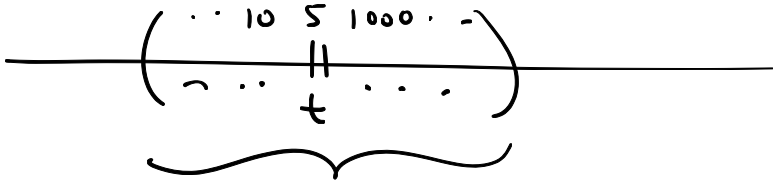
Dataset:

→ for all sequence
  for all offsets (0, 1, 2)
    S = seq of tokenized (int-) kmers

[5, 15, 0, 1000, . . .]

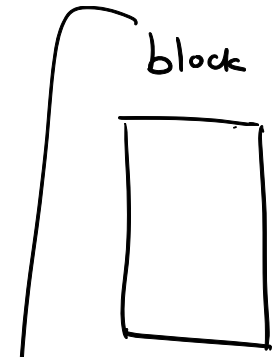


25

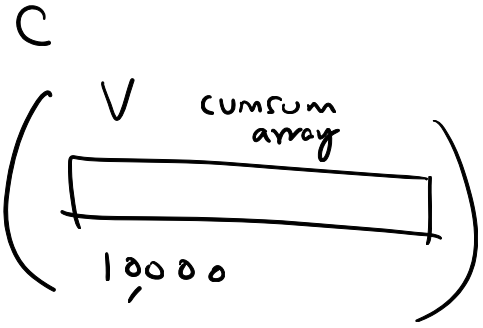
Dataloader

↳ num workers = 4

↳ batch = 1



C



np.searchsorted (C, 5x512x1024)

Worker_id = 0, 1, 2, 3 num workers = 4

array of Neg Sampler [0] = 5x512x1024

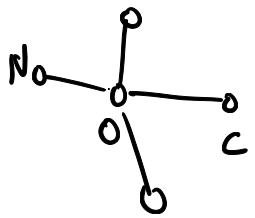
NegSampler [worker_id] = np.searchsorted (_)

