Exploring Deep Neural Network in Material Science

Week 1

19/10/2022

Exploring Deep Neural Network in Material Science

Objective: Single-sequence Protein Structure Prediction

protein sequence -> 3D structure

Protein folding

The physical process by which a protein chain is translated to its native three-dimensional structure, typically a "folded" conformation by which the protein becomes biologically functional. Via an expeditious and reproducible process, a polypeptide folds into its characteristic three-dimensional structure from a random coil.

Protein Data Bank (PDB)

Traditional determination of protein 3D structure

X-ray Diffraction (XRD)
Nuclear Magnetic Resonance (NMR)
Cryogenic (frozen tissue) Electron Mycroscopy (CryoEM)

Challenges: laborious, difficulty with large protein molecules/compounds

Traditional Modelling Method:

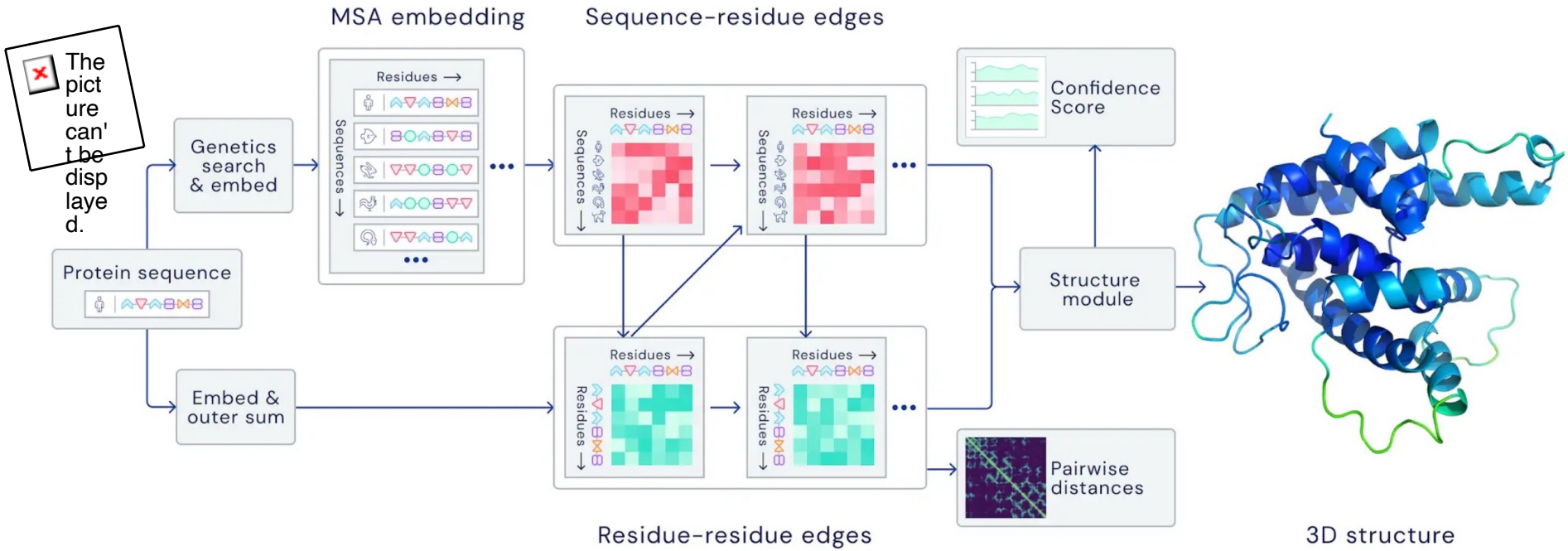
1. **Physics-based** theoretical energy functions
2. **Knowledge-based** statistical energies

Why Deep Learning?

Deeper and automated feature engineering on top of machine learning
Able to capture the complex interactions present with biological molecules

Neural-network based: CNN, RNN, GAN, GNN
Non-neural Network based: BN, VAE, diffusion

Deepmind AlphaFold 2



Multiple Sequence Alignment (MSA) is the alignment of three or more biological sequences of similar length to infer homology can be inferred and to study the evolutionary relationship between the sequences. Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

Utilise Co-evolutionary Data by Multiple Sequence Alignment

Input: protein sequence

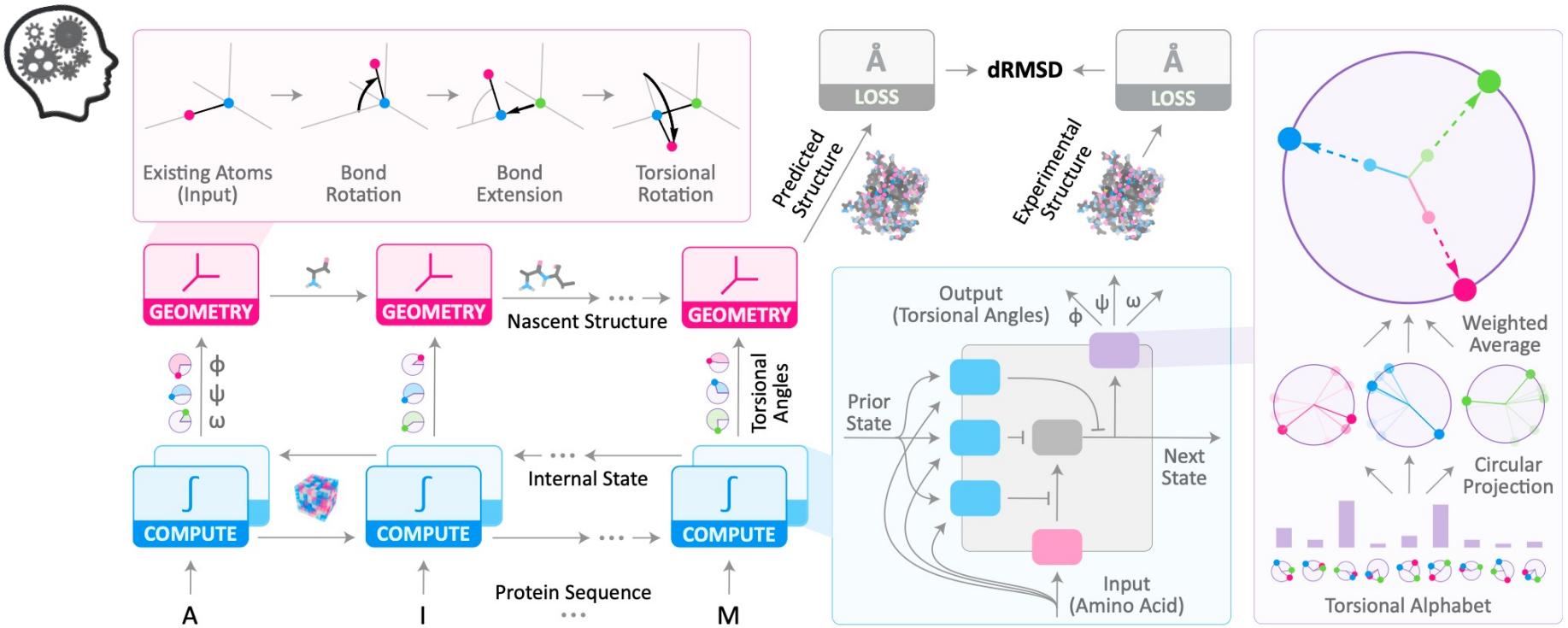
Output: protein 3D structure (x, y, z)

https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

AlphaFold 2 utilises co-evolutionary data by multiple sequence alignment
Homologous proteins come from a common evolutionary origin
MSA searches the database for protein sequences that are homologous to this protein.

