生物大數據：Protein & AlphaFold

劉 益 忠

清華大學生物資訊與結構生物研究所/國衛院群健所

博士後研究員

2022/07/19

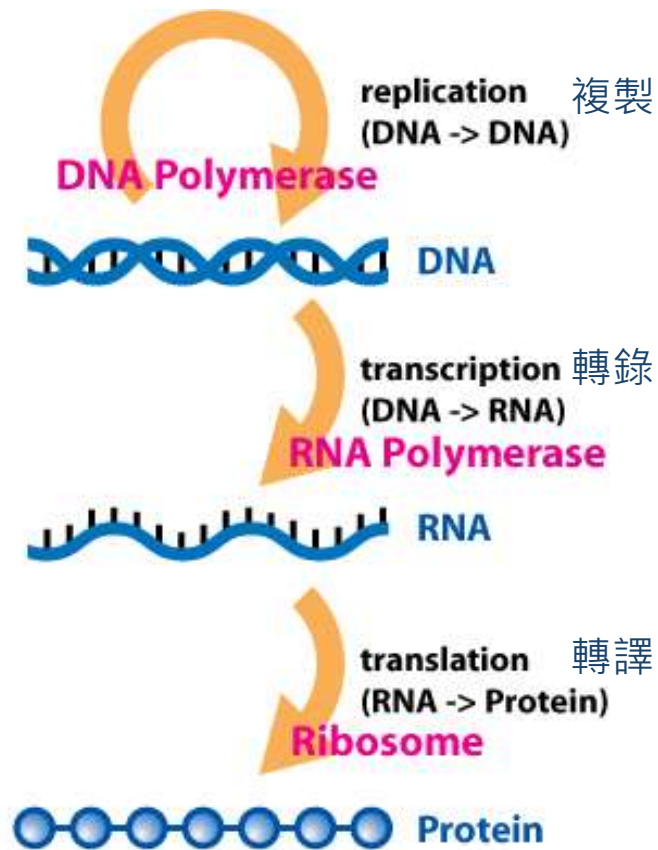# Python vs Protein

❖ 蛋白質的四級結構

❖ 蛋白質資料庫
Protein Data Bank; PDB file

❖ 蛋白質結構視覺化及PyMOL軟體

❖ 蛋白質結構模擬與預測

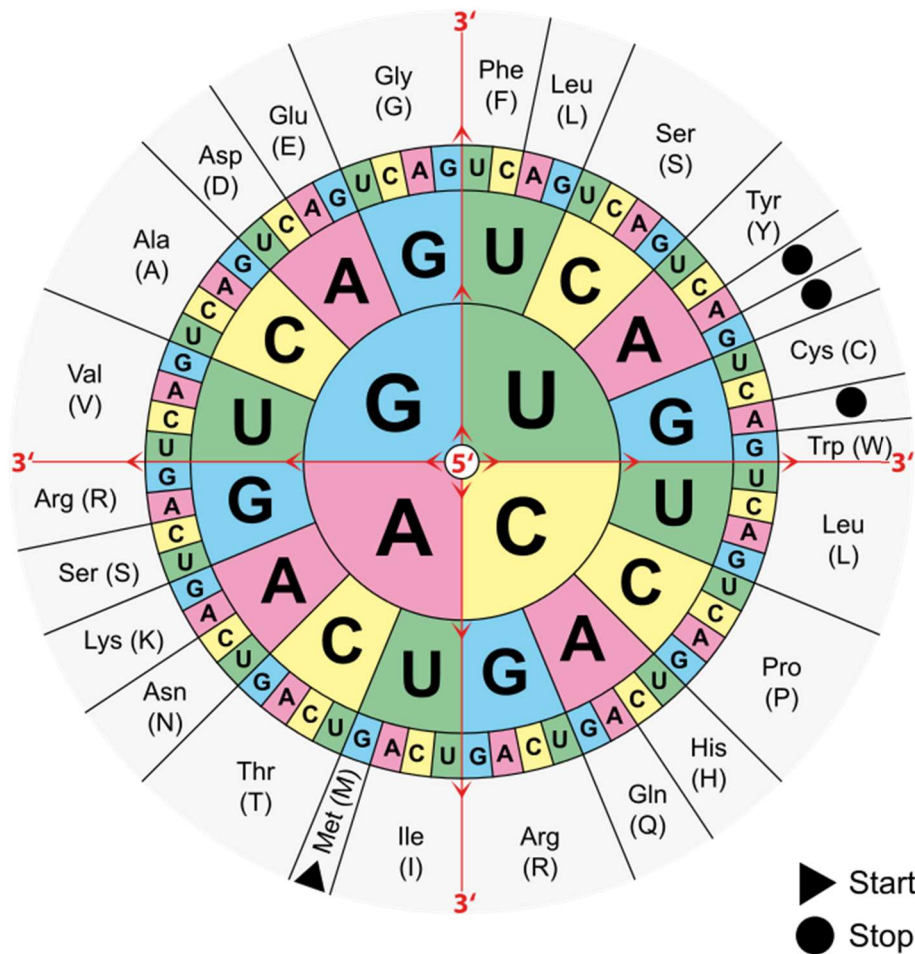❖ AlphaFold & ColabFold

蛋白質的四級結構

Start from Sequence

遺傳訊息的傳遞

replication 複製
(DNA -> DNA)
DNA Polymerase
DNA

transcription 轉錄
(DNA -> RNA)
RNA Polymerase
RNA

translation 轉譯
(RNA -> Protein)
Ribosome
Protein

中心法則 (The central dogma)

遺傳訊息傳遞

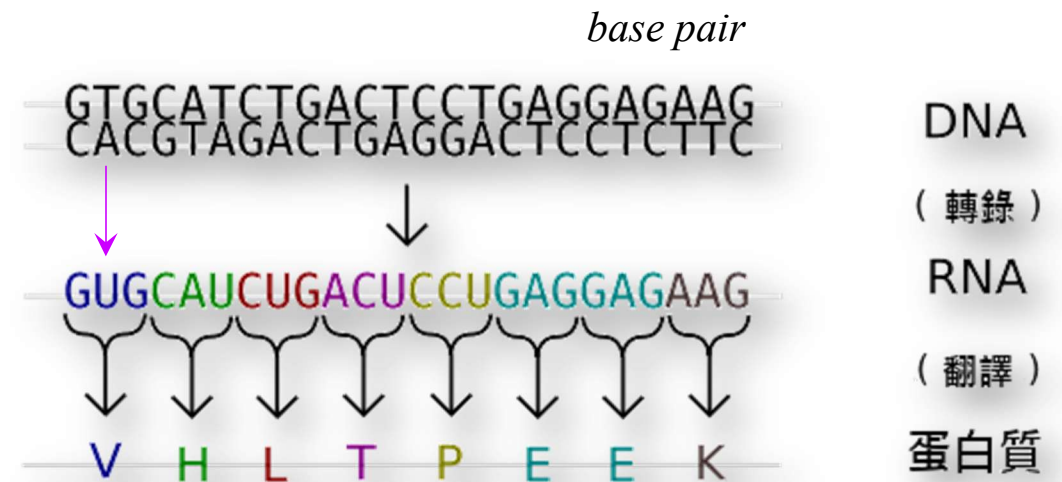*DNA → RNA → Protein*

❖ **轉錄(Transcription)**是遺傳訊息由DNA轉換到RNA的過程。

❖ **轉譯(Translation)**是根據遺傳密碼的中心法則，將成熟的mRNA分子（由DNA透過轉錄而生成）中「**鹼基的排列順序**」（核苷酸序列）解碼，並生成對應的**特定胺基酸序列**的過程。
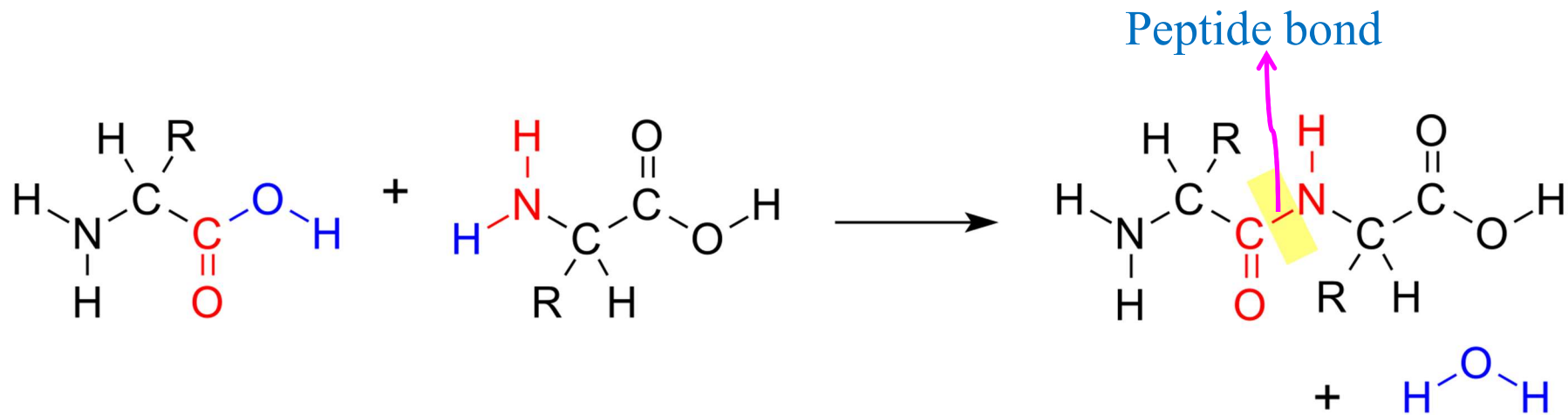
Start from Sequence

由 Mouagip - Codons aminoacids table.png, 公有領域,
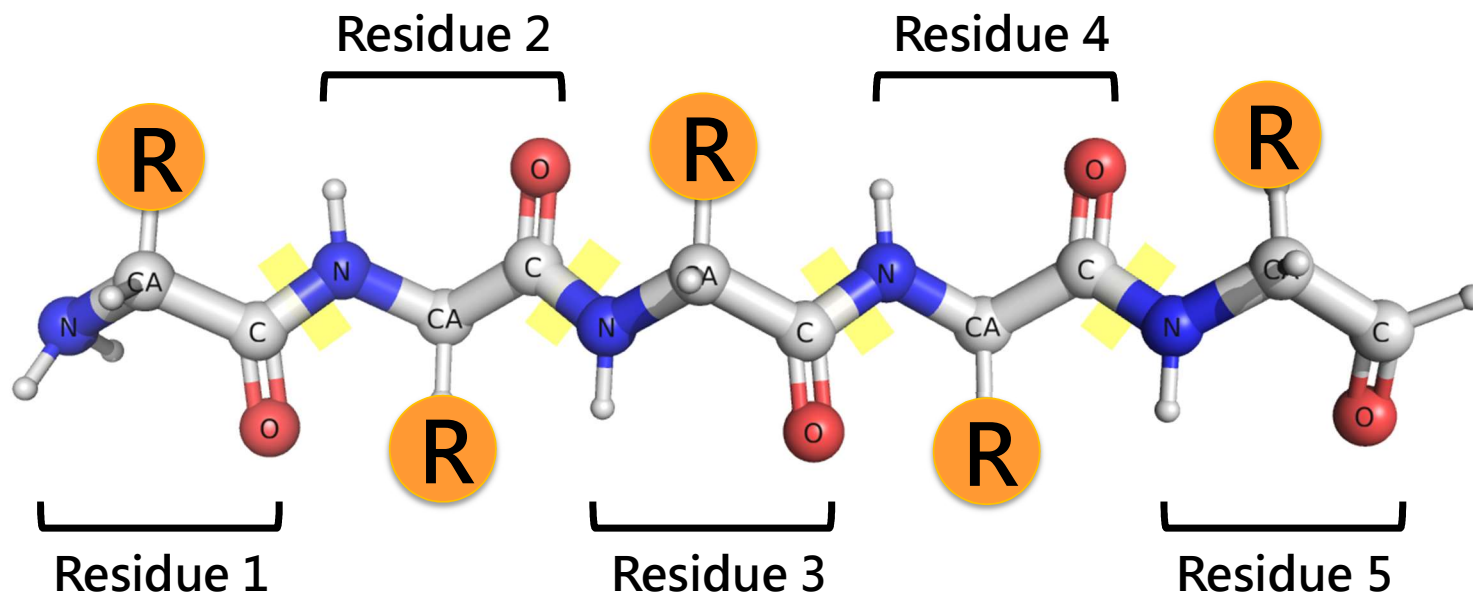https://commons.wikimedia.org/w/index.php?curid=5986132

base pair

GTGCATCTGACTCCTGAGGAGAAG
CACGTAGACTGAGGACTCCTCTTC

DNA
（轉錄）
RNA
（翻譯）
蛋白質

GUGCAUCUGACUCCUGAGGAGAAG

V H L T P E E K

(編碼一個蛋白質胺基酸序列)

Madprimederivative work, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=18365188

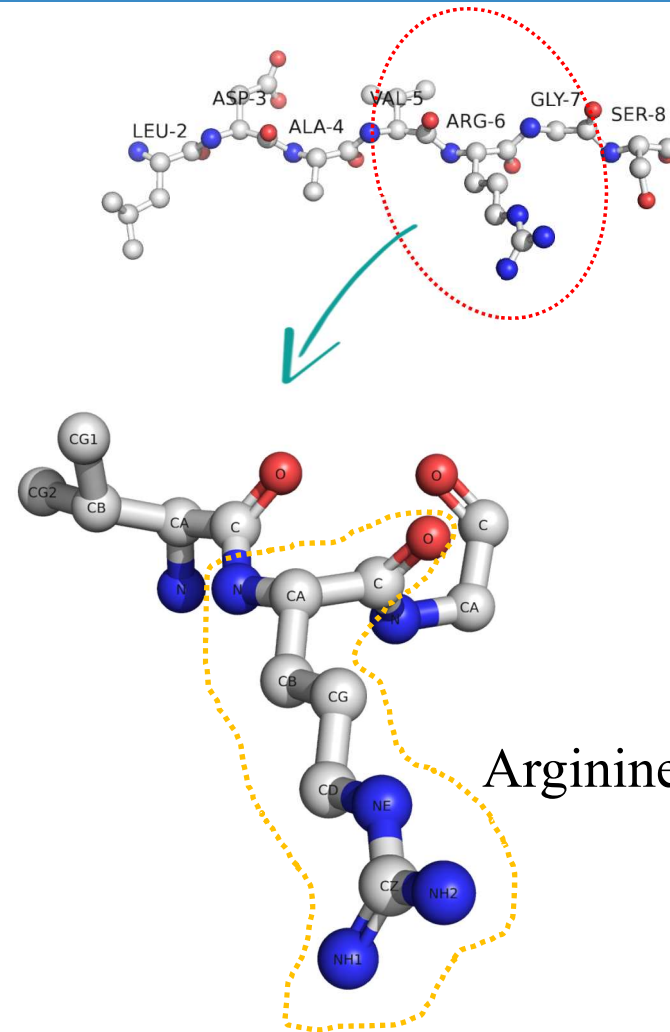# Amino acids are linked via a condensation reaction

Peptide bond

個別胺基酸以肽鍵(Peptide bond)結合

Amino acids are linked by peptide bonds

List of amino acid structures

Dan Cojocari, CC BY-SA 3.0, via Wikimedia Commons

Arginine (Arg, R)
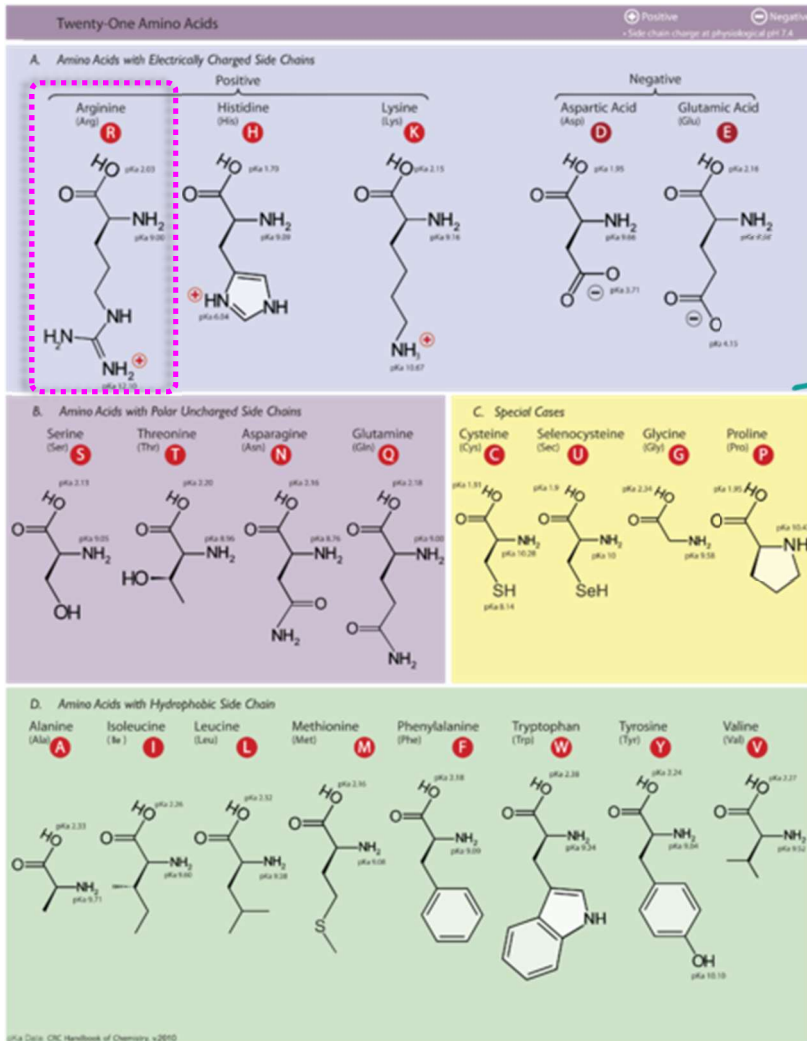
The four levels of protein structure
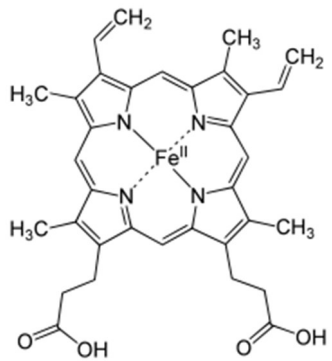

Primary structure — Amino acid — Leu – Asp – Ala – Val – Arg – Gly – Ser

LEU-2 ASP-3 ALA-4 VAL-5 ARG-6 GLY-7 SER-8

Assembly

Secondary structure — 3.6 res./turn — Hydrogen bond — β-sheet — α-helix

Folding

Tertiary structure — Polypeptide (single subunit of transthyretin)

Packing

Quaternary structure — Transthyretin, with four identical polypeptide subunits

Function-
Structure, enzymes, communication, storage, transport, defense...
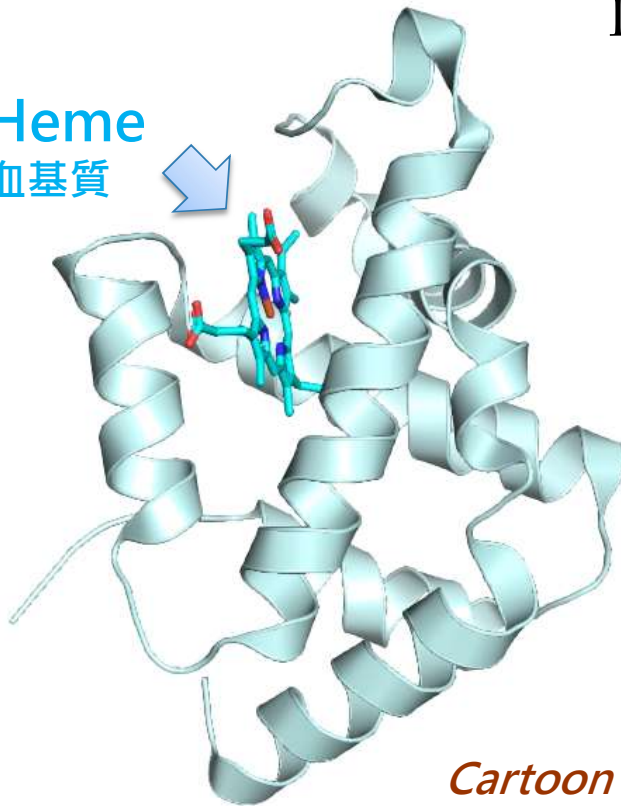
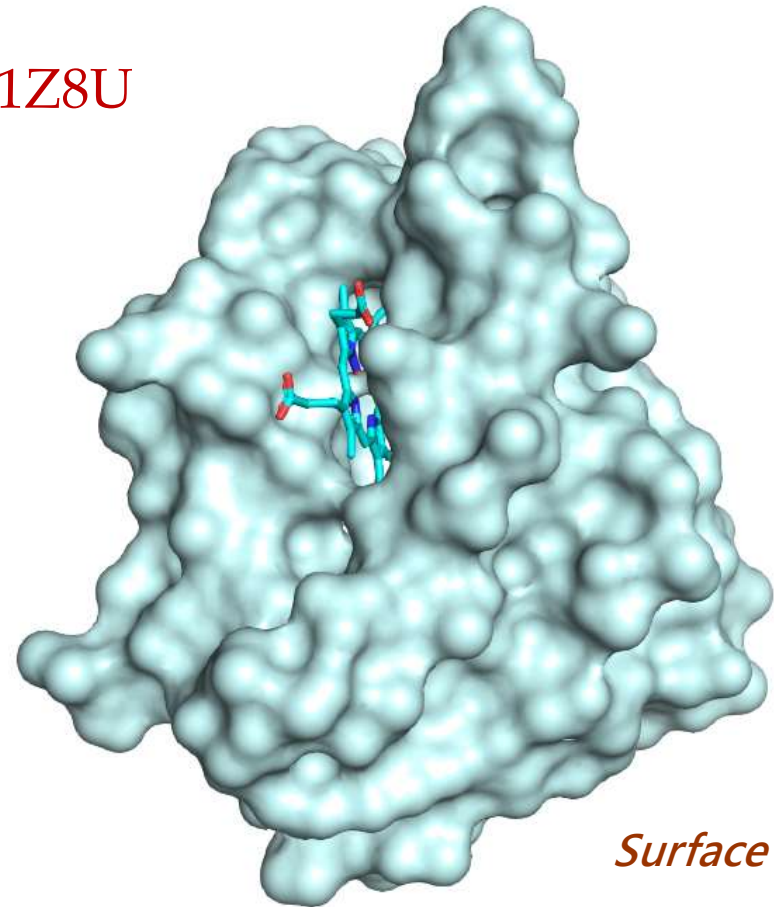*Interaction*

Hemoglobin alpha chain | *Homo sapiens*

PDB ID: 1Z8U

Heme
血基質

Cartoon 3D style

Surface 3D style

>1Z8U_2|Chains B,D|Hemoglobin (血紅素) alpha chain|Homo sapiens

About **Protein Structure**

蛋白質結構簡介

目前主要解析蛋白質結構的實驗方法

Molecular Type	X-ray	NMR	EM	Multiple methods	Neutron	Other	Total
Protein (only)	147998	11984	7772	187	72	32	168045
Protein/Oligosaccharide	8724	31	1372	5	0	0	10132
Protein/NA	7812	277	2461	3	0	0	10553
Nucleic acid (only)	2453	1416	62	12	2	1	3946
Other	154	31	5	0	0	0	190
Oligosaccharide (only)	11	6	0	1	0	4	22
Total	167152	13745	11672	208	74	37	192888

