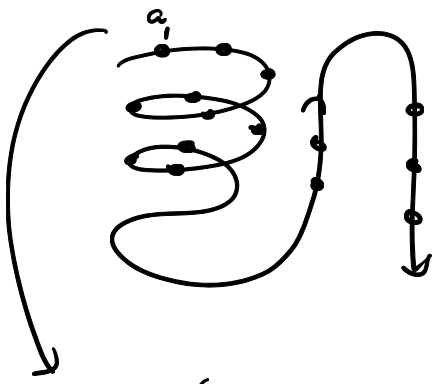
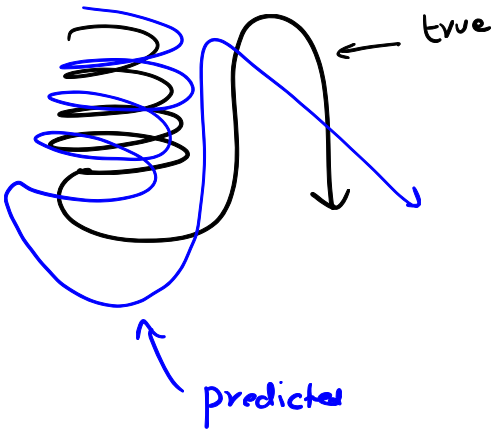
Protein structure Prediction



$a_1 \rightarrow (x_1, y_1, z_1)$ 3D coords vs $(\phi_1, \psi_1, \omega_1)$ torsion angles

↳ amino acid G ← primary structure (req)



$$\sum_{i=1}^{N} \left(\vec{T}_i - R\vec{P}_i \right)^2 = \text{obj: } \min_R$$

N: length of protein

$\vec{T}_i$: (x_i, y_i, z_i) true

$\vec{P}_i$: (x_i, y_i, z_i) predicted

R : 3x3 rotation matrix

↖ the best possible rotation matrix

$$\vec{\mu}_T = \frac{1}{N}\sum \vec{T}_i = \frac{1}{N}\sum \vec{P}_i = \vec{\mu}_P$$

means are super-imposed first

CRMSD : RMSE root mean squared error

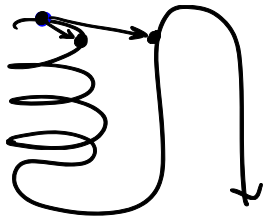
↑ Coord

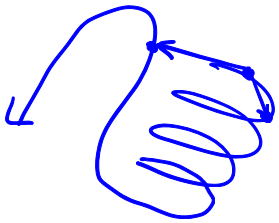
$$\sqrt{\frac{1}{N}\sum \left(\vec{T}_i - \vec{P}_i' \right)^2}$$

$P_i' = RP_i$

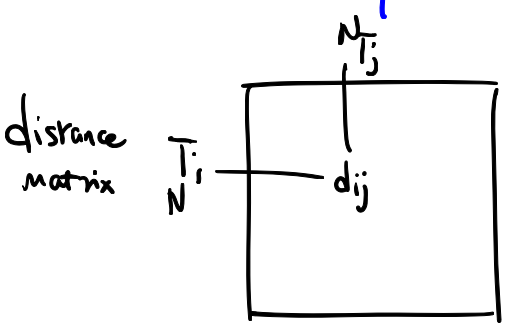
$\vec{T}_i, \vec{P}_i, \vec{P}_i' \in \mathbb{R}^3$

distance RMSD

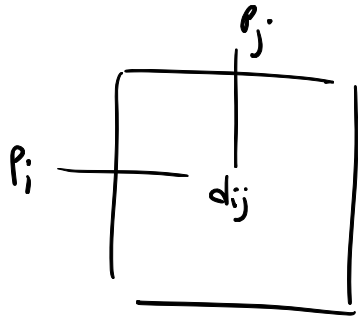




true T



predicted P



$$d_{ij} = \|\vec{T}_i - \vec{T}_j\| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

