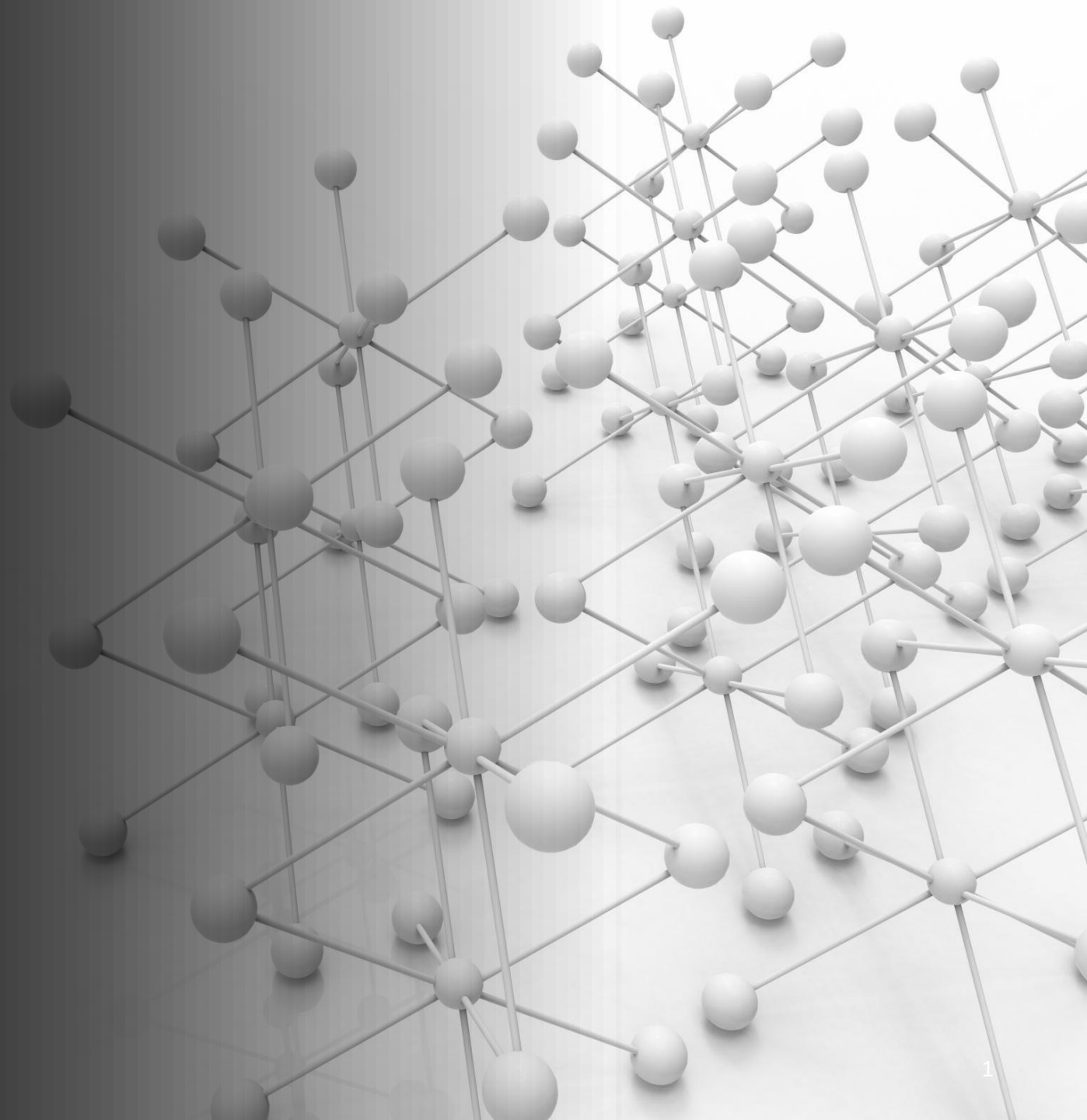
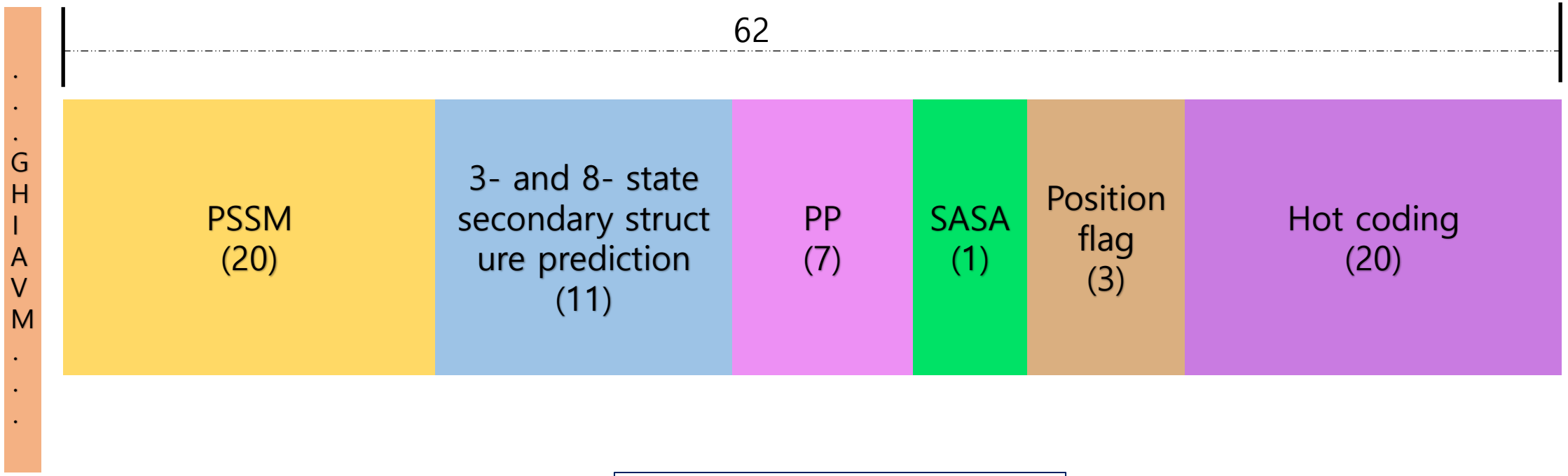
The Architecture of the LSTM code using for prediction of protein backbone geometry



