Data based approach for RNA folding

Anirudh Panigrahi (2018EE10446)
Kanishk Goyal (2018EE10471)
Supervised by Dr. Shaunak Sen, IIT Delhi
Department of Electrical Engineering

Importance of Computational Methods for RNA Folding

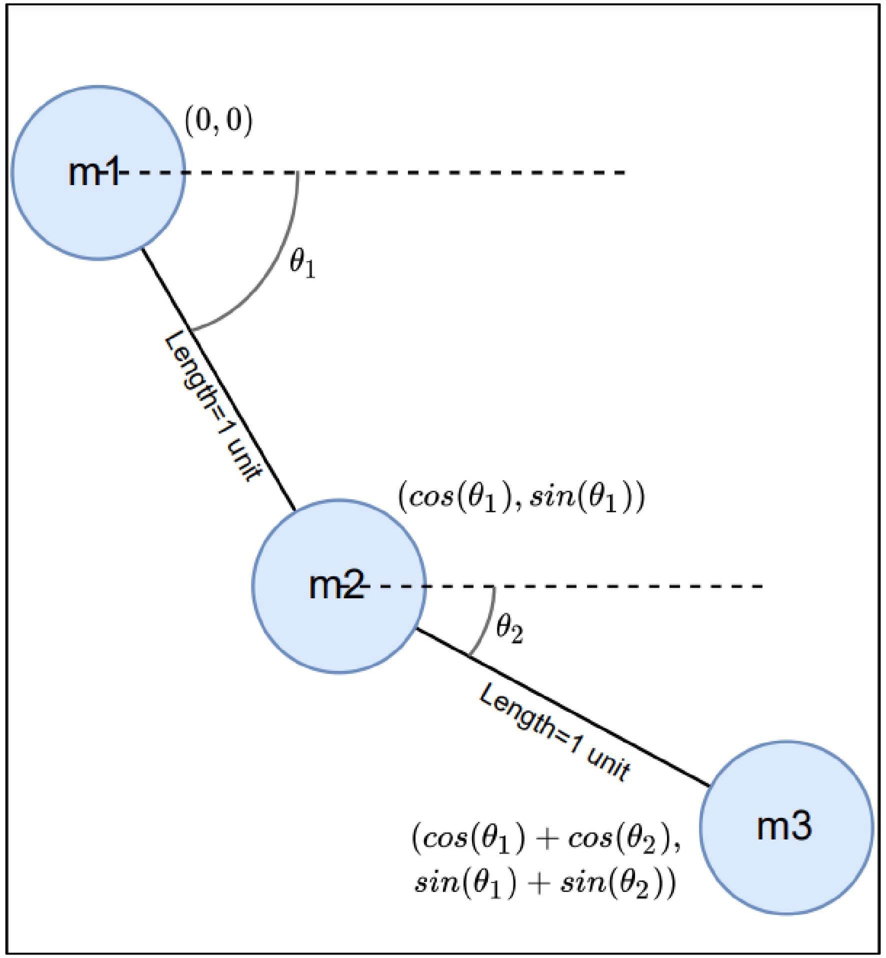
- 3D structure is important for understanding interactions of RNA
- Difficult to predict 3D structure of RNA
- Models developed for protein folding are not applicable due to limited number of known RNA structures.

Problem Statement

- To find these Structural Motifs of RNA molecules for N = 3,4,5, where N is the number of beads in the molecule using different numerical and analytical techniques. We want to find an efficient and scalable approach to solve this problem.

