Enabling Trade-offs Between Accuracy and Computational Cost: Executing Adaptive Algorithms to Reduce Time to Clinical Insight

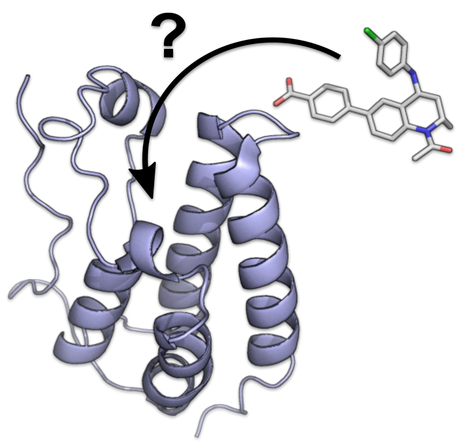
Jumana Dakka *, [1], Kristof Farkas-Pall *, [2], Vivek Balasubramanian [1] , Matteo Turilli [1], Shunzhou Wan [2], David W Wright [2], Stefan Zasada [2], Peter V Coveney [2], Shantenu Jha [1,3]

[1] **Rutgers**, the State University of New Jersey, Piscataway, NJ 08854, USA, [2] **University College London**, London, UK, WC1H 0AJ, * Contributed equally
[3] **Brookhaven National Labs,** Upton, NY 11973, USA

Motivation

The use of simulations to calculate the strength of drug binding has the potential to reduce costs and improve efficiency in drug design, as well as to underpin the development of personalised medicine.

To achieve these results we need to coordinate large simulation campaigns which:
- provide reproducible, accurate and precise binding affinities (free energies, ΔG)
- accompany results with reliable uncertainty quantification
- account for differences in behavior across varied chemical space
- complete within short turn around times



Developing workflows for use in pharmaceutical industry (or medical) scenarios requires that our solutions are robust and flexible as well as highly performant. For example, the trade off desired between the required accuracy and computational cost changes throughout the drug's discovery as the process moves from screening, to hit-to-lead and onto lead optimization. Given a large inventory of drugs, a significant number of compounds must be inexpensively screened to eliminate poor binders, before more accurate methods are needed to discriminate the best binders. This requires that progress of simulations be monitored and decisions about continued execution be predicated on scientific significance at runtime, in order to render the quickest time-to-solution. Therefore, the dependencies of the workflow are not known *a priori* and instead are learnt during execution i.e. adaptive workflows.

Producing timely clinical insight demands computational efficiency which is predicated upon advances in algorithms, scalable software systems and intelligent and efficient utilization of supercomputing resources.

We demonstrate the use of **HTBAC** (High-Throughput Binding Affinity Calculator) designed to support the requirements of computational efficiency of scalable adaptive workflows on high performance computers to produce results on reduced timescales without a loss in accuracy. We show how HTBAC leverages the real world problem of providing insight from drug candidate data using results from a collaborative project between UCL and GlaxoSmithKline to study candidates binding to the BRD4 protein.

Computational Challenges

We define the term **adaptivity** as the dependency of the future state of an entity that is contingent upon the results of its current or previous state(s). **Adaptive workflows** thereby capture workflows with partially defined dependencies. **Adaptive execution** is the capability to make runtime decisions that are based on the current or previous state of the workflow and resources.

Supporting adaptive workflows at scale require advanced computational capabilities which pose the following challenges:

- Developing simple and usable software systems to express adaptive workflows
- Efficient execution of the adaptive workflow
- Scalable execution of heterogeneous tasks

Methods

We implement two ensemble-based binding free energy protocols in HTBAC: TIES and ESMACS. We define an **ensemble** as a set of simulations in which each simulation *only* varies by a parameter. This is a broad definition that encompasses replicas, lambda windows (for alchemical protocols) or multi system simulations.

- **ESMACS** (endpoint protocol) is a computationally cheaper, but less rigorous method, it is used to directly compute the binding strength of a drug to the target protein from MD simulations (as opposed to differences in affinity).
- **TIES (**alchemical protocol) employs enhanced sampling at each lambda window to yield reproducible, accurate and precise relative binding affinities.

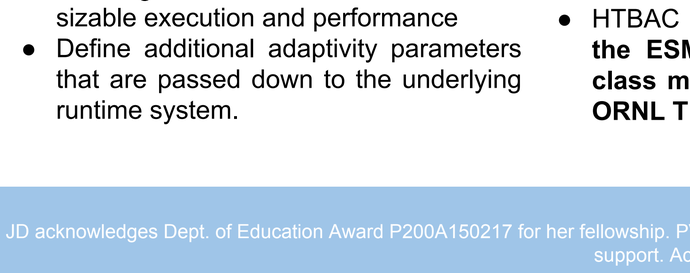


Figure 1: Schematic of the TIES protocol. The thermodynamic state of the system is changed coupled to the variable λ. The value of 0 represents **drug A**, while 1 represents **drug B** inside a **protein**. Also note, that multiple replicas of the same simulation are run, to correctly capture the error in the calculated observable.

