Prediction of Conformational Ensembles of Disordered Proteins

Md Wasi Ul Kabir (mkabir3@uno.edu), Michael Carbone (mtcarbon@uno.edu), Avdesh Mishra (avdesh.mishra@tamuk.edu), Md Tamjidul Hoque (thoque@uno.edu)
Department of Computer Science, University of New Orleans, New Orleans, LA, USA

Introduction

- Many proteins do not have a stable 3D shape under physiological conditions and are therefore referred to as **intrinsically disordered proteins** (IDPs) or disordered regions in proteins (IDRs).
- These proteins, also known as flexible proteins (FPs), lack a defined structure and have variable conformations due to their **flexible backbone atom coordinates**.
- IDPs play a **crucial** role in **functional** proteomics, and their molecular recognition functions (MoRF) regions are responsible for important biological processes such as signaling, recognition, regulation, and cell division.

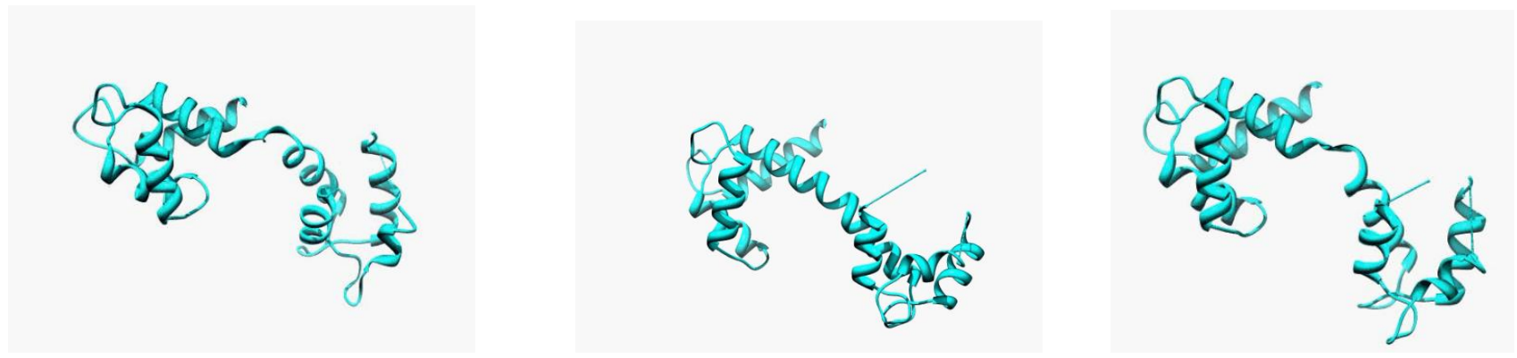


Figure 1. Three conformations of a disordered protein.