D_P : distance matrix for P

D_T : Symmetric $N \times N$ distance matrix for T

$$\|D_T - D_P\|_F = dRMSD = \sqrt{\frac{1}{N^2} \sum_i \sum_j \left(D_T(i,j) - D_P(i,j)\right)^2}$$

F ↖ Frobenius Norm

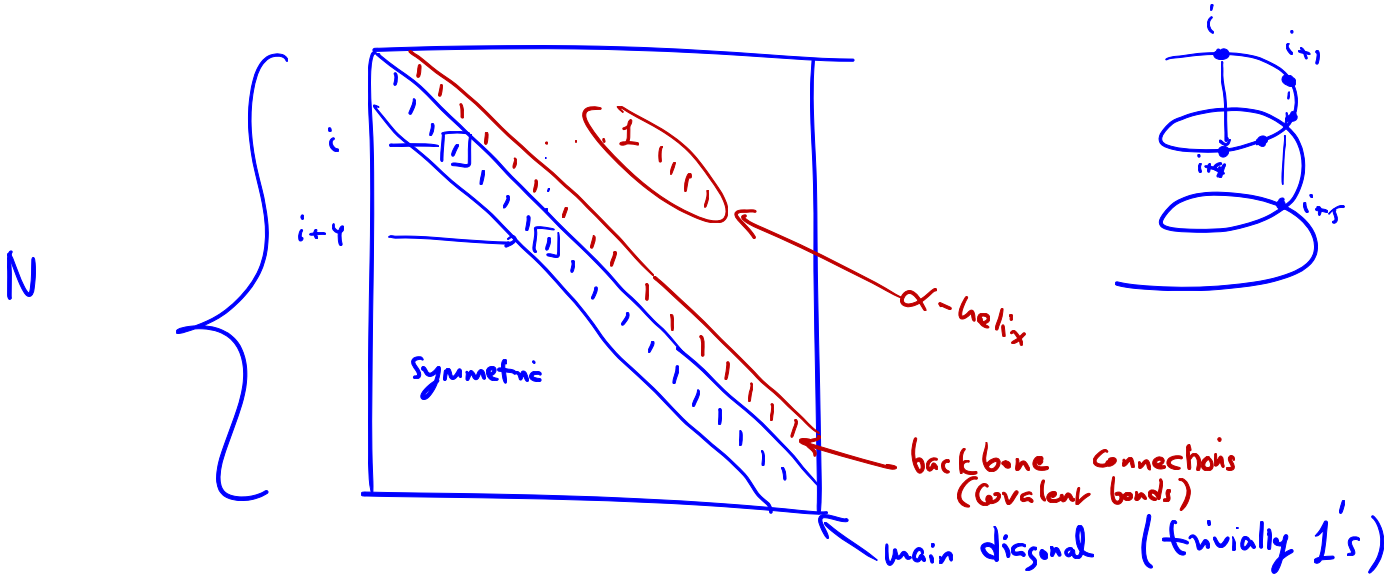
Simplified Distance Matrix : Contact Map

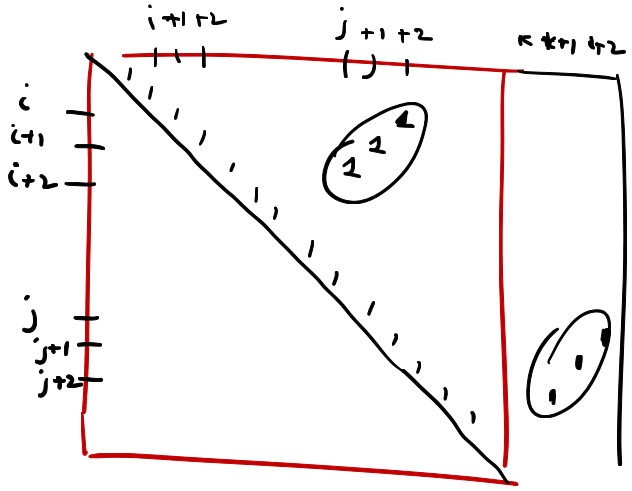
$\theta = 8\text{Å}$ $\quad$ $\text{Å} = 10^{-10}\ m = 0.1\ nm$