Feature Description of the Protein Sequence



Protein Sequence (length L a.a)

Feature Matrix Size: L X 62

PharmCADD
Pharmaceutical Computer Aided Drug Design

Feature Description in detail

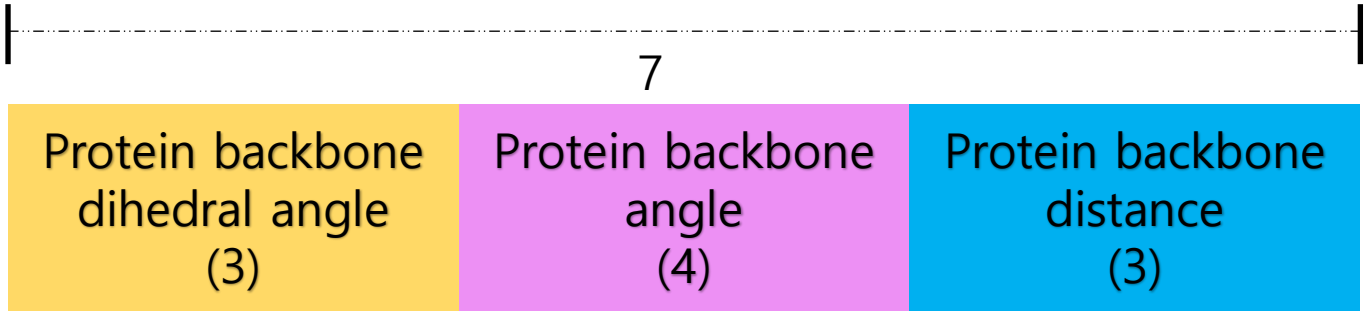
1. PSSM: **P**osition **S**pecific **S**coring **M**atrix (20)
   - Amino acid order in PSSM: A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V

