COLUMBIA UNIVERSITY

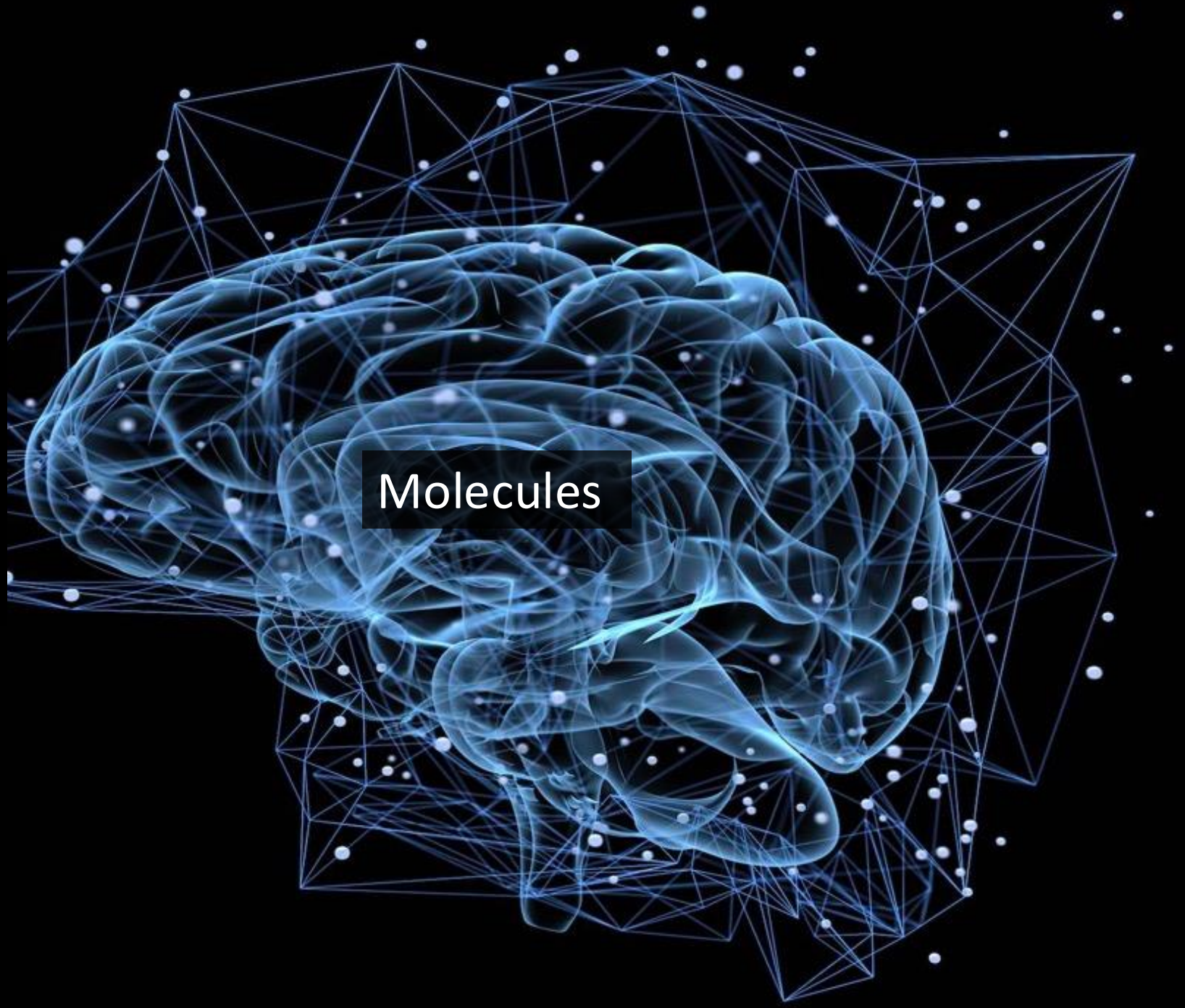
EECS 6895 Advanced Big Data and AI

Lecture 6: AI for Life Sciences I (Molecules & Proteins)

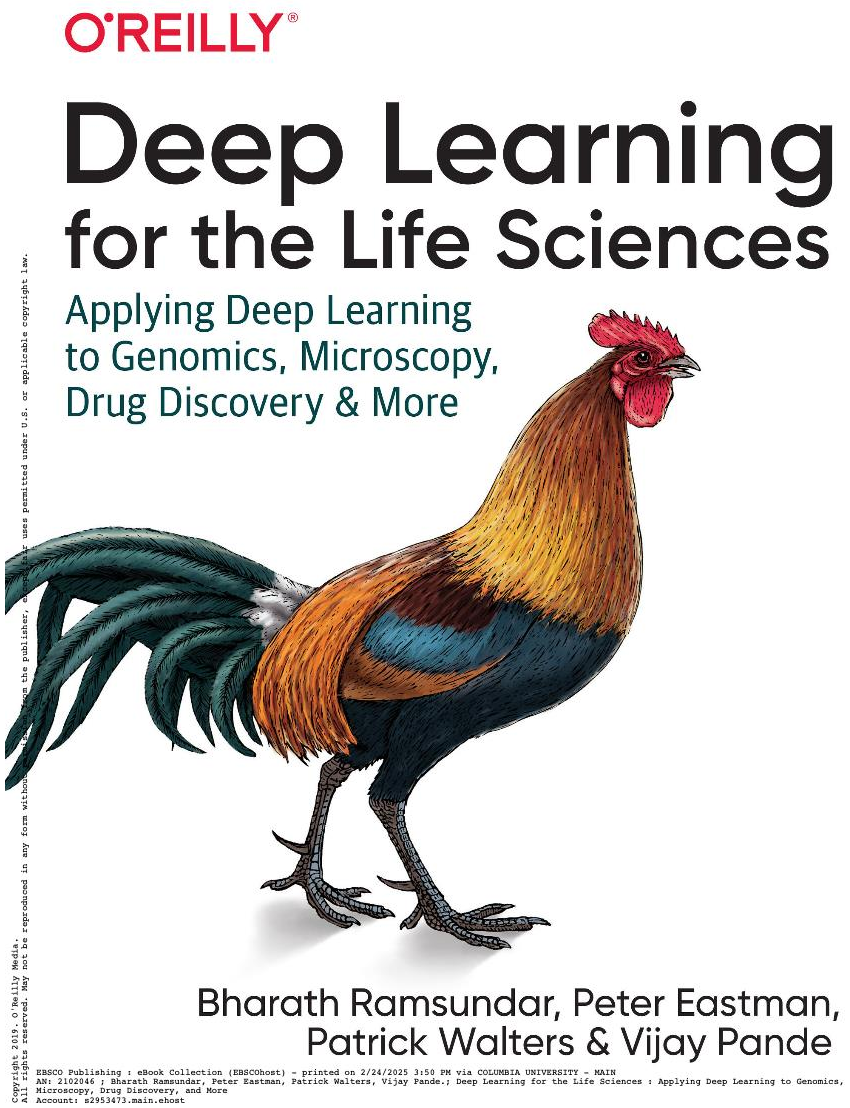
Prof. Ching-Yung Lin

Columbia University

February 25th, 2025


Molecules

Reference Book

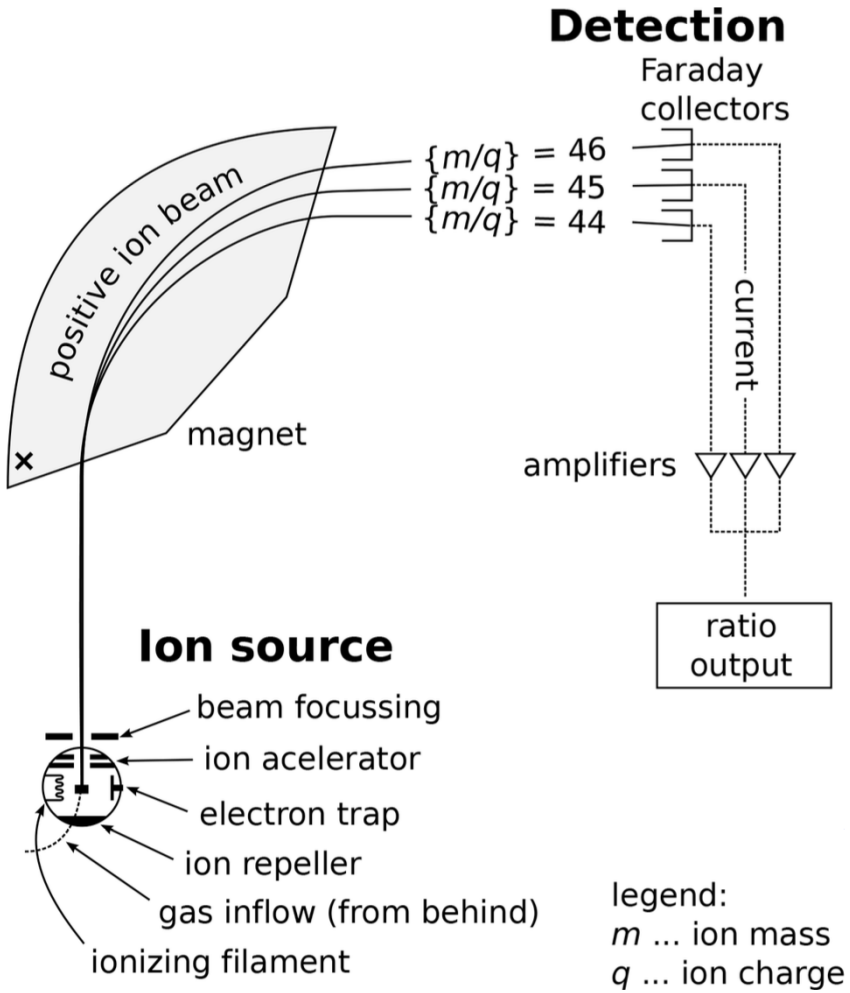


- Chapter 1: Why Life Science?
- Chapter 2: Introduction to Deep Learning
- Chapter 3: Machine Learning with DeppChem
- Chapter 4: Machine Learning for Molecules
- Chapter 5: Biophysical Machine Learning
- Chapter 6: Deep Learning for Genomics
- Chapter 7: Machine Learning for Microscopy
- Chapter 8: Deep Learning for Medicine
- Chapter 9: Generative Models
- Chapter 10: Interpretation of Deep Models
- Chapter 11: A Virtual Screening Workflow Example
- Chapter 12: Prospects and Perspectives

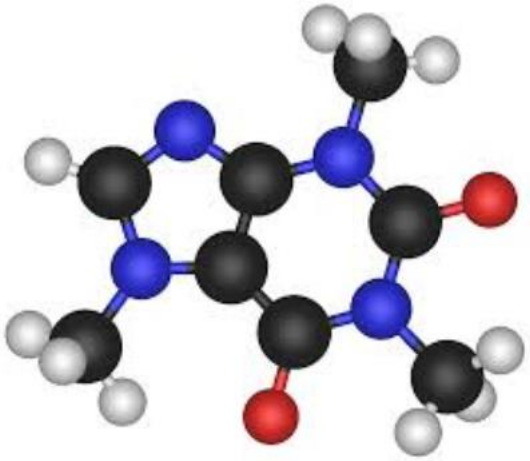
Molecules & Machine Learning

- Modern Materials Science and Chemistry is driven by the need to design new molecules that have desired properties.
- The dream of molecular machine learning is to replace the random experimentation with guided search, where machine-learned predictors can propose which new molecules might have desired properties.
- Such accurate predictors could enable the creation of radically new martials an chemicals with useful properties.
- The first step is to construct technical methods for transforming molecules into vectors of numbers that can then be passed to learning algorithms.
- These representations include chemical descriptor vectors, 2D graph representations, 3D electrostatic grid representations, orbital basis function representations, and more.
- We will review some algorithms for learning functions on molecules, including simple fully connected networks as well as more sophisticated techniques like graph convolutions.
- We'll also describe some of the limitations of graph convolutional techniques, and what we should and should not expect from them.
- We'll end the chapter with a molecular machine learning case study on an interesting dataset.

What Is a Molecule?



Mass Spectrometer



A caffeine molecule as a “ball-and-stick” diagram.

Atoms are presented as color balls.
==> Black: carbon. Red: oxygen. Blue: nitrogen. White: hydrogen.

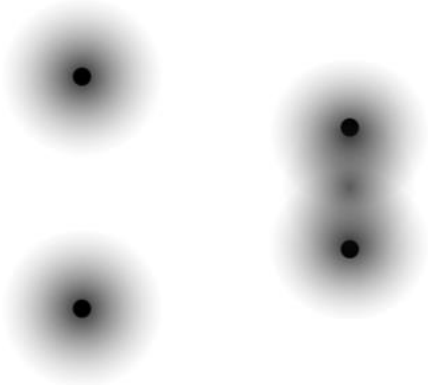
Sticks: chemical bonds

Molecules are dynamic, quantum entities

- Molecules are dynamic entities ➔ all the atoms within a given molecule are in rapid motion.
- The bonds themselves are stretching back and forth and perhaps oscillating in length rapidly.
- It's common for atoms to rapidly break off from and rejoin molecules.
- Molecules are quantum.

Covalent bonds

- Covalent bonds involve sharing electrons between two atoms.
- The same electrons spend time around both atoms.
- In general, covalent bonds are the strongest type of chemical bond.
- The are formed and broken in chemical reactions.
- Covalent bonds tend to be very stable: once they form, it takes a lot of energy to break them.
- Covalent bonds are what define molecules.
- A molecule is a set of atoms joined by covalent bonds.

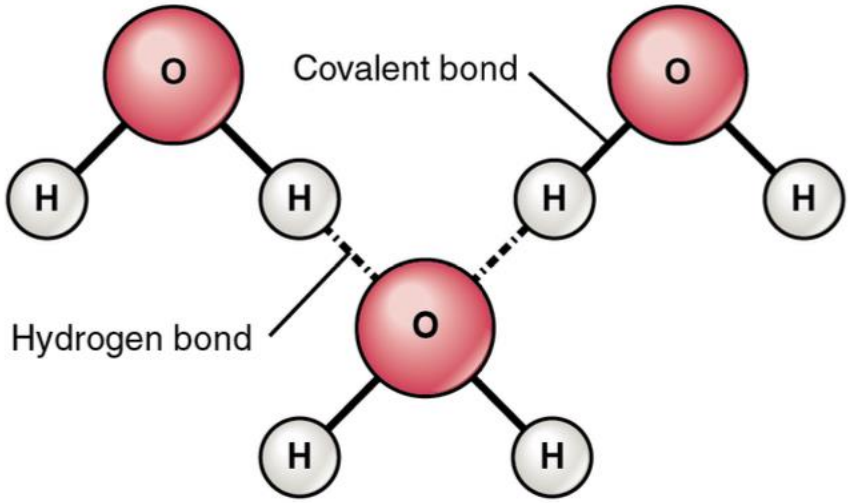


Two atomic nuclei

Two atoms with covalent bond

Noncovalent bonds

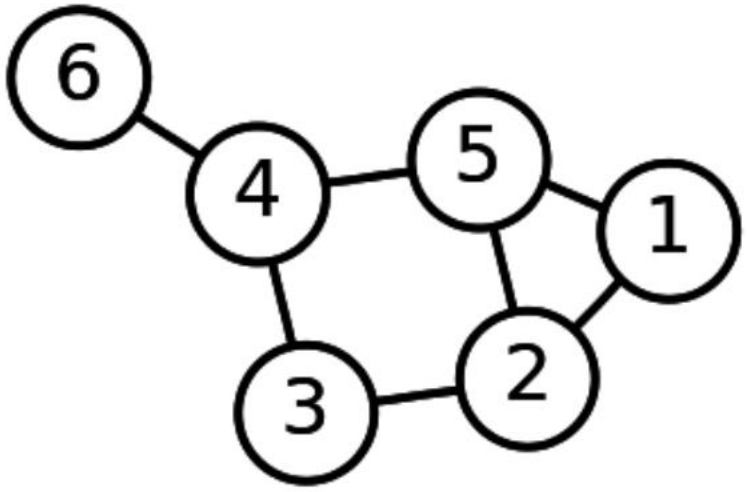
- Noncovalent bonds don't involve the direct sharing of electrons between atoms.
- They do involve weaker electromagnetic interactions.
- These bounds are more ephemeral, constantly breaking and reforming.
- Noncovalent bonds do not "define" molecules in the same sense that covalent bonds do.
- But they have a huge effect on determining the shapes molecules take on ,
- "Noncovalent bonds" is a generic term covering several different types of interactions, including hydrogen bonds, salt bridges, pi-stacking, and more.
- These types of interactions often play crucial role in drug design. ➔ most drugs interact with biological molecules in the human body through noncovalent interactions.



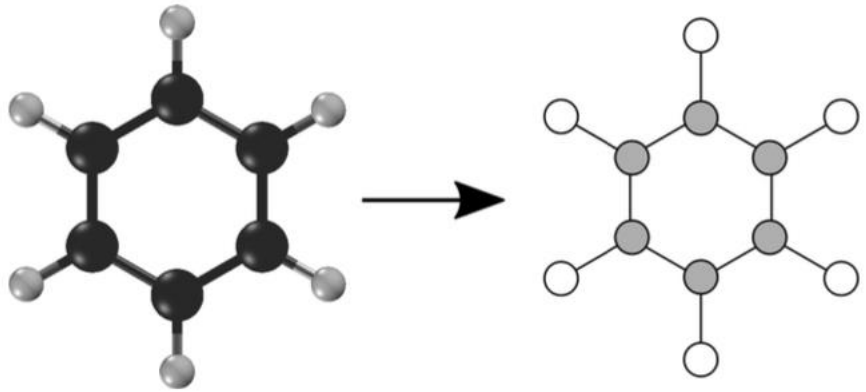
Water molecules have strong hydrogen bounding interactions between gen and oxygen on adjacent molecules

➔ *A strong network of hydrogen bonds contributes to water's power as a solvent.*

Molecular Graphs



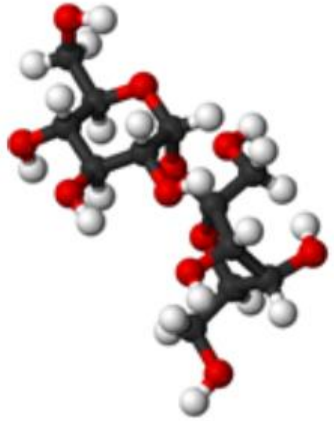
Example of mathematical graph

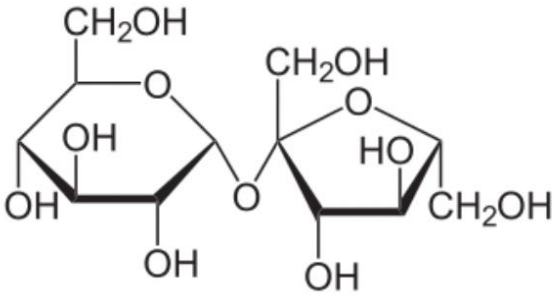


Example of converting a benzene molecule into a molecular graph.

Molecular Conformations

- A molecular graph describes the set of atoms in a molecule and how they are bonded together.
- But how are the atoms positioned relative to each other in 3D space? This is called the **molecule's conformation**.
- Atoms, bonds, and conformation are related to each other.
- If two atoms are covalently bonded, that tends to fix the distance between them, strongly restricting the possible conformations.
- The angles formed by sets of three or four bonded atoms are also often restricted.
- Sometimes there will be whole clusters of atoms that are completely rigid, all moving together as a single unit.
- But other pieces of molecules are flexible, allowing atoms to move relative to each other.
- For example, many covalent bonds allow the groups of atoms they connect to freely rotate around the axis of the bond.
- This lets the molecule take on many different conformations.

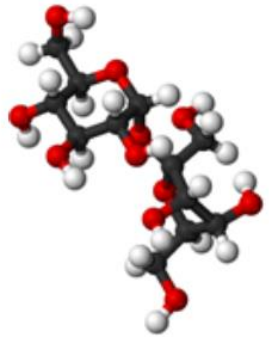


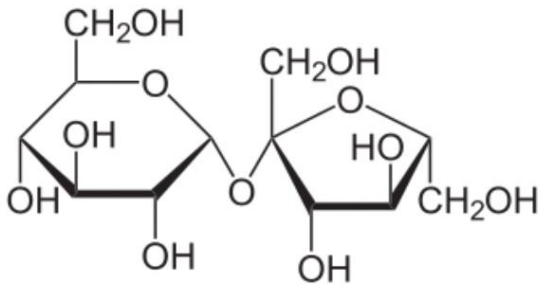


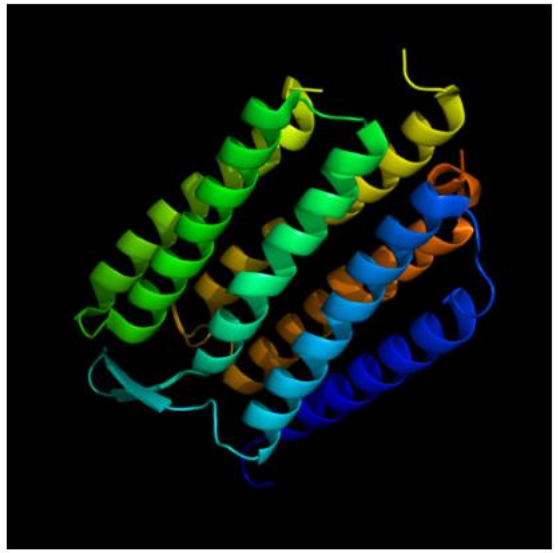
Sucrose (table sugar) represented as a 3D conformation and 2D chemical structure.

Molecular Conformations

- Rings are fairly rigid. Linker connecting them is much more flexible.
- Allowing rings to move relative to each other.
- As molecules get larger, the number of feasible conformations they can take grow enormously.
- For large macromolecules such as proteins, computationally exploring the set of possible conformations is a challenge of decades.



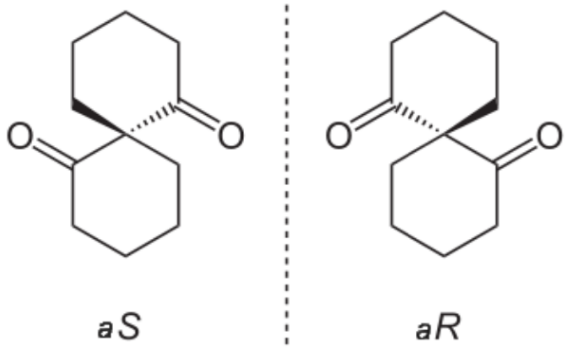




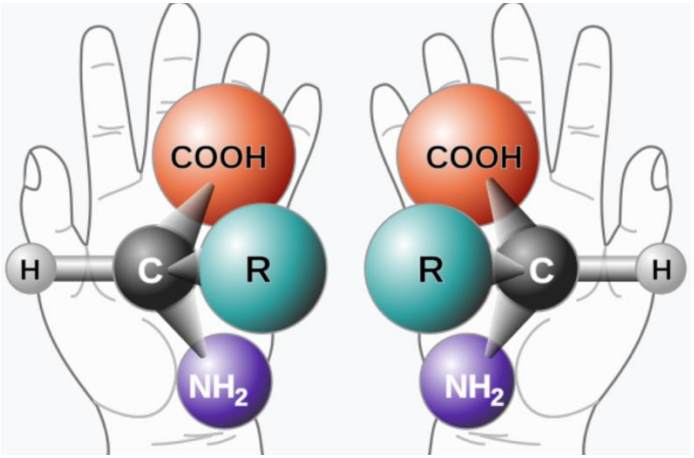
A conformation of bacteriorhodopsin(used to capture light energy) rendered in 3D. Protein conformations are particularly complex with multiple 3D geometric motifs.