Motivation

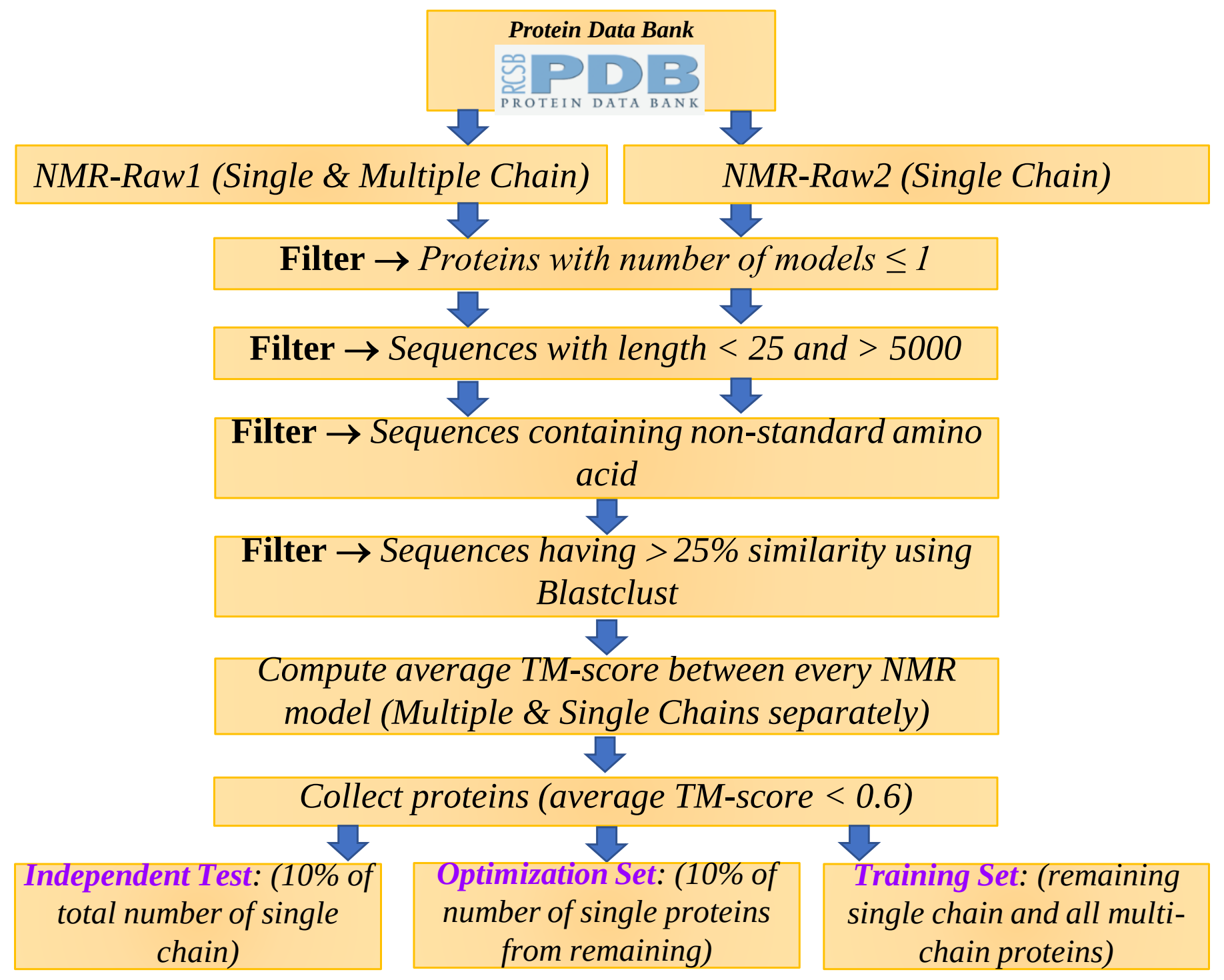
- The structural heterogeneity of IDPs is closely related to **critical human diseases** such as Parkinson's disease, Alzheimer's disease, type II diabetes, and cancer.
- Disordered Proteins are highly **abundant** in nature and their functional repertoire **complements** the functions of ordered proteins.
- To understand the function of flexible proteins, it is important to know the possible conformations that they can be found in. Thus, the generation of **possible conformational ensembles** of flexible proteins is important.
- Additionally, an **energy function** applicable to intrinsically disordered proteins is necessary to **rank** and **differentiate** low-energy conformations from the large ensemble of conformations generated during the search process.

Dataset