What about proteins lacking homologous evolution information?
Most artificially designed protein drugs and enzymes for industrial synthesis, all of which have never existed in the history of biological evolution. Like **antibodies** (produced by the body's immune system in response to antigenic stimuli stress, no evolutionary information) and **orphan proteins.**

Recurrent Geometric Network (AlQuraishi, 2019)



Utilize Multiple Homogenous Sequence

Computation units based on LSTM (recurrent neural network)

Input: amino acid residue in the protein sequence

Output: three torsion angles of the protein backbone

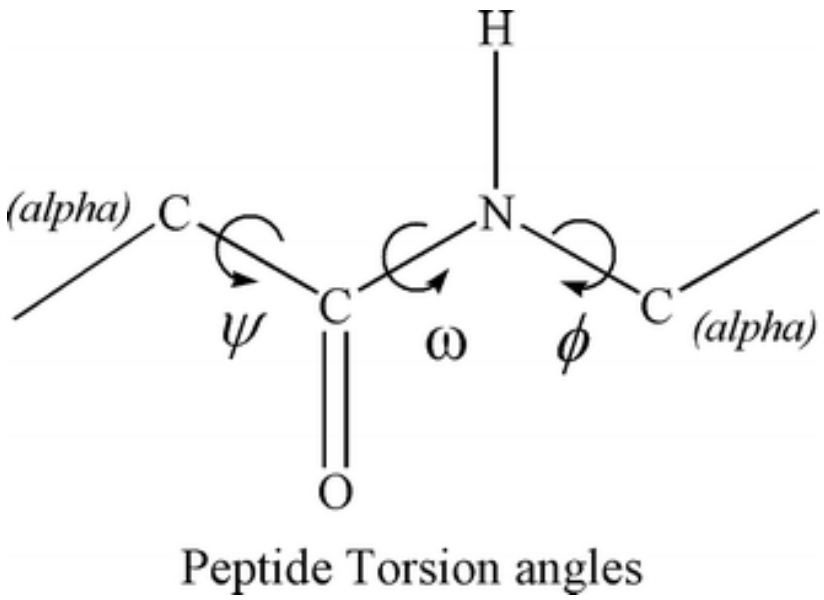
Github: https://github.com/aqlaboratory/rgn

https://www.cell.com/cell-systems/pdf/S2405-4712(19)30076-6.pdf

Recurrent Geometric Network (AlQuraishi, 2019)

Protein Structure by Position-Specific Scoring Matrices (PSSMs)

the backbone structure of a protein can be parameterized as a sequence of three dihedral angles ϕ, ψ, ω



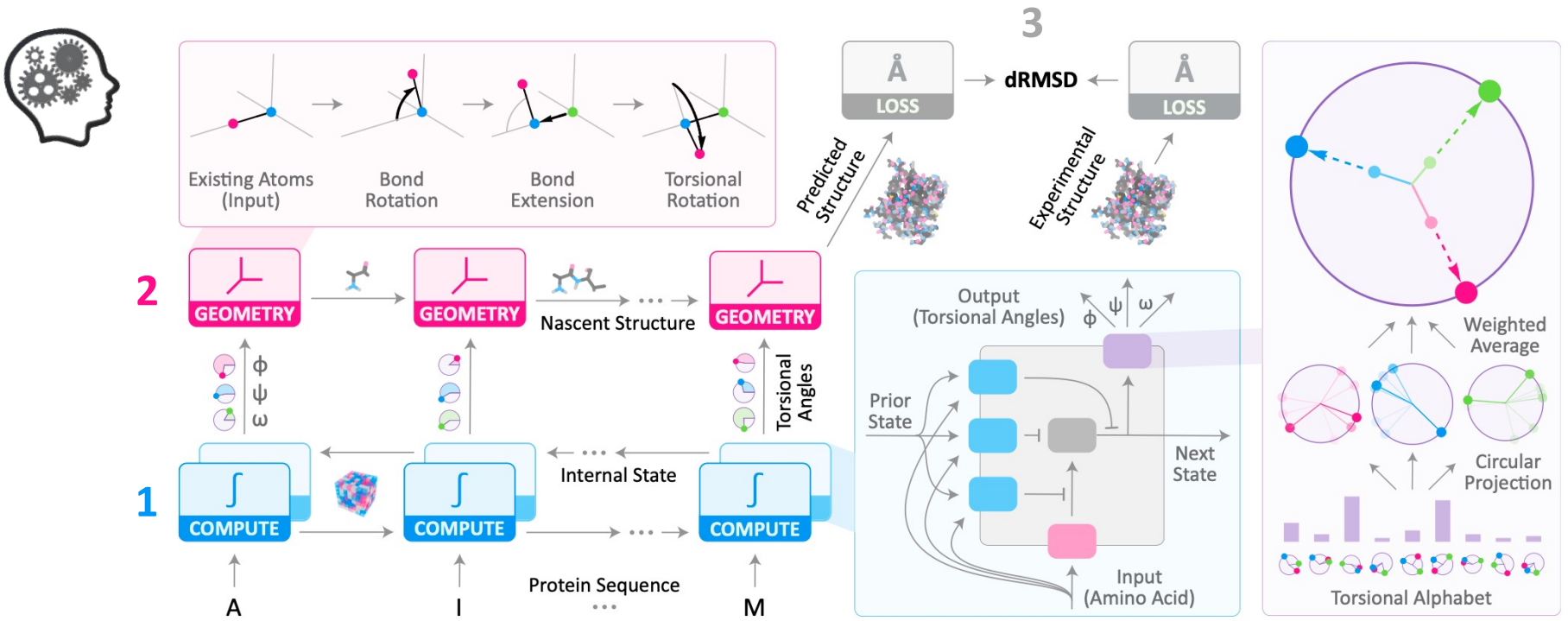
The Ramachandran Plot

Local Structural Parameters Controlling Protein Folding

ϕ (Phi)	Rotation of polypeptide backbone between N-Cα [-180,180]
ψ (Psi)	Rotation of polypeptide backbone between Cα-C [-180,180]
ω	Peptide bond between C and Ncan, assumed to be fixed at 180 (trans) or 0 (cis)