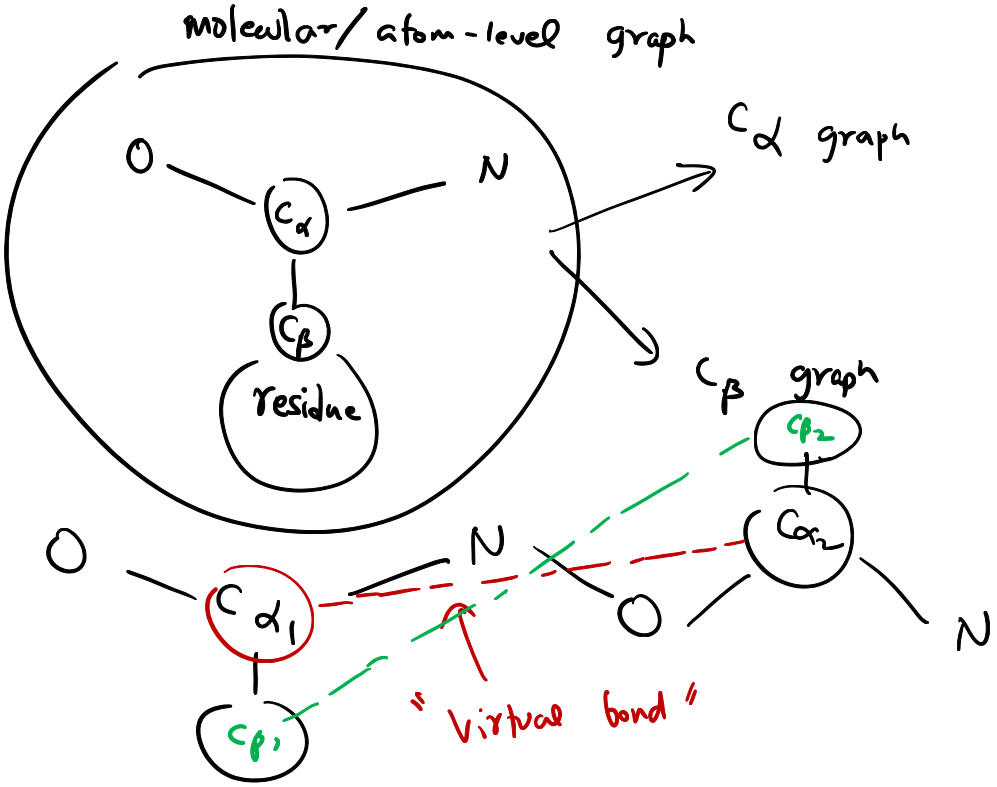
shuffle the seq idx

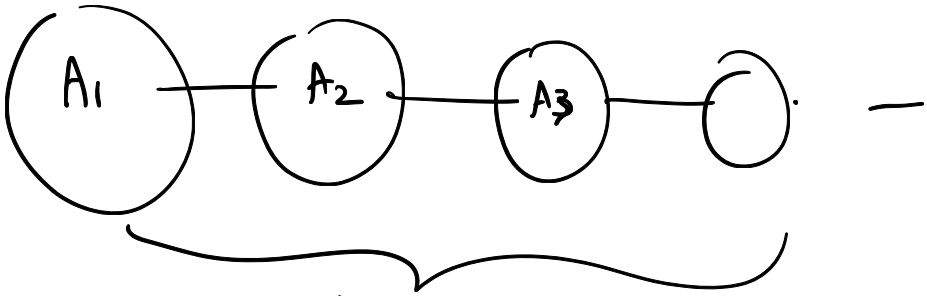
```
for (k, seq) in enumerate ( sequences )
    if k % num_workers == worker_id :
        for offset 0, 1, 2 :
            tokenize
            // generate blocks of 6 x 512 x 1024
                create 512 x 1024 pos pairs
                neg samples [worker_id] = search sorted
```

Graph neural networks

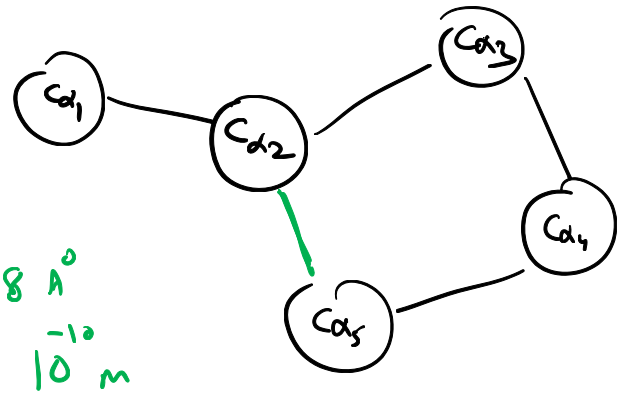
protein molecule → 3D coords of each atom







