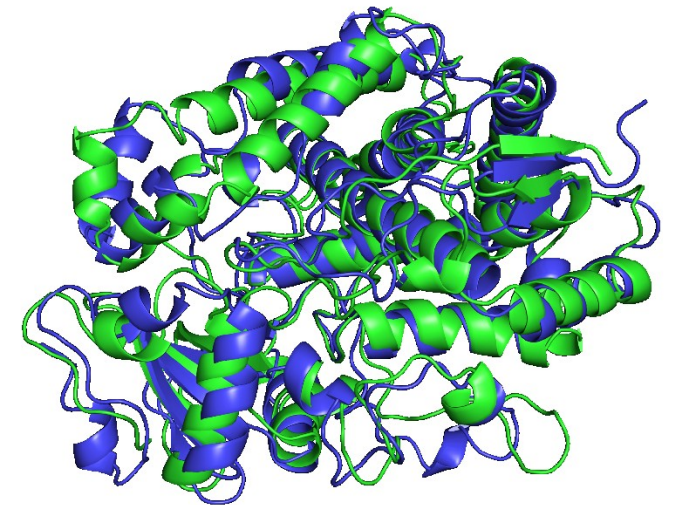
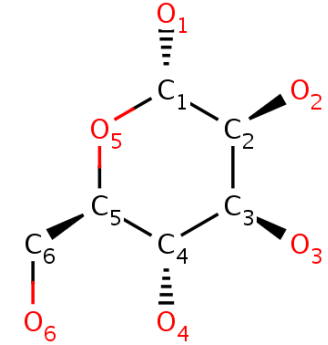


Radka Svobodová



Content

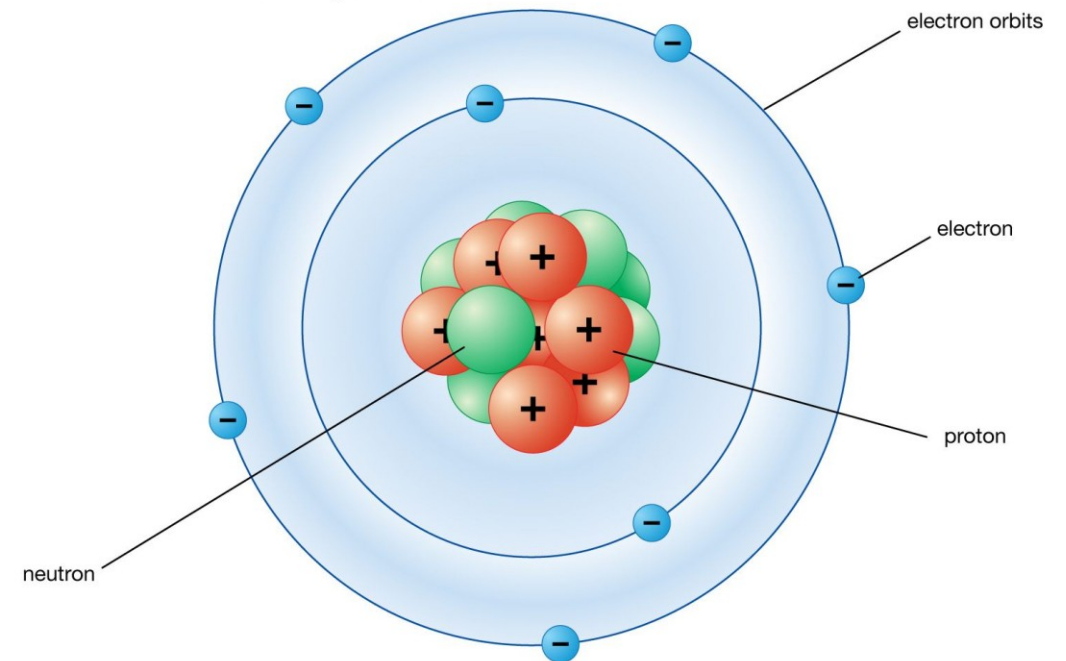
- **Introduction:** concept of chemoinformatics, content of the subject, history of the field
- **Computer model of a molecule:** 1D, 2D and 3D structure, molecule representation using graph and matrix
- **2D structure (topology) of a molecule:**
 - writing a molecule using a string (SMILES, InChi, InChiKey)
 - **Molecular graphs:** Isomorphism and canonical indexing
 - **Cycle search, fingerprints**
- **3D structure (geometry) of the molecule:**
 - representation using Cartesian and internal coordinates, data formats, geometry comparison



Basic chemical terms I

- **Atom:** basic building block from which substances are formed
- **Structure of an atom:**
 - Atom core: protons (positive charge), neutrons (no charge)
 - Electron shell: electrons (negative charge)

Bohr atomic model of a nitrogen atom



© Encyclopædia Britannica, Inc.

Basic chemical terms II

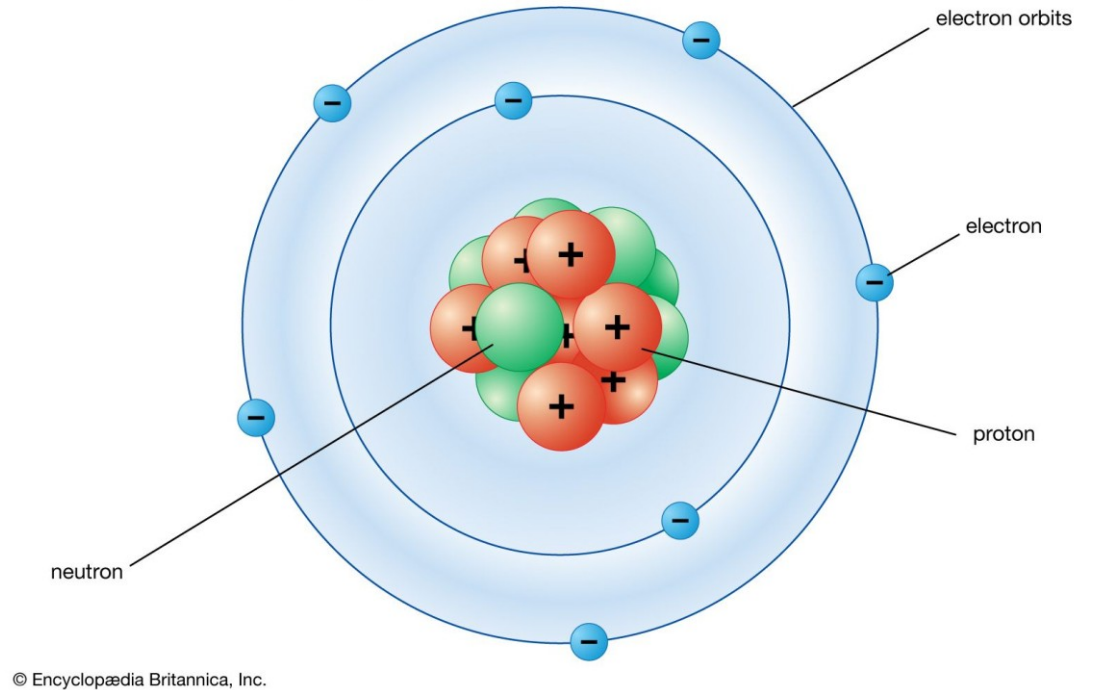
- All systems tend to occupy the state with the lowest possible total energy.



Basic chemical terms III

- The space of the electron shell can be divided into so-called **layers**.
- All electrons in a layer have the same energy value (this energy is **characteristic** of that layer).
- The further the layer is from the nucleus, the higher the energy of the electrons in it.
- The electrons in the electron shell therefore fill first the layer closest to the nucleus (the most energetically favorable), then the second closest, and so on.

Bohr atomic model of a nitrogen atom

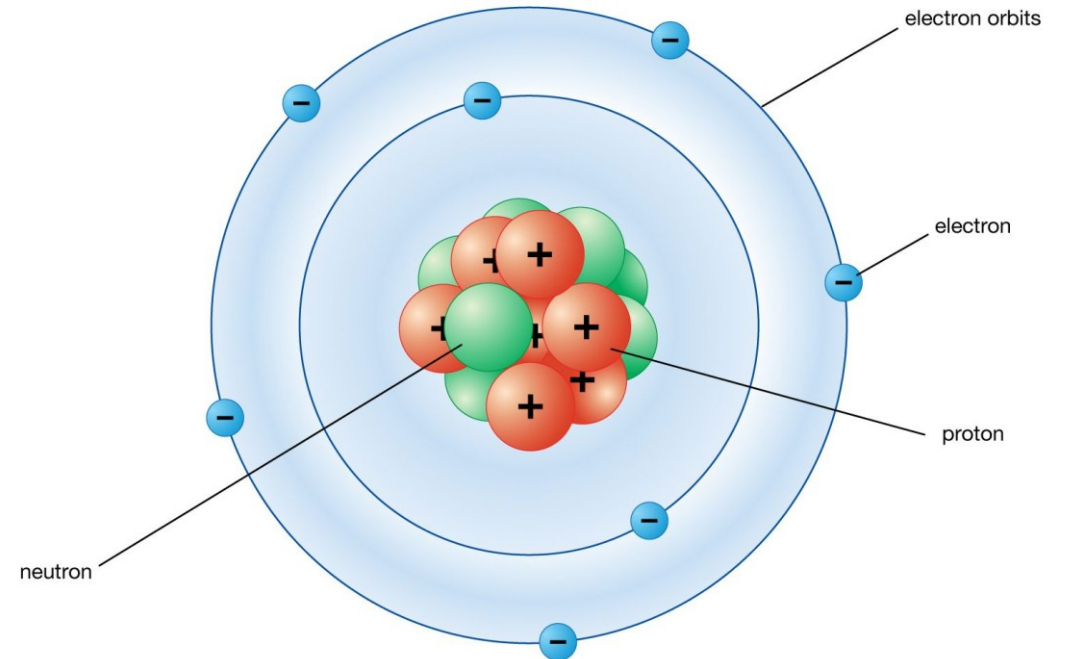


Basic chemical terms IV

- The non-empty layer that is farthest from the core is called the **valence layer**.
- In this layer are the so-called **valence electrons**.