A simple Mathematical Model for RNA folding

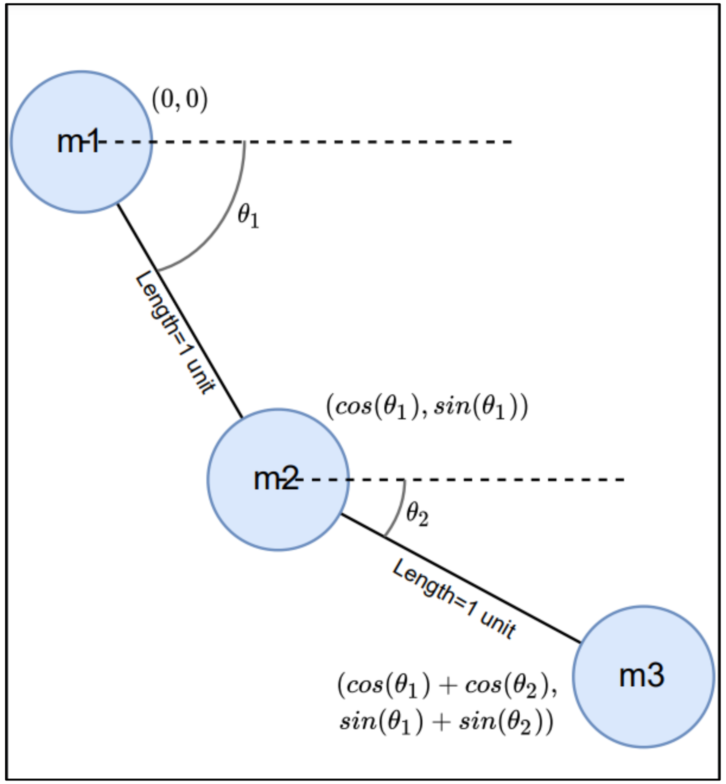
- To model a N molecule RNA chain, we can consider each molecule as a bead. We have assumed these beads to be interconnected to each other with a rigid rod of unit length.
- We can assume that the 1st bead is present at the origin. For beads 2,3....N, let, θ_{i-1} represent the angle between the bead and horizontal measured in the clockwise sense.
- The coordinate of i^{th} bead can be calculated as $(\sum_{j=1}^{j=i-1} \cos(\theta j),\ \sum_{j=1}^{j=i-1} \sin(\theta j))$ in the x-y plane.



Potential Function as per Mass-Spring model

The potential V_{ij} (as per the mass spring model)between beads i and j can be calculated as,

$$
\begin{aligned}
V_{ij} &= \frac{1}{2} d_{ij}^2 \qquad\qquad\qquad \forall i > j \ \ \text{(as per the mass spring model)} \\
&= \frac{1}{2}[(\sum_{k=1}^{k=i-1} \cos(\theta_k) - \sum_{k=1}^{k=j-1} \cos(\theta_k))^2 + (\sum_{k=1}^{k=i-1} \sin(\theta_k) - \sum_{k=1}^{k=j-1} \sin(\theta_k))^2] \\
&= \frac{1}{2}[(\sum_{k=j}^{k=i-1} \cos(\theta_k))^2 + (\sum_{k=j}^{k=i-1} \sin(\theta_k))^2] \\
&= \frac{1}{2}[(i-j) + \sum_{j \le p \le q \le i-1}\sum \cos(\theta_p - \theta_q)]
\end{aligned}
$$