2. Probability 3- and 8- state secondary structure prediction (11):
   - Obtained from spot1d program
   - Order: P(8-B), P(8-E), P(8-G), P(8-H), P(8-I), P(8-C), P(8-S), P(8-T), P(3-C), P(3-E), P(8-H),

3. PP (7): **P**hysio-chemical **P**roperties (7)
   - Obtained from **where?** [It is not clear from where these value have been taken. AAindex is a database of amino acid indices, amino acid mutation matrices, and pair-wise contact potentials. Currently, 566 Amino Acid indices are there.]
   - In order: 'Steric Param', 'Polarity', 'Volume', 'Hydrophobicity','Isoelectric Pt', 'Helix Prob', 'Sheet Prob',

4. SASA (1): solvent accessible surface area (**from where?)** [Most likely PP and SASA have been taken from fil e]
5. Position flag (3): pfm_start, pmf_middle and pfc_end
6. Hot encoding (20):
   - Order: ACDEFGHIKLMNPQRSTVWY

Parameters to be predicted (targets)



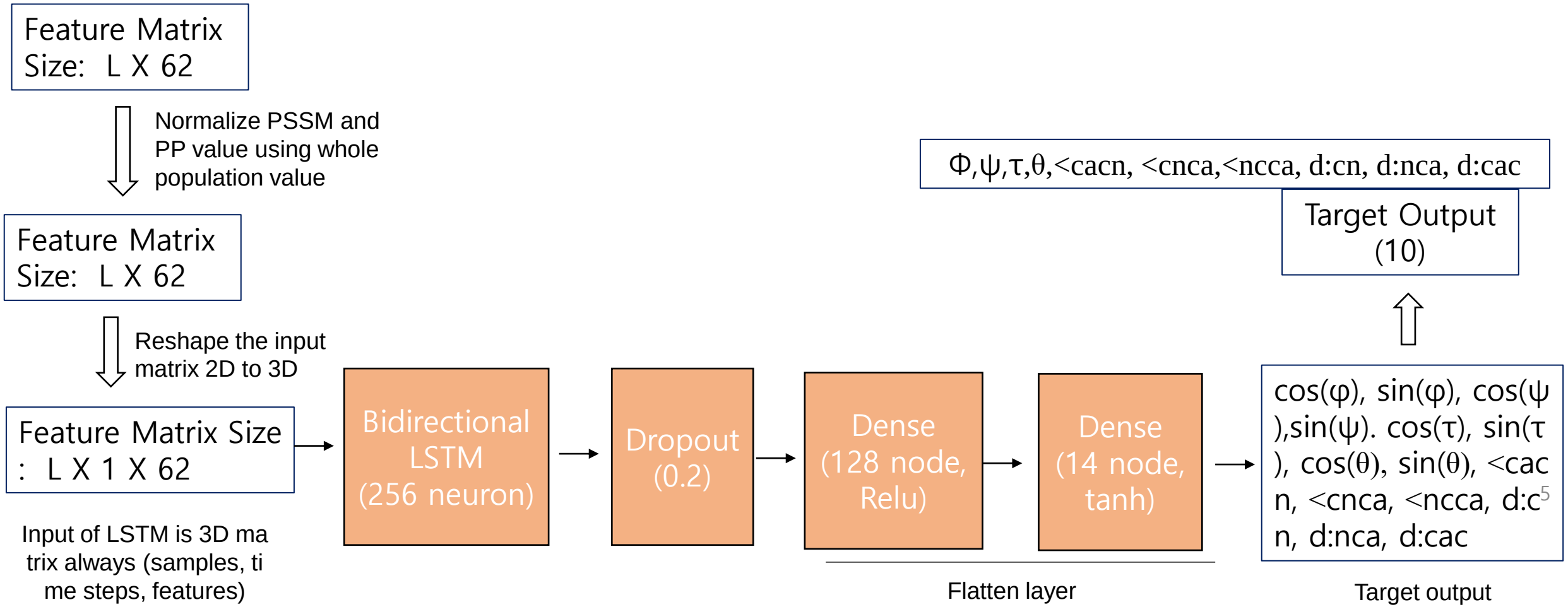
1. Protein backbone dihedral angles (3)
   - Phi (φ): C_i -N_i –CA_{i+1} –C; Psi(ψ):N_i -CA_1 –C_i –N_{i+1} and Tau (τ):CA_{i-1} -CA_{i-1} –CA_{i+1} –CA_{i+2}
   - Obtained from SPOT1D or in-house script from PDB structure (dssp). However, In-house script is not available as backup data.