- Due to their structural heterogeneity, IDPs are best described by an ensemble of structures representing the conformations.
- The **Protein Ensemble Database** (PED) only has **286 ensembles** which is very small for this study.
- We create a new dataset from **Protein Data Bank** (PDB). Figure 3 shows the steps we have used to create the datasets.

Dataset

Training, Optimization and Independent Test Datasets



Protein Data Bank

NMR-Raw1 (Single & Multiple Chain)

NMR-Raw2 (Single Chain)

Filter → *Proteins with number of models ≤ 1*

Filter → *Sequences with length < 25 and > 5000*

Filter → *Sequences containing non-standard amino acid*

Filter → *Sequences having > 25% similarity using Blastclust*

Compute average TM-score between every NMR model (Multiple & Single Chains separately)

Collect proteins (average TM-score < 0.6)

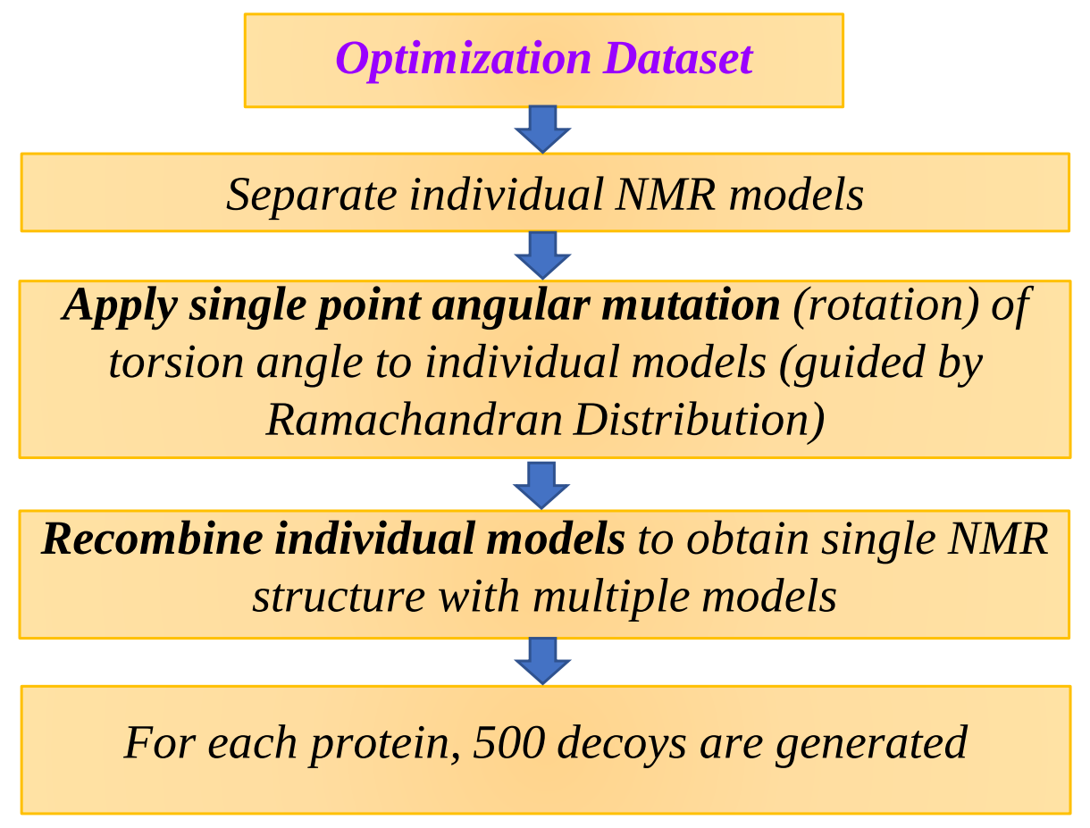
***Independent Test**: (10% of total number of single chain)*

