



THE STATE UNIVERSITY
OF NEW JERSEY



NSF REU Progress Report

By Franklin Bettencourt

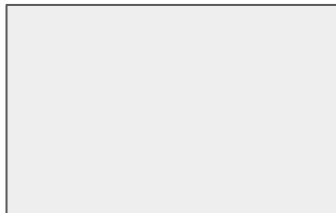


- The Problem
 - DE Shaw BPTI
- Radical Cybertools
 - EnTK
 - CoCo
- Workflow
 - ExTASY
- PCA Space
 - Data in PCA space
- Kernel Constants
 - Study
- Dimensions for PCA
 - 3D vs 4D
- Starting Conformation
 - Right vs Left
- Long Simulation
 - Barriers to cont'd exploration
- Future work
 - Github
 - Time scales
 - Re weightings
 - Atom selections

David E. Shaw et al. "Millisecond-Scale Molecular Dynamics Simulations on Anton," *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis (SC09)*, New York, NY: ACM, 2009

- 1 millisecond of protein dynamics (regular sim but long time)
- found 5 distinct "long lived" conformations of BPTI
- Agrees well with experimental data

Problem - note practical for most institutes to study a protein in this fashion. It would take far too much compute time and would have too low of a throughput for most discovery purposes.



Goal - attempt to replicate the findings of the D.E. Shaw et al. study in a much smaller fraction of compute time using *enhanced sampling methods*.



The Problem

- DE Shaw BPTI

Radical Cybertools

- EnTK
- CoCo

Workflow

- ExTASY

PCA Space

- Data in PCA space

Kernel Constants

- Study

Dimensions for PCA

- 3D vs 4D

Starting Conformation

- Right vs Left

Long Simulation

- Barriers to cont'd exploration

Future work

- Github
- Time scales
- Re weightings
- Atom selections

Radical CyberTools - <https://radical-cybertools.github.io/>

- Radical-SAGA

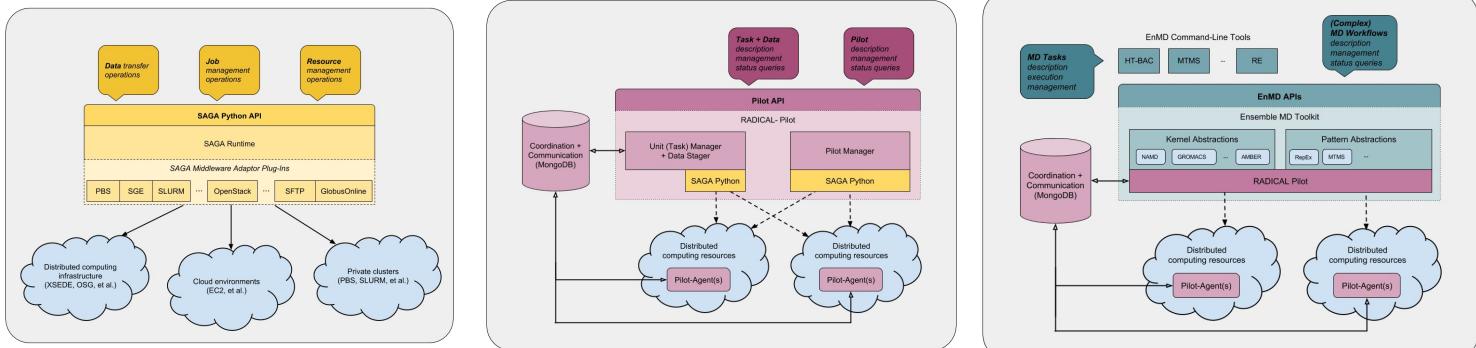
- Provides standard interface to variety of HPC clusters

- Radical-Pilot

- Uses pilots to abstract complexities of scheduling large sets interdependent jobs

- Ensemble Toolkit (EnTK)

- Enables the precise coordinating the execution of ensembles of jobs





- The Problem
 - DE Shaw BPTI
 - Radical Cybertools
 - EnTK
 - CoCo
- Workflow
 - ExTASY
- PCA Space
 - Data in PCA space
- Kernel Constants
 - Study
- Dimensions for PCA
 - 3D vs 4D
- Starting Conformation
 - Right vs Left
- Long Simulation
 - Barriers to cont'd exploration
- Future work
 - Github
 - Time scales
 - Re weightings
 - Atom selections

Ensemble Toolkit (EnTK)

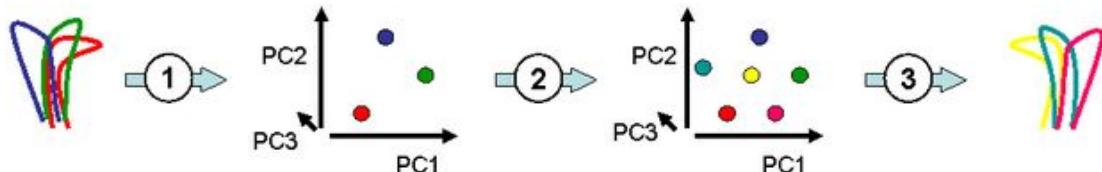
Ensemble Toolkit: Scalable and Flexible Execution of Ensembles of Tasks

V. Balasubramanian, A. Treikalis, O. Weidner, S. Jha, 2016 45th International Conference on Parallel Processing (ICPP), vol. 00, Aug. 2016

CoCo

[Laughton C.A., Orozco M. and Vranken W., COCO: A simple tool to enrich the representation of conformational variability in NMR structures, PROTEINS, 75, 206-216 \(2009\)](#)

- Uses principal component analysis (PCA) to find the conformational changes that led to largest structural changes of the protein.
- Then generate “likely” new structures from PCA results of ensembles of old structures



Step 1: The input ensemble is subjected to Principal Component Analysis and the position of each structure in a low dimensional subspace identified.