Different approaches used

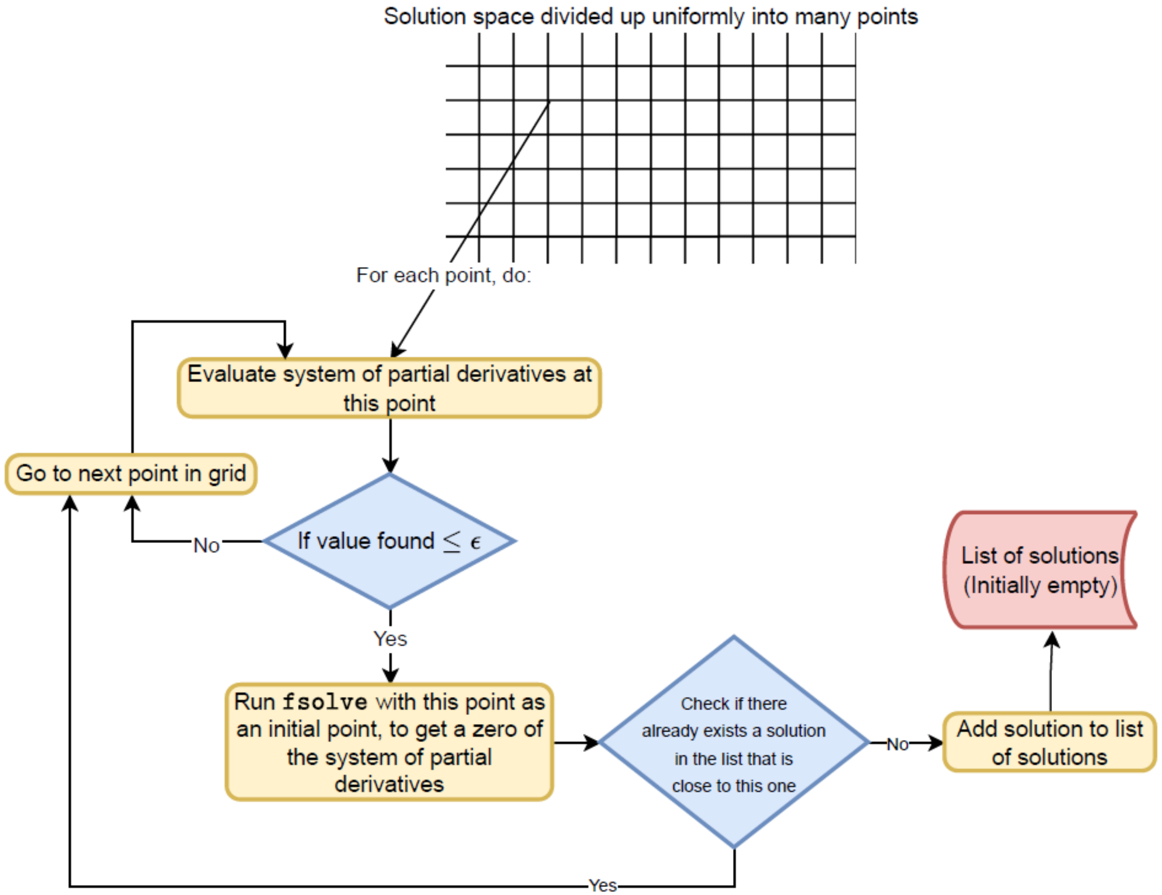
- We obtain the SMs by finding the extrema and the inflection points of the potential energy function and finding the θ_k values corresponding to those points.
- We try to find the points where the partial derivatives of the potential functions are zero. To impose rotational invariance, we assume $\theta_1 = 0$.
- We used the following three approaches to solve the problem
  - Numerical minimization of potential functions
  - Solving derivatives of potential functions numerically
  - Arcsin method

Numerical minimization of potential functions

- We find the maximas and minimas of the potential function numerically.
- We use MATLAB’s Optimization Toolbox to numerically minimize the potential function.
- To find the maximas is a simple matter of minimizing the negative of the potential function.
- Numerical optimizer of MATLAB requires an initial point as input, which we randomly generate.
- Hence, the algorithm needs to be run multiple times to find all the minima.

Solving derivatives of potential functions numerically

- We numerically solve the partial derivatives of the potential function to find all its zeros.
- Use a numerical solver to find zeros, which requires an initial point as input.
- For initial points, we divide solution space uniformly into many points, and use those points where all partial derivatives evaluate to a small value(less than a threshold)
- Double partial derivatives can tell us whether the obtained zero of the partial derivatives is a maxima, minima or saddle point.