Chirality of Molecules

- Some molecules (including many drugs) come in two forms that are mirror images of each other. ➔ this is called **Chirality.**
- Chirality is very important, and also a source of frustration both for lab chemists and computational chemists.
- The chemical reactions that produce chiral molecules often don't distinguish between the forms, producing both chiralities in equal amounts. ➔ Racemic mixtures.
- If you want to end up with just one form, your manufacturing process immediately becomes more complicated.
- Many physical properties are identical for both chiralities, so many experiments can't distinguish between chiral versions.
- Both chiralities have identical molecular graphs => many ML model that depends only on the molecular graph will be unable to distinguish.
- **The two enantiomers have the same chemical properties, except when reacting with other chiral compounds.**



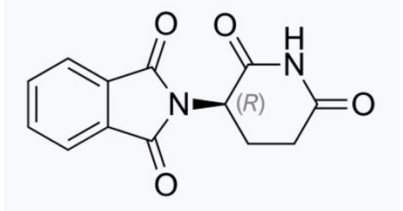
Axial chirality of a spiro compound (a compound made up of two or more rings joined together).

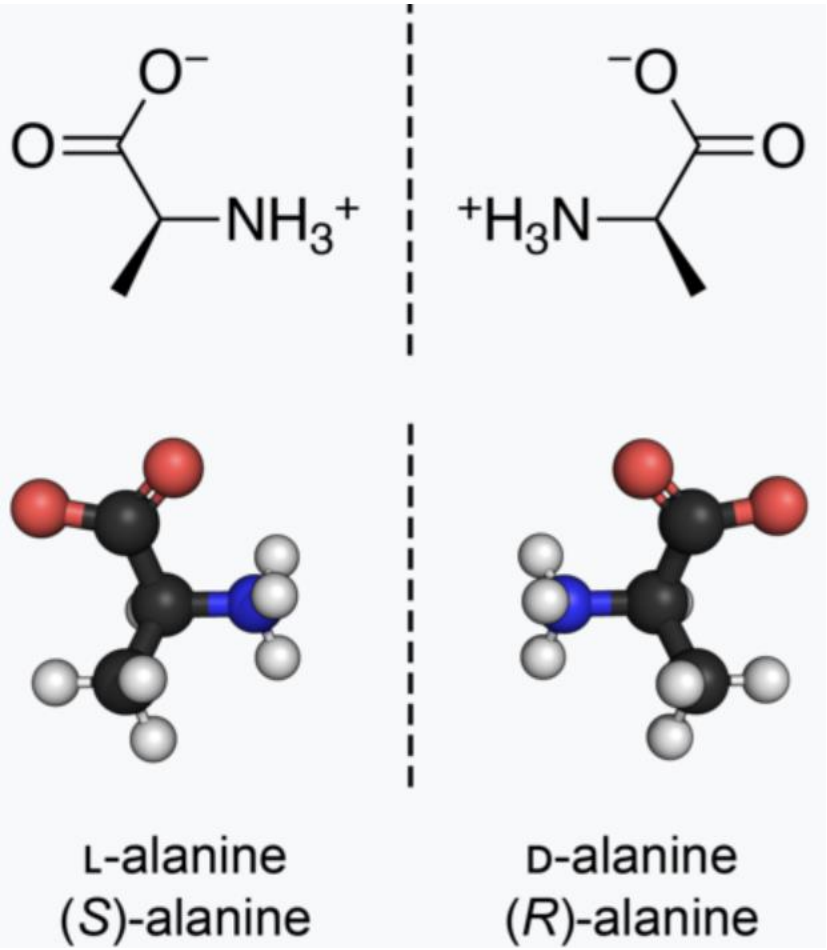


Two enantiomers of a generic amino acidthat are chiral

Chirality of Molecules

- It is possible for the two chiral forms of a drug to bind to totally different proteins. ➔ to have very different effects in the body.
- **In many cases, only one form of a drug has the desired therapeutic effect.**
- The other form just products extra side effects without having any benefit.
- Example: Thalidomide → a drug prescribed as a sedative in the 1950s and 1960s.
  - Treatment for nausea and morning sickness associated with pregnancy.
  - The R form of Thalidomide is an effective sedative.
  - The S form of of Thalidomide is teratogenic and shown to cause severe birth defects.
  - Further compounded by the fact that Thalidomide interconverts or racemizes between the two different forms in the body.



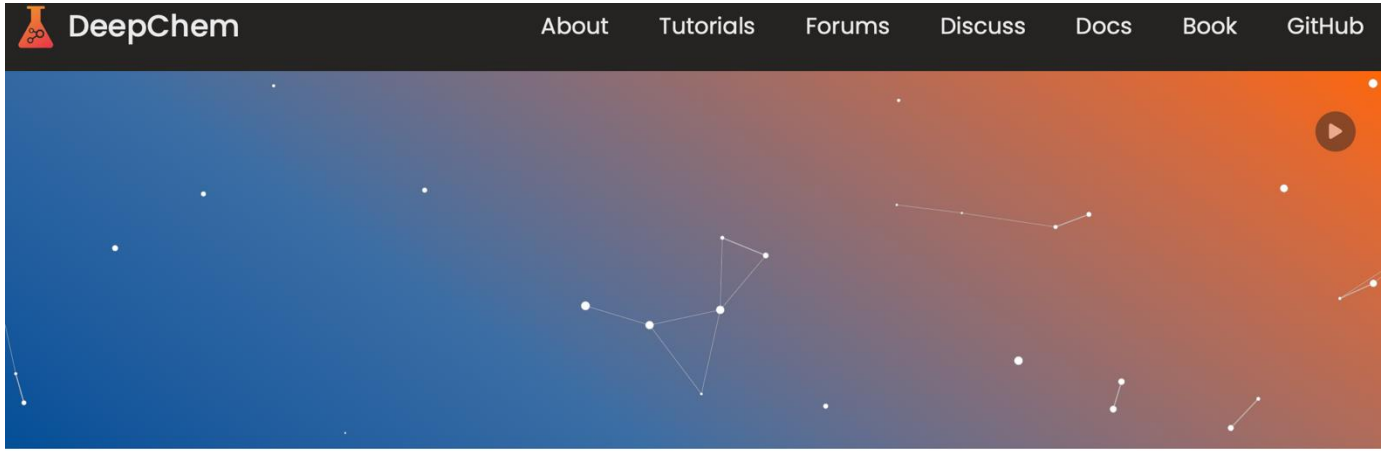


(S)-Alanine (left) and (R)-alanine (right) in zwitterionic form at neutral pH

Encoding → Featuring a Molecule

- In order to perform traditional ML on molecules, we need to transform them into feature vectors.

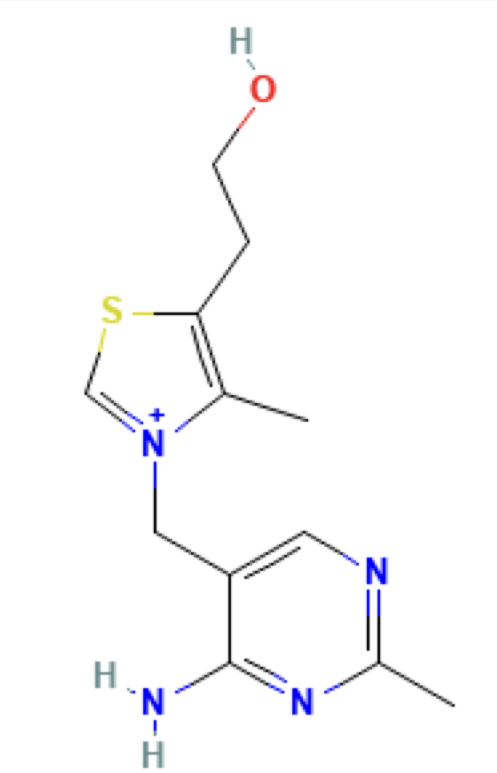
- DeepChem featuring submodule: dc.feat.



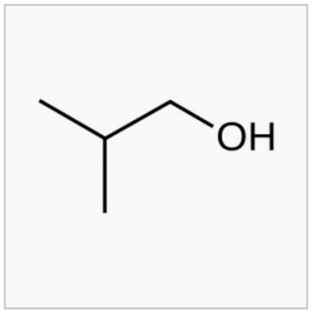
https://deepchem.io

SMILES

- SMILES is a popular method for specifying molecules with text strings.
- It stands for “Simplified Molecular-Input Line-Entry System”.
- A SMILES string describes the atoms and bonds of a molecule in a way that is both concise and reasonably intuitive for chemists.
- For instance, CC1=C(SC=[N+]1CC2=CN=C(N=C2N)C)CCO describes Thiamine, known as vitamin B1.
- While some DL models accept SMILES as input, sometimes we need to convert to a different representation, such as using RDKit in DeepChem.



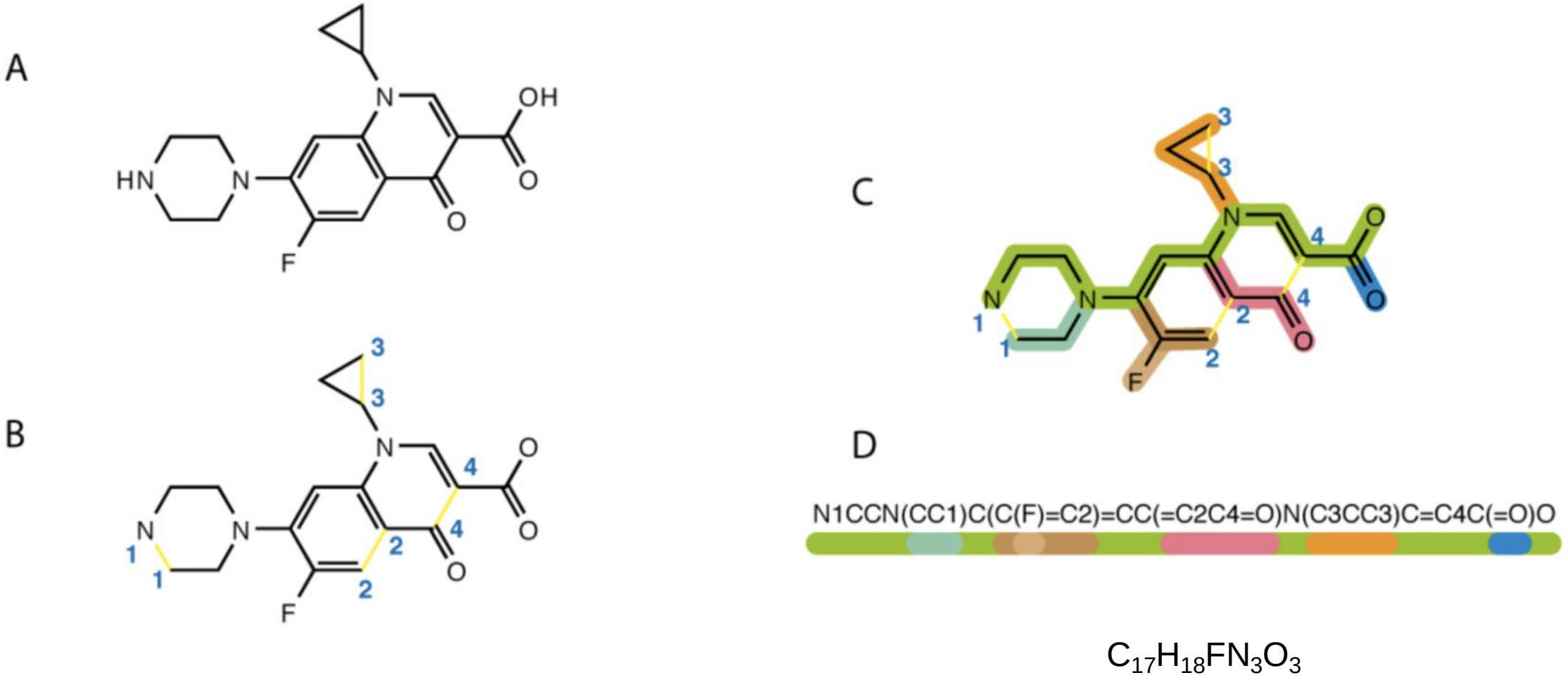
$C_{12}H_{17}N_4OS^+$



Skeletal formula of isobutanol, $(CH_3)_2CHCH_2OH$

$C_4H_{10}O$ => solvent

SMILES example



$C_{17}H_{18}FN_3O_3$

SMILES generation algorithm for ciprofloxacin, an antibiotic: break cycles, then write as branches off a main backbone.

Extended-Connectivity Fingerprints

- Chemical fingerprints are vectors of 1s and 0s that represent the presence or absence of specific features in a molecul
- Extended-connectivity fingerprints (ECFPs) are a class of featurization that combine several useful features.
- They take molecules of arbitrary size and convert them into fixed-length vectors.
- ECFPs let you take molecules of many different sizes and use them all with the same model.
- Easy to compare. The more elements that match, the more similar the molecules are.

```
smiles = ['C1CCCCC1', 'O1CCOCC1'] # cyclohexane and dioxane
mols = [Chem.MolFromSmiles(smile) for smile in smiles]
feat = dc.feat.CircularFingerprint(size=1024)
arr = feat.featurize(mols)
# arr is a 2-by-1024 array containing the fingerprints for
# the two molecules
```

Molecular Descriptors

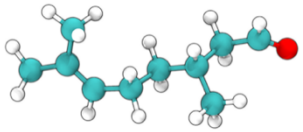
- Describe molecules with a set of physiochemical descriptors.
- Usually correspond to various computed quantities that describe the molecule's structure.
- These quantities, such as the log partition coefficient or the polar surface area, are often derived from classical physics or chemistry.

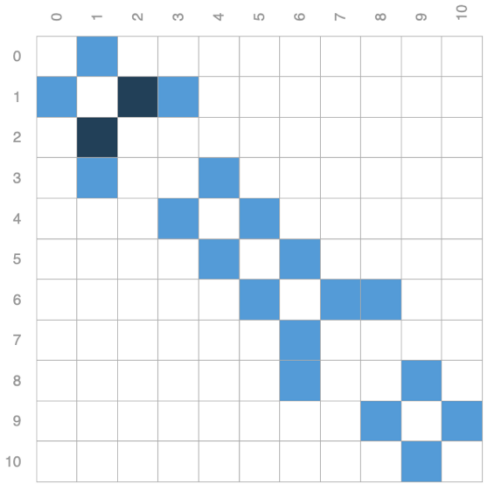
```
feat = dc.feat.RDKitDescriptors()
arr = feat.featurize(mols)
# arr is a 2-by-200 array containing properties of the
# two molecules
```

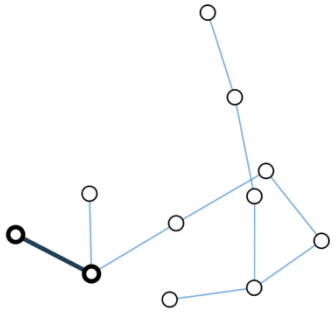
Graph Convolutions

- The featurizations were designed by human. An expert through carefully about how to represent molecules in a way that could be used as input to machine learning models.
- Deep Learning does not rely on such featurization.
- Graph Convolutional Networks begins with a vector of numbers for each node and edge.
- When the graph represents a molecule, those numbers could be high-level chemical properties of each atom, such as its element, charge, and hybridization state.
- Just as a regular convolutional layer computes a new vector for each pixel based on local region of its input, a graph convolutional layer computes a new vector for each node and/or edge.
- DeepChem includes implementations of lots of those architectures, including Graphen Convolutions, Weave models, message passing neural networks, deep tensor neural networks, etc.
- Graph convolutional networks are a powerful tool for analyzing molecules, but they have one important limitation: the calculation is based solely on the molecular graph. No information about the molecule's conformation. => cannot predict anting that is conformation –dependent.

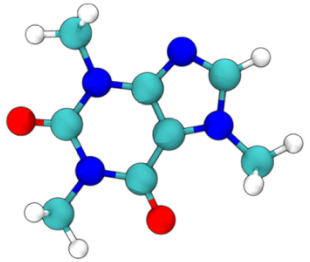
Graph representation of molecule

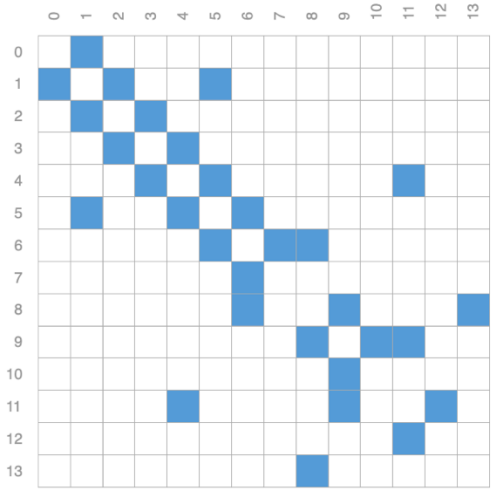


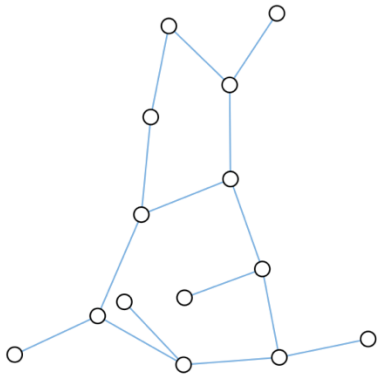




Citronella molecule

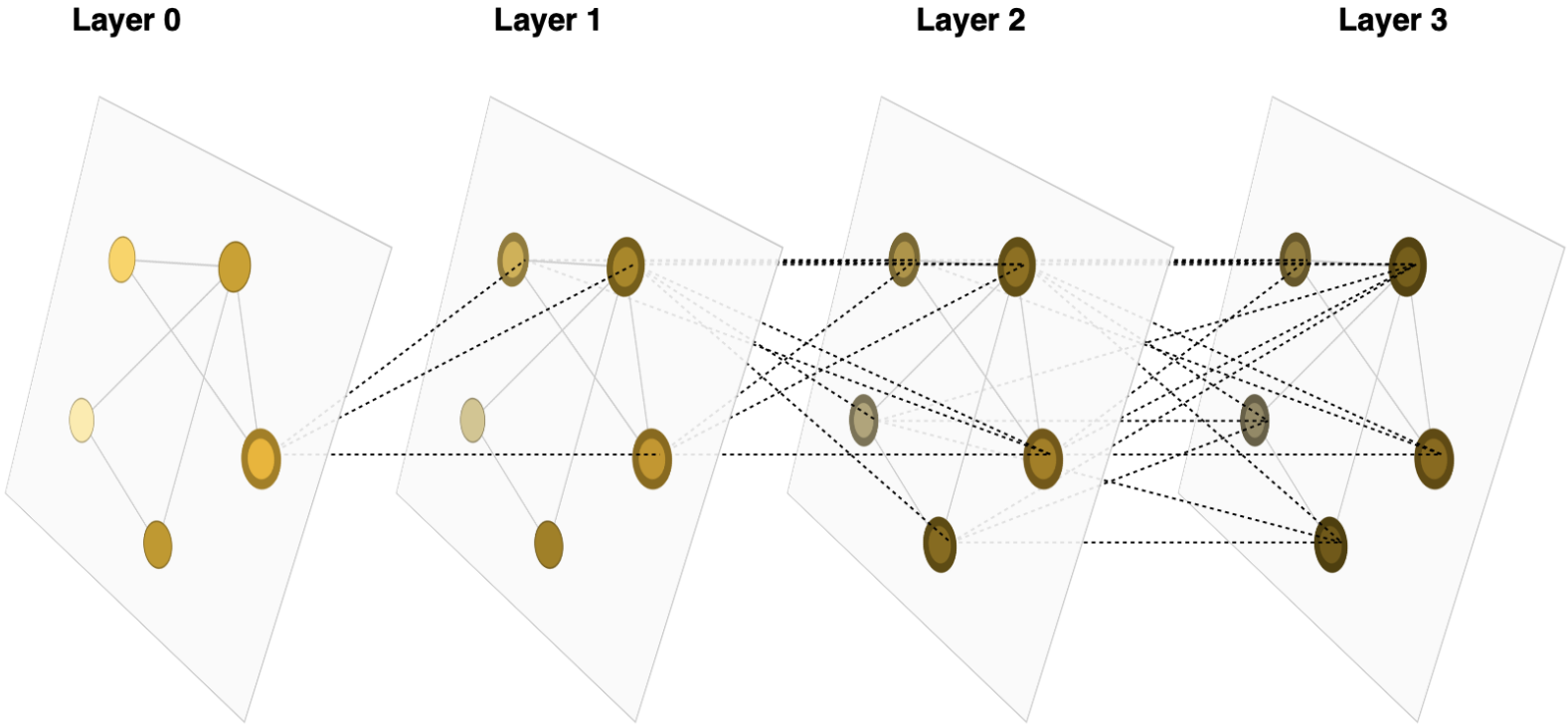






Caffeine molecule

Graph Convolutional Network



https://distill.pub/2021/gnn-intro/

Training a Model to Predict Solubility

Example Task => Train a model on a real chemical dataset to predict an important molecular property.

Load Delaney dataset

```
tasks, datasets, transformers = dc.molnet.load_delaney(featurizer='GraphConv')
train_dataset, valid_dataset, test_dataset = datasets
```

The data is related to solubility, a measure of how easily a molecule dissolves in water.

Construct and Train the model

```
model = GraphConvModel(n_tasks=1, mode='regression', dropout=0.2)
model.fit(train_dataset, nb_epoch=100)
```

```
metric = dc.metrics.Metric(dc.metrics.pearson_r2_score)
print(model.evaluate(train_dataset, [metric], transformers))
print(model.evaluate(test_dataset, [metric], transformers))
```

==> correlation coefficient of solubility prediction: 0.91 for the training data; 0.7 on the testing data.

Testing a Model to Predict Solubility

Test on a new molecule:

```
smiles = ['COC(C)(C)CCCC(C)CC=CC(C)=CC(=O)OC(C)C',
          'CCOC(=O)CC',
          'CSc1nc(NC(C)C)nc(NC(C)C)n1',
          'CC(C#C)N(C)C(=O)Nc1ccc(Cl)cc1',
          'Cc1cc2ccccc2cc1C']
```

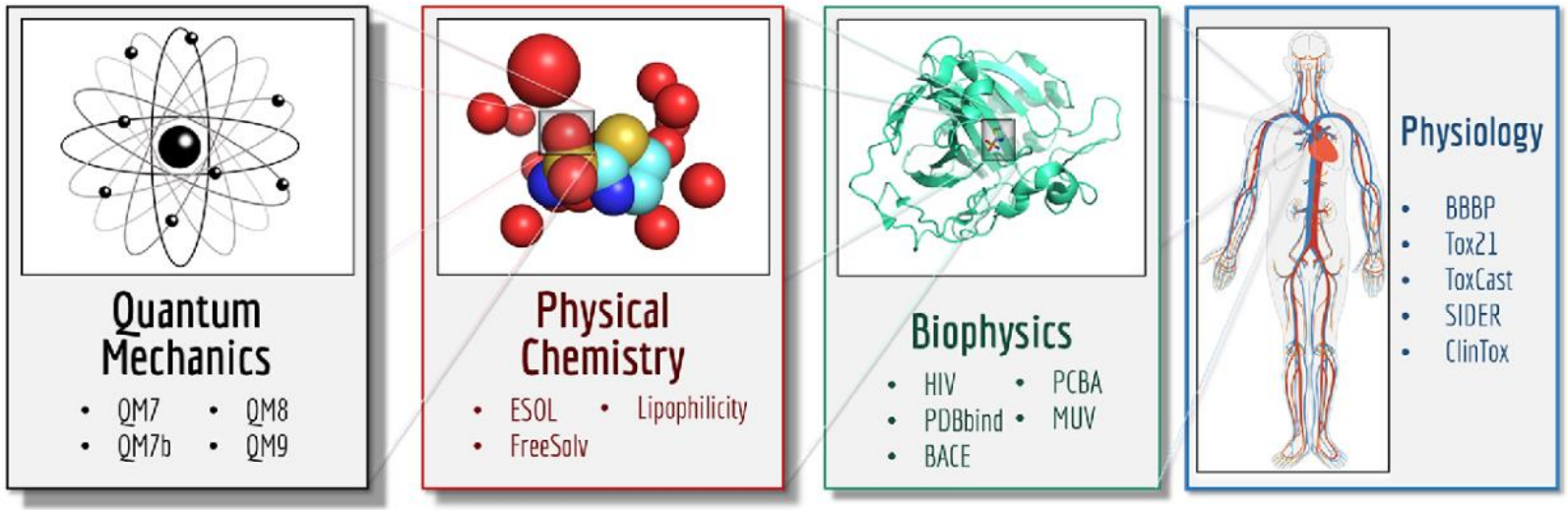
First -> using RDKit to parse SMILES strings, and use featurizer to convert them to the format expected by the graph convolution.

```
from rdkit import Chem
mols = [Chem.MolFromSmiles(s) for s in smiles]
featurizer = dc.feat.ConvMolFeaturizer()
x = featurizer.featurize(mols)
```

```
predicted_solubility = model.predict_on_batch(x)
```

MoleculeNet

- MoleculeNet is a large collection of datasets useful for molecular machine learning.
- Scientists found it useful to predict quantum, physical chemistry, biophysical, and physiological characteristics of molecules.

