Automating the Effective Fragment Potential Method

PURDUE UNIVERSITY

Rose-Hulman Institute of Technology
Undergraduate Mathematics Conference
April 21-22, 2017

Peyton Puckett and Lyudmila Slipchenko
Department of Chemistry
Department of Statistics
Puckett.peyton@gmail.com

BACKGROUND

- Computer Science
  - Software Development
  - Machine Learning
  - Hardware Prototyping
  - Robotics
  - Automation & IOT
  - Biotechnology
- Joined research project in August 2016
- Statistics Living-Learning Community
  - 20 Sophomores work on various research projects
  - Take statistics courses
  - Data Science Skills
    - "Big Data"
  - Professor Slipchenko's research group

INTRODUCTION

- Dr. Mark Ward (STAT LLC)
- Professor Lyudmila Slipchenko (Mentor)
  - Pradeep Gurunathan (Graduate Student)
  - Yongbin Kim (Graduate Student)

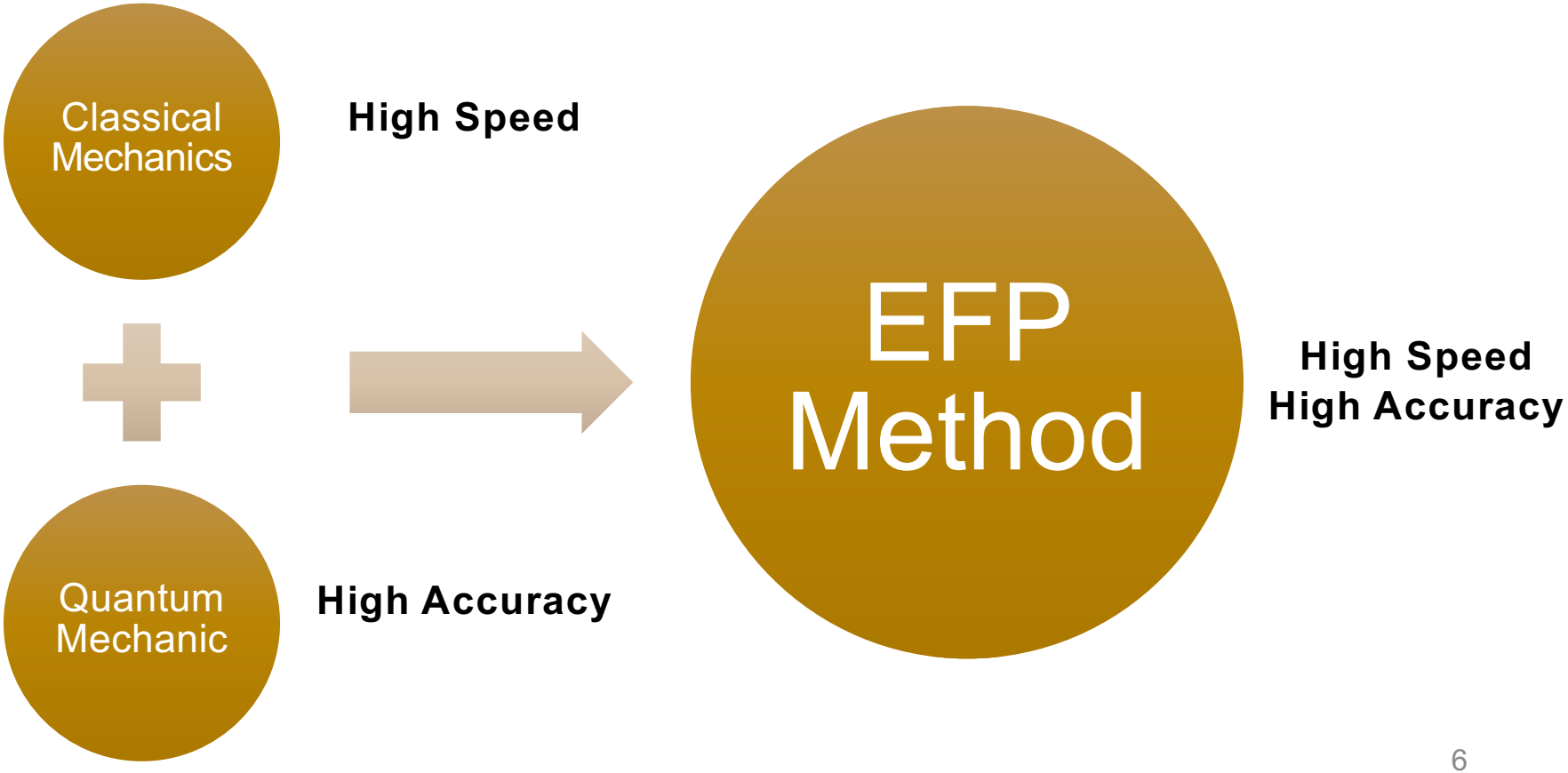
COMPUTATIONAL CHEMISTRY

- What is computational chemistry?
  - Solving chemical problems using computer simulations
  - A perspective that delves at atomistic/sub-atomistic levels