Flowchart of the algorithm used

Arcsin Method (Direct Minimization)

- Considering the partial derivative of the potential function w.r.t θ, we can convert the problem to a system of equations, that can be represented as [A][B] = [0], where B is a column matrix of differences of θ ′ s taken two at a time and A is the corresponding coefficient matrix.
- For N=4 we have,

$$\mathbf{A} = \begin{bmatrix} -a_{12} & 0 & a_{13} \\ a_{12} & -a_{23} & 0 \\ 0 & a_{23} & -a_{13} \end{bmatrix} \text{where } a_{12} = a_{23} = a_{13} = 1 \qquad \mathbf{B} = \begin{bmatrix} \sin(\theta_1 - \theta_2) \\ \sin(\theta_2 - \theta_3) \\ \sin(\theta_3 - \theta_1) \end{bmatrix}$$

- The null space is given by,

$$C \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \sin(\theta_1 - \theta_2) \\ \sin(\theta_2 - \theta_3) \\ \sin(\theta_3 - \theta_1) \end{bmatrix}$$

Arcsin Method (Direct Minimization): Solving for C

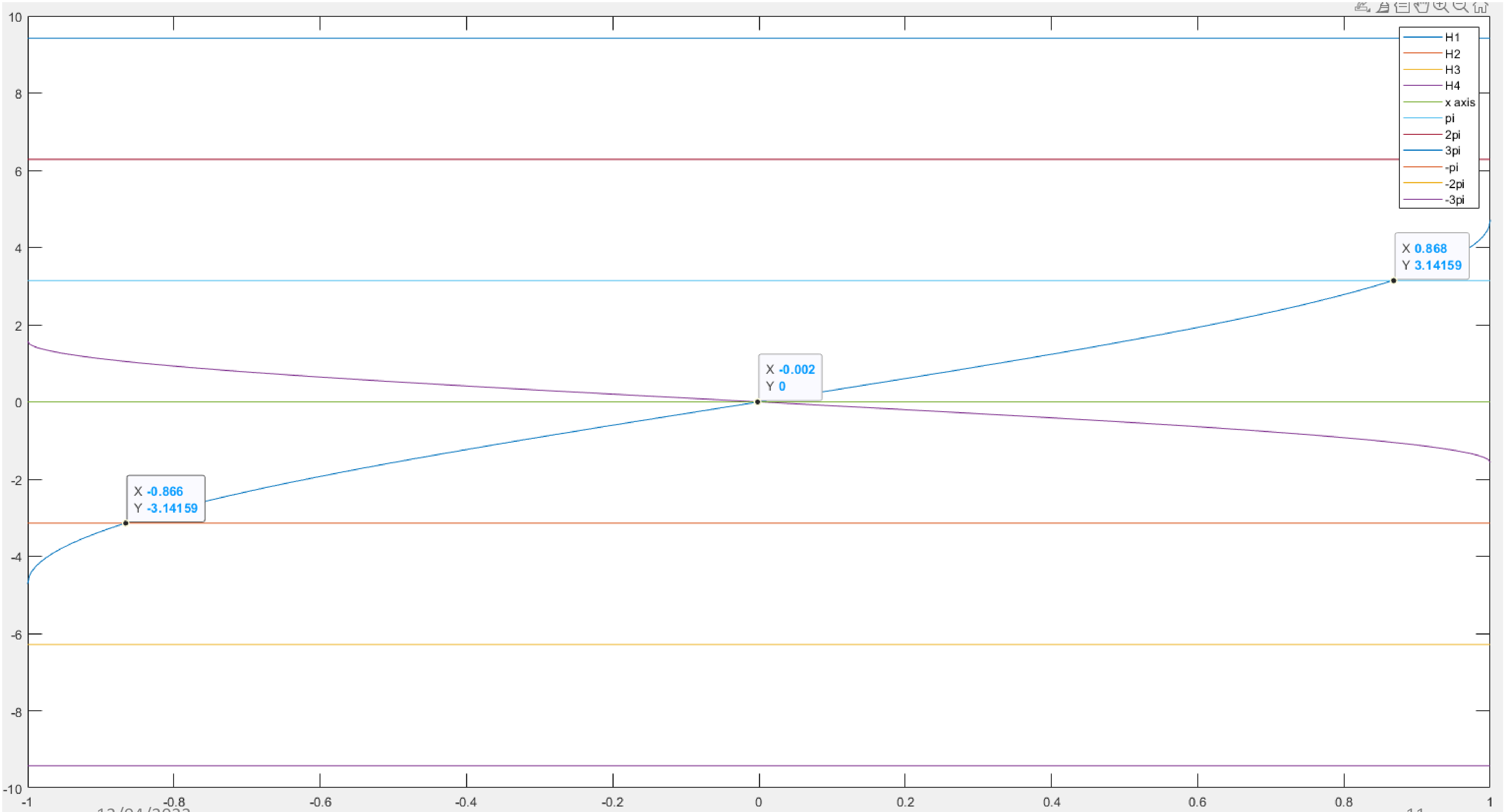
- Taking inverse we get,

$$\theta_1 - \theta_2 + \theta_2 - \theta_3 + \theta_3 - \theta_1 = p\pi + (-1)^p \sin^{-1}(C) + q\pi + (-1)^q \sin^{-1}(C\frac{a_{12}}{a_{23}}) + r\pi + (-1)^r \sin^{-1}(C\frac{a_{12}}{a_{13}})$$

- We then get 8 equations depending on the parity of p, q and r. After solving, the 8 equations reduce to the following 4 equations:

$$n\pi = (-1)^{(n+1)}[\sin^{-1}(C) + \sin^{-1}(C\frac{a_{12}}{a_{23}}) + \sin^{-1}(C\frac{a_{12}}{a_{13}})]$$
$$n\pi = (-1)^{(n+1)}[-\sin^{-1}(C) - \sin^{-1}(C\frac{a_{12}}{a_{23}}) + \sin^{-1}(C\frac{a_{12}}{a_{13}})]$$
$$n\pi = (-1)^{(n+1)}[-\sin^{-1}(C) + \sin^{-1}(C\frac{a_{12}}{a_{23}}) - \sin^{-1}(C\frac{a_{12}}{a_{13}})]$$