These **valence electrons** are the subject of the study of chemistry because they can participate in chemical bonding.

Bohr atomic model of a nitrogen atom



© Encyclopædia Britannica, Inc.

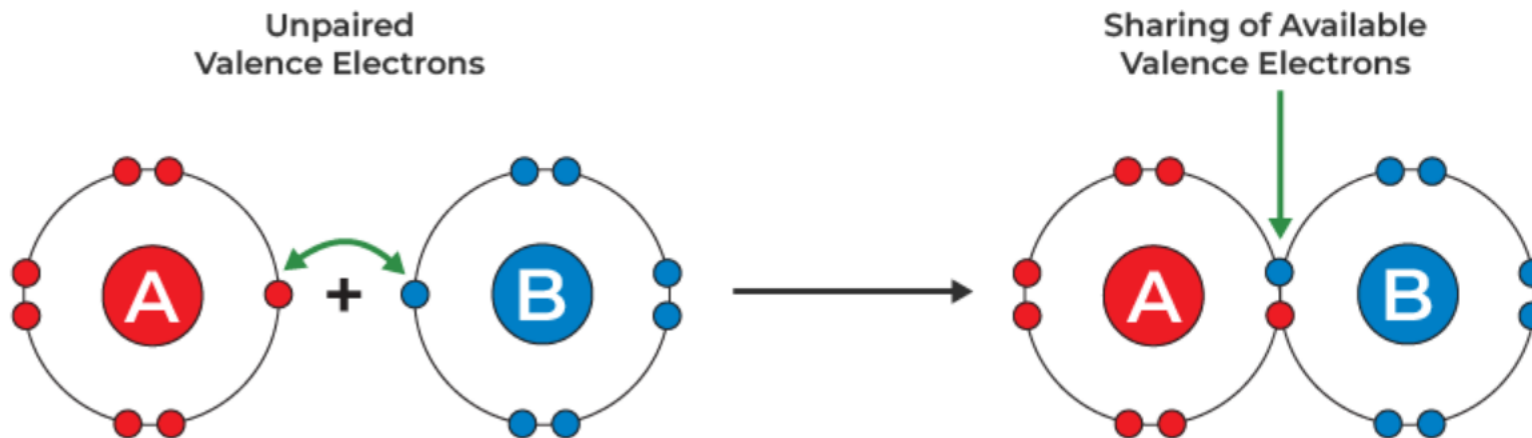
Basic chemical terms V

Chemical bond:

Two atoms come together at a sufficiently small distance (**bonding distance**) => overlapping of their electron shells.

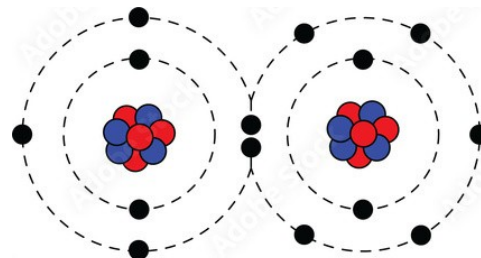
The valence electrons of both atoms change their trajectories.

If the resulting system has a lower energy than the original, the **atoms remain at bonding distance** => chemical bonding is formed.

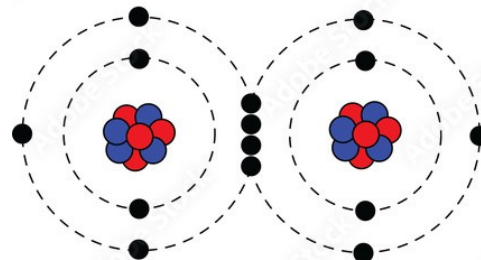


Basic chemical terms VI

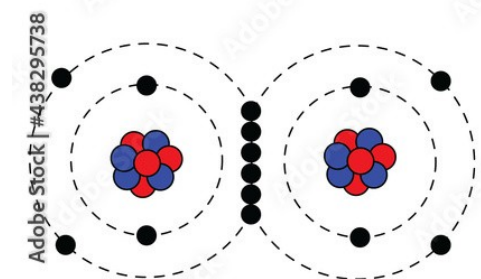
- **Bond order:**
 - **single bond:** two valence electrons are involved (binding electron pair)
 - **double bond:** two bonding electron pairs are involved
 - **Triple bond:** analogous
 - higher multiplicities do not occur in real chemical environments
 - **Aromatic bond:** when single and double bonds alternate, the electrons are delocalised among them. These bonds have properties between single and double bonds.



Single Covalent Bond



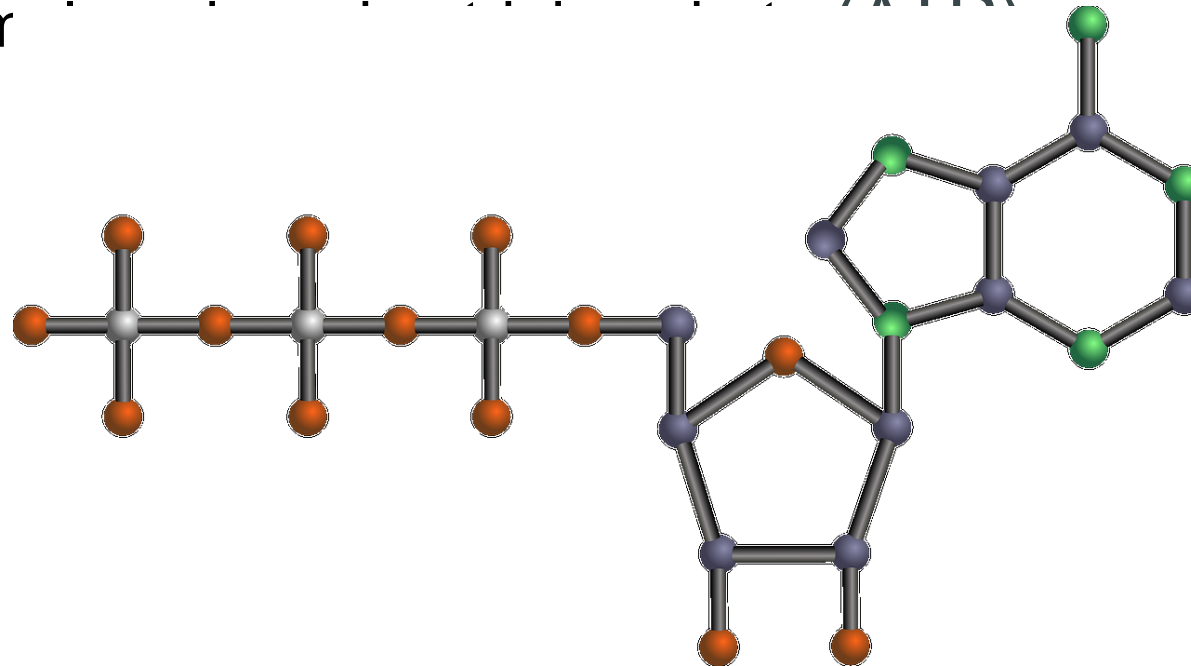
Double Covalent Bond



Triple Covalent Bond

Basic chemical terms VII

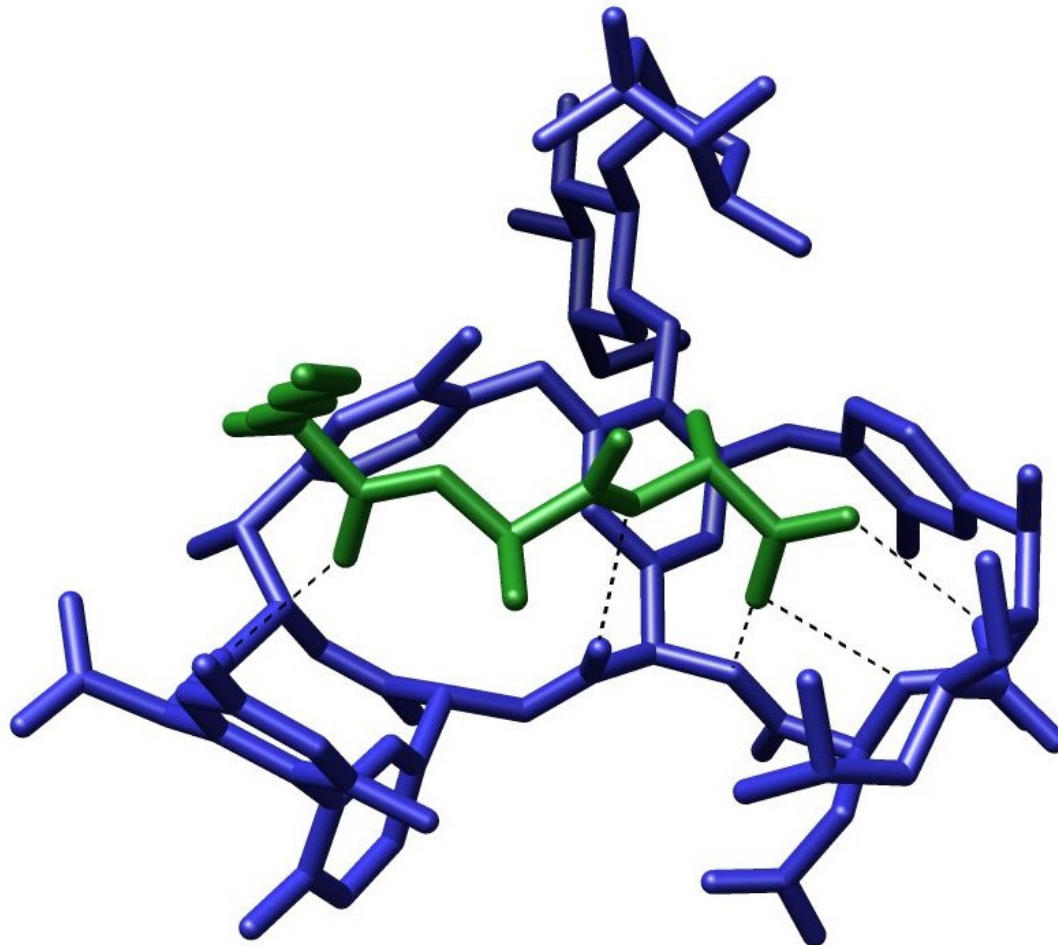
- **Molecule:** A system of atoms joined together by bonds to form a single unit. The basic structural unit of a substance. The carrier of the chemical properties of a substance.
- Exam



Basic chemical terms VIII

- **Molecular system:** A system containing one or more molecules.