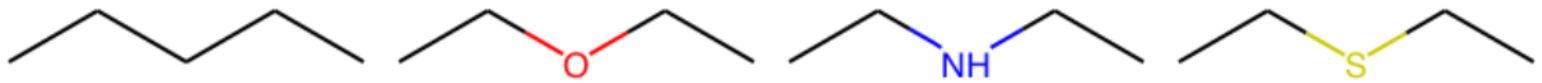


SMART Strings

- We encounter situations where we want to determine whether atoms in a molecule match a particular pattern.

- Examples:
  - Searching a database of molecules to identify molecules containing a particular substructure.
  - Aligning a set of molecules on a common substructure to improve visualization
  - Highlighting a substructure in a plot.
  - Constraining a substructure during a calculation.

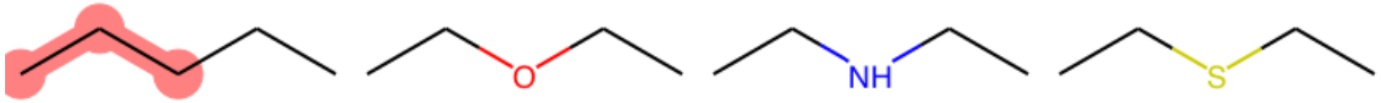
```
from rdkit import Chem
from rdkit.Chem.Draw import MolsToGridImage

smiles_list = ["CCCCC","CCOCC","CCNCC","CCSCC"]
mol_list = [Chem.MolFromSmiles(x) for x in smiles_list]
```



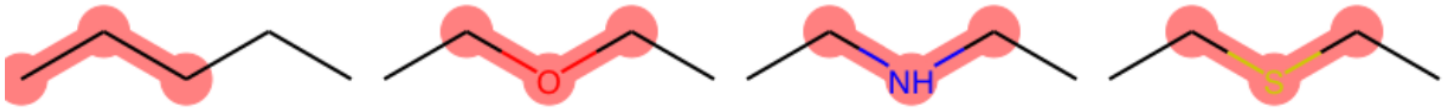
SMART Strings

- SMARTS is an extension of SMILES that can be used to create queries.

```
query = Chem.MolFromSmarts("CCC")
match_list = [mol.GetSubstructMatch(query) for mol in
mol_list]
MolsToGridImage(mols=mol_list, molsPerRow=4,
highlightAtomLists=match_list)
```

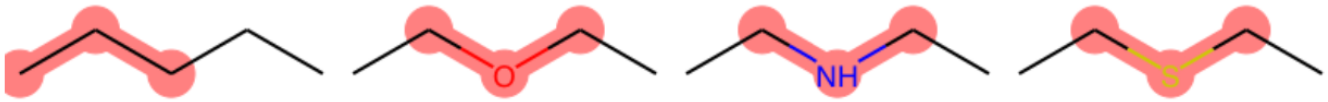



```
query = Chem.MolFromSmarts("C*C")
match_list = [mol.GetSubstructMatch(query) for mol in
mol_list]
MolsToGridImage(mols=mol_list, molsPerRow=4,
highlightAtomLists=match_list)
```

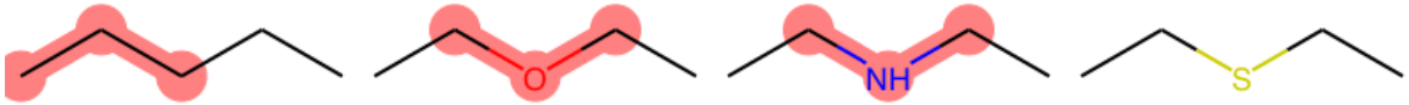


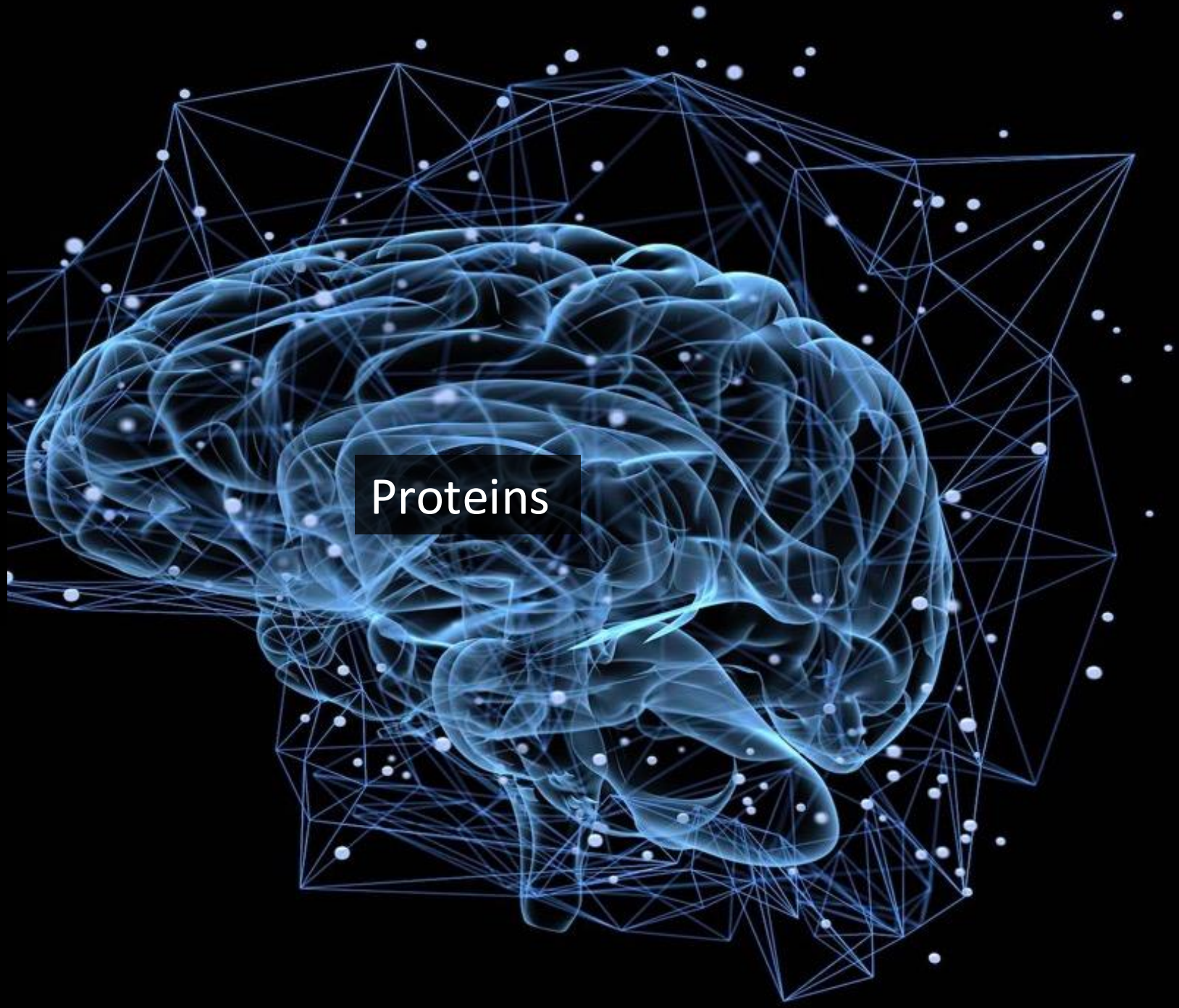
SMART Strings

```
query = Chem.MolFromSmarts("C*C")
match_list = [mol.GetSubstructMatch(query) for mol in
mol_list]
MolsToGridImage(mols=mol_list, molsPerRow=4,
highlightAtomLists=match_list)
```




```
query = Chem.MolFromSmarts("C[C,N,O]C")
match_list = [mol.GetSubstructMatch(query) for mol in
mol_list]
MolsToGridImage(mols=mol_list, molsPerRow=4,
highlightAtomLists=match_list)
```



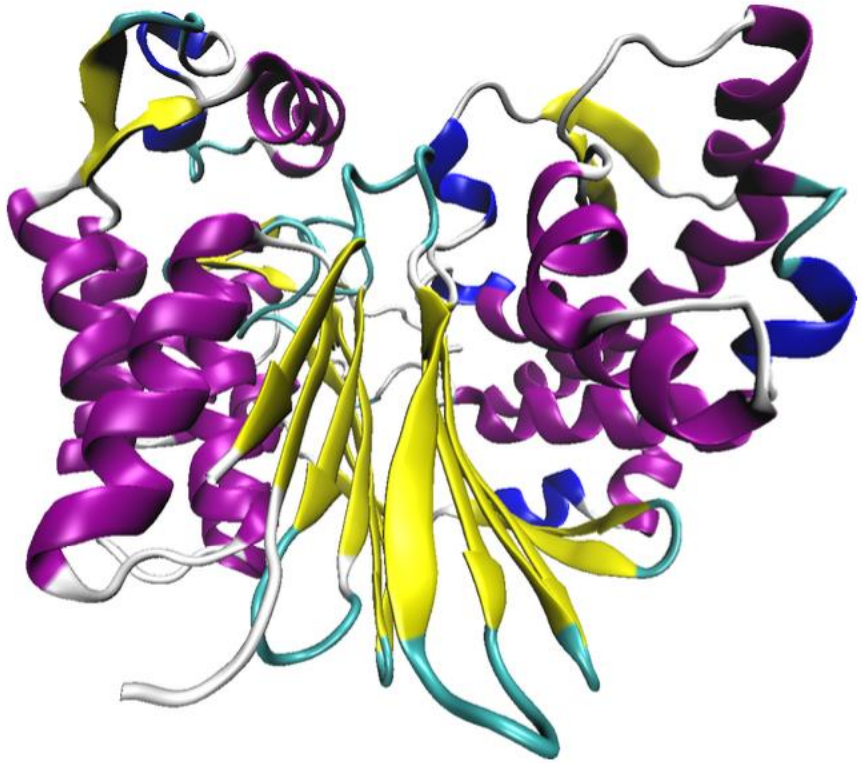

Proteins

Importance of Protein

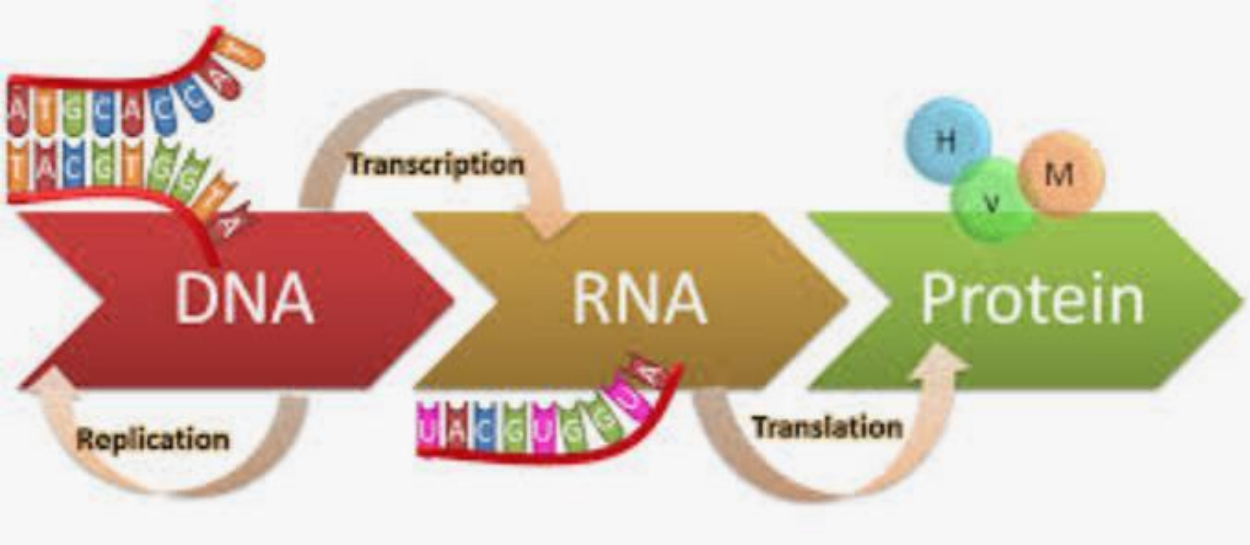
- Proteins are tiny machines that do most of the work in a cell.
- Despite their small size, they can be very complicated.
- A typical protein is made of thousands of atoms arranged in precise ways.
- To understand any machine, you must know what parts it is made of and how they are put together.
- You need to know how it interacts with other molecules.
- They act on molecules, are acted upon by others, and draw energy from others.
- All these interactions depend on the specific positioning of atoms in the two molecules.
- To understand them, you must know how the atoms are arranged in 3D space.
- Unfortunately, you can't just look at a protein under a microscope. Far too small for that. Right now, there are three major methods: ***X-ray crystallography***, ***Nuclear Magnetic Resonance*** (NMR), and ***Cryo-Electron Microscopy*** (cryo-EM).

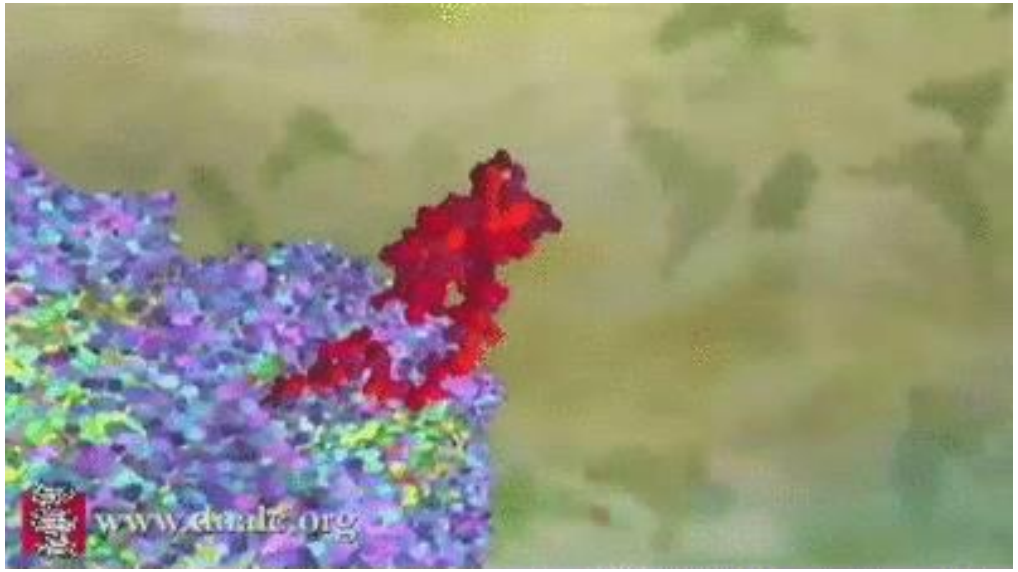
Protein Database (PDB)

- PDB is the primary repository for known protein structures.
- It contained over 142,000 structures by 2019.
- For any protein you really want to study, there is good chance that its structure was still unknown.
- You also want to know many structures of each protein.
- Many proteins can exist in multiple functionally different states (e.g., “active” and “inactive” states0.
- If a protein binds to other molecules, you want a separate structure with the protein bound to each one so you can see exactly how they bind.
- (2019) PDB was still in a “low data” stage. Had far less data than we wanted. That might be true for decades. ➔ AI changed it…

