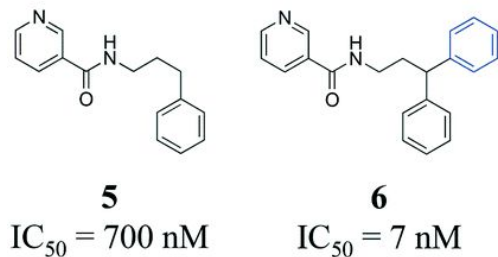
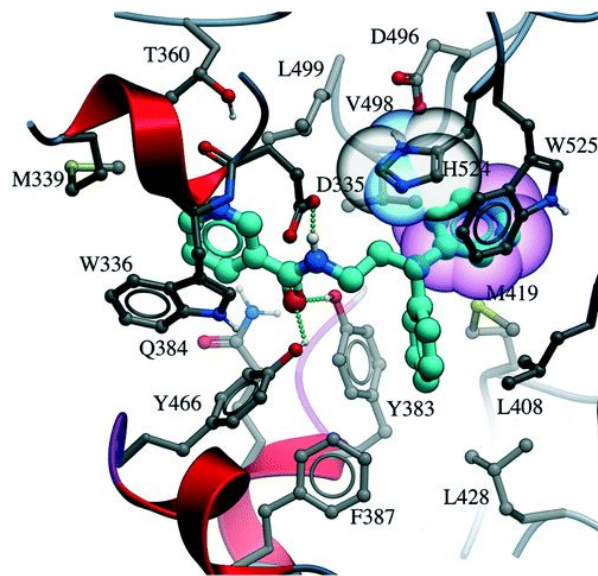


# AMIDD 2023 Lecture 8: Protein-ligand binding



(a)



(b)

**Dr. Jitao David Zhang, Computational Biologist**

<sup>1</sup> *Pharmaceutical Sciences, Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche*

<sup>2</sup> *Department of Mathematics and Informatics, University of Basel*

# Today's goals

1. Biological sequence analysis is fundamental to characterize protein functions.
2. Target-based drug discovery is about to find and make molecules for specific and high-affinity protein-ligand interactions.
3. Basic concepts of structure-based and ligand-based drug design

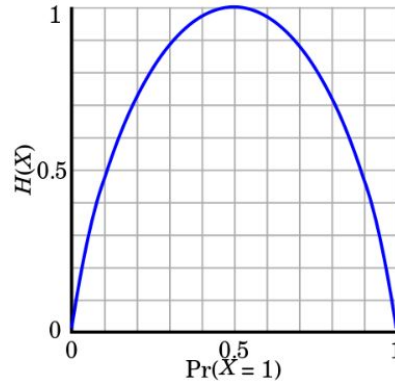
# WebLogo: a transition from the deterministic view to a probabilistic view

```

malEpKp4  GACCTCGGTT
crp       GAAGGCGACC
cytR     CGATGCGAGG
fur       AAATGTAAGC
araB2    TGCCGTGATT
ompR     TAACGTGATC
glpACB   TTGTTTGATT
rot      AGAGGTGATT
cya      AGGTGTTAAA
rhaS     AATGTGAAC
glpFK    TTTTATGACG
cdd      ATTTGCGATG
tdcA     ATTTGTGAGT
deoP2    TTATTTGAAC
nupG1    TTATTTGCCA
crp      TAATGTGACG
aldB     ATTCGTGATA
malEpKp1 TTGTTGTGATC
nag      TTTTGTGAGT
malEpKp3 TTTT6CAAGC
malEpKp2 TAATGTGGAG
dadAX    AGATGTGATT
gut      TTTT6CGATC
glpFK    AAGTTCGATA
dadAX    AGATGTGAGC
lac      TAATGTGAGT
cdd      TAATGAGATT
mtl      TCTTGTGATT
cytR     AAATTCAATA
p1nACB  AAACGTGATT
  
```

Aligned sequences

Information entropy of a Bernoulli trial

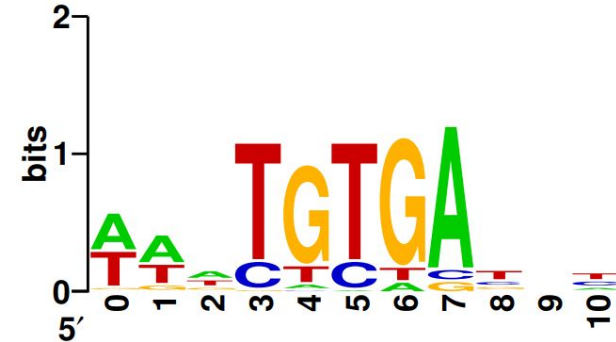


[Brona, Wikimedia](#), shared with CC BY-SA 3.0

$$R_{seq} = S_{max} - S_{obs} = \log_2 N - \left( - \sum_{n=1}^N p_n \log_2 p_n \right)$$

Conservation per site defined as difference between maximal and observed information

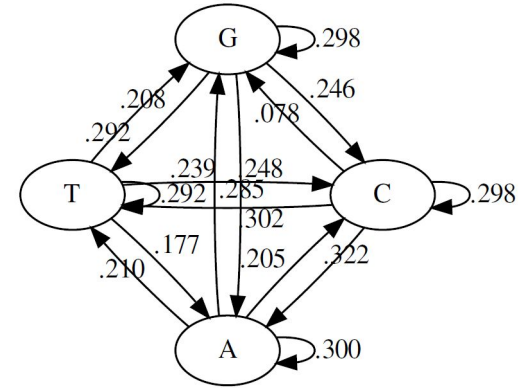
1. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18, 6097–6100 (1990).
1. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: A Sequence Logo Generator. *Genome Res.* 14, 1188–1190 (2004).



WebLogo

# A probabilistic view of biological sequence analysis with Markov chains

- A **discrete-time** Markov chain is a sequence of random variables with the Markov property, namely that the probability of moving to the next state depends only on the present state and not on the previous states.
- A Markov chain is often represented by either a **directed graph** or a **transition matrix**.



## Applications

- Given a string, assuming that the Markov chain model is suitable, we can construct a Markov chain, for instance by counting transitions and normalize the count matrix.
- Given a Markov chain model and a string, we can calculate the probability that the string is generated by the specific model with the **chain rule of conditional probability**.

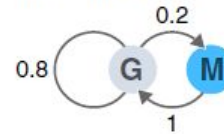
	A	C	G	T
A	.300	.205	.285	.210
C	.322	.298	.078	.302
G	.248	.246	.298	.208
T	.177	.239	.292	.292

Data: <https://web.stanford.edu/class/stats366/exs/HMM1.html>

# Stationary distribution exist for ergodic (irreducible and aperiodic) Markov Chains

- A Markov Chain has stationary  $n$ -step transition probabilities, which are the  $n$ th power of the one-step transition probabilities. Namely,  $P_n = P^n$ .
- A stationary distribution  $\pi$  is a row vector whose entries are non-negative and sum to 1. It is unchanged by the operation matrix  $P$  on it, and is defined by  $\pi P = \pi$ .
  - Note that it has the form of the left eigenvector equation,  $uA = \kappa u$ , where  $\kappa$  is a scalar and  $u$  is a row vector. In fact,  $\pi$  is a normalised (sum to 1) multiple of a left eigenvector  $e$  of the transition matrix  $P$  with an eigenvalue of 1.

**a** Two-state model

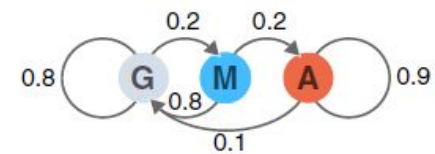


$$T = \begin{bmatrix} p_{GG} & p_{GM} \\ p_{MG} & p_{MM} \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 \\ 1 & 0 \end{bmatrix}$$

$$T^2 = \begin{bmatrix} 0.84 & 0.16 \\ 0.80 & 0.20 \end{bmatrix}$$

$$T^4 = \begin{bmatrix} 0.83 & 0.17 \\ 0.83 & 0.17 \end{bmatrix} \approx \lim_{n \rightarrow \infty} T^n$$

**b** Three-state model



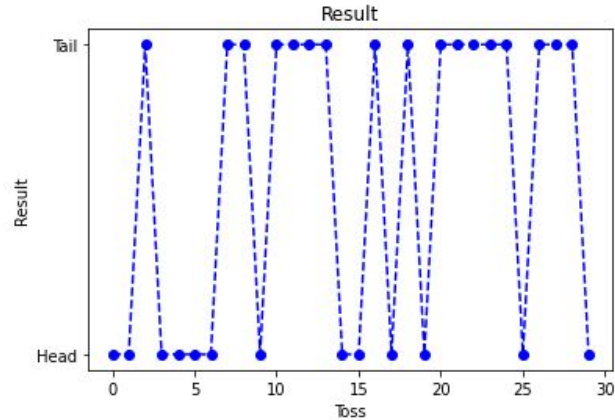
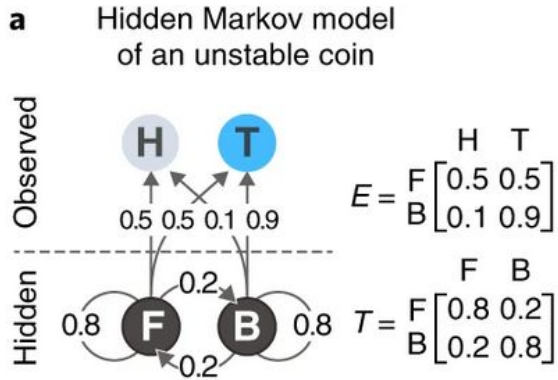
$$T = \begin{bmatrix} p_{GG} & p_{GM} & p_{GA} \\ p_{MG} & p_{MM} & p_{MA} \\ p_{AG} & p_{AM} & p_{AA} \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 & 0 \\ 0.8 & 0 & 0.2 \\ 0.1 & 0 & 0.9 \end{bmatrix}$$

$$T^{50} = \begin{bmatrix} 0.625 & 0.125 & 0.25 \\ 0.625 & 0.125 & 0.25 \\ 0.625 & 0.125 & 0.25 \end{bmatrix} \approx \lim_{n \rightarrow \infty} T^n$$

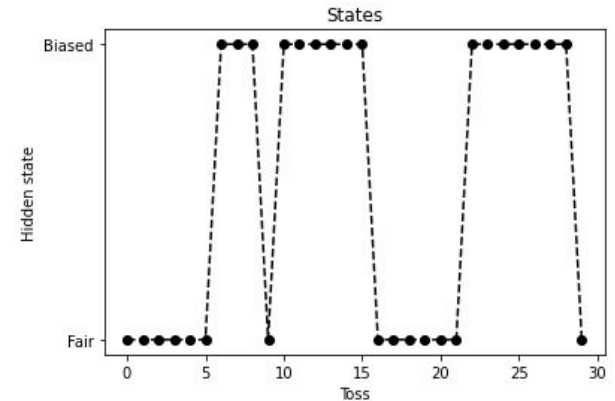
G=Growth, M=Mitosis, A=Arrest

Grewal, Jasleen K., Martin Krzywinski, and Naomi Altman. 2019. "[Markov Models—Markov Chains](#)." Nature Methods 16 (8): 663–64.