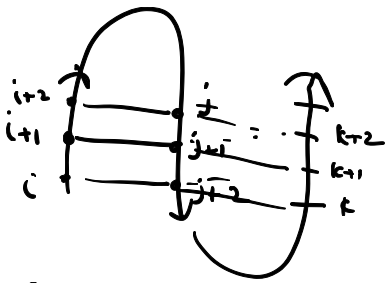
	r_1	r_2	r_3
r_1	0	6Å	10Å
r_2	6Å	0	6Å
r_3	10Å	6Å	0

$\Rightarrow_{\theta}$

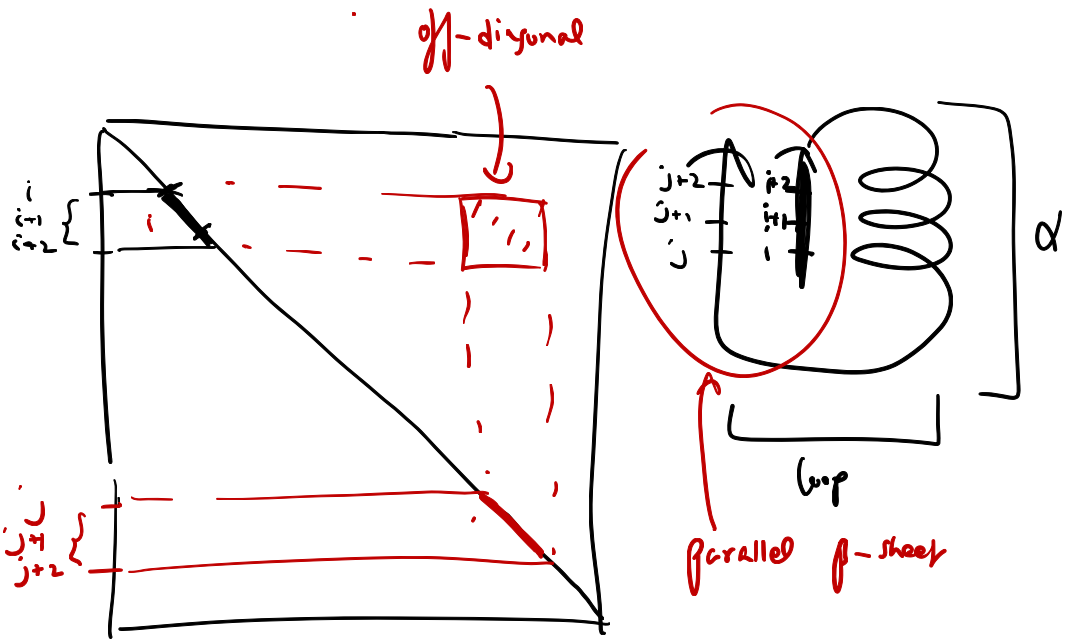
1	1	0
1	1	1
0	1	1

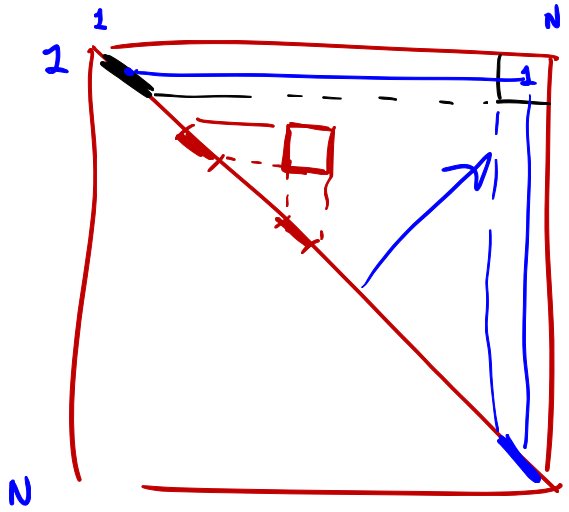






Anti-parallel
β-sheet





Very Sparse

O(N) Contacts

Distance map

$d_{ij} \in \mathbb{R}$

Contact map (0)

