Clustering Data for the Docking@Home Project

By: Brenda Medina

Intern from University of Texas at El Paso(UTEP)
As part of the CRA-W DMP at the University of Delaware (UDel)

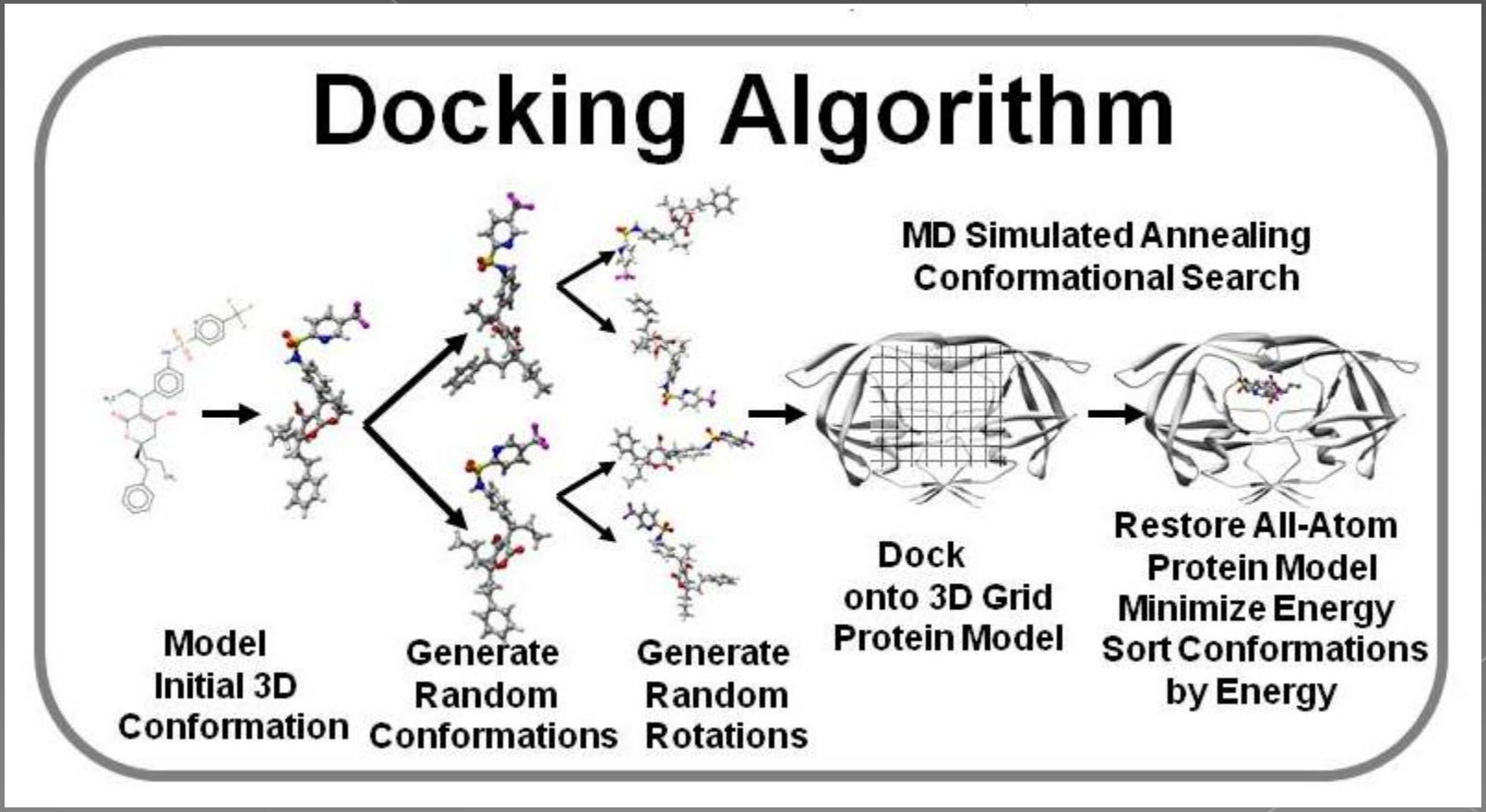
OUTLINE

- **Overview**
- Purpose
- Program
  - Clustering algorithm
  - Implementation overview
- Limitations
- Parameters
  - Data
  - HIV protein
  - Ligands
- Results
  - Complex 1hvi
  - Complex 1hvj
- Future work
- Acknowledgments