***Optimization Set**: (10% of number of single proteins from remaining)*

***Training Set**: (remaining single chain and all multi-chain proteins)*

Figure 2. Illustrate the steps to create the dataset for training, optimization, and testing.

Construction of Optimization and Independent Test Decoys



Optimization Dataset

Separate individual NMR models

***Apply single point angular mutation** (rotation) of torsion angle to individual models (guided by Ramachandran Distribution)*

***Recombine individual models** to obtain single NMR structure with multiple models*

For each protein, 500 decoys are generated

***Similar** approach is used to obtain 500 decoys per protein *for **Independent test dataset**.*

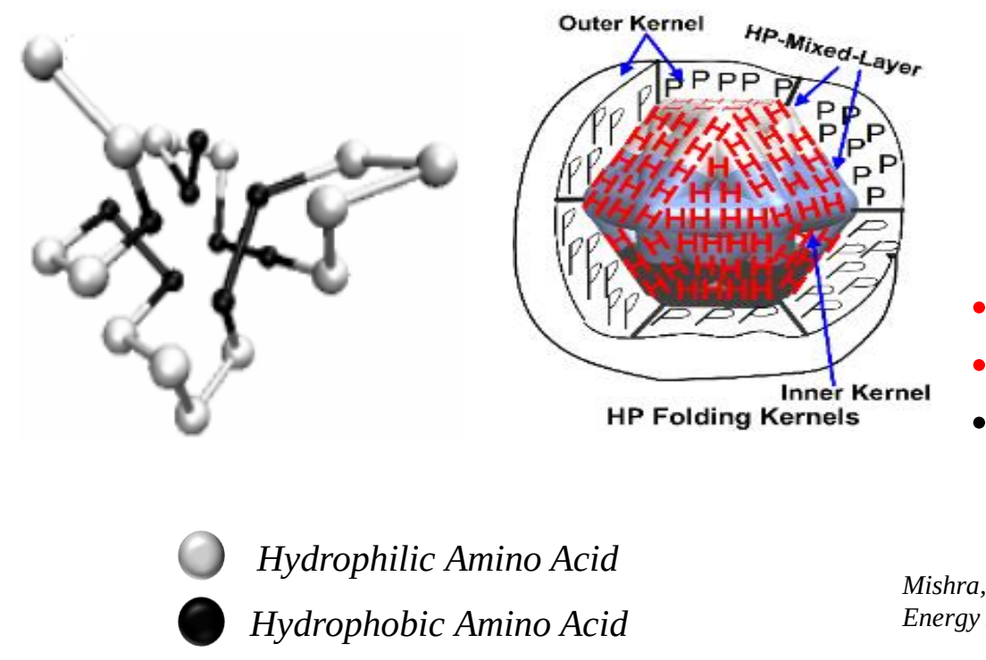
Figure 3. Generating decoy set by applying single point angular mutation.