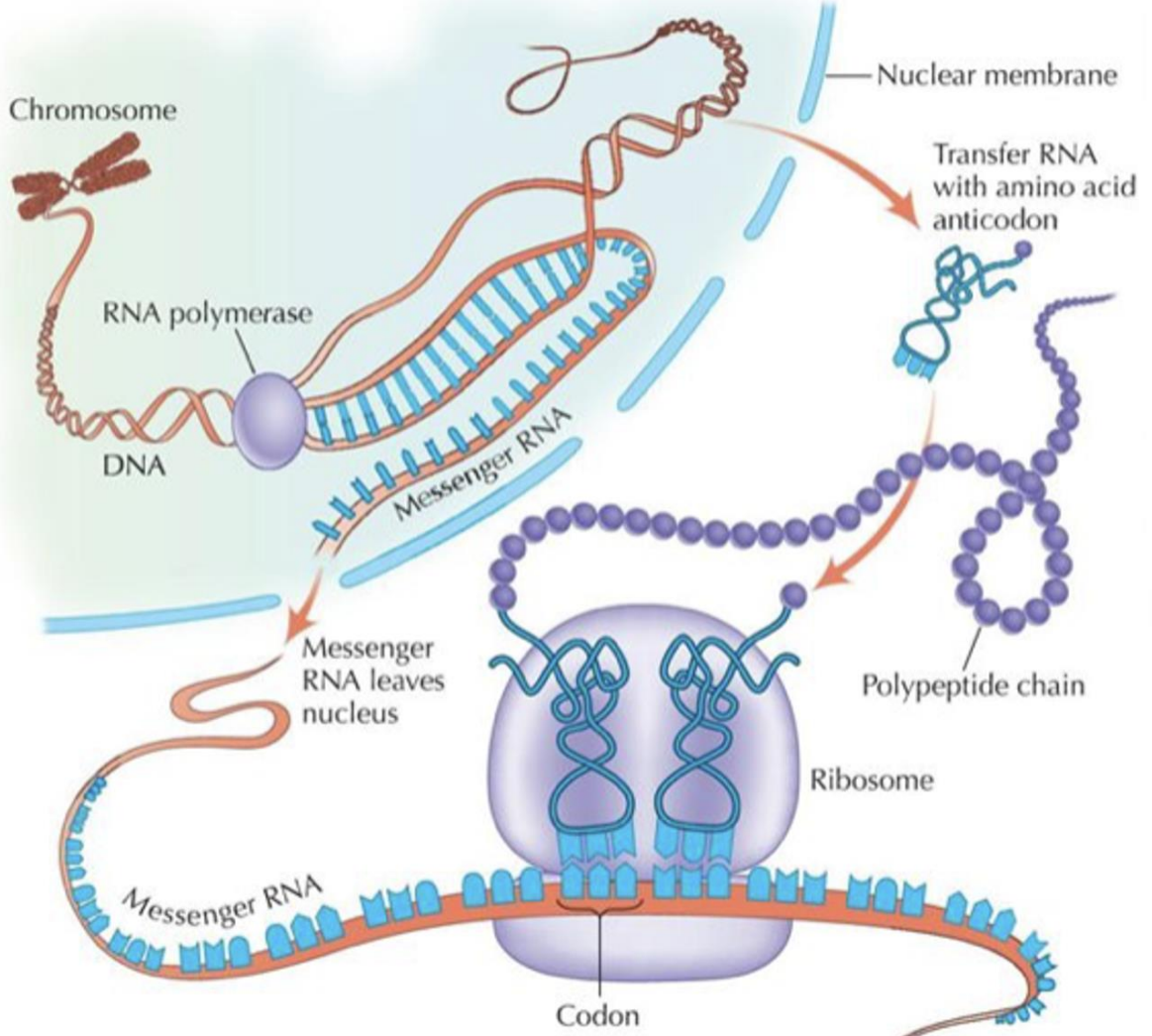


Central Dogma of (Molecular) Biology





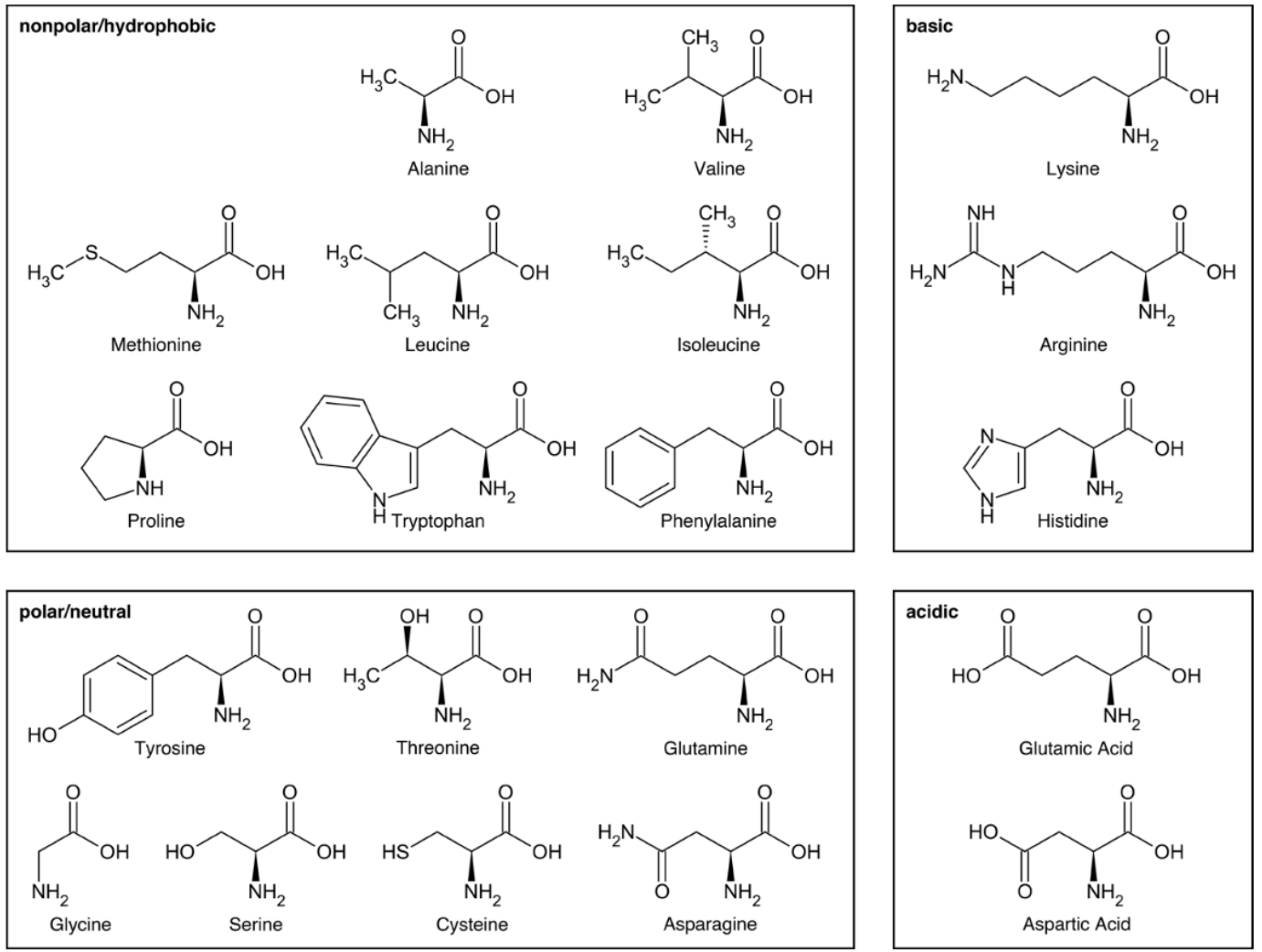
Forming Protein



Protein is the Foundation of All Life

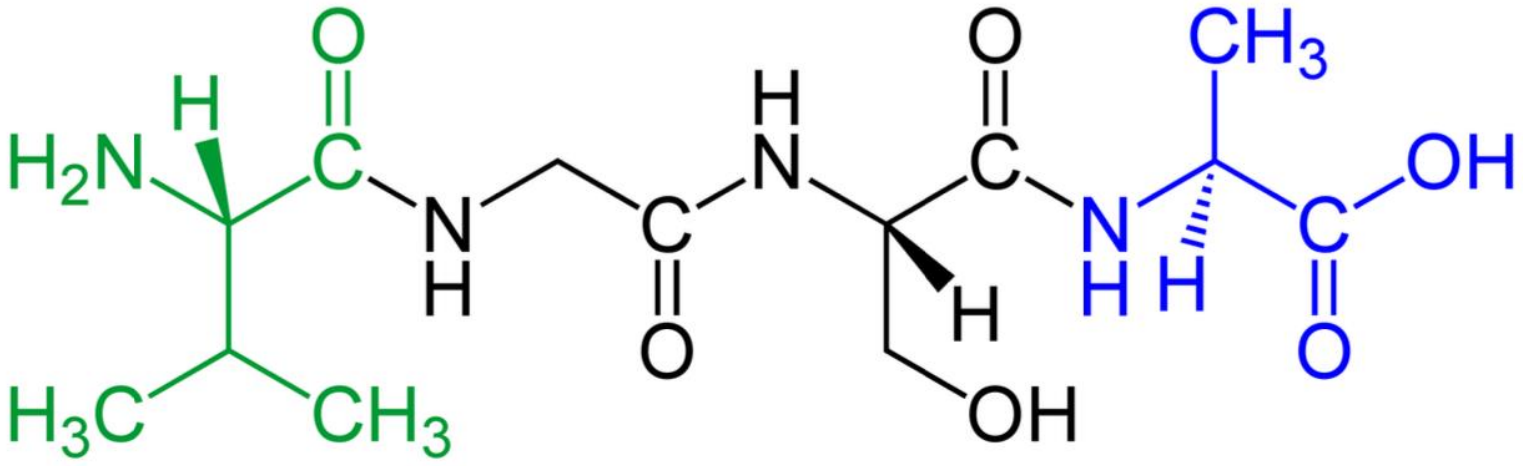
A Protein == A Biological Machine Part

Amino Acid Sequences



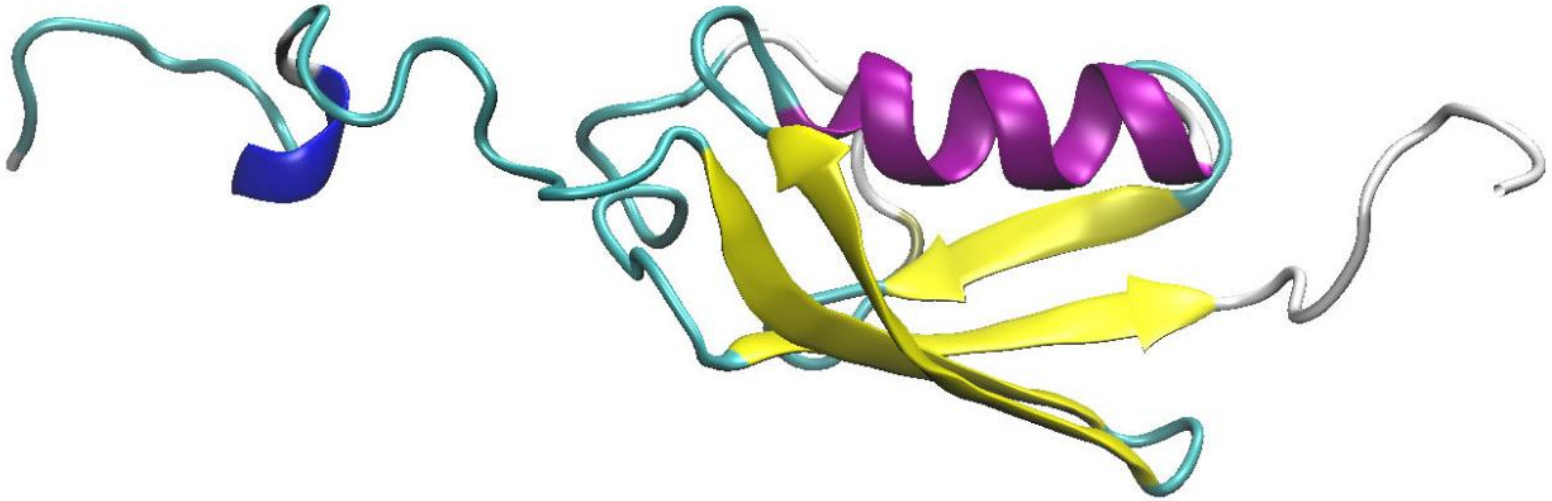
Chain of Amino Acids

- A protein is a chain of amino acids linked one to the next to the next.
- The start of the amino acid chain is typically referred to as the N-terminus.
- The end of the chain is called the C-terminus.
- Small chains of amino acids are commonly called peptides, while longer chains are called proteins.
- Peptides are too small to have complex 3D structures, but the structure of proteins can be very complicated.



Snapshot of a disordered protein

- Most proteins take a rigid shape.
- But, there are also intrinsically disordered proteins which have regions that refuse to take right shapes.

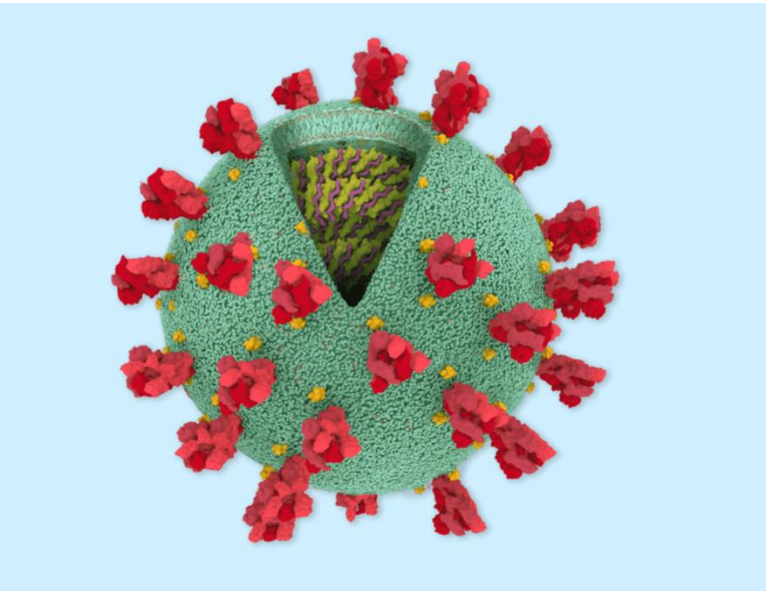


SUMO-1 protein. The central core of the protein has structure, while the N-terminal and C-terminal regions are disordered. Intrinsically disordered proteins such as SUMO-1 are challenging to handle computationally.

RNA of COVID-19 Virus

COLUMBIA UNIVERSITY

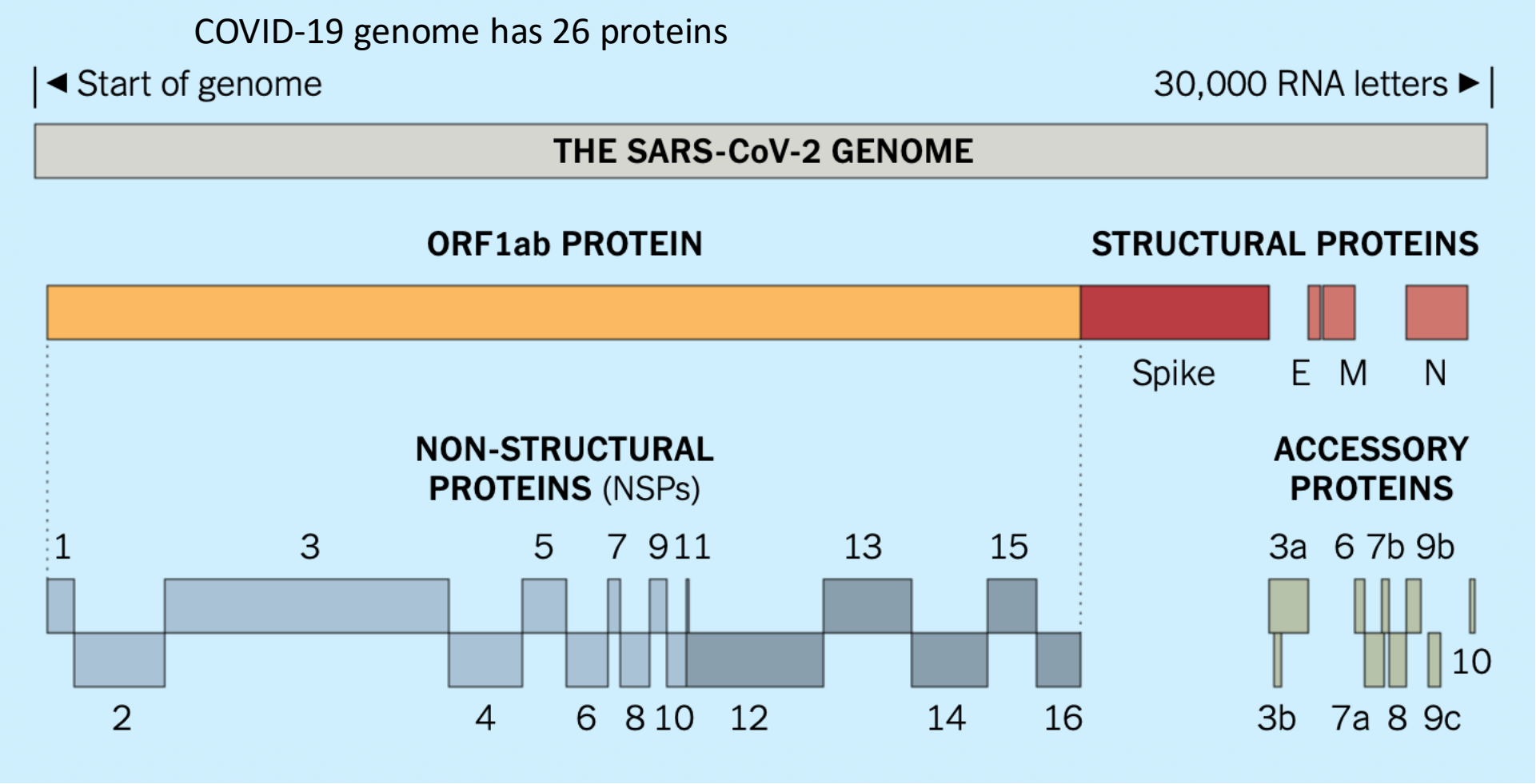
guucucuaaacgaacuuuaaaaucuguguggcugucacucggcugcaugcuuagugcacu
cacgcaguauaauuaauaacuaauuacugucguugacaggacacgaguaacucgucuauc
uucugcaggcugcuuacgguuucguccguguugcagccgaucaucagcacaucuagguuu
cguccgggugugaccgaaagguaagauggagagccuugucccuggguuucaacgagaaaac
acacguccaacucaguuugccuguuuuacagguucgcgacgugcucguacguggcuuugg
agacuccguggaggaggucuuaucagaggcacgucaacaucuuaaagauggcacuugugg
cuuaguagaaguugaaaaaggcguuuugccucaacuugaacagcccuauguguucaucaa
acguucggaugcucgaacugcaccucauggucauguuaugguugagcugguagcagaacu
cgaaggcauucaguacggucguaguggugagacacuuggugucccuugucccucauguggg
cgaaauaccaguggcuuaccgcaagguucuucuucguaagaacgguaauaaaggagcugg
uggccauaguuacggcgccgaucuaaagucauuugacuuaggcgacgagcuuggcacuga
uccuuaugaagauuuucaagaaaacuggaacacuaaacauagcaguggguguuacccguga



Genome is the Parts of Machine

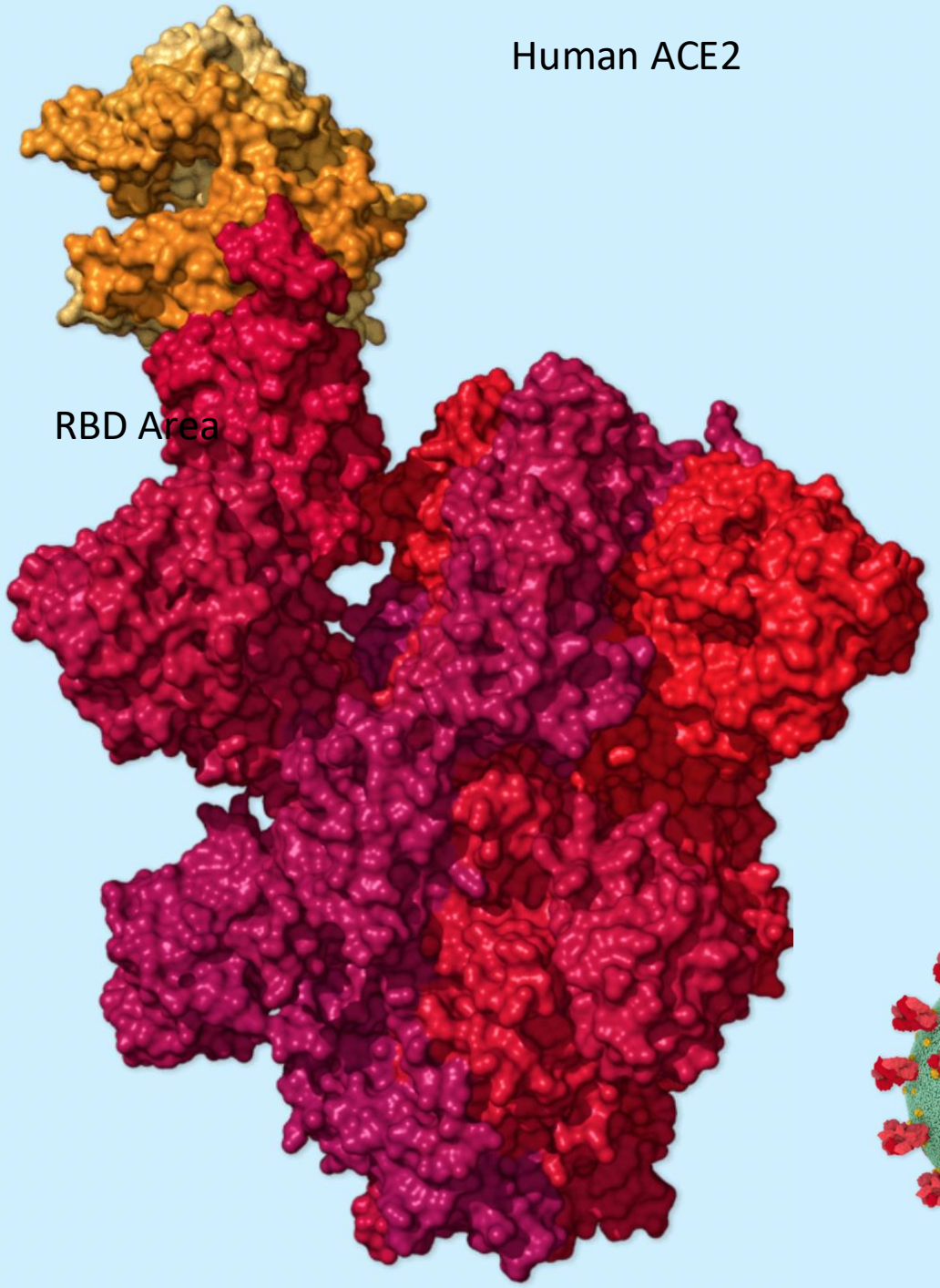
Columbia University

Genome is the Software. Genome is the Hardware.

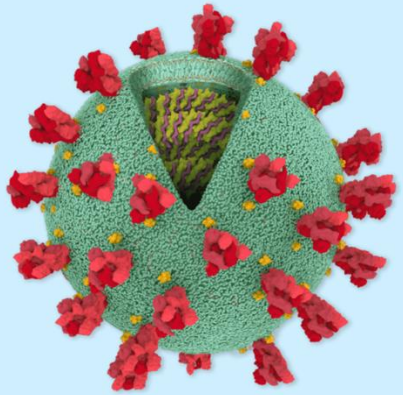


Human ACE2

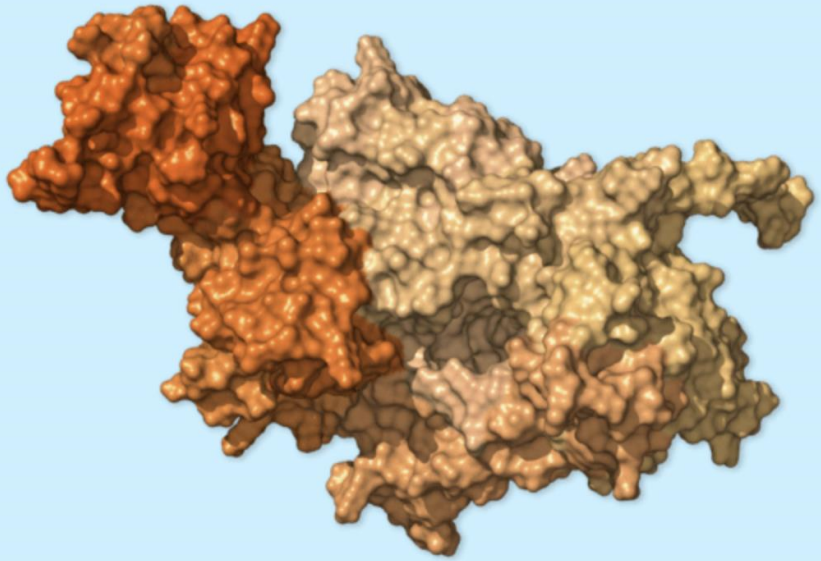
RBD Area



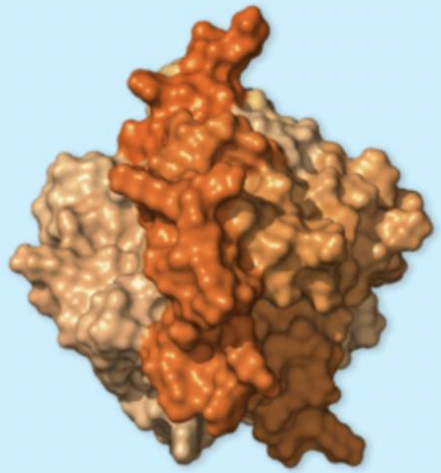
Each Spike is a combination of 3 Spike Proteins



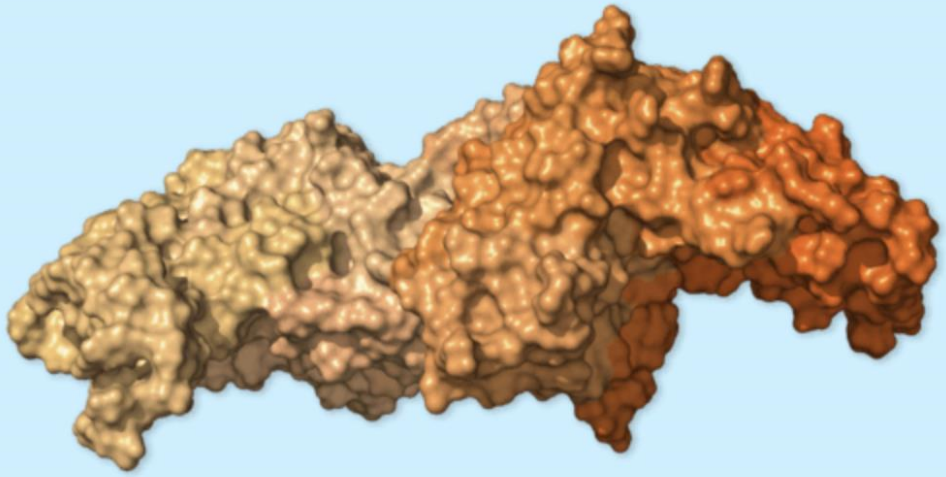
Copy Machine · NSP12



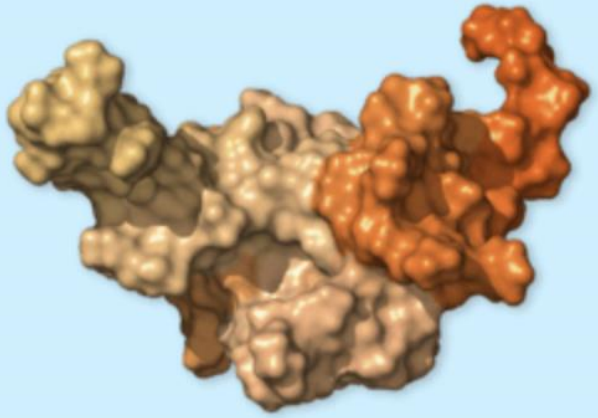
Escape Artist · ORF3a



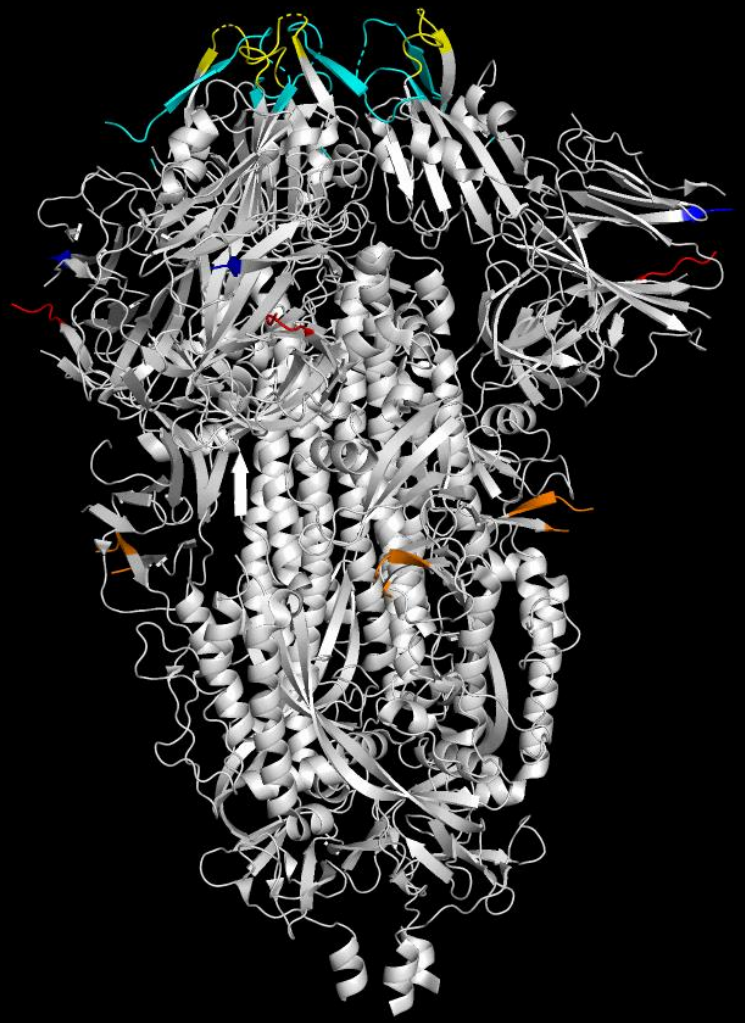
Viral Proofreader · NSP14



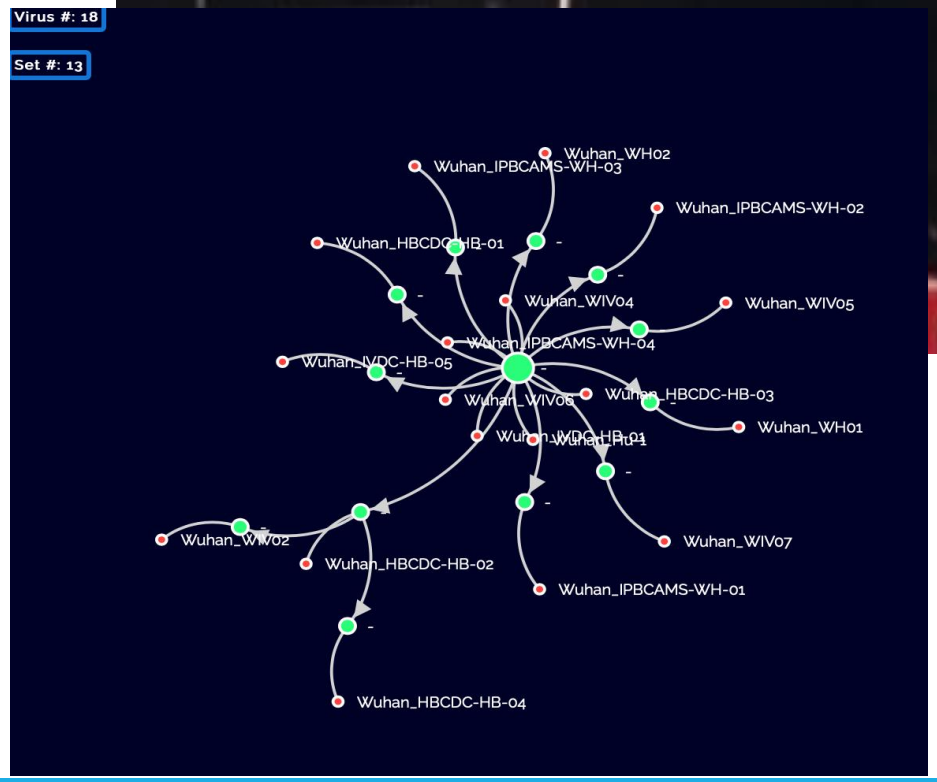
Genetic Camouflage · NSP10



S protein Structure



Colored areas are 6 significant different sections comparing the standard Wuhan HU-1 with two prior coronaviruses found in bats (ZC45 and ZXC21)