- Why use computational chemistry?
  - Understand chemical phenomena that are too complex to be described using experiments
  - Comprehend why we get certain experimental results
- Methods used to describe systems
  - Classical Mechanics: Newtonian Equations (Fast but less accurate)
  - Quantum Mechanics: Schrodinger Equation (Accurate but not as fast)
  - **EFP = Classical Mechanics + Quantum Mechanics (fast and accurate)**

SOFTWARE USED

- Packages
  - Q-Chem
  - GAMESS
  - LibEFP
- Scripting
  - Bash
  - R
  - Python
- Visualizations
  - R Plotly
  - Jmol
- Machine Learning (Future)
  - Scikit-Learn

EFFECTIVE FRAGMENT POTENTIAL METHOD

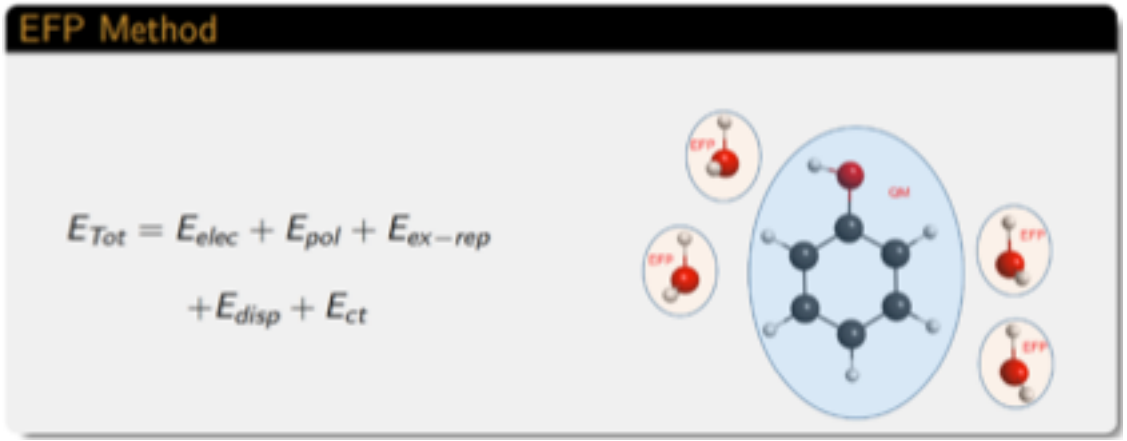


EFFECTIVE FRAGMENT POTENTIAL METHOD

- Effective Fragment Potential Method (EFP)
  - Provides an accurate description of intermolecular interactions
  - Interactions between main and surrounding systems
- Compute parameters based on chemical fragments
- Applications
  - Biology / Chemistry
  - Biomedicine
  - Material and Drug Design
  - Photovoltaics