(20220716 updated)

X-ray: X-ray crystallography X-光晶體學
NMR: Nuclear magnetic resonance 核磁共振
cryo-EM: cryo-electron microscopy 冷凍電顯

https://www.rcsb.org/stats/summary

Convert a PDB file to a Structure

>6ANE_1|Chains A, B, C|Poly(ethylene terephthalate) hydrolase
MQTNPYARGPNPTAASLEASAGPFTVRSFTVSRPSGYGAGTVYYPTNAGGTVGAIAIVPGYTARQSSIK
WWGPRLASHGFVVITIDTNSTLDQPSSRSSQQMAALRQVASLNGTSSSPIYGKVDTARMGVMGWSMGGG
GSLISAANNPSLKAAAPQAPWDSSTNFSSVTVPTLIFACENDSIAPVNSSALPIYDSMSRNAKQFLEIN
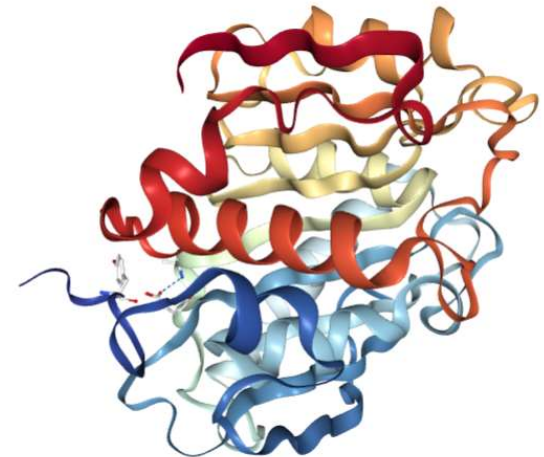GGSHSCANSGNSNQALIGKKGVAWMKRFMDNDTRYSTFACENPNSTRVSDFRTANCSLEHHHHHH

FASTA
format

Sequence
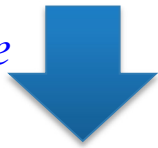
PDB structure file
記錄每個原子的3D座標

ORIGX2		0.000000	1.000000	0.000000		0.00000			
ORIGX3		0.000000	0.000000	1.000000		0.00000			
SCALE1		0.018915	0.000000	0.000000		0.00000			
SCALE2		0.000000	0.004276	0.000000		0.00000			
SCALE3		0.000000	0.000000	0.006058		0.00000			
ATOM	1	N	THR A	2	-38.844	-36.139	26.075	1.00 43.20	N
ATOM	2	CA	THR A	2	-38.028	-36.236	24.823	1.00 42.15	C
ATOM	3	C	THR A	2	-37.851	-34.874	24.086	1.00 38.29	C
ATOM	4	O	THR A	2	-38.695	-34.429	23.294	1.00 39.45	O
ATOM	5	CB	THR A	2	-38.559	-37.317	23.881	1.00 44.81	C
ATOM	6	OG1	THR A	2	-37.647	-37.461	22.783	1.00 45.87	O
ATOM	7	CG2	THR A	2	-39.990	-37.009	23.388	1.00 45.59	C
ATOM	8	N	ASN A	3	-36.718	-34.247	24.370	1.00 31.48	N
ATOM	9	CA	ASN A	3	-36.367	-32.903	23.903	1.00 26.55	C
ATOM	10	C	ASN A	3	-36.100	-32.887	22.404	1.00 24.30	C
ATOM	11	O	ASN A	3	-35.074	-33.415	21.952	1.00 24.90	O
ATOM	12	CB	ASN A	3	-35.117	-32.423	24.672	1.00 23.36	C
ATOM	13	CG	ASN A	3	-34.672	-31.001	24.298	1.00 22.22	C
ATOM	14	OD1	ASN A	3	-35.184	-30.378	23.371	1.00 19.60	O
ATOM	15	ND2	ASN A	3	-33.706	-30.494	25.041	1.00 22.68	N
ATOM	16	N	PRO A	4	-36.950	-32.197	21.628	1.00 23.95	N
ATOM	17	CA	PRO A	4	-36.754	-32.262	20.171	1.00 23.73	C
ATOM	18	C	PRO A	4	-35.513	-31.536	19.636	1.00 23.47	C
ATOM	19	O	PRO A	4	-35.170	-31.710	18.468	1.00 22.30	O
ATOM	20	CB	PRO A	4	-38.030	-31.646	19.613	1.00 24.93	C
ATOM	21	CG	PRO A	4	-38.510	-30.712	20.682	1.00 26.00	C
ATOM	22	CD	PRO A	4	-38.061	-31.301	21.999	1.00 25.07	C
ATOM	23	N	TYR A	5	-34.862	-30.723	20.469	1.00 23.26	N

**軟體轉換成
立體結構**

Visualized by
Molecular Graphical Tools

Structure Summary : 6ANE

*6ANE: Crystal Structure of Ideonella sakaiensis(大阪堺菌) **PET Hydrolase**(聚對苯二甲酸乙二酯-水解酶)*

6ANE.pdb

PDB ID

Display Files

https://www.rcsb.org/structure/6ane

PDB檔案內容

Display Files ▾

close

FASTA Sequence

PDB File

PDB File (Header)

mmCIF File

mmCIF File (Header)

PDBML/XML File

PDBML/XML File (Header)

線上顯示PDB檔案內容,
前面部分是對結構的註解

```
HEADER    HYDROLASE                               12-AUG-17   6ANE
TITLE     CRYSTAL STRUCTURE OF IDEONELLA SAKAIENSIS PET HYDROLASE
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: POLY(ETHYLENE TEREPHTHALATE) HYDROLASE;
COMPND   3 CHAIN: A, B, C;
COMPND   4 FRAGMENT: RESIDUES 28-290;
COMPND   5 SYNONYM: PETASE;
COMPND   6 EC: 3.1.1.101;
COMPND   7 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: IDEONELLA SAKAIENSIS (STRAIN 201-F6);
SOURCE   3 ORGANISM_TAXID: 1547922;
SOURCE   4 STRAIN: 201-F6;
SOURCE   5 GENE: ISF6_4831;
SOURCE   6 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   7 EXPRESSION_SYSTEM_TAXID: 511693;
SOURCE   8 EXPRESSION_SYSTEM_STRAIN: BL21;
SOURCE   9 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE  10 EXPRESSION_SYSTEM_PLASMID: PET24B
KEYWDS    PETASE, AB HYDROLASE, CUTINASE, PLASTIC, HYDROLASE
EXPDTA    X-RAY DIFFRACTION
AUTHOR    P.GALAZ-DAVISON,M.SOTOMAYOR,L.P.PARRA,C.A.RAMIREZ-SARMIENTO
REVDAT   2   25-APR-18 6ANE    1       REMARK
REVDAT   1   18-APR-18 6ANE    0
JRNL        AUTH   T.FECKER,P.GALAZ-DAVISON,F.ENGELBERGER,Y.NARUI,M.SOTOMAYOR,
JRNL        AUTH 2 L.P.PARRA,C.A.RAMIREZ-SARMIENTO
JRNL        TITL   ACTIVE SITE FLEXIBILITY AS A HALLMARK FOR EFFICIENT PET
JRNL        TITL 2 DEGRADATION BY I. SAKAIENSIS PETASE.
JRNL        REF    BIOPHYS. J.                  V. 114  1302 2018
```

http://www.wwpdb.org/documentation/file-format-content/format33/sect2.html

PDB檔案內容- ATOM

Text-based format with 3D information

```
CRYST1   52.868  233.852  165.076  90.00    90.00    90.00 C 2 2 21      24
ORIGX1       1.000000  0.000000  0.000000        0.00000
ORIGX2       0.000000  1.000000  0.000000        0.00000
ORIGX3       0.000000  0.000000  1.000000        0.00000
SCALE1       0.018915  0.000000  0.000000        0.00000
SCALE2       0.000000  0.004276  0.000000        0.00000
SCALE3       0.000000  0.000000  0.006058        0.00000
ATOM      1  N   THR A   2     -38.844 -36.139  26.075  1.00 43.20           N
ATOM      2  CA  THR A   2     -38.028 -36.236  24.823  1.00 42.15           C
ATOM      3  C   THR A   2     -37.851 -34.874  24.086  1.00 38.29           C
ATOM      4  O   THR A   2     -38.695 -34.429  23.294  1.00 39.45           O
ATOM      5  CB  THR A   2     -38.559 -37.317  23.881  1.00 44.81           C
ATOM      6  OG1 THR A   2     -37.647 -37.461  22.783  1.00 45.87           O
ATOM      7  CG2 THR A   2     -39.990 -37.009  23.388  1.00 45.59           C
ATOM      8  N   ASN A   3     -36.718 -34.247  24.370  1.00 31.48           N
ATOM      9  CA  ASN A   3     -36.367 -32.903  23.903  1.00 26.55           C
ATOM     10  C   ASN A   3     -36.100 -32.887  22.404  1.00 24.30           C
ATOM     11  O   ASN A   3     -35.074 -33.415  21.952  1.00 24.90           O
ATOM     12  CB  ASN A   3     -35.117 -32.423  24.672  1.00 23.36           C
ATOM     13  CG  ASN A   3     -34.672 -31.001  24.298  1.00 22.22           C
ATOM     14  OD1 ASN A   3     -35.184 -30.378  23.371  1.00 19.60           O
ATOM     15  ND2 ASN A   3     -33.706 -30.494  25.041  1.00 22.68           N
ATOM     16  N   PRO A   4     -36.950 -32.197  21.628  1.00 23.95           N
ATOM     17  CA  PRO A   4     -36.754 -32.262  20.171  1.00 23.73           C
ATOM     18  C   PRO A   4     -35.513 -31.536  19.636  1.00 23.47           C
ATOM     19  O   PRO A   4     -35.170 -31.710  18.468  1.00 22.30           O
ATOM     20  CB  PRO A   4     -38.030 -31.646  19.613  1.00 24.93           C
ATOM     21  CG  PRO A   4     -38.510 -30.712  20.682  1.00 26.00           C
ATOM     22  CD  PRO A   4     -38.061 -31.301  21.999  1.00 25.07           C
ATOM     23  N   TYR A   5     -34.862 -30.723  20.469  1.00 23.26           N
```

Display Files ▼ — FASTA Sequence, PDB File, PDB File (Header), mmCIF File, mmCIF File (Header), PDBML/XML File, PDBML/XML File (Header)

x y z

PDB檔案內容- HETATM

結構檔案中的金屬離子或化合物會以**HETATM**的形式註解。像本例中的MG

ATOM- Record Format

```
ATOM    5809  CB   CYS C 262      12.178 -10.029 -19.765  1.00 32.71        C
ATOM    5810  SG   CYS C 262      10.381 -10.019 -19.990  1.00 33.77        S
TER     5811       CYS C 262
HETATM  5812 MG     MG C 301      -0.050  -7.496 -40.983  0.50 23.98       MG
HETATM  5813  O    HOH A 301     -18.774  -8.400  42.837  1.00 28.05        O
```

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"ATOM "	
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	AChar	iCode	Code for insertion of residues.
31 - 38	Real(8.3)	x	Orthogonal coordinates for X in Angstroms.
39 - 46	Real(8.3)	y	Orthogonal coordinates for Y in Angstroms.
47 - 54	Real(8.3)	z	Orthogonal coordinates for Z in Angstroms.
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	tempFactor	Temperature factor.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

http://www.wwpdb.org/documentation/file-format-content/format33/sect9.html#ATOM

Check protein data file from CoLab

Upload

01_Protein Data Bank.ipynb

To

Colab Notebooks

Protein Structure Visualization

蛋白質結構視覺化

工欲善其事　必先利其器

Visualization of Protein Structure

Protein sequence

```
>3TE4:A|PDBID|CHAIN|SEQUENCE
GPLGSPYTIELIQPEDGEAVIAMLKTFFFKDE      EKYSLKPLPD
PSPDDVPEKAADSCEHPKFKKILSLMDHVEEQ      DGKILSVDTN
VYHVLCSSHYSARVMEKLGFHEVFRMQFADYK      GIQVMAKEV
```
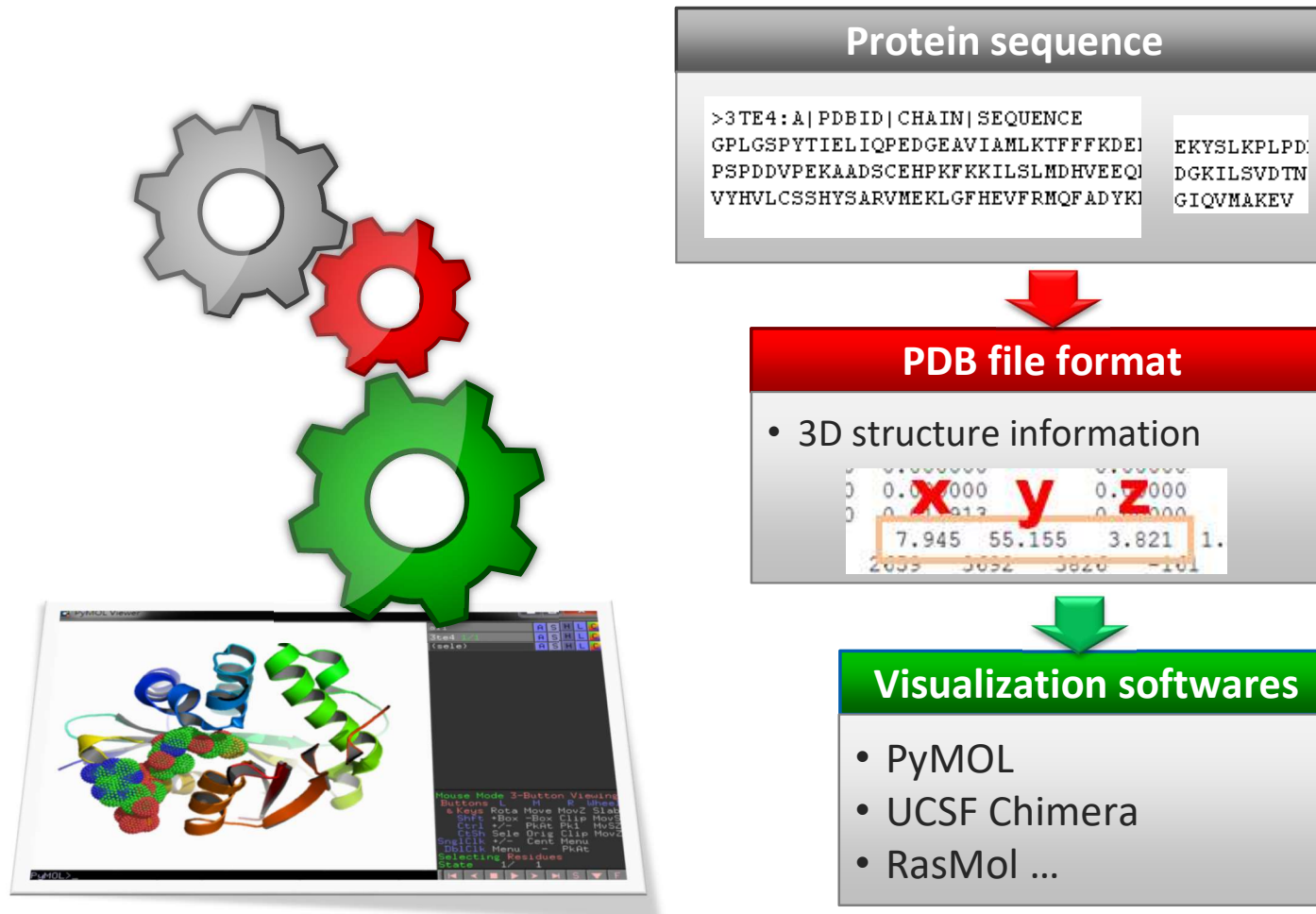
From Sequence to Structure

PDB file format

- 3D structure information

```
  0.0 000     0.0 000
  0.0  000     0.z 000
7.945  55.155  3.821  1.
```

每個原子的3D座標

Visualization softwares

- PyMOL
- UCSF Chimera
- RasMol ...

轉換成立體結構

*結構視覺化*軟體工具

Download PyMOL

https://pymol.org/2/

PyMOL-2.5.2_293-Win64-portable-py37.zip

*在zip檔案上按滑鼠右鍵，選擇
解壓縮全部(T)…

開啟(O)
在新視窗開啟(E)
PC Edit with PyCharm
解壓縮全部(T)…
7-Zip   >
CRC SHA   >

✕

📁 解壓縮壓縮 (Zipped) 資料夾

選取目的地並解壓縮檔案

檔案會解壓縮至這個資料夾(F):

E:\PyMOL-2.5.2_293-Win64-portable-py37   瀏覽(R)...

☐ 完成時顯示解壓縮的檔案(H)

由於PyMOL讀取中文資料夾會有問題, 請不要直接解壓縮到"桌面"上,建議放在其它的硬碟 D or E槽, 或是英文資料夾下面

🖥 > 本機 > 桌面 ❌

解壓縮(E)   取消
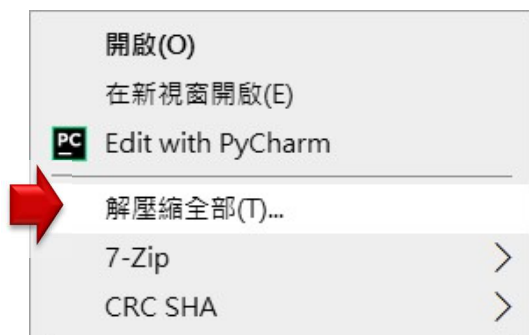
安裝 PyMOL

On Windows

- **ZIP Archive**: **Unzip and run PyMOLWin.exe**
  (**ZIP解壓縮後，直接執行PyMOLWin.exe**)

On a Mac

1. Double-click on *.dmg file
2. Drag PyMOL icon to Applications folder

On Linux

1. Unpack the .tar.bz2 file, e.g.: tar -jxf PyMOL-2.2.0_0-Linux-x86_64.tar.bz2
2. Run PyMOL with: cd pymol; ./pymol

https://pymol.org/2/support.html?#installation
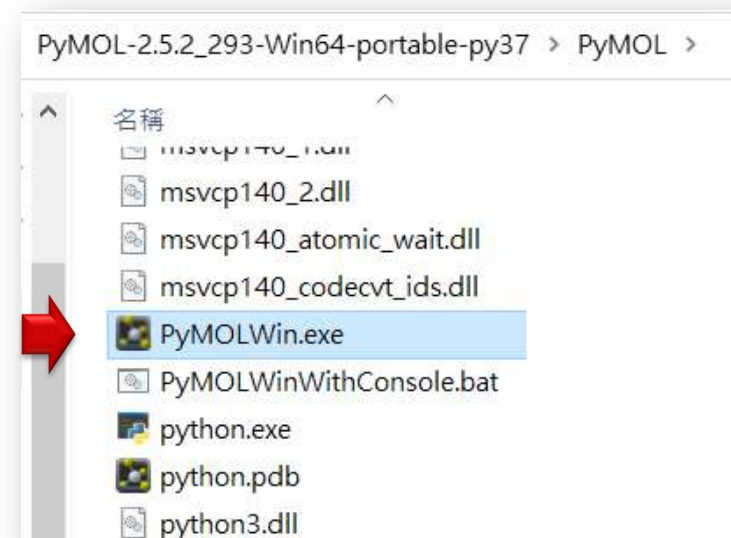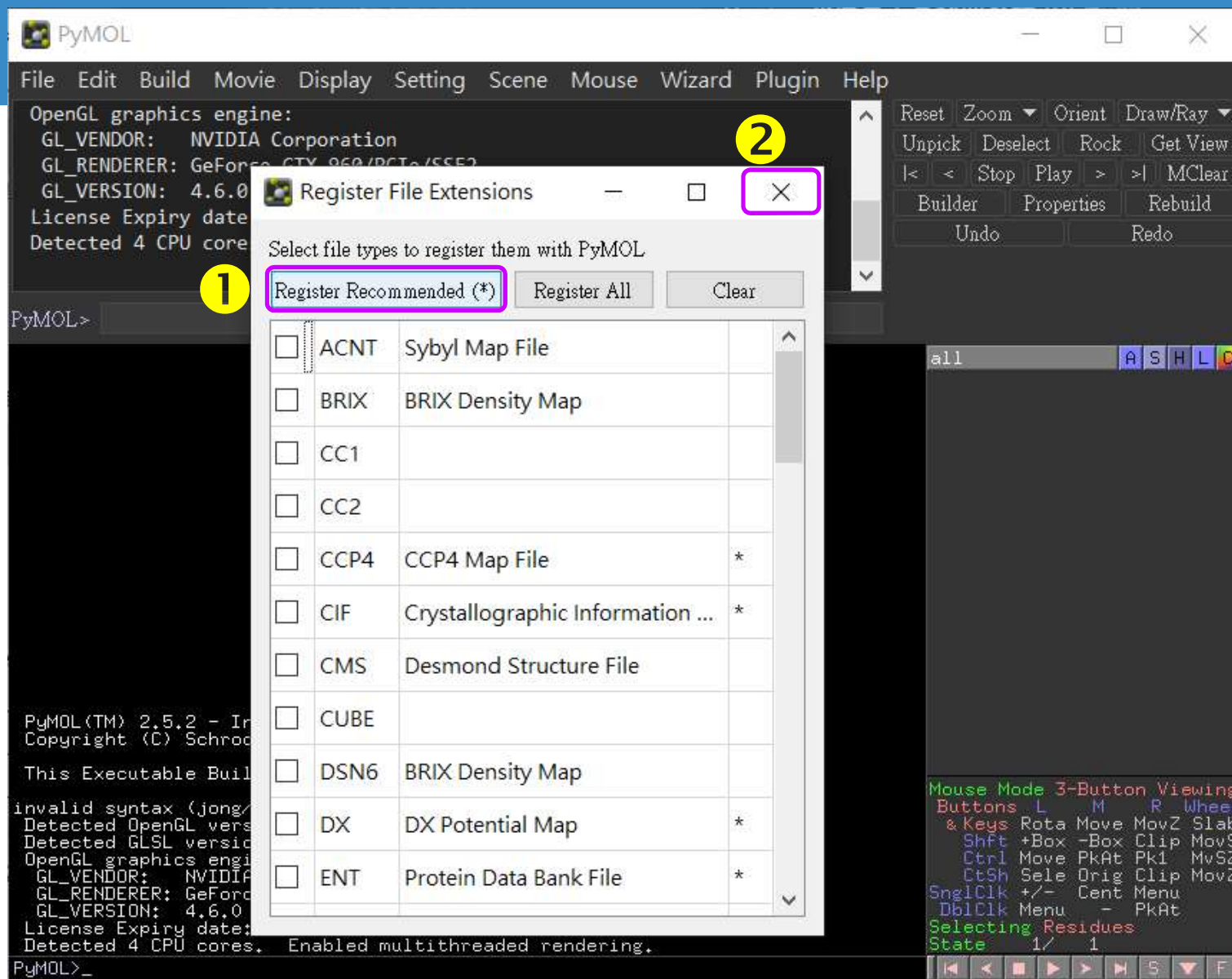
PyMOL 軟體操作介面
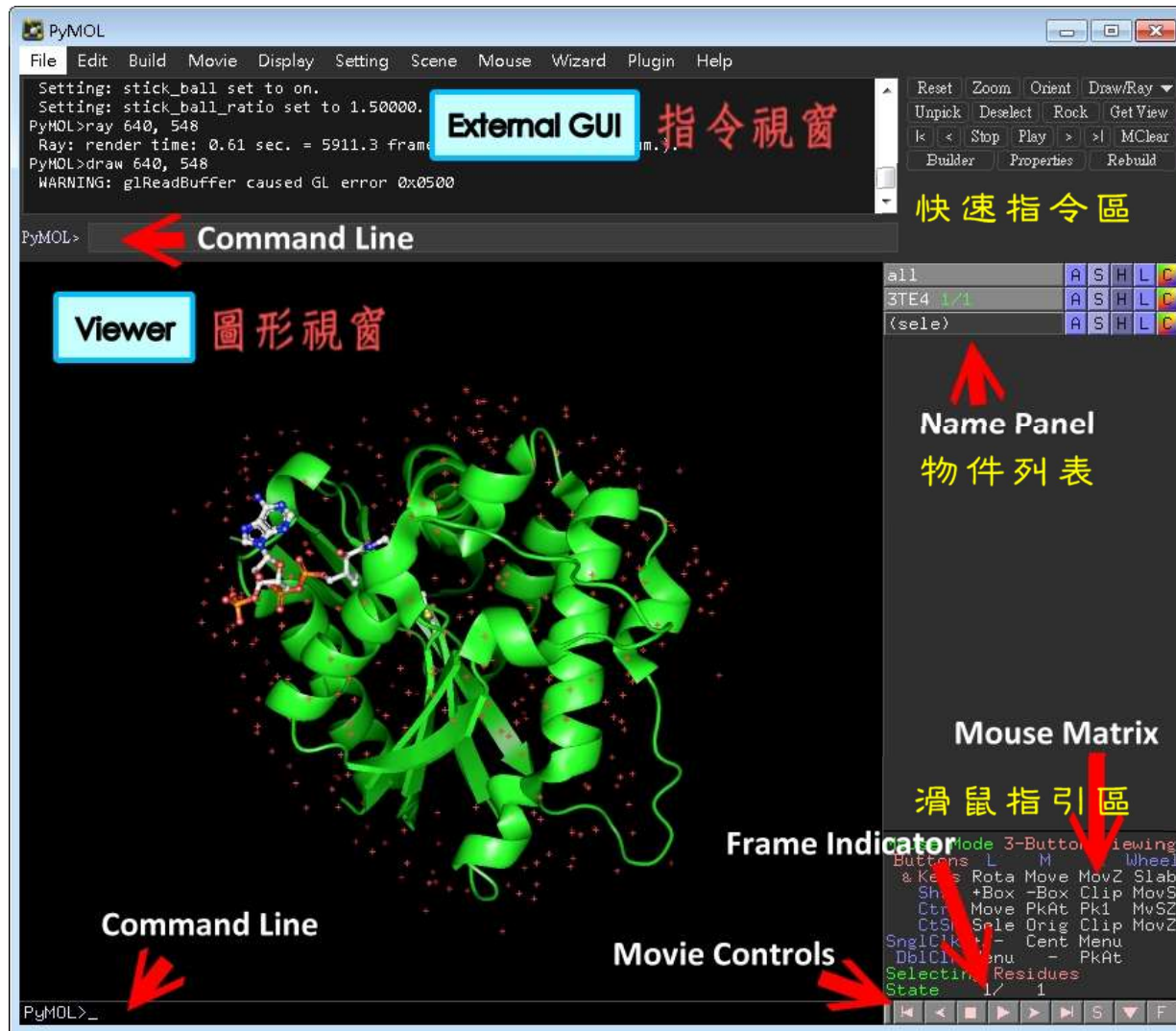
*控制個別物件按鈕