Github: https://github.com/aqlaboratory/rgn
https://www.cell.com/cell-systems/pdf/S2405-4712(19)30076-6.pdf

Three-stages of Recurrent Geometric Network



1. **Computation:** Each amino acid (residue) in the protein sequence is entered as a computing unit. Output three torsion angles for one residue
2. **Geometry:** Take three torsional angles for each residue to predict partial backbone, output a new backbone extended by one residue until 3D structure is completed
3. **Assessment:** Distance-based root mean square deviation between predicted and experimental structures

Github: https://github.com/aqlaboratory/rgn
https://www.cell.com/cell-systems/pdf/S2405-4712(19)30076-6.pdf

Utilize Multiple Homogenous Sequence

Still rely on availability of homogenous sequence

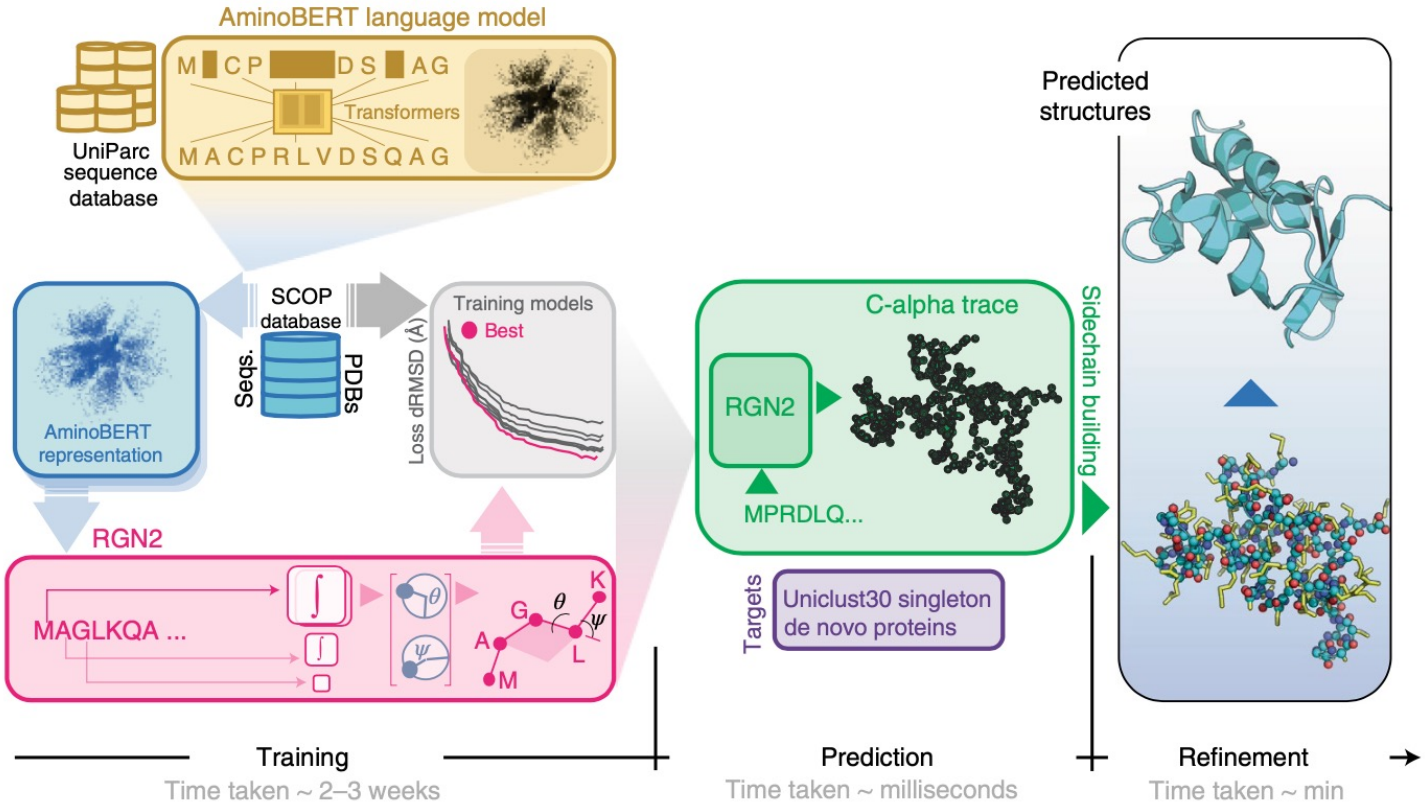
Enough input information?

PSSMs(ϕ, ψ, ω) are much weaker than multiple sequence alignments. They are based on single residue mutation frequencies and ignore how each residue mutates in response to all other residues.

Github: https://github.com/aqlaboratory/rgn

https://www.cell.com/cell-systems/pdf/S2405-4712(19)30076-6.pdf

Recurrent Geometric Network 2 (Chowdhury, 2022)



Use pretrained language model **AminoBERT** to extract latent information on protein sequence

Describe protein geometry based on the Frenet–Serret formulas, which is translationally and rotationally invariant

https://doi.org/10.1038/s41587-022-01432-w

Experimental Design

Input
protein sequence (i.e. k-mers/n-grams)
geometry (PSSMs),
molecular structure (i.e. SMILES string)

Encoding/Embedding
Word2Vec, PLM, GNN

Output
protein 3D structure (x, y, z)

Metrics
Global Distance Test–Total Score (GDT_TS)
distance-based Root Mean Squared Deviation (dRMSD)

