Alphafold 1

q: Protein sequence, length = L

MLIGAWS...

Homologous sequence

related by evolution

PSSM: position specific scoring matrix

MSA: Multiple sequence alignment

1) query the sq DB
2) extract similar sequence

BLAST

3) MSA
4) PSSM

CASP 15

Critical Assessment of Structure Prediction

(2022!)

FM

Free Modeling

TBM

Template-based Modeling

PDB

Protein data Bank

~ 200k structure

Uniprot sequence database

10^6+ sequence

q → MLIGAWSPIM

→ MLGAPISTW

→ MGAWP

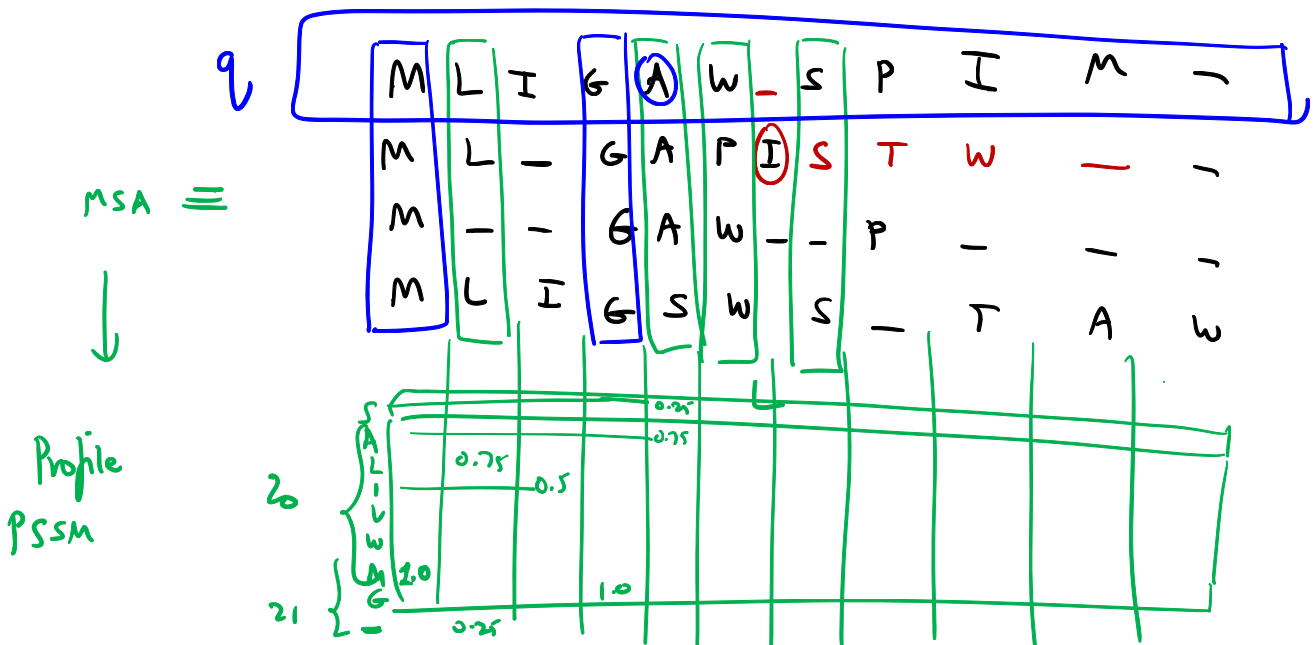