# Hidden Markov Models model hidden states based on observations



```
[0 0 1 0 0 0 0 1 1 0
 1 1 1 1 0 0 1 0 1 0
 1 1 1 1 1 0 1 1 1 0]
0=Head, T=Tail
```



```
[0 0 0 0 0 0 1 1 1 0 1 1
 1 1 1 1 0 0 0 0 0 0 1 1 1
 1 1 1 1 0]
0=Fair, 1=Biased
```

A Hidden Markov Model of an unstable coin that has a 20% chance of switching between a fair state (F) and a biased state (B). Source: Grewal, Jasleen K., Martin Krzywinski, and Naomi Altman. 2019. "[Markov Models — Hidden Markov Models.](#)" Nature Methods 16 (9): 795–96.

# The Viterbi algorithm estimates transmission and emission matrices

A Hidden Markov Model consists of two graphs (matrices): one of **hidden states**, which corresponds to the **transmission matrix**, and one of **observed states**, which corresponds to the **emission matrix**.

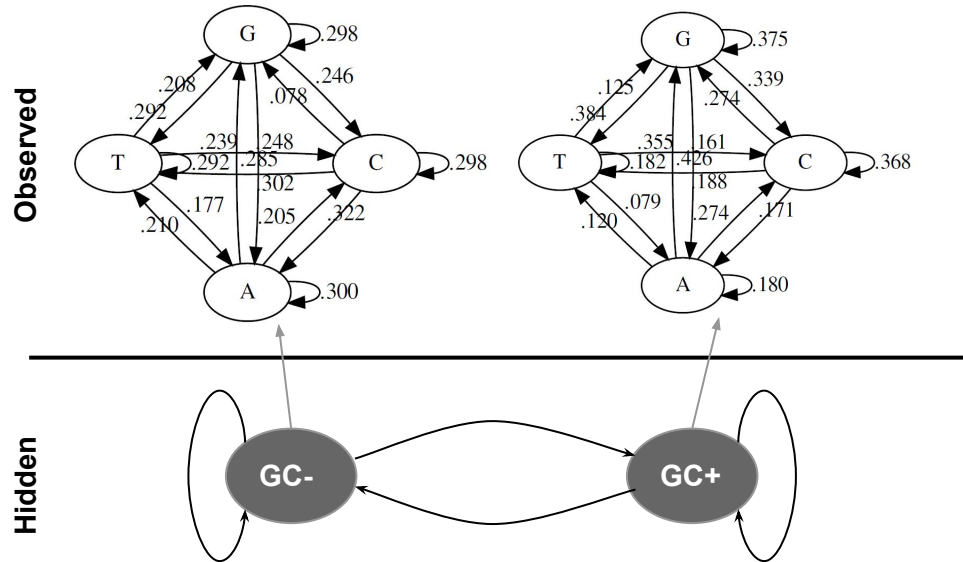


Illustration of a Hidden Markov Model predicting CpG islands in genomic sequences

The **Viterbi algorithm** (based on dynamic programming), or the **Baum-Welch algorithm** (a special case of EM algorithms) is used to estimate its parameters.

```
Transmission Matrix Generated:
[[0.8 0.2]
 [0.2 0.8]]
```

```
Transmission Matrix Recovered:
[[0.774 0.226]
 [0.104 0.896]]
```

```
Emission Matrix Generated:
[[0.5 0.5]
 [0.1 0.9]]
```

```
Emission Matrix Recovered:
[[0.539 0.461]
 [0.152 0.848]]
```

The transmission and emission matrices estimated by the Viterbi algorithm from 1000 observations generated by the HMM model in the last slide. [Source code](#)

# Profile Hidden Markov models capture evolutionary changes in homologous sequences

**M:** match states. In the match state, the probability distribution is the frequency of the amino acids in that position.

**I:** insert states, which model highly variable regions in the alignment

**D:** delete states, which allows gaps and deletion.

Profile HMMs belongs to **generative models**.

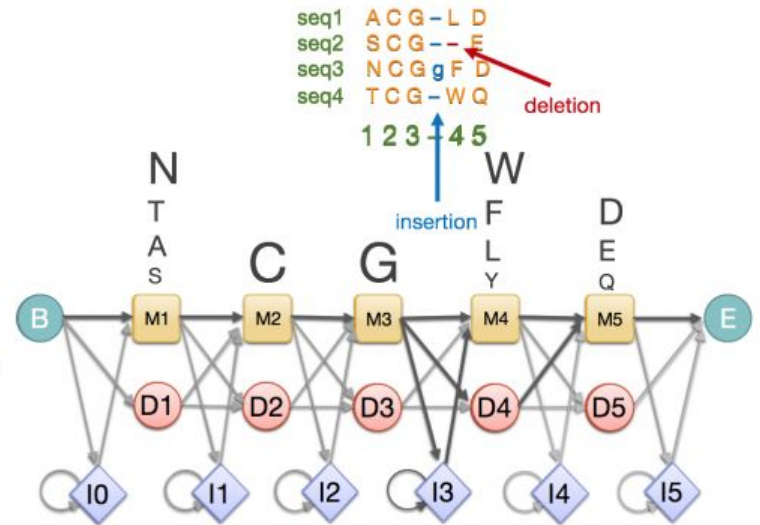
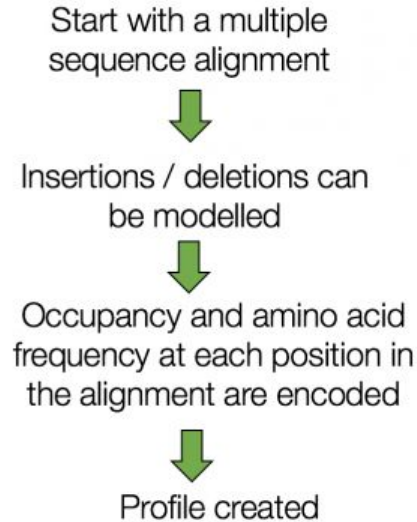
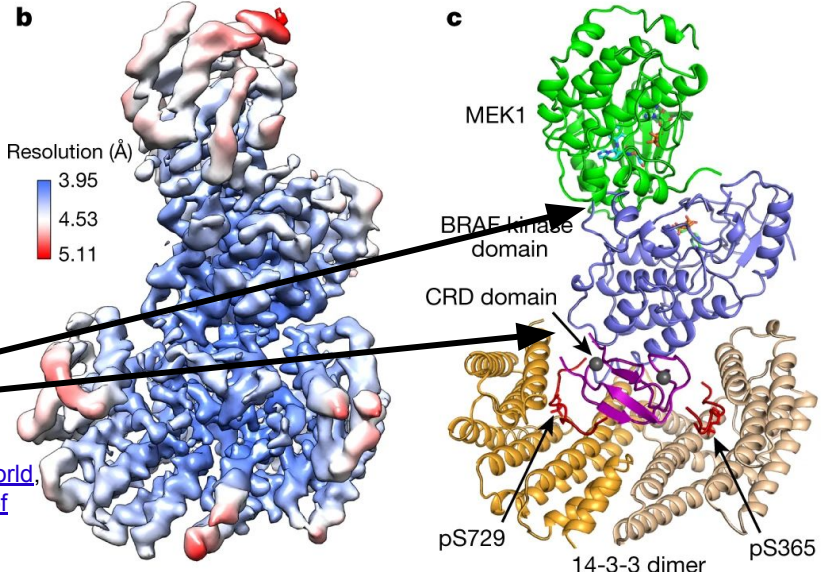
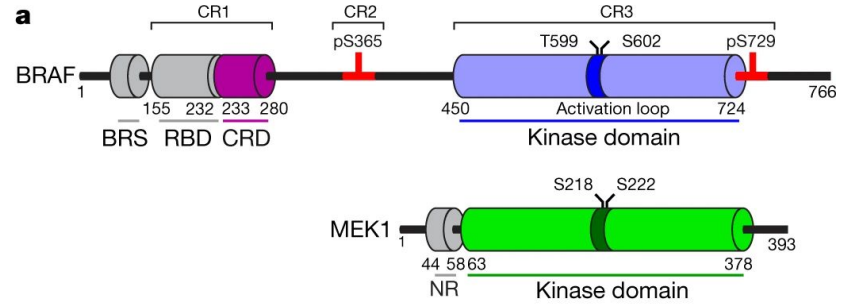
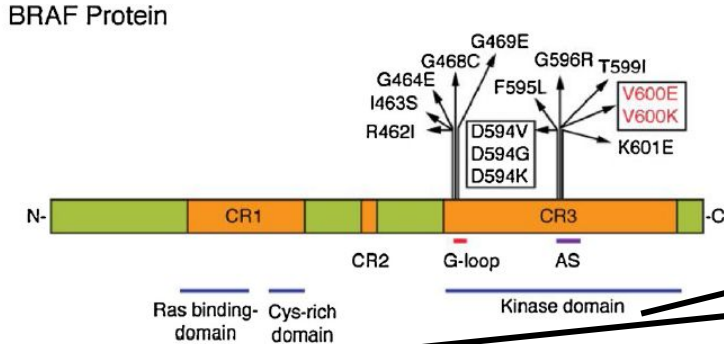
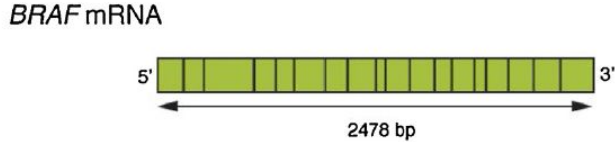
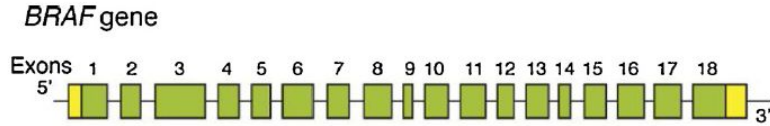