SE(3)-equivariant Networks: From Equiformer to EquiFold

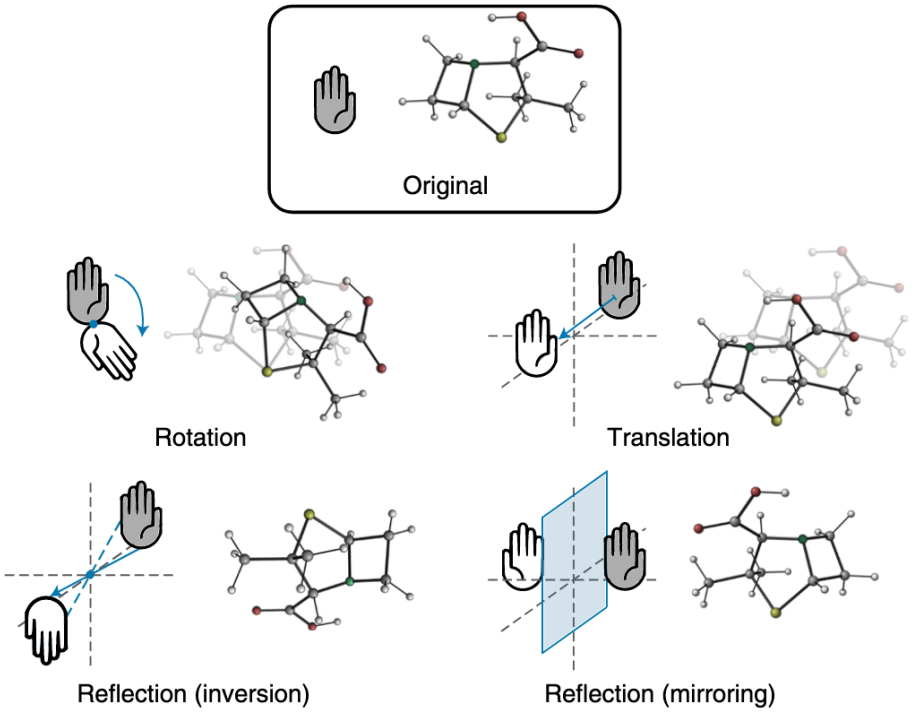
Week 2

26/10/2022

E(3) Equivariance

Molecular Coordinate System via the Symmetry of 3D Euclidean Space

3D Euclidean group E(3): 3D translation, 3D rotation, and inversion



"The laws of physics are invariant to the choice of coordinate systems and therefore properties of atomistic systems are **equivariant**."

Energy of an atomistic system should be constant regardless of how we shift the system

https://www.nature.com/articles/s42256-021-00418-8.pdf?proof=t
https://arxiv.org/pdf/2206.11990.pdf

SE(3)-equivariant Networks: From Equiformer to EquiFold

Equiformer

- Equivariant feature based on irreducible representations of the SO(3) group (the group of rotations in R3)
- Molecular property prediction

EquiFold

- Equivariant feature based on geometric tensors
- Molecular structure prediction

https://arxiv.org/pdf/2206.11990.pdf
https://www.biorxiv.org/content/10.1101/2022.10.07.511322v1.full.pdf

Group Theory and Equivariant Models

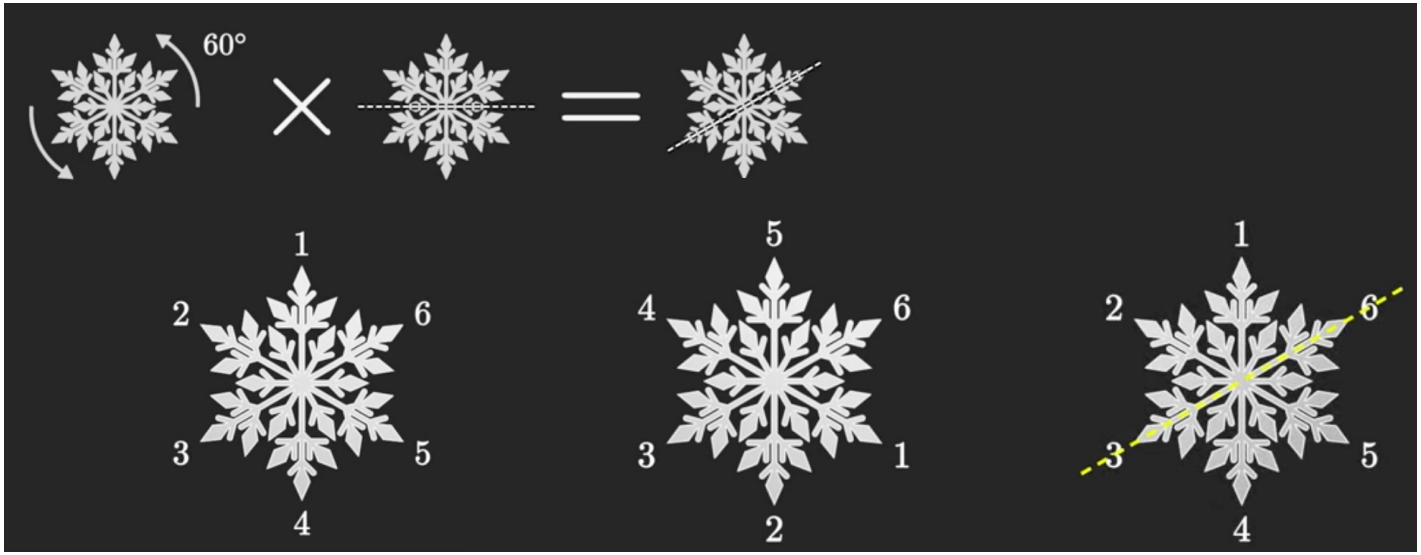
Week 3

02/11/2022

Group and Group Action

Group is an abstraction of **Symmetry Actions**
Number is an abstraction of **Counts**

Symmetry emerges when different ways of representing something “mean” the same thing. Symmetry of geometric object is about the actions (translation, rotation and inversion) to leave it look the same.



Rotate by 60 degrees and flip by x-axis == flip about its diagonal
A collection of all these actions taken together constitute a Group G

Encoding Symmetry

Deep Neural Networks are designed for different data types

Arrays ⇨ Dense NN
Components are independent. (No symmetry)

2D images ⇨ Convolutional NN
Spatial data. The same features can be found anywhere in an image. Locality. (2D-translation symmetry)

Text ⇨ Recurrent NN
Sequential data. Next input/output depends on input/output that has come before. (time-translation symmetry)

Graph ⇨ Graph (Conv.) NN
Topological data. Network passes messages between neighbouring nodes connected via edges. (permutation symmetry)

3D physical data ⇨ Euclidean NN
Data in 3D Euclidean space. Freedom to choose coordinate system. (3D Euclidean symmetry E(3): 3D rotations, translations, and inversion)

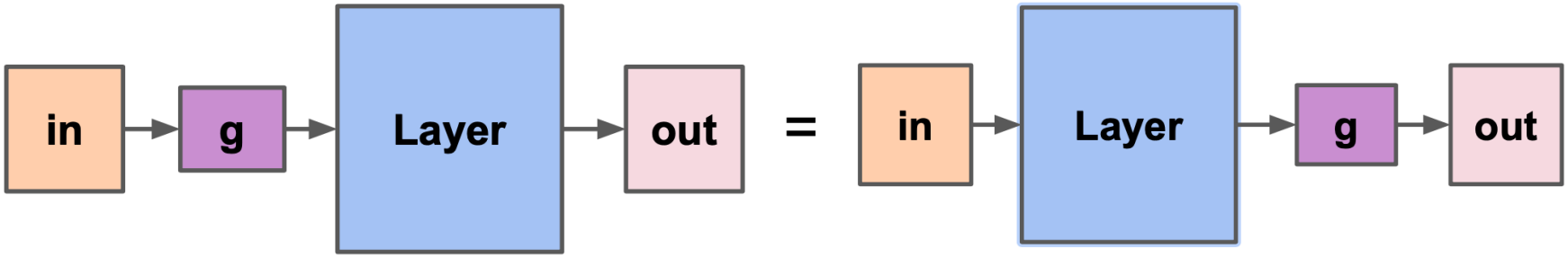
How to make models “symmetry-aware” for 3D data ?

