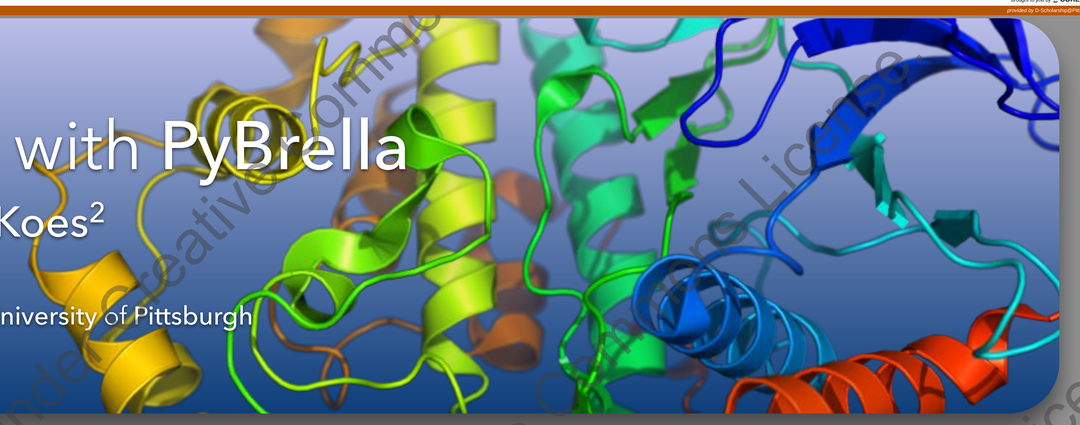
Computing Free Energies with PyBrella

Charles Yuan[1] and David Koes[2]

[1] Mt. Lebanon High School

[2] Department of Systems and Computational Biology, University of Pittsburgh

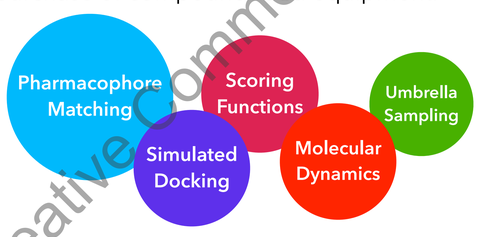


Drug Discovery

Modern **drug discovery** is a tedious process that is often limited by the **expense** and **time** of screening a target.

Computational drug discovery provides an alternative method for **generating hits** to a target, by allowing:

- **rapid screening**, since computational technology and speed is still advancing;
- **enormous numbers of compounds**, by searching huge databases and using combinatorial chemistry for new molecules;
- and **less effort and cost**, by requiring less purchase of compounds and equipment.



Computational analysis usually consists of:

- **spacial and electrostatic docking and matching** of a potential compound;
- brief **scoring and ranking** of a compound's predicted efficacy;
- and **in-depth evaluation** of individual molecules deemed of sufficient quality.

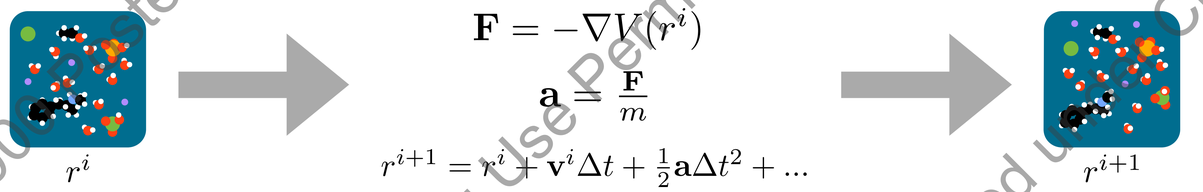
This last step is particularly **resource and time-intensive**, making it often the **limiting factor** in drug screening.

We seek to develop an **effective and economical** method to reliably correlate the actual **rate of dissociation** with a calculated trajectory of system energy, or **potential of mean force (PMF).**

Choosing to avoid simple single-point calculations allows us to consider the **full protein-water system** and increase overall prediction accuracy.

The study uses the dataset **CSAR2012**, which includes numerous **test protein targets** and associated **known active/inactive compounds.**

Steered Molecular Dynamics



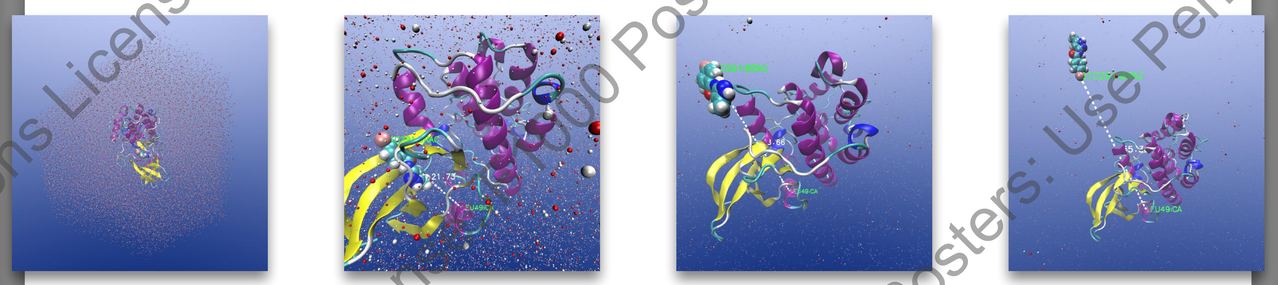
Molecular dynamics (MD) is the **simulation** of molecular **motion** that accounts for **positions** and **trajectories**.