Step 2: New points are placed to fill gaps in the distribution.

Step 3: The new points are converted back into new structures, which are returned to the user.



ExTASY: Scalable and flexible coupling of MD simulations and advanced sampling techniques

V. Balasubramanian, I. Bethune, A. Shkurti, E. Breitmoser, E. Hruska, C. Clementi, C. Laughton, S. Jha

<https://bitbucket.org/extasy-project/extasy-workflows>

- ***Extensible Toolkit for Advanced Sampling and analYsis (ExTASY)***

ExTASY Iteration (K stands for kernel)

Sim (Gromacs)

K1 / K2 - Energy Min

Minimize energy of bpti to coco output

K3 / K4 - Equilibrium

Equilibrate k1/k2 output

K5 / K6 - Production MD

Use k3/k4 output for MD sim

K7 / K8 - Post Process k5/k6 output trajs

Analysis (CoCo)

PCA analysis and Gen new conformations

- The Problem
 - DE Shaw BPTI
- Radical Cybertools
 - EnTK
 - CoCo
- Workflow
 - ExTASY
- PCA Space
 - Data in PCA space
- Kernel Constants
 - Study
- Dimensions for PCA
 - 3D vs 4D
- Starting Conformation
 - Right vs Left
- Long Simulation
 - Barriers to cont'd exploration
- Future work
 - Github
 - Time scales
 - Re weightings
 - Atom selections

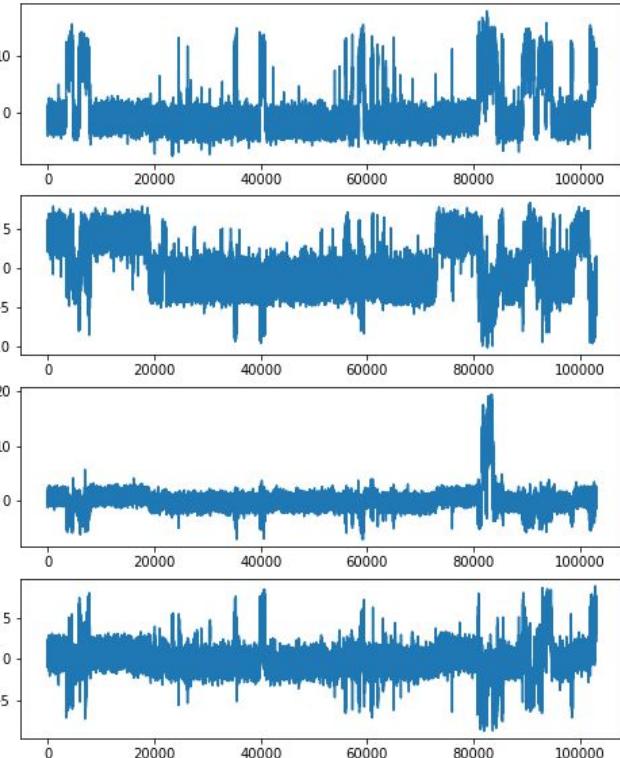
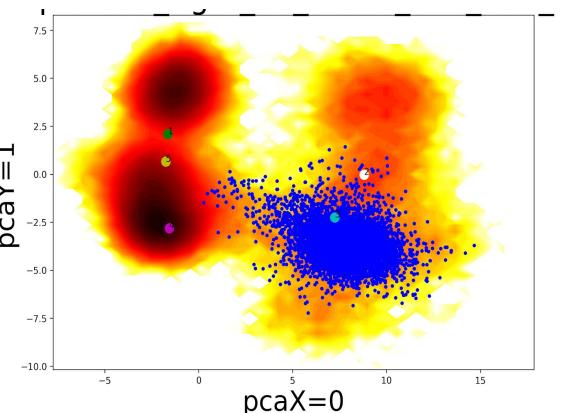
DE Shaw Data in PCA Space

PCA

- Calculates the four conformational changes which lead to the most structural variance
- See plots on right side

Mapping to DE Shaw Space

- One ExTASY simulation (blue) mapped onto a 2D PCA DE Shaw Space.
- Five distinct DE Shaw conform. shown as colored points



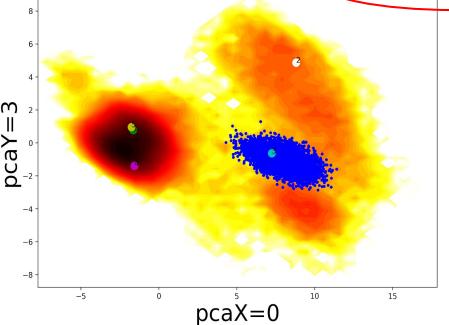
The Problem
 - DE Shaw BPTI
 Radical Cybertools
 - EnTK
 - CoCo
 Workflow
 - ExTASY
 PCA Space
 - Data in PCA space
 Kernel Constants
 - Study
 Dimensions for PCA
 - 3D vs 4D
 Starting Conformation
 - Right vs Left
 Long Simulation
 - Barriers to cont'd exploration
 Future work
 - Github
 - Time scales
 - Re weightings
 - Atom selections