Flexible Energy Function

Hydrophobic-Hydrophilic Properties

Hypothesis 1: ***Hydrophobic and hydrophilic interaction*** plays vital role in ***determining the conformation of the protein***

- ***Hydrophobic*** amino acid form the ***inner core*** of the protein
- ***Hydrophilic*** amino acid form the ***outer surface*** of the protein
- There exist a ***mixed hydrophobic-hydrophilic interaction layer*** in between inner core and outer surface



Therefore, to individually capture these interactions, we created three different interaction libraries:

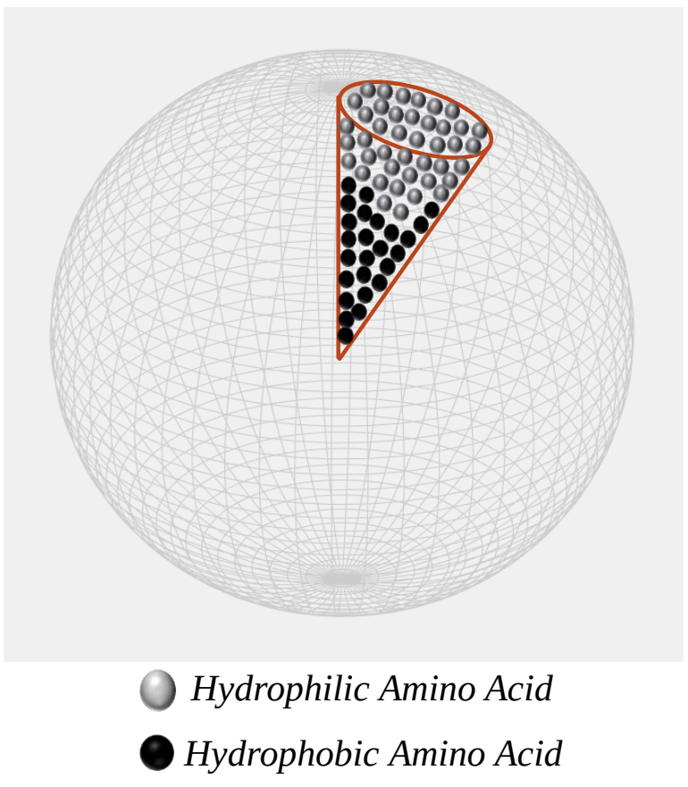
- Hydrophobic – Hydrophilic Interaction Library
- Hydrophobic – Hydrophobic Interaction Library
- Hydrophilic – Hydrophilic Interaction Library

Mishra, A. and M. T. Hoque (2017). "Three-Dimensional Ideal Gas Reference State Based Energy Function." Current Bioinformatics **12**(2): 171-180.

Ideal gas reference state

Hypothesis 2: Extends the concept of finite ideal gas reference state.

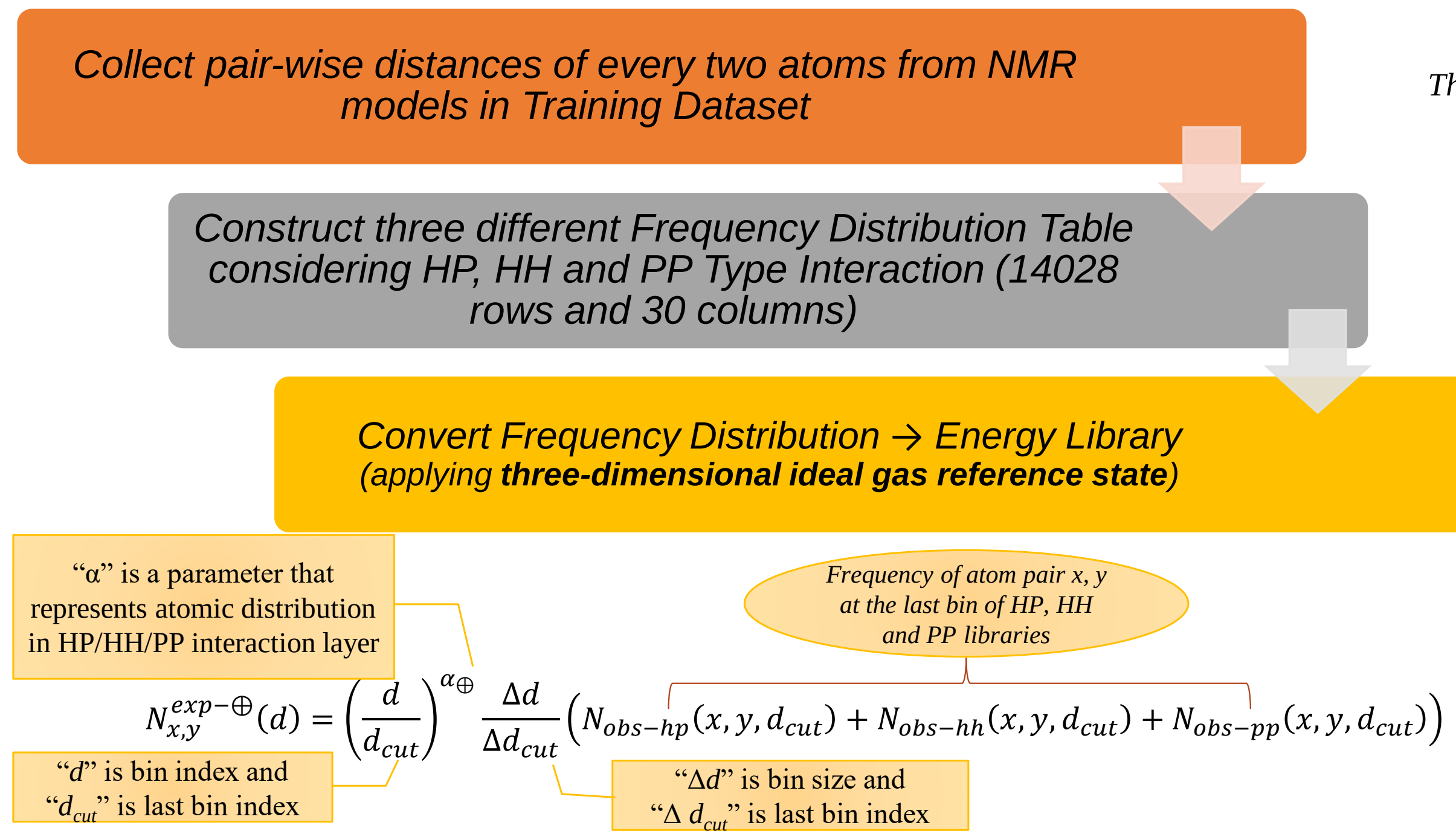
- *The well known "**finite ideal gas reference state**" **considers protein as a sphere and assumes that the expected number of atoms at distant bin "d" will increase in d^{α}*** where "α" is a constant determined through experiments.
- *Besides, we assume that the **expected number of atoms in the spherical environment increase gradually as we move from the center of the sphere to the surface.***
- Therefore, we derive ***three different reference state*** for three different interaction types (HP, HH and PP). Each reference state has its ***individual alphas (α_{hp}, α_{hh}, and α_{pp}).***
- Variable alphas is optimized using GA.