Example:



Basic chemical terms IX

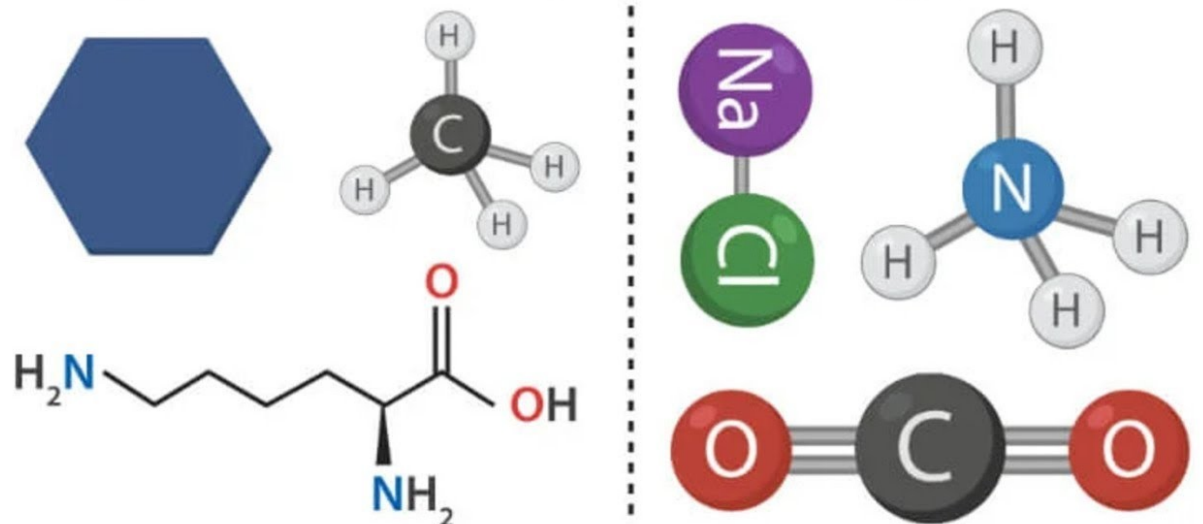
- **Organic molecules:**

- Their main component is carbon, the only element that is able to form longer chains of the $(-C-)n$ type, $n > 10$.
- This property of carbon allows the formation of complex molecules - the building blocks of living systems.
- Organic molecules also contain elements: H, O, S, N, F, Cl, Br, I

- **Inorganic molecules:**

- All molecules, that are not organic.

Organic vs Inorganic Compounds



How to describe a molecule in a computer?

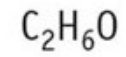
- Find out which information describes the molecule
- Write them into the computer



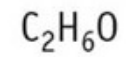
Which information describes the molecule?

Number of atoms?

Ethanol



Dimethyl
ether



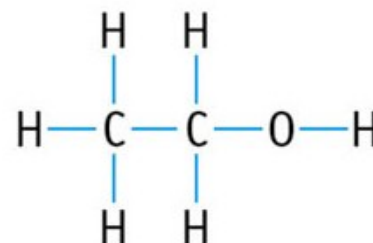
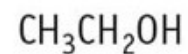
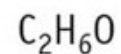
Which information describes the molecule?

Number of atoms?

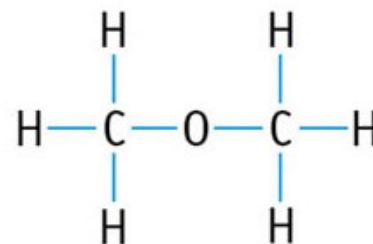
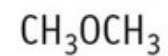
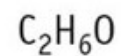
Not enough

Number of atoms and positions of bonds?

Ethanol



Dimethyl
ether



Which information describes the molecule?

Number of atoms?

Not enough

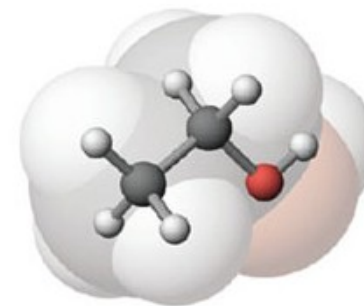
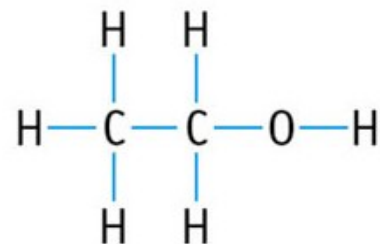
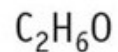
Number of atoms and positions of bonds?

Better

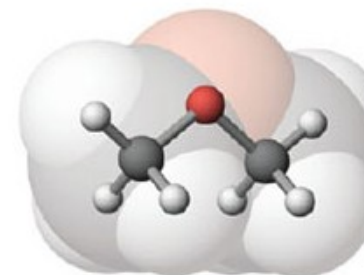
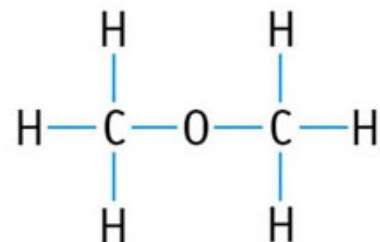
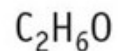
Number of atoms, positions of bonds and positions of atoms in 3D space?

Yes

Ethanol



Dimethyl ether



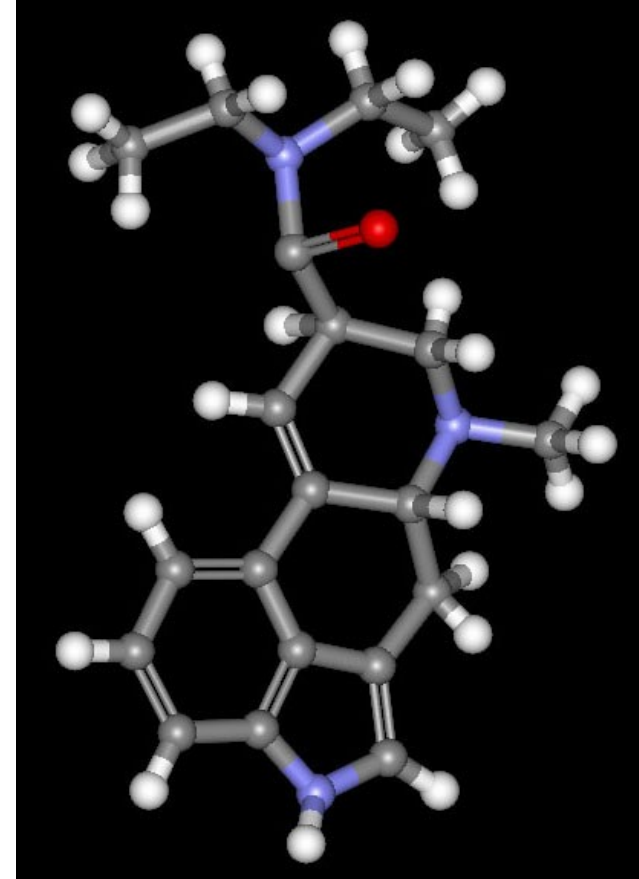
Model of molecule for computer processing

Atoms:

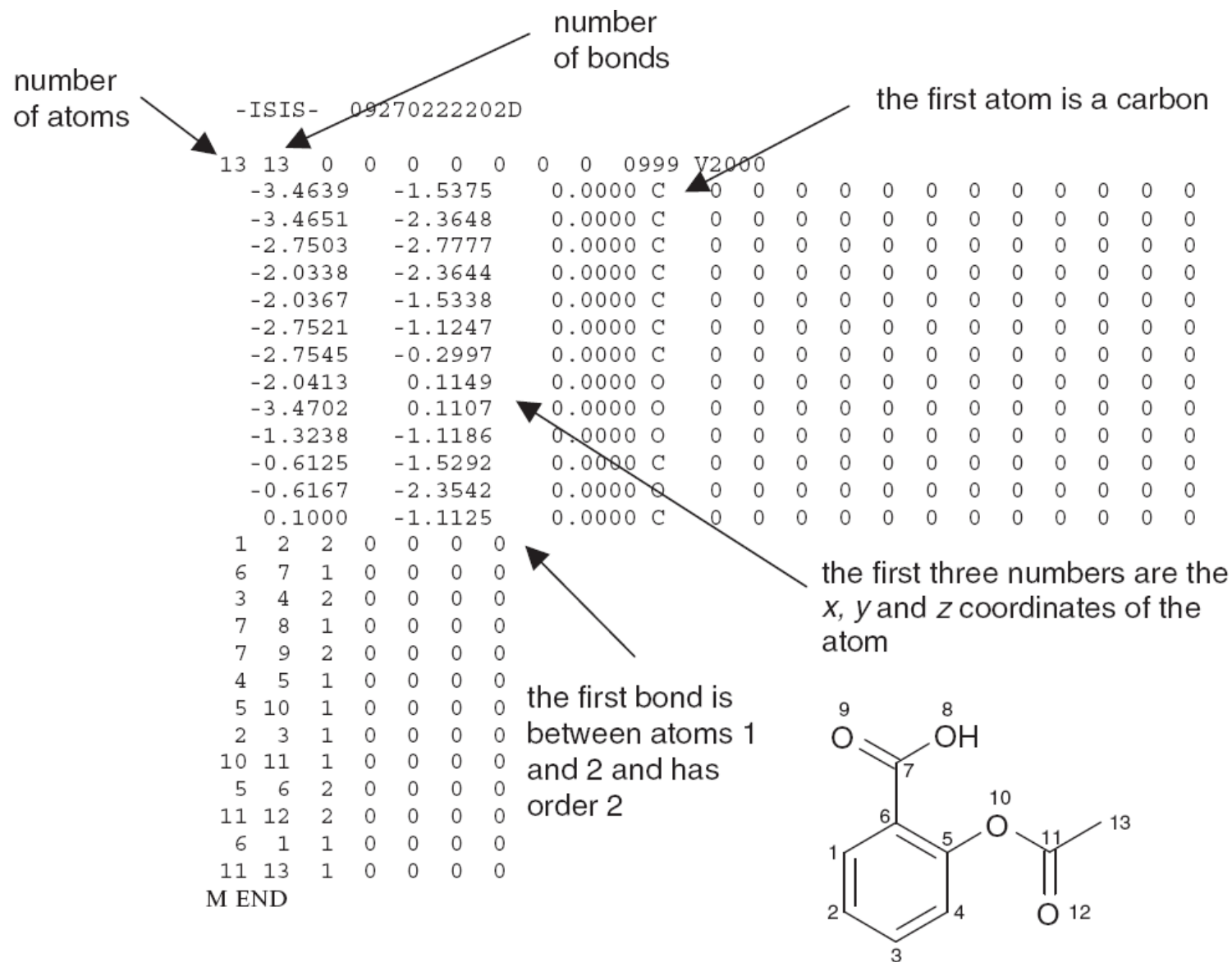
- Points in space
- Chemical symbol of the element listed for each