1/0

bins = 2

$\theta = 8\text{Å}$

discretized distance map

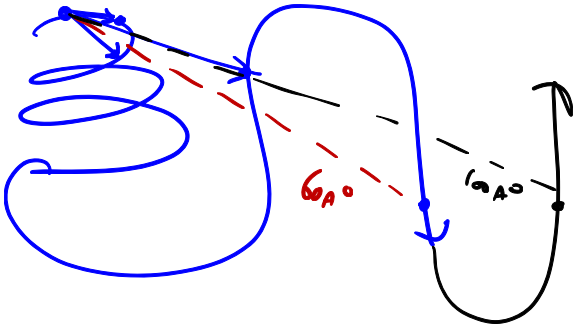
d_{ij} =

0 2Å 4Å 6Å ...

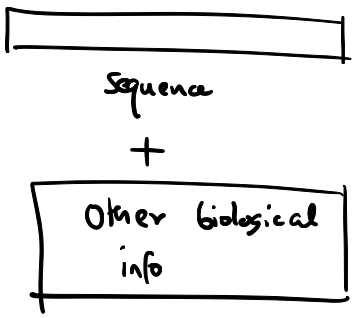
bin size = 2Å

0Å 2Å 22Å

64 bins



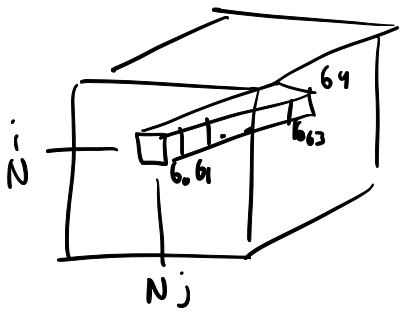
Input



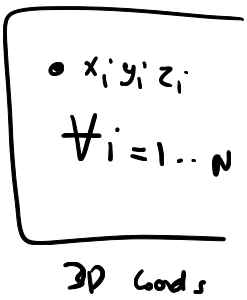
Sequence only based

Output

64 - bin distance map



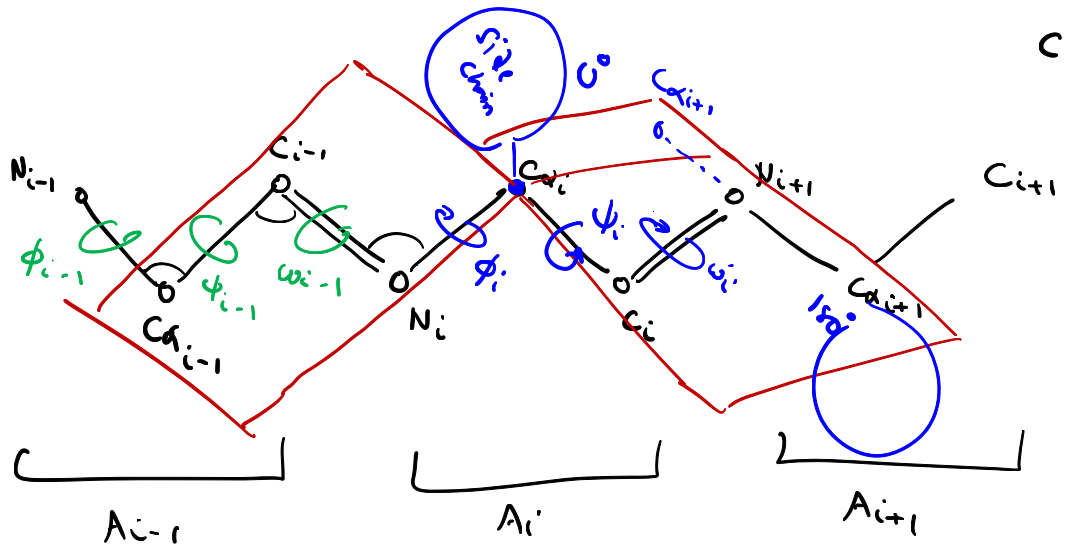
final output

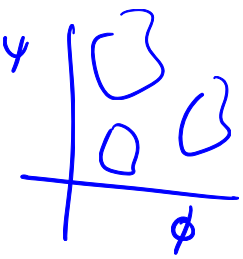


Torsion Angle (dihedral angle)

N-terminus

C-terminus





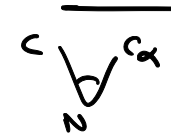
bond-length