A: action
S: show
H: hide
L: label
C: color
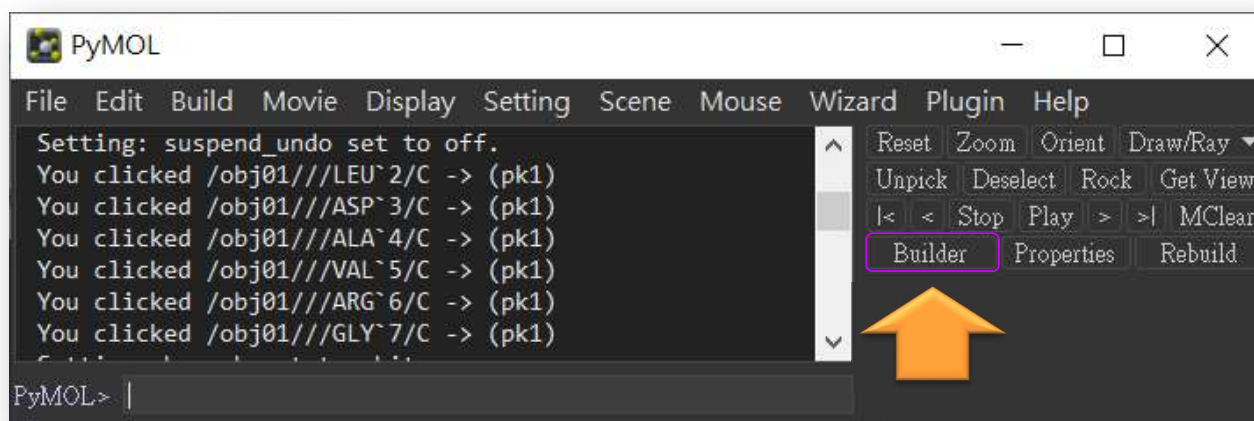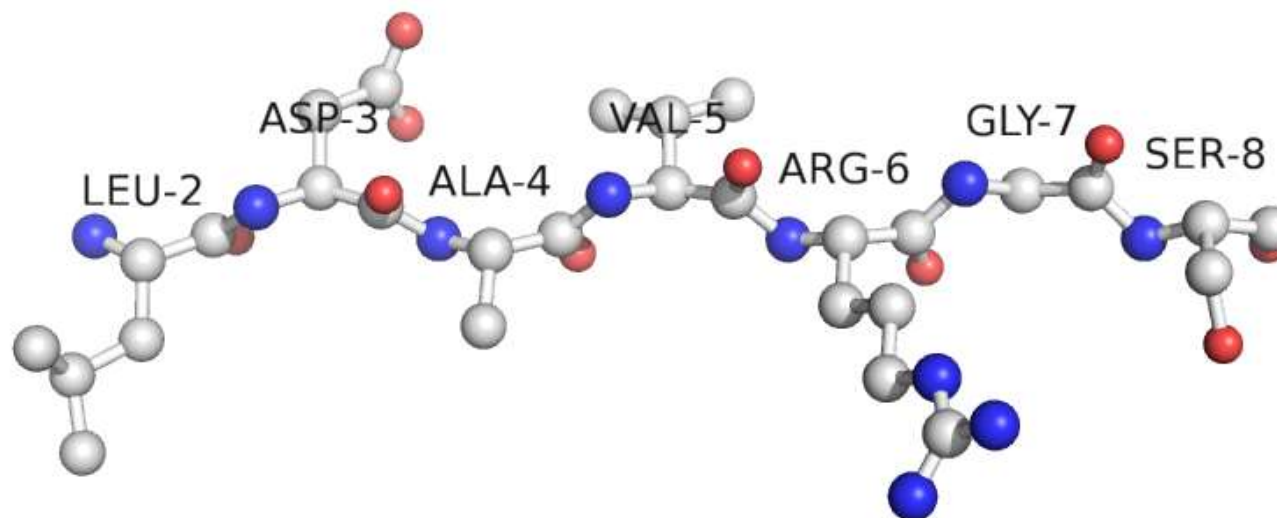
開啟 **Builder,** 切換 **Protein**
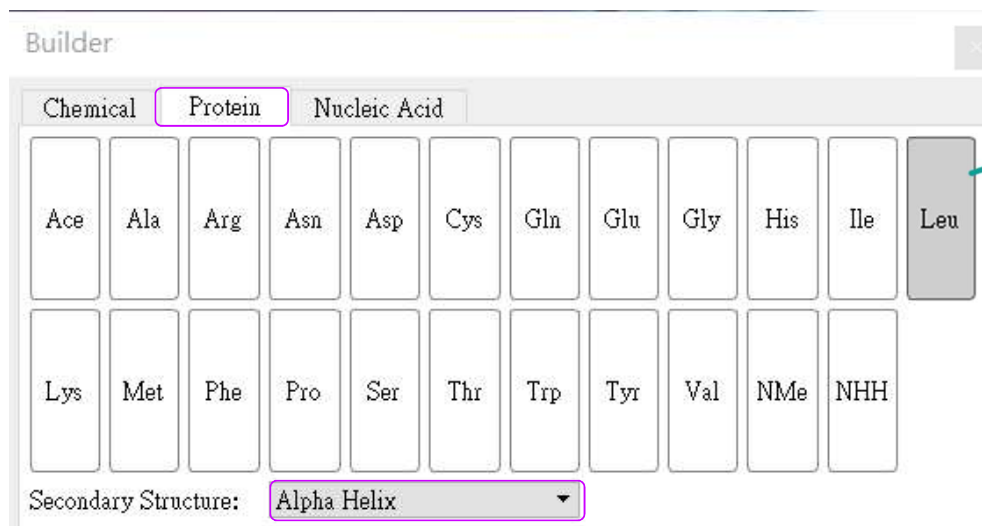
選擇**Alpha Helix二級結構**
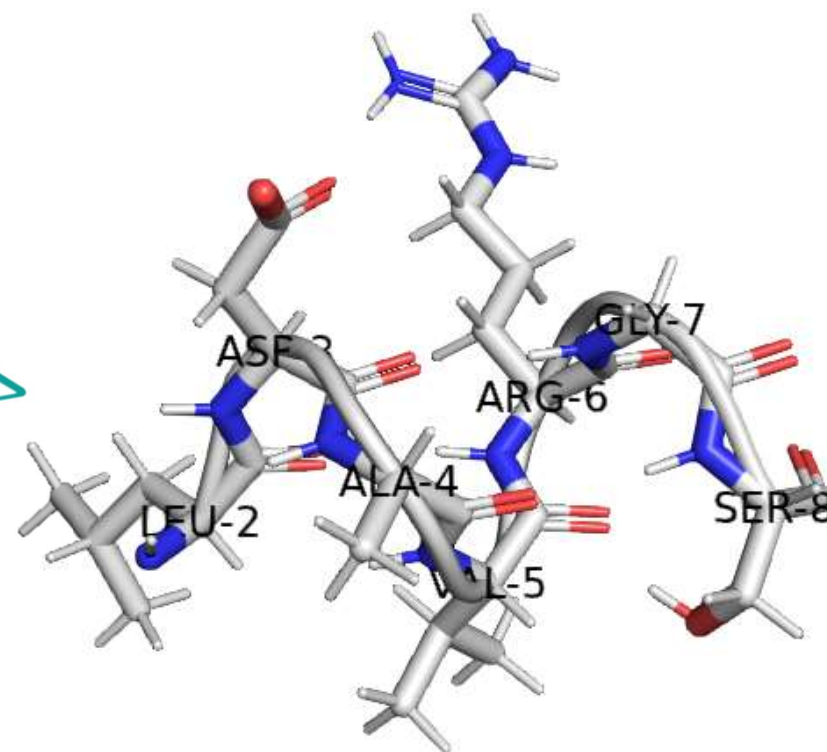
*可檢視backbone, side chain

蛋白質的3D立體結構
可以提供我們什麼訊息?

what's the **NEXT**?

For example,

ACE2

Angiotensin-Converting Enzyme 2 (ACE2)

血管緊張素轉換酶

ACE2: The Receptor for SARS-CoV-2
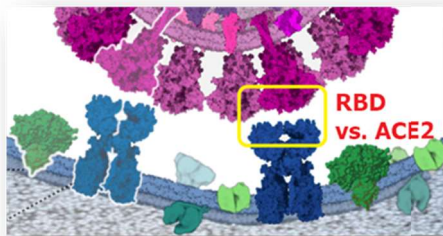
SARS-CoV-2與靶細胞的表面蛋白相互作用
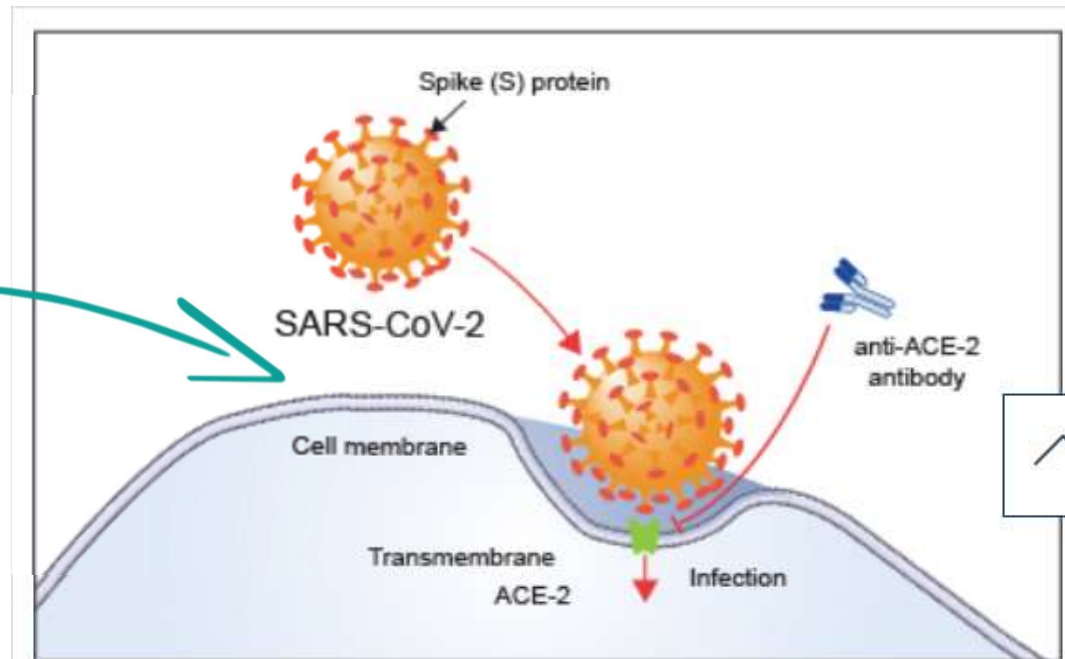
Figure 1. ACE-2 is the host cell receptor responsible for mediating infection by SARS-CoV-2, the novel coronavirus responsible for coronavirus disease 2019 (COVID-19). Treatment with anti-ACE-2 antibodies disrupts the interaction between virus and receptor.

https://www.rndsystems.com/resources/articles/ace-2-sars-receptor-identified

Structure Alignment

RBD
223 aa
ACE2

814 aa

6M17 (China)

(2020) Science **367**: 1444-1448

RBD

Identities
186/217(86%)

ACE2

Identities
597/597(100%)

RBD
217 aa
RBD
ACE2
597 aa

6VW1 (USA)

RMSD = 1.290
(root-mean-square deviation)

(2020) Nature **581**: 221-224

RBD
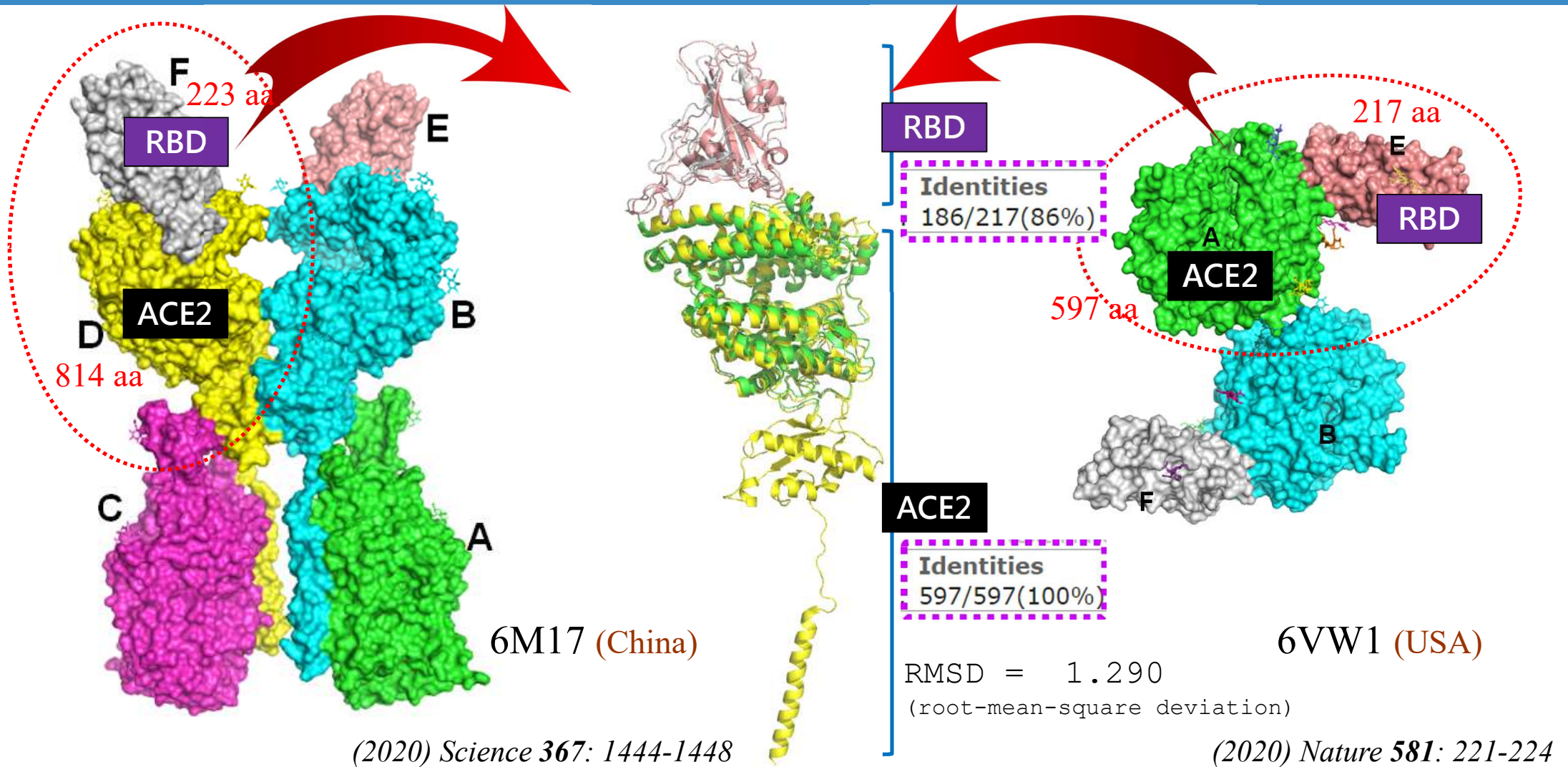
ACE2

The
coronavirus recognition sites?

識別位點
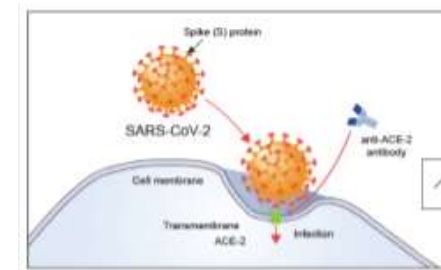
顯示氫鍵結合的位置
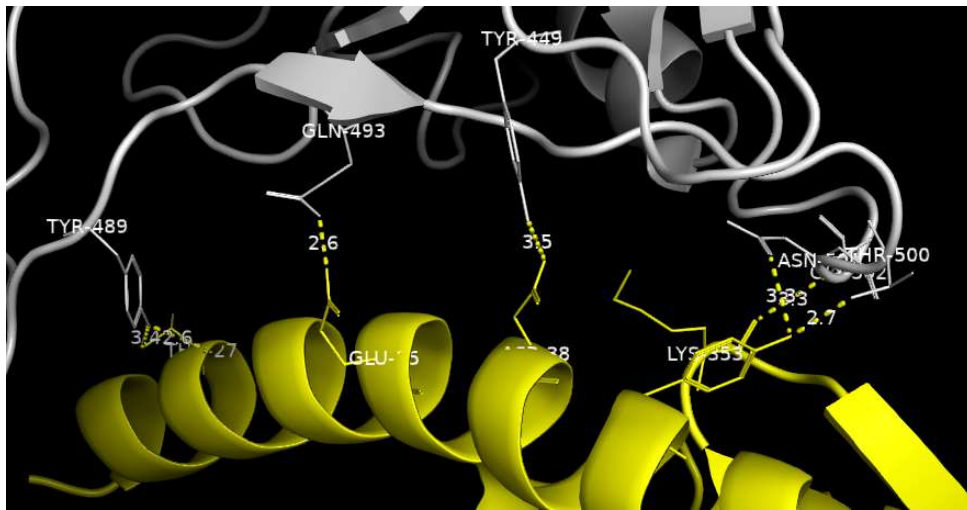
6M17 (China)

RBD	Tyr489, Gln493, Tyr449, Thr500, Asn501*, Gly502
ACE2	Thr27, Glu35, Asp38, Tyr41, Lys353

*(2020) Science **367**: 1444-1448*
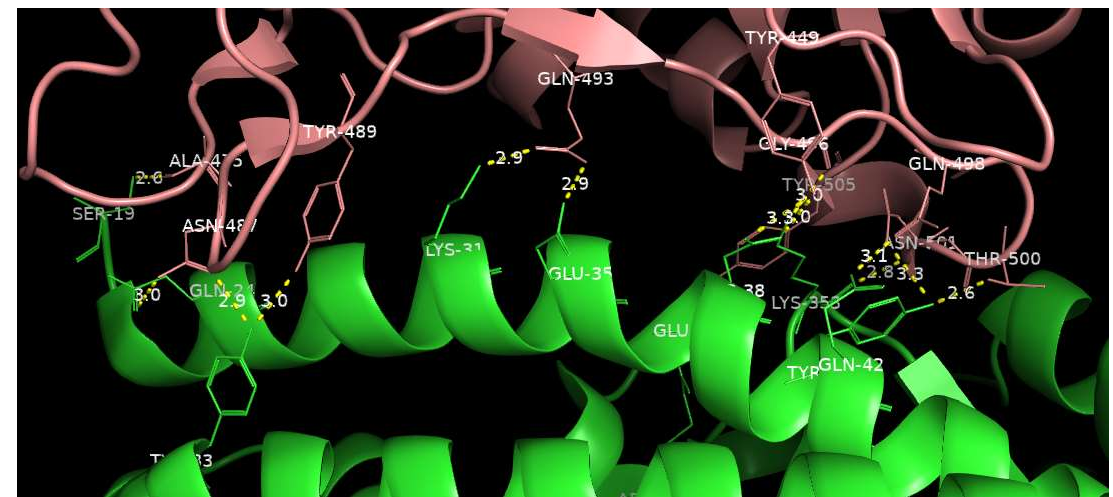
6VW1 (USA)

RBD	Ala475, Asn487, Tyr489, Gln493, Gly496, Tyr449, Gln498, Thr500, Asn501*, Try505
ACE2	Ser19, Gln24, Tyr83, Lys31, Glu35, Glu37, Asp38, Tyr41, Gln42, Lys353, Arg393