Figure from [Pfam](#)

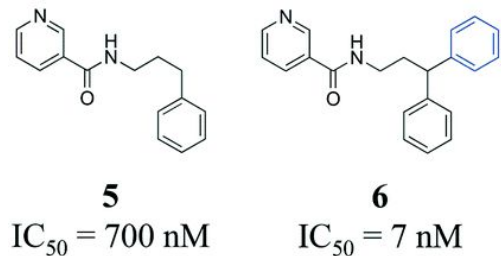


# Protein domains: self-stabilizing and folding independently from the rest

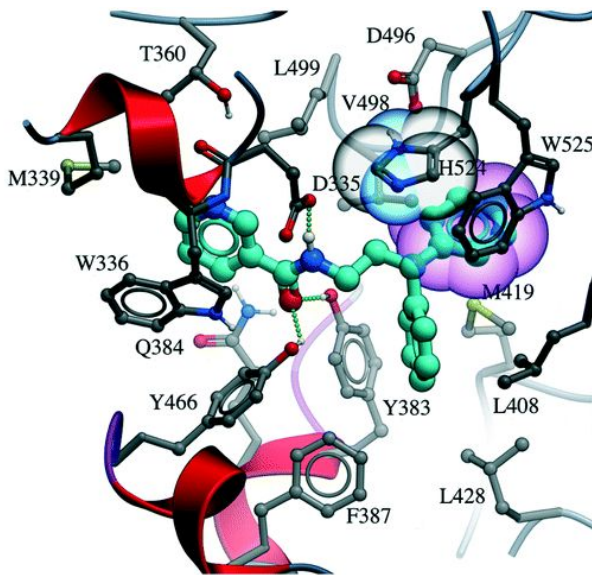


CR=conserved region. CRD=Cys-rich domain  
 Left: Frisone, *et al.*, [A BRAF New World](#), Critical Reviews in Oncology/Hematology (2020);  
 Right: Park, *et al.*, [Architecture of Autoinhibited and Active BRAF-MEK1-14-3-3 Complexes](#), Nature (2019)..

# Goal of target-based drug discovery: to make a molecule that binds specifically and strongly to the target protein domain



(a)



(b)

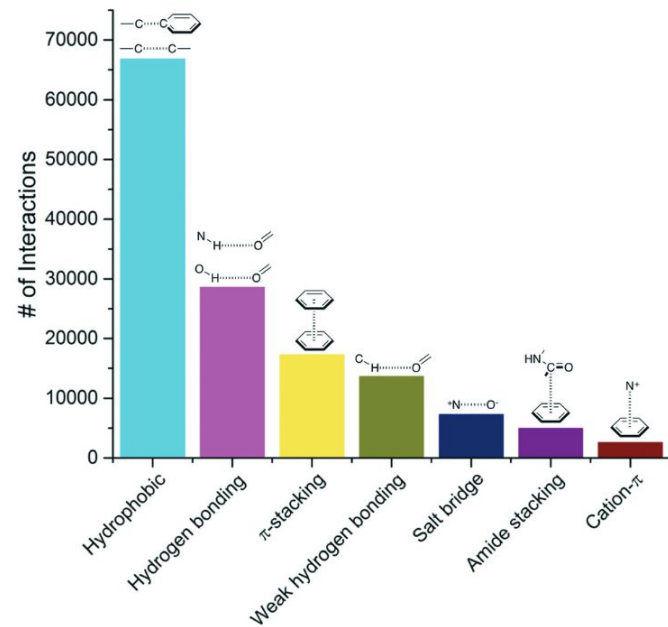


Fig. 1 Frequency distribution of the most common non-covalent interactions observed in protein-ligands extracted from the PDB.

# Protein Data Bank (PDB) contains solved structures of proteins and protein-ligand interactions



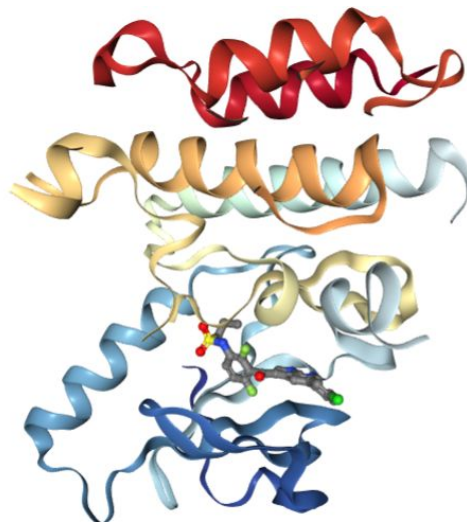
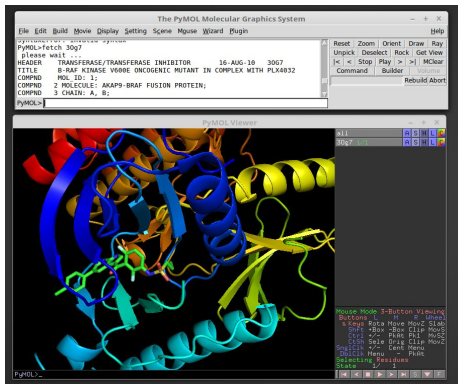
RCSB PDB PROTEIN DATA BANK  
157145 Biological Macromolecular Structures  
Enabling Breakthroughs in Research and Education  
Search  
Advanced

Structure Summary 3D View Annotations Sequence Sequence

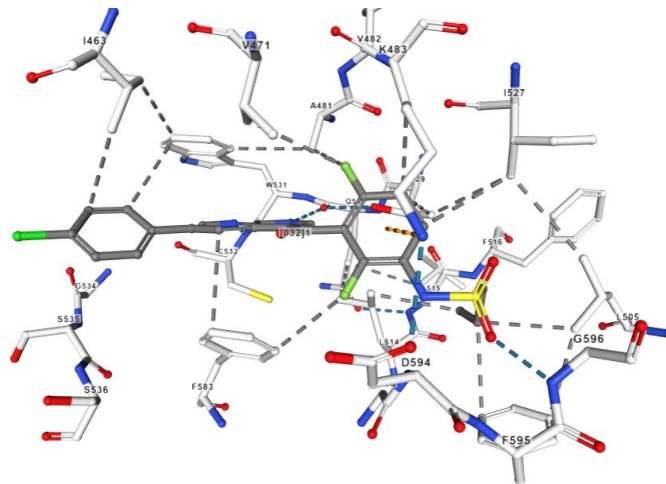
3OG7

B-Raf Kinase V600E oncogenic mutant in complex with PLX4032

<http://www.rcsb.org/3d-view/3OG7>



Structural view



Ligand view

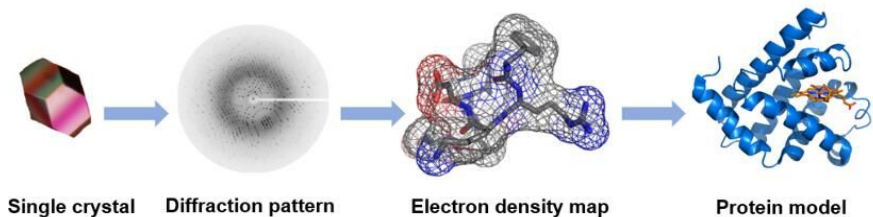
Balls and sticks: protein V600E and ligand (PLX4032)

**Blue dashes:** hydrogen bonds (<3.5 Angstrom)

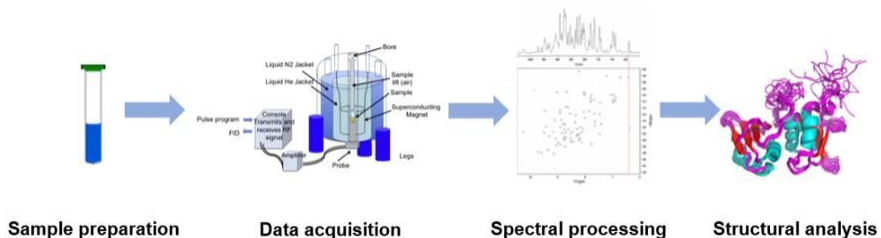
**Gray dashes:** hydrophobic interactions (<4 Angstrom)

Working with PDB files with *PyMol* from the command-line

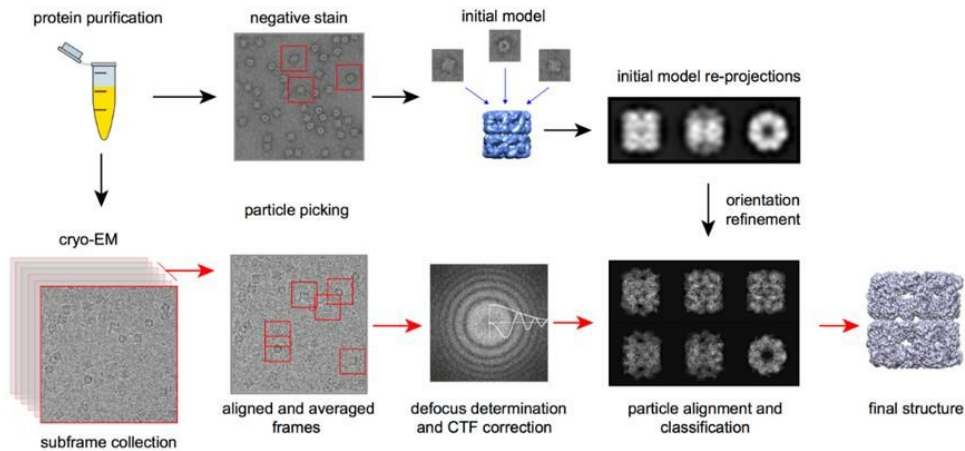
# X-ray, NMR, and CryoEM are major experimental approaches to determining protein structures