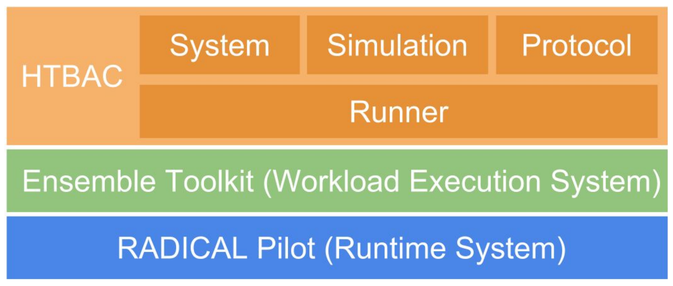
HTBAC

HTBAC is a Python library that enables:

- Unprecedented throughput by allowing concurrent screening of drug binding affinities from multiple compounds
- Support for **intra-protocol adaptivity**, leading to resources being assigned at runtime which provides the efficient execution of individual protocols

Protocols can be expressed and implemented easily with the HTBAC user-facing API:

- Allowing domain scientists to scale to a sizable execution and performance
- Define additional adaptivity parameters that are passed down to the underlying runtime system.



- HTBAC rests on RADICAL-Cybertools (RCT)
  - The Ensemble Toolkit decouples the description of ensemble-based applications from their execution by abstracting resource and execution management.
  - The runtime system, RADICAL-Pilot enables the submission of container jobs to the resource manager of the HPC systems.
- HTBAC has shown **performance and scalability of the ESMACS and TIES protocols on leadership class machines including NCSA Blue Waters and ORNL Titan**.

Adaptivity at Scale

HTBAC uses adaptive schemes to enhance the scientific accuracy or convergence rate of simulations. We define two types of adaptive schemes:

1. **Concurrent simulation count adaptivity:** results analysed at run time inform decisions about the number of concurrent simulations. In Figure 2 the adaptive quadrature algorithm adds additional simulations to reduce the error on the predicted binding affinity.
2. **Simulation order adaptivity:** a user-defined threshold criterion (e.g. convergence tolerance), is used to determine whether to add (or remove) simulation steps from the protocol. Figure 3 shows how error monitoring of the observable can reduce time to convergence.

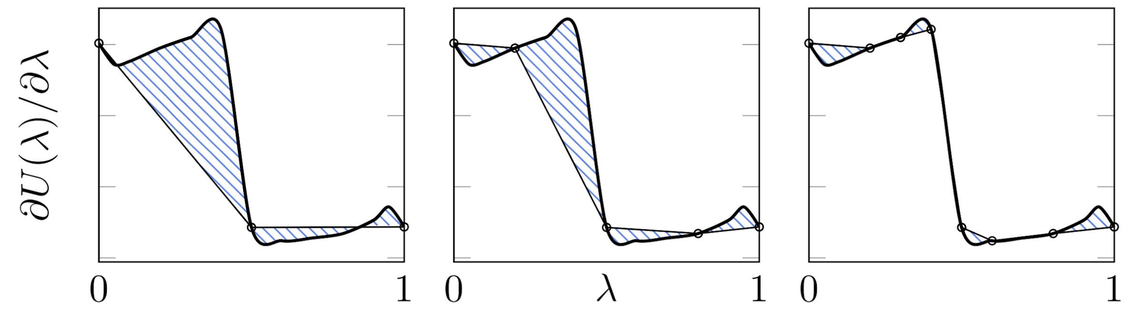
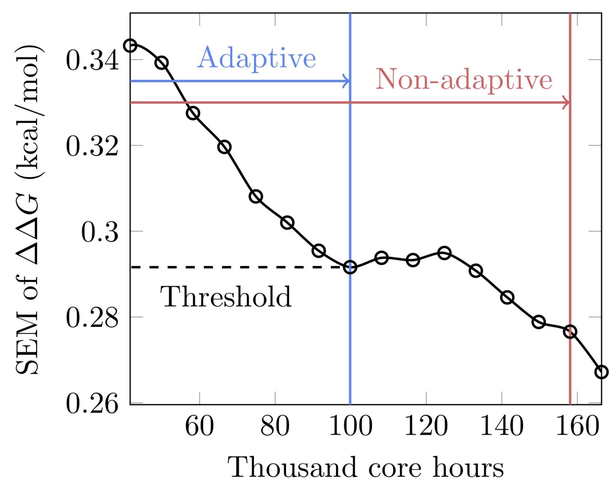


Figure 2: Adaptive quadrature of the function f(λ) = ∂U/∂λ in the interval [0, 1] using the trapezoidal rule. Each circle represents a λ value at which a simulations may be executed. From left to right the simulations are increased to increase fidelity, with extra runs bisecting points where deviation between existing points is above a set threshold.The true integration error is the difference between the interpolated function and the actual function (shaded area).