Overview

- Why understanding protein-ligand interactions is important?
  - Development of new pharmaceutical drugs
  - Determining protein function
- Why simulate protein-ligand interactions?
  - Wet lab approach is expensive in terms of: people, money, resources, and time

Overview


Docking Algorithm
MD Simulated Annealing
Conformational Search
Model
Initial 3D
Conformation
Generate
Random
Conformations
Generate
Random
Rotations
Dock
onto 3D Grid
Protein Model
Restore All-Atom
Protein Model
Minimize Energy
Sort Conformations
by Energy

Overview

- Problem with simulations:
  - Large sample of docked protein-ligand complexes
- Solution: uncover patterns through clustering

OUTLINE

- Overview
- **Purpose**
- Program
  - Clustering algorithm
  - Implementation overview
- Limitations
- Parameters
  - Data
  - HIV protein
  - Ligands
- Results
  - Complex 1hvi
  - Complex 1hvj
- Future work
- Acknowledgments

Purpose

- Clustering attempts to uncover the a correlation between the following:
  - The force field used and the docking convergence
    - This will aid in developing a method to automatically cluster ligands
  - The lowest energy and the root mean square deviation
    - This will aid in developing a method which automatically selects the ligand(s), conformation(s) and rotation(s), which minimizes protein-ligand complex energy

OUTLINE

- Overview
- Purpose
- **Program**
  - Clustering algorithm
  - Implementation overview
- Limitations
- Parameters
  - Data
  - HIV protein
  - Ligands
- Results
  - Complex 1hvi
  - Complex 1hvj
- Future work
- Acknowledgments

Program: Clustering algorithm

- Clustering method: K-Mean
  1. Randomly select K centroids
  2. Compute distance from all data points to every centroid
  3. Assign cluster membership: minimize data point-centroid distance
  4. Repeat steps 2 and 3 until no data point switches clusters
- Distance: Root Mean Square Deviation (RMSD)

Program: Implementation overview

Calculate distances among all docking results in one complex

Apply clustering algorithm to each and every complex

Output results of clustering

OUTLINE

- Overview
- Purpose
- Program
  - Clustering algorithm
  - Implementation overview
- **Limitations**
- Parameters
  - Data
  - HIV protein
  - Ligands
- Results
  - Complex 1hvi
  - Complex 1hvj
- Future work
- Acknowledgments

Limitations

- Challenges of K-Mean:
  - Randomness
  - Calculating centroids
  - Choosing K
- Every clustering algorithm has challenges
- BUT… these challenges drive future work: accuracy improvement