→ MLIGSWSTAW

MSA

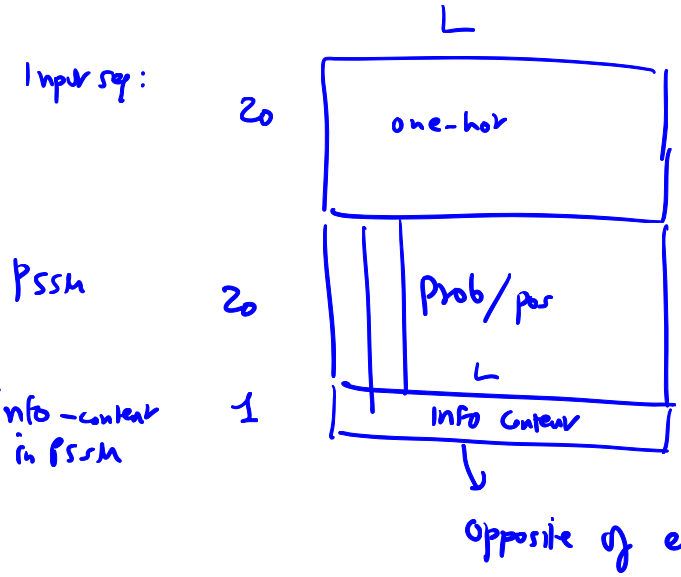
↓

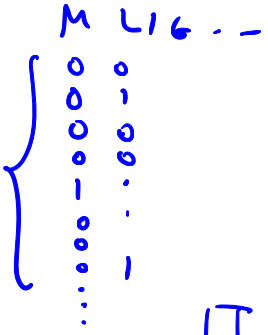
weighted version of LCS

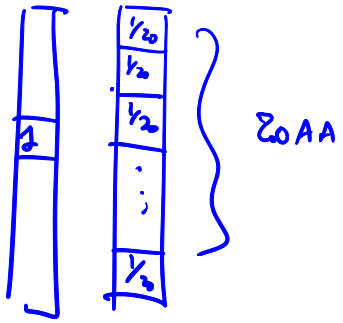
(longer common subsequence)



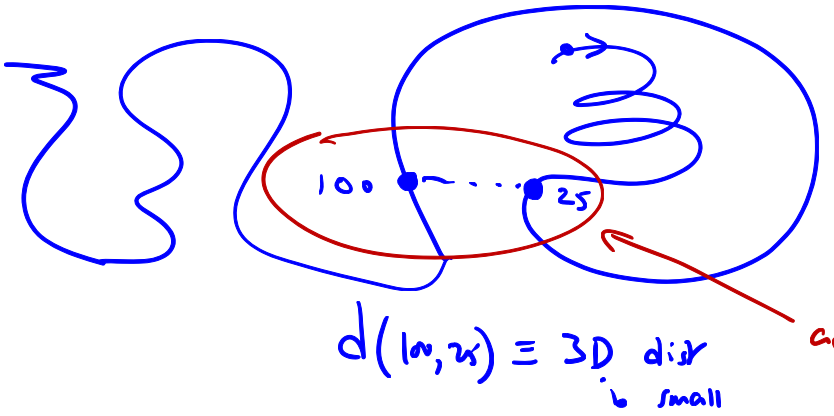
Ungapped PSSM : disallow gaps

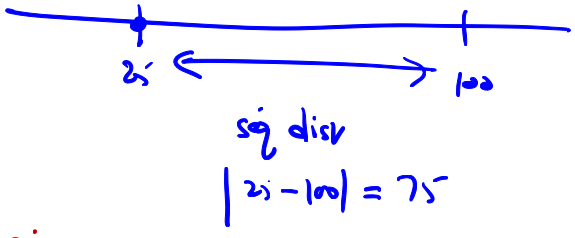






Derive pair-wise feature from seq





Co-evolution info (Potts model)

a_i vs a_j $\rightarrow$ f_{ij}

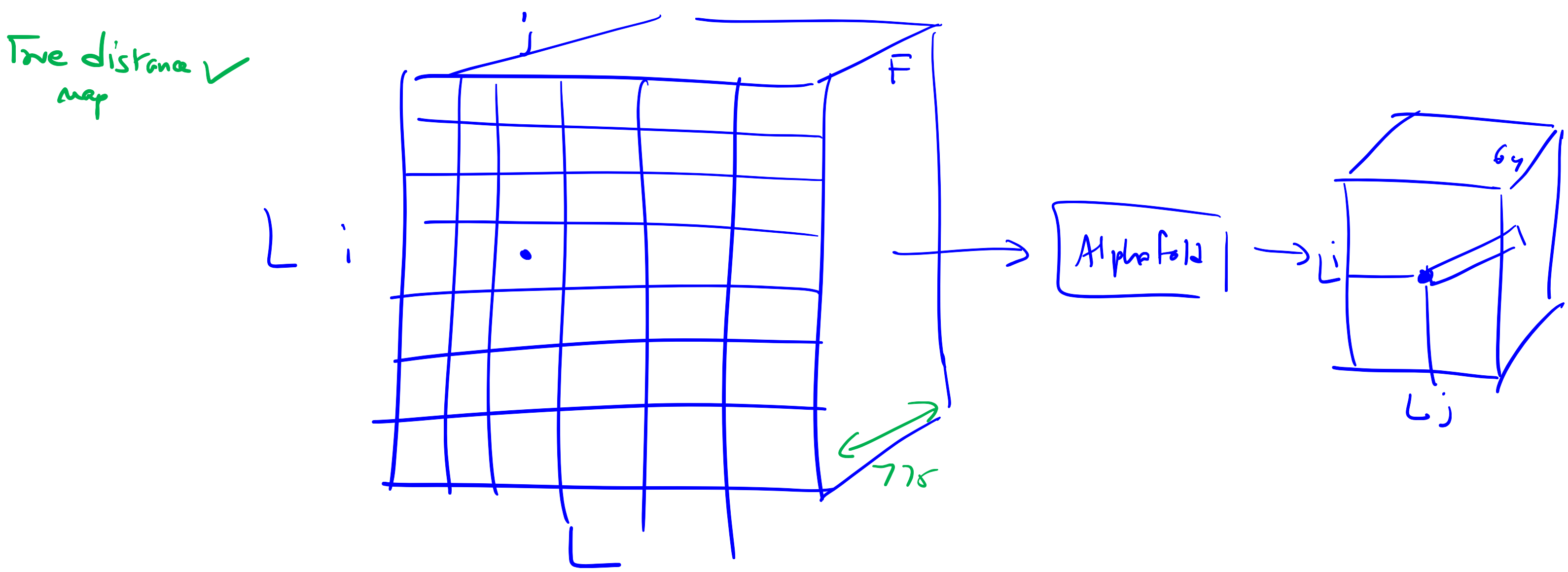
22 [ ] x [ ] 22 = 484 + 1 = 485 dim feature vector