1. Data Augmentation
Feed data to straight to the model and see what happens at the problem. But computationally expensive for 3D data

2. Invariant Inputs
Convert the data to invariant representations

3. Invariant models & Equivariant models
Construct the model that can handles coordinates

Invariant Models & Equivariant Models



For a function to be equivariant means that we can act on our inputs with **g**, OR act our outputs with **g** and we get the same answer (for every operation).
For a function with invariant input (e.g. invariant models) means **g** is the identity (no change).

Benefit of using equivariant models: substantially shrink the space of functions that need to be optimized over. Needs fewer data to constrain the function.

Group Theory and Irreducible Representations

Week 4

09/11/2022

Group Theory Recap

Group defines a set of operations and how these operations compose together
- collection of elements
- combine two element to produce another element

1. Closure:
If a, b are elements of a group G, and their composition a*b=c, c must also be an element of G
2. Associativity:
For any elements a, b, c… in the group G, (ab)c=a(bc)
3. Identity:
There exist a unique identity e, such that $e \cdot a = a$ and $a \cdot e = a$
4. Inverse:
•Every element a in G has a unique inverse a^{-1} , which is also in G, and satisfies $a \cdot a^{-1} = a^{-1} \cdot a = e$

Class is a subset of elements that are conjugate to each other.
Two elements a and b of a group G are said to be conjugate if there is a group element g, called the conjugating element, such that $a = gbg^{-1}$.
Theorem: all of the elements in a conjugate class have the same order/character of trace

Representation

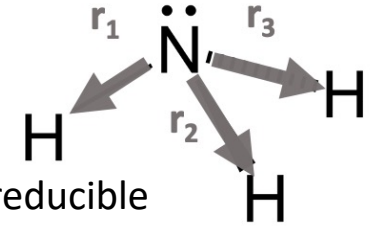
Let **G ={ e, a, b, c, …}** be a finite group of order **g** with **e** as identity element.
Let **R ={ R(e), R(a), R(b), R(c),** …} be the collection of non singular matrices all of the same order with the property

R(a)R(b)=R(c) if **a*b =c** in the group **G**

The collection of R of matrices is said to be representation of group.

Representation defines how the group acts on vector space
Any reducible representation can be broken down into some combination of irreducible representations.

Similarity transformation: $XRX^{-1} = S$



R - the transformation matrices of the basis (bond) vectors r_1, r_2, r_3 (matrices for reducible representation)

S - the block diagonal matrices (each "block" correspond to irreducible representation).

R and S are called conjugate matrices

Take C_3^1 as an example:

R Transformation matrices for the basis vectors r_1, r_2, r_3

E	C_3^1	C_3^2	$\sigma_v(1)$	$\sigma_v(2)$	$\sigma_v(3)$
$\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$	$\begin{bmatrix}0&0&1\\1&0&0\\0&1&0\end{bmatrix}$	$\begin{bmatrix}0&1&0\\0&0&1\\1&0&0\end{bmatrix}$	$\begin{bmatrix}1&0&0\\0&0&1\\0&1&0\end{bmatrix}$	$\begin{bmatrix}0&0&1\\0&1&0\\1&0&0\end{bmatrix}$	$\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix}$

S Transformation matrices for the basis vectors x, y, z

$$\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} \quad \begin{bmatrix}-\frac{1}{2}&-\frac{\sqrt{3}}{2}&0\\\frac{\sqrt{3}}{2}&-\frac{1}{2}&0\\0&0&1\end{bmatrix} \quad \begin{bmatrix}-\frac{1}{2}&\frac{\sqrt{3}}{2}&0\\-\frac{\sqrt{3}}{2}&-\frac{1}{2}&0\\0&0&1\end{bmatrix} \quad \begin{bmatrix}1&0&0\\0&-1&0\\0&0&1\end{bmatrix} \quad \begin{bmatrix}-\frac{1}{2}&-\frac{\sqrt{3}}{2}&0\\-\frac{\sqrt{3}}{2}&\frac{1}{2}&0\\0&0&1\end{bmatrix} \quad \begin{bmatrix}-\frac{1}{2}&\frac{\sqrt{3}}{2}&0\\\frac{\sqrt{3}}{2}&\frac{1}{2}&0\\0&0&1\end{bmatrix}$$

$XRX^{-1} = S$ in form of

$$\begin{bmatrix}-x&\frac{x}{2}&\frac{x}{2}\\0&-\sqrt{3}\frac{x}{2}&\sqrt{3}\frac{x}{2}\\-y&-y&-y\end{bmatrix}\begin{bmatrix}0&0&1\\1&0&0\\0&1&0\end{bmatrix}\begin{bmatrix}-\frac{2}{3x}&0&-\frac{1}{3y}\\\frac{1}{3x}&-\frac{1}{\sqrt{3}x}&-\frac{1}{3y}\\\frac{1}{3x}&\frac{1}{\sqrt{3}x}&-\frac{1}{3y}\end{bmatrix}=\begin{bmatrix}-\frac{1}{2}&-\frac{\sqrt{3}}{2}&0\\\frac{\sqrt{3}}{2}&-\frac{1}{2}&0\\0&0&1\end{bmatrix}$$

Similarity transformation: $XRX^{-1} = S$

Objective: Perform similarity transformation to find which symmetry operations belong to the same class

Similar matrices (conjugate matrices) have the same character (trace)

If matrices S and R are conjugate, there exists some other matrix X such that: $tr(S)=tr(XRX^{-1})$