**X-ray crystallography**



**Nuclear Magnetic Resonance (NMR)**



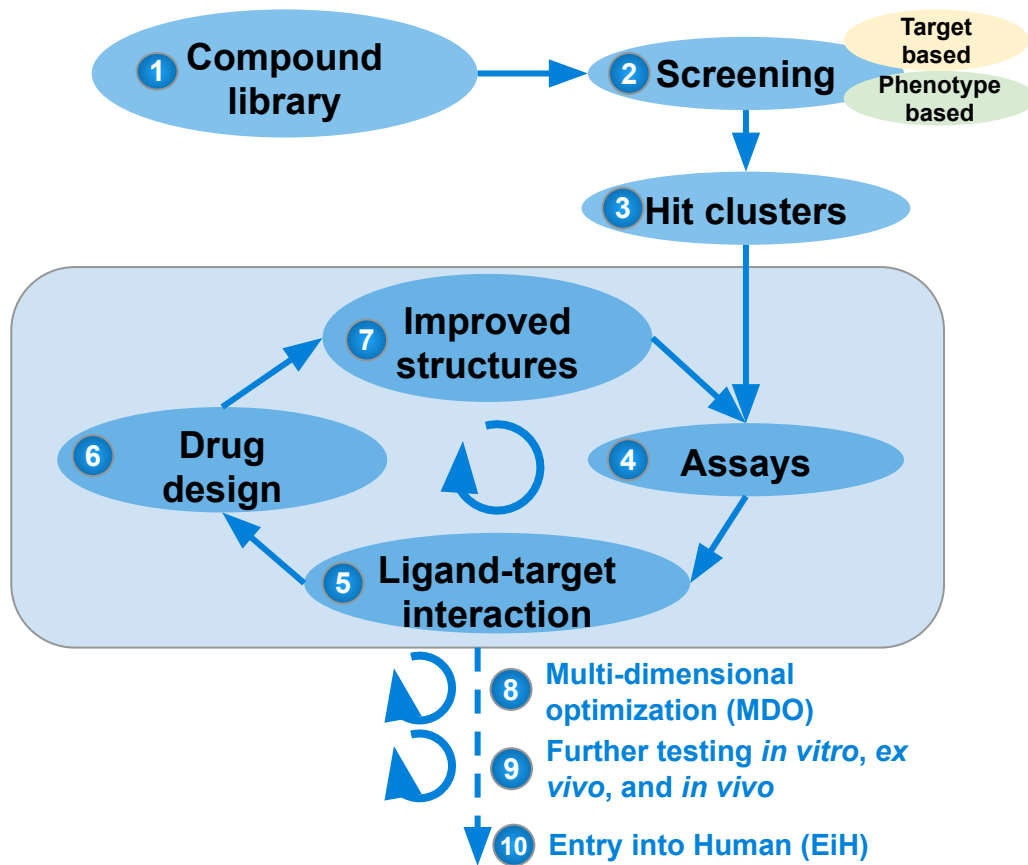
**Cryo-electron microscopy (CryoEM)**

Figure sources:

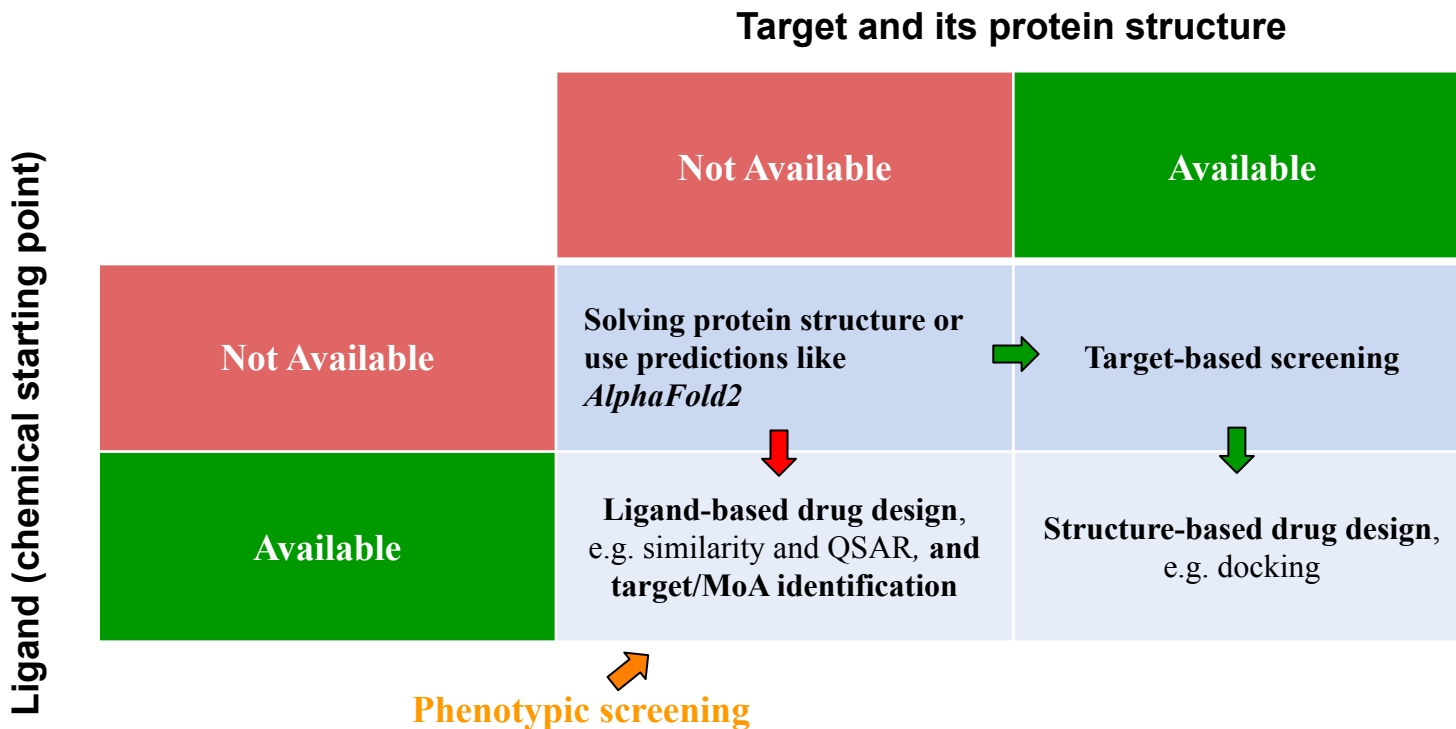
[https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em\\_6.htm](https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm)

# Workflow in a typical target-based drug-discovery program

1. Compound library construction (small molecules, large molecules, RNA therapeutics, or other modalities)
2. Screening compounds with *bioassays*, or *assays*, which determine potency of a chemical by its effect on biological entities: proteins, cells, *etc*;
3. Hit identification and clustering;
4. More assays, complementary to the assays used in the screening, maybe of lower throughput but more biologically relevant;
5. Analysis of ligand-target interactions, for instance by getting the co-structure of both protein (primary target, and off-targets if necessary) and the hit;
6. *Drug design*, namely to modify the structure of the drug candidate;
7. Analog synthesis and testing (back to step 4);
8. Multidimensional Optimization (MDO), with the goal to optimize potency, selectivity, safety, bioavailability, *etc*;
9. Further *in vitro*, *ex vivo*, and *in vivo* testing, and preclinical development;
10. Entry into human (Phase 0 or phase 1 clinical trial).