Zhou, H. and Y. Zhou (2002). "Distance-scaled, Finite Ideal-gas Reference State Improves Structure-derived Potentials of Mean Force for Structure Selection and Stability Prediction." Protein Sci. **11**: 2714–2726.

Mishra, A. and M. T. Hoque (2017). "Three-Dimensional Ideal Gas Reference State Based Energy Function." Current Bioinformatics **12**(2): 171-180.

Construction of 3D HP Libraries



Collect pair-wise distances of every two atoms from NMR models in Training Dataset

Construct three different Frequency Distribution Table considering HP, HH and PP Type Interaction (14028 rows and 30 columns)

Convert Frequency Distribution → Energy Library
*(applying **three-dimensional ideal gas reference state**)*

"α" is a parameter that represents atomic distribution in HP/HH/PP interaction layer

Frequency of atom pair x, y at the last bin of HP, HH and PP libraries

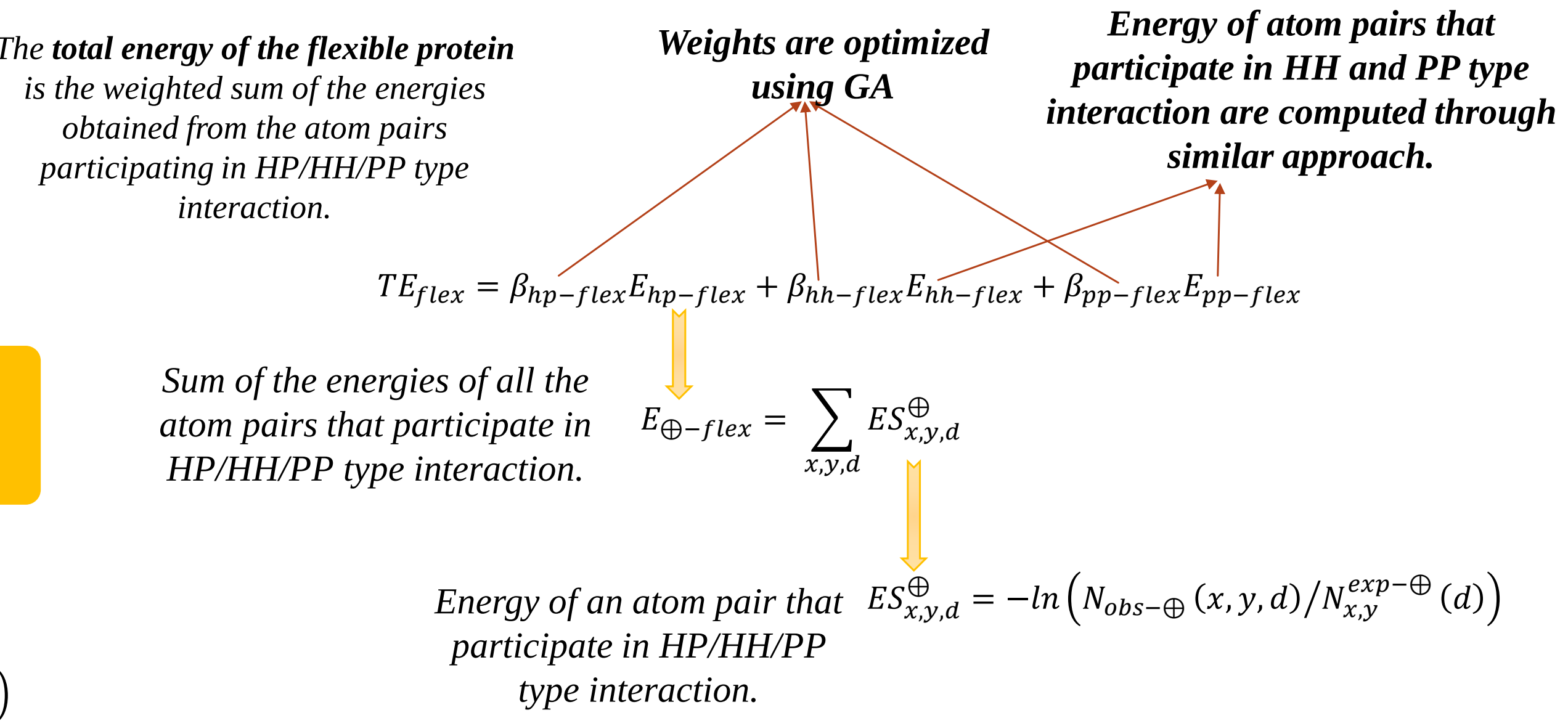
$$N_{x,y}^{exp-\oplus}(d) = \left(\frac{d}{d_{cut}}\right)^{\alpha_{\oplus}} \frac{\Delta d}{\Delta d_{cut}} \left(N_{obs-hp}(x,y,d_{cut}) + N_{obs-hh}(x,y,d_{cut}) + N_{obs-pp}(x,y,d_{cut})\right)$$

"d" is bin index and "d_{cut}" is last bin index

"Δd" is bin size and "Δ d_{cut}" is last bin index

Figure 4. Construction of energy library. $N_{x,y}^{exp-\oplus}(d)$ represents the expected number of atom pairs at distance d for ⊕ group, $N_{obs-\oplus}(x,y,d_{cut})$ represents the number of observation of atom pairs xth and yth observed at cutoff distance obtained from ⊕ library (specifically, the ⊕ represents HP or, HH or, PP) and $\alpha_{\oplus}$, is the parameter that belongs to ⊕ group, which is optimized by GA.

Scoring Technique



*The **total energy of the flexible protein** is the weighted sum of the energies obtained from the atom pairs participating in HP/HH/PP type interaction.*

Weights are optimized using GA

Energy of atom pairs that participate in HH and PP type interaction are computed through similar approach.

$$TE_{flex} = \beta_{hp-flex}E_{hp-flex} + \beta_{hh-flex}E_{hh-flex} + \beta_{pp-flex}E_{pp-flex}$$

Sum of the energies of all the atom pairs that participate in HP/HH/PP type interaction.