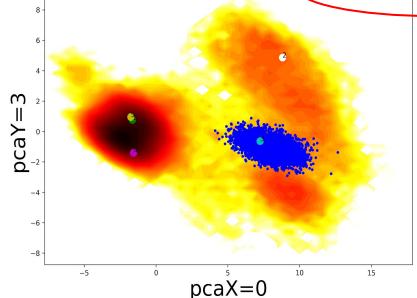
ExTASY K12 and K34 Value Selection

- The Problem
 - DE Shaw BPTI
- Radical Cybertools
 - EnTK
 - CoCo
- Workflow
 - ExTASY
- PCA Space
 - Data in PCA space
- Kernel Constants
 - Study
- Dimensions for PCA
 - 3D vs 4D
- Starting Conformation
 - Right vs Left
- Long Simulation
 - Barriers to cont'd exploration
- Future work
 - Github
 - Time scales
 - Re weightings
 - Atom selections

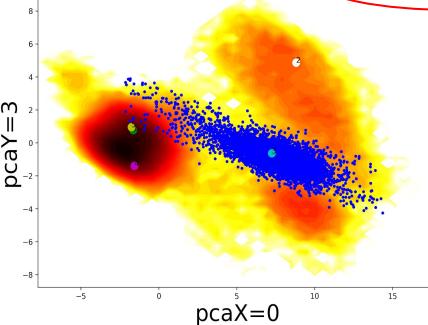
ExTASY pcax=0, pcaY=3, iters=29
 Agg time=69600 (ps), frac DE time=6.96e-05
~~p12b03_right_d4_itrs48_k12_100_k34_100~~



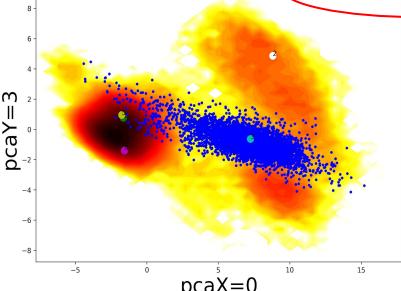
ExTASY pcax=0, pcaY=3, iters=29
 Agg time=69600 (ps), frac DE time=6.96e-05
~~p12b03_right_d4_itrs48_k12_1000_k34_100~~



ExTASY pcax=0, pcaY=3, iters=29
 Agg time=69600 (ps), frac DE time=6.96e-05
~~p12b03_right_d4_itrs48_k12_100_k34_1000~~



ExTASY pcax=0, pcaY=3, iters=29
 Agg time=69600 (ps), frac DE time=6.96e-05
~~p12b03_right_d4_itrs48_k12_1000_k34_1000~~



ExTASY 3D vs 4D

The Problem

- DE Shaw BPTI

Radical Cybertools

- EnTK
- CoCo

Workflow

- ExTASY

PCA Space

- Data in PCA space

Kernel Constants

- Study

Dimensions for PCA

- 3D vs 4D

Starting Conformation

- Right vs Left

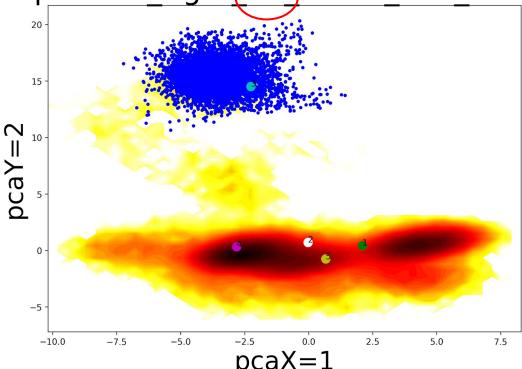
Long Simulation

- Barriers to cont'd exploration

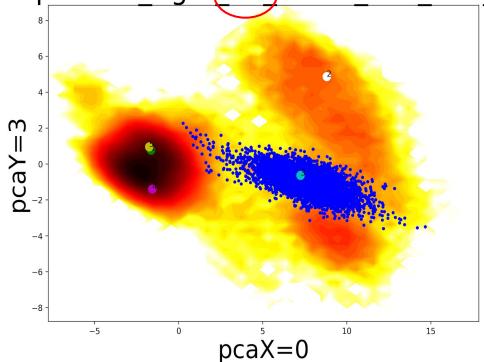
Future work

- Github
- Time scales
- Re weightings
- Atom selections

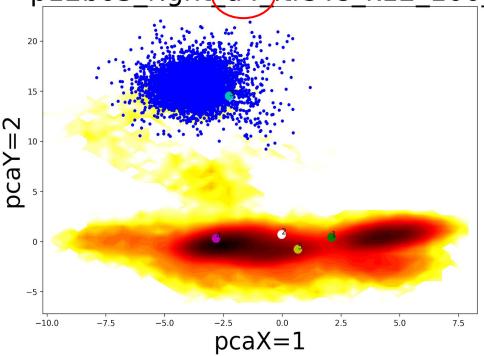
ExTASY pcax=1, pcaY=2, iters=29
Agg time=69600 (ps), frac DE time=6.96e-05
p12b03_right_d3_itrs48_k12_100_k34_1000