We used the MD package **AMBER12** to simulate the **removal** of molecules from proteins by **steered molecular dynamics (SMD):**

Proteins from the CSAR dataset were prepared with various ligands encapsulated in a periodic **20 Å water box,** then were allowed to **minimize** and **equilibrate** for 1 nanosecond of simulation time.

A **pair of atoms** was selected from the protein and the ligand in preparation for applying a pushing force to ensure a **free exit path** for the ligand.

Constraints were added to the protein **backbone** to disallow distortion, and a **simulated force** was applied on the protein and ligand until they were **20 Å** apart, allowing the ligand to escape into free solution.



Images of the simulation process, from the starting water box to the final exit of the molecule

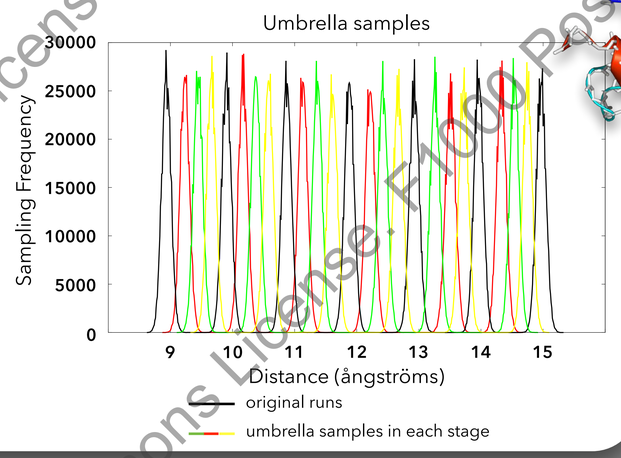
Umbrella Sampling with PyBrella

We used the technique of **umbrella sampling,** which:

- uses **snapshots** throughout the SMD trajectory as starting points for **new simulations;**
- requires **local force restraints,** to weakly hold the compound in place, and **coverage of all distances;**
- and is **slower** if implemented in a naïve fashion.

To implement the method, we developed the program **PyBrella**, which extracts **frames** from SMD output, assigns **constraints**, controls **simulation**, then dynamically adds **new runs** so that all distances receive sampling to a **minimum threshold**, avoiding **extraneous sampling.** Then, it compares the **first 80% to the last 20%** of each run to determine **convergence,** and **extends runs** that have not converged.

Finally, we use the **Weighted Histogram Analysis Method (WHAM)** to recover the PMF of the reaction.



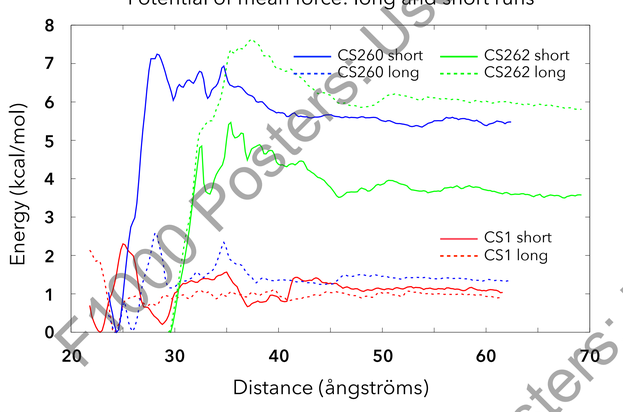
Results and Analysis

Two main parameters have been under investigation:

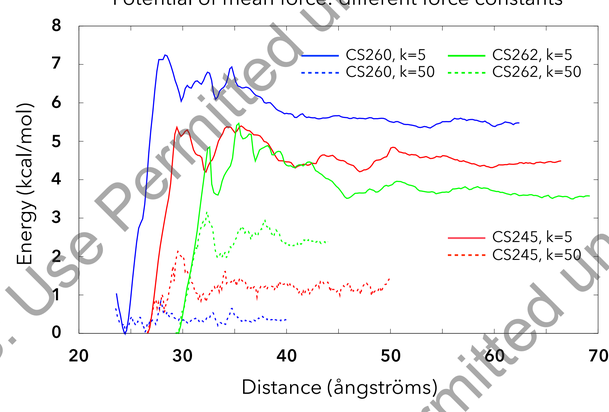
- the usage of **longer runs** (10 ns) or **collections of shorter runs** (1 ns) for every distance position;
- and the choice of **force constant** for the umbrella sampling restraint, from a rigid **k=50** to a loose **k=5.**

Analysis is shown for a limited set of CSAR. Generally, long runs produce **unsatisfactorily flat and uninformative** PMF curves. Shorter runs yielded higher PMFs and more appropriate energy curves. Likewise, the smaller force constant k=5 showed **drastically improved** results over k=50, with much higher PMFs. Since CS245, CS260, and CS262 all have **high actual affinities,** a **higher PMF** is expected.