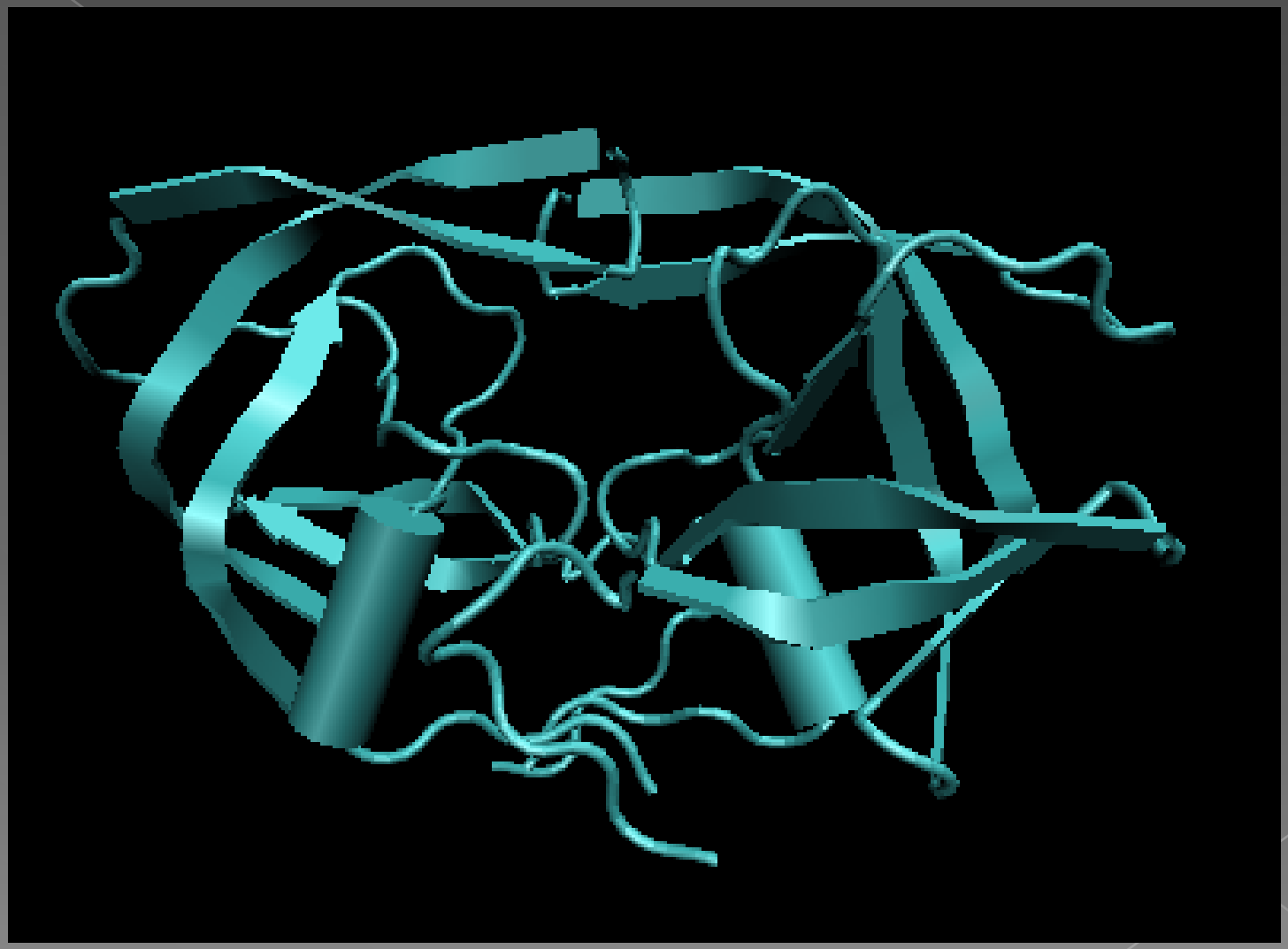
OUTLINE

- Overview
- Purpose
- Program
  - Clustering algorithm
  - Implementation overview
- Limitations
- **Parameters**
  - Data
  - HIV protein
  - Ligands
- Results
  - Complex 1hvi
  - Complex 1hvj
- Future work
- Acknowledgments