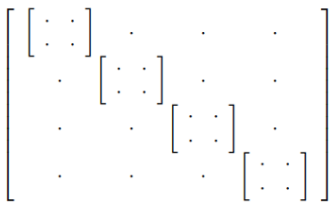
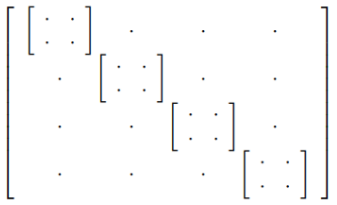
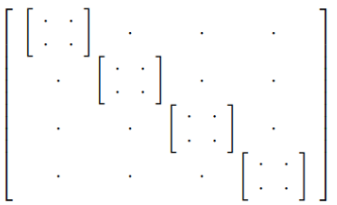
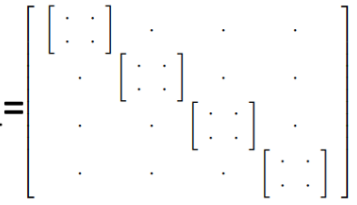
$$\chi(R) = \sum_j a_j \chi_j(R)$$

Given a set of matrices which constitute a representation

$$R_1=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \quad R_2=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \quad R_3=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \quad R_4=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$$X = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$$XRX^{-1} = S$$

$S_1=$  $S_2=$  $S_3=$  $S_4=$ 

* Block diagonal matrix: diagonal elements are square matrices of any size and the off-diagonal elements are 0

Which irreps are present in the reducible representation and how many of them?

Given the characters of matrices for a reducible representation:

	1E	$2C_3(z)$	$3\sigma_v$
τ	12	0	2

C_{3v}	E	$2C_3(z)$	$3\sigma_v$	linear functions, rotations
A_1	+1	+1	+1	z
A_2	+1	+1	-1	R_z
E	+2	-1	0	(x, y) (R_x, R_y)

$$a_i = \left(\frac{1}{h}\right)\sum_{p=1}^{k} n_p\, \chi(R_p)\chi_i(R_p)$$

$$a(A_1) = \frac{1}{6}[(1\cdot 12\cdot 1) + (2\cdot 0\cdot 1) + (3\cdot 2\cdot 1)] = 3$$

$$a(A_2) = \frac{1}{6}[(1\cdot 12\cdot 1) + (2\cdot 0\cdot 1) + (3\cdot 2\cdot (-1))] = 1$$

$$a(E) = \frac{1}{6}[(1\cdot 12\cdot 2) + (2\cdot 0\cdot (-1)) + (3\cdot 2\cdot 0)] = 4$$

Calculating H2O molecule , use group theory to calculate what phonons,6, symmetry operation

$$\tau = 3A_1 + A_2 + 4E$$

a_i – the number of times the irreducible representation i is present in the reducible representation
h – number of elements in the group (order of the group)
k – the total number of classes in a group
n_p – the number of elements in the p^{th} class
χ – character of reducible representation
χ_i – character of irreducible representation i
R_p – any one of the symmetry operations in the p^{th} class
$\chi(R_p)$ – character of reducible representation corresponding to any one of the operations R in the p^{th} class
$\chi_i(R_p)$ – character of irreducible representation i corresponding to any one of the operations R in the p^{th} class

Use characters of matrices to convert into irreps instead of full matrices

$$\tau = 3A_1 + A_2 + 4E$$

12 by 12 matrix X

12 by 12 matrix for reducible representation

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} =$$

$$\begin{bmatrix} A_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & e_{11} & e_{12} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & e_{21} & e_{22} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & e_{11} & e_{12} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & e_{12} & e_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & e_{11} & e_{12} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & e_{12} & e_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & e_{11} & e_{12} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & e_{12} & e_{22} \end{bmatrix}$$

Each block is a irrep

Reduction Formula to convert into irreps

$$a_i = \left(\frac{1}{h}\right)\sum_{p=1}^{k} n_p\ \chi(R_p)\chi_i(R_p)$$

a_i – the number of times the irreducible representation i is present in the reducible representation
h – number of elements in the group (order of the group)
χ – character of reducible representation
χ_i – character of irreducible representation i
R_p – any one of the symmetry operations in the p^{th} class

The sum of the squares of the characters in any irrep equals h

$$\sum_R [\chi_i(R)]^2 = h$$

The vectors whose components are the characters of two different irreducible representations are orthogonal

$$\sum_R \chi_i(R)\chi_j(R) = 0 \quad if\ i \neq j$$

Combine the two above:

$$\sum_R \chi_i(R)\chi_j(R) = h\delta_{ij}$$

matrix representation -> change of basis -> block diagonal matrix
(independent subspace transformation)

$$\rho(g) = Q^{-1}\left(\rho_1(g) \oplus \rho_2(g)\right) Q = Q^{-1}\begin{pmatrix} \rho_1(g) & 0 \\ 0 & \rho_2(g) \end{pmatrix} Q$$

Irreps of SO(3): Wigner-D matrices

Wigner-D matrices of **type l** are the irreducible matrix representations of SO(3). Denoted as $\mathbf{D}^{l}(\mathbf{R})$ with dimension of $(2l+1) \times (2l+1)$

$\mathbf{D}^{0}(\mathbf{R})\ v = 1v = v$ *scalar*

$\mathbf{D}^{1}(\mathbf{R})\ v = \mathbf{R}\ v$ *3D vector (velocity, force, displacement) transforms directly via the rotation matrix* $\boldsymbol{R} \in SO(3)$

Wigner-D functions form a complete orthogonal basis for functions on SO(3)

Wigner-D and Spherical Harmonics

The central column $\mathbf{D}^{0}$ is invariant to rotations $\mathbf{R}_{\alpha}$

Relative Spatial(Structural) Encodings

Given backbone coordinates $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^3 : 1 \leq i \leq N\}$

Backbone geometry is defined as：

$$\boldsymbol{O}_i = [\boldsymbol{b}_i \;\; \boldsymbol{n}_i \;\; \boldsymbol{b}_i \times \boldsymbol{n}_i],$$

where $\boldsymbol{b}_i$ is the negative bisector of angle between the rays $(\boldsymbol{x}_{i-1} - \boldsymbol{x}_i)$ and $(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)$, and $\boldsymbol{n}_i$ is a unit vector normal to that plane. Formally, we have

$$\boldsymbol{u}_i = \frac{\boldsymbol{x}_i - \boldsymbol{x}_{i-1}}{||\boldsymbol{x}_i - \boldsymbol{x}_{i-1}||}, \quad \boldsymbol{b}_i = \frac{\boldsymbol{u}_i - \boldsymbol{u}_{i+1}}{||\boldsymbol{u}_i - \boldsymbol{u}_{i+1}||}, \quad \boldsymbol{n}_i = \frac{\boldsymbol{u}_i \times \boldsymbol{u}_{i+1}}{||\boldsymbol{u}_i \times \boldsymbol{u}_{i+1}||}.$$