1 ↑ info-content

$f_{ij} \in \mathbb{R}^{485}$

$f_i \in \mathbb{R}^{145} \quad \forall i = 1 \ldots L$

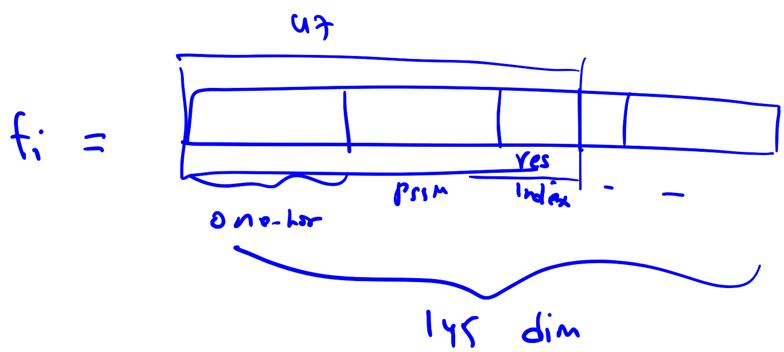
$f_{ij} \in \mathbb{R}^{485} \quad \forall i,j = 1 \ldots L$



$F_{ij} \equiv$ feature vector for (i,j) cell.

$= \vec{f}_{ij} \,||\, \vec{f}_i \,||\, \vec{f}_j \,||\, |\vec{f}_i - \vec{f}_j| \,||\, \vec{f}_i \odot \vec{f}_j \,||$

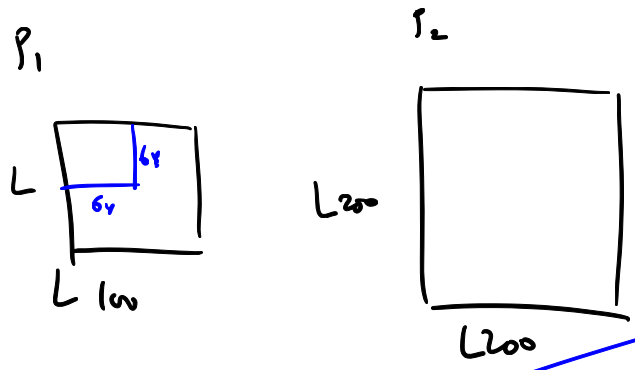
Co-evolution vector: 485; $\vec{f}_i$: 145; $\vec{f}_j$: 145; $|\vec{f}_i - \vec{f}_j|$: 145; $\vec{f}_i \odot \vec{f}_j$: 145 (extra!)