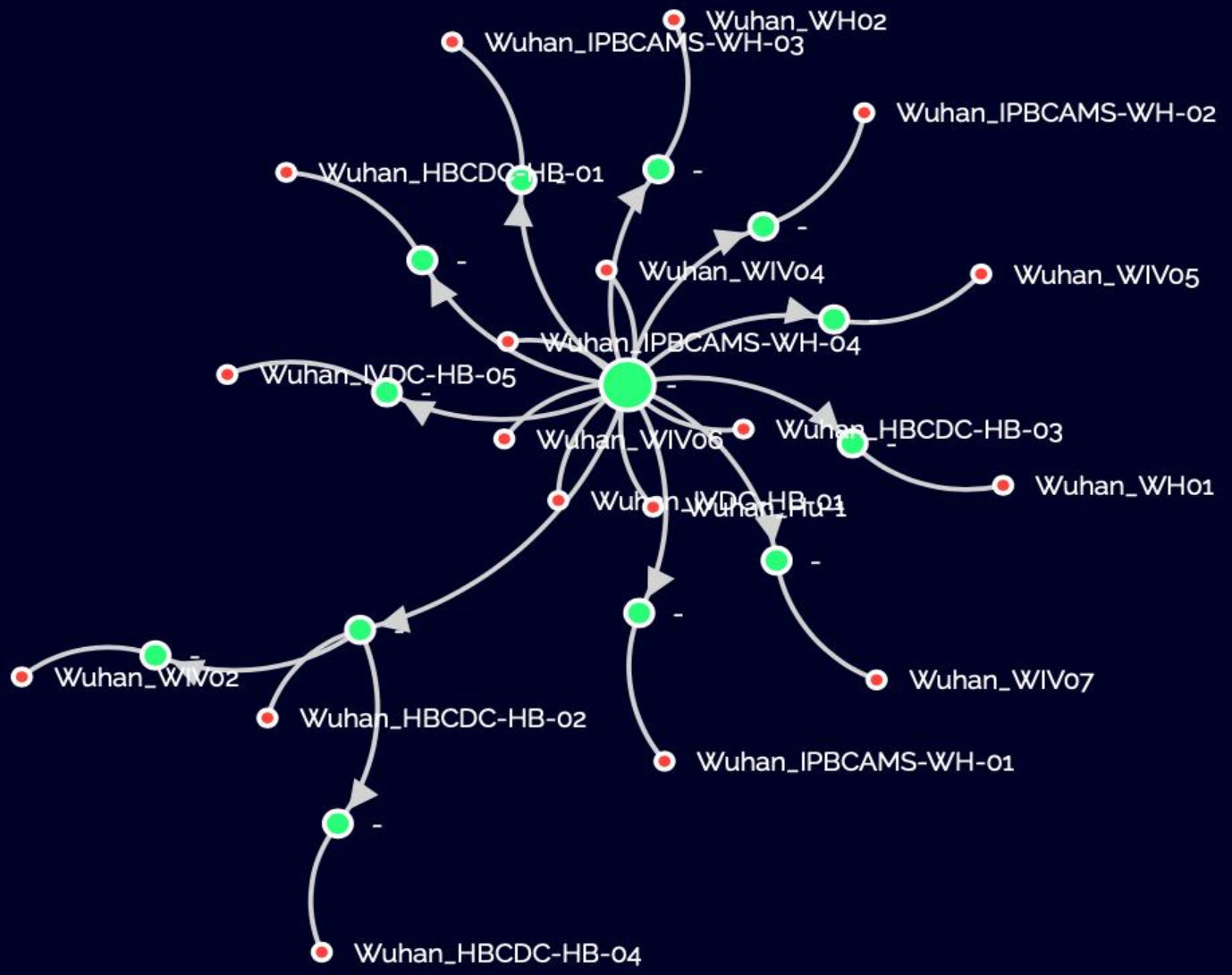
By 12/24/2019, there were 3-5 variants in various genomes ➔ Scientific evidence can almost certain the most widely spread COVID-19 started at late Nov or early Dec.

On average, SARS-CoV-2 adds a stable mutated point per 1-2 weeks

Virus #: 18

Set #: 13

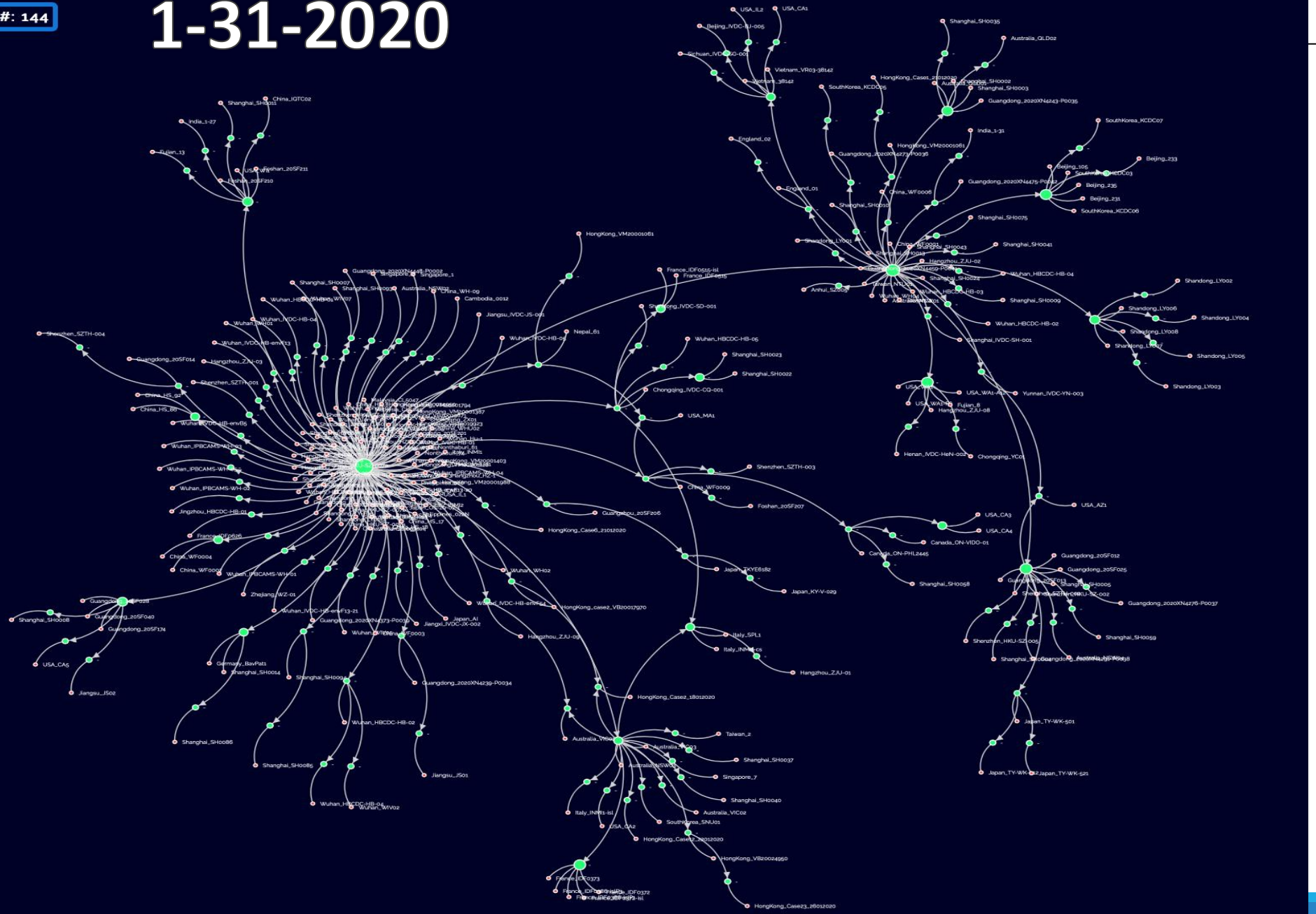
12-31-2019



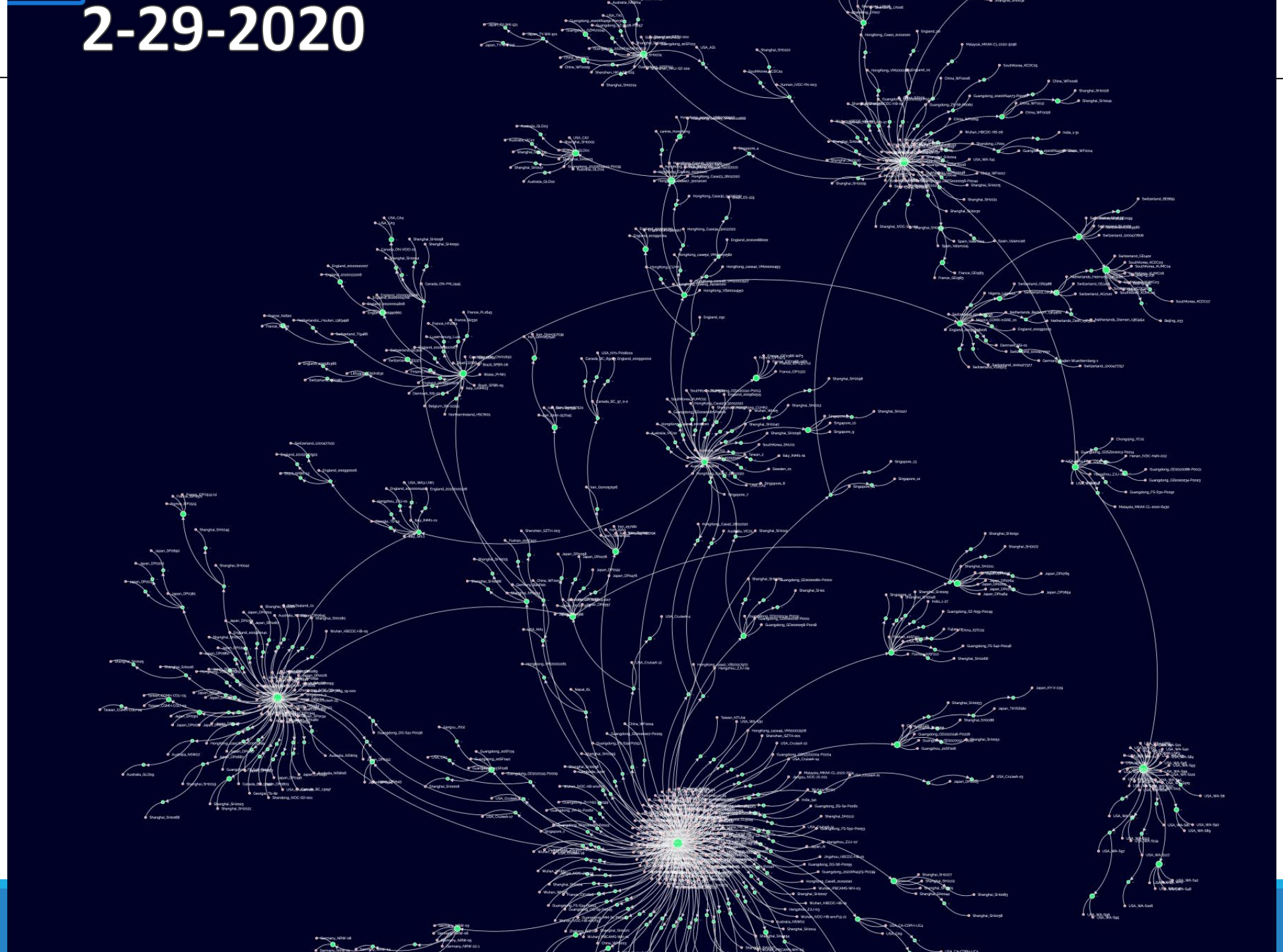
Virus #: 275

Set #: 144

1-31-2020

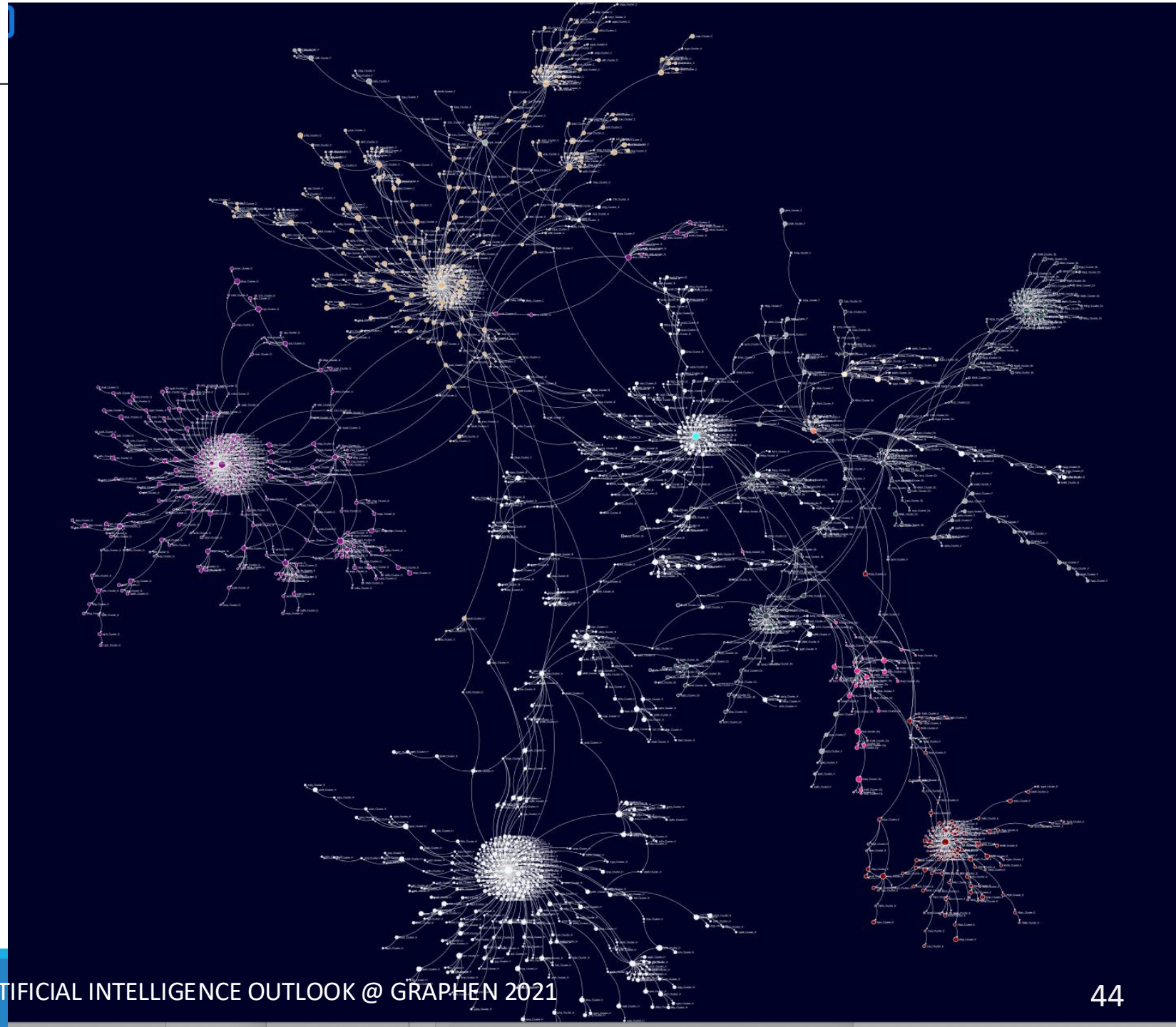


2-29-2020

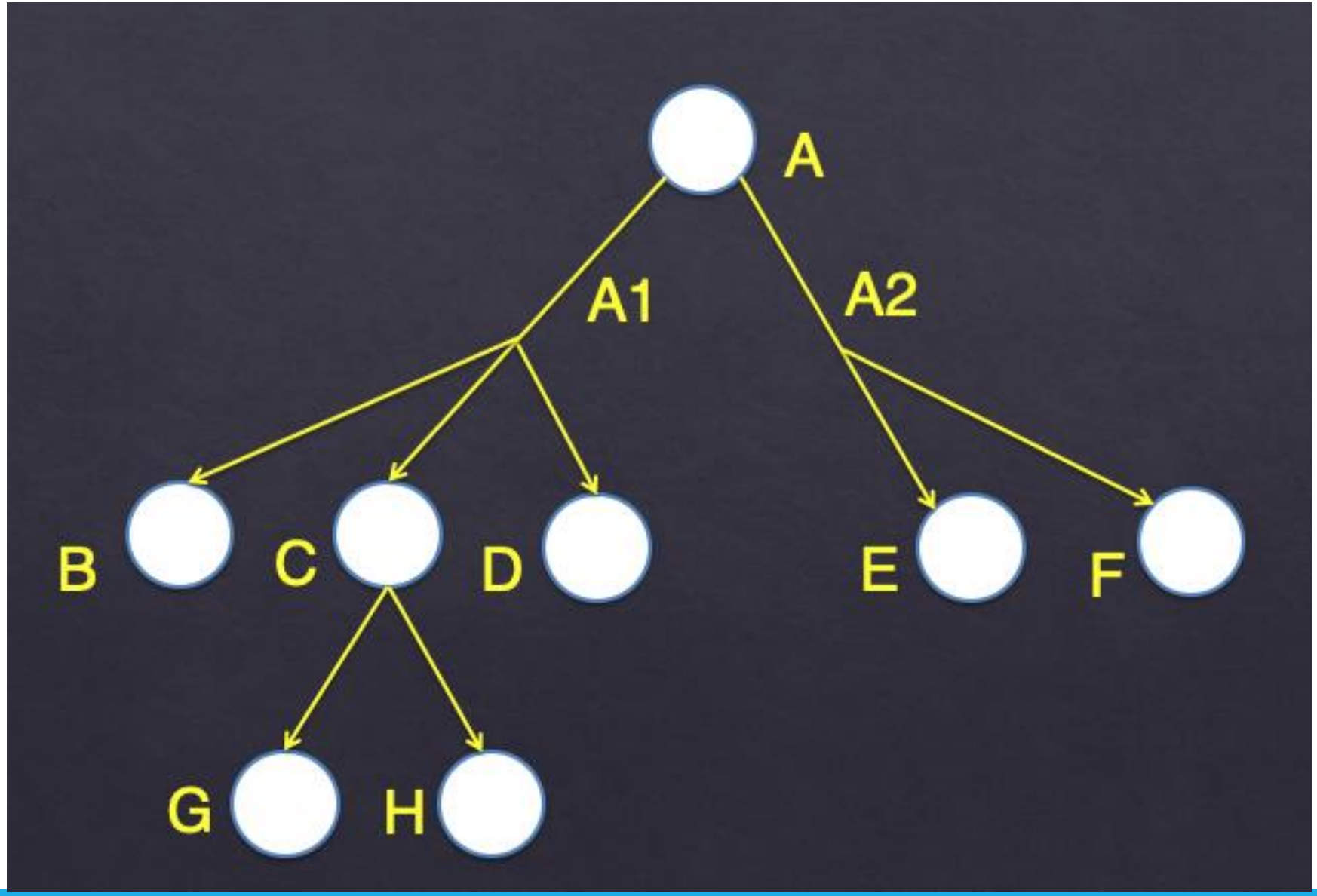


Live COVID-19 Mutation Monitoring

By 1/27/2021: The full genome mutations of 408989 SARS-CoV-2 viruses were analyzed by Graphen



Eight Major Families



Distribution and Start of Eight Major Families

Cluster ID	Cluster Properties		
	Distribution region	*First Appearance Date*	*Features*
A	China	12/30/2019	All COVID-19 viruses may have evolved from here, among which the A2 virus is the closest to the virus genes on bats and pangolins.
B	China and other Asian countries	12/24/2019	Various virus strains evolved from the A1 virus in the A family.
C	Europe	1/28/2020	S protein variation.
D	United Kingdom, the Netherlands, Hong Kong	1/21/2020	NSP2, NSP3 and ORF3a protein variants.
E	United states of America West Coast, Canada	1/19/2020	Most of the ORF8 protein variants have two NSP13 protein variants.
F	Spain, Australia, South Korea, China	1/10/2020	All are viruses that have evolved directly from A2.
G	Europe, South America	2/16/2020	Variations of three consecutive sites on S protein and N protein.
H	France, United States of America East Coast	2/21/2020	S protein variation and ORF3a protein variation.

Protein

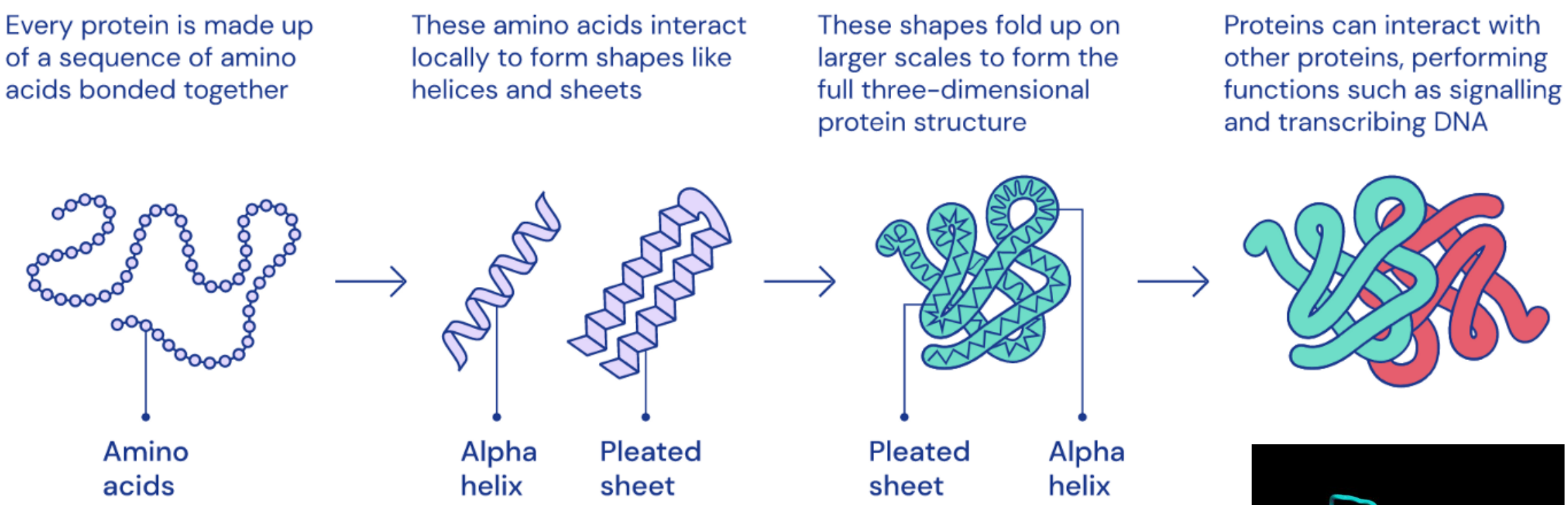
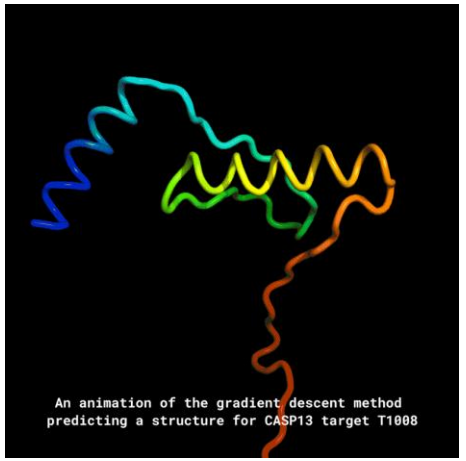


FIGURE 1: COMPLEX 3D SHAPES EMERGE FROM A STRING OF AMINO ACIDS.