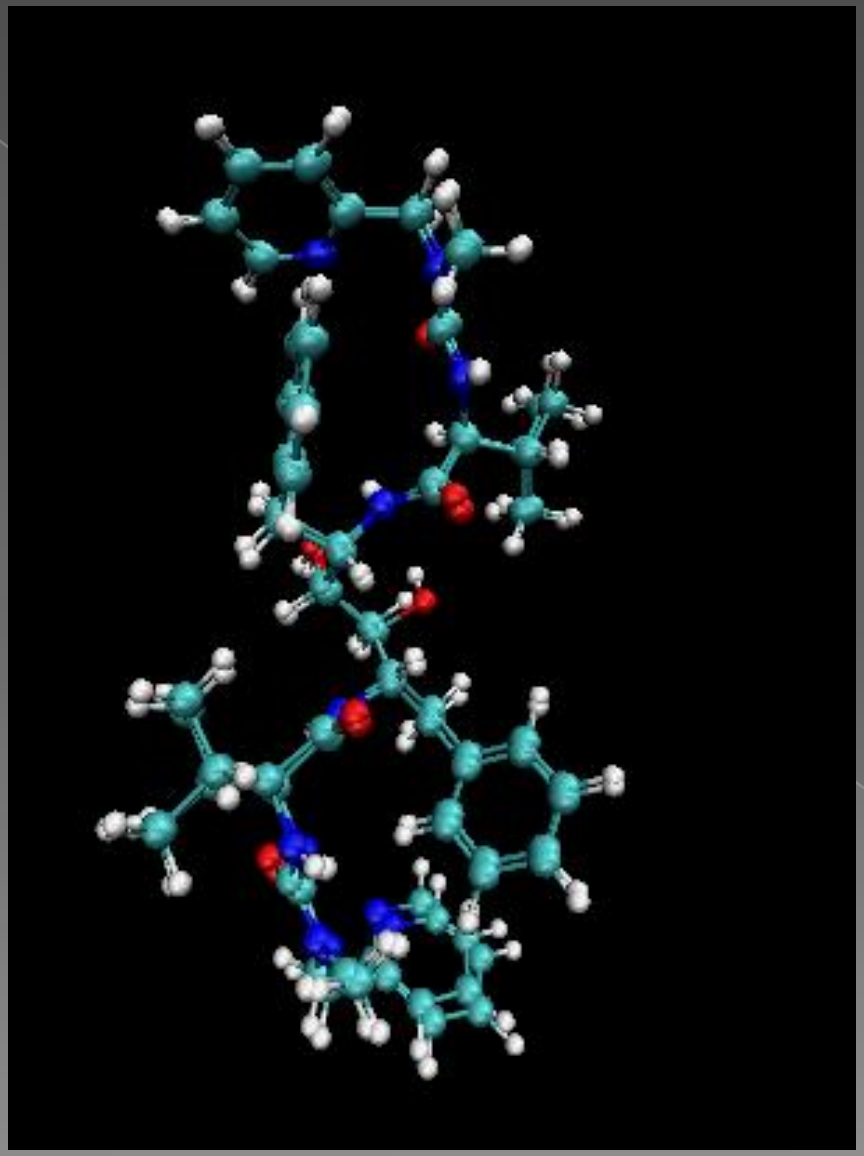
Parameters: Data

- Analysis of 2 complexes: 1hvi, 1hvj
- Each complex with 300 docking results
- Tested k-Mean clustering with k=7
- Data obtained from volunteers across the world through the Docking@Home project