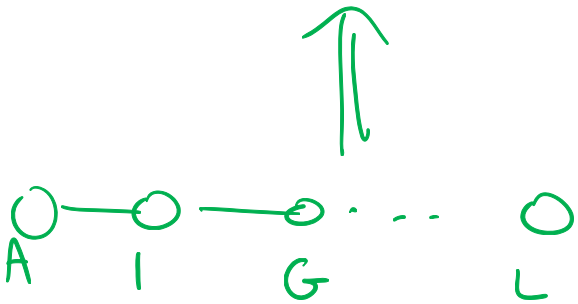
Sequence



6-8 A°

10^{-10} m

C_α - graph



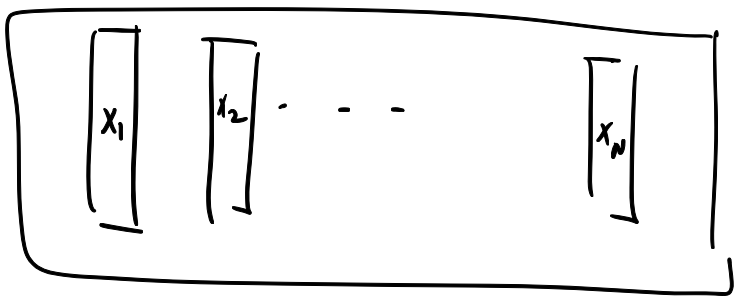
Structure prediction

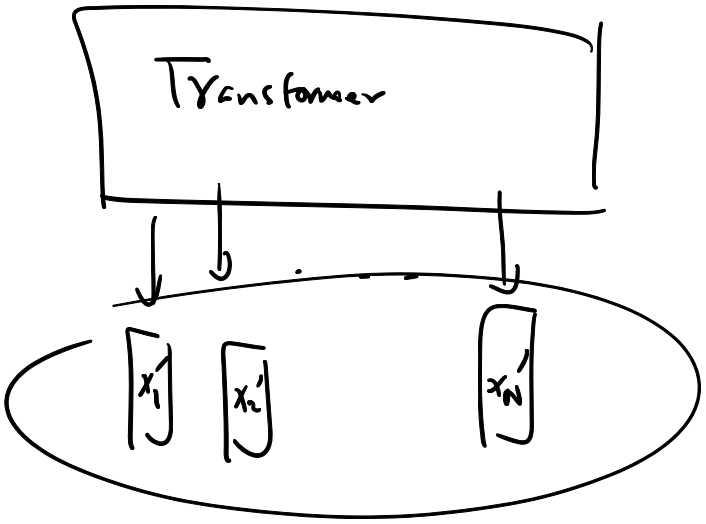
(Machine translation)

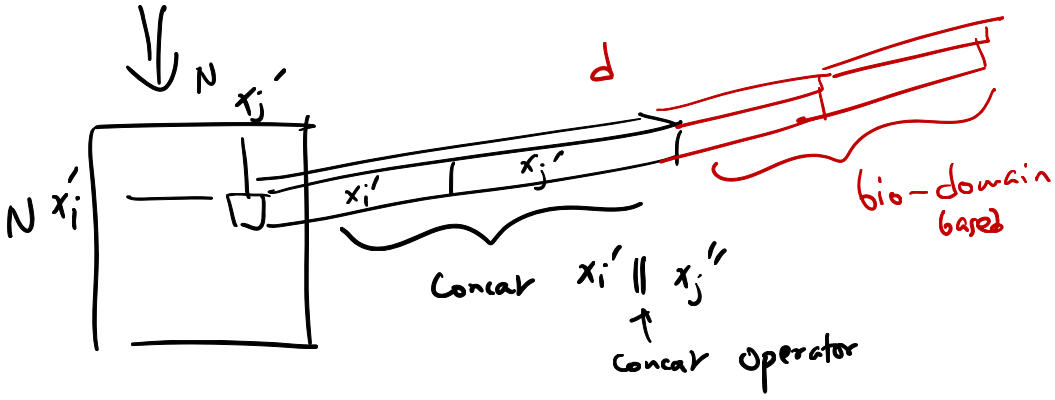
Alphafold:

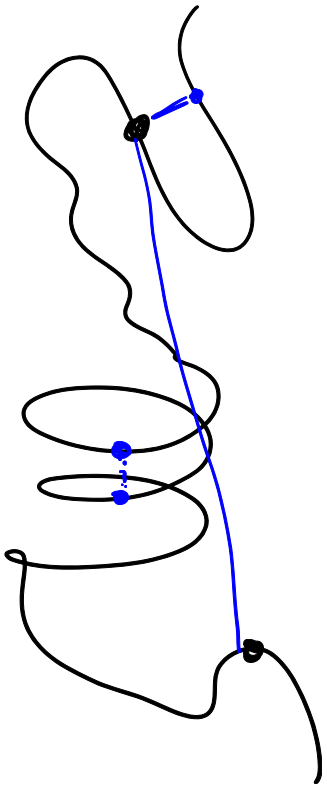
A I G L . . .