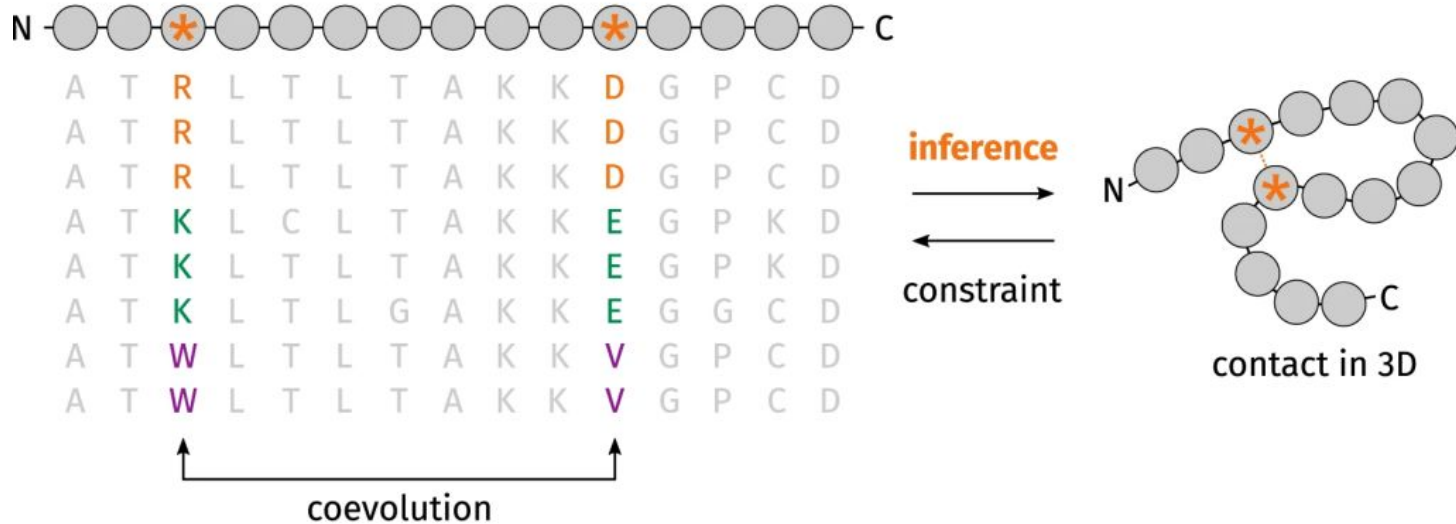


# Ligand-based and structure-based drug design



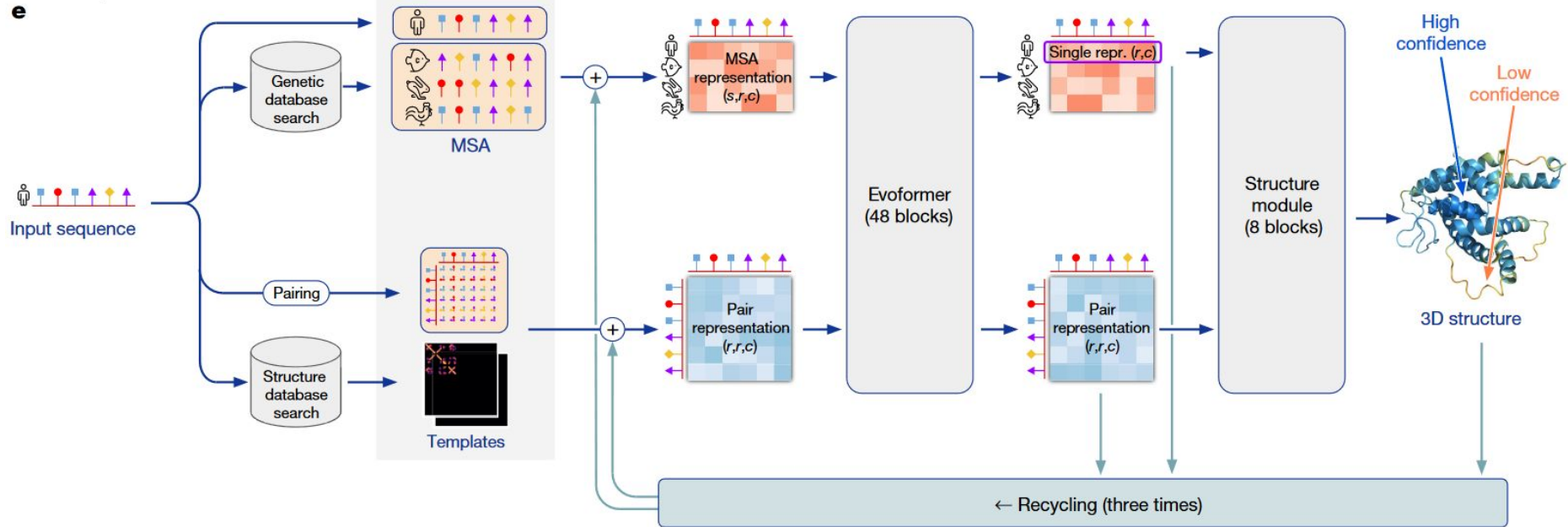
QSAR= quantitative structure activity relationship; MoA= mechanism of action, or mode of action

# One of the key ideas of AlphaFold2: learning from evolutionary constraints



Marks, Debora S., Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. "Protein 3D Structure Computed from Evolutionary Sequence Variation." PLOS ONE 6, no. 12 (December 7, 2011): e28766. <https://doi.org/10.1371/journal.pone.0028766>.

# AlphaFold2 uses co-evolution of residues, determined structures, and neural networks to achieve the high performance



Jumpe et al. "Highly Accurate Protein Structure Prediction with AlphaFold." Nature 596, no. 7873 (August 2021): 583–89.

<https://doi.org/10.1038/s41586-021-03819-2>. A blog post that explains how AlphaFold2 works: [blogiq.com](https://blogiq.com)



# ChEMBL as information source of small molecules

## Nomenclature

*caffeine*  
*1,3,7-trimethylxanthine*  
*methyltheobromine*

## Bioactivity

*Affinity to human  
 proteins and drug  
 targets*

## Chemical data

*Formula: C<sub>8</sub>H<sub>10</sub>N<sub>4</sub>O<sub>2</sub>*  
*Charge: 0*  
*Mass: 194.19*

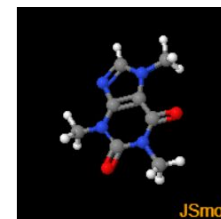
## Database Xrefs

*PubChem: CID2519*  
*BindingDB: 1849*

## Chemical Informatics

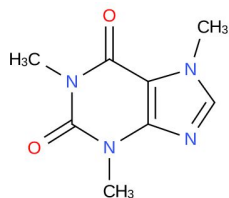
*InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)  
 12(3)8(14)11(6)2/h4H,1-3H3*  
*SMILES: CN1C(=O)N(C)c2ncn(C)c2C1=O*

## Visualisation



A subset of available information from EBI ChEBI/ChEMBL,  
 inspired by EBI's roadshow *Small Molecules in Bioinformatics*

# Representation of small molecules



**Molfile:** [View Raw](#) [Download](#) [Editor](#) [Copy](#)

**Canonical SMILES:** CN1C(=O)N(C)c2ncn(C)c2C1=O

**Standard InChI:** InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H, 1-3H3

**Standard InChI Key:** RYYVLZVUVIJVGH-UHFFFAOYSA-N

CHEMBL113

SciTegic12231509382D

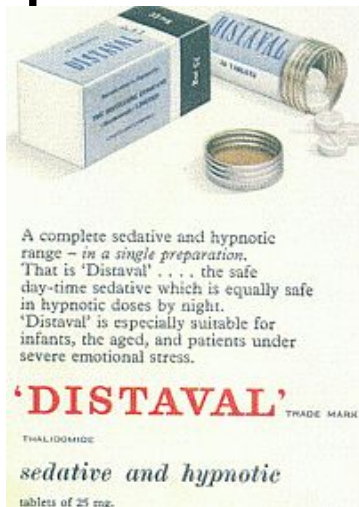
```

14 15 0 0 0 0 999 V2000
-1.1875 -9.6542 0.0000 C 0 0
-1.1875 -8.9625 0.0000 C 0 0
-1.8125 -10.0292 0.0000 N 0 0
-2.4167 -8.9625 0.0000 N 0 0
-2.4167 -9.6542 0.0000 C 0 0
-1.8125 -8.6000 0.0000 C 0 0
-0.5000 -9.8917 0.0000 N 0 0
-0.5000 -8.7625 0.0000 N 0 0
-0.1125 -9.3042 0.0000 C 0 0
-3.0250 -10.0375 0.0000 O 0 0
-1.8125 -7.8917 0.0000 O 0 0
-1.8125 -10.7417 0.0000 C 0 0
-3.0250 -8.6000 0.0000 C 0 0
-0.2917 -8.0750 0.0000 C 0 0
2 1 2 0
3 1 1 0
4 5 1 0
5 3 1 0
6 2 1 0
7 1 1 0
8 2 1 0
9 7 2 0
10 5 2 0
11 6 2 0
12 3 1 0
13 4 1 0

```

- Simplified Molecular-Input Line-Entry System (SMILES)
- IUPAC International Chemical Identifier (InChI)
- InChiKey: a 27-character, hash version of InChI
- Molfile: a type of [chemical table files](#)

# The tragedy of thalidomide and the importance of representation



Frances Oldham Kelsey received the President's Award for Distinguished Federal Civilian Service from President John F. Kennedy, 1962

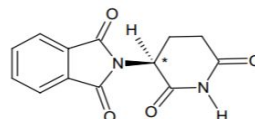
## Canonic SMILES of thalidomide

C1CC(=O)NC(=O)C1N2C(=O)C3=CC=CC=C3C2=O



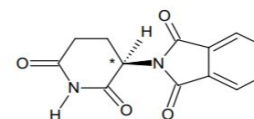
(1957)

I thank Manuela Jacklin for her help preparing this slide.



(-)(S)-thalidomide

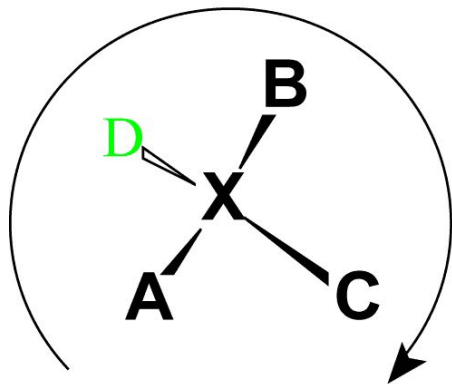
Isomeric SMILES of (-)(S)-thalidomide  
C1CC(=O)NC(=O)[C@H]1N2C(=O)C3=CC=CC=C3C2=O



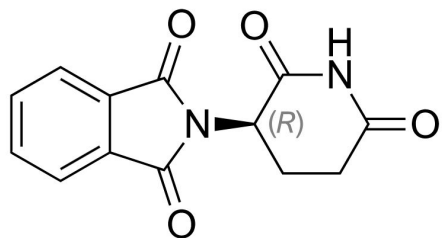
(+)(R)-thalidomide

Isomeric SMILES of (+)(R)-thalidomide  
C1CC(=O)NC(=O)[C@@H]1N2C(=O)C3=CC=CC=C3C2=O

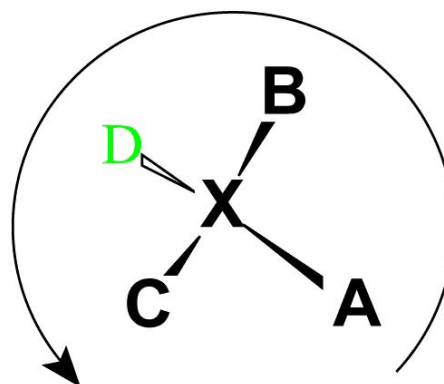
# Absolute configuration of atoms within a chiral molecule