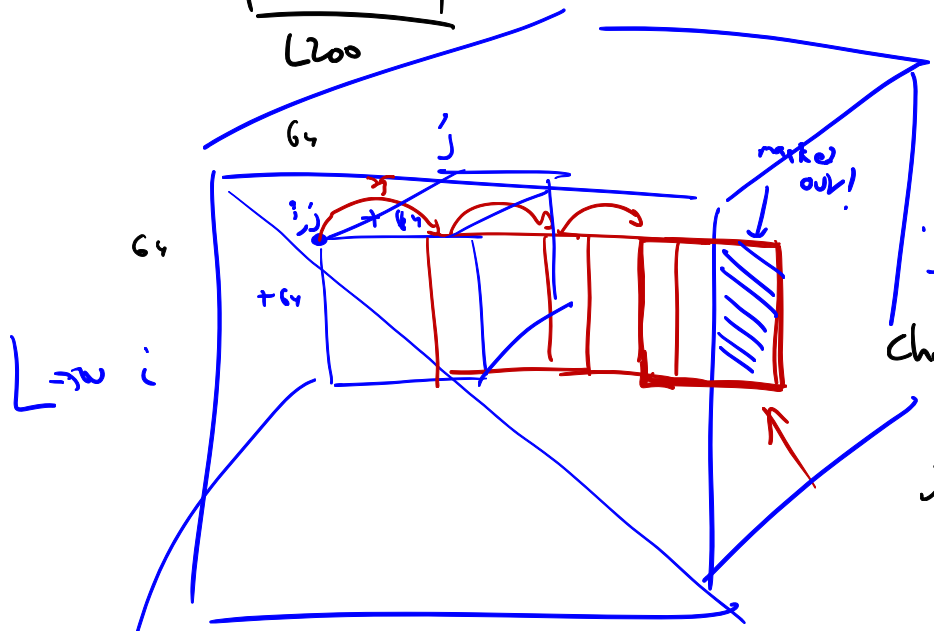


$$
\begin{array}{r}
485 \\
145 \\
+\ 145 \\
\hline
775
\end{array}
$$

dim features!