BENEFITS TO USING EFP

- Collects data about a chemical system using QM/EFP interactions
  - Parameterization is done by using Quantum Mechanics (QM)
- Total Energy can be decomposed
  - EFP Focuses on Polarization Energy



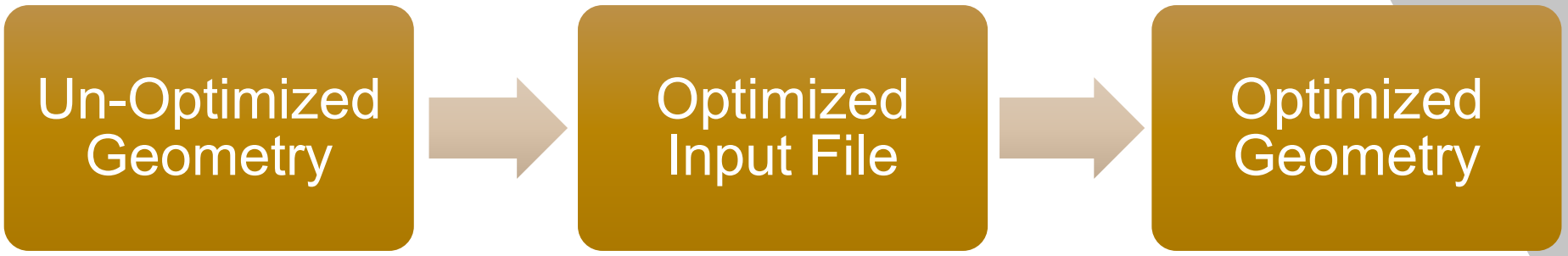
OVERALL GOAL

- Develop a simple yet fast method for EFP calculations
- Determine whether it is possible to use the same parameters in different geometries
  - Is it necessary to compute parameters multiple times?
  - What distribution of structures can a molecule exhibit?
- Analyze structures of amino acids in various proteins

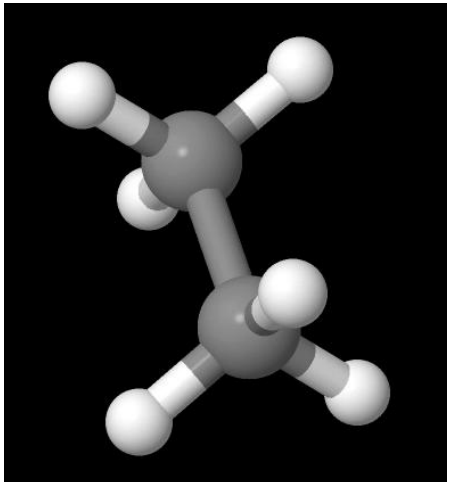
SSI PROTEIN DATABASE

- Contains pairs of interacting simple molecules (amino acid residues) extracted from protein data bank
- Used as a basis for testing purposes
- Available reference data for this database that can be used to compare with our calculations
- 6768 Data Points (Files)
- Each file contains XYZ Coordinates for a pair of molecules

COMPUTE OPTIMIZED GEOMETRIES



Q-Chem Software Package

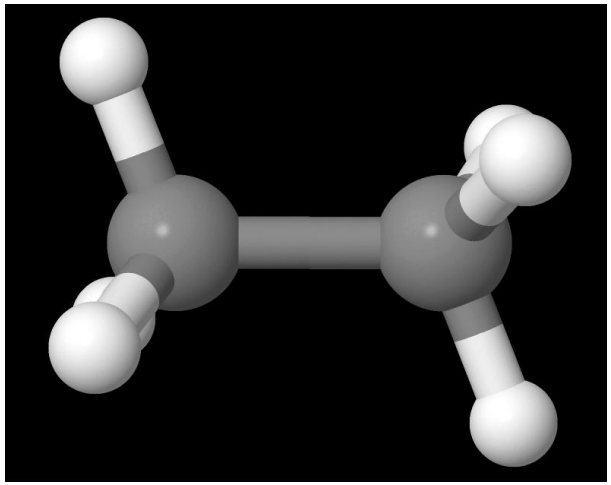


UN-OPTIMIZED

peytonpuckett — ssh ppucket@carter.rcac.