Protein is the Foundation of All Life

A Protein == A Biological Machine Part



Predicting 3D Protein Structure

- There were two main approaches to predict protein structures.

- **Homology Modeling:**
  - protein sequences and structures are the product of billions of years of evolution.
  - If two proteins are near relatives ("homologs") that only recently diverged from each other, they probably have similar structures.
  - You first look for a homolog whose structure is already known, then try to adjust it based on differences between the sequences of the two proteins.
  - Homology modeling works reasonably well for determining the overall shape of a protein.
  - But, it often gets details wrong.

- **Physical Modeling:**
  - Using knowledge of the laws of physics, you try to explore many different conformations that protein night take on and predict which one will be most stable.
  - This method required enormous amounts of computing time.
  - It was considered too difficult to do.
  - But, now it works, such as Graphen ATOM platform.

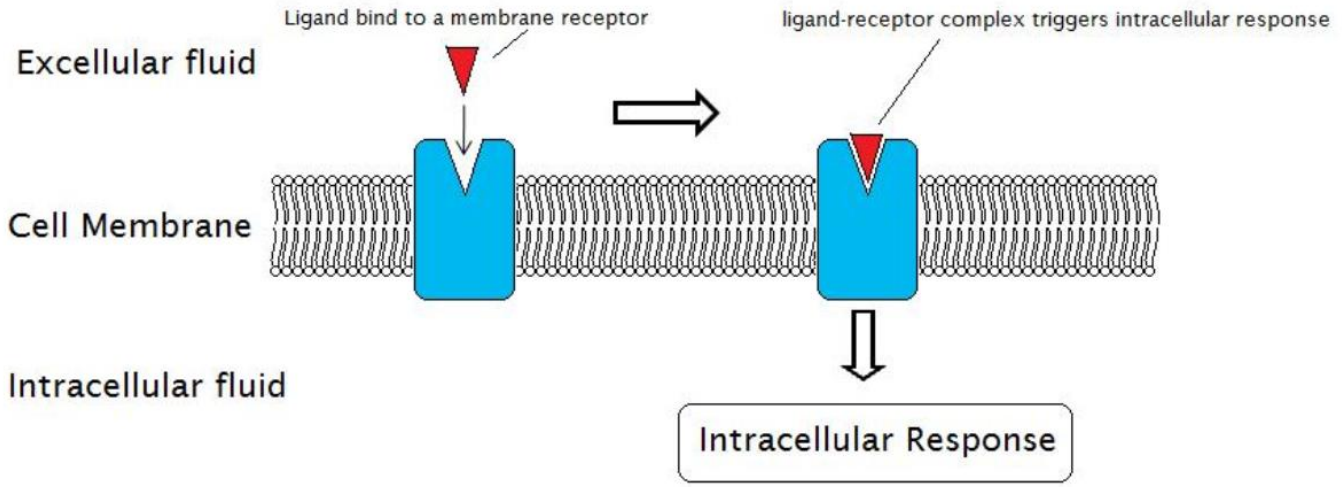
Short Primer on Protein Binding

Proteins often bind to small molecules.

Sometimes, the binding behavior is central to the protein's function: the main role for a given protein can involve binding to particular molecules.

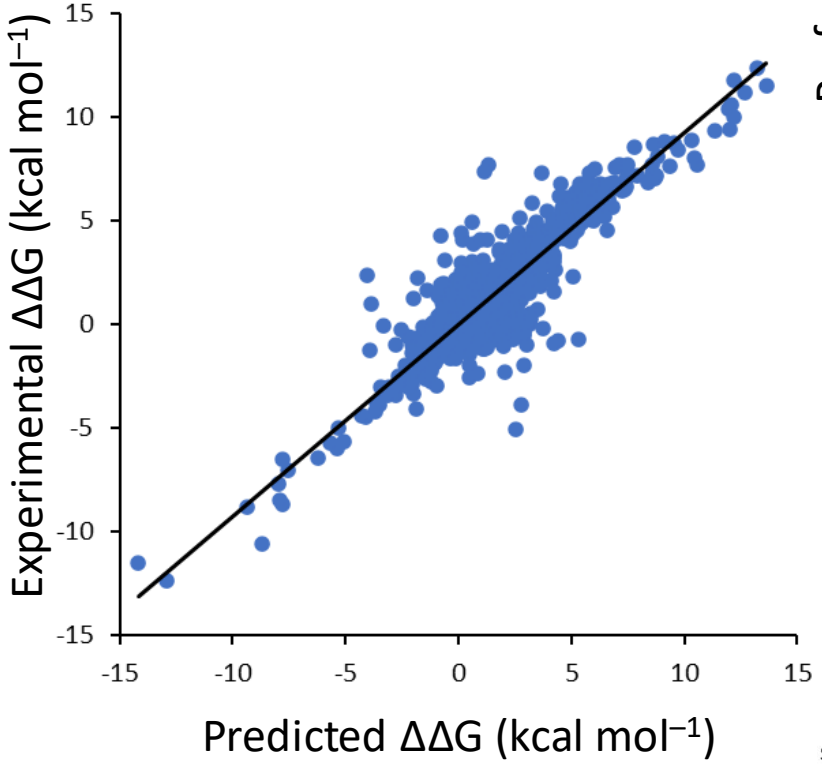
For example, signaling transduction in cells often passes messages via the mechanism of a protein binding to another molecule.

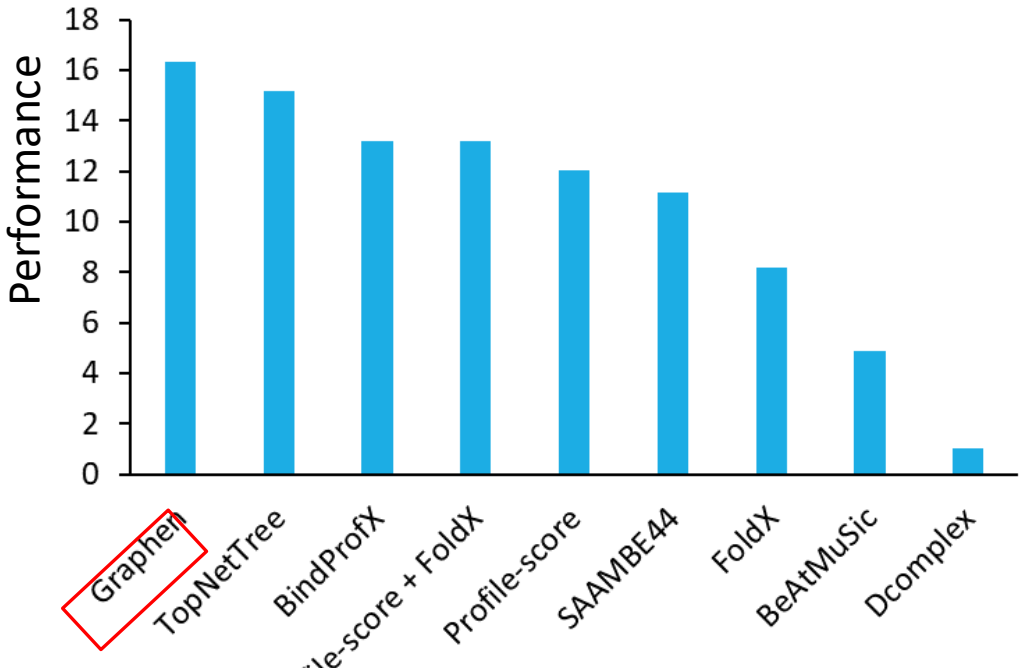
Other times, the molecule binding to the protein is foreign: possibly a drg we've created to manipulate the protein, possibly a toxin that interferes with its function.

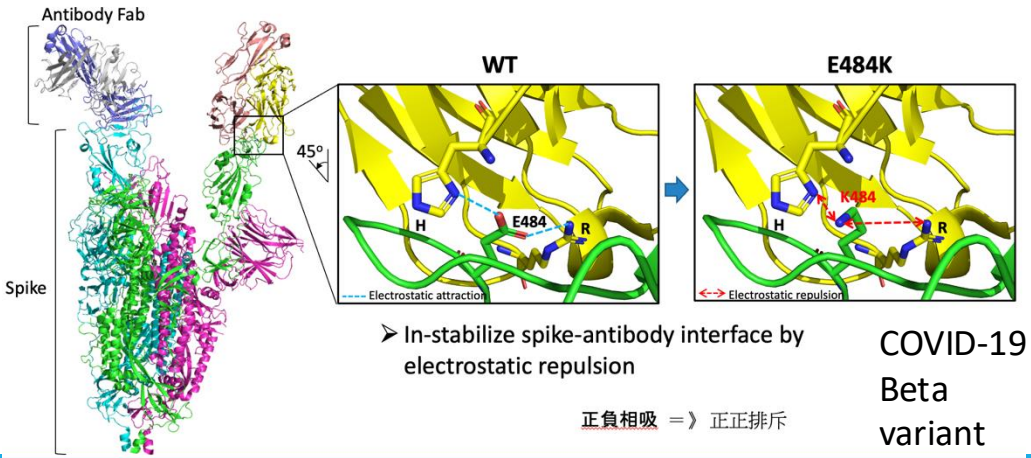


Example: Graphen binding energy prediction model is better than the current state-of-the-art in predicting ΔΔG

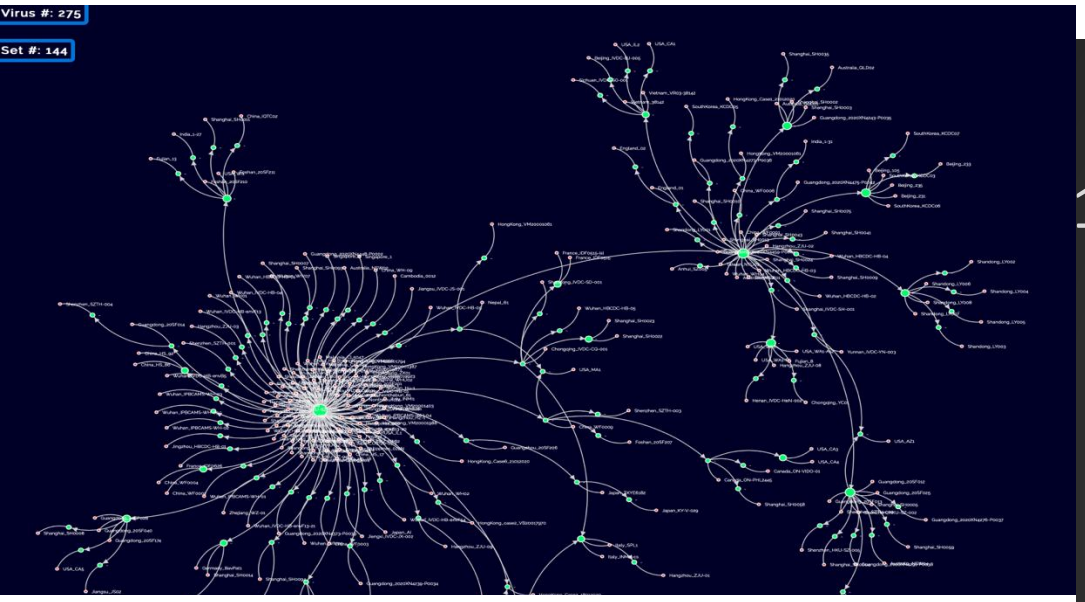
Models	*Pearson's r*
Graphen-Atom	0.914
TopNetTree	0.850
BindProfX	0.738
Profile-score + FoldX	0.738
Profile-score	0.675
SAAMBE44	0.624
FoldX	0.457
BeAtMuSic	0.272
Dcomplex	0.056

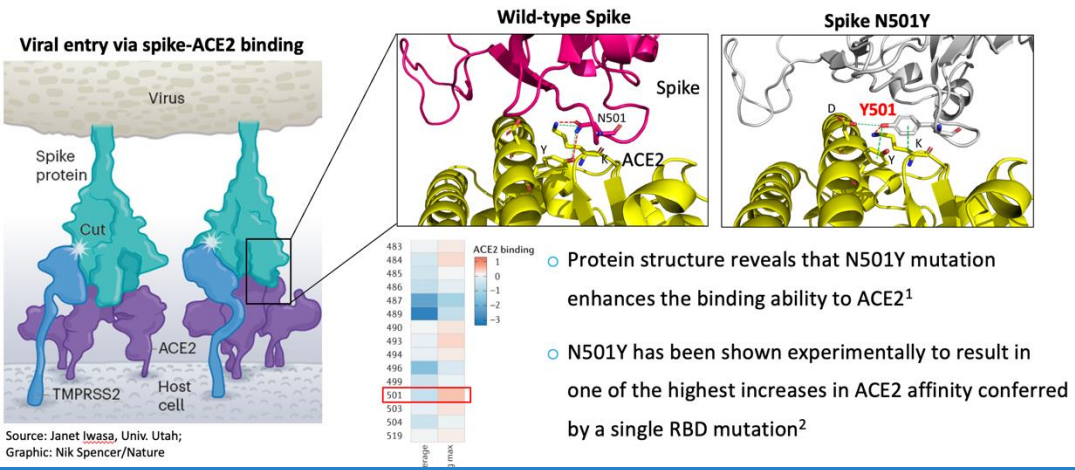


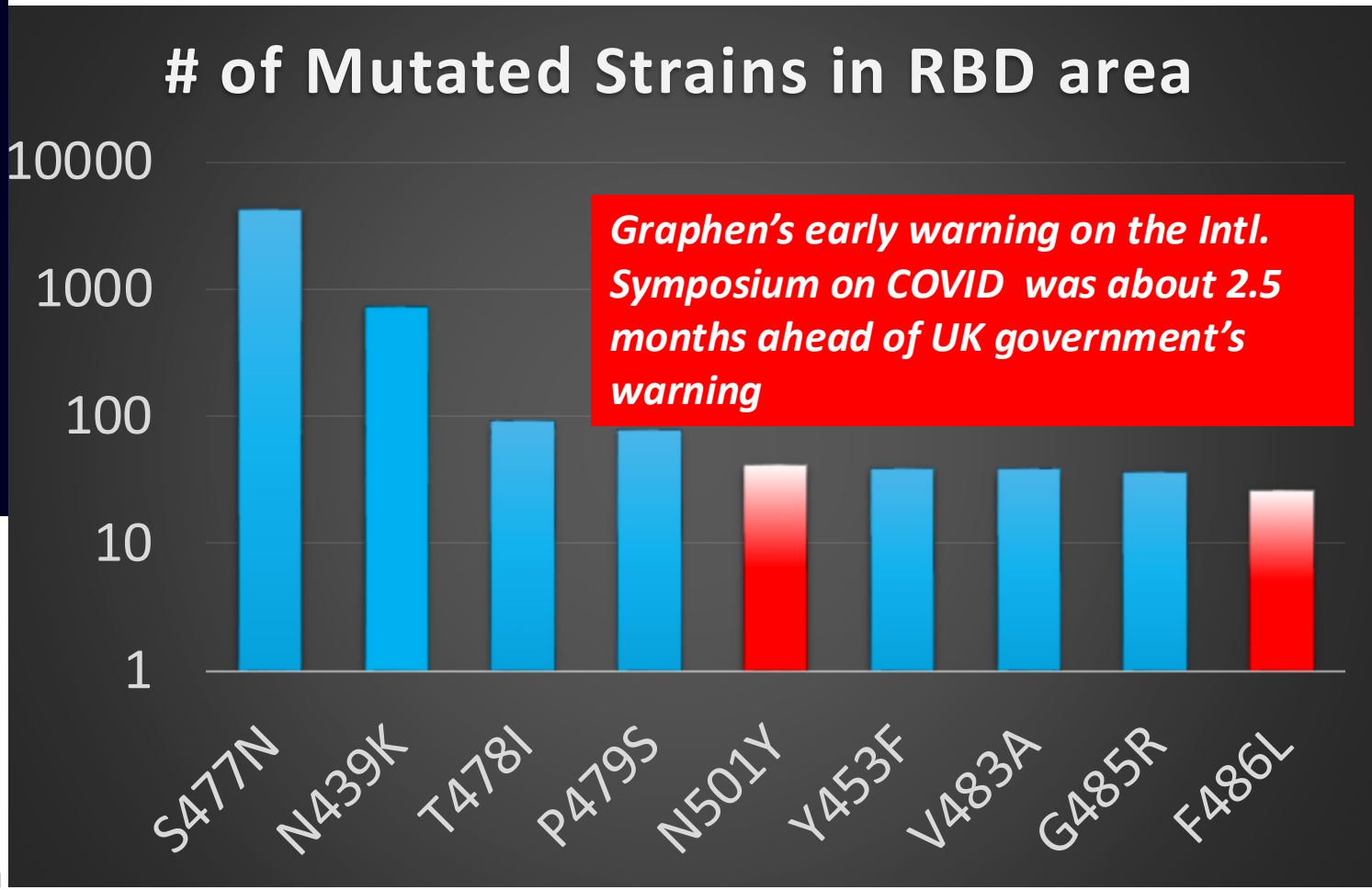




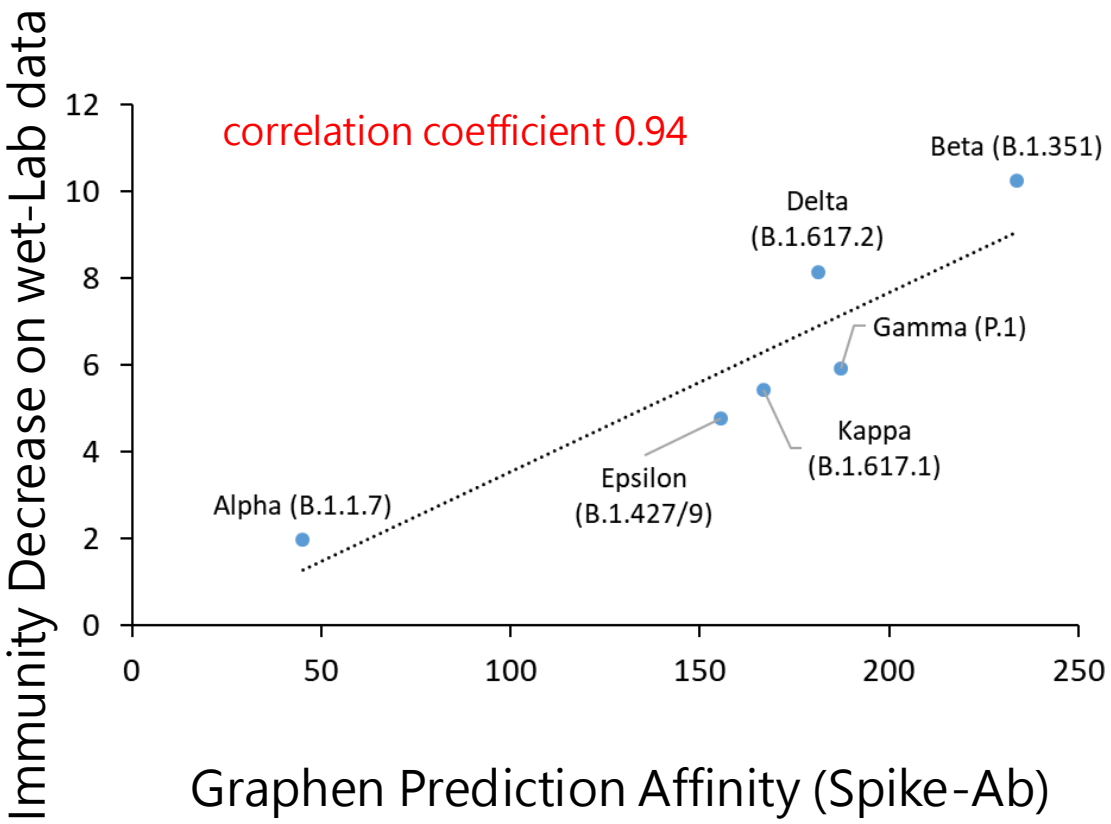
Graphen AI provided early warning of Alpha variants in September 2020, about 2.5 months ahead of first warning issued by the UK government

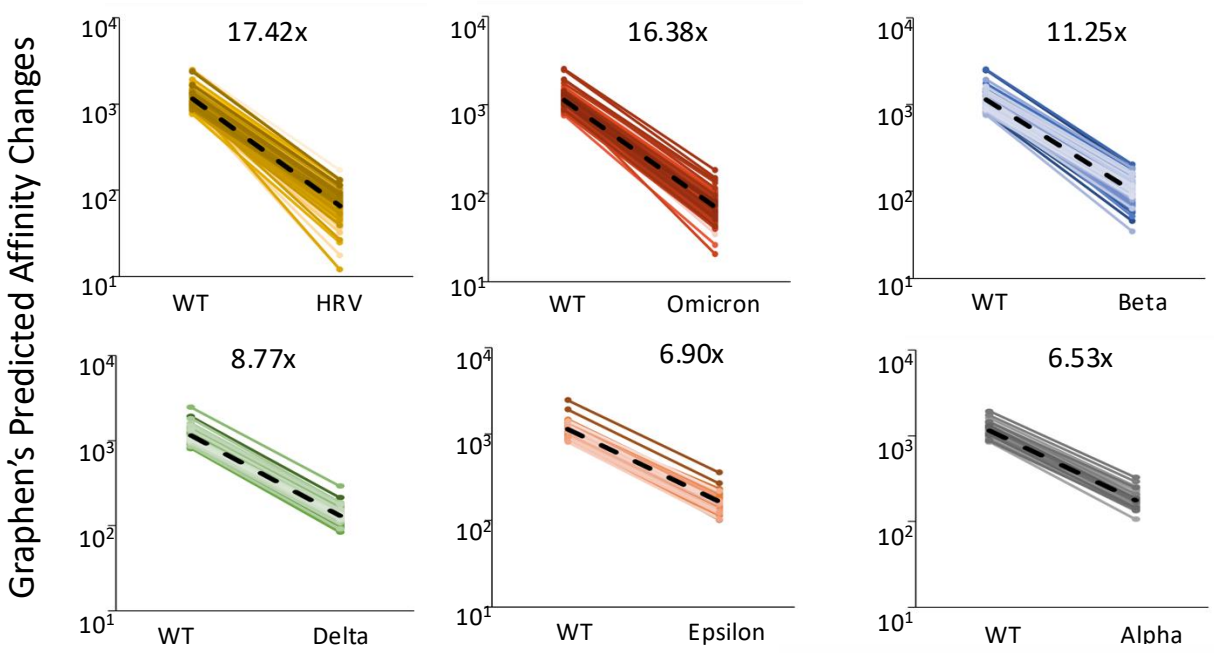




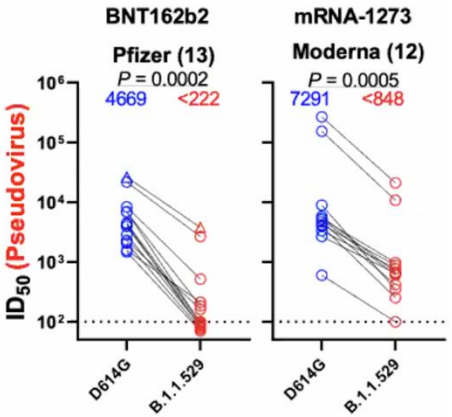


Graphen's prediction of variants' functions (perfectively) matched real-world wet-lab data