**R**

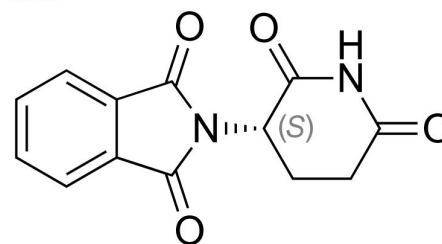


Sedative  
Non-teratogenic



**S**

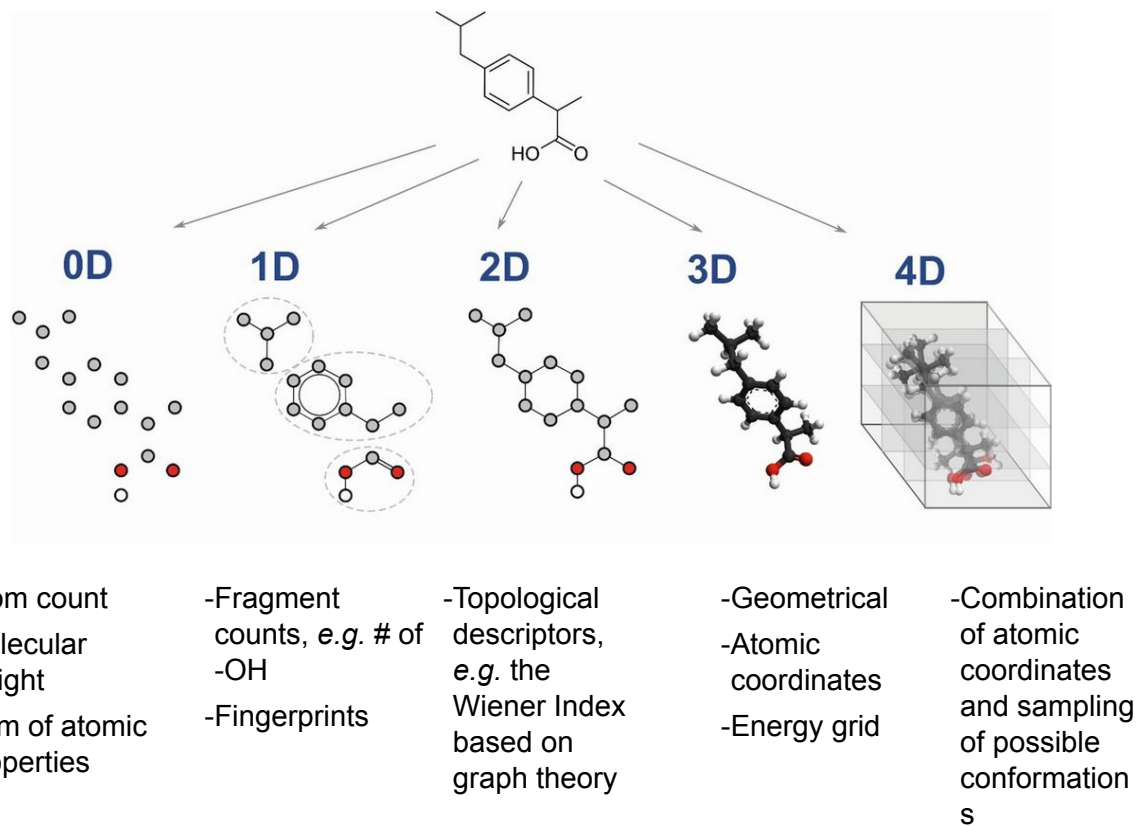
Embryo-toxic  
Teratogenic



# Molecular descriptors: numeric values that describe chemical molecules.

In contrast to symbolic representations, molecular descriptors enable **quantification of molecular properties.**

Molecular descriptors allows mathematical operations and statistical analysis that associate biophysical or biochemical properties with molecule structures.



# Conclusions

- **Sequence analysis is fundamental for many tasks in drug discovery.**
- **Target-based drug discovery is about to find specific interactions between a new drug molecule and its primary protein target.**
- **Small molecules can be presented by molecular descriptors in order to be analyzed with mathematical and statistical tools.**

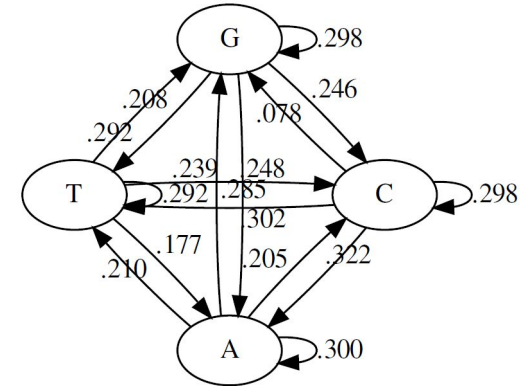
# Offline exercises for lecture 8

Please submit your results to [the Google Form](#).

- Compare  $p(\text{ACGTGGT}|\text{M})$  and  $p(\text{ACCTGGT}|\text{M})$ , where M stands for the model on the right side. Report the ratio of the two values below. For instance,  $p(\text{ACG})/p(\text{ACC})=p(\text{A})p(\text{C}|\text{A})p(\text{G}|\text{C})/(p(\text{A})p(\text{C}|\text{A})p(\text{C}|\text{C}))=p(\text{G}|\text{C})/p(\text{C}|\text{C})=0.078/0.298=0.261$ . Note that the pipe means 'given' or 'conditional on'.
- We have got a RNA sequence by sequencing sputum from a patient (see below). How can we know the original genome of the sequence, and ideally the gene encoding the sequences? Tips: go to the NCBI BLAST tool ([https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE\\_TYPE=BlastSearch&LINK\\_LOC=blasthome](https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome)), copy and paste the sequence as the query sequence, and try your luck. Default parameters are okay.

ATGTTTGT TTTTCTTGT TTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACAACCAGA  
 ACTCAATTACCCCTGCATACACTAATTCTTTTCACACGTGGTGT TTTATTACCCTGACAAAGTT  
 TTCAGATCCTCAGT

- Required reading:** Selected pages of *Evaluation of the Biological Activity of Compounds: Techniques and Mechanism of Action Studies* by Dougall and Unitt and answer questions. To answer offline-activity questions, it is required to read pages 15-22 (1-8 of the 29 pages in total, before section '4. Types of Enzyme Inhibition and Their Analysis'), page 27 (section 6A), and pages 34-37 (Assay Biostatistics). The rest is optional reading.



Following alphabet

<i>p-value</i>	Following alphabet			
	A	C	G	T
A	.300	.205	.285	.210
C	.322	.298	.078	.302
G	.248	.246	.298	.208
T	.177	.239	.292	.292

Preceding alphabet

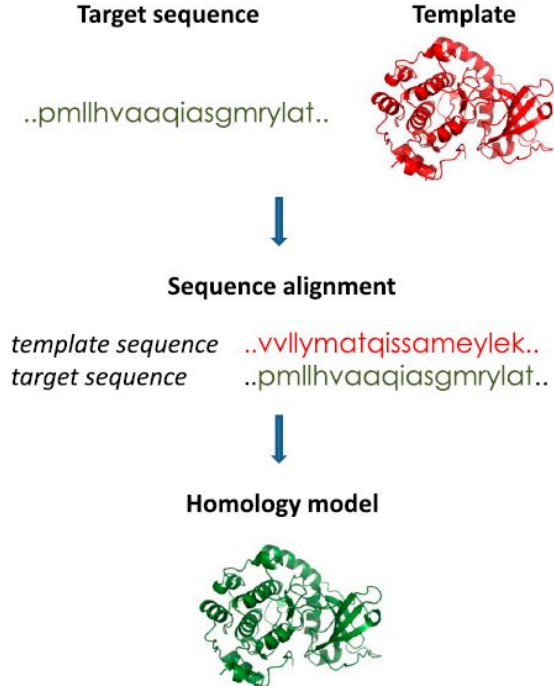
# Backup slides



# Three major experimental approaches to determining protein structures

Method	Underlying physical properties	Main mathematical technique used	Advantages	Limitations
<b>X-ray crystallography</b>	The crystalline structure of a molecule causes a beam of incident X-rays to diffract into many specific directions.	Fourier series and Fourier transform	<ul style="list-style-type: none"> <li>Established</li> <li>Broad molecular weight range</li> <li>High resolution</li> </ul>	<ul style="list-style-type: none"> <li>Crystallization</li> <li>Static model</li> </ul>
<b>Nuclear Magnetic Resonance (NMR)</b>	Nuclei with odd number of protons and/or neutrons in a strong constant magnetic field, when perturbed by a weak oscillating magnetic field, produce an electromagnetic signal with a frequency characteristic of the magnetic field at the nucleus.	Distance geometry (the study of matrices of distances between pairs of atoms) of and discrete differential geometry of curves	<ul style="list-style-type: none"> <li>3D structure in solution</li> <li>Dynamic study possible</li> </ul>	<ul style="list-style-type: none"> <li>High sample purity needed</li> <li>Molecular weight limit (~&lt;40-50 kDa)</li> <li>Sample preparation and computational simulation</li> </ul>
<b>Cryo-electron microscopy</b>	An electron microscope using a beam of accelerated electrons (instead of protons) as a source of illumination. Samples are cooled to cryogenic temperatures and embedded in an environment of vitreous water (amorphous ice).	An inverse problem of reconstruction - the estimation of randomly rotated molecule structure from a projection with noise; Fourier transform; iterative refinement	<ul style="list-style-type: none"> <li>Easy sample preparation</li> <li>Native-state structure</li> <li>Small sample size</li> </ul>	<ul style="list-style-type: none"> <li>Costly EM equipment</li> <li>Challenging for small proteins</li> </ul>

# If no structure is available, homology model building and *in silico* prediction may help



W296–W303 *Nucleic Acids Research*, 2018, Vol. 46, Web Server issue  
doi: 10.1093/nar/gky427