$$n\pi = (-1)^{(n+1)}[\sin^{-1}(C) - \sin^{-1}(C\frac{a_{12}}{a_{23}}) - \sin^{-1}(C\frac{a_{12}}{a_{13}})]$$

- We then plot the above 4 equations and solve them graphically, to find values of C.


H1
H2
H3
H4
x axis
pi
2pi
3pi
-pi
-2pi
-3pi
X 0.868
Y 3.14159
X -0.002
Y 0
X -0.866
Y -3.14159
10
8
6
4
2
0
-2
-4
-6
-8
-10
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

Arcsin Method (Direct Minimization)

Proceeding in a similar manner for N=5 we get,

$$\mathbf{A} = \begin{bmatrix} -1' & -1 & 0 & 0 & 0 & 1 \\ 1 & 0 & -1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 1 & -1 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} \sin(\theta_1 - \theta_2) \\ \sin(\theta_1 - \theta_3) \\ \sin(\theta_2 - \theta_3) \\ \sin(\theta_2 - \theta_4) \\ \sin(\theta_3 - \theta_4) \\ \sin(\theta_4 - \theta_1) \end{bmatrix}$$

$$C_1 \begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + C_2 \begin{bmatrix} -1 \\ 1 \\ 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} + C_3 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \sin(\theta_1 - \theta_2) \\ \sin(\theta_1 - \theta_3) \\ \sin(\theta_2 - \theta_3) \\ \sin(\theta_2 - \theta_4) \\ \sin(\theta_3 - \theta_4) \\ \sin(\theta_4 - \theta_1) \end{bmatrix}$$

- We considered 3 angles at a time and solved four set of equations and took the intersection of their solutions.

$$C_1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + C_2 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} + C_3 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \sin(\theta_1 - \theta_2) \\ \sin(\theta_2 - \theta_3) \\ \sin(\theta_3 - \theta_1) \end{bmatrix}$$