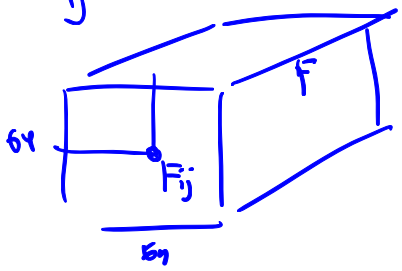
Alphafold

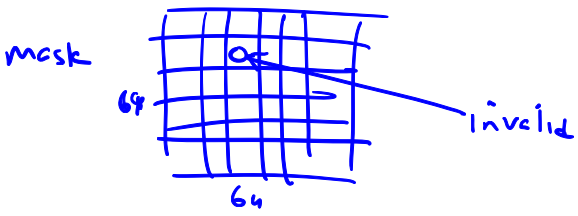




L = 500

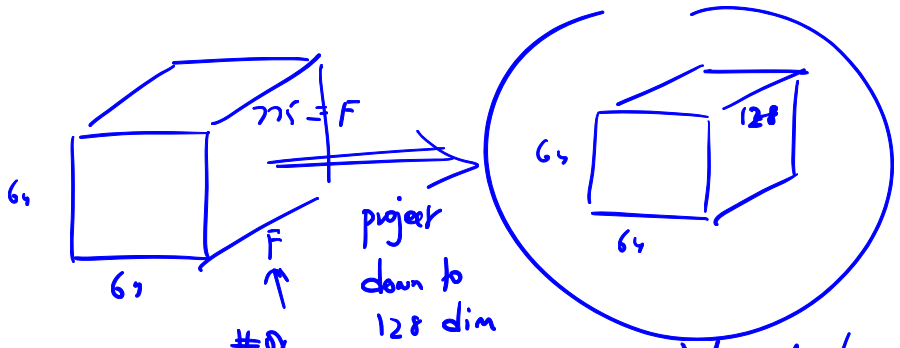
Use padding of 32 all around





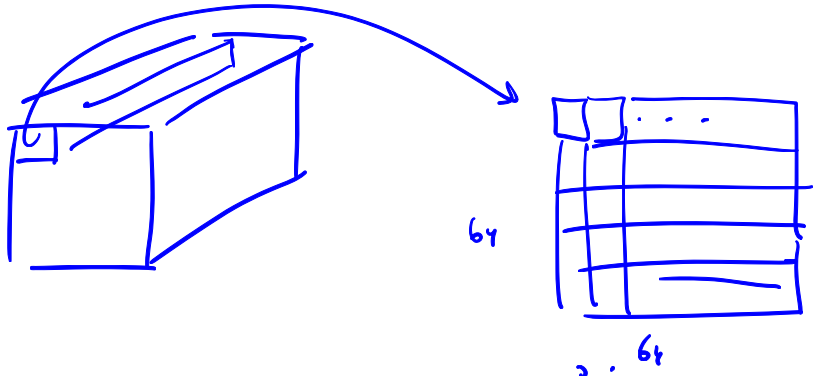
For same protein

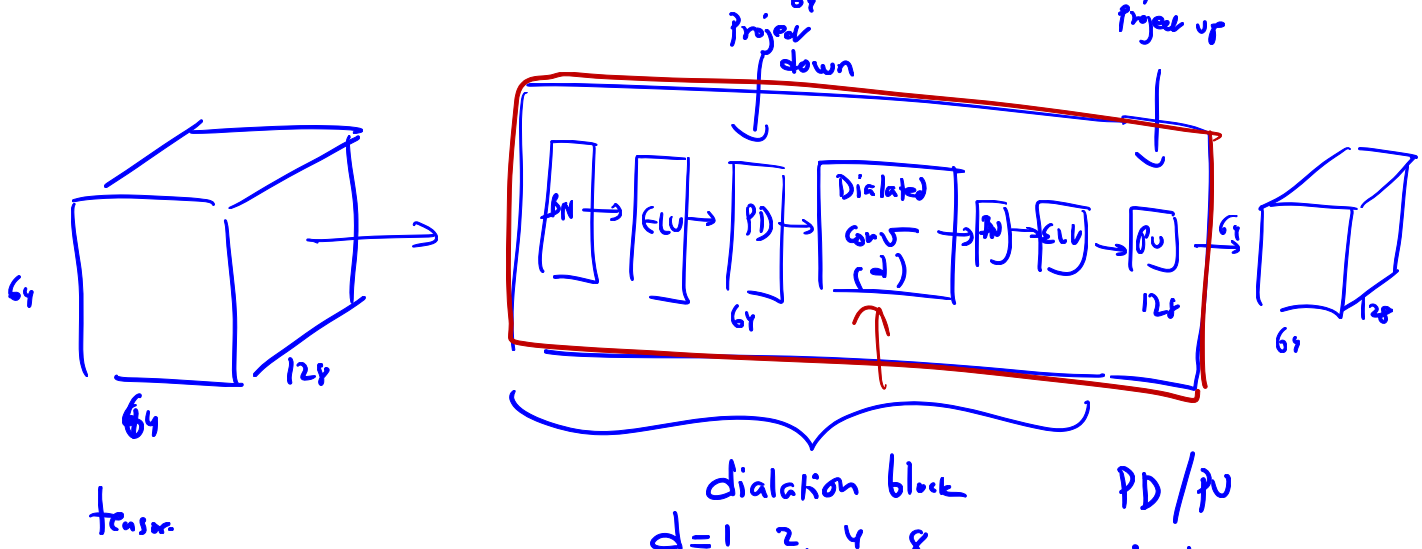
↳ multiple crops with i,j starting + stride



use 1x1 (xF) | x 128 filters (out_channels)