*(2020) Nature **581**: 221-224*

PyMOL 操作環境介紹

PyMOL 教學網頁
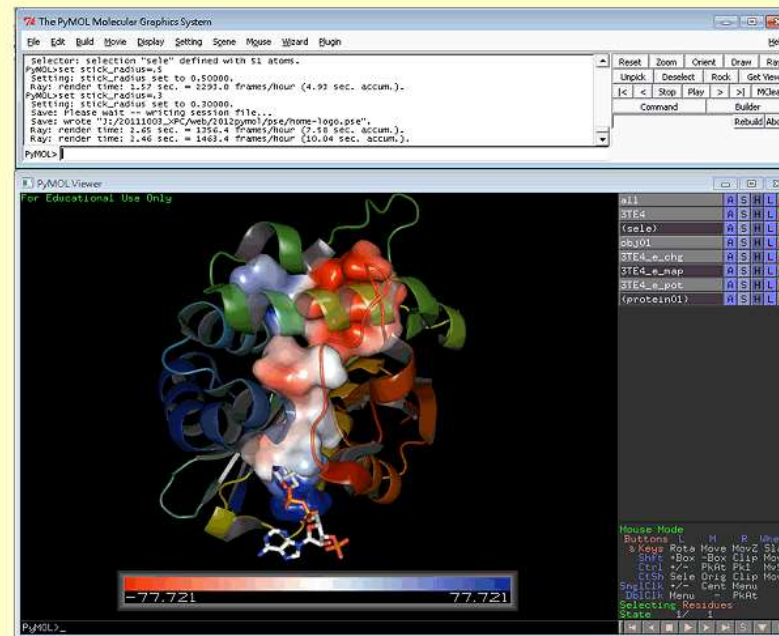
A Tutorial for the PyMOL Basics

http://140.114.98.75/sbio/pymol/

PyMOL Tutorial List

List

PyMOL Experience

Browse

PyMOL
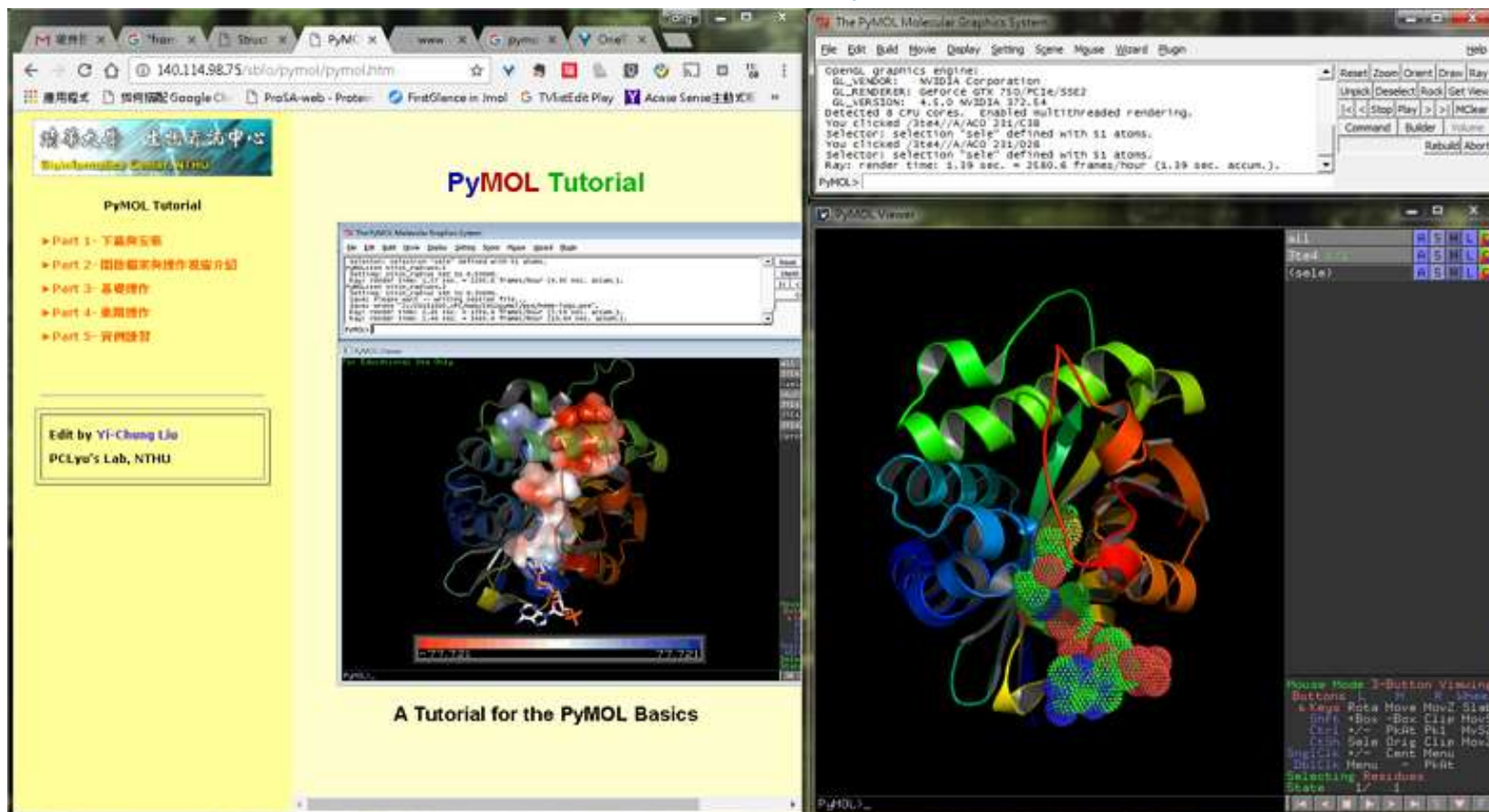
*建議可以將教學網頁放左邊參考, PyMOL程式放右邊操作

Try PyMOL Demo

Select 12 Demo Show

PyMOL Demo:
Wizard > Demo > …

Click **End** *to end*

輝瑞-新冠抗病毒口服藥 Paxlovid

衛 生 福 利 部
Ministry of Health and Welfare
促 進 全 民 健 康 與 福 祉

請輸入關鍵字 🔍 進階

熱門關鍵字： COVID-19 防疫補償 疫苗 長期照顧

本部簡介▾ 最新消息▾ 便民服務▾ 法令規章▾ 衛教視窗▾ 重大事件▾ 本部各單位及所屬機關▾

首頁 / 最新消息 / 焦點新聞

衛福部食藥署同意核准輝瑞公司COVID-19口服抗病毒藥品 Paxlovid 專案輸入

● 資料來源：食品藥物管理署　● 建檔日期：111-01-15　● 更新時間：111-01-17

Paxlovid 建議使用對象

我國診治指引
針對「Paxlovid」建議使用之對象

◆具以下任一**重症風險因子，未使用氧氣**且於**發病五天內**之 ≥ **12歲且體重 ≥ 40公斤**病患。

◆重症風險因子：年齡 ≥ 65歲、糖尿病、慢性腎病、心血管疾病(含高血壓)、慢性肺疾、BMI ≥ 25（或12-17歲兒童青少年BMI 超過同齡第85百分位)、其他影響免疫功能之疾病或已知重症風險因子者。

中央流行疫情指揮中心   2022/01/27

SARS-CoV-2 main protease (Mpro) + Paxlovid

Biological Assembly 1 ❓
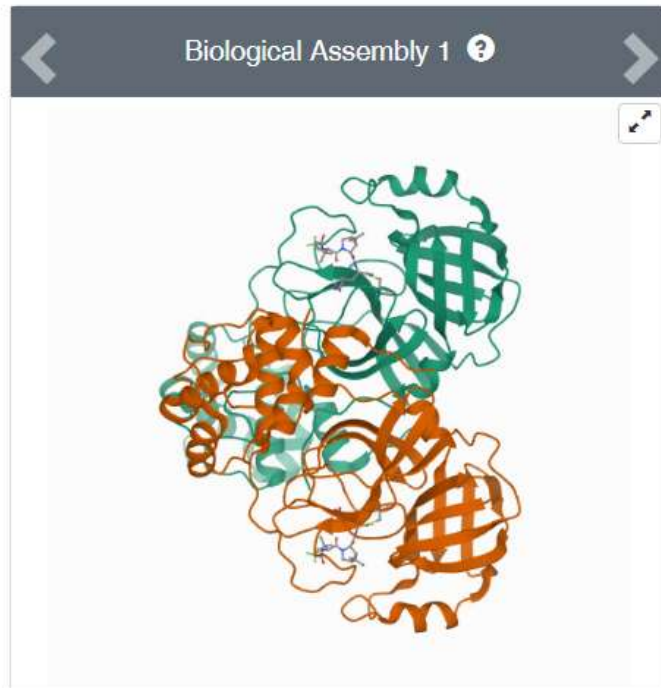
*Paxlovid是Mpro蛋白酶抑制劑

7SI9

Room temperature X-ray structure of SARS-CoV-2 main protease (Mpro) in complex with PF-07321332

PDB DOI: 10.2210/pdb7SI9/pdb

Classification: HYDROLASE
Organism(s): Severe acute respiratory syndrome coronavirus 2
Expression System: Escherichia coli
Mutation(s): No ⓘ

Deposited: 2021-10-12 **Released:** 2021-10-20
Deposition Author(s): Kovalevsky, A., Kneller, D.W., Coates, L.
Funding Organization(s): Not funded

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.00 Å

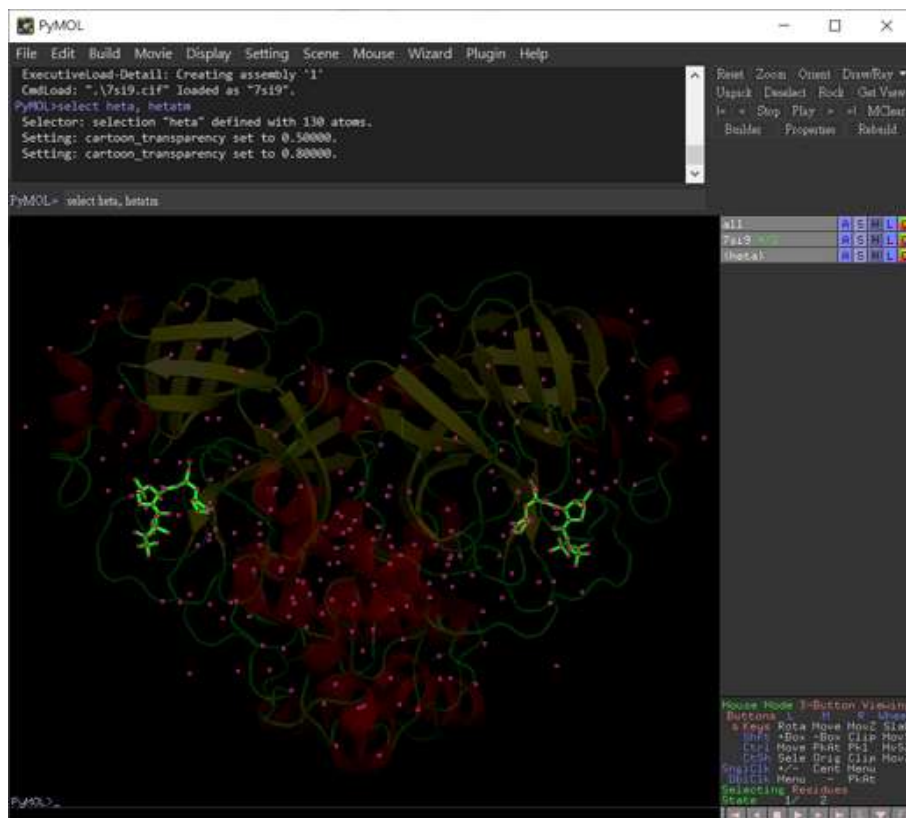📄 Display Files ▾   ⊕ Download Files ▾

https://www.rcsb.org/structure/7SI9

1. 開啟 7SI9
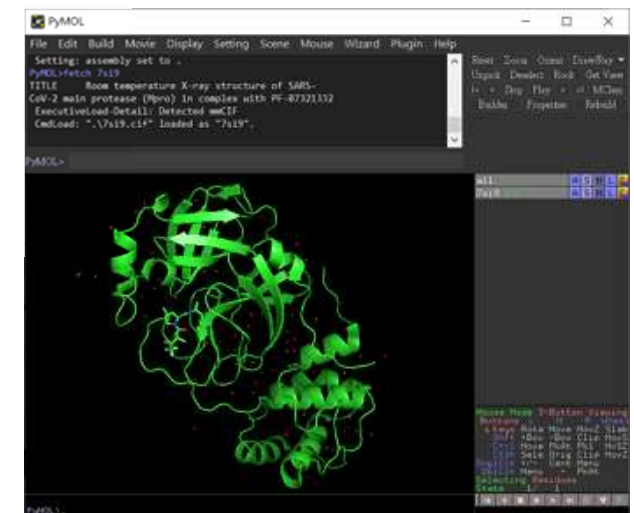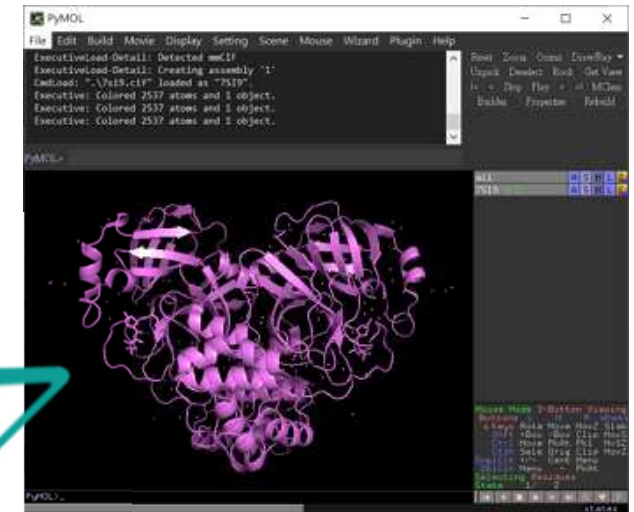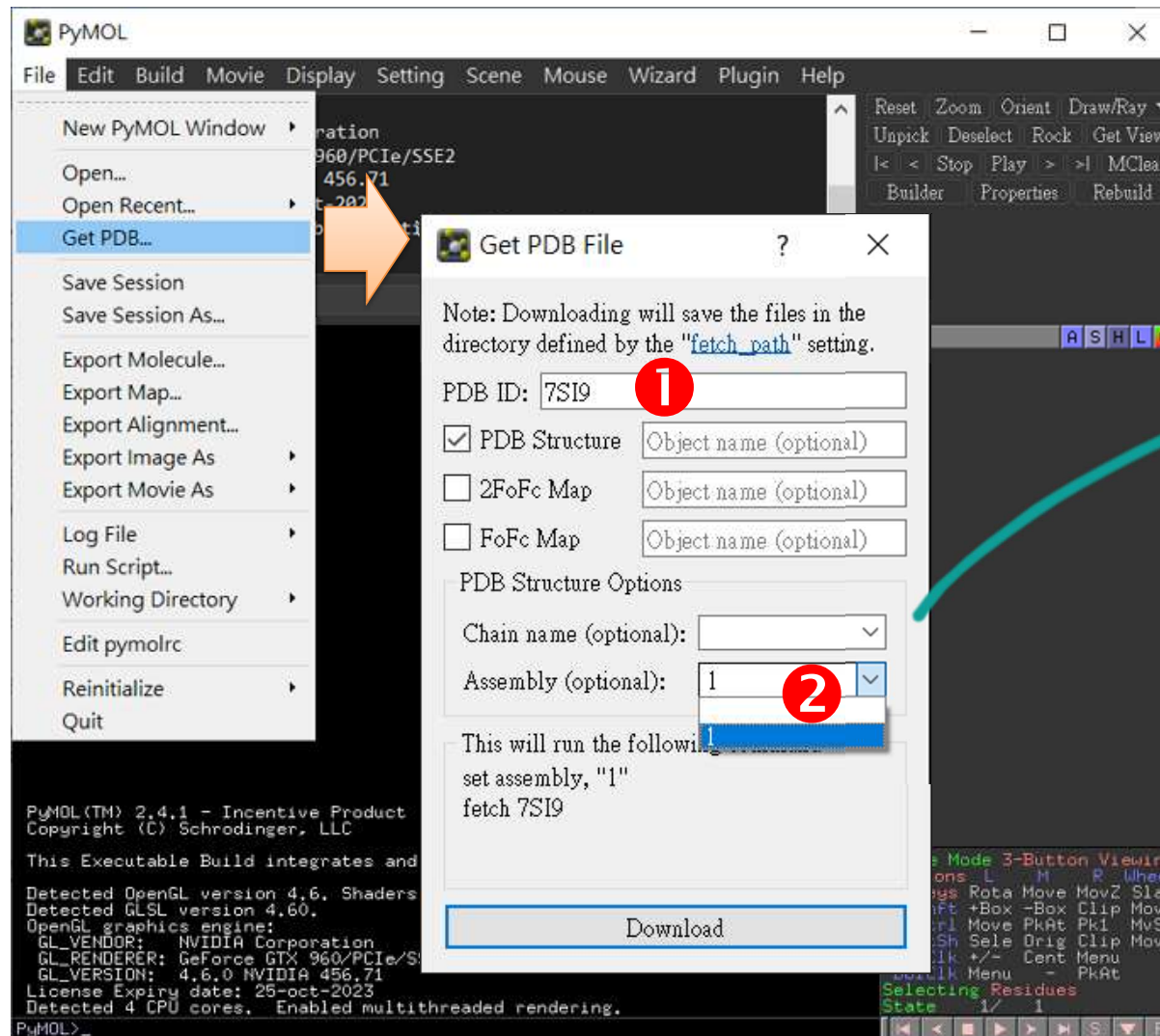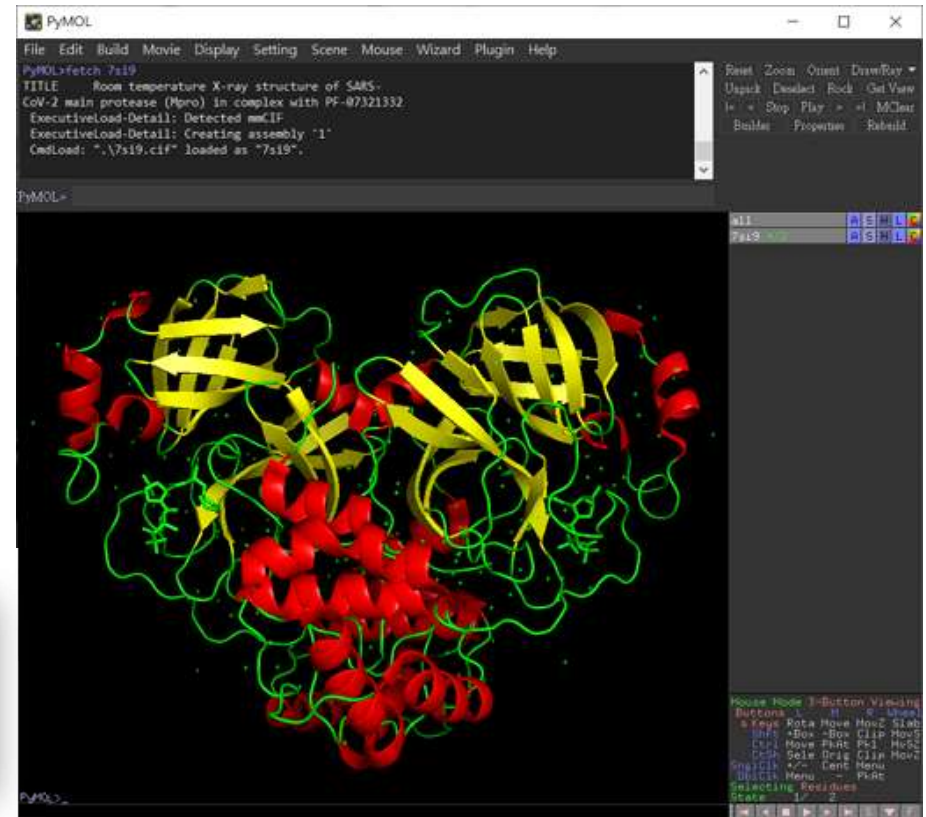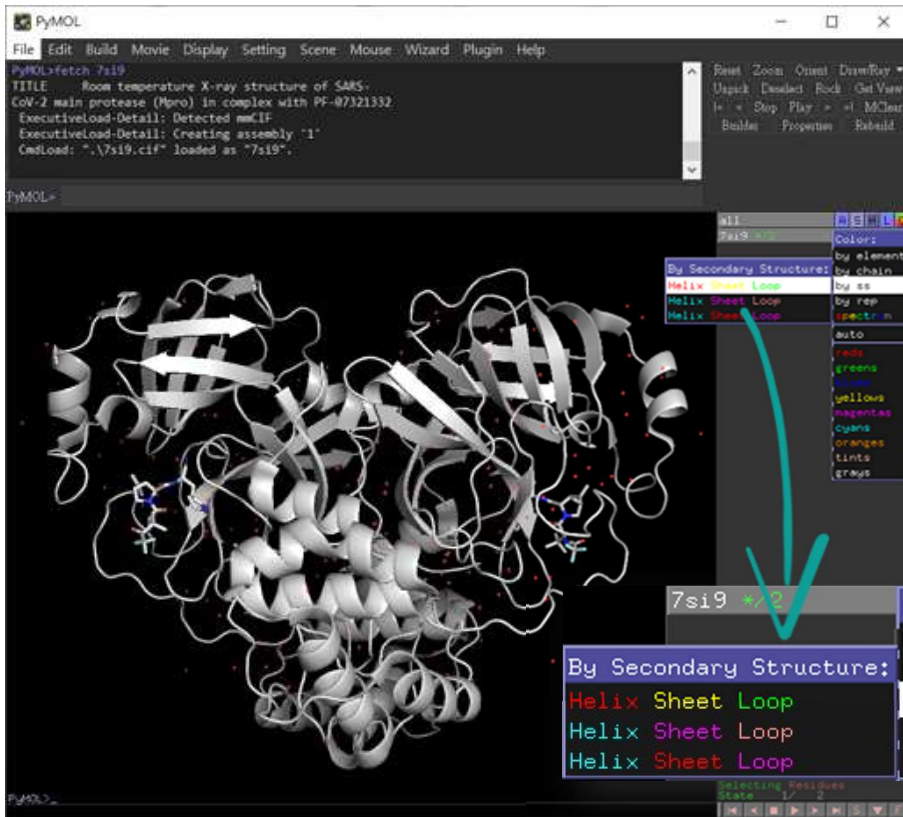
依**Secondary Structure** 著色，Helix, Sheet, Loop 共有**3**種套色可以選擇(見上圖)
Color by ss (By Secondary Structure)

3. 除了 protein chains之外，有沒有其它的東西?