2. Protein backbone angle (4):
   - CA_i-C_i-N_{i+1} (<cacn); C_{i-1}-N_i-CA_i (<cnca) and N_i-CA_i-C_i(<ncac); CA_{i-1}-CA_i-CA_{i+1} (θ)
   - Calculated from PDB structure. We must write in-house code as both data and code are not availa ble

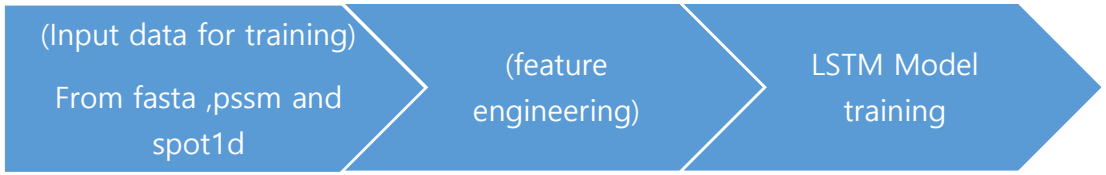
3. Protein backbone distance
   - Distance between C_{i-1}-N_i (d:cn); N_i-CA_i (d:nca) and CA_i-C_i(d:cac);
   - Calculated from PDB structure. (In-house code)

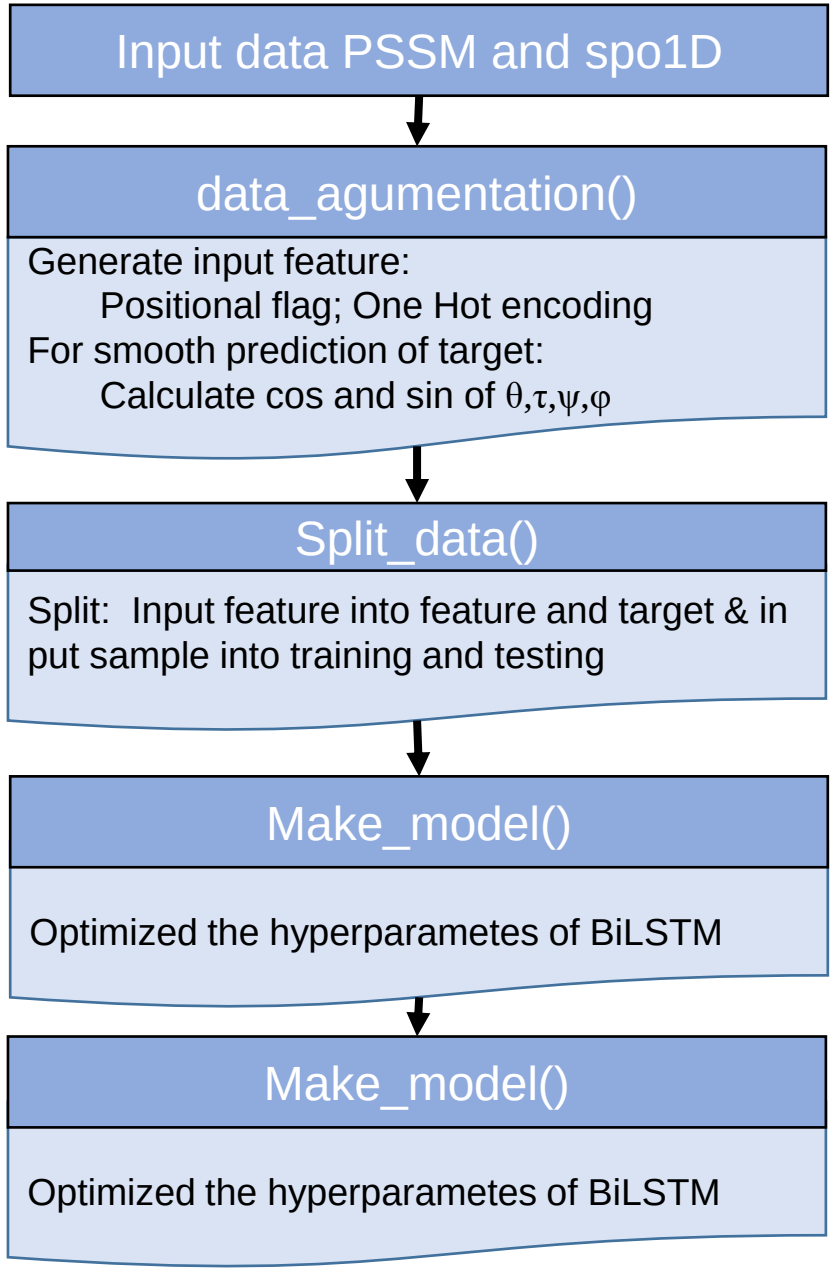
PharmCADD
Pharmaceutical Computer Aided Drug Design