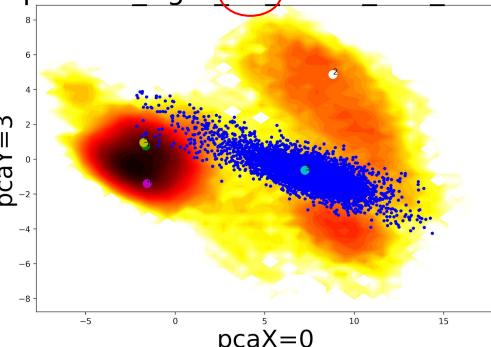
ExTASY pcax=0, pcaY=3, iters=29
Agg time=69600 (ps), frac DE time=6.96e-05
p12b03_right_d3_itrs48_k12_100_k34_1000



ExTASY pcax=1, pcaY=2, iters=29
Agg time=69600 (ps), frac DE time=6.96e-05
p12b03_right_d4_itrs48_k12_100_k34_1000



ExTASY pcax=0, pcaY=3, iters=29
Agg time=69600 (ps), frac DE time=6.96e-05
p12b03_right_d4_itrs48_k12_100_k34_1000



ExTASY 3D vs 4D cont'd

The Problem

- DE Shaw BPTI

Radical Cybertools

- EnTK
- CoCo

Workflow

- ExTASY

PCA Space

- Data in PCA space

Kernel Constants

- Study

Dimensions for PCA

- 3D vs 4D

Starting Conformation

- Right vs Left

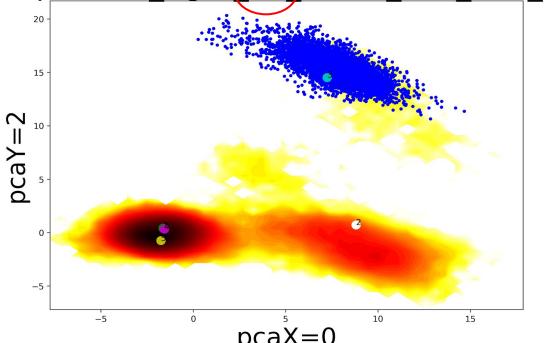
Long Simulation

- Barriers to cont'd exploration

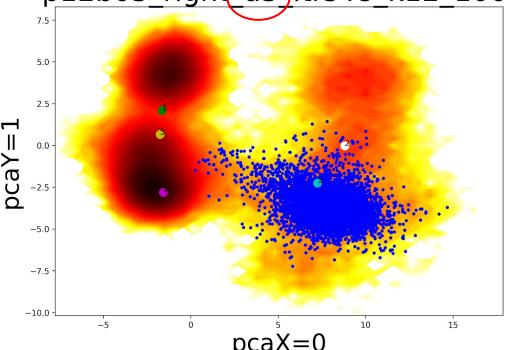
Future work

- Github
- Time scales
- Re weightings
- Atom selections

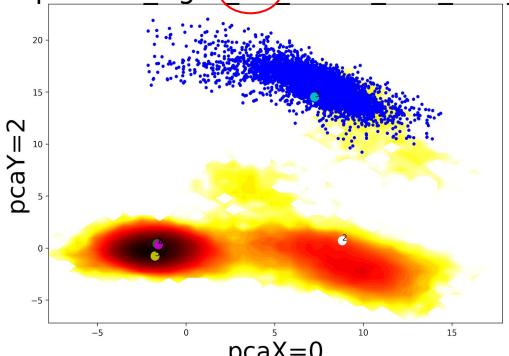
ExTASY pcax=0, pcaY=2, iters=29
 Agg time=69600 (ps), frac DE time=6.96e-05
 p12b03_right d3 itrs48_k12_100_k34_1000



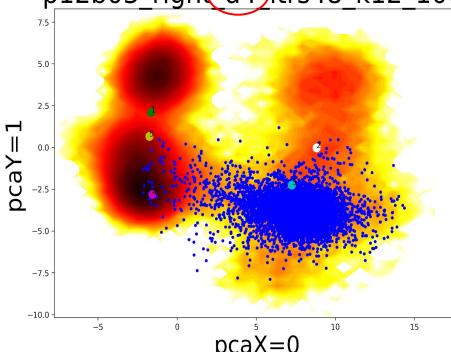
ExTASY pcax=0, pcaY=1, iters=29
 Agg time=69600 (ps), frac DE time=6.96e-05
 p12b03_right d3 itrs48_k12_100_k34_1000



ExTASY pcax=0, pcaY=2, iters=29
 Agg time=69600 (ps), frac DE time=6.96e-05
 p12b03_right d4 itrs48_k12_100_k34_1000



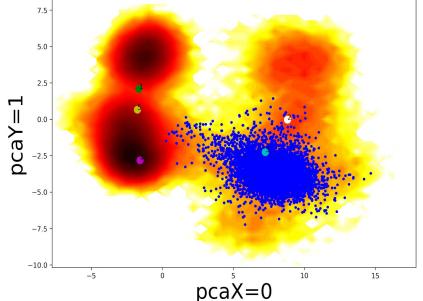
ExTASY pcax=0, pcaY=1, iters=29
 Agg time=69600 (ps), frac DE time=6.96e-05
 p12b03_right d4 itrs48_k12_100_k34_1000