>select heta, hetatm
選擇所要的東西，並命名為 heta (命名以方便辨識為宜)

****

>select resn HOH
選擇水 (resn= residue name)

或是**加以命名:**
>select water, resn HOH
選擇水，並命名為 water

****

>select hetatm and not resn HOH
選擇除了水之外的 hetro atoms

*也可以直接從視窗介面選取與展示

4. 呈現疏水性的胺基酸

>select heta, hetatm
選擇所要的東西，並命名為 heta (命名以方便辨識為宜)

>hide everything (或是用 hide all 亦可)

>show_as ribbon
\# show everything as a ribbon, 這是一個非常好用的指令, 常常我們讓結構顯示了多種3D樣式, 但是最後只要顯示一種時, 只要使用show_as, 就可以直接完成。

>select hydrophobes, resn ala+leu+val+ile+trp+phe+met+pro

>show sticks, hydrophobes

>color magenta, hydrophobes

Ligands

Ligand sites

Double Chains — Complex

One Chain — Receptor

Compound — Ligand
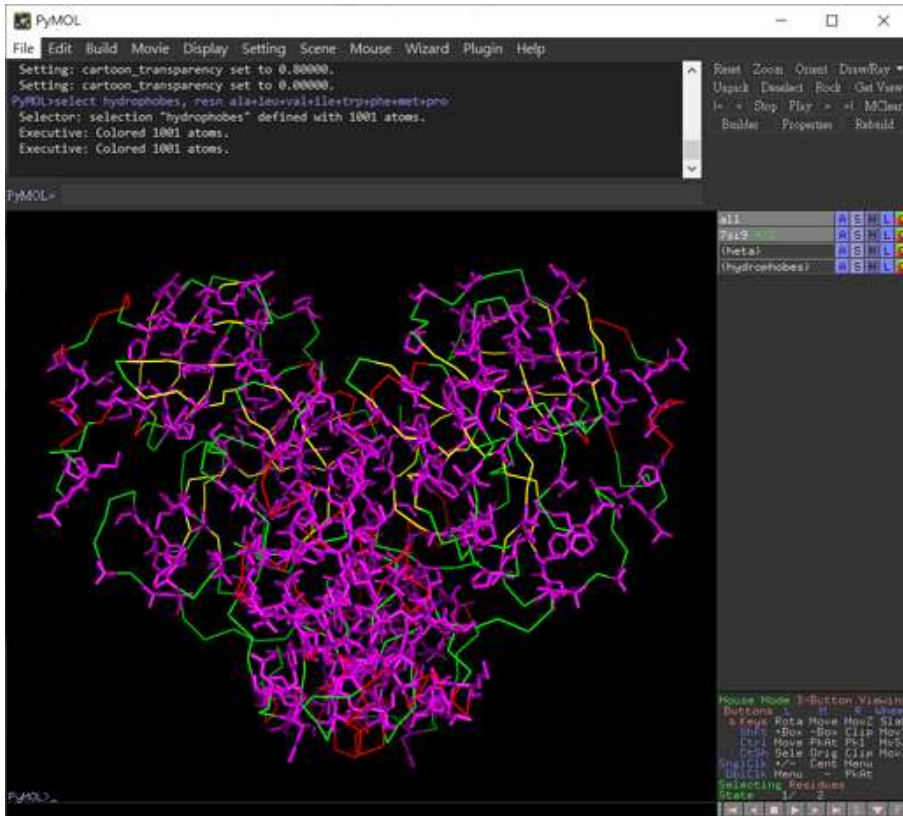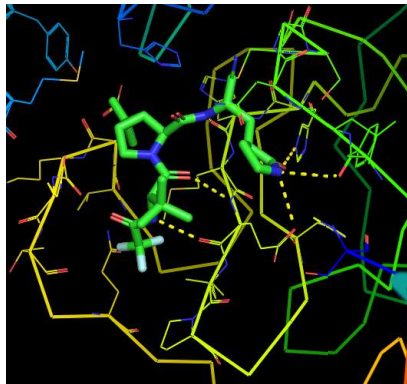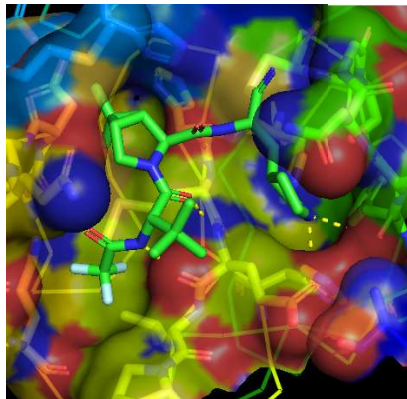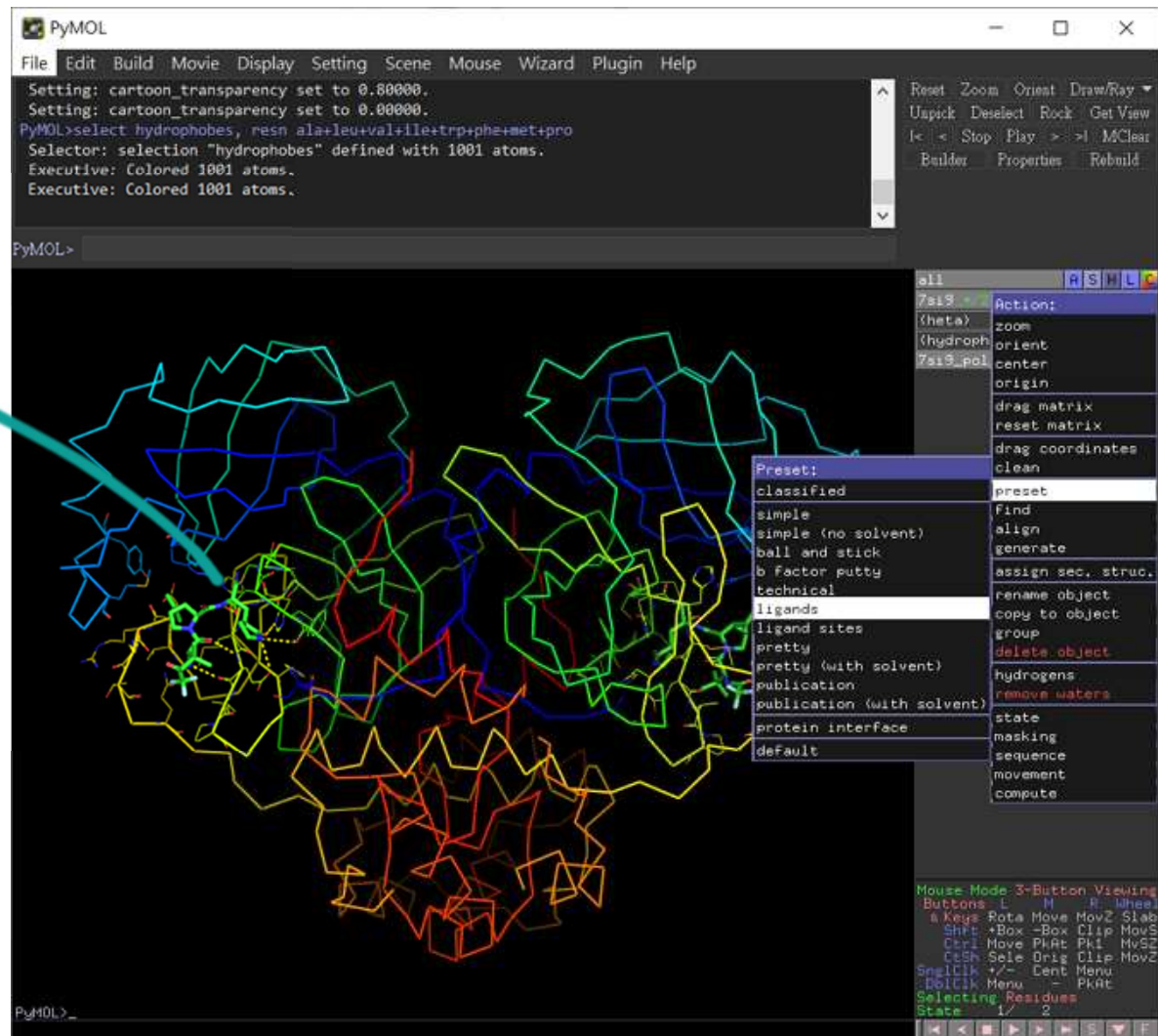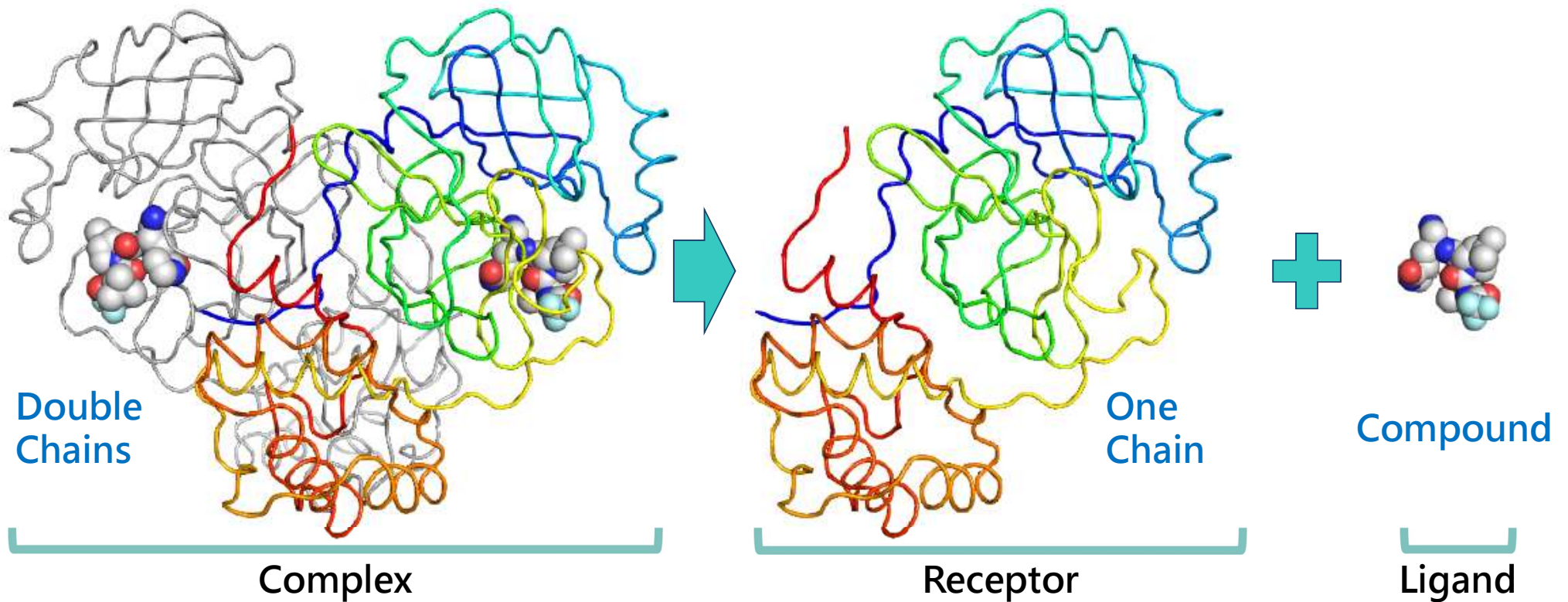
*Receptor與Ligand間主要會探討非共價鍵結作用力，包括靜電力（Ionic Interactions）、偶極力（Dipole Interactions, 如氫鍵）、凡得瓦力（Van Der Waals Interactions）、疏水性（Hydrophobic Effect）
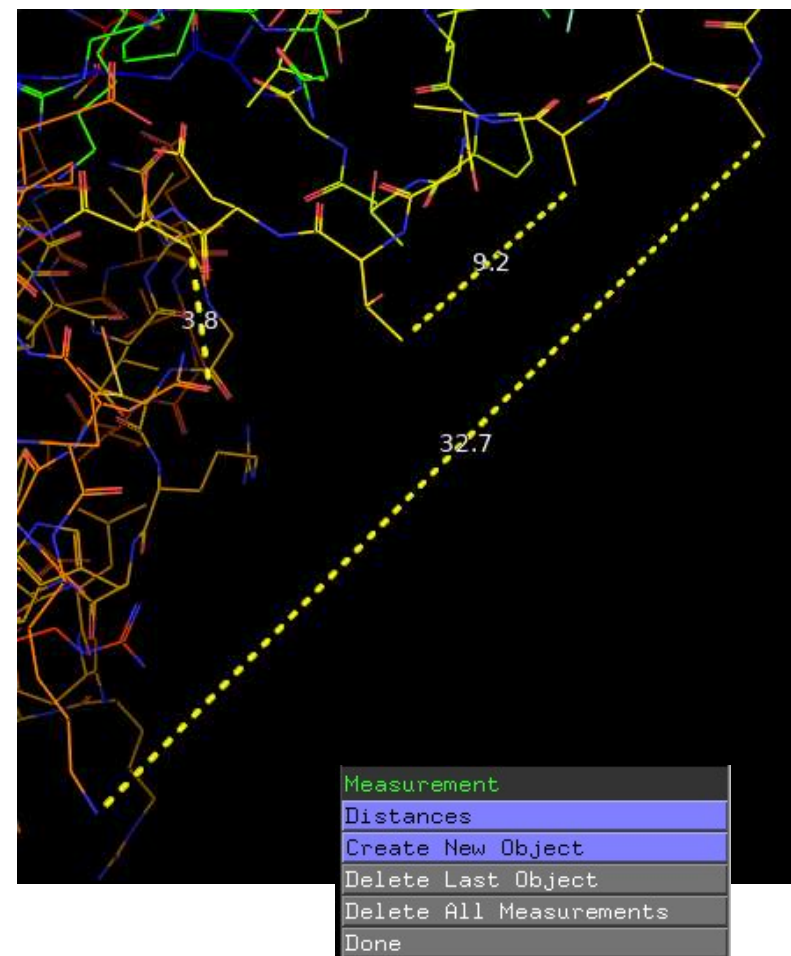
6. 儲存影像
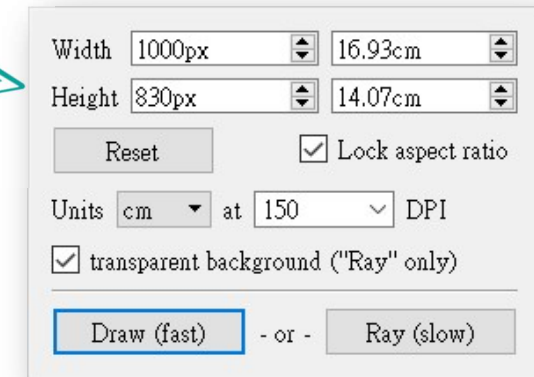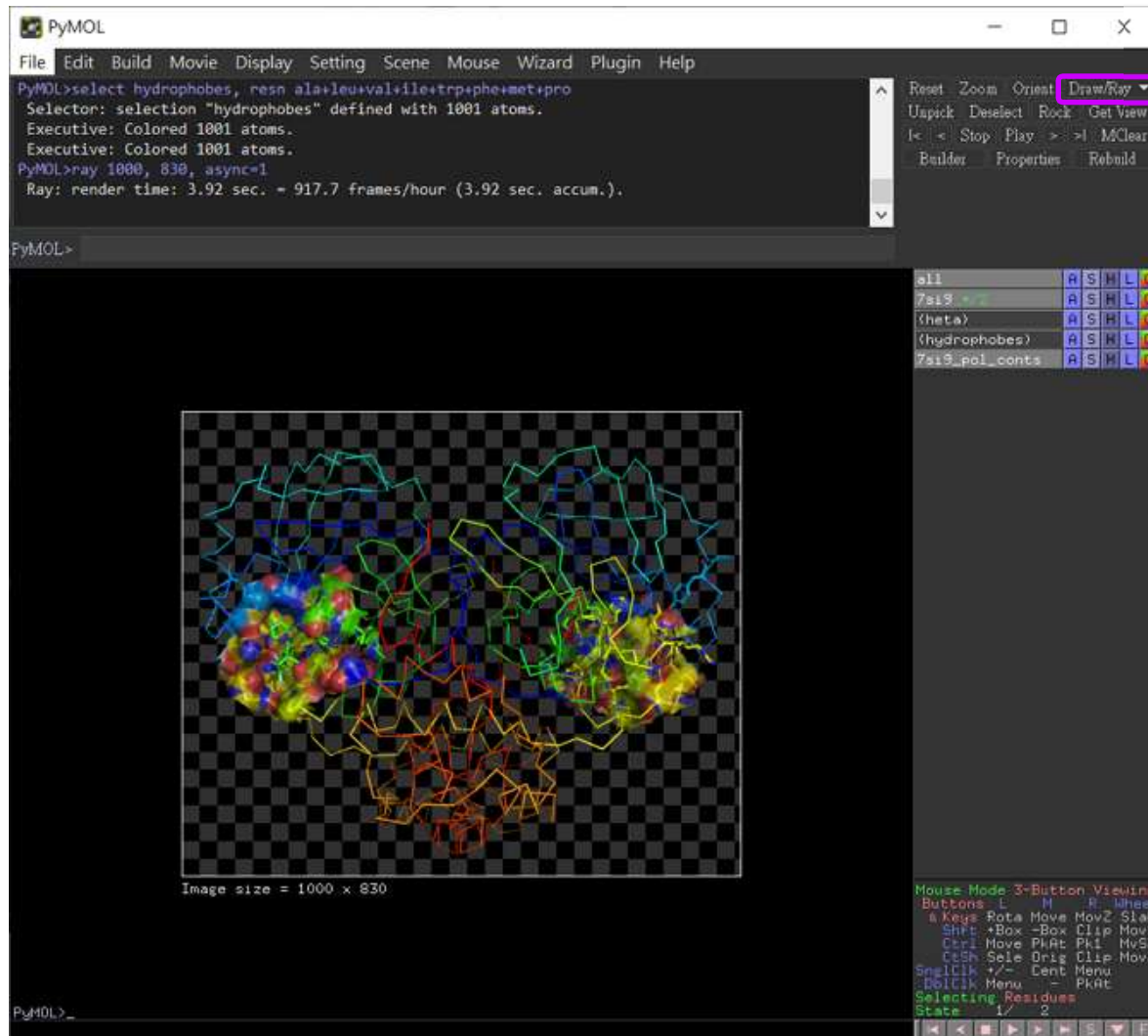
- 一般預設開啟的解析度是 **640 x 480 pixels**，所以 PyMOL 另外提供了 "ray tracer" 功能，可提高影像品質。

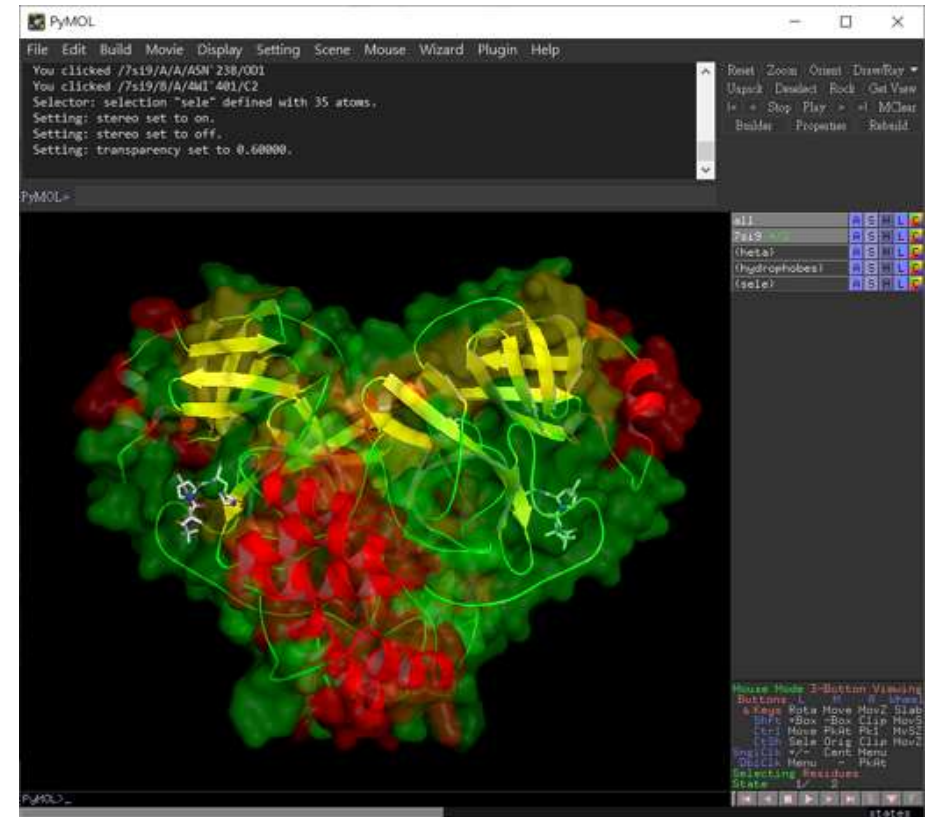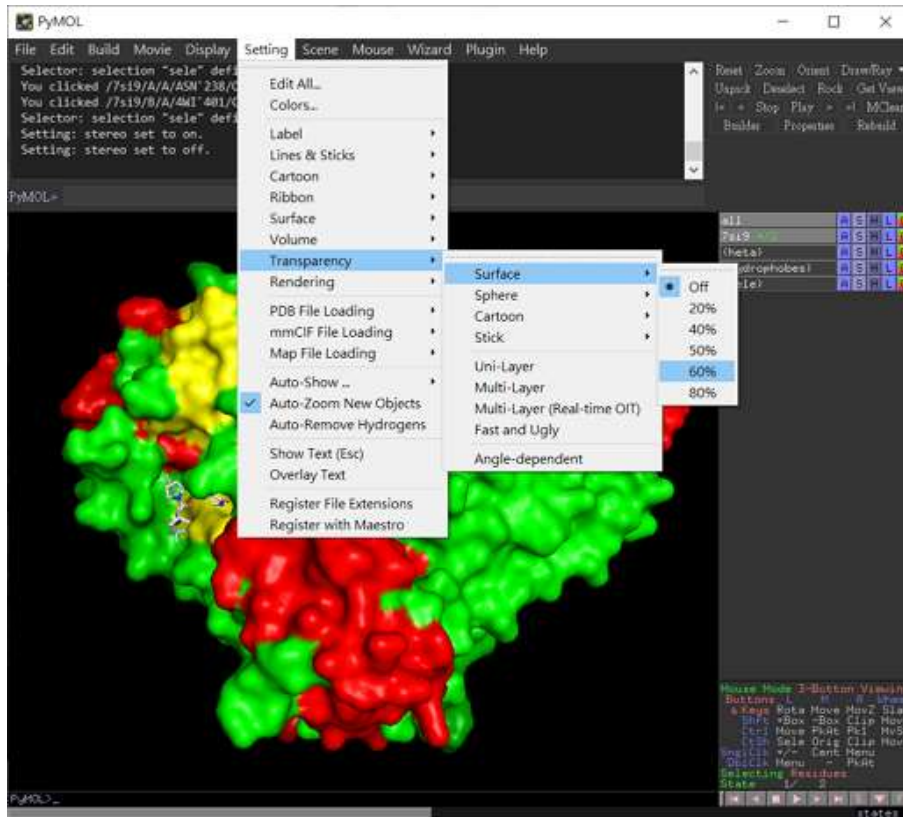- 也可以直接指定影像的解析度: (在指令視窗鍵入) **ray 2000,2000**

- 儲存影像: 存成**png**圖檔

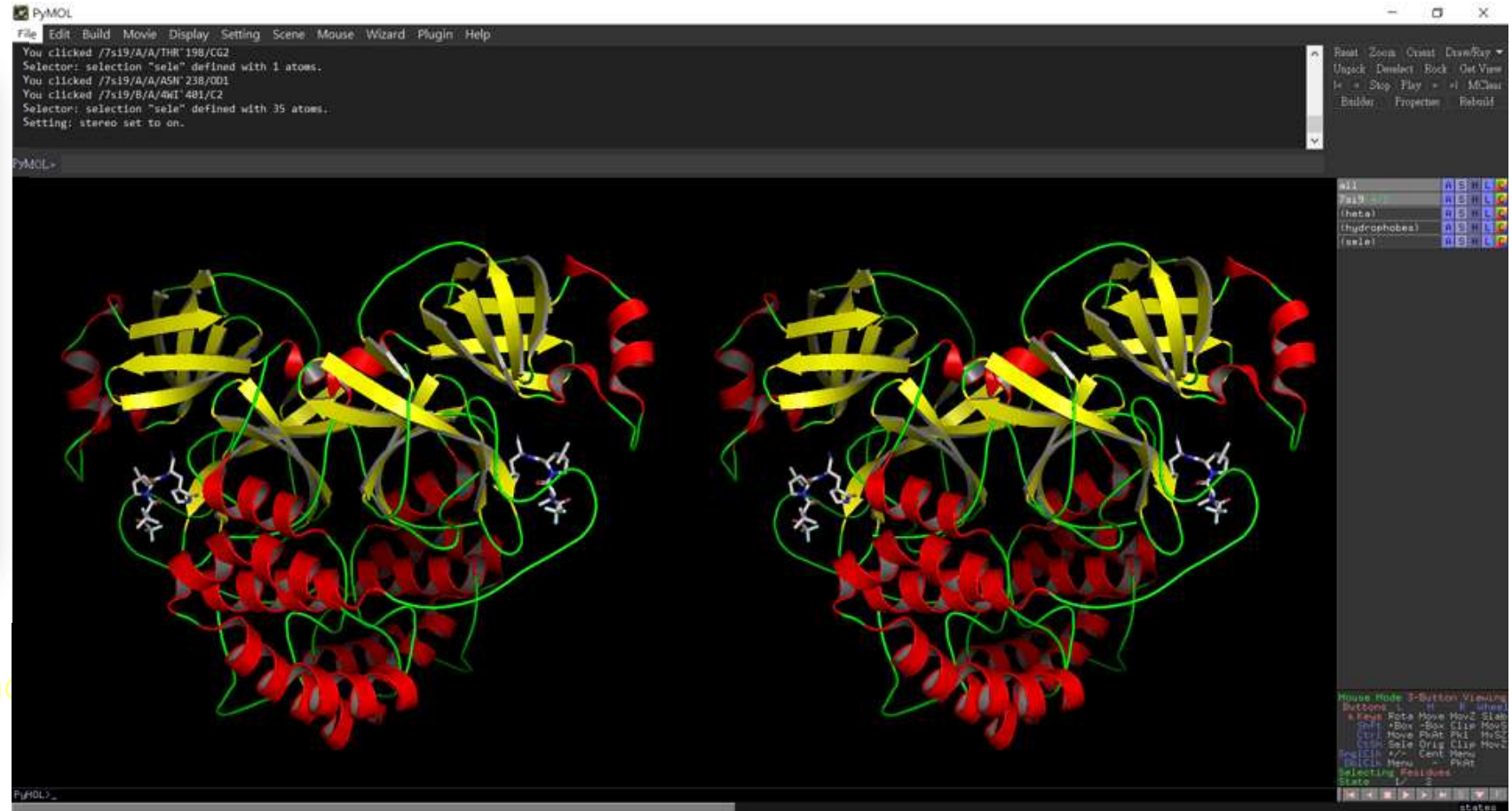# 7. 如何測量兩個原子間的距離?