2d conv



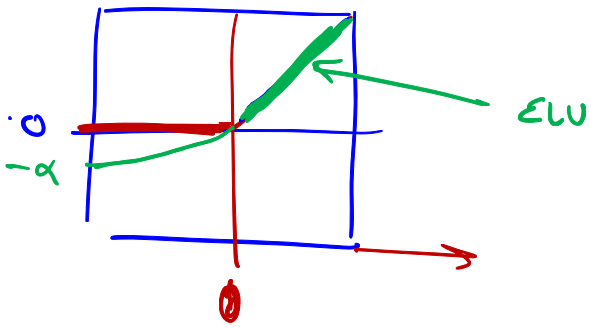


tensor

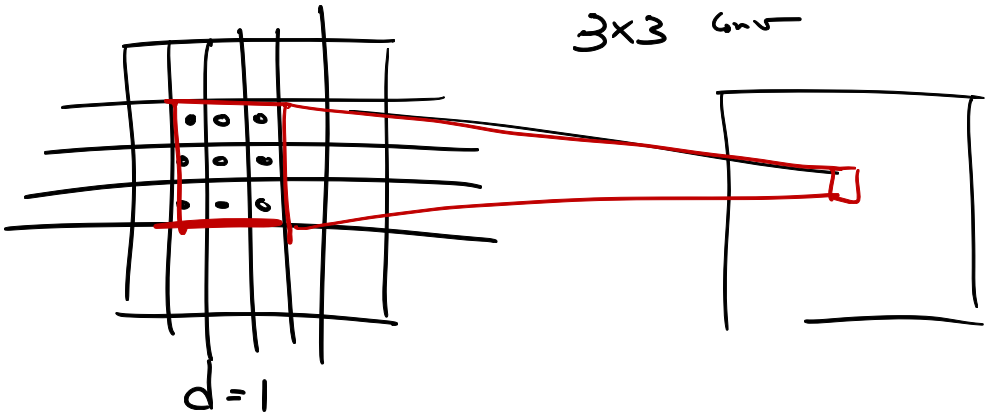
dialation block

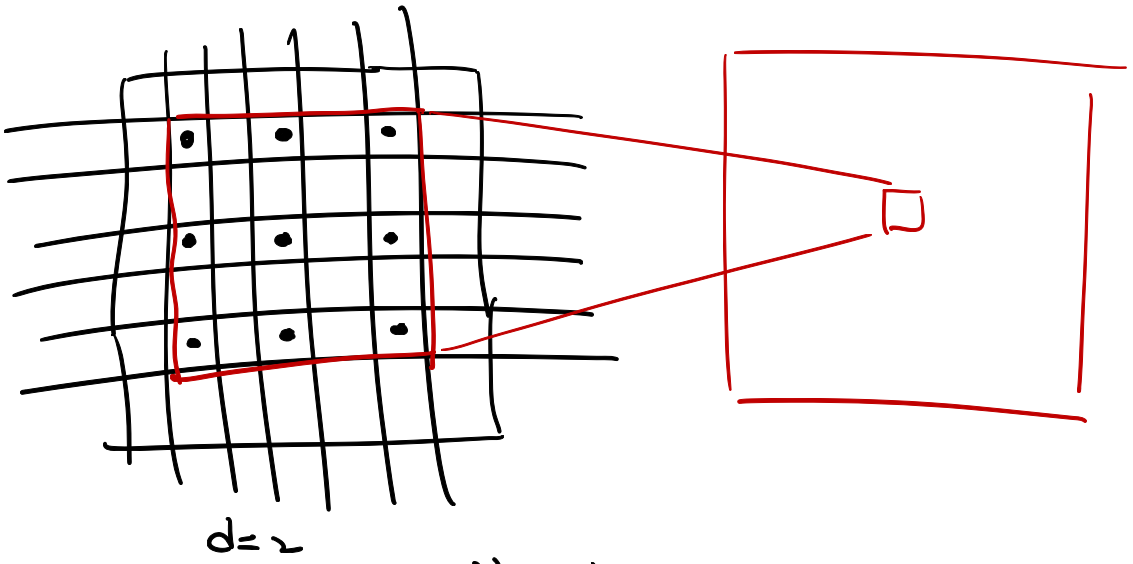
d = 1, 2, 4, 8

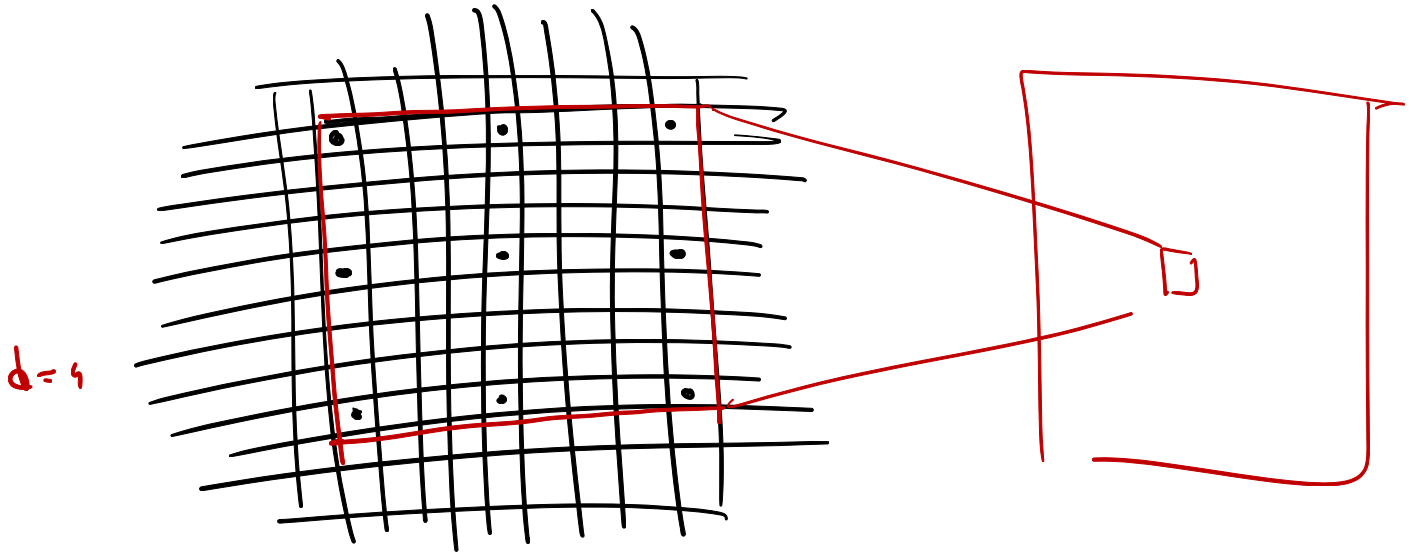
PD / PU : 1x1 conv (out_channels / in_channels)