~ — ssh ppucket@carter.rcac.purdue.edu

```
8
0 1 set SSI index 008ALA-034ALA-1 Residue 008ALA and 034ALA interaction No. 1
C        19.07900000     37.47900000     51.22800000
C        19.51100000     36.22400000     51.99000000
H        19.90800000     38.18000000     51.14200000
H        18.73700000     37.20100000     50.23200000
H        18.25600000     37.96400000     51.74800000
H        18.69000000     35.51500000     52.04000000
H        19.82548000     36.48481000     53.01131000
H        20.33665000     35.71992000     51.46636000
```



OPTIMIZED

peytonpuckett — ssh ppucket@carter.rcac

~ — ssh ppucket@carter.rcac.purdue.edu

```
8
0 1 set SSI index 008ALA-034ALA-1 Residue 008ALA and 034ALA interaction No. 1
C        0.7629092728     -0.0039387919     -0.0015316133
C        -0.7663370977     -0.0014680009     -0.0011165315
H        1.1678538452     0.4478972271     0.9194697687
H        1.1648513650     -1.0285583622     -0.0703699074
H        1.1673785265     0.5670024494     -0.8540021771
H        -1.1712736014     -0.4544493277     -0.9215359511
H        -1.1682175676     1.0232751751     0.0664076709
H        -1.1707526886     -0.5712281319     0.8521617345
```

ROOT-MEAN-SQUARE DEVIATION (RMSD)

- RMSD measures similarity between atomic coordinates
- Based on Kabsch Algorithm
- Implemented in Python
- Input xyz coordinates for two molecules
- Original coordinates (A)
- Optimized coordinates (A')
- Using RMSD value, we can make inferences about the protein