放大局部區域, 以便測量兩原子間的距離

8. 分子內部圖 (Slab mode)

Show> as> Surface

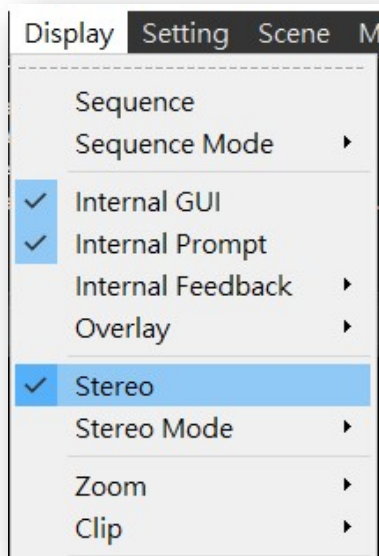滾動**滑鼠中鍵滾輪**，執行平板面模式(slab mode)，呈現**剖面**圖，在剖面之前的原子將看不到，只看到剖面之後的物質。
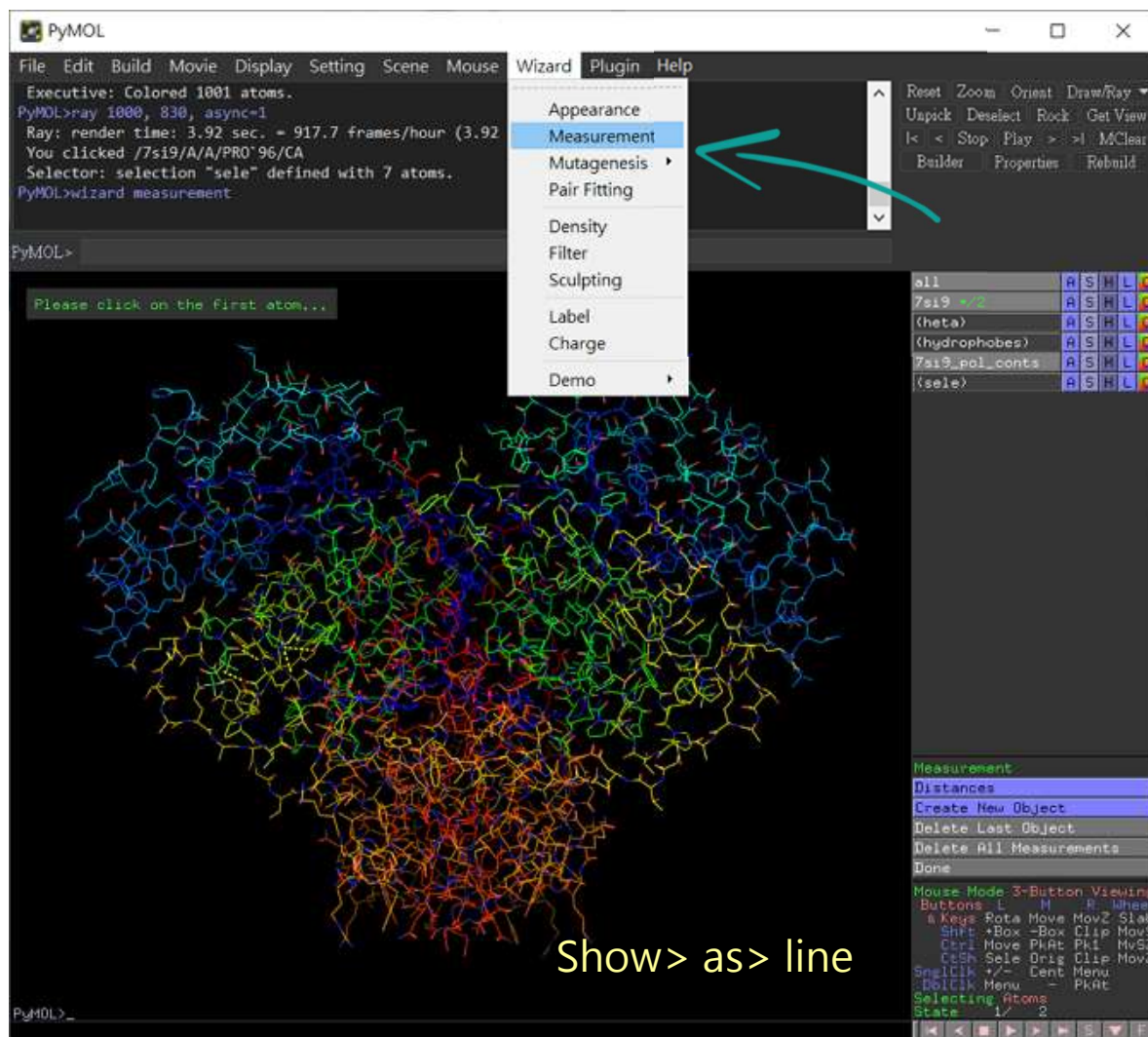
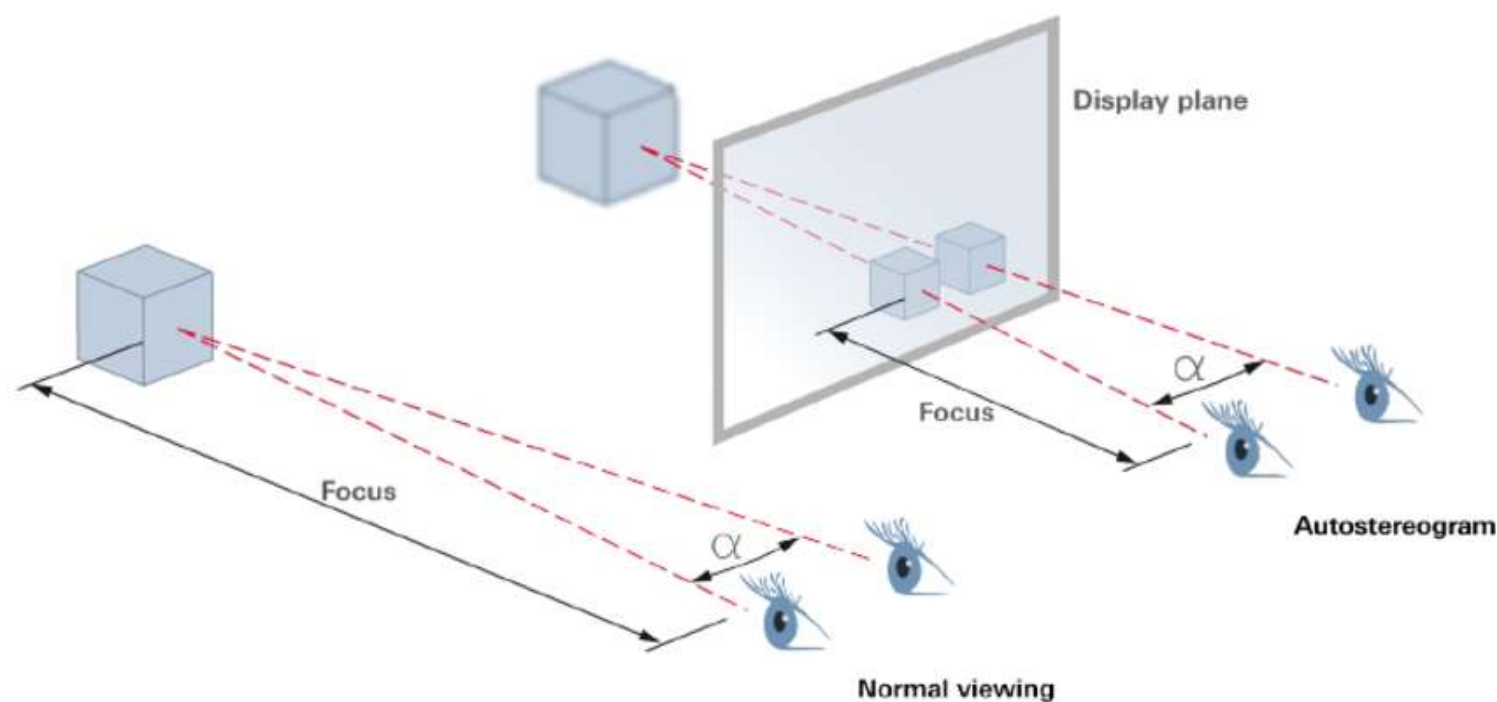# 9. 設定透明surface (Transparency)

- 在功能表Setting的下拉選單中選取**Transparency> Surface> 60%**
- **這是surface透明度60%的效果, 同時也顯露原本被包覆的cartoon樣式**

10. 我要如何看到分子的3D立體圖 (stereo view)

設定立體顯示
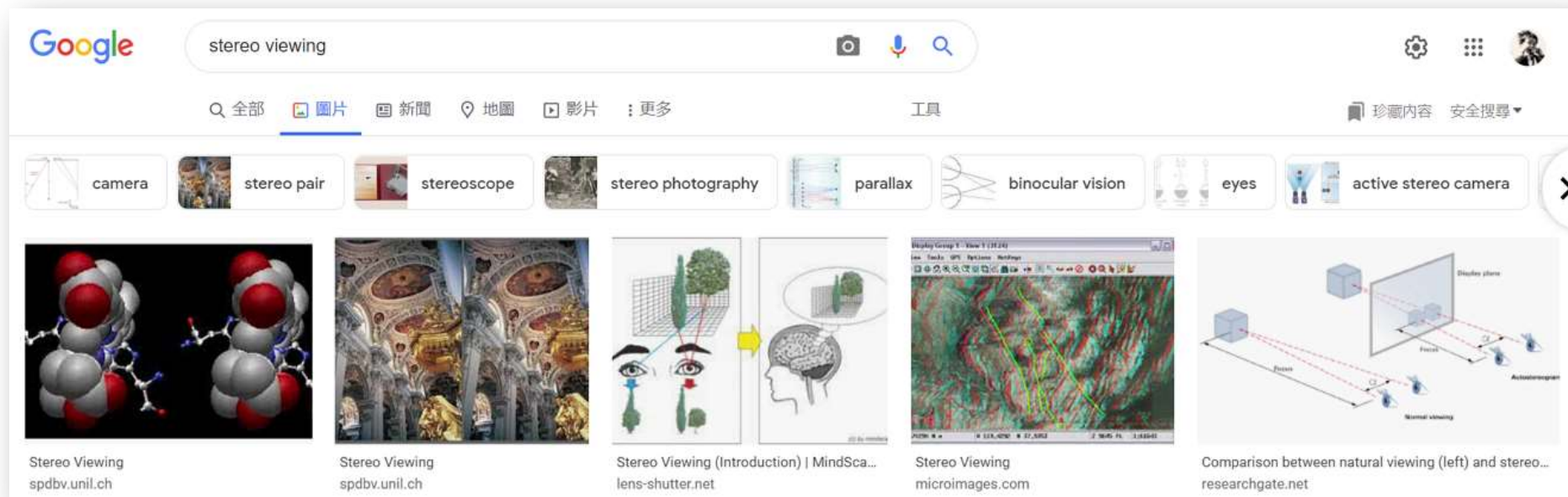**Display> Stereo**

10. 我要如何看到分子的3D立體圖 (stereo view)

10. 我要如何看到分子的3D立體圖 (stereo view)

對於大部分的研究而言，此一步驟並非必要。在沒有立體圖的情形下，旋轉(Rotation)亦有助於了解3D結構的關係。然而，假如你常常看分子圖形，學習如何看立體圖將有助於你對於複合體空間結構的研究。

此外, 大家可以利用Google Search去搜尋"stereo viewing", 可找到許多相關資訊參考。

Colab-OpenSource-PyMOLpySnips

colabOpenSourcePyMOLpySnips

🔗 Library of PyMOL Python snippets for installing and running Open Source PyMOL on Google Colab via a Colab notebook

The installPyMOL snippet will install PyMOL with one click of a button.

CO Open in Colab

Check protein data file from CoLab

Upload

02_pymol on colab.ipynb

To

Colab Notebooks colab

Protein Structure Modeling and Prediction

蛋白質結構模擬與預測

目前主要解析蛋白質結構的實驗方法

Molecular Type	X-ray	NMR	EM	Multiple methods	Neutron	Other	Total
Protein (only)	147998	11984	7772	187	72	32	168045
Protein/Oligosaccharide	8724	31	1372	5	0	0	10132
Protein/NA	7812	277	2461	3	0	0	10553
Nucleic acid (only)	2453	1416	62	12	2	1	3946
Other	154	31	5	0	0	0	190
Oligosaccharide (only)	11	6	0	1	0	4	22
Total	167152	13745	11672	208	74	37	192888

(20220716 updated)

X-ray: X-ray crystallography X-光晶體學

NMR: Nuclear magnetic resonance 核磁共振

cryo-EM: cryo-electron microscopy 冷凍電顯

https://www.rcsb.org/stats/summary

GenBank and WGS Statistics

Sequences

10,000,000,000

Oct 15 2021
WGS: **1,721,064,101**

WGS (Whole Genome Shotgun)

GenBank

100,000,000

1,000,000

10,000

0

1990   2000   2010   2020

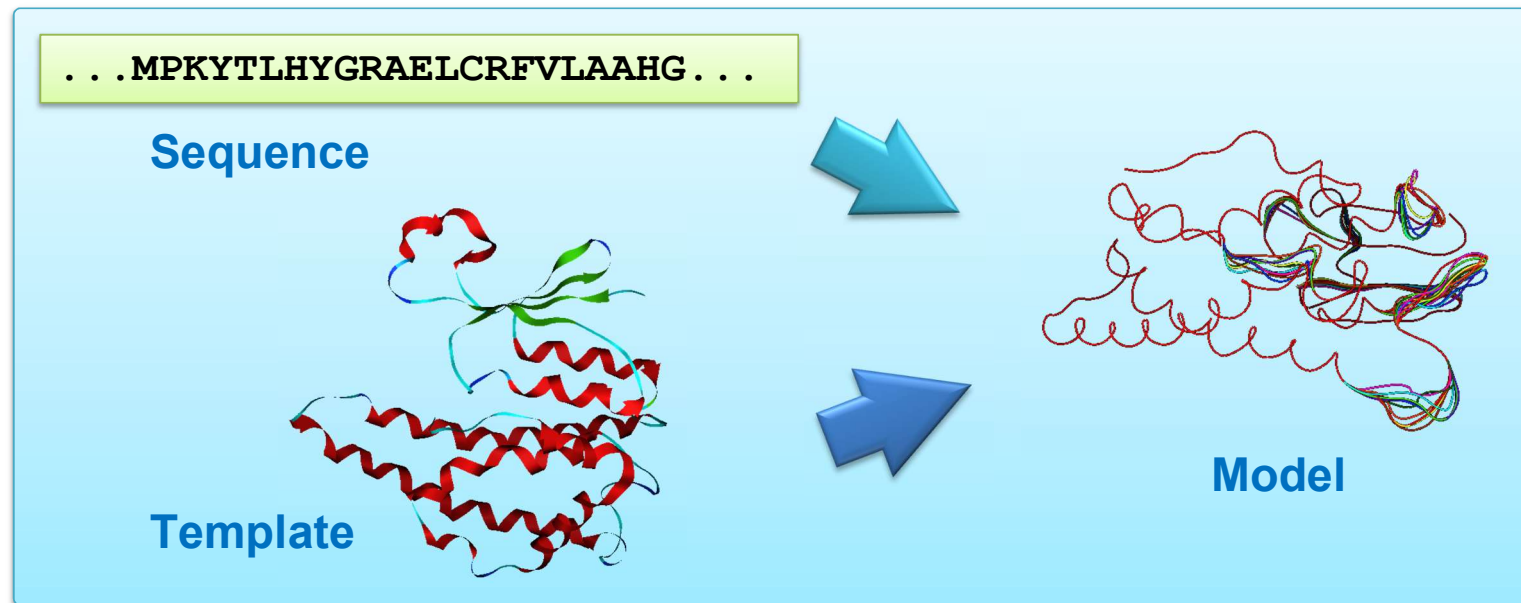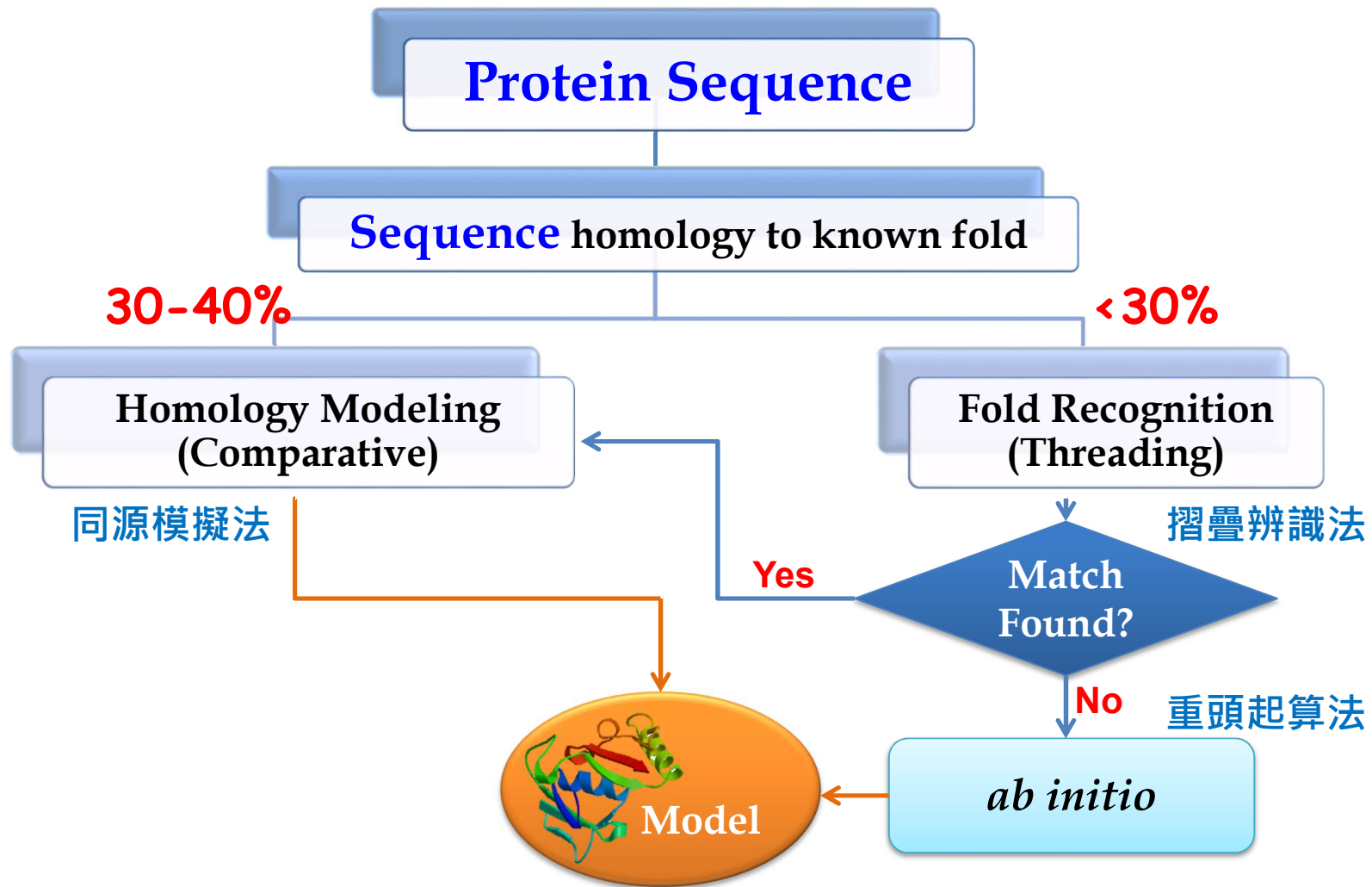*https://www.ncbi.nlm.nih.gov/genbank/statistics/*
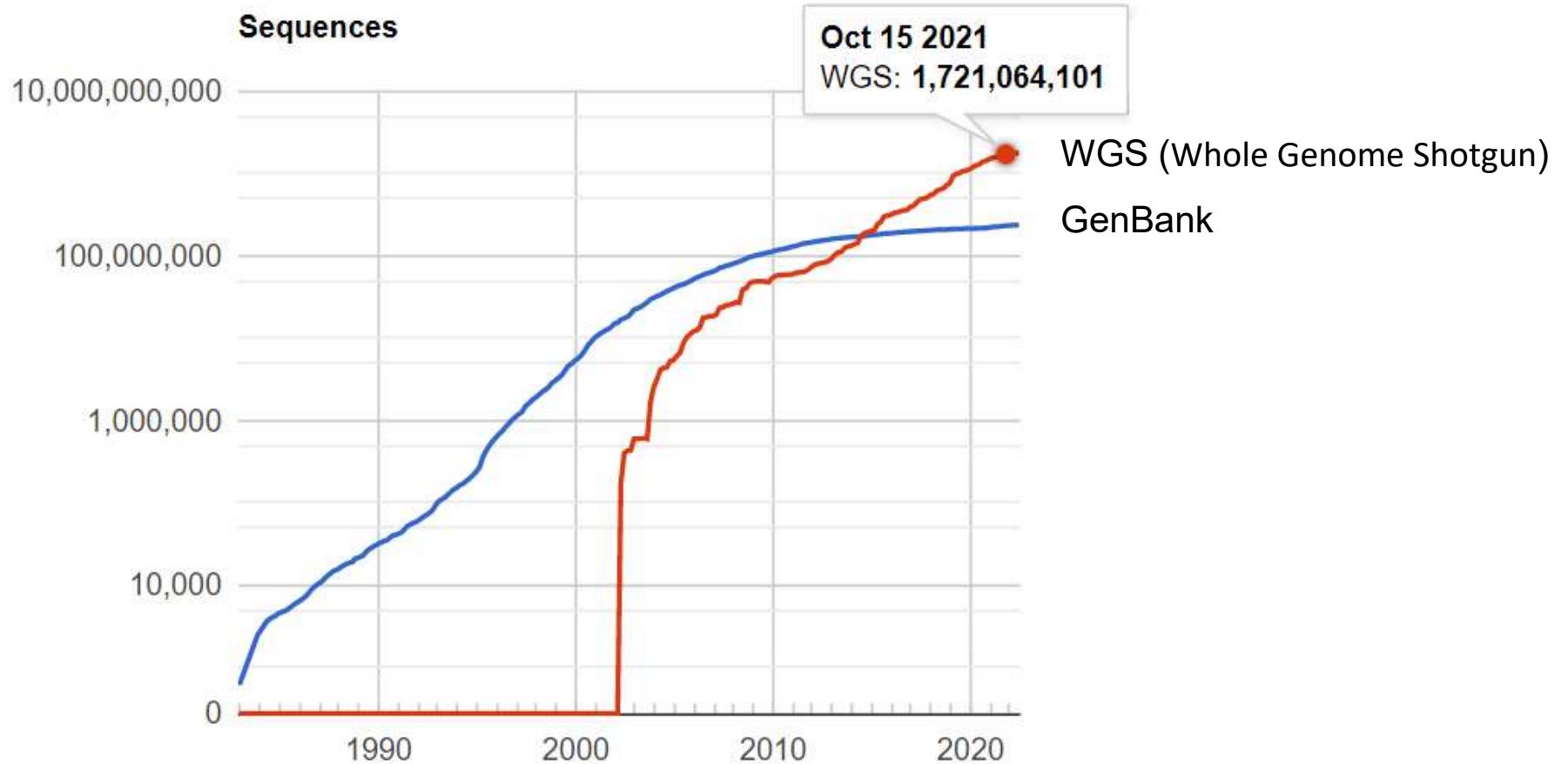
Theoretical Model

- 由於目前已知資料庫中，完整的蛋白質序列數量已遠遠超過所解出的蛋白質結構數量，而且利用實驗方法解出蛋白質結構不是太耗時間，就是受到一些限制而無法定出結構，於是, 科學家想到利用**電腦模擬計算**方式來獲得蛋白質結構。

- 結構模擬大致可分為下列三種方式：

  - Homology or comparative modeling
  - Fold recognition (threading)
  - *ab initio* techniques.