Relative Structural Encodings

$$\boldsymbol{e}_{ij}^{(s)} = \left(\mathbf{r}\left(||\boldsymbol{x}_j - \boldsymbol{x}_i|| \right), \quad \boldsymbol{O}_i^T \frac{\boldsymbol{x}_j - \boldsymbol{x}_i}{||\boldsymbol{x}_j - \boldsymbol{x}_i||}, \quad \mathbf{q}\left(\boldsymbol{O}_i^T \boldsymbol{O}_j \right) \right)$$

$$\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^3 : 1 \leq i \leq N\}$$

$$\boldsymbol{O}_i = [\boldsymbol{b}_i \;\; \boldsymbol{n}_i \;\; \boldsymbol{b}_i \times \boldsymbol{n}_i]$$

Distance: distance between the N, Cα, C, O and a virtual Cβ atom, r(·) lifted into a radial basis with r[2, 22]

Direction: Cα − Cα − Cα frame orientation

Rotation: Cα − Cα − Cα frame rotation, q(·) of the quaternion representation of the spatial rotation matrix $\boldsymbol{O}_i^T \boldsymbol{O}_j$ Quaternions represent 3D rotations as four-element vectors

Relative Positional Encodings

Transformer positional encoding

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}})$$

Class

Class ProteinFeatures:
    """
    edge features
    node features
    """

Class PositionalEncodings:
    """
    Transformer features
    """

Exploring Deep Neural Network in Material Science

Protein folding

The physical process by which a protein chain is translated to its native three-dimensional structure, typically a "folded" conformation by which the protein becomes biologically functional. Via an expeditious and reproducible process, a polypeptide folds into its characteristic three-dimensional structure from a random coil.

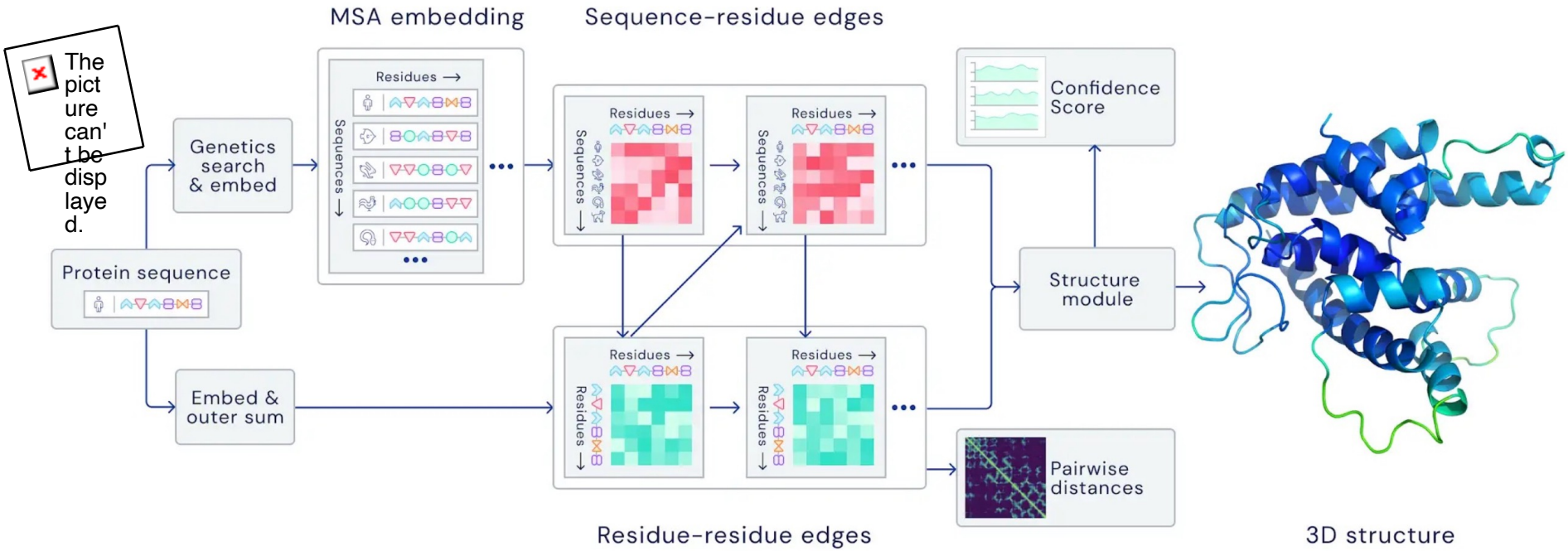
Protein Data Bank (PDB)

Traditional determination of protein 3D structure

X-ray Diffraction (XRD)
Nuclear Magnetic Resonance (NMR)
Cryogenic (frozen tissue) Electron Mycroscopy (CryoEM)

Challenges: laborious, difficulty with large protein molecules/compounds

Deepmind AlphaFold 2



Multiple Sequence Alignment (MSA) is the alignment of three or more biological sequences of similar length to infer homology can be inferred and to study the evolutionary relationship between the sequences. Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

Utilise Co-evolutionary Data by Multiple Sequence Alignment

Input: protein sequence

Output: protein 3D structure

https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Computational Protein Design: Inverse Protein Folding

Antibodies and binders

Designed binders to bind with specific viruses or receptors.

Shin, JE., Riesselman, A.J., Kollasch, A.W. et al. “Protein design and variant prediction using autoregressive generative models.” Nat Commun 12, 2403 (2021).

Cancer cell therapy

Cells with engineered receptors target and kill cancer cells.

Sockolosky, Jonathan T., et al. "Selective targeting of engineered T cells using orthogonal IL-2 cytokine-receptor complexes." Science 359.6379 (2018): 1037-1042.

Gene editing

Engineered enzymes target and edit specific genetic sequences.