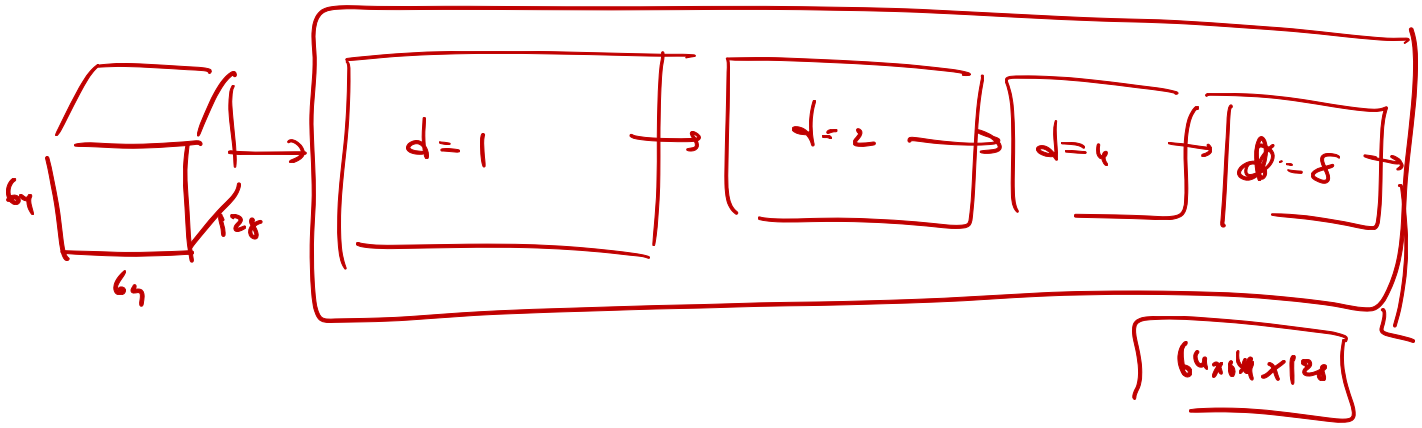


Dialated conv (d ≡ dialation)