- Homology or comparative modeling is currently the most accurate method to predict the three-dimensional structure of proteins.
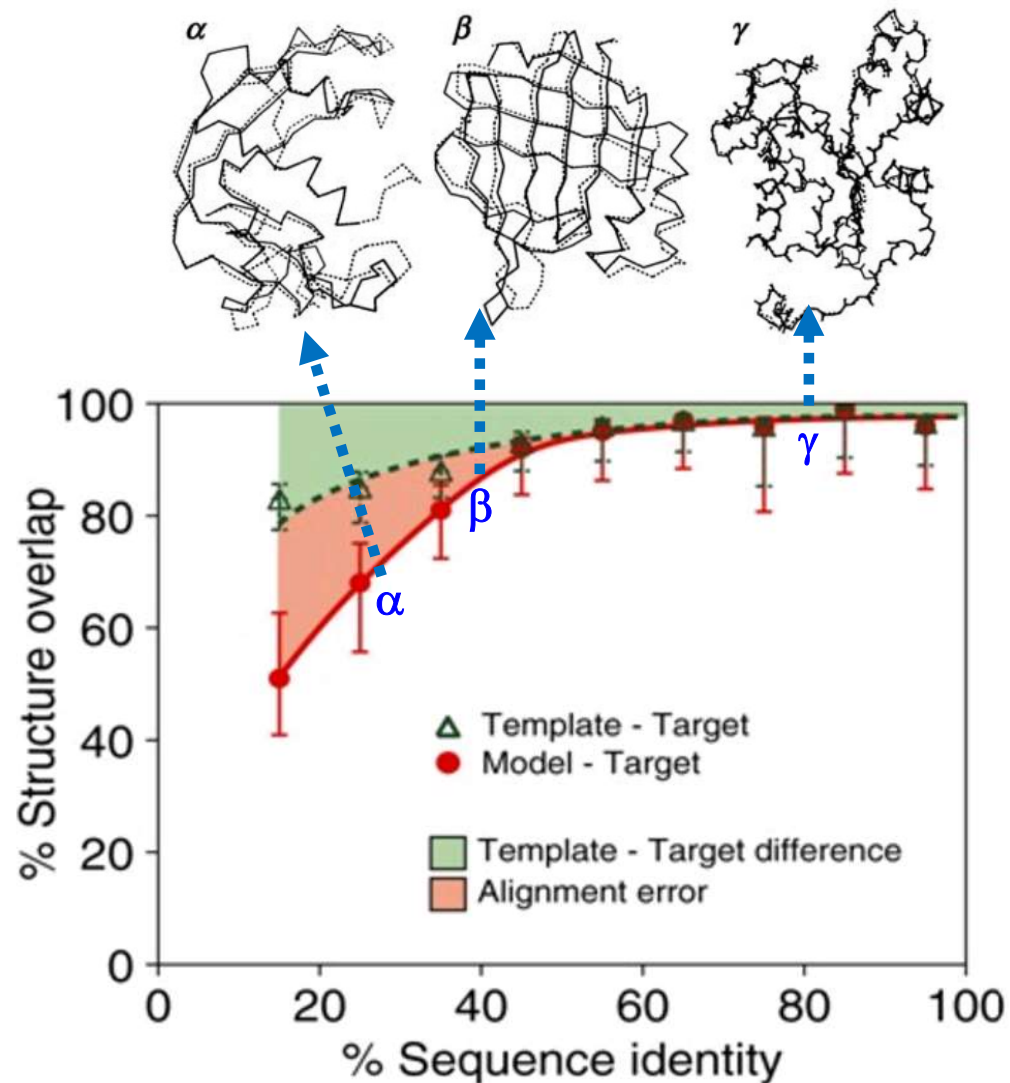
Homology Modeling (同源模擬法)

- 同源模擬法的基本假設是特定的胺基酸序列會構成特定的蛋白質結構。

- 主要是利用現存已解出的結構為模板(由NMR或是X-ray diffraction所解出的結構)，模擬出未知結構蛋白質序列(protein sequence)的三度空間結構。

- 一般來說，當所要解的蛋白質序列(target protein)和模板(template)之間的序列相似度愈高，所模擬出來的結構愈正確也愈可信。

...MPKYTLHYGRAELCRFVLAAHG...

Sequence

Template

Model

Protein Structure Prediction

Protein Sequence

Sequence homology to known fold

30-40%   <30%

Homology Modeling (Comparative)   **Fold Recognition (Threading)**

同源模擬法   摺疊辨識法

Yes   **Match Found?**

No   重頭起算法

Model   *ab initio*

Model Evaluation

The accuracy of the model depends on its sequence identity with the template

In general, 30% sequence identity is required for generating useful models.

Model Building Steps

Part 1 — Target

Part 2 — Structures Templates / Modelling

Part 3 — Evaluation
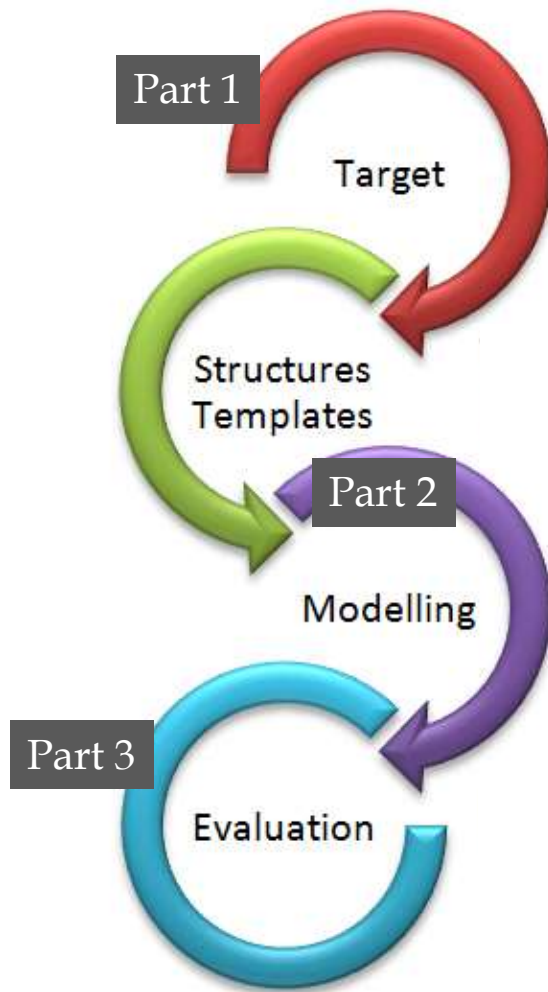
- Target sequence (primary sequence)

- Searching for structures of templates in databases

- Construction of the model

- Evaluation of model quality

Start Your Modeling

SARS-CoV-2與靶細胞的表面蛋白相互作用

結構功能的蛋白質
Structural Proteins

包膜蛋白
Envelope (E) protein

膜蛋白
Membrane (M) protein

核殼蛋白
Nucleocapsid

棘蛋白(S蛋白)
Spike (S) protein

非結構功能的蛋白質
Non-Structural Proteins

COVID-19 virus

直徑大小介於
50-200nm

RBD vs. ACE2

Transmembrane protease serine 2 (TMPRSS2)

Angiotensine converting enzyme 2 (ACE2)

靶細胞的受體

Marta Palma

https://pdb101.rcsb.org/news/2020#5ed7c17cdab5c9354c274f56
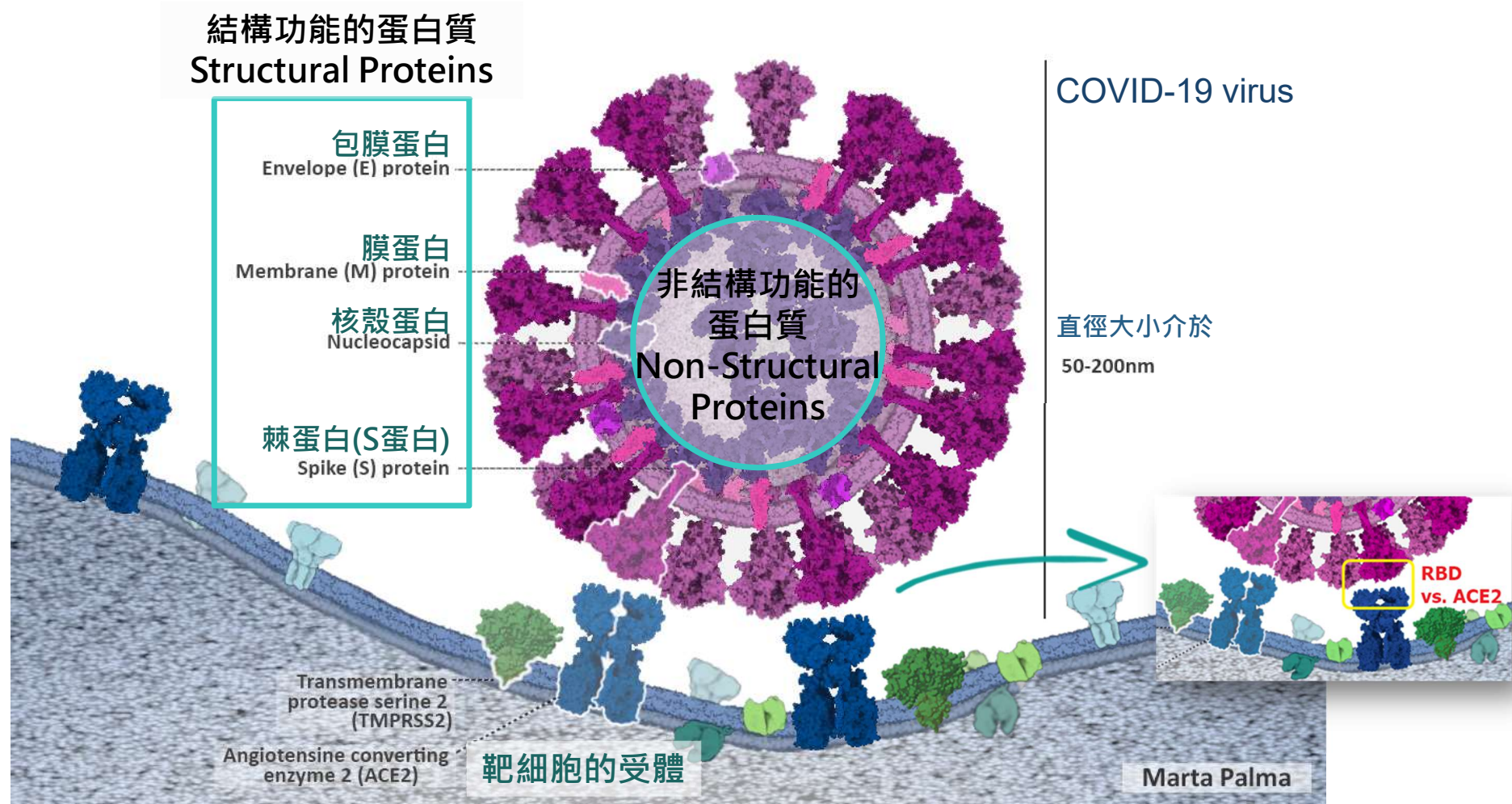
SARS-CoV-2的基因組與蛋白質結構組成
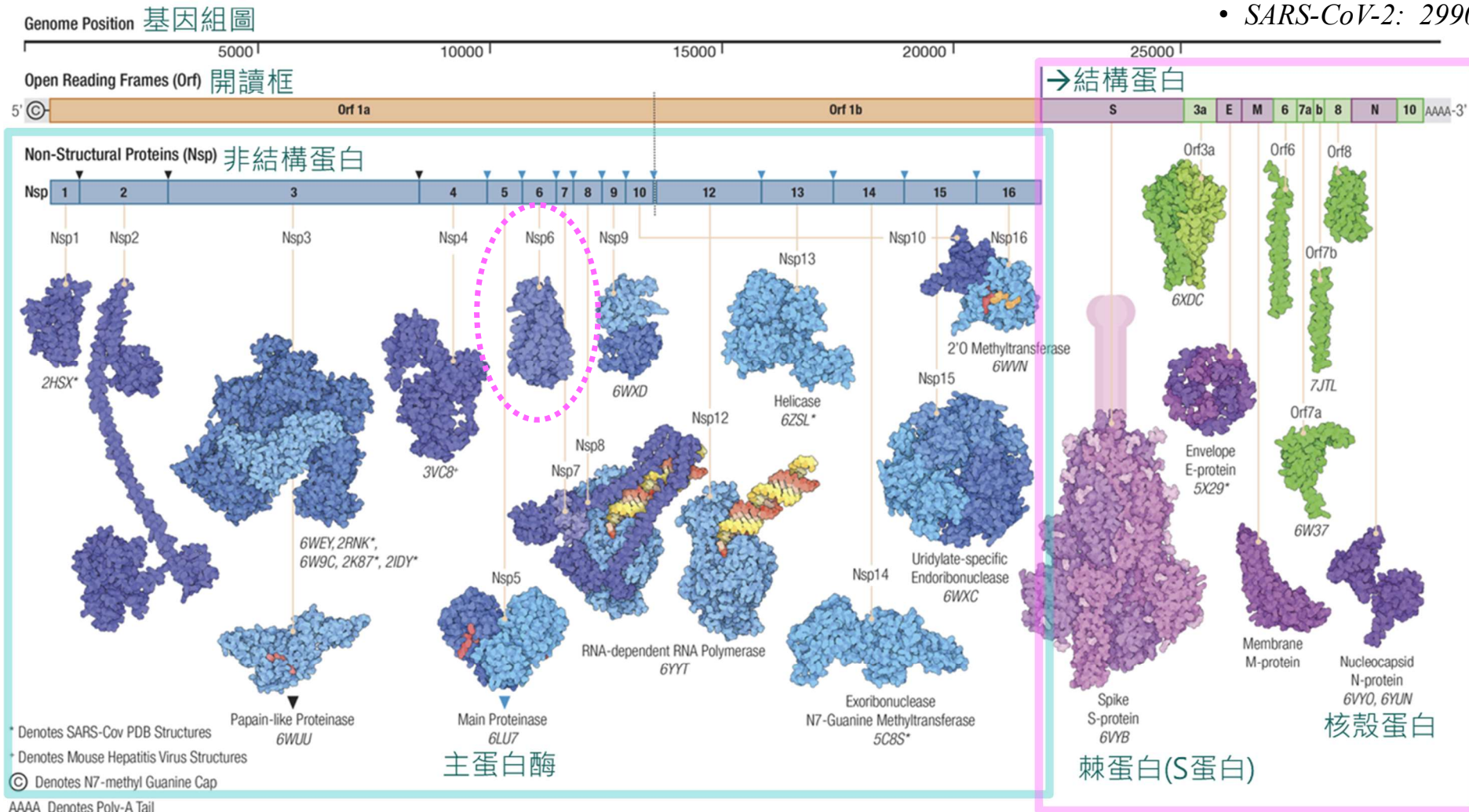
- *SARS-CoV-2: 29903 base pair (bp)*
  29.9 Kb
- *E. Coli*
  4.6 Mb
- *Homo*
  3.2 Gb

Architecture of the SARS-CoV-2 genome and proteome from *bioRxiv* 2020.12.01.406637

NCBI SARS-CoV-2 Resources

SARS-CoV-2 next-generation sequencing runs in **SRA**

View in SRA

Coronaviridae family-containing **SRA** runs

Download from SRA Run Selector

SARS-CoV-2 protein structures, domains, and sequences available through **NCBI Structure**

View Structure Data

SARS-CoV-2 related compounds, substances, pathways, bioassays, and more in **PubChem**

View in PubChem
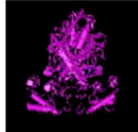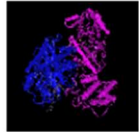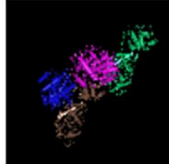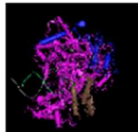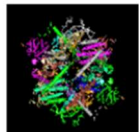
Genome expression studies related to SARS-CoV-2 in **GEO**

View in GEO

https://www.ncbi.nlm.nih.gov/sars-cov-2/

Select modeling target : Nsp6

SARS-CoV-2-related data provided by the **Protein Domains** resource

ORF name	CDD model	Model name (mouse-over name for description)	SARS Cov-2 structure	SARSr-CoV structure	Other coronavirus structure	RefSeq protein	Length
Nsp5	cd21666	betaCoV_Nsp5_Mpro	7BQY [all]	1WOF [all]	3D23 [all]	YP_009725301 [domain architecture]	306
Nsp6	cd21560	betaCoV-Nsp6	*No structure data available!*			YP_009725302 [domain architecture]	290
Nsp7	cd21827	betaCoV_Nsp7	7BV2 [all]	2AHM [all]	-	YP_009725303 [domain architecture]	83

https://www.ncbi.nlm.nih.gov/Structure/SARS-CoV-2.html

Sequence of Nsp6

National Library of Medicine
National Center for Biotechnology Information

Log in

Protein

| Protein ∨ | Nsp6 SARS-CoV-2 | ⊗ | Search |

Create alert   Advanced

Help

```
>nsp6
SAVKRTIKGTHHWLLLTILTSLLVLVQSTQWSLFFFLYENAFLPFAMGIIAMSAFAMMFVKHKHAFLCLF
LLPSLATVAYFNMVYMPASWVMRIMTWLDMVDTSLSGFKLKDCVMYASAVVLLILMTARTVYDDGARRVW
TLMNVLTLVYKVYYGNALDQAISMWALIISVTSNYSGVVTTVMFLARGIVFMCVEYCPIFFITGNTLQCI
MLVYCFLGYFCTCYFGLFCLLNRYFRLTLGVYDYLVSTQEFRYMNSQGLLPPKNSIDAFKLNIKLLGVGG
KPCIKVATVQ
```

Model by SWISS-MODEL

http://swissmodel.expasy.org/

Start a New Modelling Project

1. Copy & paste your target sequence;   2. Title: *nsp6 (optional)*   3. Click "*Search For Templates*"

Model Results

Template Results ⊙

| | Templates | Quaternary Structure | Sequence Similarity | Alignment | More ▾ |

↓↑Sort	⬍Coverage	⬍GMQE	⬍QSQE	⬍Identity	⬍Method	⬍Oligo State	⬍Ligands
☑	7nbu.1.P 30S ribosomal protein S16 Structure of the HigB1 toxin mutant K95A from Mycobacterium tuberculosis (Rv1955) and its target, the cspA mRNA, on the E. coli Ribosome.						
⌄		0.00	-	7.69	EM	hetero-50-mer ⚠	1 x C-A ⤢, 21 x MG ⤢, 1 x MS6 ⤢, 2 x ZN ⤢
☑	5t7v.1.F 30S ribosomal protein S16 Methicillin Resistant, Linezolid resistant Staphylococcus aureus 70S ribosome (delta S145 uL3)						
⌄		0.00	-	12.00	EM	hetero-38-mer ⚠	None
☐	6q97.1.p 30S ribosomal protein S16 Structure of tmRNA-SmpB bound in A site of E. coli 70S ribosome						
⌄		0.00	-	7.69	EM	hetero-53-mer ⚠	1 x A-U-G-G-U-C, 24 x MG ⤢, 2 x ZN ⤢
☐	4v5h.1.P 30S RIBOSOMAL PROTEIN S16 E.Coli 70s Ribosome Stalled During Translation Of TnaC Leader Peptide.						
⌄		0.00	-	8.00	EM, 5.8Å	hetero-51-mer ⚠	None
☐	5li0.1.P 30S ribosomal protein S16 70S ribosome from Staphylococcus aureus						
⌄		0.00	-	11.54	EM	hetero-47-mer ⚠	1 x O ⤢, 1 x ZN ⤢
☐	4v7t.2.P 30S ribosomal protein S16 Crystal structure of the E. coli ribosome bound to chloramphenicol.						

Build Models ②

Clear Selection
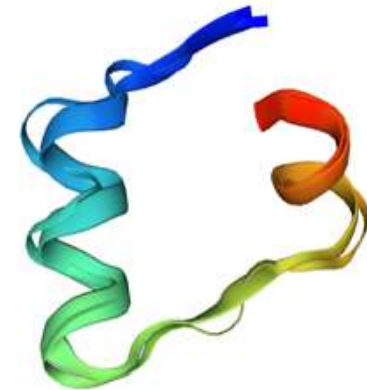
⚙ Cartoon ▲ 📷 ▲ ▶ ⏏ ⟳

| 7nbu.1.P | | ✖ |
| 5t7v.1.F | | ✖ |

AlphaFold2 應用實例

DeepMind/ AlphaFold

❖ AlphaFold是Alphabet/Google旗下DeepMind開發的一款進行**蛋白質結構預測**的人工智慧程式，AlphaFold有兩個主要版本，AlphaFold 1 (2018)在2018年12月第13屆蛋白質結構預測技術的關鍵測試（**Critical Assessment of protein Structure Prediction，CASP**）的總體排名中名列第一，該程序成功預測了被競賽組織者評為最困難目標的最準確結構，其中沒有來自具有部分相似序列蛋白質的現有模板結構。

❖ 在2020年11月的第14屆**CASP**競賽中，AlphaFold 2 (2020)表現良好，中位分數為92.4（滿分100分），準確度遠遠高於其他任何程式。

AlphaFold in CASP

Median Free–Modelling Accuracy

- The main metric used by CASP to measure the accuracy of predictions is the Global Distance Test (GDT) which ranges from 0-100.