$C - N : 1.33\text{Å}$

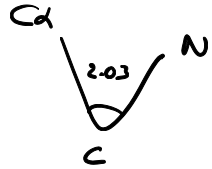
$N - C_\alpha : 1.46\text{Å}$

$C_\alpha - C : 1.52\text{Å}$

bond angle



2.12 rads



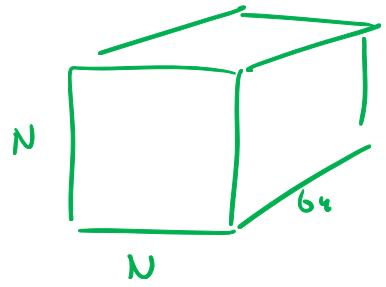
$\approx 110° - 115°$

$\omega_i \in \{0°, 180°\}$

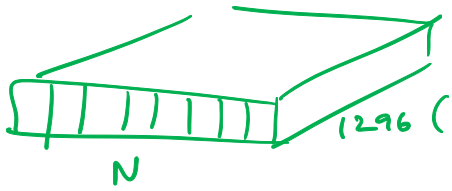
gs — trans

pre-dominant one

$\omega_i = 180°$ in practice with rare exceptions!



+



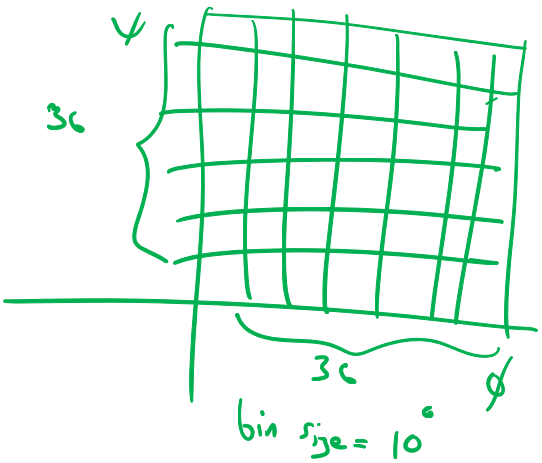
discrete distance map

torsion angle bins

$(CE_d + CE_a)$

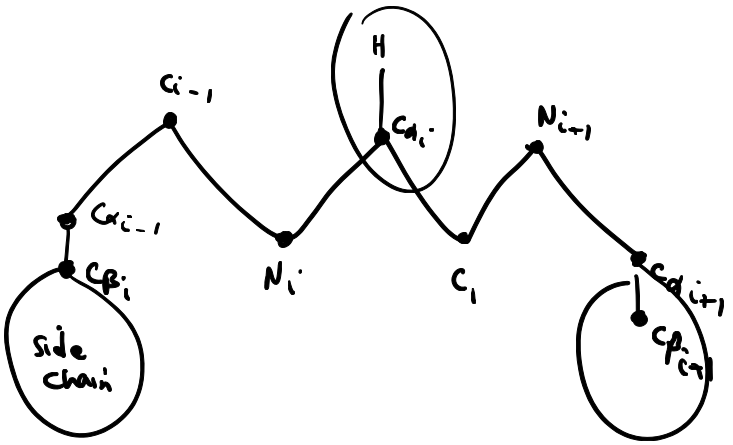
final loss

(ASA: accessible surface area, etc.)