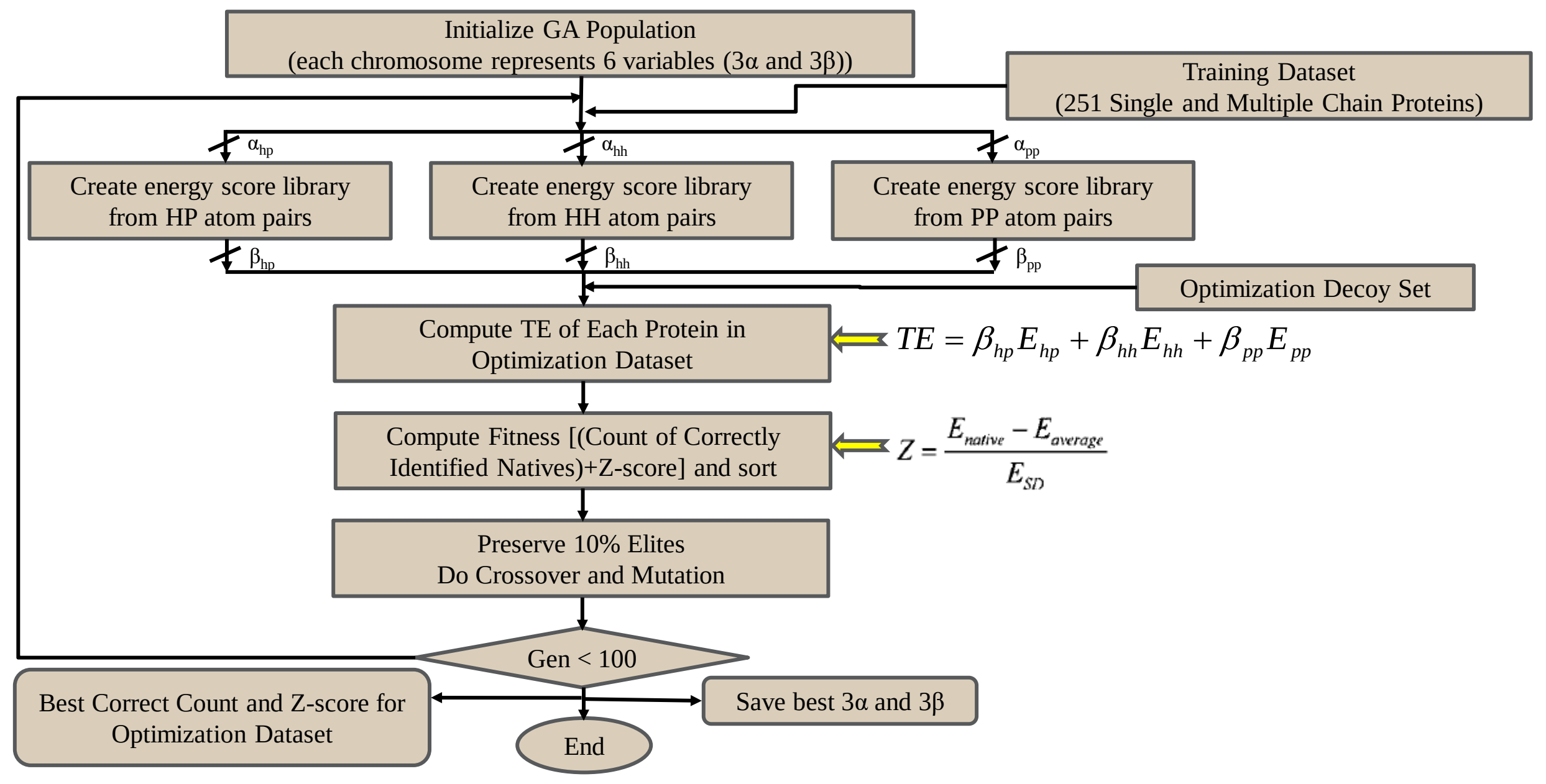
$$E_{\oplus-flex} = \sum_{x,y,d} ES_{x,y,d}^{\oplus}$$

Energy of an atom pair that participate in HP/HH/PP type interaction.

$$ES_{x,y,d}^{\oplus} = -ln\left(N_{obs-\oplus}(x,y,d)/N_{x,y}^{exp-\oplus}(d)\right)$$

Figure 5. Total energy score of flexible proteins.

Optimization using Genetic Algorithm



Initialize GA Population
(each chromosome represents 6 variables (3α and 3β))

Training Dataset
(251 Single and Multiple Chain Proteins)