Published online 21 May 2018

## SWISS-MODEL: homology modelling of protein structures and complexes

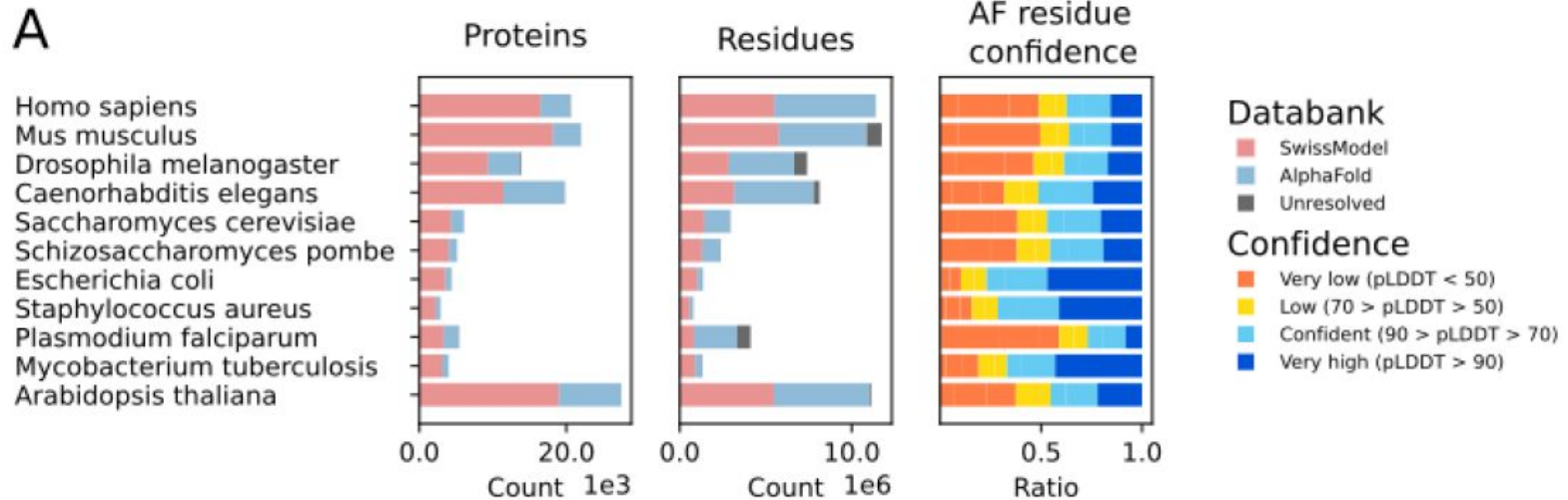
Andrew Waterhouse<sup>1,2,†</sup>, Martino Bertoni<sup>1,2,†</sup>, Stefan Bienert<sup>1,2,†</sup>, Gabriel Studer<sup>1,2,†</sup>, Gerardo Tauriello<sup>1,2,†</sup>, Rafal Gumienny<sup>1,2</sup>, Florian T. Heer<sup>1,2</sup>, Tjaart A. P. de Beer<sup>1,2</sup>, Christine Rempfer<sup>1,2</sup>, Lorenza Bordoli<sup>1,2</sup>, Rosalba Lepore<sup>1,2</sup> and Torsten Schwede<sup>1,2,\*</sup>

<sup>1</sup>Biozentrum, University of Basel, Klingelbergstrasse 50–70, CH-4056 Basel, Switzerland and <sup>2</sup>SIB Swiss Institute of Bioinformatics, Biozentrum, University of Basel, Klingelbergstrasse 50–70, CH-4056 Basel, Switzerland

Received February 09, 2018; Revised May 01, 2018; Editorial Decision May 02, 2018; Accepted May 07, 2018

- **Levinthal's paradox:** *It would take a protein the present age of the universe to explore all possible configurations and find the minimum energy configuration. Yet proteins fold in microseconds.*
- **CASP: Critical Assessment of Techniques for Protein Structure Prediction**
- A thought-provoking blog from Mohammed AlQuraishi: [AlphaFold @ CASP13: "What just happened?"](#), with an informal but good overview of history of protein structure prediction, and his indictment (criminal accusations) of both academia and pharma.
- By 2021 AlphaFold2 and RoseTTAfold reach experiment-level accuracy in some predictions of protein static structure. By 2023 AlphaMissing has been used to predict the consequence of mutations.

# AlphaFold2 & RoseTTAfold extend our understanding of protein biology, while their impact on drug discovery remains to be seen



Akdel, Mehmet, Douglas EV Pires, Eduard Porta-Pardo, Jurgen Janes, Arthur O. Zalevsky, Balint Meszaros, Patrick Bryant, et al. "A Structural Biology Community Assessment of AlphaFold 2 Applications," September 26, 2021. <https://doi.org/10.1101/2021.09.26.461876>.

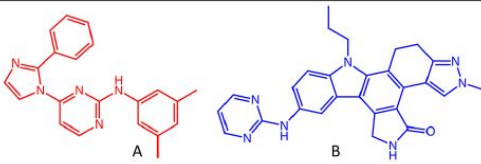
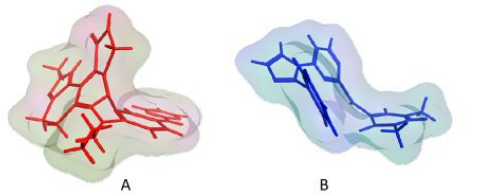
# Brief introduction to AlphaFold (2) and RoseTTAFold

- AlphaFold (available in 2018, relevant research since ~2010s)
  - Key assumption: **a distance map**, created by following the observation that co-evolving amino acids have close physical interactions.
  - Key algorithm: graph neural networks that predict distances between distances, as well as  $\varphi$  (Psi, dihedral angle of the N-C $\alpha$  bond) and  $\psi$  (Phi, C-C $\alpha$  bond) angles for each amino acid. Trained with amino-acid and structural data of 29,000 proteins, with neural network and gradient descent.
- AlphaFold2 (available in 2020)
  - Improving drawback of AlphaFold1, which overwrites interactions between nearby residues over residues further apart.
  - Major changes
    - Transformers that refine a vector representation of each relationship between two amino acids in the protein. Attention mechanism is used to learn from data.
    - A single differentiable end-to-end model instead of modular models
    - Local physicals is applied only at the final refinement step.
- [RoseTTAFold](#) (Science 2021): a three-track network integrating 1D (sequence), 2D (distance), and 3D (coordinate) level information. Possible to model protein-protein complexes. Code and server available.

# Molecular similarity and similarity measures

Chemical similarity	Mol. weight	LogP	Rotatable bonds	Aromatic rings	Heavy atoms
	A	341.4	5.23	4	4
B	463.5	4.43	4	5	35

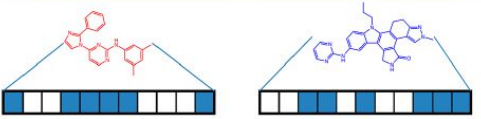
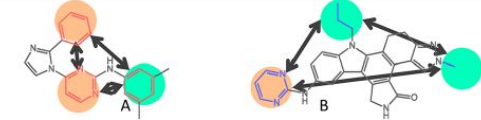
  

Molecular similarity		
2D similarity		
3D similarity		

Biological similarity	Vascular endothelial growth factor receptor 2	Tyrosine-protein kinase TIE-2
	A	active
B	active	active

Global similarity		
Local similarity		

**Table 2 Formulas for the various similarity and distance metrics**

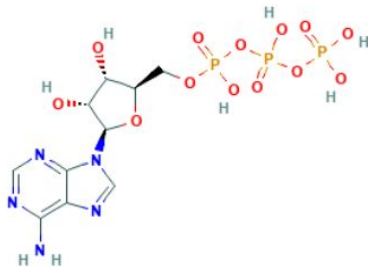
Distance metric	Formula for continuous variables <sup>a</sup>	Formula for dichotomous variables <sup>b</sup>
Manhattan distance	$D_{A,B} = \sum_{j=1}^n  x_{jA} - x_{jB} $	$D_{A,B} = a + b - 2c$
Euclidean distance	$D_{A,B} = \left[ \sum_{j=1}^n (x_{jA} - x_{jB})^2 \right]^{1/2}$	$D_{A,B} = [a + b - 2c]^{1/2}$
Cosine coefficient	$S_{A,B} = \left[ \frac{\sum_{j=1}^n x_{jA} x_{jB}}{\sum_{j=1}^n (x_{jA})^2 \sum_{j=1}^n (x_{jB})^2} \right]^{1/2}$	$S_{A,B} = \frac{c}{[ab]^{1/2}}$
Dice coefficient	$S_{A,B} = \left[ \frac{2 \sum_{j=1}^n x_{jA} x_{jB}}{\sum_{j=1}^n (x_{jA})^2 + \sum_{j=1}^n (x_{jB})^2} \right]$	$S_{A,B} = 2c/[a + b]$
Tanimoto coefficient	$S_{A,B} = \frac{\left[ \sum_{j=1}^n x_{jA} x_{jB} \right]}{\left[ \sum_{j=1}^n (x_{jA})^2 + \sum_{j=1}^n (x_{jB})^2 - \sum_{j=1}^n x_{jA} x_{jB} \right]}$	$S_{A,B} = c/[a + b - c]$
Soergel distance <sup>b</sup>	$D_{A,B} = \left[ \frac{\sum_{j=1}^n  x_{jA} - x_{jB} }{\sum_{j=1}^n \max(x_{jA}, x_{jB})} \right]$	$D_{A,B} = 1 - \frac{c}{[a + b - c]}$

$S$  denotes similarities, while  $D$  denotes distances. The two can be converted to each other by *similarity* = 1/(1+*distance*).  $x_{jA}$  means the  $j$ -th feature of molecule A.  $a$  is the number of *on* bits in molecule A,  $b$  is number of *on* bits in molecule B, while  $c$  is the number of bits that are *on* in both molecules.

(Left) Maggiora, Gerald, Martin Vogt, Dagmar Stumpfe, und Jürgen Bajorath. „[Molecular Similarity in Medicinal Chemistry](#)“. *Journal of Medicinal Chemistry* 57, Nr. 8 (24. April 2014): 3186–3204. (Right) Bajusz, Dávid, Anita Rácz, and Károly Héberger. 2015. “[Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations?](#)” *Journal of Cheminformatics* 7 (1): 20.