Bonds:

- Pairs of atoms that are bonded
- Bond order



Description of a molecule in a computer



```

21 21 0 0 0 0 0 0 0 0 0 1 V2000
 18.7769 -15.2504 -0.1032 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 18.7571 -16.6359 -0.1252 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 17.5868 -14.5409 -0.1114 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 17.5465 -17.3106 -0.1545 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 16.3767 -15.2158 -0.1421 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 16.3559 -16.6013 -0.1633 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 17.6081 -13.0313 -0.0880 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 20.0592 -14.5322 -0.0715 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 17.5247 -18.7799 -0.1764 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 15.1150 -14.4620 -0.1527 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 20.0742 -13.3140 -0.0089 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 21.1073 -15.1564 -0.0523 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 16.4750 -19.3759 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 18.5697 -19.4030 -0.2650 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 14.0496 -15.0560 -0.1515 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 15.1330 -13.2425 -0.1568 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 19.7111 -17.2054 -0.1194 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 15.3860 -17.1427 -0.1873 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 17.6136 -12.6451 -1.1298 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 16.7057 -12.6567 0.4410 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 18.5209 -12.6823 0.4410 H 0 0 0 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
 3 7 1 0 0 0 0
 1 8 1 0 0 0 0
 4 9 1 0 0 0 0
 5 10 1 0 0 0 0
 8 11 2 0 0 0 0
 8 12 2 0 0 0 0
 9 13 2 0 0 0 0
 9 14 2 0 0 0 0
 10 15 2 0 0 0 0
 10 16 2 0 0 0 0
 17 2 1 0 0 0 0
 18 6 1 0 0 0 0
 19 7 1 0 0 0 0
 20 7 1 0 0 0 0
 21 7 1 0 0 0 0
M END

```

Challenge:

Draw this molecule.
What is the name of
the molecule?

Databases of small organic molecules

> 1 M structures of small molecules

- Small molecule: < 100 atoms
- Small molecules = “drug-like” molecules
- Experimental structures
- Predicted (computed) structures

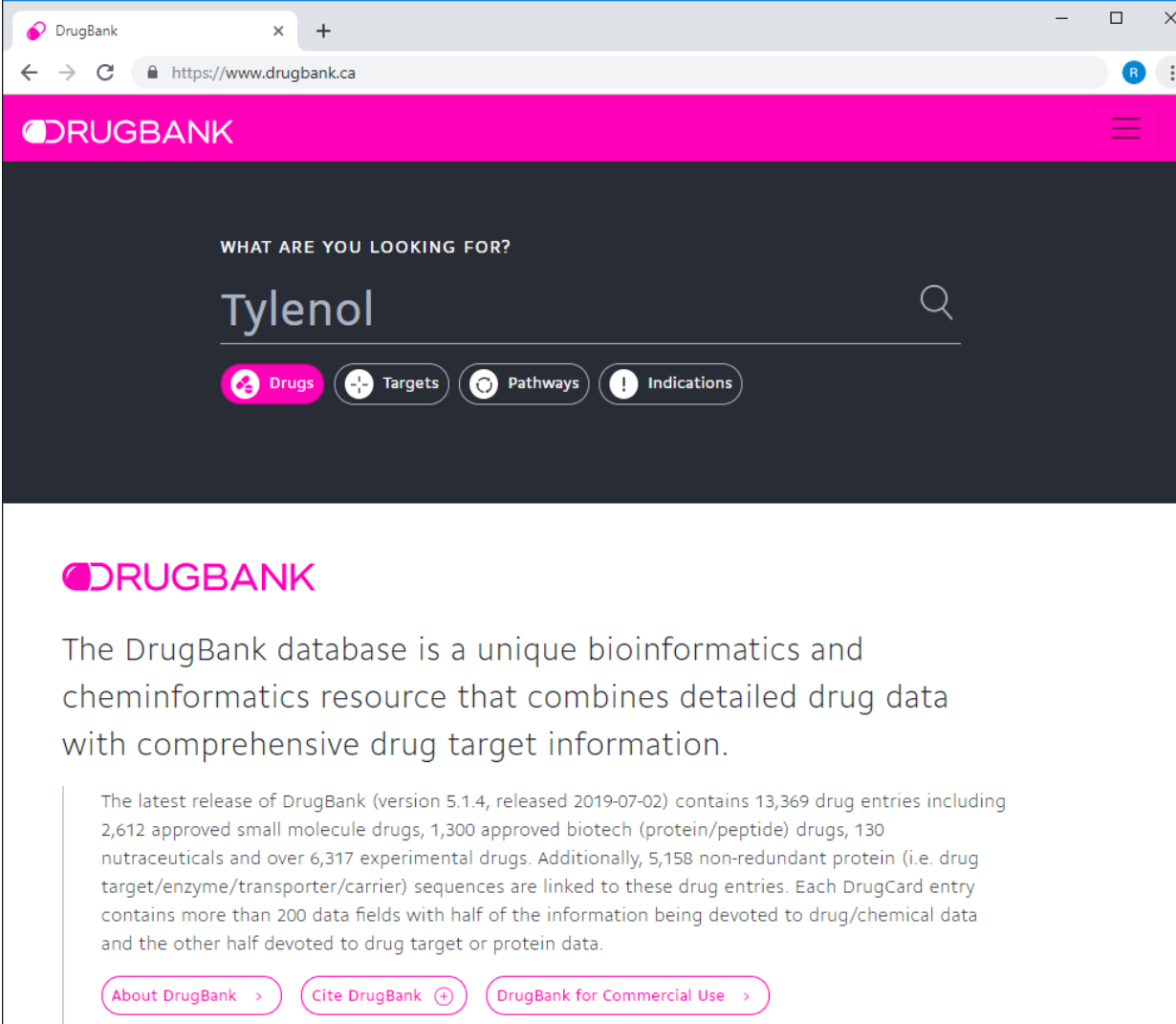
Pub  hem

ZINC

 ChEMBL

 DRUGBANK

DrugBank – database of drugs



The screenshot shows a web browser window with the DrugBank website. The browser's address bar shows the URL <https://www.drugbank.ca>. The website has a pink header with the DrugBank logo and a search bar. The search bar contains the text "Tylenol" and a magnifying glass icon. Below the search bar are four buttons: "Drugs" (highlighted in pink), "Targets", "Pathways", and "Indications".

DRUGBANK

WHAT ARE YOU LOOKING FOR?