Results

Figure 3 (right): The uncertainty in the computed observable - measured using the standard error of the mean (SEM) - is monitored as the simulation progresses. HTBAC adapts the simulation protocol by adding or removing stages. Once the convergence criteria is met, the next drug candidate is considered. HTBAC optimizes the time to insight by removing redundant stages, while obtaining robust, converged results and increasing the throughput of drug candidates investigated.



Weak and strong scaling properties of HTBAC on NCSA Blue Waters
- **TTX** measures total time to execution, while
- **Overhead** captures HTBAC and RCT

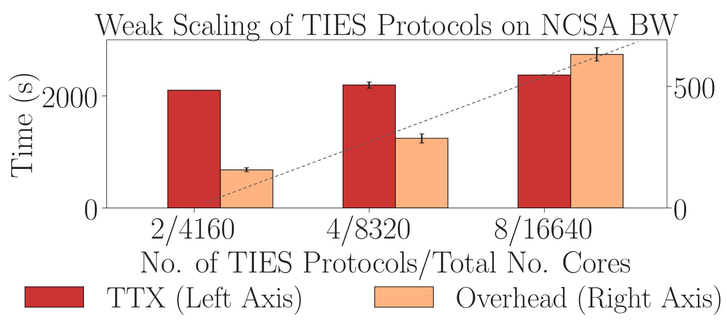


Figure 4 (Left) weak scaling:
The ratio of the number of TIES protocol instances to the amount of resources is kept constant. We ran three trials at each protocol configuration. The overhead scales linearly with the number of protocol instances. Error bar in 2 and 8 protocol instances TTX are insignificant.

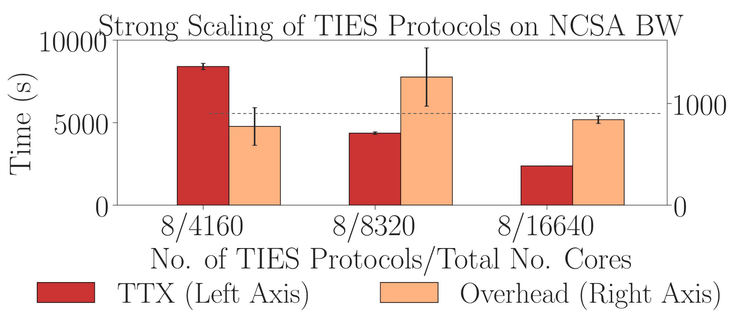


Figure 5 (Left) strong scaling:
We fix the number of protocol instances and vary the amount of resource. For a fixed workload, we find linear decrease in TTX. To within errors, the overhead remains constant. Two trials at each protocol configuration. Error bar at 16,640 cores TTX is insignificant.

Impact on drug discovery

The drug design process involves the filtering of millions of compounds to smaller number of "hits" that bind the target protein and then further refine to "leads" that form the basis of candidate drugs. While experiment automation reduces completion time, *in silico* approaches have the potential to cut time to completion and cost by an order of magnitude. This uptake of computational techniques in industrial settings fits into the wider vision of building a **digital twin,** where the target protein is based on the patient's genetic sequence, enabling **personalized medicine**.

HTBAC provides a flexible middleware layer enabling the execution of simulation protocols with the scale and turnaround time required for decision making in an industrial profit driven setting.

Adaptivity
- Automation of calculation optimization (e.g. simulation duration and lambda placement in TIES)
- Ensures efficient resource usage for calculations involving highly variable chemical and convergence properties
- Allows strategic allocation of resources (computation distribution informed by lower fidelity results)

Encapsulation
- Facilitates use by non computational experts
- Enhanced repeatability and reproducibility
- Facilitates sensitivity analyses

Scale
- Lead optimization considers of the order of 10k small molecules
- TIES predictions currently use ~25k core hours
- Pharmaceutical campaign could use a quarter billion core hours (over a two week duration)
- Huge savings possible through adaptivity

JD acknowledges Dept. of Education Award P200A150217 for her fellowship. PVC would like to acknowledge the support of the EU H2020 Comp-BioMed (675451), EUDAT2020 (654065) and ComPat (671564) projects. We thank Andre Merzky and other members of the RADICAL team for support. Access to Blue Waters was made possible by NSF 1713749. The software capabilities were supported by RADICAL- Cybertools (NSF 1440677)