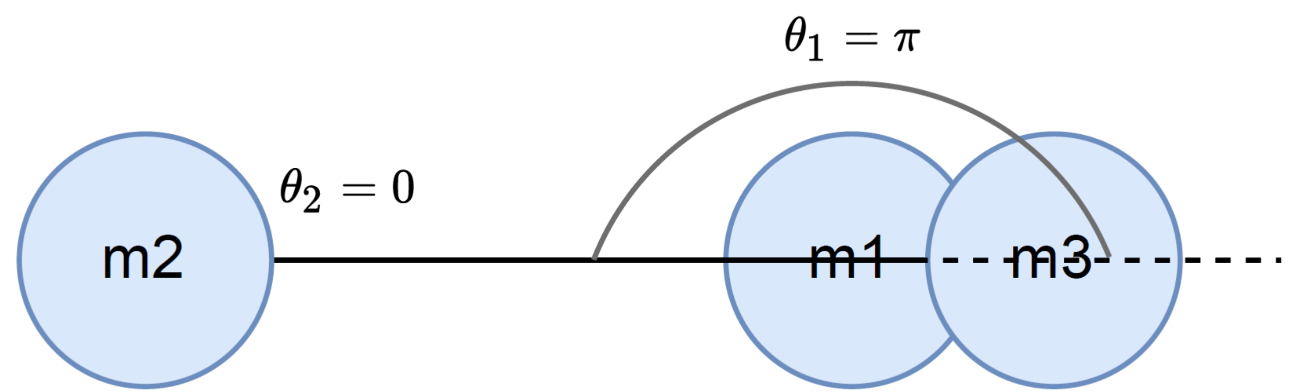
- This is the set of rows with θ_1, θ_2 and θ_3; rows containing θ_4 are omitted. Similarly, we constructed 3 other sets of equations with (θ_1, θ_2, θ_4), (θ_1, θ_3, θ_4) and (θ_2, θ_3, θ_4) terms.

Results obtained

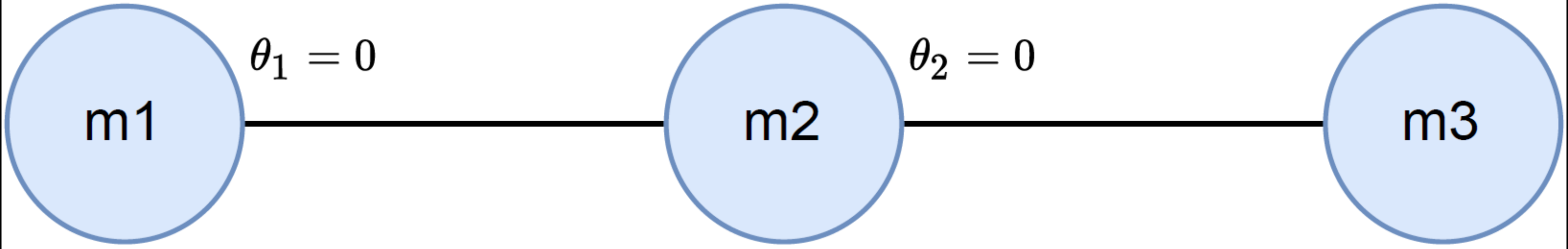
- All 3 methods yielded identical results.

N	Solution point #	Type of point	θ_1	θ_2	θ_3	θ_4
3	1	Minima	0	π	-	-
	2	Maxima	0	0	-	-
4	3	Minima	0	$4\pi/3$	$2\pi/3$	-
	4	Maxima	0	0	0	-
	5	Saddle pt	0	π	0	-
5	6	Minima	0	$3\pi/2$	π	$\pi/2$
	7	Maxima	0	0	0	0
	8	Saddle pt	0	π	$\pi/2$	$3\pi/2$