The architecture of the LSTM code



Flow chart of the LSTM code





PharmCADD
Pharmaceutical Computer Aided Drug Design

Coding error in position flag calculation

The LSTM code: function data_augmentation()

```
# Designate position flag class: 0 at start, 2 at end, 1 in between
conditions = [data[:, COL['phi']] == '0', data[:, COL['psi']] == '0']
classes = [0, 2]
data = np.hstack((
    data,
    np.select(conditions, classes, default=1).reshape(-1, 1)
))

# Create and append One Hot Encoder columns for position flag
pfc = data[:, -1].astype(int)
ohef_matrix = np.zeros((pfc.size, pfc.max() + 1))
ohef_matrix[np.arange(pfc.size), pfc] = 1
data = np.hstack((
    data,
    ohef_matrix
))
```

Positional flag to identify start and end of a chain:
0 at start [1 0 0]; 2 at end [0 1 0] and 1 in between [0 0 1]
pfc[0], pfc[1] and pfc[2]

COL['phi'] is int, however comparing with string '0'

Therefore, it does not encode the terminal of protein properly

The LSTM code: function split_data()

```
# Array columns to be used as features for LSTM model
window_cols = [
    'index', 'A', 'R', 'N', 'D', 'C', 'Q', 'E', 'G', 'H', 'I', 'L', 'K',
    'M', 'F', 'P', 'S', 'T', 'W', 'Y', 'V', 'ss8_B', 'ss8_E', 'ss8_G',
    'ss8_H', 'ss8_I', 'ss8_C', 'ss8_S', 'ss8_T', 'ss3_C', 'ss3_E', 'ss3_H',
    'Steric Param', 'Polarity', 'Volume', 'Hydrophobicity',
    'Isoelectric Pt', 'Helix Prob', 'Sheet Prob', 'SASA', 'pfc_start',
    'pfc_middle', 'pfc_end', 'aa_A', 'aa_C', 'aa_D', 'aa_E', 'aa_F',
    'aa_G', 'aa_H', 'aa_I', 'aa_K', 'aa_L', 'aa_M', 'aa_N', 'aa_P',
    'aa_Q', 'aa_R', 'aa_S', 'aa_T', 'aa_V', 'aa_W', 'aa_Y',
    'position_flag'
]
```

Data_augmentation() created three positional flags. pfc[0], pfc[1] and pfc[2].
However, in time for splitting feature and target (split_data ()) four positional flags are mentioned.