# Selected commonly used molecular descriptors

**Molecular Weight (MW).**  
for example, adenosine triphosphate (ATP), the *energy molecule*, has a MW of 507.

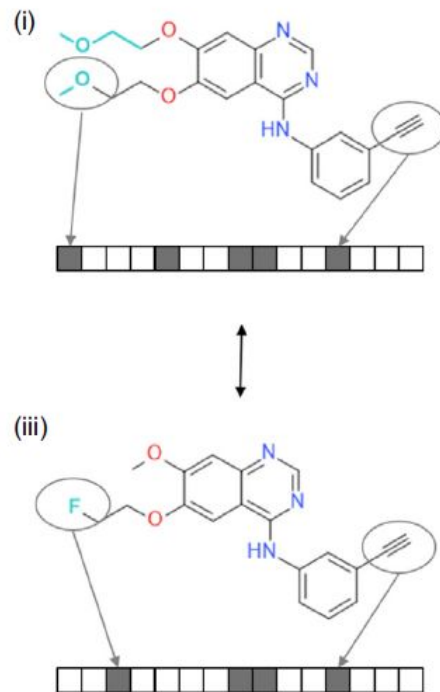


**logP** (partition coefficient) quantifies the hydrophilicity and hydrophobicity of a molecule. The calculated version (cLogP) exists as well.

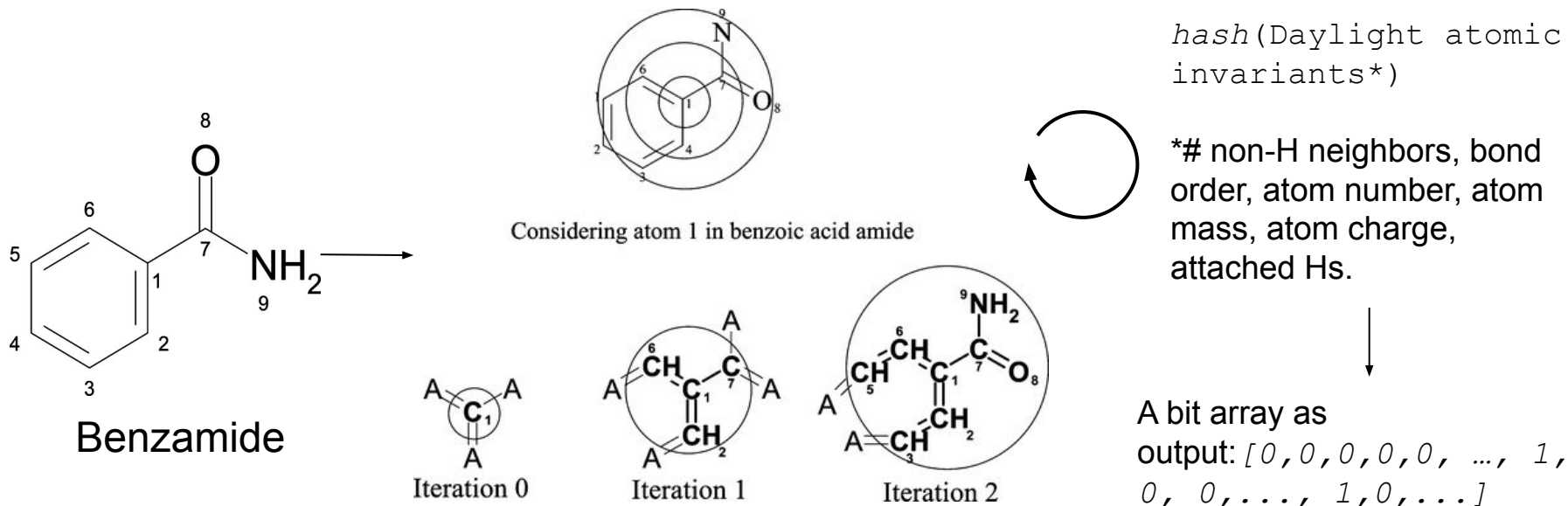


$$\log P_{\text{oct/wat}} = \log \left( \frac{[\text{solute}]_{\text{octanol}}^{\text{un-ionized}}}{[\text{solute}]_{\text{water}}^{\text{un-ionized}}} \right)$$

**Molecular fingerprints:** a set of techniques to represent molecules in a bit array.



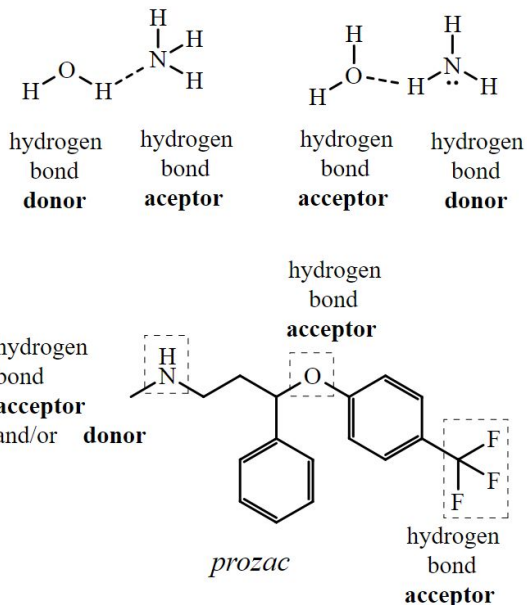
# Extended-connectivity fingerprints (ECFPs) and Functional-class fingerprints (FCFPs) extract and compare (multi-)sets of subgraphs



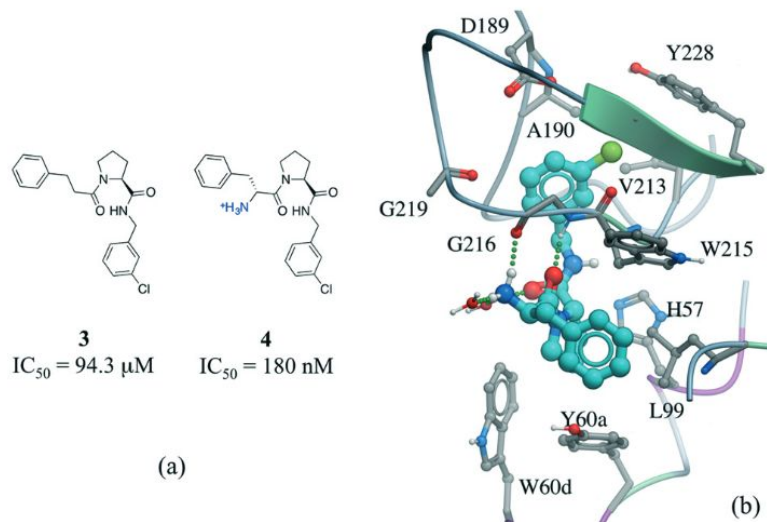
Implemented in [RDKit](#) and other software. Publication and tutorials: (1) Rogers, David, and Mathew Hahn. “[Extended-Connectivity Fingerprints](#).” Journal of Chemical Information and Modeling (2010). (2) Tutorial by [Manish Kumar](#) and (3) Tutorial by [Leo Klärner](#).

# Number of hydrogen bond acceptors and donors are important descriptors, too

A **hydrogen bond**: an electrostatic force of attraction between a hydrogen (H) atom which is covalently bonded to a more electronegative "donor" atom or group (Dn), and another electronegative atom bearing a lone pair of electrons—the hydrogen-bond acceptor.



Hydrogen bonds (H-bonds) both influence the structure of the molecule and its binding to the target.

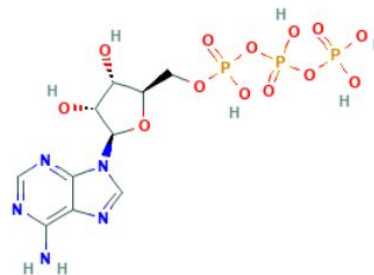


Effect of adding a hydrogen bond in a thrombin inhibitor: a) chemical structure of a pair of thrombin inhibitors; b) crystal structure of molecule 4 (cyan carbons) in complex with thrombin (PDB: 2ZC9). Hydrogen bonds are displayed in dotted green lines.

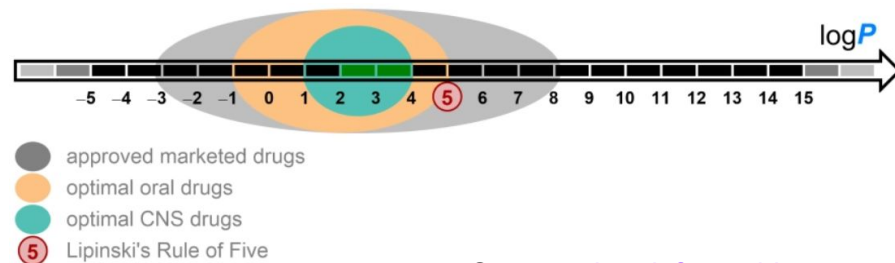


# Lipinski's Rule of Five of small-molecule drugs

- **HBD $\leq$ 5**: No more than **5 hydrogen-bond donors**, e.g. the total number of nitrogen–hydrogen and oxygen–hydrogen bonds.
- **HBA $\leq$ 10**: No more than **10 hydrogen-bond acceptors**, e.g. all nitrogen or oxygen atoms
- **MW $<$ 500**: A molecular weight less than **500 Daltons**, or 500 g/mol.
- **logP $\leq$ 5**: An octanol-water partition coefficient (**log P**) that does not exceed **5**. (10-based)



ATP (MW=507)



Source: [cheminfographic.com](http://cheminfographic.com)

# Rules are made to be broken: more drugs are now beyond the space of Ro5

Table 1. New FDA Approvals (2014 to Present)<sup>a</sup> of Oral bRo5 Drugs

drug	year approved	therapeutic area	MW	cLogP	HBD	N+O
velpatasvir	2016	HCV	883.02	2.5	4	16
venetoclax	2016	oncology	868.44	10.4	3	14
elbasvir	2016	HCV	882.0	2.6	4	16
grazoprevir	2016	HCV	766.90	-2.0	3	15
cobimetinib	2015	oncology	531.31	5.2	3	5
daclatasvir	2015	HCV	738.88	1.3	4	14
edoxaban	2015	cardiovascular	548.06	-0.9	3	11
ombitasvir	2014	HCV	894.13	1.3	4	15
paritaprevir	2014	HCV	765.89	1.1	3	14
netupitant	2014	nausea from chemotherapy	578.59	6.8	0	5
ledipasvir	2014	HCV	889.00	0.9	4	14
ceritinib	2014	oncology	558.14	6.5	3	8

DeGoey, *et al.*. 2018. "[Beyond the Rule of 5: Lessons Learned from AbbVie's Drugs and Compound Collection.](#)" *Journal of Medicinal Chemistry* 61 (7): 2636–51.