Structures of the solution points found: N=3

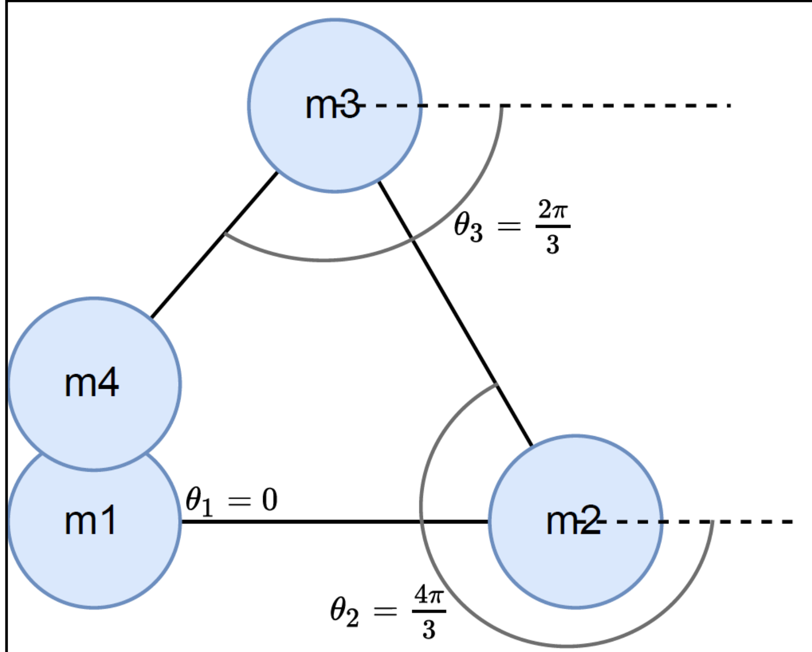


Minima structure

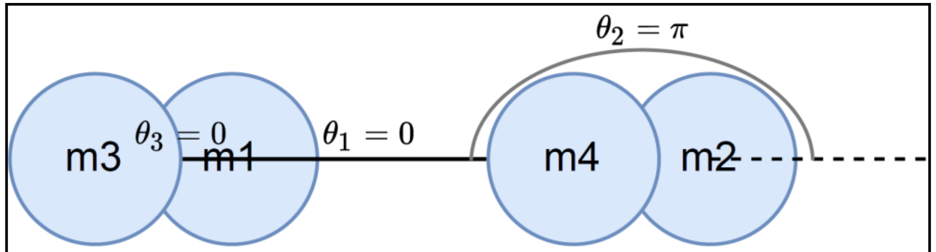


Maxima structure

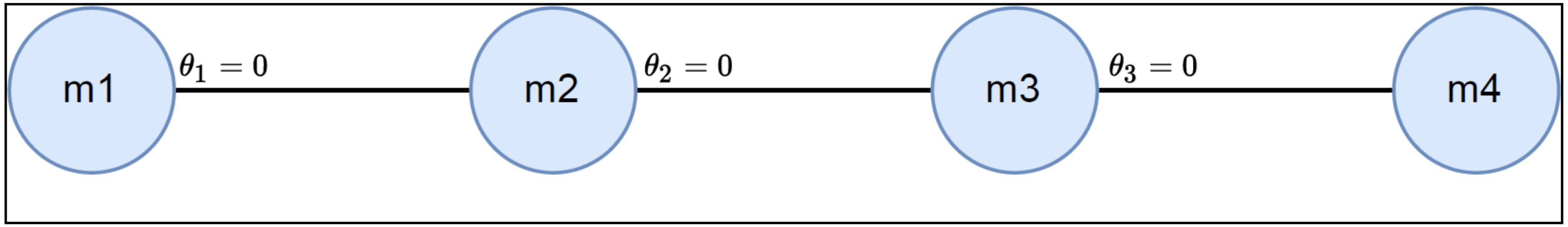
Structures of the solution points found: N=4