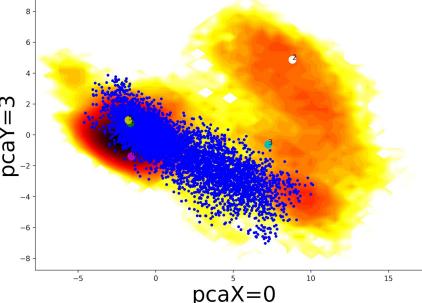
ExTASY Starting Right vs Left

- The Problem
 - DE Shaw BPTI
- Radical Cybertools
 - EnTK
 - CoCo
- Workflow
 - ExTASY
- PCA Space
 - Data in PCA space
- Kernel Constants
 - Study
- Dimensions for PCA
 - 3D vs 4D
- Starting Conformation
 - Right vs Left
- Long Simulation
 - Barriers to cont'd exploration
- Future work
 - Github
 - Time scales
 - Re weightings
 - Atom selections

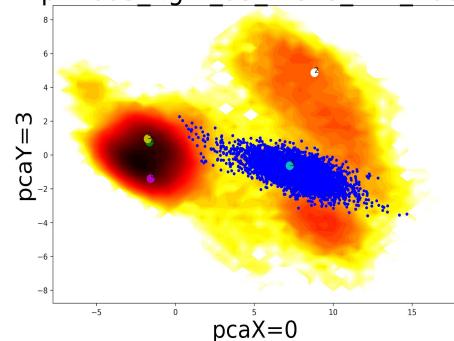
ExTASY pcax=0, pcaY=1, iters=29
 Agg time=69600 (ps), frac DE time=6.96e-05
 p12b03_right d3_itrs48_k12_100_k34_1000



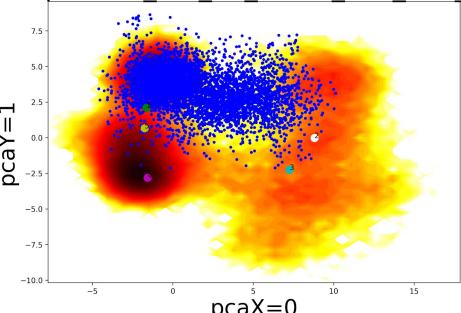
ExTASY pcax=0, pcaY=3, iters=27
 Agg time=64800 (ps), frac DE time=6.48e-05
 p12b01_left d3_itrs48_k12_100_k34_1000



ExTASY pcax=0, pcaY=3, iters=29
 Agg time=69600 (ps), frac DE time=6.96e-05
 p12b03_right_d3_itrs48_k12_100_k34_1000

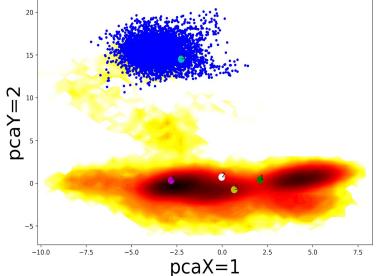


ExTASY pcax=0, pcaY=1, iters=27
 Agg time=64800 (ps), frac DE time=6.48e-05
 p12b01_left_d3_itrs48_k12_100_k34_1000

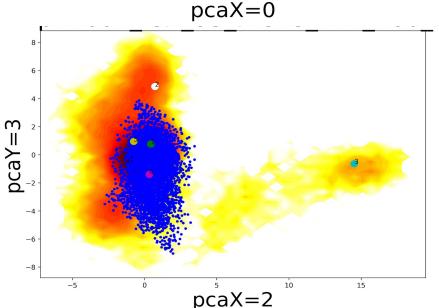
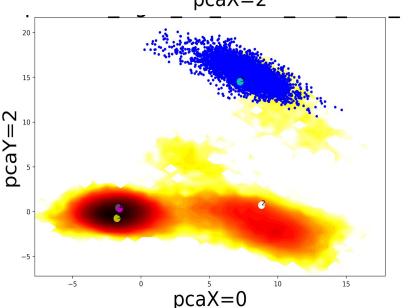
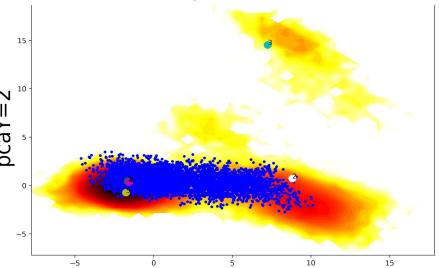
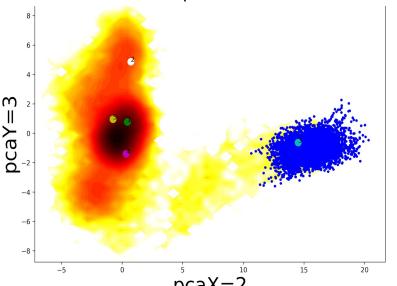
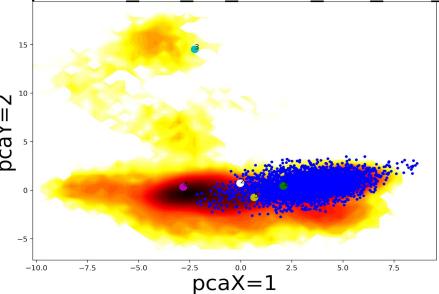


ExTASY Starting Right vs Left

ExTASY pcaX=1, pcaY=2, iters=29
 Agg time=69600 (ps), frac DE time=6.96e-05
 p12b03_right_d3_itrs48_k12_100_k34_1000