Tylenol

Drugs Targets Pathways Indications

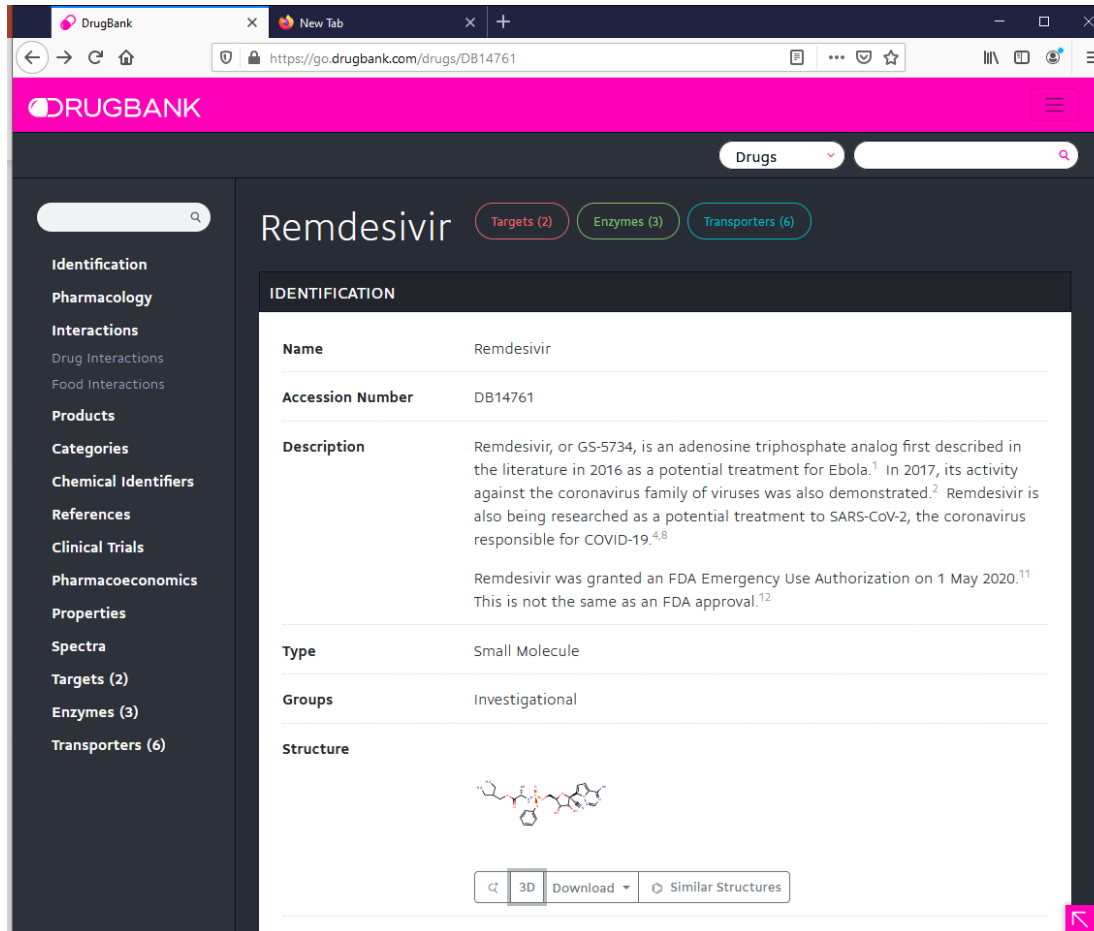
DRUGBANK

The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information.

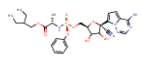
The latest release of DrugBank (version 5.1.4, released 2019-07-02) contains 13,369 drug entries including 2,612 approved small molecule drugs, 1,300 approved biotech (protein/peptide) drugs, 130 nutraceuticals and over 6,317 experimental drugs. Additionally, 5,158 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. Each DrugCard entry contains more than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.

[About DrugBank >](#) [Cite DrugBank +](#) [DrugBank for Commercial Use >](#)

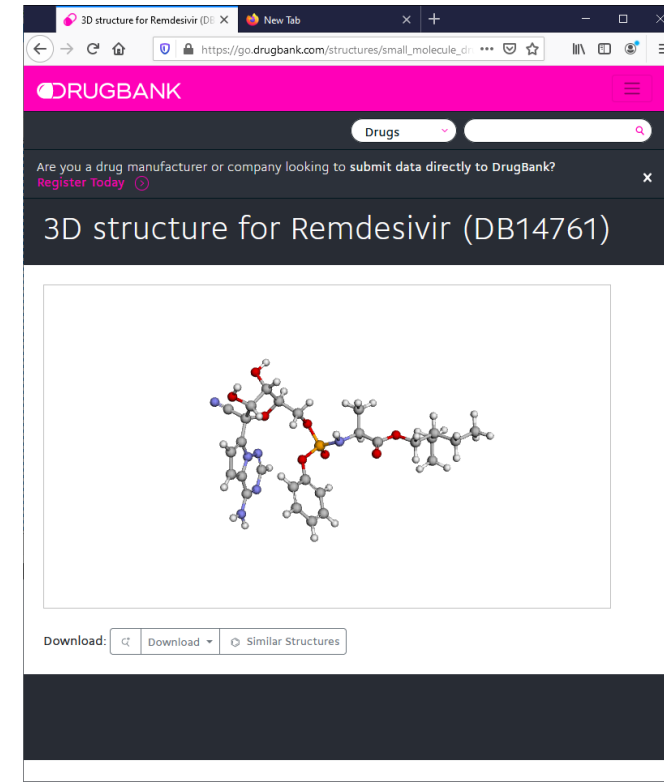
DrugBank – database of drugs



The screenshot shows the DrugBank website interface. The top navigation bar is pink with the DrugBank logo and a search bar. Below the navigation bar, there is a search bar and a dropdown menu set to "Drugs". The main content area is titled "Remdesivir" and includes tabs for "Targets (2)", "Enzymes (3)", and "Transporters (6)". A sidebar on the left lists various categories such as Identification, Pharmacology, Interactions, Products, Categories, Chemical Identifiers, References, Clinical Trials, Pharmacoeconomics, Properties, Spectra, Targets (2), Enzymes (3), and Transporters (6). The main content area displays the following information:

IDENTIFICATION	
Name	Remdesivir
Accession Number	DB14761
Description	<p>Remdesivir, or GS-5734, is an adenosine triphosphate analog first described in the literature in 2016 as a potential treatment for Ebola.¹ In 2017, its activity against the coronavirus family of viruses was also demonstrated.² Remdesivir is also being researched as a potential treatment to SARS-CoV-2, the coronavirus responsible for COVID-19.^{4,8}</p> <p>Remdesivir was granted an FDA Emergency Use Authorization on 1 May 2020.¹¹ This is not the same as an FDA approval.¹²</p>
Type	Small Molecule
Groups	Investigational
Structure	

At the bottom of the structure section, there are buttons for "3D", "Download", and "Similar Structures".



The screenshot shows the DrugBank website interface for the 3D structure of Remdesivir. The top navigation bar is pink with the DrugBank logo and a search bar. Below the navigation bar, there is a search bar and a dropdown menu set to "Drugs". The main content area is titled "3D structure for Remdesivir (DB14761)". A notification banner at the top asks if the user is a drug manufacturer or company looking to submit data directly to DrugBank, with a "Register Today" link. The main content area displays a 3D ball-and-stick model of the Remdesivir molecule. Below the model, there are buttons for "Download" and "Similar Structures".

PubChem – database of organic molecules

The image shows a screenshot of the PubChem website homepage. At the top, the browser address bar shows the URL <https://pubchem.ncbi.nlm.nih.gov/>. The page header includes the NIH logo and the text "U.S. National Library of Medicine National Center for Biotechnology Information". Below this is the PubChem logo and navigation links for "About", "Blog", "Submit", and "Contact". A promotional banner for the American Chemical Society National Meeting in San Diego (August 25-29, 2019) is also visible.

The main content area features a large "Explore Chemistry" heading with the tagline "Quickly find chemical information from authoritative sources". A search bar is prominently displayed, with a magnifying glass icon on the right. Below the search bar, a row of "Try" examples includes "aspirin", "EGFR", "C9H8O4", "57-27-2", "C1=CC=C(C=C1)C=O", and "InChI=1S/C3H6O/c1-3(2)4/h1-2H3". Below these examples are radio buttons for "Use Entrez", "Compounds", "Substances", and "BioAssays", with "Compounds" selected.

Four icons represent key features: "Draw Structure" (a pencil and chemical structure), "Upload ID List" (an upward arrow), "Browse Data" (a grid of squares), and "Periodic Table" (a grid of dots).

At the bottom of the page, a statistics bar displays: "96M Compounds", "236M Substances", "268M Bioactivities", "30M Literature", "3M Patents", and "693 Data Sources". Links for "See More Statistics >" and "Explore Data Sources >" are provided.

Below the screenshot, the text "What is PubChem?" is centered.

PubChem – database of organic molecules

The screenshot shows a web browser window with the URL <https://pubchem.ncbi.nlm.nih.gov/compound/5288826>. A red banner at the top contains a warning icon and text: "COVID-19 is an emerging, rapidly evolving situation. Get the latest public health information from CDC: <https://www.coronavirus.gov>. Get the latest research from NIH: <https://www.nih.gov/coronavirus>." Below this is the NIH logo and "National Library of Medicine National Center for Biotechnology Information". The main header features the PubChem logo, navigation links (About, Blog, Submit, Contact), and a search bar.