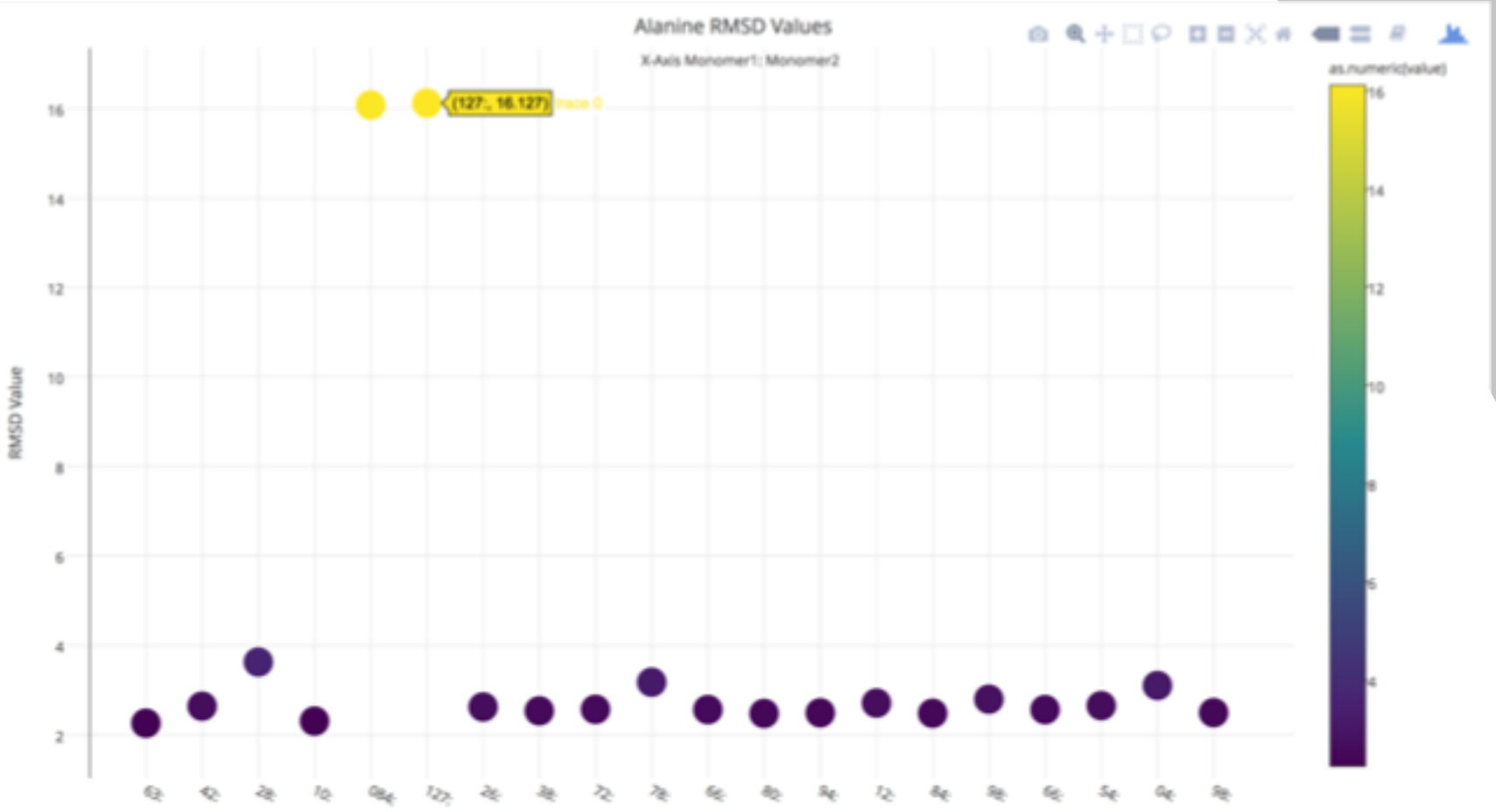
3958_02waterdimer.xyz

```
6

O   -0.066999140   0.000000000   1.494354740
H    0.815734270   0.000000000   1.865866390
H    0.068855100   0.000000000   0.539142770
O    0.062547750   0.000000000  -1.422632080
H   -0.406965400  -0.760178410  -1.771744500
H   -0.406965400   0.760178410  -1.771744500
```