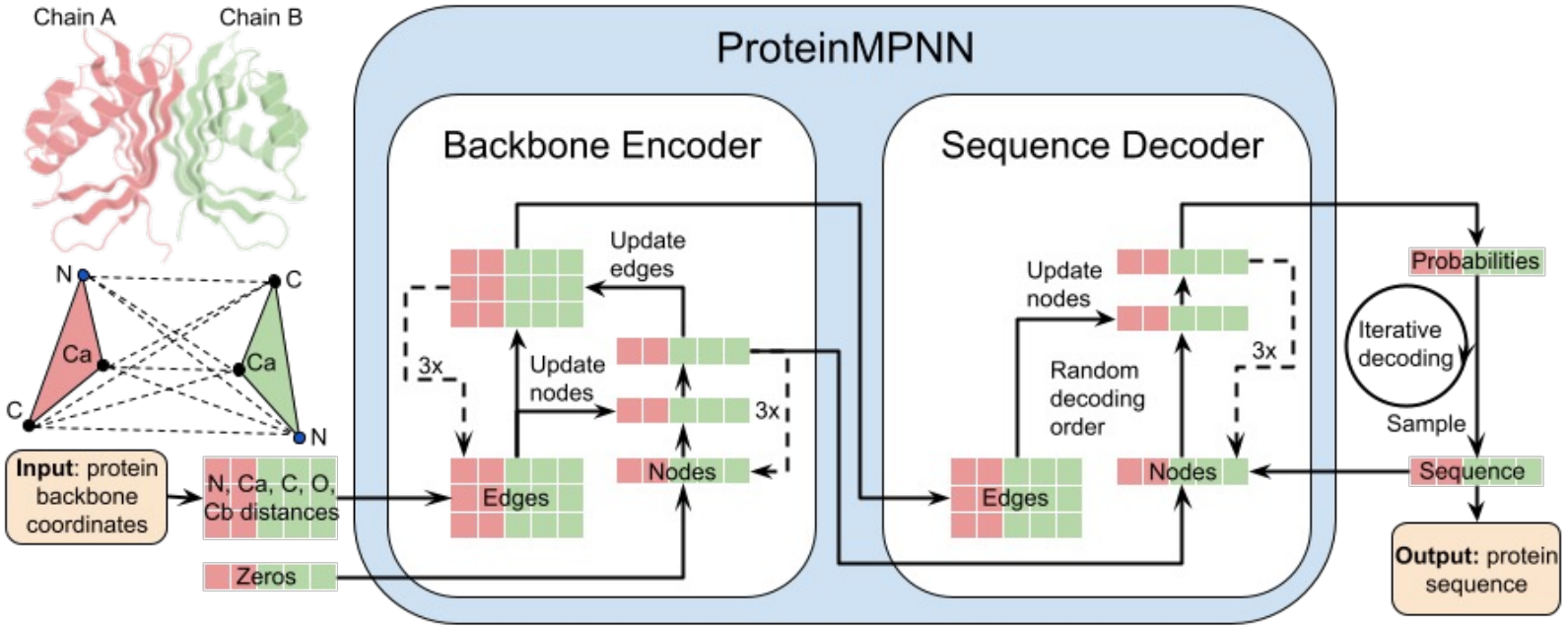
Thean, Dawn GL, et al. "Machine learning-coupled combinatorial mutagenesis enables resource-efficient engineering of CRISPR-Cas9 genome editor activities." Nature Communications 13.1 (2022): 1-14.

Plastic degradation

PETase enzymes eat plastic by catalyzing chemical reactions.

Lu, Hongyuan, et al. "Machine learning-aided engineering of hydrolases for PET depolymerization." Nature 604.7907 (2022): 662-667.

Inverse Protein Folding: ProteinMPNN



ProteinMPNN is a message passing neural network that aims to find an amino acid sequence that will fold into a given structure. The full network is composed of an encoder and a decoder with 3 layers each. The network takes as inputs the 3D coordinates and computes the following information for each residue:

(i) the distance between the N, Cα, C, O and a virtual Cβ atom,
(ii) the Cα − Cα − Cα frame orientation and rotation
(iii) the backbone dihedral angles,
(iv) the distances to the 48 closest residues.

Relative Structural Encodings

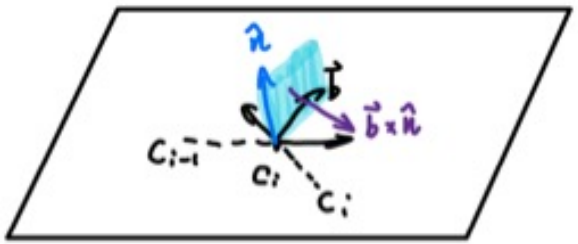
Given backbone coordinates $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^3 : 1 \leq i \leq N\}$

Backbone geometry is defined as:

$$\boldsymbol{O}_i = [\boldsymbol{b}_i \;\; \boldsymbol{n}_i \;\; \boldsymbol{b}_i \times \boldsymbol{n}_i],$$

where $\boldsymbol{b}_i$ is the negative bisector of angle between the rays $(\boldsymbol{x}_{i-1} - \boldsymbol{x}_i)$ and $(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)$, and $\boldsymbol{n}_i$ is a unit vector normal to that plane. Formally, we have

$$\boldsymbol{u}_i = \frac{\boldsymbol{x}_i - \boldsymbol{x}_{i-1}}{||\boldsymbol{x}_i - \boldsymbol{x}_{i-1}||}, \quad \boldsymbol{b}_i = \frac{\boldsymbol{u}_i - \boldsymbol{u}_{i+1}}{||\boldsymbol{u}_i - \boldsymbol{u}_{i+1}||}, \quad \boldsymbol{n}_i = \frac{\boldsymbol{u}_i \times \boldsymbol{u}_{i+1}}{||\boldsymbol{u}_i \times \boldsymbol{u}_{i+1}||}.$$



$O_i \simeq [\vec{b} \;\; \hat{n} \;\; \vec{b}\times\hat{n}]$

$\vec{b}$ unit plane vector

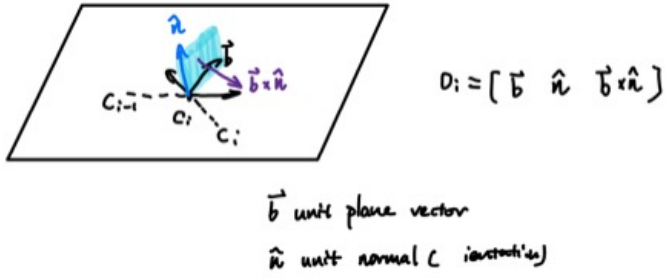
$\hat{n}$ unit normal (perpendicular)

Relative Structural Encodings

$$e_{ij}^{(s)} = \left(\mathbf{r}\left(||\boldsymbol{x}_j - \boldsymbol{x}_i|| \right), \quad \boldsymbol{O}_i^T \frac{\boldsymbol{x}_j - \boldsymbol{x}_i}{||\boldsymbol{x}_j - \boldsymbol{x}_i||}, \quad \mathbf{q}\left(\boldsymbol{O}_i^T \boldsymbol{O}_j \right) \right)$$

$$\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^3 : 1 \leq i \leq N\}$$

$$\boldsymbol{O}_i = [\boldsymbol{b}_i \;\; \boldsymbol{n}_i \;\; \boldsymbol{b}_i \times \boldsymbol{n}_i]$$



Distance: distance between the N, Cα, C, O and a virtual Cβ atom, r(·) lifted into a radial basis with r[2, 22]

Direction: Cα − Cα − Cα frame orientation

Rotation: Cα − Cα − Cα frame rotation, q(·) of the quaternion representation of the spatial rotation matrix $\boldsymbol{O}_i^T \boldsymbol{O}_j$ Quaternions represent 3D rotations as four-element vectors

Invariant Representation Spherical Harmonics