ExTASY pcaX=1, pcaY=2, iters=27
 Agg time=64800 (ps), frac DE time=6.48e-05
 p12b01_left_d3_itrs48_k12_100_k34_1000



The Problem

- DE Shaw BPTI

Radical Cybertools

- EnTK
- CoCo

Workflow

- ExTASY

PCA Space

- Data in PCA space

Kernel Constants

- Study

Dimensions for PCA

- 3D vs 4D

Starting Conformation

- Right vs Left

Long Simulation

- Barriers to cont'd exploration

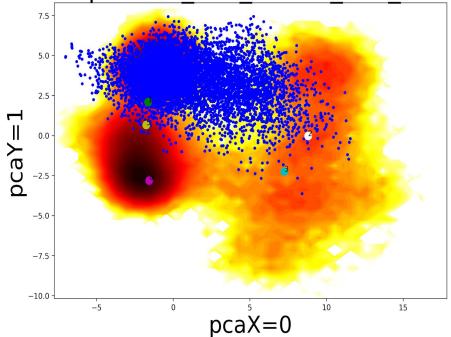
Future work

- Github
- Time scales
- Re weightings
- Atom selections

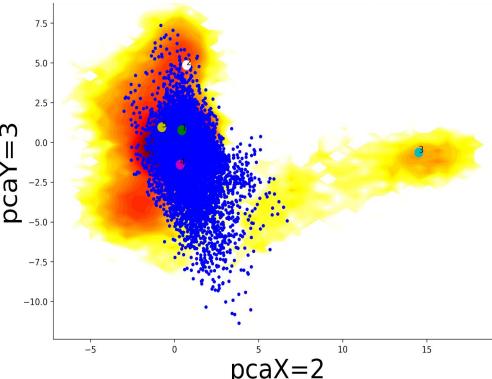
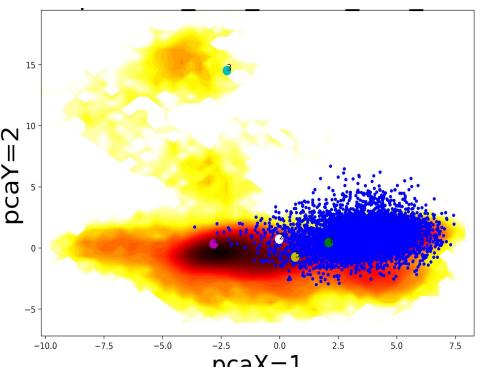
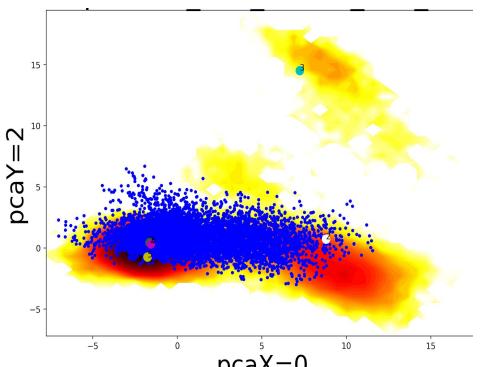
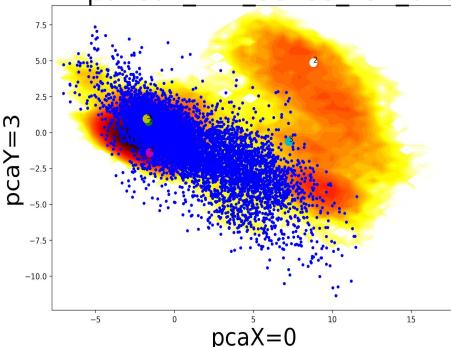
ExTASY Longest Sim So Far...

- The Problem**
 - DE Shaw BPTI
- Radical Cybertools**
 - EnTK
 - CoCo
- Workflow**
 - ExTASY
- PCA Space**
 - Data in PCA space
- Kernel Constants**
 - Study
- Dimensions for PCA**
 - 3D vs 4D
- Starting Conformation**
 - Right vs Left
- Long Simulation**
 - Barriers to cont'd exploration
- Future work**
 - Github
 - Time scales
 - Re weightings
 - Atom selections

ExTASY pcax=0, pcaY=1, iters=37
 Agg time=88800 (ps), frac DE time=8.88e-05
 p07e01_k12_00100_k34_01000_itx2



ExTASY pcax=0, pcaY=3, iters=37
 Agg time=88800 (ps), frac DE time=8.88e-05
 p07e01_k12_00100_k34_01000_itx2





So Far...

- ExTASY shows great potential for more quickly exploring DE Shaw Space
- Github has been setup for next person to pick up where left off and finish project

The Problem

- DE Shaw BPTI

Radical Cybertools

- EnTK
- CoCo

Workflow

- ExTASY

PCA Space

- Data in PCA space

Kernel Constants

- Study

Dimensions for PCA

- 3D vs 4D

Starting Conformation

- Right vs Left

Long Simulation

- Barriers to cont'd exploration

Future work

- Github
- Time scales
- Re weightings
- Atom selections

What Next (parameters to study)

ExTASY simulations seemed to initially explore new space very quickly but then eventually slow down. This could be because the large numbers of trajectories generated “dilute” any new ones so CoCo doesn’t capture newer evolving trends. It could also be that we need to study all heavy atoms instead of just the alpha carbons of the amino acids. There are several parameters we should examine in the future to improve exploration of new space.