Create energy score library from HP atom pairs

Create energy score library from HH atom pairs

Create energy score library from PP atom pairs

Optimization Decoy Set

Compute TE of Each Protein in Optimization Dataset

$$TE = \beta_{hp}E_{hp} + \beta_{hh}E_{hh} + \beta_{pp}E_{pp}$$

Compute Fitness [(Count of Correctly Identified Natives)+Z-score] and sort

$$Z = \frac{E_{native} - E_{average}}{E_{SD}}$$

Preserve 10% Elites
Do Crossover and Mutation

Gen < 100

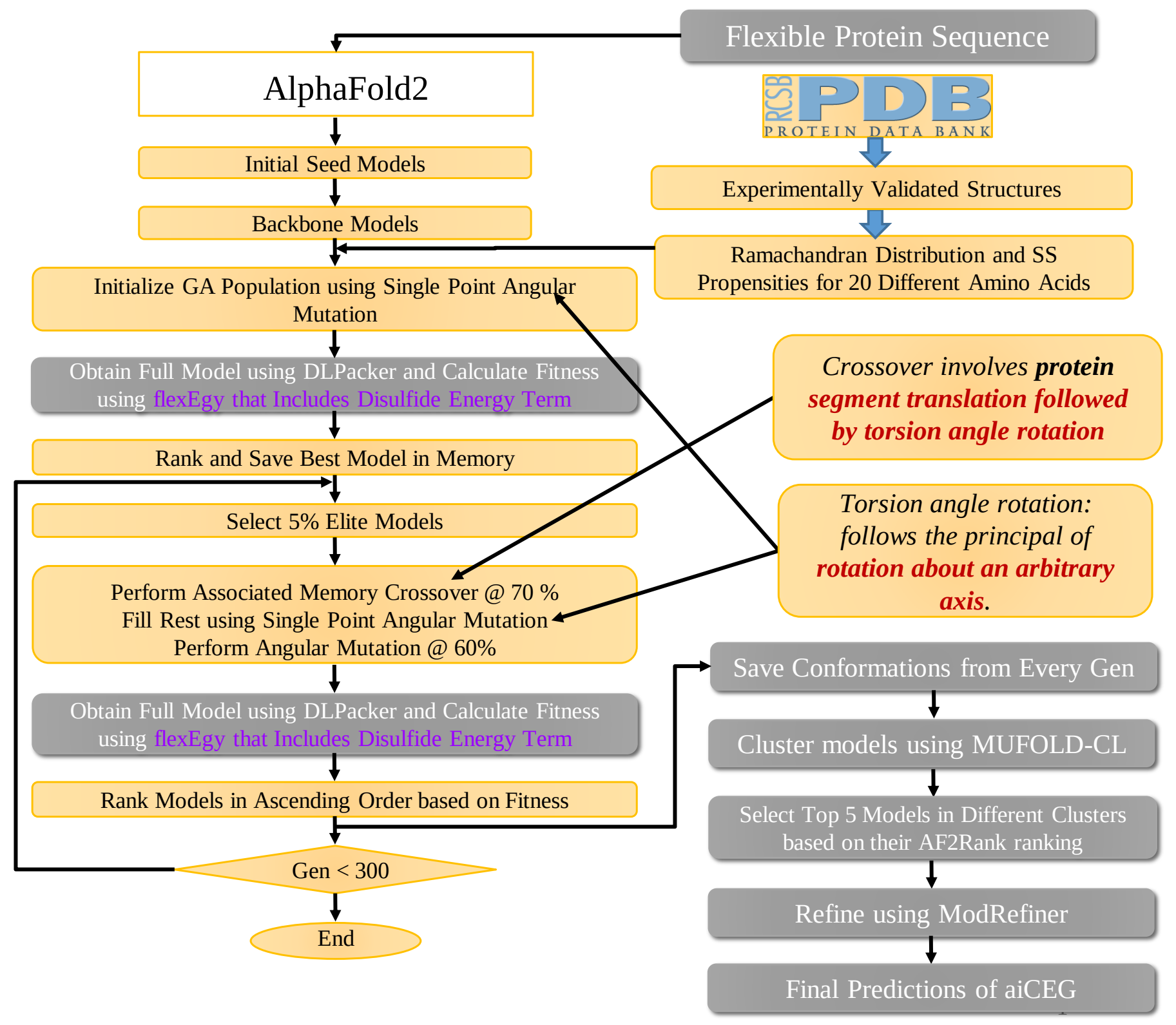
Best Correct Count and Z-score for Optimization Dataset

Save best 3α and 3β

End

Figure 6. Optimizing the 3α and 3β parameters using Genetic Algorithm (GA).

Ab Initio Conformational Ensemble Generator



Flexible Protein Sequence

AlphaFold2

Initial Seed Models

Backbone Models

Experimentally Validated Structures

Ramachandran Distribution and SS Propensities for 20 Different Amino Acids

Initialize GA Population using Single Point Angular Mutation

Obtain Full Model using DLPacker and Calculate Fitness using flexEgy that Includes Disulfide Energy Term

Rank and Save Best Model in Memory

Select 5% Elite Models

Perform Associated Memory Crossover @ 70 %
Fill Rest using Single Point Angular Mutation
Perform Angular Mutation @ 60%

Obtain Full Model using DLPacker and Calculate Fitness using flexEgy that Includes Disulfide Energy Term

Rank Models in Ascending Order based on Fitness

Gen < 300

End

*Crossover involves **protein segment translation followed by torsion angle rotation***

*Torsion angle rotation: follows the principal of **rotation about an arbitrary axis**.*

Save Conformations from Every Gen

Cluster models using MUFOLD-CL

Select Top 5 Models in Different Clusters based on their AF2Rank ranking

Refine using ModRefiner

Final Predictions of aiCEG

Figure 7. Conformational Ensembles Generator of Disordered Proteins using Genetic Algorithm.

Conclusion and Future Plans

- In this study, we propose a flexible energy function and an ab initio conformational ensemble generator to predict the conformational ensembles of disordered proteins.
- We are currently implementing the proposed method. Our main challenge is the computational resources needed to run the ensemble generator pipeline.
- We intend to improve the flexible energy function with the prediction of the disulfide bond.
- We also plan to incorporate the disordered predictor tool, Dispredict3.0, in the Genetic Algorithm to apply crossover and mutation in disordered regions.
- Furthermore, we will investigate Molecular dynamics (MD) simulation for disordered proteins.
- This work would be novel because, to the best of our knowledge, no computational method exists to predict the conformations of disordered proteins.