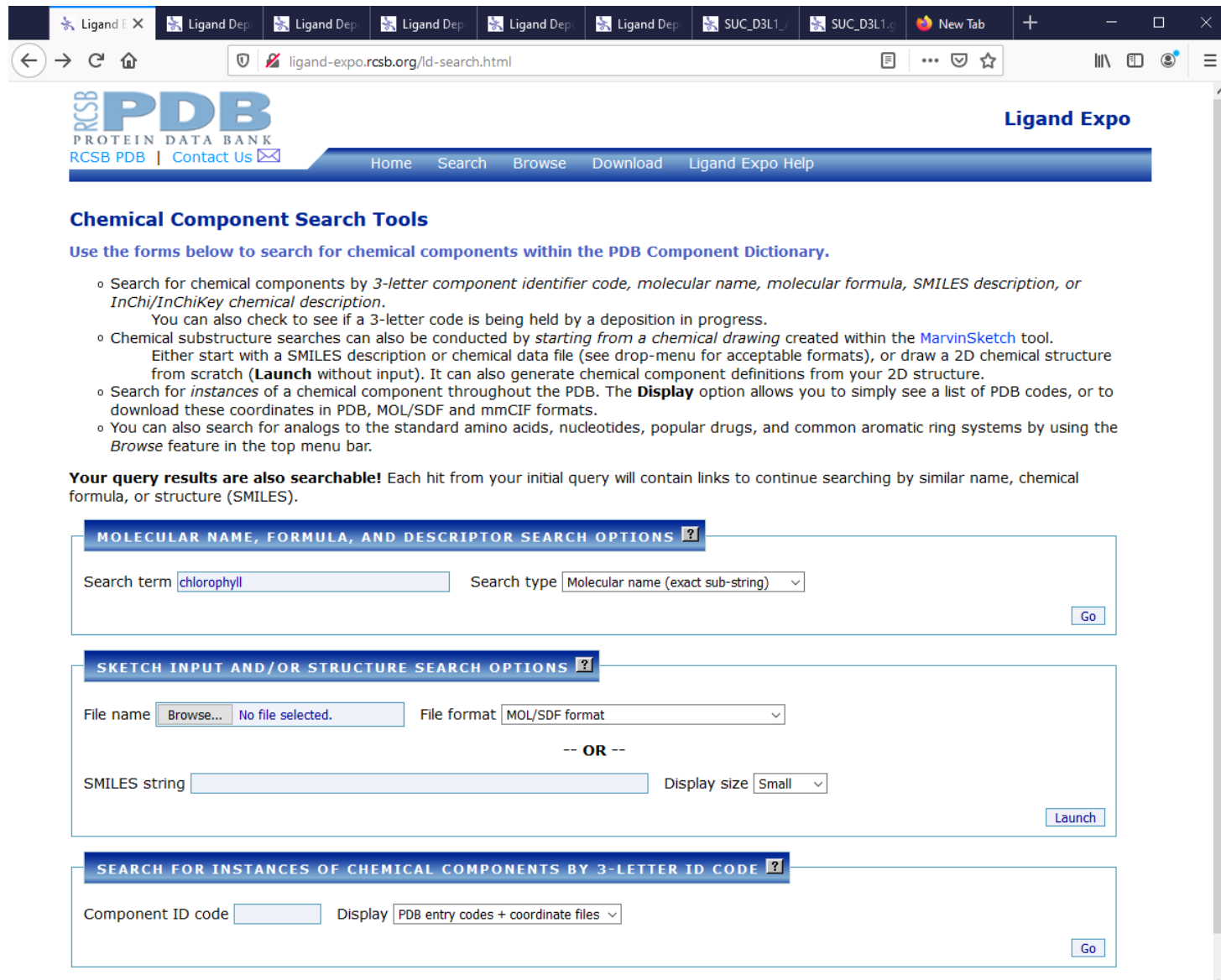
The main content area is titled "COMPOUND SUMMARY" and features the name "Morphine" in large text. To the right are social media sharing options (Share, Tweet, Email) and buttons for "Cite" and "Download". A "CONTENTS" sidebar on the right lists sections: Title and Summary (selected), 1 Structures, 2 Names and Identifiers, 3 Chemical and Physical Properties, 4 Spectral Information, 5 Related Records, 6 Chemical Vendors, 7 Drug and Medication Information, 8 Pharmacology and Biochemistry, and 9 Use and Manufacturing.

The "Structure" section is highlighted with a blue border and contains the following information:

- PubChem CID: 5288826
- Structure: Three molecular representations are shown: 2D (a skeletal structure), 3D (a ball-and-stick model), and Crystal (a ball-and-stick model with a unit cell). Below these is a link to "Find Similar Structures".
- Chemical Safety: A red diamond warning icon with an exclamation mark is shown, with the text "Irritant" below it.

Ligand Expo – database of ligands

Ligand = molecule bound in a protein



The screenshot shows a web browser window with the URL `ligand-expo.rcsb.org/ld-search.html`. The page header includes the RCSB PDB logo and a navigation menu with links for Home, Search, Browse, Download, and Ligand Expo Help. The main content area is titled "Chemical Component Search Tools" and provides instructions on how to use the search forms.

Chemical Component Search Tools

Use the forms below to search for chemical components within the PDB Component Dictionary.

- Search for chemical components by 3-letter component identifier code, molecular name, molecular formula, SMILES description, or InChi/InChiKey chemical description.
You can also check to see if a 3-letter code is being held by a deposition in progress.
- Chemical substructure searches can also be conducted by starting from a chemical drawing created within the MarvinSketch tool.
Either start with a SMILES description or chemical data file (see drop-menu for acceptable formats), or draw a 2D chemical structure from scratch (**Launch** without input). It can also generate chemical component definitions from your 2D structure.
- Search for instances of a chemical component throughout the PDB. The **Display** option allows you to simply see a list of PDB codes, or to download these coordinates in PDB, MOL/SDF and mmCIF formats.
- You can also search for analogs to the standard amino acids, nucleotides, popular drugs, and common aromatic ring systems by using the Browse feature in the top menu bar.

Your query results are also searchable! Each hit from your initial query will contain links to continue searching by similar name, chemical formula, or structure (SMILES).

MOLECULAR NAME, FORMULA, AND DESCRIPTOR SEARCH OPTIONS ?

Search term Search type

SKETCH INPUT AND/OR STRUCTURE SEARCH OPTIONS ?

File name No file selected. File format

-- OR --

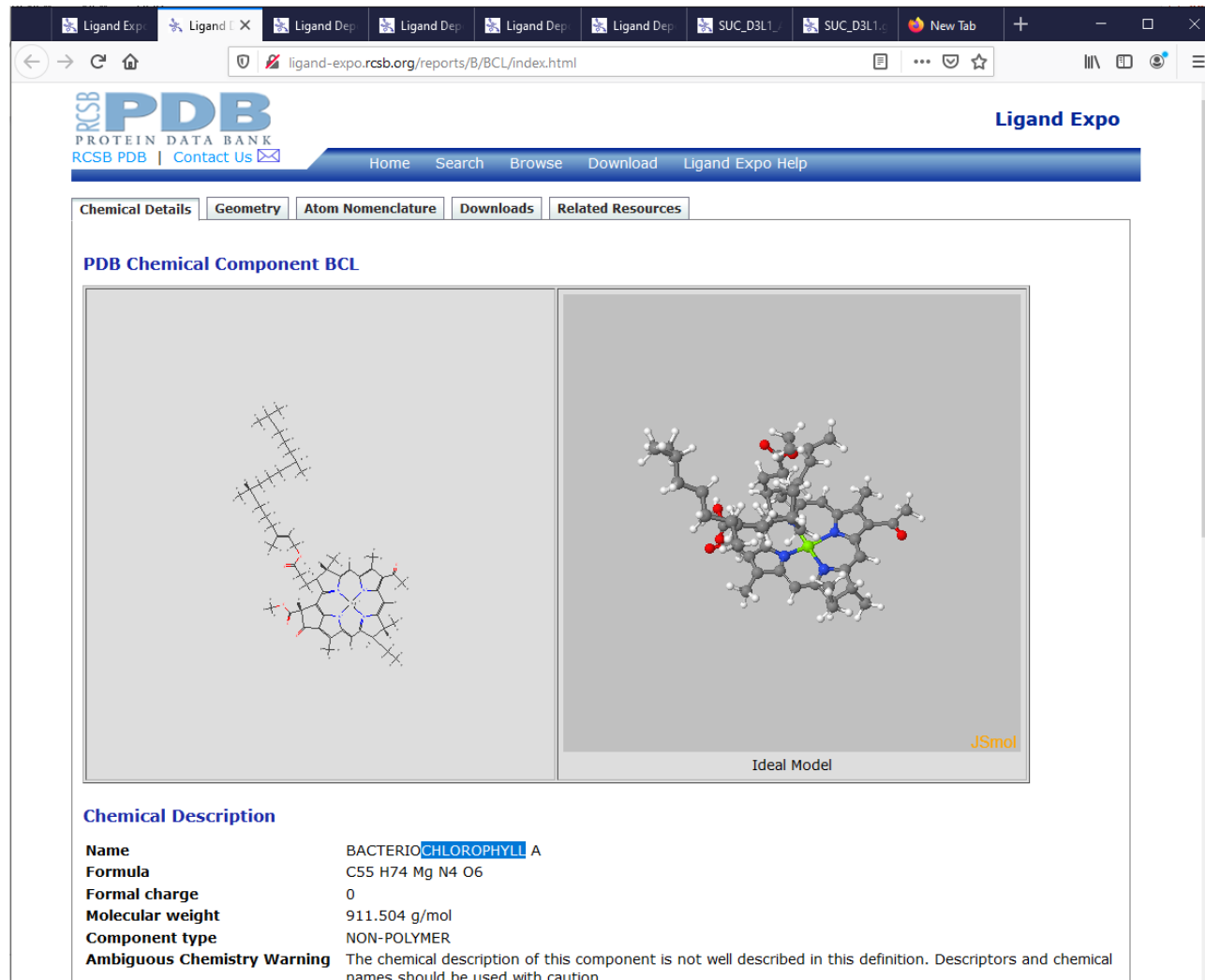
SMILES string Display size

SEARCH FOR INSTANCES OF CHEMICAL COMPONENTS BY 3-LETTER ID CODE ?

Component ID code Display

Ligand Expo – database of ligands

Ligand = molecule bound in a protein



The screenshot shows the PDB Ligand Expo interface for the chemical component BCL. The browser address bar shows the URL ligand-expo.rcsb.org/reports/B/BCL/index.html. The page features the PDB logo and navigation links. Below the navigation, there are tabs for 'Chemical Details', 'Geometry', 'Atom Nomenclature', 'Downloads', and 'Related Resources'. The main content area displays the title 'PDB Chemical Component BCL' and two molecular models: a 2D chemical structure on the left and a 3D ball-and-stick model on the right, labeled 'Ideal Model' with a 'JSmol' viewer control. Below the models is a 'Chemical Description' section with the following details:

Name	BACTERIO CHLOROPHYLL A
Formula	C55 H74 Mg N4 O6
Formal charge	0
Molecular weight	911.504 g/mol
Component type	NON-POLYMER
Ambiguous Chemistry Warning	The chemical description of this component is not well described in this definition. Descriptors and chemical names should be used with caution.