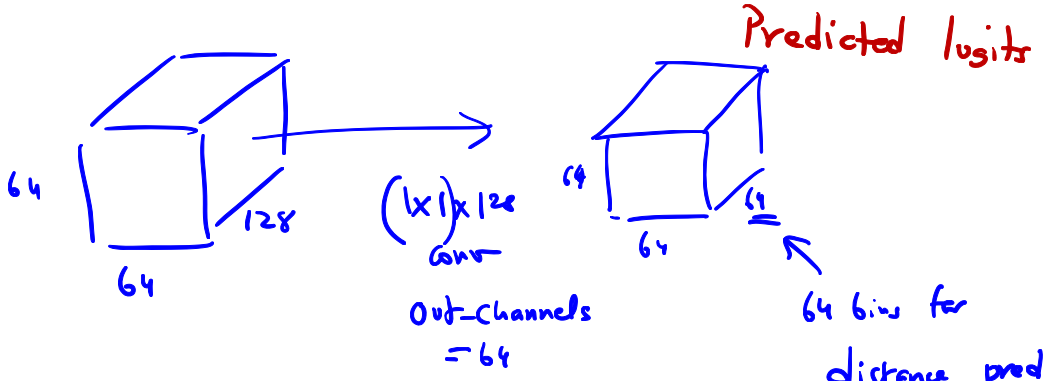








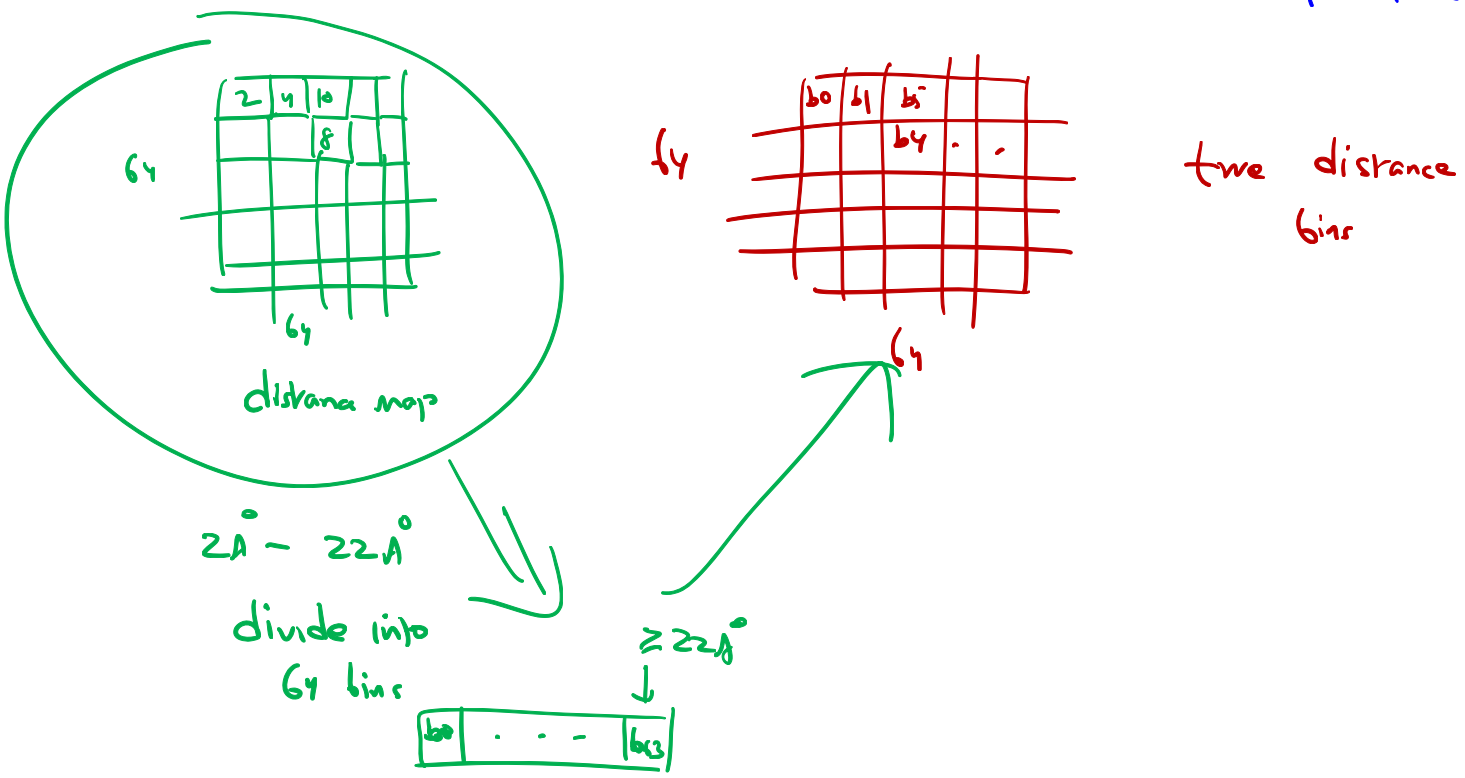
22 dilation blocks!



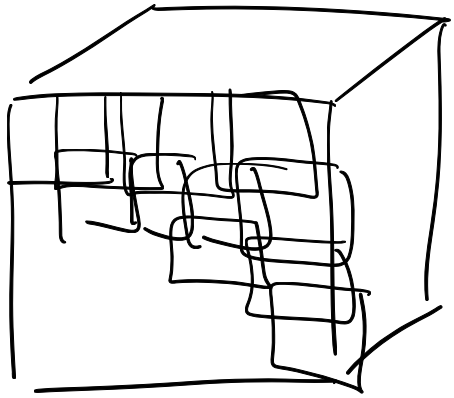
CE - with logits (dim=2)

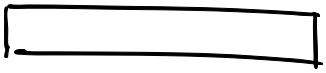
for softmax

Vs



Testing



$i,j =$ 

multiple predictions

$$\max_k \left\{ P(b_k \mid i,j) \right\}$$

$k = 0 \ldots 63$