PharmCADD
Pharmaceutical Computer Aided Drug Design

Phi and Psi are different in SPOT1D and DSSP

SPOT1D: 1ELK

jobc@ip-10-219-35-50: ~/pratiti/LSTM/complete

```
(base) jobc@ip-10-219-35-50:~/pratiti/LSTM/complete$ head -5 1ELKA.spot1d
#       AA      SS3     SS8     ASA     HSEa-u  HSEa-d  CN13    theta   tau     phi     psi     P(3-C)  P(3-E)  P(3-H)  P(8-C)  P(8-S)  P(8-T)  P(8-H)  P(8-G)  P(8-I)  P
(8-E)   P(8-B)
0       S       C       C       57.76   9.18    13.60   23.53   117.41  -161.90 -97.63  143.3   99.96   0.03    0.01    99.87   0.07    0.04    0.00    0.00    0.00    0
.01     0.00
1       D       C       C       104.03  4.29    13.55   19.26   104.35  -142.48 -85.81  21.95   88.32   1.18    10.49   78.89   1.83    9.17    3.95    4.85    0.00    0
.63     0.67
2       F       C       C       107.58  9.44    11.27   21.36   106.67  142.99  -87.22  55.11   80.01   2.34    17.65   42.03   16.43   22.06   4.53    12.47   0.00    1
.10     1.37
3       L       C       C       82.82   10.42   11.61   22.53   108.87  106.41  -90.04  100.7   79.05   2.83    18.13   43.70   16.20   19.35   4.42    13.41   0.00    1
.39     1.52
```

DSSP: 1ELK

#	RESIDUE	AA	STRUCTURE	BP1	BP2	ACC	N-H-->O	O-->H-N	N-H-->O	O-->H-N	TCO	KAPPA	ALPHA	PHI	PSI	X-CA	Y-CA	Z-CA
1	1	A	S	0	0	138	0, 0.0	0, 0.0	0, 0.0	0, 0.0	0.000	360.0	360.0	360.0	168.7	9.0	-7.5	27.7
2	2	A	D +	0	0	154	1,-0.1	2,-2.2	2,-0.1	0, 0.0	0.276	360.0	114.1	-89.5	5.6	11.5	-4.9	28.4
3	3	A	F +	0	0	165	2,-0.0	2,-0.4	0, 0.0	-1,-0.1	-0.499	44.5	162.9	-84.0	76.1	14.4	-7.4	27.9
4	4	A	L -	0	0	100	-2,-2.2	-2,-0.1	1,-0.1	0, 0.0	-0.765	22.8	-179.5	-98.4	138.2	15.8	-5.6	24.8
5	5	A	L +	0	0	147	-2,-0.4	2,-0.2	2,-0.1	-1,-0.1	0.556	44.6	107.1	-114.9	-11.0	19.3	-6.3	23.5
6	6	A	G S S-	0	0	49	1,-0.1	5,-0.1	2,-0.1	-2,-0.0	-0.546	79.1	-108.3	-74.1	132.8	19.8	-4.1	20.5
7	7	A	N > -	0	0	81	-2,-0.2	3,-2.4	1,-0.2	4,-0.4	-0.431	21.1	-139.5	-62.6	123.0	22.2	-1.2	21.1
8	8	A	P G > S+	0	0	19	0, 0.0	3,-0.6	0, 0.0	45,-0.2	0.780	100.1	57.4	-53.5	-28.4	20.1	2.0	21.2

I think we should calculate 10 target value from PDB structure, as PDB structure a re experimental value and SPOT1D

PharmCADD
Pharmaceutical Computer Aided Drug Design

Spot1D program

- There was some conflict to use the common spot1d programmer. Therefore, we have created our own folder “PSP” and make a run properly.

- Datasets:
  - PSSM_2.txt and fasta files of ~6500 and ~7000 are present in the backup data. However, spot1d files ar e not in the backup data. Therefore, we are generating spot1d files for all fasta.
  - “Clean_traning.pbz2” is not in the backup data. Therefore, we had to create it our-self

Interpretation and future direction:

- As we identify some issues in the LSTM code and unavailable of the training dataset. (Some feature and target value are missing) , we can **re-construct the LSTM model for protein structure prediction and try to improve performance of the model.**

OR

- We can model **the loop structure prediction model on base of the LSTM**

Need Dr. Wu's suggestion at the point

PharmCADD
Pharmaceutical Computer Aided Drug Design