Databases of biomacromolecules

Mainly proteins

> 200 k experimental structures

> 200 M computed structures



AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

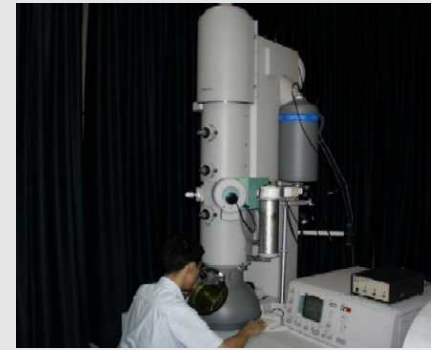
Protein Data Bank – sources of data



89% X-ray
crystallography

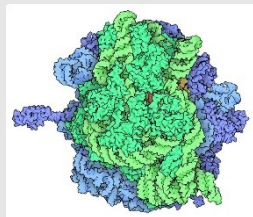
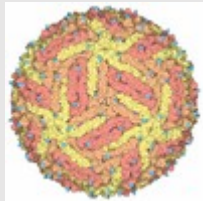


10% NMR
spectroscopy

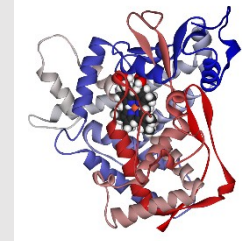


1% cryoelectron
microscopy

3D struktura

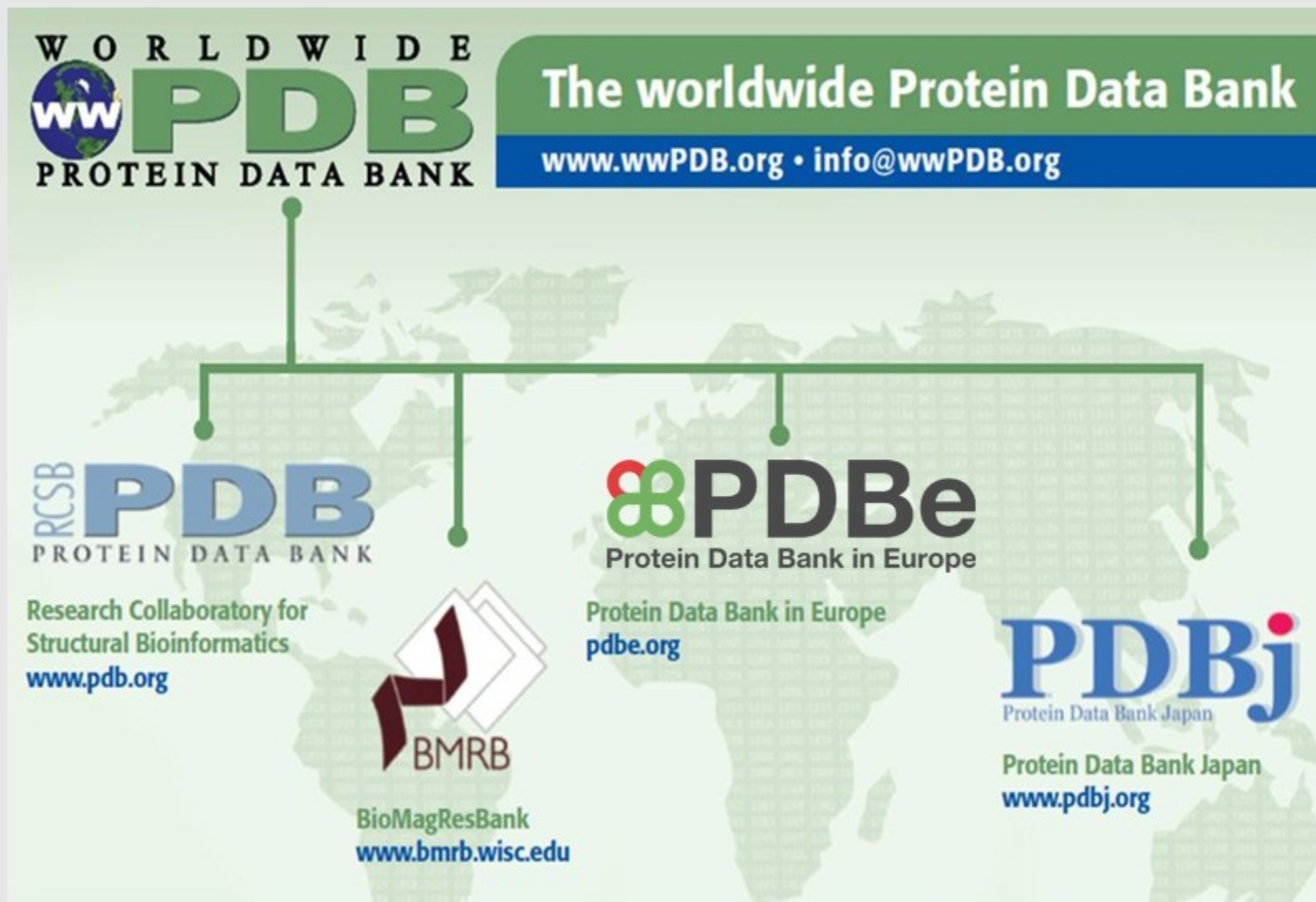


...								
ATOM	46	C	GLY	A	70	51.536	23.360	40.507
ATOM	47	O	GLY	A	70	50.947	22.279	40.325
ATOM	48	N	ILE	A	71	50.965	24.532	40.270
ATOM	49	CA	ILE	A	71	49.595	24.644	39.786
...								



WORLDWIDE
PDB
PROTEIN DATA BANK

Protein Data Bank



> 225 000 biomacromolecular structures

Protein Data Bank

The screenshot shows the Protein Data Bank in Europe (PDBe) homepage. The browser address bar displays "European Bioinformatics Institute [GB] | https://www.ebi.ac.uk/pdbe/". The website header includes the EMBL-EBI logo, the PDBe logo, and the text "Protein Data Bank in Europe Bringing Structure to Biology". A search bar is located in the top right corner with a "Search" button and "Advanced search" link. Below the header is a navigation menu with links for "PDBe home", "Deposition", "PDBe services", "PDBe training", "Documentation", and "About PDBe".

The main content area is divided into several sections:

- Featured structure:** "Pygmalion 5x2g" (1st September 2019). The text states: "The September image in our 2019 calendar is inspired by a molecular system that can edit DNA and the story of a statue coming to life." A "Read more..." link is provided.
- News:** Three recent news items are listed:
 - "Links added to raw experimental data at PDBe" (12 August, 2019)
 - "Improve your previously released PDB coordinates with OneDep" (1 August, 2019)
 - "A celebration of the PDB Art project" (26 July, 2019)
 - "Mandatory mmCIF format for crystallographic depositions to the PDB" (1 July, 2019)
- Events:** Two events are listed:
 - "Art Exhibition: Molecules of Life" by **Kendrew Foyer, EBI South Building Wellcome Genome Campus** (10 Sep 2019 to 27 Sep 2019)
 - "EBI Structural bioinformatics course" by **EMBL-EBI, Cambridge, UK** (16 Sep 2019 to 20 Sep 2019)
 - "EBI Exploring Biological Sequence course" by **EMBL-EBI, Cambridge, UK** (10 Oct 2019)
- Popular:** A list of popular links including PDBe-KB, EMsearch, PDBeFold, PDBePISA, PDBeChem, Sequence search, PDBe REST API, EM resources, NMR resources, EMPIAR, Coordinate Server, and PDB Component Library.
- Latest archive statistics:** "As of 11 September 2019 the PDB contains 155830 entries (latest PDB entries, chemistry, biology) and EMDB contains 9016 entries (latest map releases, latest header releases, latest updates)." (Note: The original image contains a typo '155830' which has been corrected to '155830').
- Tweets by @PDBEurope:** A tweet from David Armstrong (@DaveASci) is shown, dated 13h, mentioning a training course in Medellin, Colombia.

A footer banner at the bottom of the page states: "This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our [Privacy Notice](#) and [Terms of Use](#). I agree, dismiss this banner".

Protein Data Bank

PDBe > 3hyu

Crystal structure of the altitude adapted hemoglobin of guinea pig.
Source organism: *Cavia porcellus*

Primary publication:
Structure of the altitude adapted hemoglobin of guinea pig in the R2-state.
Pairet B, Jaenicke E
PLoS ONE 5 e12389 (2010)
PMID: 20811494

X-ray diffraction
1.67Å resolution
Released: 23 Jun 2010
Model geometry: [Progress bar]
Fit model/data: [Progress bar]

Quick links

- 3hyu overview
- Citations
- Structure analysis
- Function and Biology
- Ligands and Environments
- Experiments and Validation
- View
- Downloads
- 3D Visualisation

Function and Biology [Details]

Biochemical function: heme binding

Biological process: oxygen transport

Cellular component: hemoglobin complex

Sequence domains:

- Haemoglobin, alpha-type
- Haemoglobin, beta-type
- Globin
- Globin/Protoglobin
- Globin-like superfamily

Ligands and Environments

2 bound ligands:

- 2 x HEM
- 4 x PO4

No modified residues

Structure analysis [Details]

Assembly composition: hetero tetramer (preferred)

Entry contents: 2 distinct polypeptide molecules

Macromolecules (2 distinct):

- Hemoglobin subunit alpha

Experiments and Validation [Details]

Metric	Percentile Ranks	Value
Rfree	[Bar chart]	0.201
Clashscore	[Bar chart]	3
Ramachandran outliers	[Bar chart]	0
Sidechain outliers	[Bar chart]	0.4%
RSRZ outliers	[Bar chart]	3.1%