Parameters: HIV protein



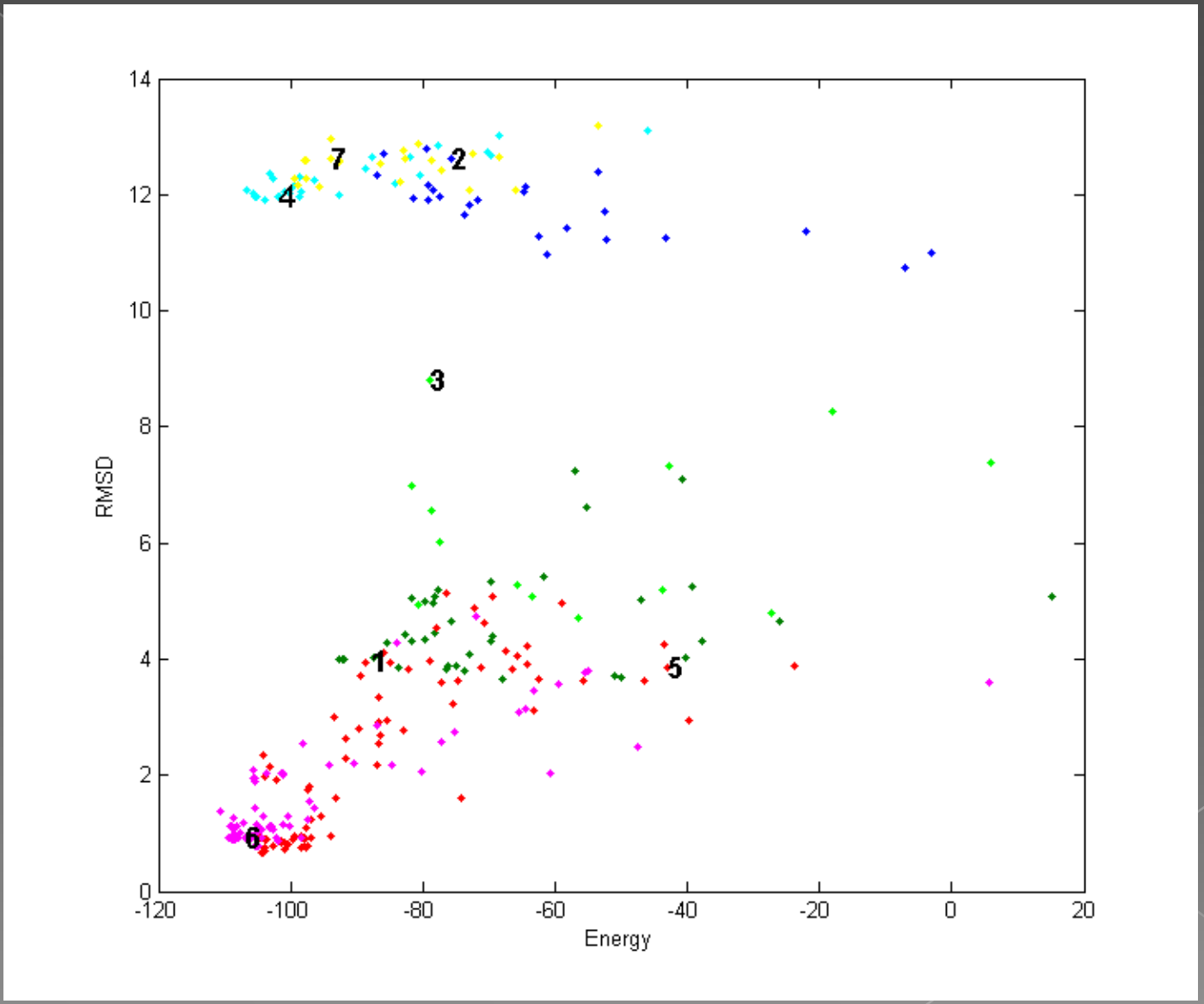
Parameters: Ligands



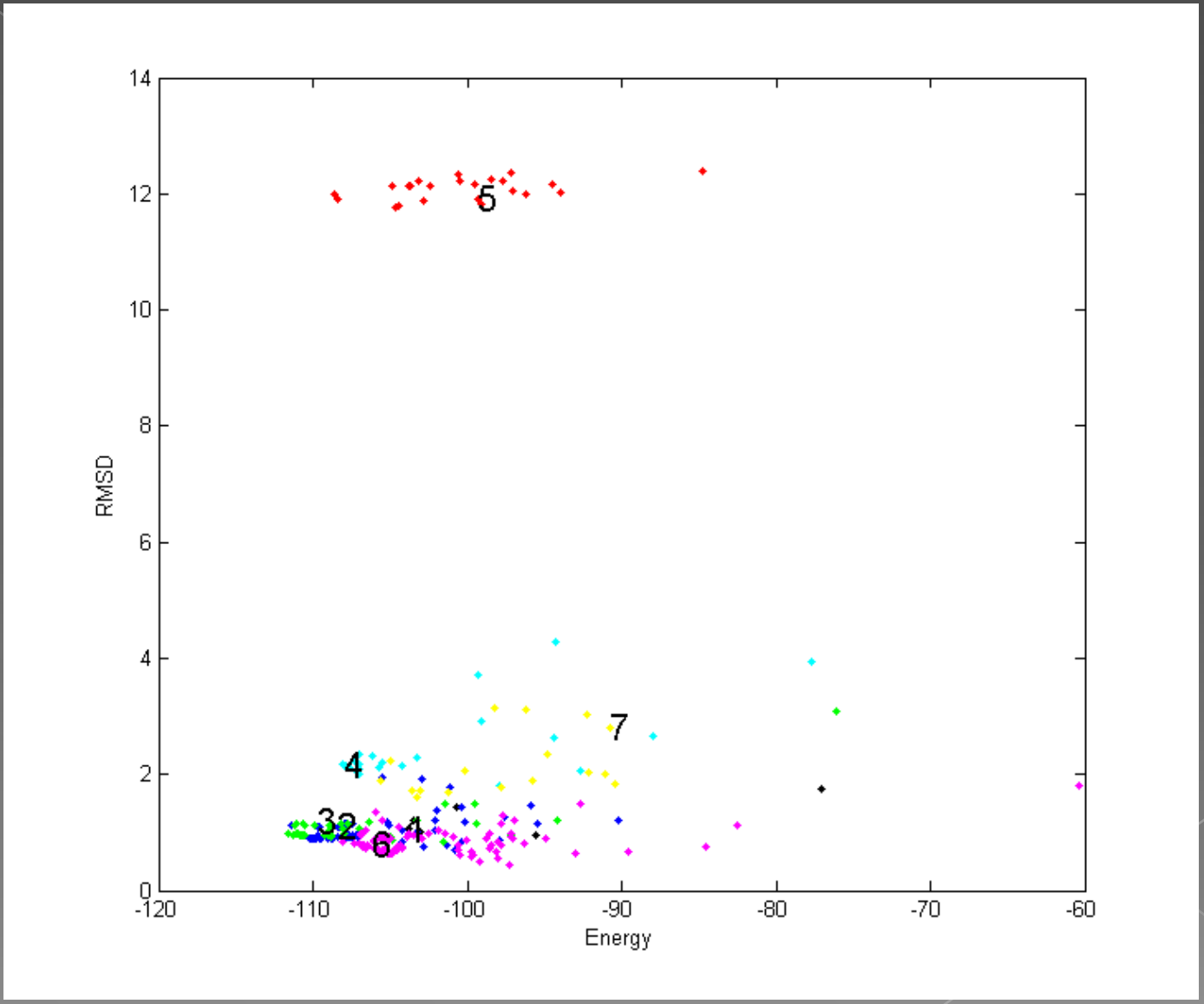
OUTLINE

- Overview
- Purpose
- Program
  - What it does
  - Implementation details
- Limitations
- Parameters
  - Data
  - HIV protein
  - Ligands
- **Results**
  - Complex 1hvi
  - Complex 1hvj
- Future work
- Acknowledgments

Results: Complex 1hvi



Results: Complex 1hvj



OUTLINE

- Overview
- Purpose
- Program
  - Clustering algorithm
  - Implementation overview
- Limitations
- Parameters
  - Data
  - HIV protein
  - Ligands
- Results
  - Complex 1hvi
  - Complex 1hvj
- **Future work**
- Acknowledgments

Future Work

- Compare accuracies when using different:
  - K values in the case of the K-Mean algorithm
  - Types of distances e.g. maximum value distance
  - Methods of computing the centroids in the case of the K-Mean algorithm
  - Clustering methods
- Take into account force fields and energies when clustering as opposed to only complexes

Future Work

- Improve efficiency by calculating only the distances needed at the time of clustering instead of all distances between every data point
- End goal: Method to automatically select ligand(s) which minimize(s) protein-ligand complex energy

OUTLINE

- Overview
- Purpose
- Program
  - Clustering algorithm
  - Implementation overview
- Limitations
- Parameters
  - Data
  - HIV protein
  - Ligands
- Results
  - Complex 1hvi
  - Complex 1hvj
- Future work
- **Acknowledgments**

Acknowledgments

- Dr. Michela Taufer
- Trilce Estrada
- CRA-W Distributed Mentorship program (CRA-W DMP)
- UDel
- Grant Funding:
  - NSF OCI #0506429, DAPLDS - a Dynamically Adaptive Protein-Ligand Docking System based on Multi-Scale Modeling