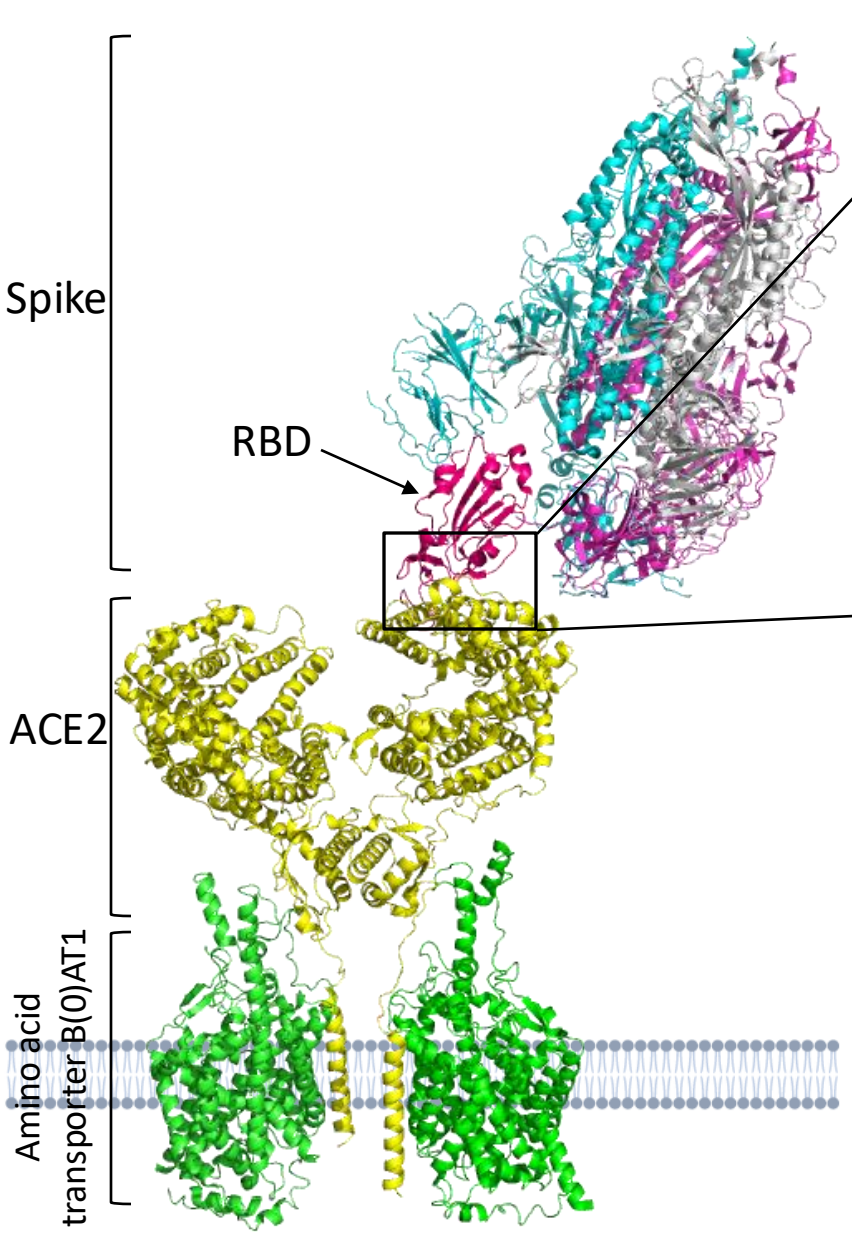




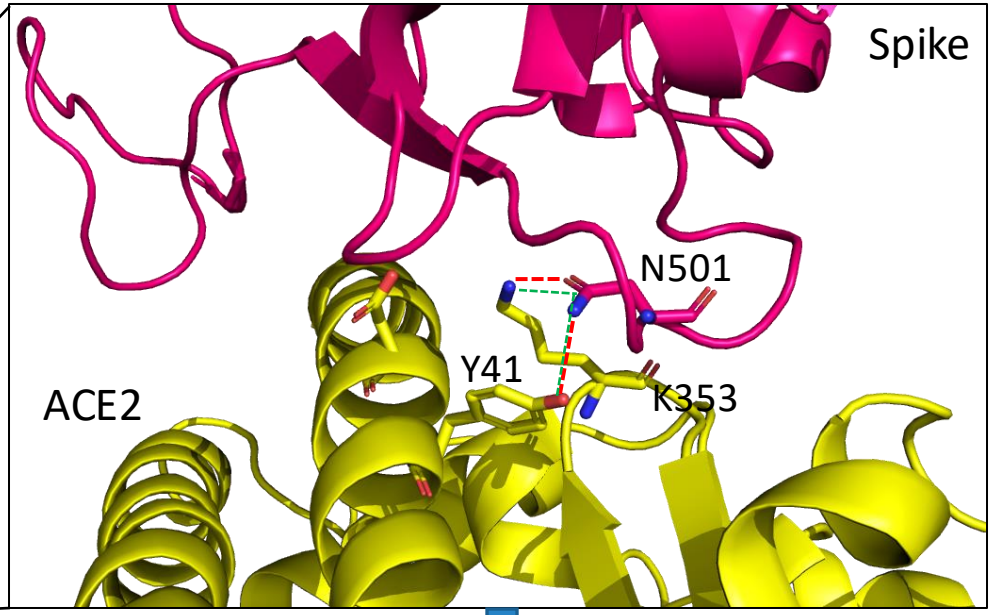
(Nov 26, 2021) Graphen predicted Omicron's immunity escape is 16.38x. (Dec 15, 2021) Columbia Univ Med School presented the vaccine efficacy is down 20x by Pfizer and 9x by Moderna.



Spike N501Y increase the binding ability to ACE2 (by Graphen ATOM)



90°



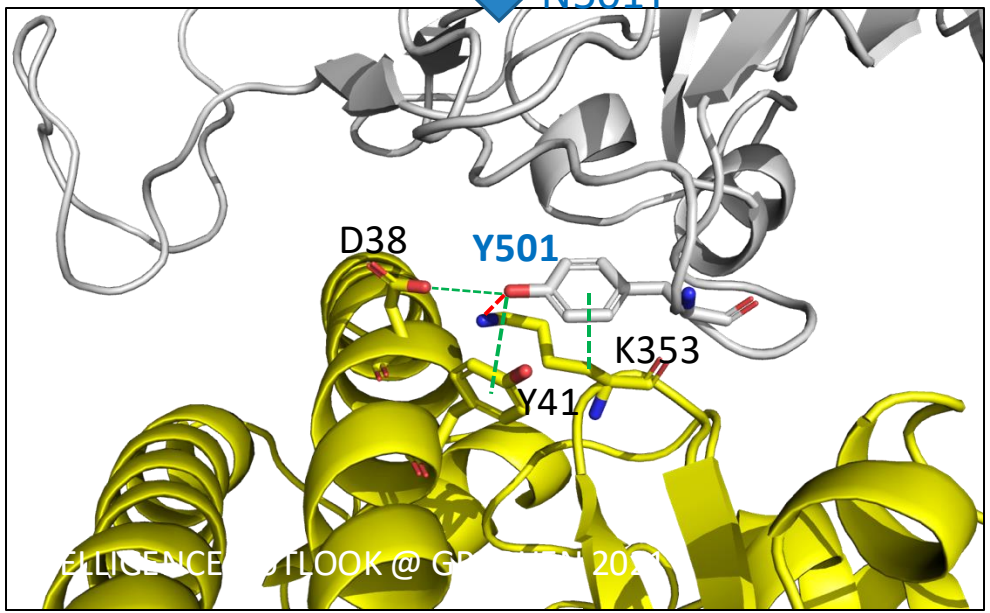
Energy

ACE2	Spike	VDW	HB
Y41	N501	-3.82	-2.46
K353	N501	-5.64	-3.5

Total energy: -15.42

--- Hydrogen bond
--- Van der Waals

N501Y

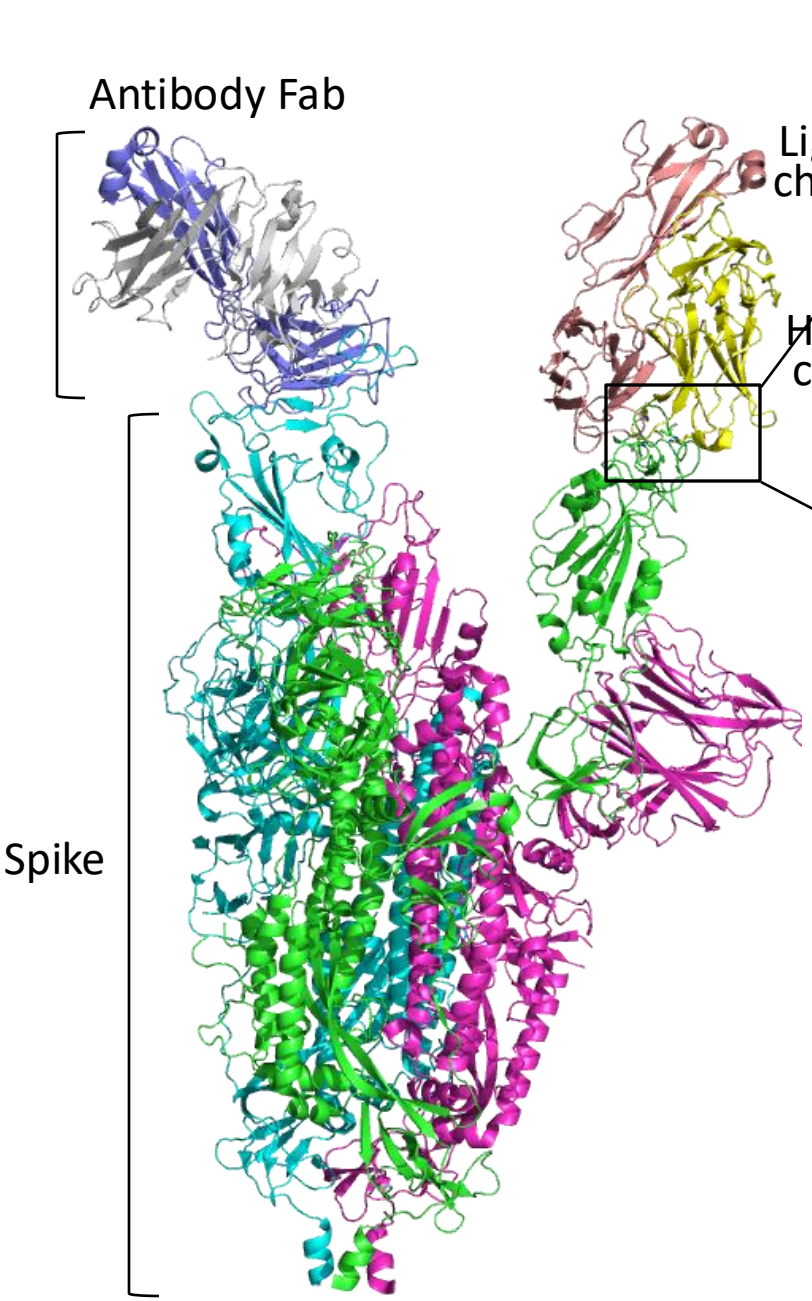


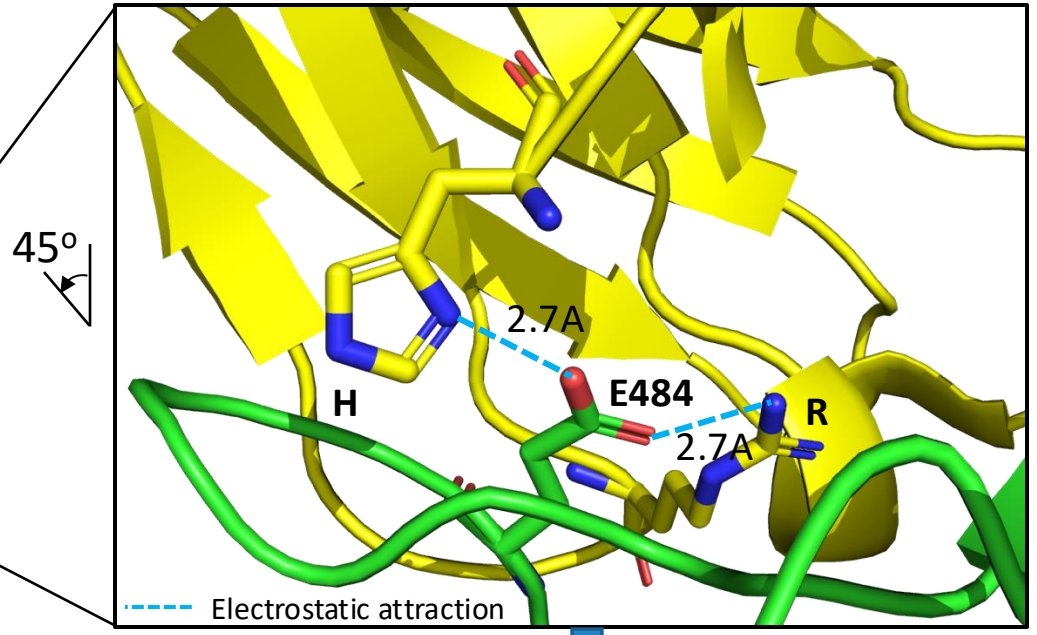
ACE2	Spike	VDW	HB
Y41	Y501	**-7.33**	0
K353	Y501	**-14.15**	-3.5
D38	Y501	-0.32	0

Total energy: -25.3

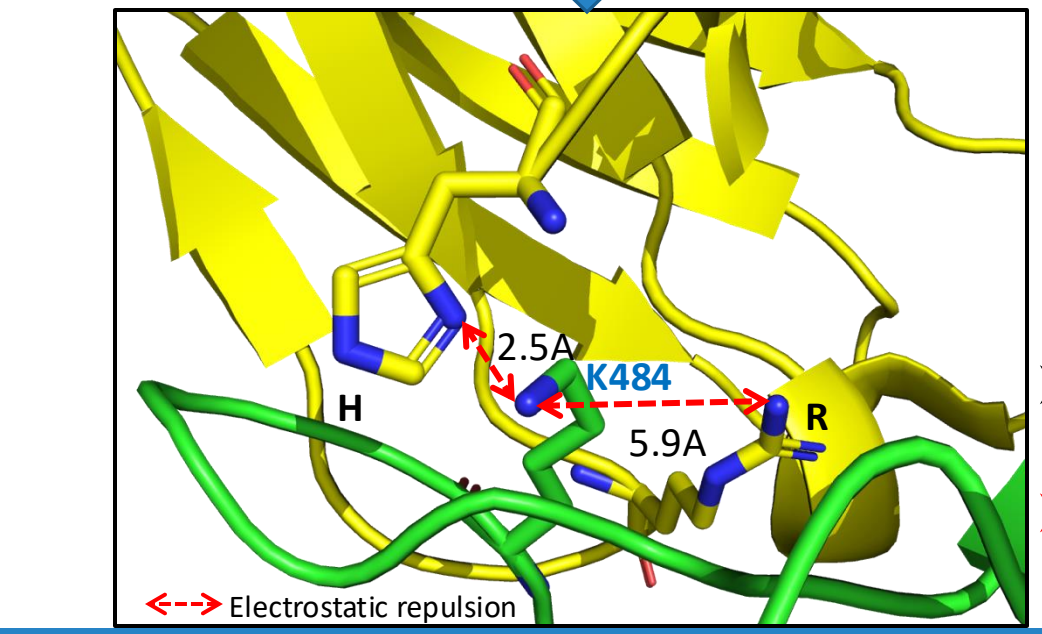
- Stabilize interface by increasing binding energy (1.6-fold) and number of contacting residues

Vaccine binding strength decreased on the 'South African' variants (by Graphen ATOM)





E484K



Spike	Anti-body	VDW	HB	Elec.
E484	H	-4.44	-4.20	-2.92
E484	R	-8.98	-8.31	-2.90

Total energy: -25.79

Spike	Anti-body	VDW	HB	Elec.
K484	H	-6.68	-1.36	4.15
K484	R	-10.03	0	0.71

Total energy: -13.21

- In-stabilize interface by electrostatic repulsion
- Binding strength ↓ 51%

Drug targets

Drugs Don't Just Target a Single Protein

As we've discussed, it can be extraordinarily useful to reduce the problem of designing a drug for a disease to the problem of designing a drug that interacts tightly with a given protein. But it's extremely important to realize that in reality, any given drug is going to interact with many different subsystems in the body. The study of such multifaceted interactions is broadly called polypharmacology.

At present, computational methods for dealing with polypharmacology are still relatively undeveloped, so the gold standard for testing for polypharmacological effects remains animal and human experimentation. As computational techniques mature, this state of affairs may shift over the next few years.

Graphen solved many of such issues.

Example: Graphen's High Selective and effective drug candidate

Graphen Drug Inhibitor efficacy

Target	IC50 (nM)
FLT3	1.99
CSF1R (FMS)	5.89
KIT	27.39
PDGFRA	23.44
PDGFRB	36.49
RET	27.27
KDR (VEGFR2)	159.94
PHKG1	>100
PHKG2	>100
CK1α1	>30000

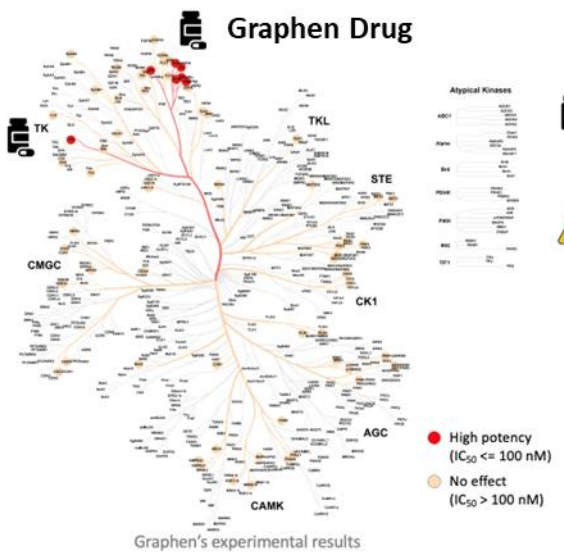
Selectivity

Target Name	Graphen Drug	Sunitinib
CK1α1	inactive	active
VEGFR2	inactive	active
LCK	inactive	active
PHKG1	inactive	active
PDGFRB	active	active
RET	active	active

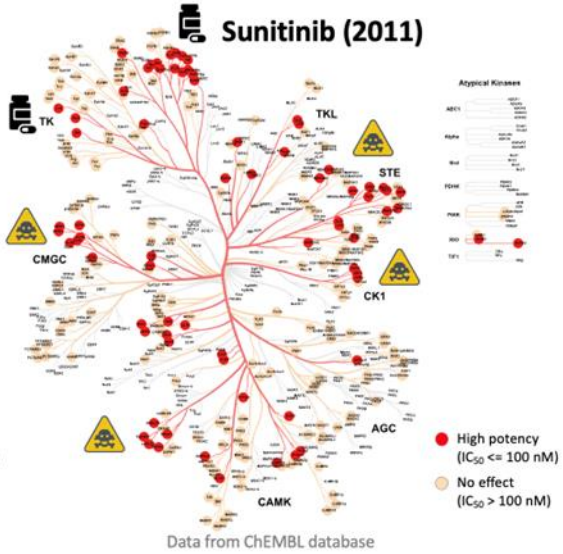
Side Effect

- Cardiovascular side effects
- Hypothyroidism
- Hepatotoxicity & Oxidative Stress

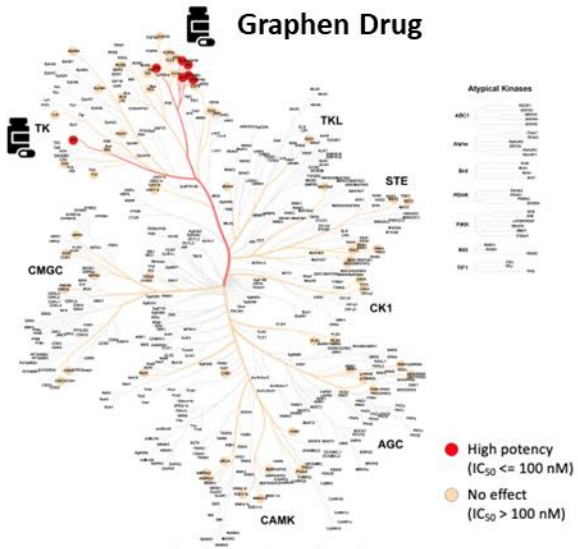
High Efficacy. Low Side-Effect. No Toxicity



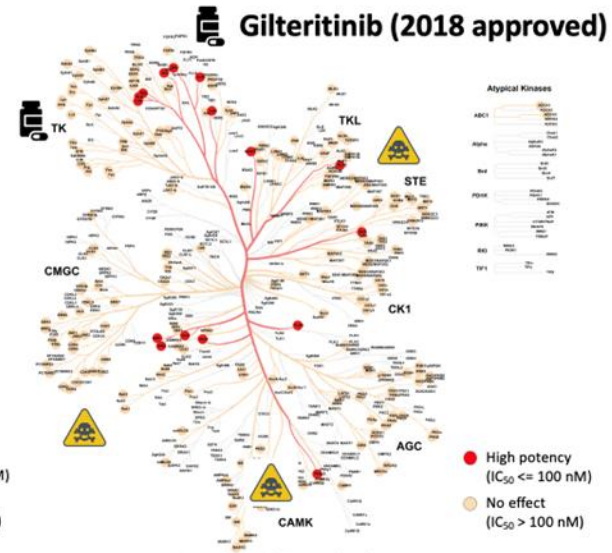
Graphen's experimental results



Data from ChEMBL database



Graphen's experimental results



Data from ChEMBL database

Biophysical Featurizations

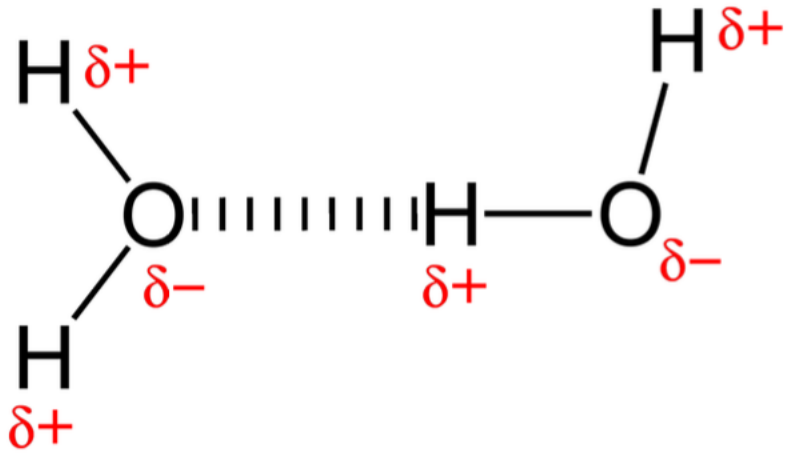
- If applying traditional machine learning methods, the first step is to transform (or featurize) training data to a format suitable for learning algorithms.
- The behaviors of biophysical systems are critically constrained by their 3D structures, so the 2D techniques miss crucial information.
- The alternative featurization technique is the atomic featurization, which simply provides a processed representation of the 3D positions and identities of all atoms in the system.
- This makes the challenge for the learning algorithm considerably harder.
- But it also makes it feasible for learning algorithms to detect new patterns of interesting behavior.

Grid Featurization

- By converting biophysical structures into vectors, we can use ML algorithms to make predictions about them.
- Ideally, a featurization technique would need to have significant knowledge about the chemistry of such systems.
- Those features might include, e.g., counts of noncovalent bonds between the protein and ligand, such as hydrogen bonds or other interactions. (Most protein-ligand systems don't have covalent bonds between the protein and ligand.
- The grid featurizer searches for the presence of chemical interactions within a given structure and constructs a feature vector that contains counts of these interactions.