← sequence









a) 0/1 prediction — edge exists or not | binary logistic regression

b) $d(A_i, A_j)$ = real value ← regression problem

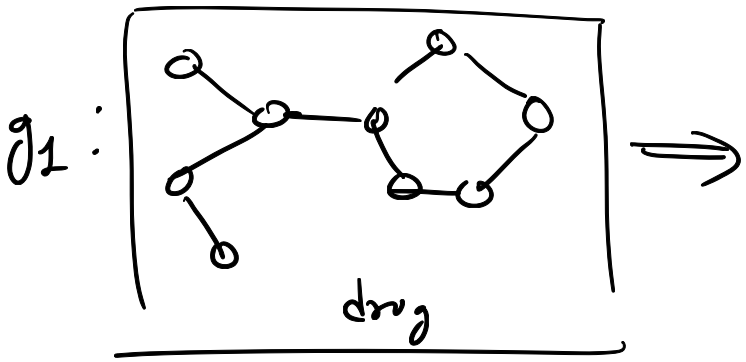
c) Create k-bins

$d(A_i A_j)$ →
- 0 – 2 A
- 2 – 4 A
- 4 – 6 A
- ⋮
- > 22 A°

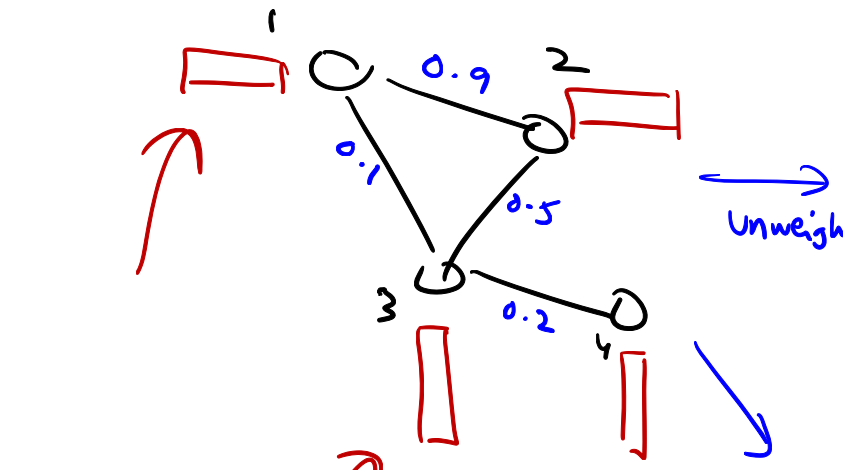
} k-class problem

Molecular graphs: small chemical compounds

Atom-level graphs ⟶ labels/value per graph



$\Longrightarrow \quad y_1 = 1$



Unweighted

A + I

	1	2	3	4
1	0	1	1	0
2	1	0	1	0
3	1	1	0	1
4	0	0	1	0

$a_{ij} = a_{ji}$

Identity

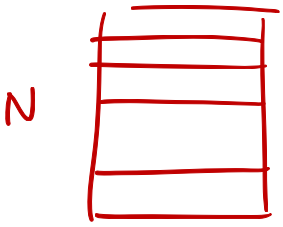
adjacency value between v_i and v_j

Weighted adjacency matrix

0	0.5	0.1	0
0.9	0	0.5	0
0.1	0.5	0	0.2
0	0	0.2	0

A

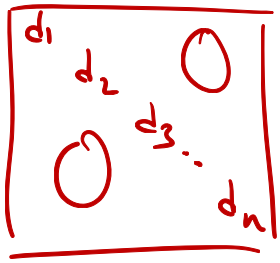
Input feature vectors



d-dim input features per node

D: degree matrix

$$d(v_i) = \sum_{j=1}^{n} a_{ij}$$



diagonal matrix

Unweighted:

$D =$

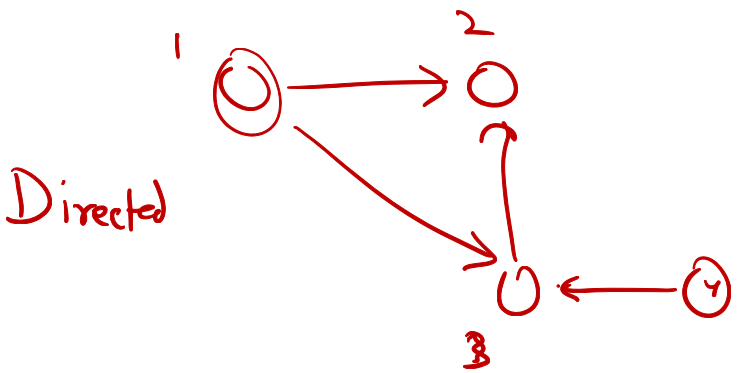
2	0	0	0
0	2	0	0
0	0	3	0
0	0	0	1

weighted

1.0			
	1.4		
		0.8	
			0.2

Undirected graphs: A is symmetric

$$a_{ij} = a_{ji}$$



Directed

A

	1	2	3	4	outdegree
1	0	1	1	0	2
2	0	0	0	0	0
3	0	1	0	0	1
4	0	0	1	0	1
in-degree	0	2	2	0	

$D^{in} =$

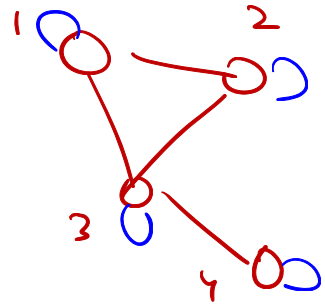
0			
	2		
		2	
			0

D^{out}

2			
	0		
		1	
			1

Transition Matrix

Undirected



$A+I$

1	1	1	0
1	1	1	0
1	1	1	1
0	0	1	1

D

3			
	3		
		4	
			2

prob transition matrix

$$M = m_{ij} = \frac{a_{ij}}{d_i} =$$

1/3	1/3	1/3	0
1/3	1/3	1/3	0
1/4	1/4	1/4	1/4
0	0	1/2	1/2

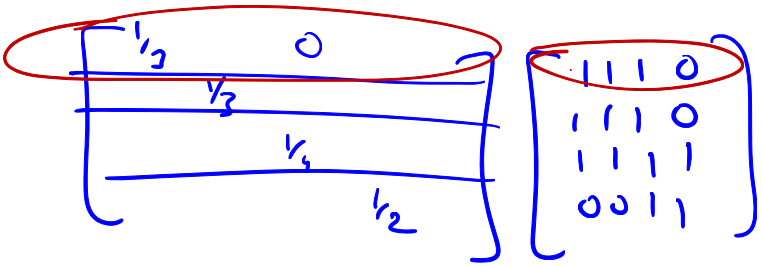
asymmetric

$$= D^{-1} A \quad \begin{bmatrix} 1&1&1&0 \\ 1&1&1&0 \\ 1&1&1&1 \\ 0&0&1&1 \end{bmatrix}$$

$$D = \begin{bmatrix} 3 & & & \\ & 3 & & \\ & & 4 & \\ & & & 2 \end{bmatrix} \qquad \tilde{D}^{-1} = \begin{bmatrix} 1/3 & & & \\ & 1/3 & & \\ & & 1/4 & \\ & & & 1/2 \end{bmatrix}$$

$$D^{-1} D = I$$

$$\tilde{D}^{-1} D = I$$



$\tilde{D}^{-1}$ A

graph Laplacian

$$L = D - A$$

$$\begin{bmatrix} 3 & & & \\ & 3 & & \\ & & 4 & \\ & & & 2 \end{bmatrix} - \begin{bmatrix} 1&1&1&0 \\ 1&1&1&0 \\ 1&1&1&1 \\ 0&0&1&1 \end{bmatrix} = \begin{bmatrix} 2&-1&-1&0 \\ -1&2&-1&0 \\ -1&-1&3&-1 \\ 0&0&-1&1 \end{bmatrix}$$

Normalized Laplacian

$$L_a = D^{-1} L = D^{-1}(D - A)$$

$$= D^{-1} D - D^{-1} A$$

$$= I - M$$