XYZ FILE FORMAT WATER DIMER

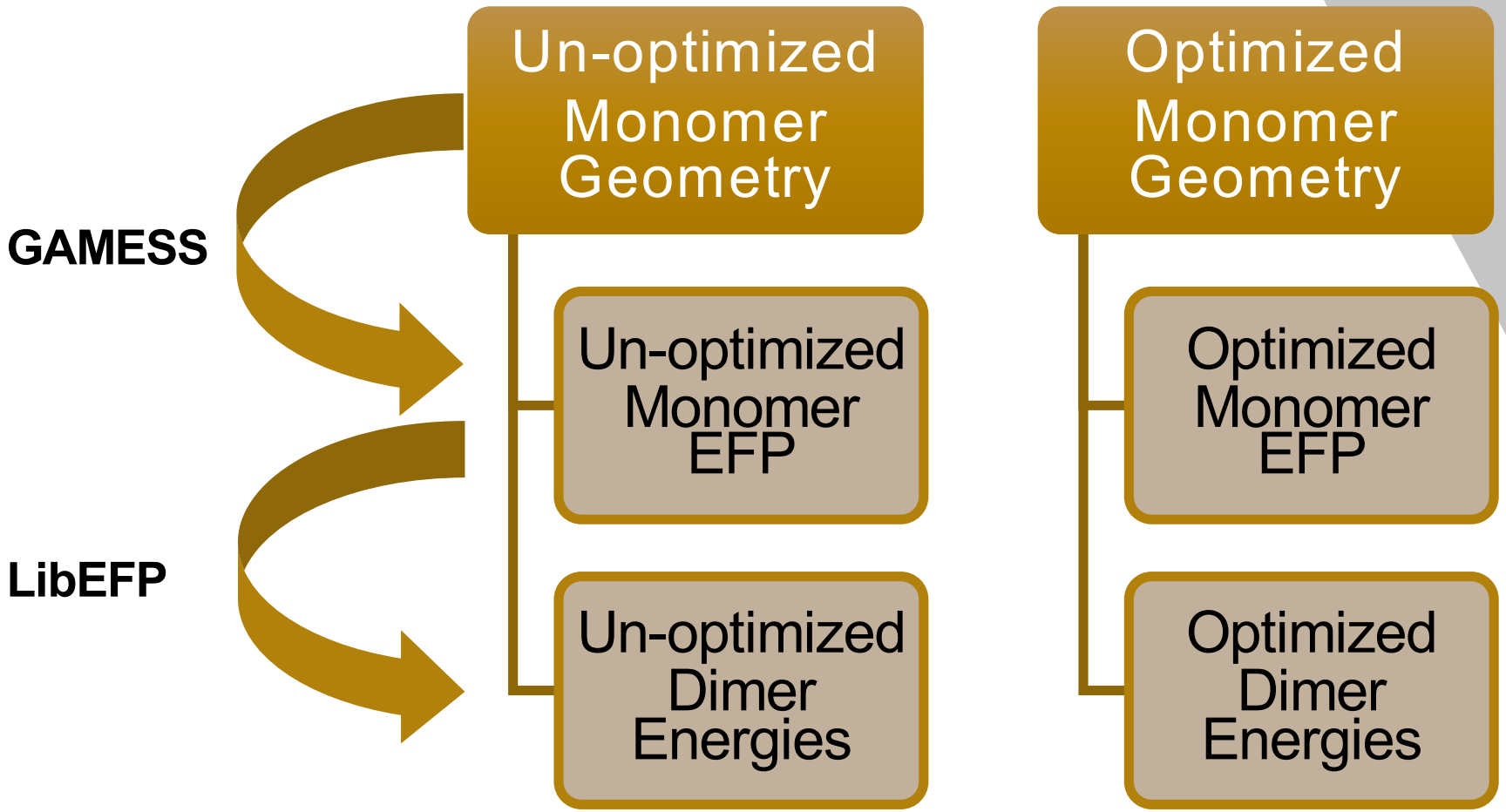
RMSD VALUES: ALANINE RESIDUES



DATA EXTRACTION

- File Manipulation & Parsing (*filename.*xyz.inp.out)
- Each file contains more information regarding the calculations
  - Here we are just looking for the new coordinates
- Extract XYZ coordinates that correspond to newly generated optimized geometry of the initial molecule
- Extracted data is formatted into a new .xyz file

COMPUTE ENERGY FROM PARAMETERS



TOTAL ENERGY CALCULATIONS

- Can be used to describe most stable form of a system
- Explained graphically using R Plotly and other software packages
- Collaborating with other group members to test a new method to calculate these energies
  - Other method uses less resources yet produces similar calculations with minimal error
  - If this is successful, it is potentially more efficient for larger systems

IN THE NEAR FUTURE

- Machine Learning Implementation
  - Predict parameters if Geometry changes
  - Compute fewer parameters with higher accuracy
  - Less computational time and resources
- Apply to larger molecules
  - Improve script efficiency
  - Implement other databases
- Streamline entire process
  - Be able to provide estimated computation time based on size of inputs
  - Combine data into readable format

RESOURCES AND ACKNOWLEDGEMENTS

- National Science Foundation
- Purdue Department of Chemistry
- Purdue Department of Statistics
- Purdue ITAP Research Computing
- Statistics Living-Learning Community