- **Time scales** - perhaps fewer trajector files put out over larger time scales will capture more important long term trends
- **Traj frame re-weighting** - adjusting CoCo to emphasize more newly acquired trajectories may allow it to reach into untouched spaces more quickly.
- **Atom Selection** - Setting CoCo to study all heavy atoms or some combination of select heavy atoms may allow CoCo to capture conformational variance

Future Work:

- Integration of COCO & PCA based analysis capabilities into **MDAnalysis**, and ExTASY middleware into **MIDAS**.



The Problem

- DE Shaw BPTI

Radical Cybertools

- EnTK
- CoCo

Workflow

- ExTASY

PCA Space

- Data in PCA space

Kernel Constants

- Study

Dimensions for PCA

- 3D vs 4D

Starting Conformation

- Right vs Left

Long Simulation

- Barriers to cont'd exploration

Future work

- Github
- Time scales
- Re weightings
- Atom selections

Thank You!

Thank you to **everyone on the Radical team!** You were all immensely helpful and patient. I couldn't have come nearly as far without you all.

Thank you **Professor Jha and Professor Laughton** for all of your guidance in pursuing this project and for investing in me.

Thank you to the **NSF 1443054** for the REU which enable me to pursue this research project



Introduction

- Radical Cybertools
- EnTK
- CoCo

Workflow

- ExTASY

Early Phase Dev

- asfd

Phase 1 Results

- asdf

Residue Selection

- 1-55

Phase 2 Results

- Slide 1
- Slide 2

Phase 2 Interpretations

- hmmm

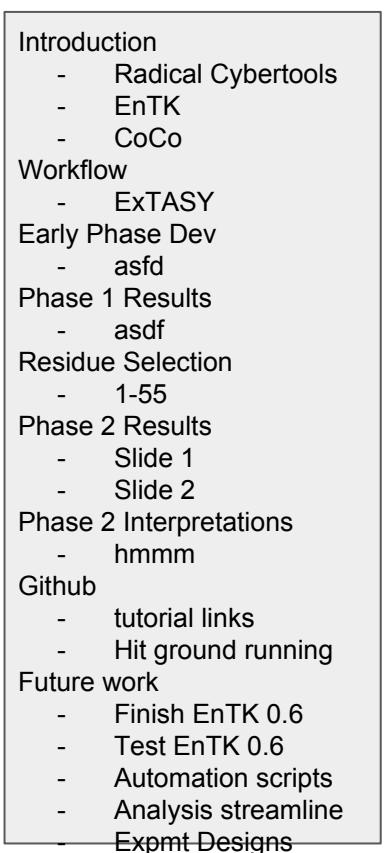
Github

- tutorial links
- Hit ground running

Future work

- Finish EnTK 0.6
- Test EnTK 0.6
- Automation scripts
- Analysis streamline
- Expm Design

SLIDE BACK
UPS START
HERE



- Radical CyberTools - <https://radical-cybertools.github.io/>

- Radical-SAGA

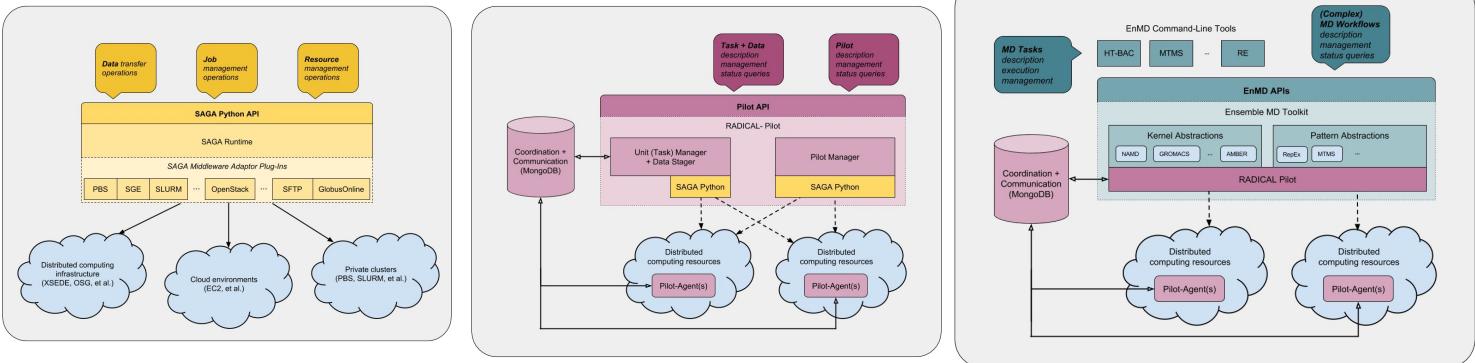
- Provides standard interface to variety of HPC clusters
- Reduce overhead of learning how to use different HPC cluster

- Radical-Pilot

- Uses pilots to abstract complexities of scheduling interdependent jobs
- allows the user greater flexibility in operating HPC clusters

- Ensemble Toolkit (EnTK)

- Enables the scheduling and coordinating of large ensembles of jobs. Allows user to focus on scientific development rather than HPC scripting