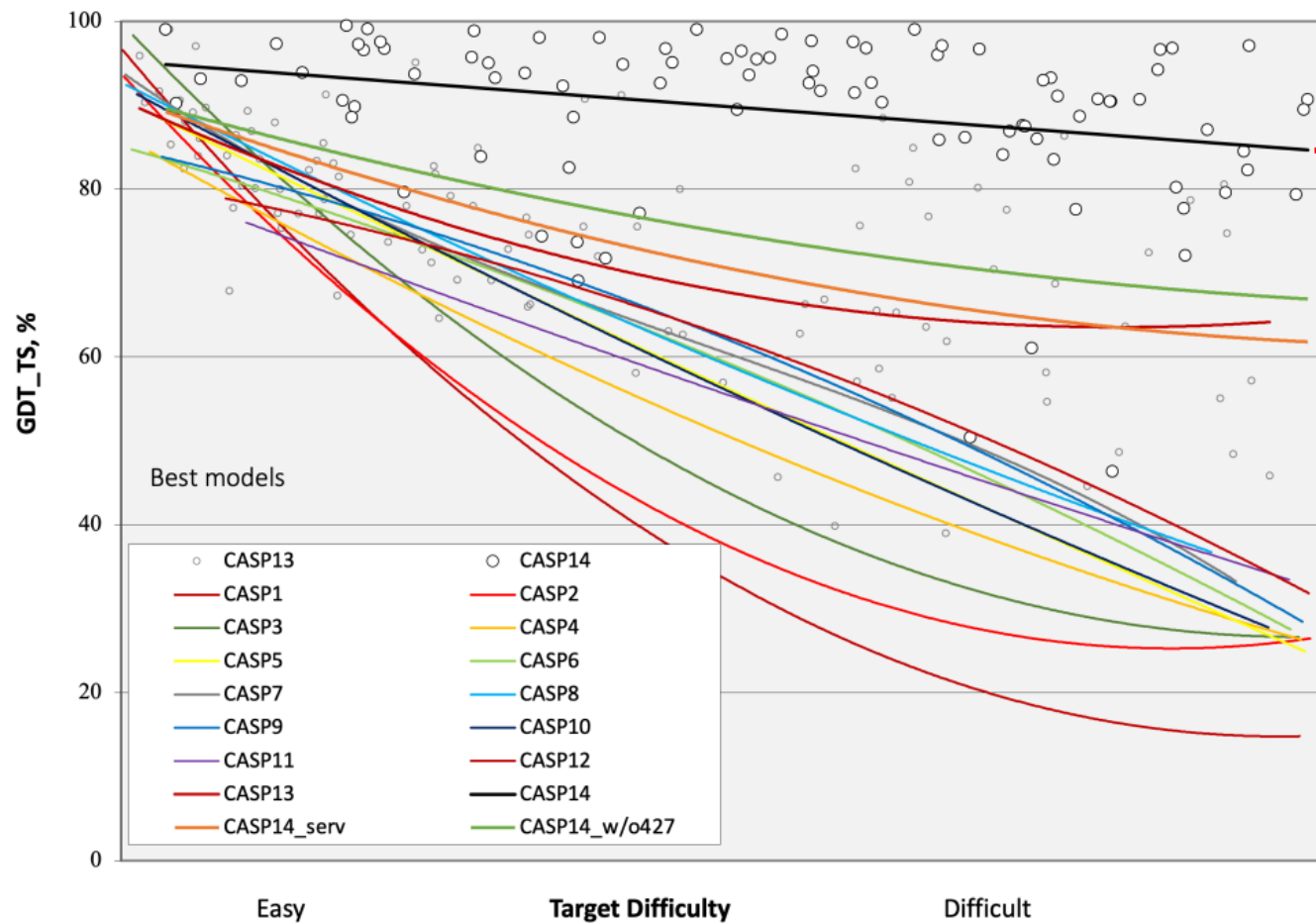
Protein Data Bank

Scabin
6vv4



Prediction of protein structures by AlphaFold

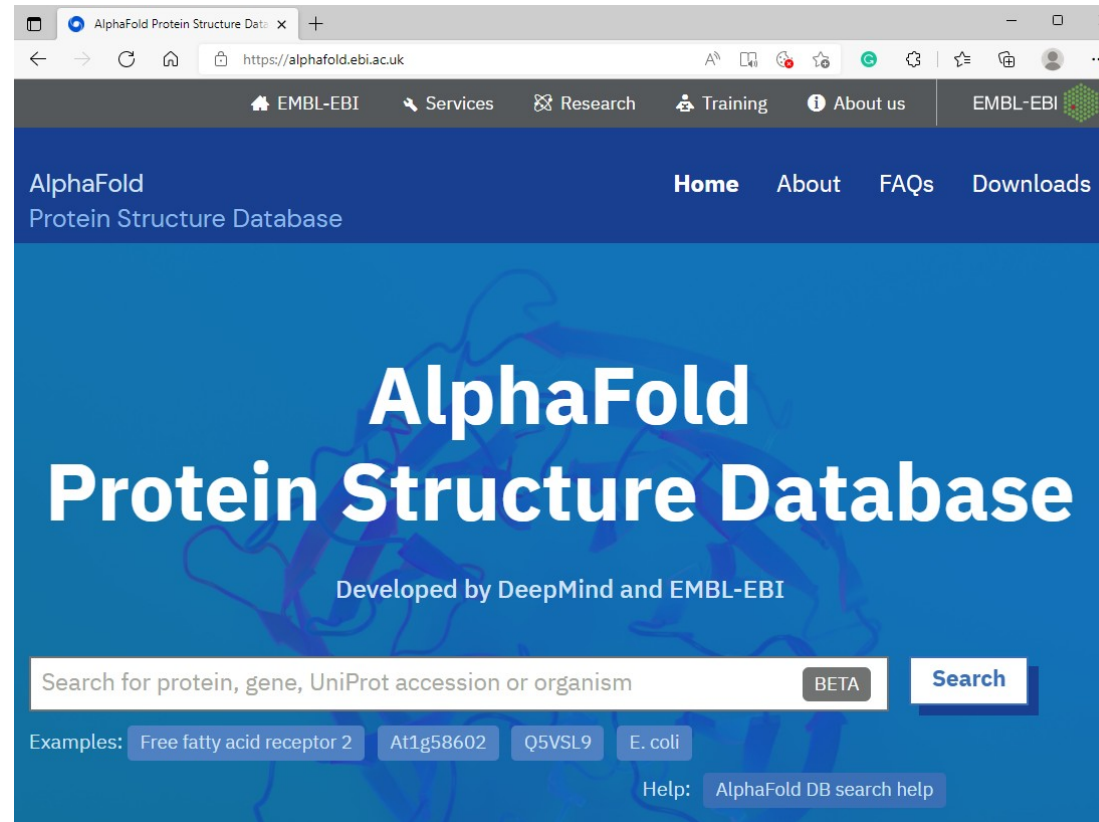
Structures generated by artificial intelligence



Structure prediction challenge 2020: AlphaFold2 wins

Prediction of protein structures by AlphaFold

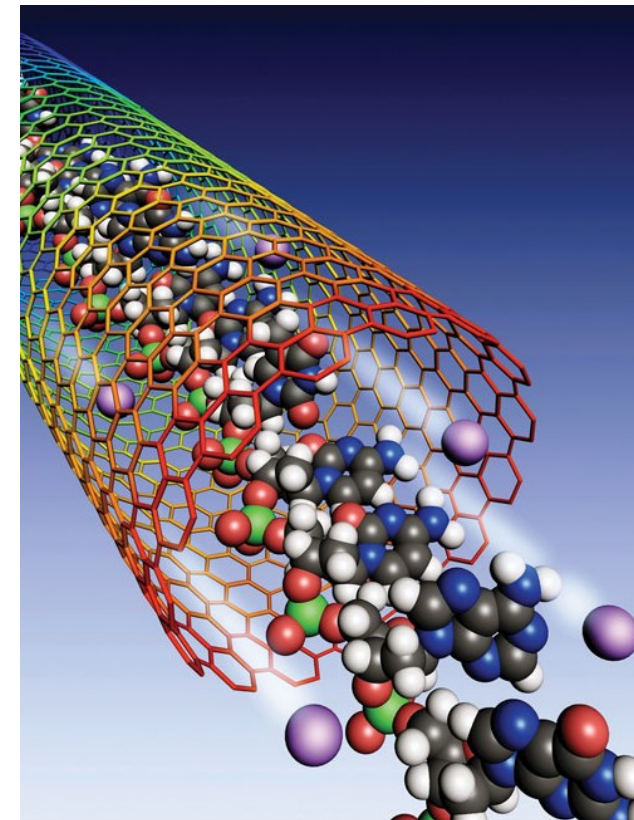
Structures generated by artificial intelligence



> 200M protein structures

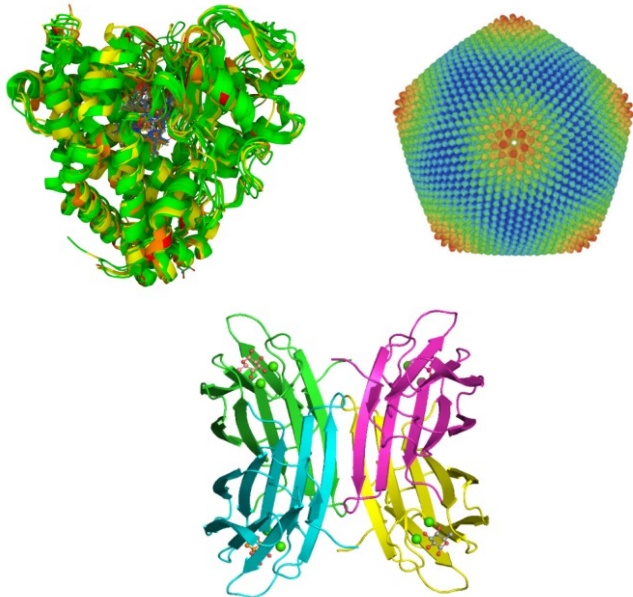
Exercises

- Search the PubChem database for a testosterone molecule:
 - See its 2D structure. How many O's does it have?
 - Look at its 3D structure. Is any of its cycles planar?
 - Look at its SDF file. What are the x, y, and z coordinates of the first atom?
- Search the DrugBank database for a penicillin molecule:
 - How many S atoms does it have?
 - Are any of its cycles planar?
 - Look at its SDF file. What 2 atoms form the first bond?
- Search the LigandExpo database for a fructose molecule:
 - How many double bonds does it has?
 - How many C's are off cycle?
 - Look at its SDF file. What are the coordinates of the first hydrogen?
- Look up the green mamba venom molecule in the Protein Data Bank:
 - How many beta-sheets does it have?
 - Look at the PDB file. Which amino acid is the first?



Thank you for your attention

Bioinformatics



Tools
Databases
AI

Chemoinformatics