- Improvements in the median accuracy of predictions in the free modelling category for the best team in each CASP, measured as best-of-5 GDT.

https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology
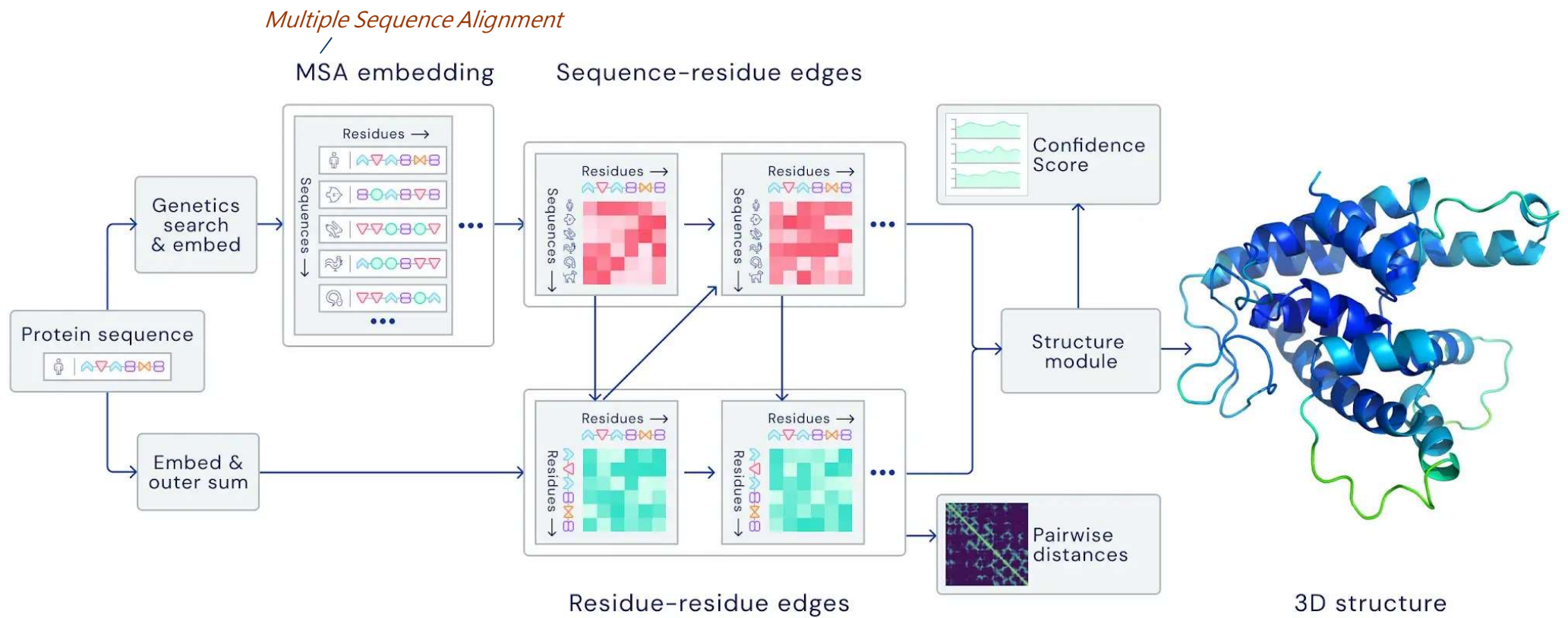
An overview of the main neural network model architecture.

https://alphafold.ebi.ac.uk/

AlphaFold Protein Structure Database

- There are **995,411 structures** available on the AlphaFold DB website.

- AlphaFold DB currently provides predicted structures for the **48 organisms**:
  - Model organism proteomes
  - Global health proteomes
  - Swiss-Prot

KLHL23

Length:558
Mass (Da):63,923
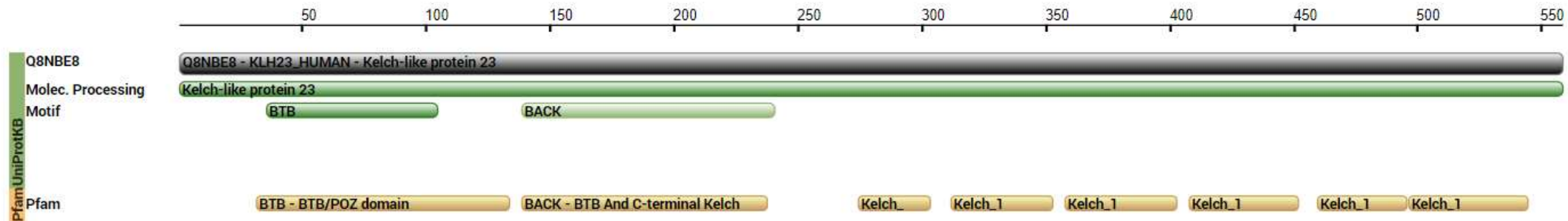
KLHL23 structure??
UniProtKB - Q8NBE8 (KLH23_HUMAN)
https://www.uniprot.org/uniprot/Q8NBE8

Related template structures of KLHL23

Q9NVR0
- Q9NVR0 - KLH11_HUMAN - Kelch-like protein 11
- Molec. Processing: si Kelch-like protein 11
- Motif: BTB BACK
- UP Sites
- Secstruc
- PDB Validation
- 3I3N.B
- 3I3N.A

3I3N.B

Q14145
- Q14145 - KEAP1_HUMAN - Kelch-like ECH-associated protein 1
- Molec. Processing: Kelch-like ECH-associated protein 1
- Motif: BTB BACK
- UP Sites
- Secstruc
- PDB Validation
- 2FLU.X

2FLU.X

A0A077LQB4
- A0A077LQB4 - A0A077LQB4_PLAFA - Putative kelch protein K13
- Motif: coiled coi BTB
- Secstruc
- PDB Validation
- 4YY8.B
- 4YY8.A

4YY8.B

KLHL23: Kelch-like protein 23

Structure[i]

Model Confidence:

■ Very high (pLDDT > 90)

□ Confident (90 > pLDDT > 70)

□ Low (70 > pLDDT > 50)

□ Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions with low pLDDT may be unstructured in isolation.

pLDDT corresponds to the model's prediction of its score on the local Distance Difference Test

https://www.uniprot.org/uniprot/Q8NBE8
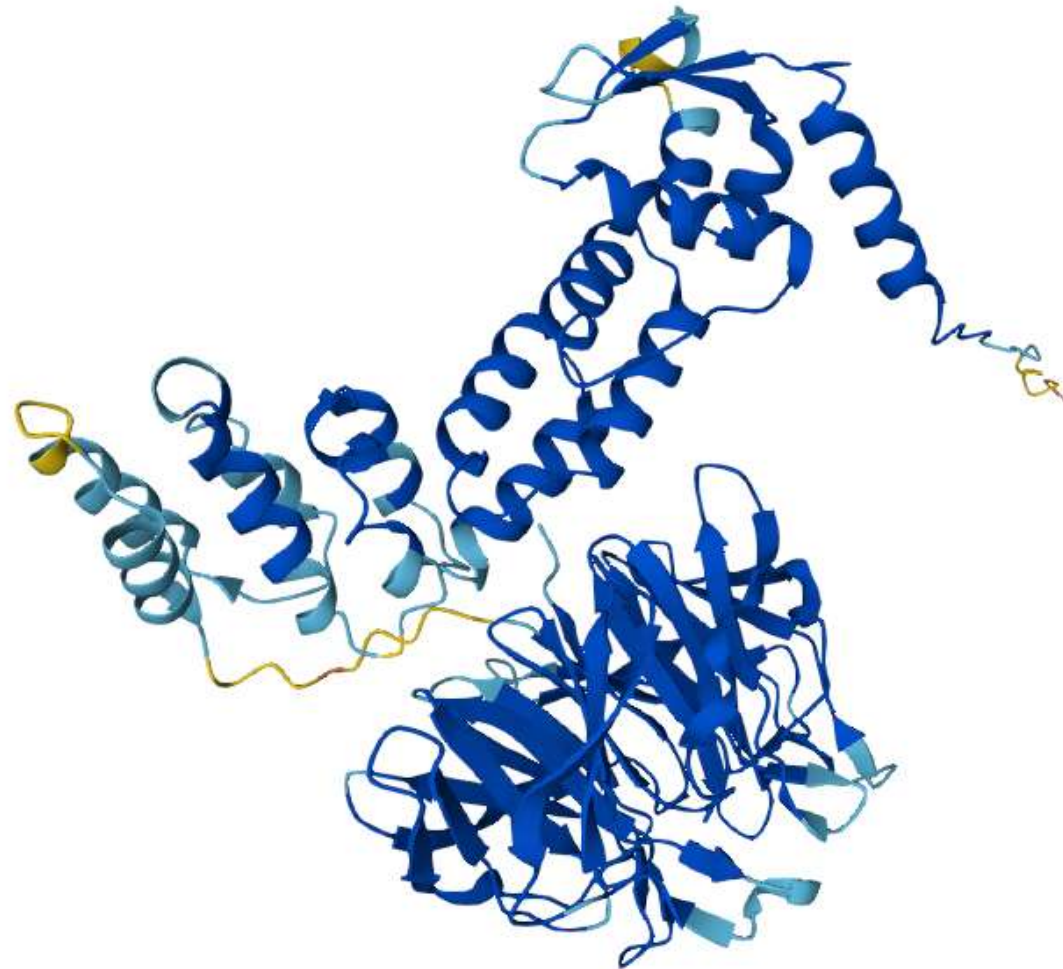
KLHL23

AlphaFold structure prediction
https://alphafold.ebi.ac.uk/entry/Q8NBE8
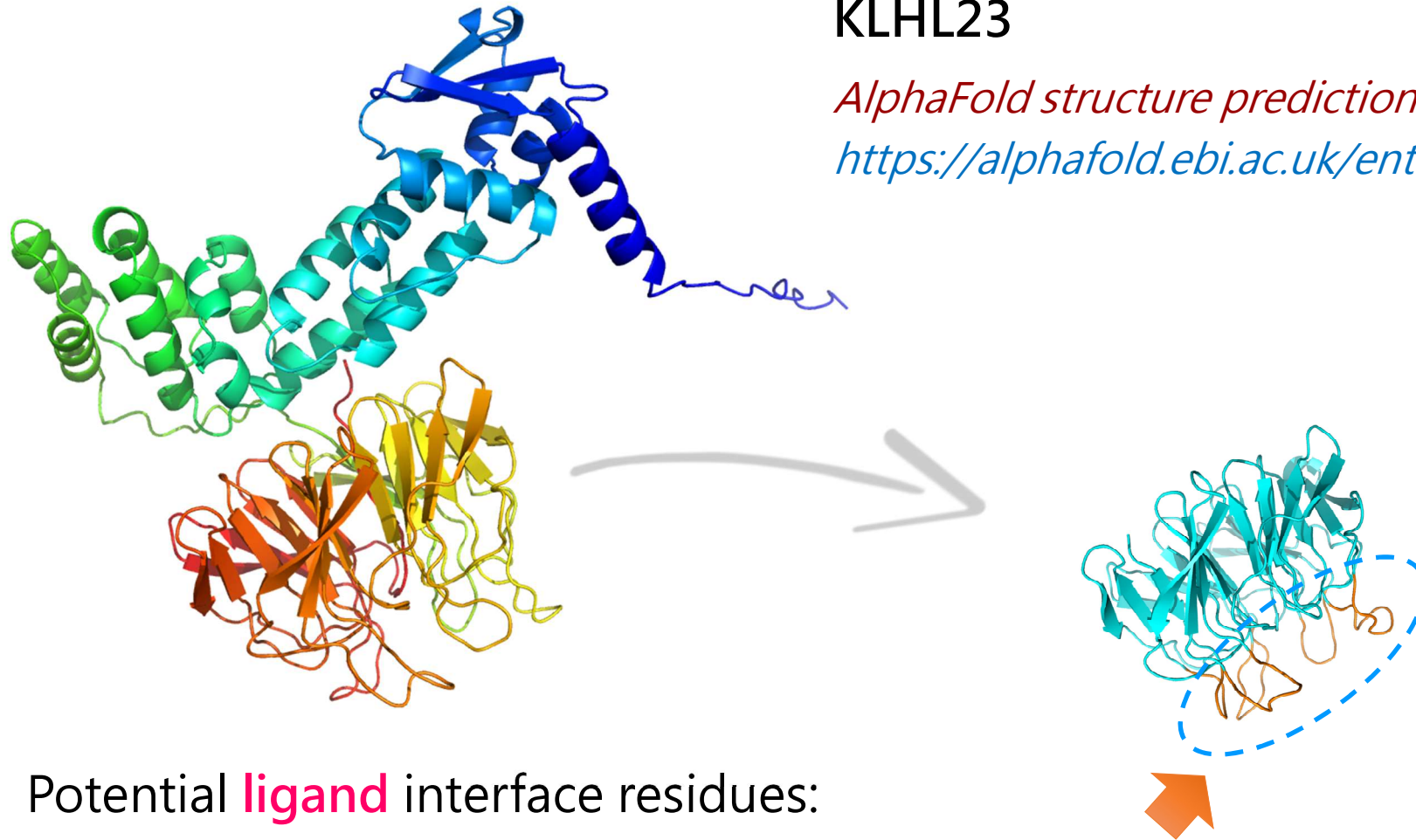
- Potential **ligand** interface residues:

  281-286:B,328-336:B,377-383:B,424-433:B,516-523:B

ColabFold

Making protein folding accessible to all

ColabFold

❖ ColabFold offers accelerated prediction of protein structures and complexes by combining the fast homology search of MMseqs2 with AlphaFold2 or RoseTTAFold.

Many-against-Many sequence searching
多對多序列搜尋

❖ ColabFold's 40−60-fold faster search and optimized model utilization enables prediction of close to 1,000 structures per day on a server with one graphics processing unit.

❖ Coupled with Google Colaboratory, ColabFold becomes a free and accessible platform for protein folding.

❖ ColabFold is open-source software available at https://github.com/sokrypton/ColabFold and its novel environmental databases are available at https://colabfold.mmseqs.com.

ColabFold

.gitignore	gitignore pytestemp	7 months ago
★ AlphaFold2.ipynb	Update news that server migration and db update is comp...	4 days ago
★ AlphaFold2_complexes.ipynb	Update license from CC-BY-NC to CC-BY of AlphaFold2_co...	6 months ago
Contributing.md	Update Contributing.md	8 months ago
LICENSE	Create LICENSE	12 months ago
README.md	Update README.md	4 days ago
★ RoseTTAFold.ipynb	adding link to pyrosetta version	2 months ago
colabfold_search.sh	Add message to colabfold_search.sh to avoid usage	5 days ago

https://github.com/sokrypton/ColabFold

ColabFold/AlphaFold2.ipynb

https://github.com/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

Input protein sequence,
then hit **Runtime -> Run all**

+ 程式碼 + 文字

顯示程式碼

```
adding: config.json (deflated 48%)
adding: nsp6_80823_predicted_aligned_error_v1.json (deflated 89%)
adding: nsp6_80823_unrelaxed_rank_1_model_3_scores.json (deflated 70%)
adding: nsp6_80823_unrelaxed_rank_2_model_5_scores.json (deflated 70%)
adding: nsp6_80823_unrelaxed_rank_3_model_4_scores.json (deflated 69%)
adding: nsp6_80823_unrelaxed_rank_4_model_1_scores.json (deflated 69%)
adding: nsp6_80823_unrelaxed_rank_5_model_2_scores.json (deflated 70%)
adding: nsp6_80823.a3m (deflated 75%)
adding: nsp6_80823_relaxed_rank_1_model_3.pdb (deflated 78%)
adding: nsp6_80823_relaxed_rank_2_model_5.pdb (deflated 78%)
adding: nsp6_80823_relaxed_rank_3_model_4.pdb (deflated 78%)
adding: nsp6_80823_relaxed_rank_4_model_1.pdb (deflated 78%)
adding: nsp6_80823_relaxed_rank_5_model_2.pdb (deflated 78%)
adding: nsp6_80823_unrelaxed_rank_1_model_3.pdb (deflated 78%)
adding: nsp6_80823_unrelaxed_rank_2_model_5.pdb (deflated 78%)
adding: nsp6_80823_unrelaxed_rank_3_model_4.pdb (deflated 78%)
adding: nsp6_80823_unrelaxed_rank_4_model_1.pdb (deflated 78%)
adding: nsp6_80823_unrelaxed_rank_5_model_2.pdb (deflated 78%)
adding: cite.bibtex (deflated 52%)
adding: nsp6_80823_coverage.png (deflated 8%)
adding: nsp6_80823_PAE.png (deflated 1%)
adding: nsp6_80823_plddt.png (deflated 4%)
2022-07-17 14:12:46,731 file_cache is unavailable when using oauth2client >= 4.0.0 or google-auth
```
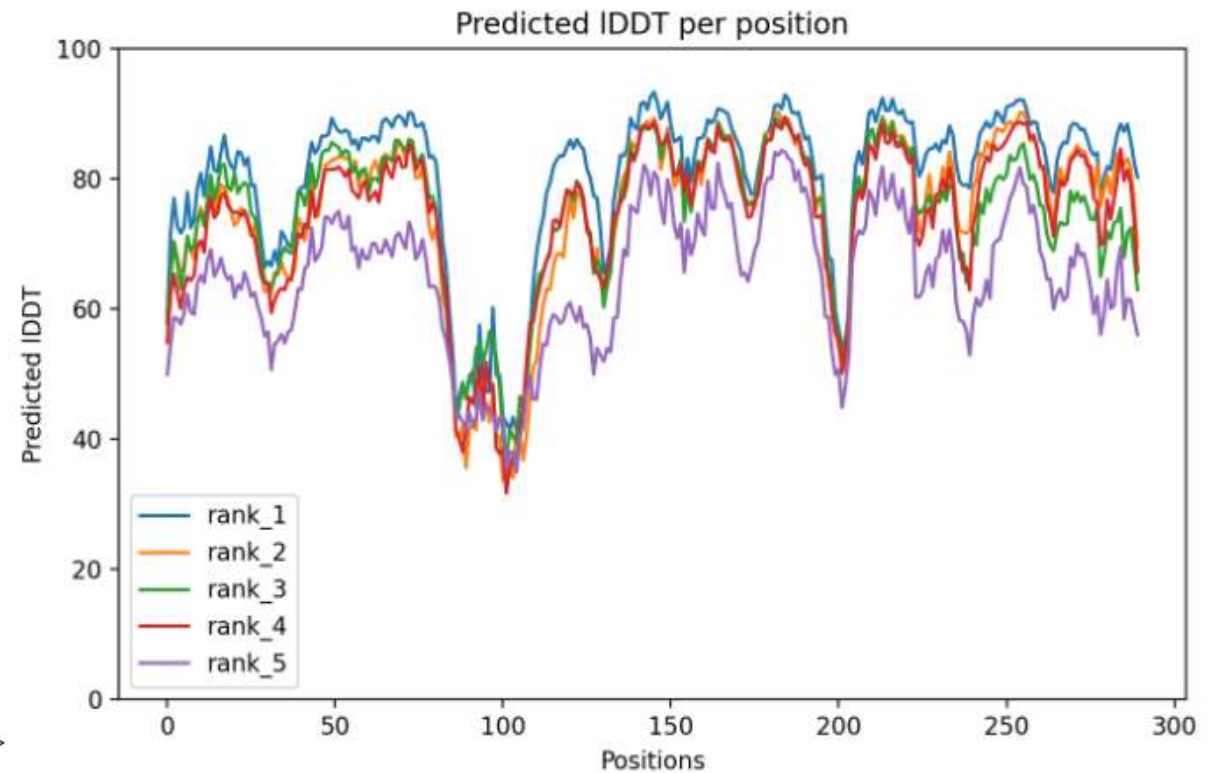
nsp6-AlphaFold2.ipynb

nsp6_80823_relaxed_rank_1_model_3.pdb

Model Confidence:

Very high (pLDDT > 90)

Confident (90 > pLDDT > 70)

Low (70 > pLDDT > 50)

Very low (pLDDT < 50)

Predicted lDDT per position

名稱	修改日期	類型	大小
cite.bibtex	2022/7/17 下午 09:42	BIBTEX 檔案	4 KB
config.json	2022/7/17 下午 09:42	JSON 檔案	1 KB
nsp6_80823.a3m	2022/7/17 下午 09:43	A3M 檔案	93 KB
nsp6_80823_coverage.png	2022/7/17 下午 10:12	PNG 檔案	104 KB
nsp6_80823_PAE.png	2022/7/17 下午 10:12	PNG 檔案	587 KB
nsp6_80823_plddt.png	2022/7/17 下午 10:12	PNG 檔案	221 KB
nsp6_80823_predicted_aligned_error_v1.json	2022/7/17 下午 10:12	JSON 檔案	959 KB
nsp6_80823_relaxed_rank_1_model_3.pdb	2022/7/17 下午 10:12	Protein Data Bank...	373 KB
nsp6_80823_relaxed_rank_2_model_5.pdb	2022/7/17 下午 10:12	Protein Data Bank...	373 KB
nsp6_80823_relaxed_rank_3_model_4.pdb	2022/7/17 下午 10:12	Protein Data Bank...	373 KB
nsp6_80823_relaxed_rank_4_model_1.pdb	2022/7/17 下午 10:12	Protein Data Bank...	373 KB
nsp6_80823_relaxed_rank_5_model_2.pdb	2022/7/17 下午 10:12	Protein Data Bank...	373 KB
nsp6_80823_unrelaxed_rank_1_model_3.pdb	2022/7/17 下午 10:12	Protein Data Bank...	184 KB
nsp6_80823_unrelaxed_rank_1_model_3_scores.json	2022/7/17 下午 10:12	JSON 檔案	512 KB
nsp6_80823_unrelaxed_rank_2_model_5.pdb	2022/7/17 下午 10:12	Protein Data Bank...	184 KB
nsp6_80823_unrelaxed_rank_2_model_5_scores.json	2022/7/17 下午 10:12	JSON 檔案	517 KB
nsp6_80823_unrelaxed_rank_3_model_4.pdb	2022/7/17 下午 10:12	Protein Data Bank...	184 KB
nsp6_80823_unrelaxed_rank_3_model_4_scores.json	2022/7/17 下午 10:12	JSON 檔案	522 KB
nsp6_80823_unrelaxed_rank_4_model_1.pdb	2022/7/17 下午 10:12	Protein Data Bank...	184 KB
nsp6_80823_unrelaxed_rank_4_model_1_scores.json	2022/7/17 下午 10:12	JSON 檔案	521 KB
nsp6_80823_unrelaxed_rank_5_model_2.pdb	2022/7/17 下午 10:12	Protein Data Bank...	184 KB
nsp6_80823_unrelaxed_rank_5_model_2_scores.json	2022/7/17 下午 10:12	JSON 檔案	538 KB

nsp6_80823.result.zip

nsp6_80823.result

Thank You

Yi-Chung Liu

2022/07/19