Phase 1 Studies



- Introduction
 - Radical Cybertools
 - EnTK
 - CoCo
- Workflow
 - ExTASY
- Early Phase Dev
 - asfd
- Phase 1 Results
 - asdf
- Residue Selection
 - 1-55
- Phase 2 Results
 - Slide 1
 - Slide 2
- Phase 2 Interpretations
 - hmmm
- Github
 - tutorial links
 - Hit ground running
- Future work
 - Finish EnTK 0.6
 - Test EnTK 0.6
 - Automation scripts
 - Analysis streamline
 - Exptm Designs

Select All Alpha Carbons of Protein Residues

- CoCo would analyze the movements of the Alpha carbons in all trajectory files
- New structures generated would be based on the Alpha carbon movements which led to the most variance in conformational change

Varied

- Constant for Energy minimization step
- Constant for Equilibration step

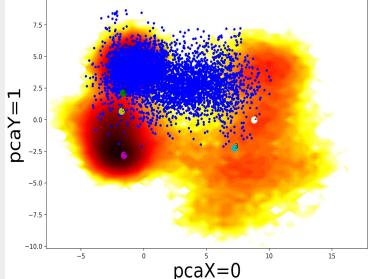
Goal

- Find which combinations explored the largest area of the DE Shaw conformation space

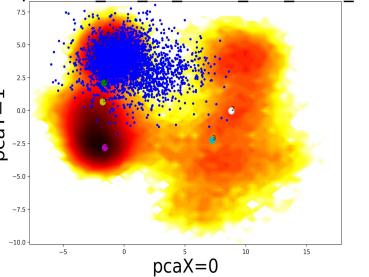


- Residue Selection
- DE Shaw Data
- DE Shaw PCAs
- ExTASY Kernel Constants
- ExTASY 3D vs 4D
- ExTASY Right vs Left
- ExTASY Exploring Non-DE Shaw Space
- ExTASY Longest Sim So Far
- Issue: Data Staging
- Issue: Many Tmux Sessions
- Issue: SuperMIC Utilization
- Other HPC Resources
- Next Steps

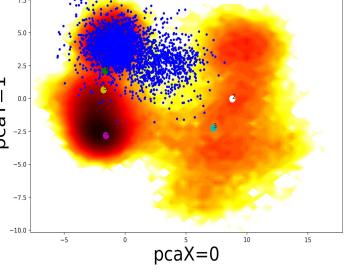
ExTASY pcax=0, pcaY=1, iters=27
Agg time=64800 (ps), frac DE time=6.48e-05
p12b01_left_d3_itrs48_k12_100_k34_1000



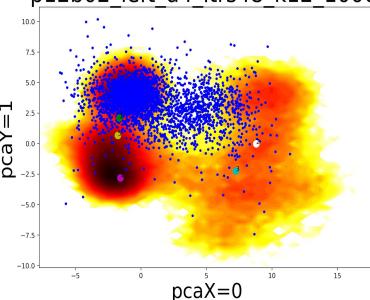
ExTASY pcax=0, pcaY=1, iters=17
Agg time=40800 (ps), frac DE time=4.08e-05
p12b02_left_d3_itrs48_k12_1000_k34_1000



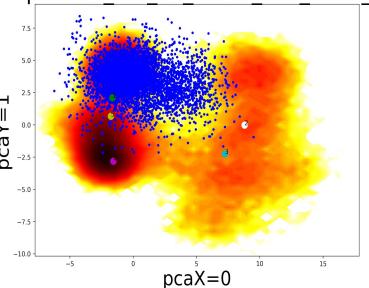
ExTASY pcax=0, pcaY=1, iters=15
Agg time=36000 (ps), frac DE time=3.6e-05
p12b02_left_d4_itrs48_k12_100_k34_1000



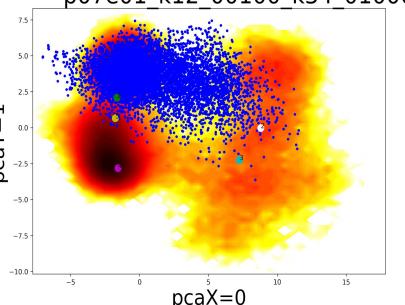
ExTASY pcax=0, pcaY=1, iters=15
Agg time=36000 (ps), frac DE time=3.6e-05
p12b02_left_d4_itrs48_k12_1000_k34_1000



ExTASY pcax=0, pcaY=1, iters=6
Agg time=72000 (ps), frac DE time=7.2e-05
p12b04_left_d4_itrs48_k12_1000_k34_1000



ExTASY pcax=0, pcaY=1, iters=37
Agg time=88800 (ps), frac DE time=8.88e-05
p07e01_k12_00100_k34_01000_itx2





How to proceed ... test ...

- Introduction
 - Radical Cybertools
 - EnTK
 - CoCo
- Workflow
 - ExTASY
- Early Phase Dev
 - asfd
- Phase 1 Results
 - asdf
- Residue Selection
 - 1-55
- Phase 2 Results
 - Slide 1
 - Slide 2
- Phase 2 Interpretations
 - hmmm
- Github
 - tutorial links
 - Hit ground running
- Future work
 - Finish EnTK 0.6
 - Test EnTK 0.6
 - Automation scripts
 - Analysis streamline
 - Exptm Designs

Analysis methods

Review data / see it change vs time ... (charlie script)
Quantitative decision making method?

DOE

Time scales
Traj frames considered
Traj frame re-weighting
Frame outputs on different time scales
Heavy atom vs carbons vs residue heavies and CA ... etc ...
DOE studies to finish ...