Potential of mean force: long and short runs



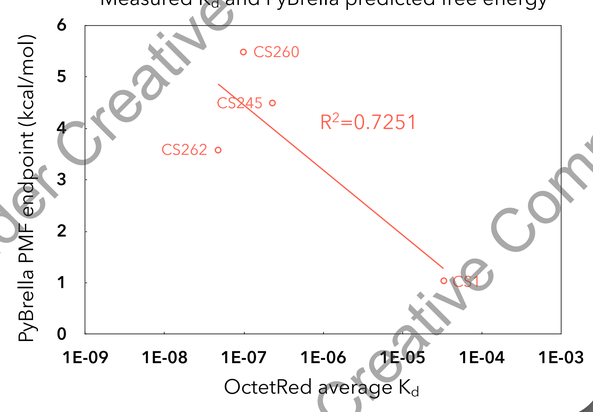
Potential of mean force: different force constants



The **correlation** between the **measured K_d** and the **predicted free energy** is the determinant of success, and tentatively there is an **expected correlation**, using the short runs with k=5.

A **lower dissociation constant,** corresponding to a stronger attraction, does indeed cause a **higher PyBrella PMF** across four compounds.

Measured K_d and PyBrella predicted free energy



Conclusions

So far, the link between actual efficacy and predicted free energy by PyBrella is insufficient since **more data are necessary** to confirm a correlation, though there is an **expected trend.**

There are still large **sources of error,** for example in bias inherent in the **SMD trajectory** or **distortion** of the protein or drug **structure**.

There is a **surprisingly large reliance** on the choice for the umbrella sampling **force constant,** suggesting odd behavior by WHAM since the PMF **should ideally be independent** of the force constant.

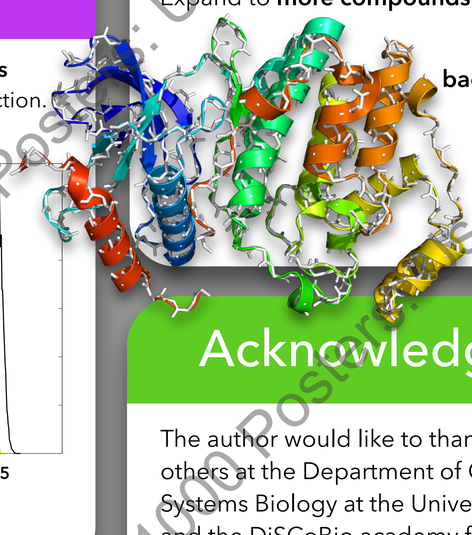
The availability of **more data** in the longer runs is **not necessarily better,** since several compounds received uncharacteristic PMFs, but this determination is still inconclusive.

Next Steps

Investigate use of **other force constants** and begin to develop a **general method** to determine force constants for each compound

Test **even larger collections** of shorter runs instead of longer runs to improve speed and reduce error

Expand to **more compounds** in the CSAR dataset



Apply **constraints on backbone** to ensure that the compound exits instead of bending the protein

Structure: PDB 4G6N

Acknowledgements

The author would like to thank Dr. David Koes and others at the Department of Computational and Systems Biology at the University of Pittsburgh and the DiSCoBio academy for guidance and resources.

This project is supported by National Institutes of Health grant GM108340-01.

References

A. Grossfield. WHAM: the Weighted Histogram Analysis Method, version 2.0.7.

A. W. Goetz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand, R. C. Walker. Routine microsecond molecular dynamics simulations with AMBER - Part I: Generalized Born. *J. Chem. Theory Comput.* **8** (5), 1542-1555 (2012). DOI: 10.1021/ct200909j

D. A. Case, T. A. Darden, T. E. Cheatham et al. AMBER 12, University of California, San Francisco. (2012).

D. R. Koes, C. J. Camacho. ZINCPharmer: pharmacophore search of the ZINC database. *Nucl. Acids Res.* **40** (W1): W409-W414. (2012). DOI: 10.1093/nar/gks378

J. J. Irwin, B. K. Shoichet, M. M. Mysinger et al. Automated docking screens: a feasibility study. *J. Med. Chem.* **52** (18), 5712-5720. (2009).

H. Carlson, J. Dunbar, J. Gestwicki, J. Stuckey, H. Showalter, S. Wang. Community Structure-Activity Resource. (2014).