Minima structure

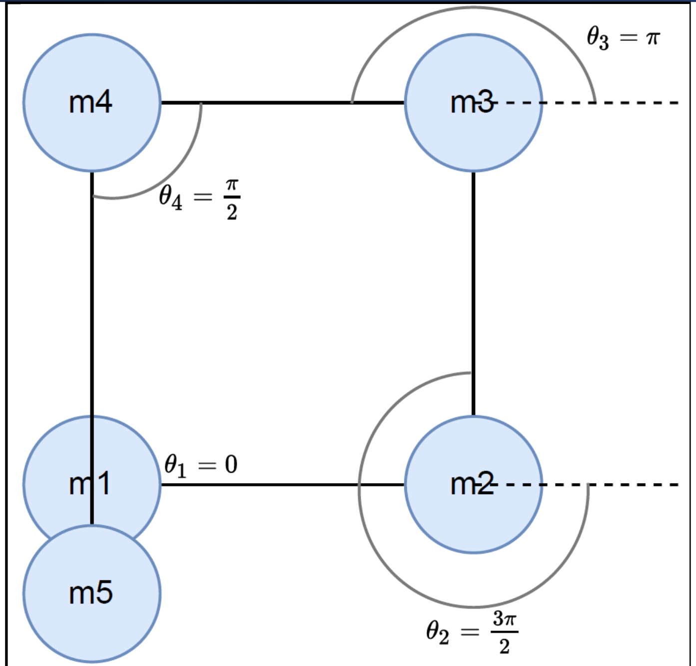


Saddle point structure

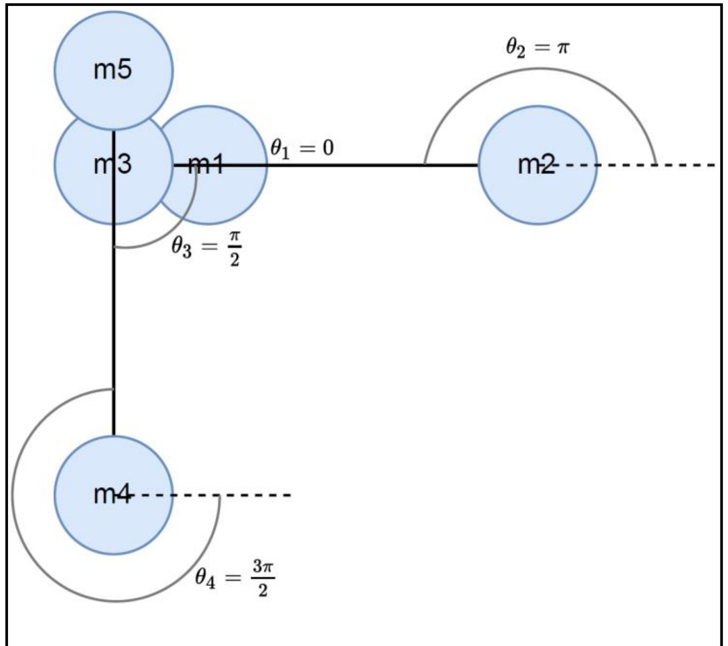


Maxima structure

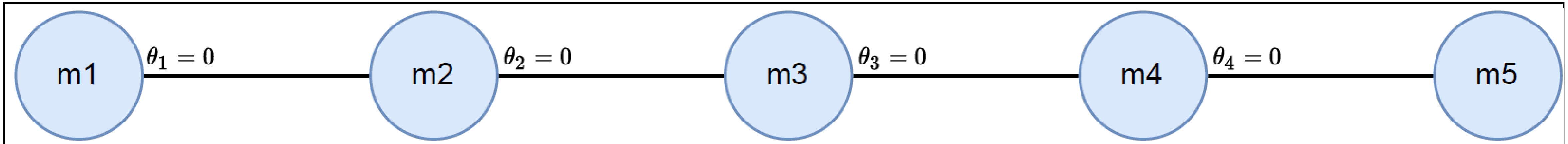
Structures of the solution points found: N=5



Minima structure



Saddle point structure



Maxima structure

Comparison of the 3 approaches

- All 3 approaches gave same result.
- Numerical minimization is general and efficient but it cannot find the saddle points of the potential function.
- Solving derivatives of potential functions is very general and it works for all N. However, it becomes very computationally intensive for higher values of N. Another issue with this approach is that it outputs the same solution point multiple times.
- The Arcsin method cannot be generalized to higher number of particles. It becomes increasingly impractical to graphically look for solutions for N>4

Future work

- The method of solving derivatives numerically becomes sluggish at higher values of N. It can be further optimized using dynamic programming and related approaches
- The potential function can be refined by incorporating external forces and effects of changes in external conditions like pressure, temperature etc.

THANK YOU