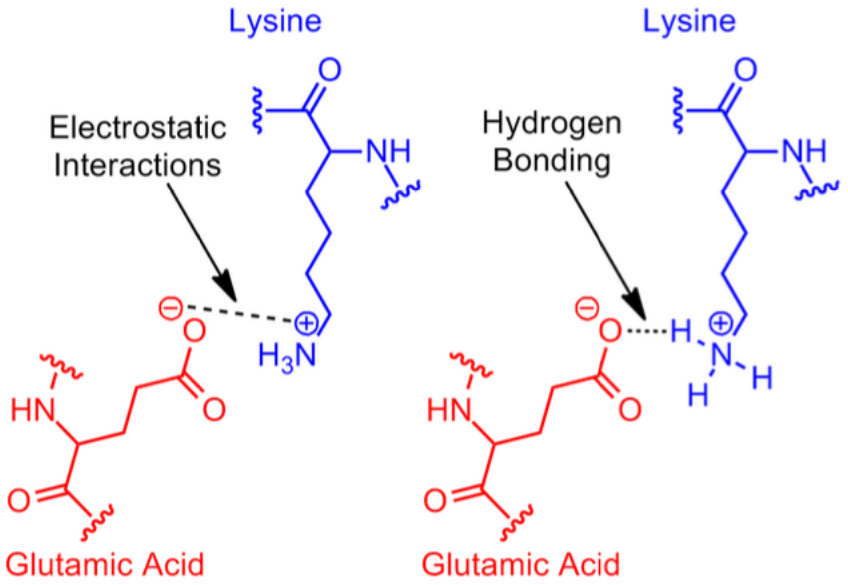
Hydrogen bonds

- When a hydrogen atom is covalently bonded to a more electronegative atom such as oxygen or nitrogen, the sharded electrons spend most of their time closer to the more electronegative atom.
- This leaves the hydrogen with a net positive charge.
- If that positively charged hydrogen then gets close to another atom with a net negative change, they are attracted to each other.
- This forms a hydrogen bond.
- Hydrogen atoms are so small, they can get very close to other atoms, leading to a strong electrostatic attraction.



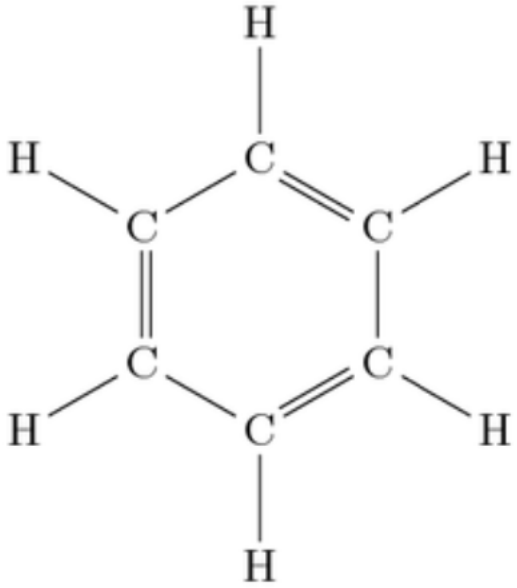
Salt bridges

- A salt bridge is a noncovalent attraction between two amino acids, where one has a positive charge and the other has a negative change.
- It combines both ionic bonding and hydrogen bonding.
- Although these bonds are relatively weak, they can help stabilize the structure of a protein by providing an interaction between distant amino acids in the proteins sequence.

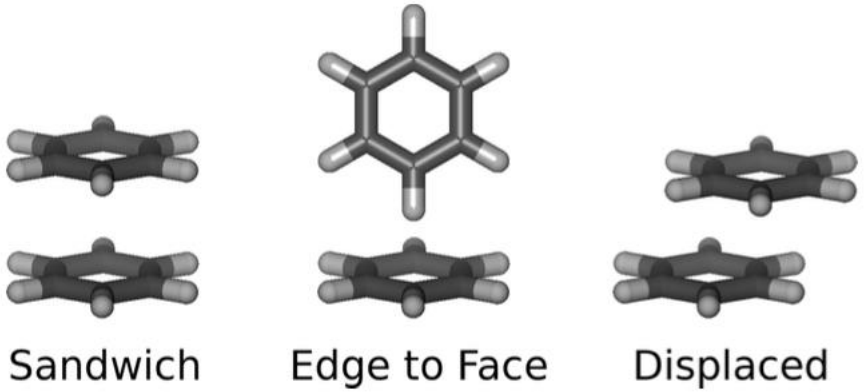


Pi-stacking interactions

- Pi-stacking interactions are a form of noncovalent interaction between aromatic rings.
- These are flat, ring-shaped structures that appear in many biological molecules, inclyding DNA and RNA.
- They also appear in the side chains of some amino acids, including phenylalanine, tyrosine, and tryptophan.



An aromatic ring in the benzene molecule.



Various noncovalent aromatic ring interactions.

PDBBind Case Study

- The PDBBind dataset contains a large number of biomolecular srystal structures and their binding affinities.
- A biomolecule is any molecule of biological interest.
- That includes not just proteins, but also nucleic acids (such as DNA and RNA), lipids, and smaller drug-like molecules
- Much of the richness of biomolecular systems results from the ineractions of various biomolecules with one anothe
- A binding affinity is the experimentally measure affinity of two molecules to form a complex, with the two molecule
- The PDBBind dataset has gathered structures of a number of biomolecular complexes.
- They include protein-ligand complexes, protein-protein, protein-nuclelic acid, and nucleic acid-ligan complexes.
- For instance, the protein-ligand dataset includes 15000 such complexes.
- The learning challenge for the PDBBind dataset is to predict the binding affinity for a complex given the protein-liga

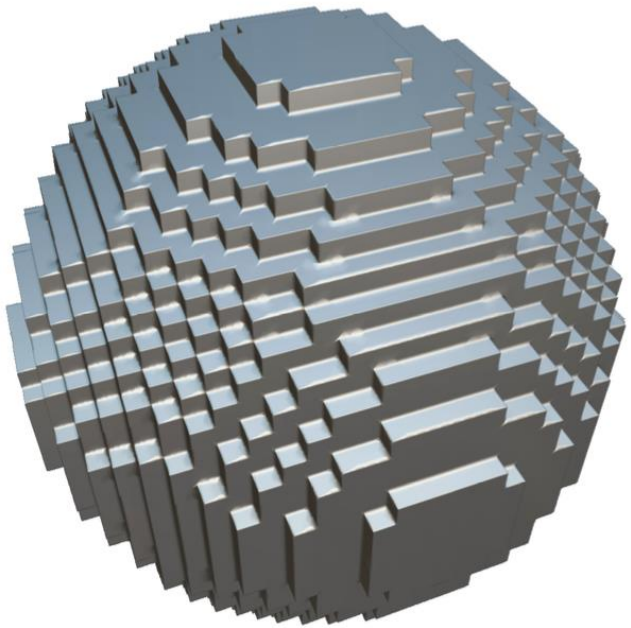
PDBBind

A visualization of the 2D3U protein-ligand complex from the PDBBind dataset. The protein is represented in cartoon format for ease of visualization, and that the ligand (near the top-right corner) is represented in ball-and-stick format for full detail.

Featuring PDBBind Dataset

```
import deepchem as dc
featurizer = dc.feat.RdkitGridFeaturizer(
        voxel_width=2.0, sanitize=True, flatten=True,
        feature_types=['hbond', 'salt_bridge', 'pi_stack',
                       'cation_pi', 'ecfp', 'splif'])
```

```
tasks, datasets, transformers = dc.molnet.load_pdbbind(
        featurizer=featurizer, splitter="random", subset="core")
train_dataset, valid_dataset, test_dataset = datasets
```



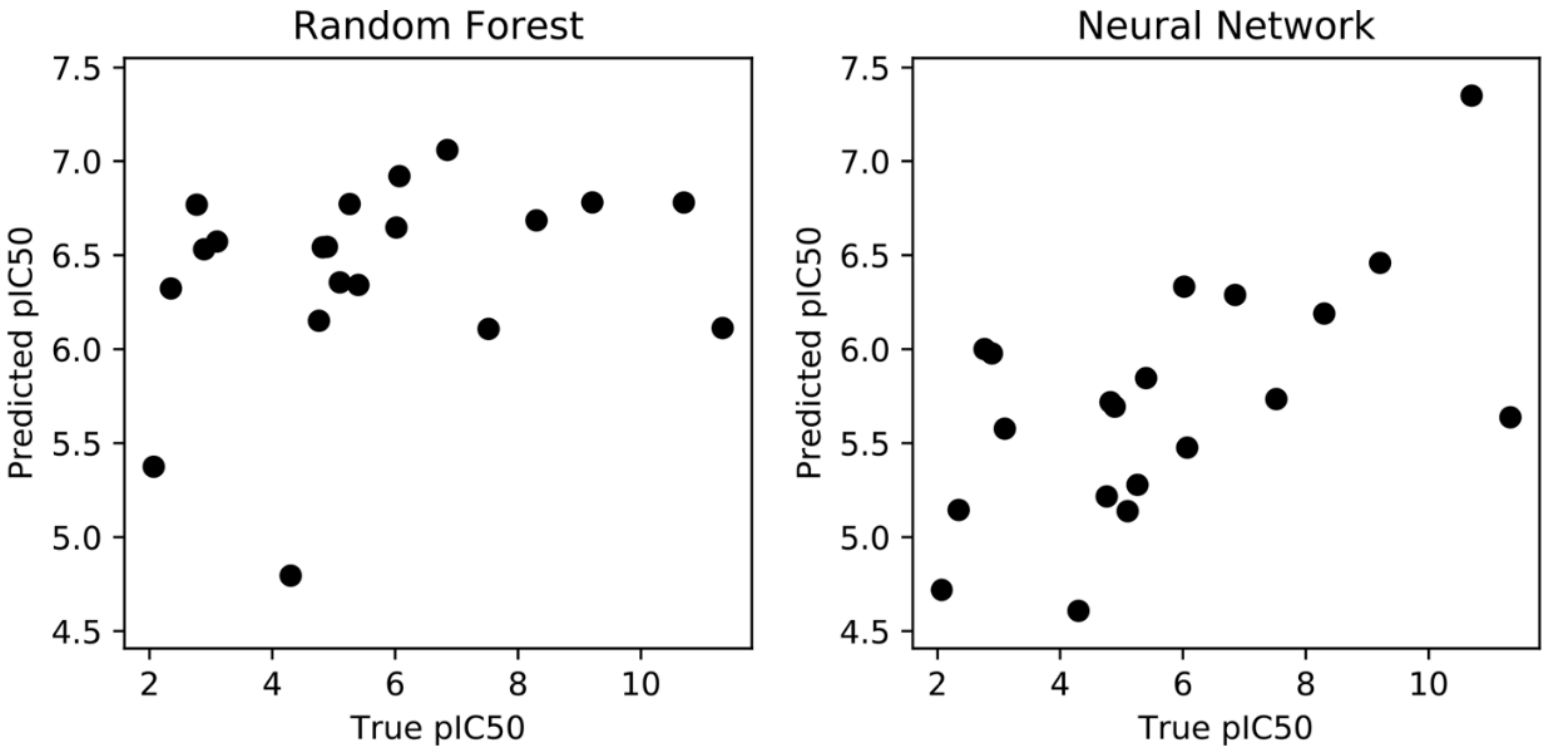
PDBBind Model

```
from sklearn.ensemble import RandomForestRegressor
sklearn_model = RandomForestRegressor(n_estimators=100)
model = dc.models.SklearnModel(sklearn_model)
model.fit(train_dataset)
```

```
n_features = train_dataset.X.shape[1]
model = dc.models.MultitaskRegressor(
        n_tasks=len(tasks),
        n_features=n_features,
        layer_sizes=[2000, 1000],
        dropouts=0.5,
        learning_rate=0.0003)
model.fit(train_dataset, nb_epoch=50)
```

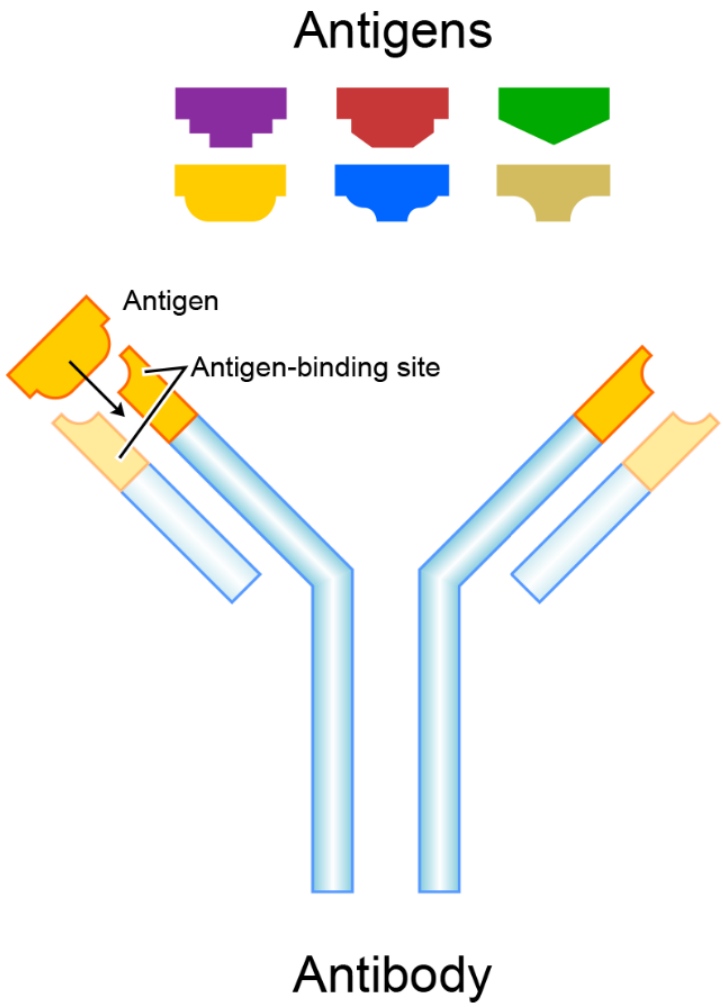
PDBBind Model Evaluation

```
metric = dc.metrics.Metric(dc.metrics.pearson_r2_score)


print("Evaluating model")
train_scores = model.evaluate(train_dataset, [metric], transformers)
test_scores = model.evaluate(test_dataset, [metric], transformers)

print("Train scores")
print(train_scores)

print("Test scores")
print(test_scores)
```



Ongoing Protein Modeling & Structure Prediction tasks

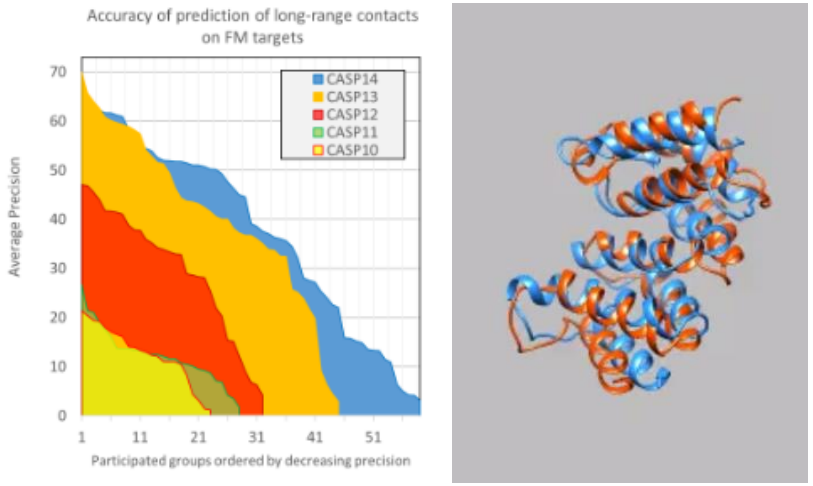
- Protein-protein and protein-DNA systems follow the same basic physics as protein-ligand systems at a high level.
- Many of the physical interactions that drive protein-ligand interactions are driven by charged dynamics.
- Protein-protein dynamics, may be driven more by bulk hydrophobic interactions.

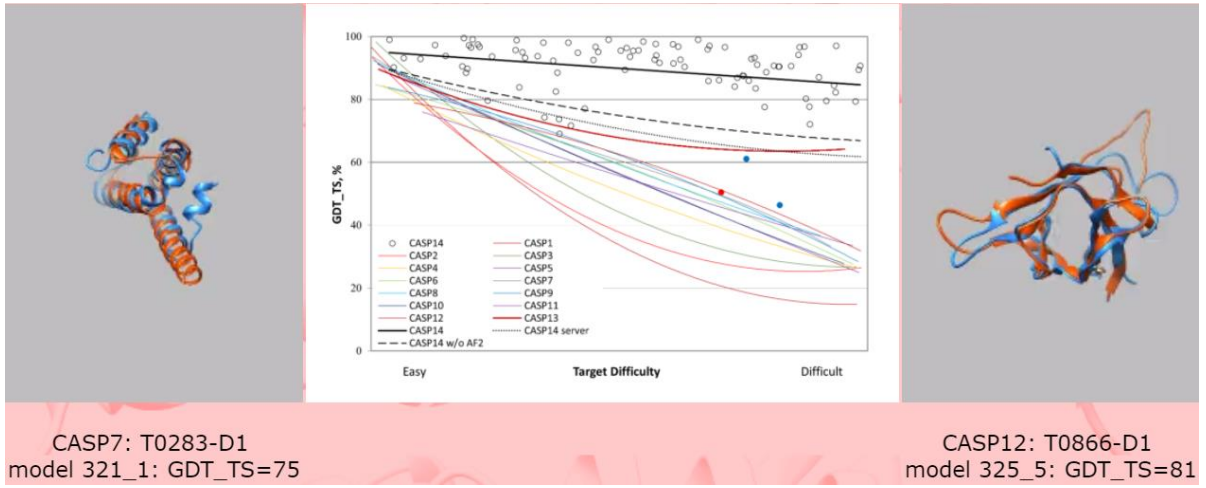


CASPs : Critical Assessment of protein Structure Prediction

- CASP is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994, the primary goal of CASP is to help advance the methods of identifying protein three-dimensional structure from its amino acid sequence.

- Google AlphaFold 2 scored above 90 for around 2/3 of the proteins in CASP14's global distance test (GDT). Since this, applying predicted protein structure to precise analysis of target research, mechanism revealing, and even drug developments became convincingly



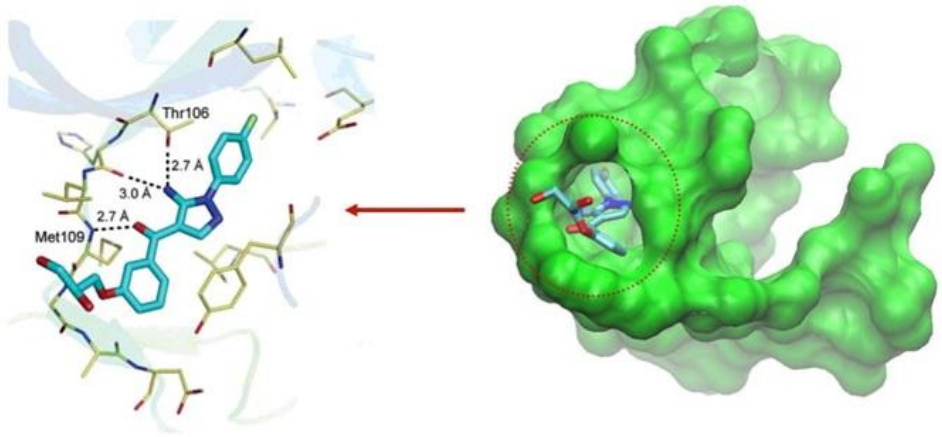


CASP7: T0283-D1
model 321_1: GDT_TS=75

CASP12: T0866-D1
model 325_5: GDT_TS=81

CASP16 and Beyond : Next stage of structural prediction

- For the outstanding records that Google Alphafold2 has achieved in CASP14 in **single protein structural prediction**, this year the competition goes into the new stage: more complex conformation prediction. Majorly including:
  - **Multi-chain protein structures**
  - **Protein-Ligand complexes**
  - **RNA structures and complexes**
  - **Protein conformational ensembles**

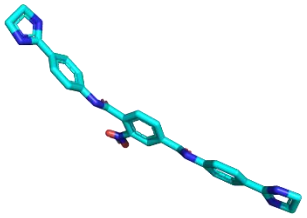


Protein-Ligand Interaction

Protein-Nucleotide complexing

Drug-Target Interactions and complexes

- In small molecular drug development, to reach viable drug efficacy, it's crucial to clarify how the **drug molecule interacts with designed targets**, as well as the consequences induced by the conjugation of drug-target pairs .
- Drug-target interaction can be predicted with majorly two ways in traditional: ligand docking and quantum-physic simulation. These methods are limited with lacking flexibility to know the induced conformational change and extremely high cost of time and computational resources, respectively.

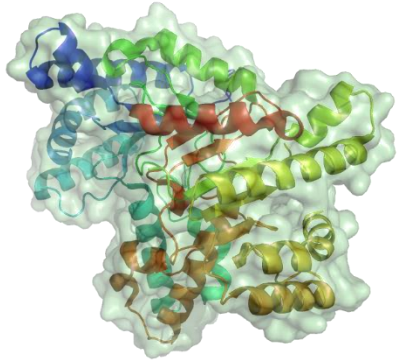


Ligand

+

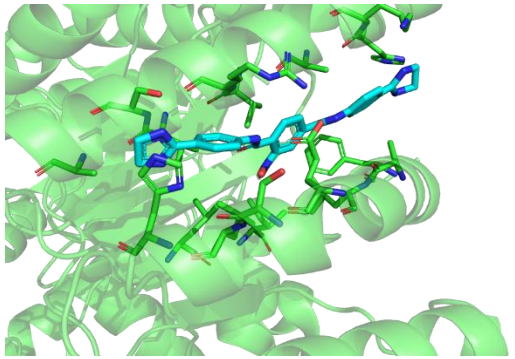
>Protein A
SGFRKMAFPSGKVEGCMVQVTCG
TTTLNGLWLDDVVYCPRHVICTSE
DMLNPNYEDLLIRKSNHNFLVQAG
NVQLRVIGHSMQNCVLKLKVDTA
NPKTPKYKFVRIQPGQTFSVLACY
NGSPSGVYQCAMRPNFTIKGSFLN
GSCGSVGFNIDYDCVSFCYMH….

Protein sequence



Protein structure

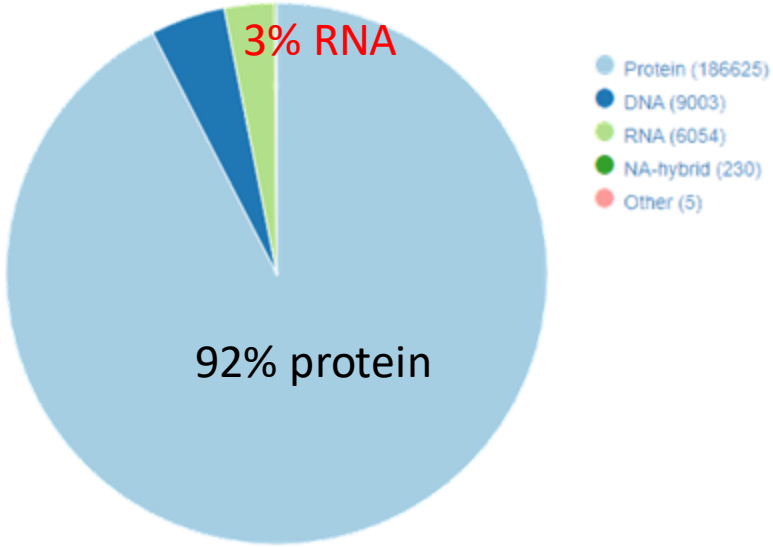
+



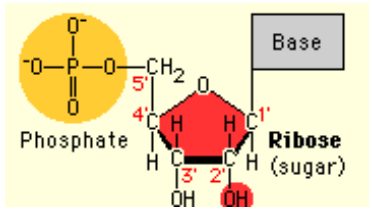
Protein-ligand interaction

RNA structures and complexes predictions

- RNA molecules play fundamental roles in cellular processes. Their function and interactions with other biomolecules are dependent on the ability to form complex three-dimensional (3D) structures
- Experimental determination of **RNA atomic 3D structures** is laborious and challenging
  - Only **3%** of known RNA 3D structures in public database



RNA basic unit:
Nucleotide



Primary Sequence
UGCUCCUAGUACGAGAGGACCGGAGUG…

How RNA folds?

3D structure