1296 bins

Representation: C_β G



C_{β_i} $\forall i = 1..N$

except for G which uses C_α

$$d_{ij} = \| \vec{C}_{\beta_{i-1}} - \vec{C}_{\beta_i} \|$$

(x_{y2}) (x_{y2})

Alphafold 1 : Very deep CNN

|N| = length

MILGPAWS...

actual input seq

+

Seq database

feature vectors

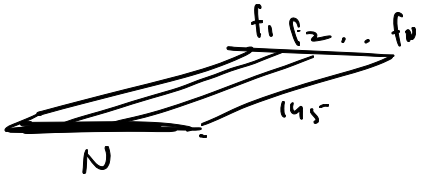
f_i : feature vector for pos i in seq (N)

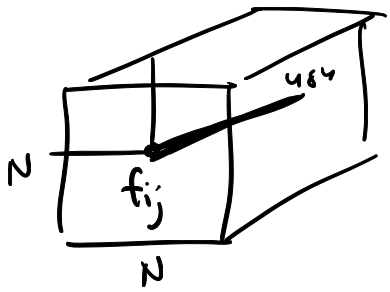
$\in \mathbb{R}^{145}$

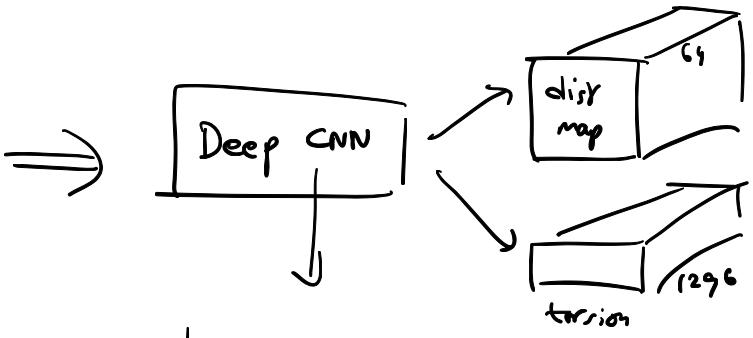
f_{ij} : pair-wise features (N^2)

for pos i & pos j

$\in \mathbb{R}^{484}$







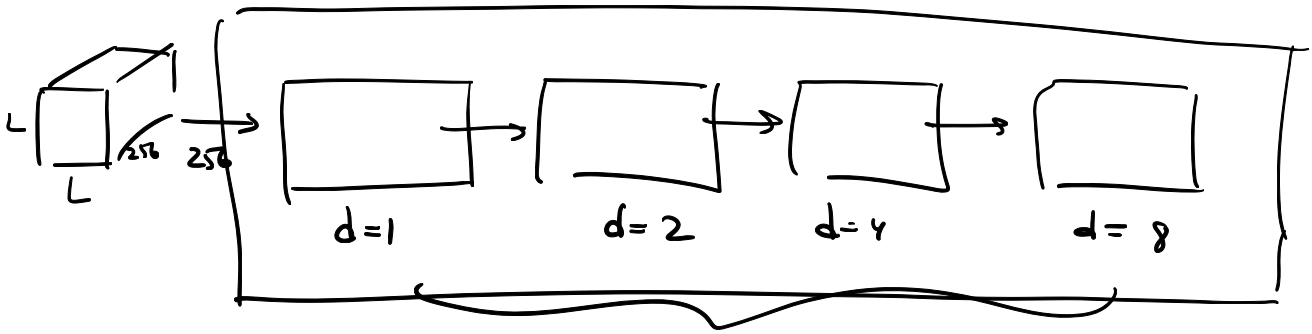
dialations

dialated convolutions

220 blocks of dialated CNNs

7x7 blocks of dim 256

48x4 blocks of 128 dim features



d ≡ dialation

L: 64 ⇒ 64x64 crop of the distance matrix !

220 x 6 layers = 1320 layers!