Concurrent Drug Binding Affinity Calculations using Adaptive Simulations

No Author Given

No Institute Given

Abstract. The vision and potential of precise, accurate and rapid computation of ligand binding strengths has driven interest in the use of molecular simulations for applications in both computer-aided drug design and patient specific medicine. Simulation protocols based on ensembles of multiple runs of the same system provide an efficient method for producing robust free energy estimate, and equally important statistical uncertainties. Variations in the chemical and biophysical properties of different systems impact the optimal protocol choice for different proteins and classes of drugs targetting them. However, these are difficult to determine *a priori*, thus requiring adaptive execution of a simulation workflows. To support the scalable, adaptive and automated calculation of the binding free energy on high-performance computing resources, we introduce the High-throughput Binding Affinity Calculator (HTBAC). HTBAC uses readily available building blocks to attain both workflow flexibility and performance. We demonstrate close to perfect weak scaling over hundreds of concurrent multi-stage binding affinity calculation pipelines. This permits a rapid time-to-solution that is essentially invariant of the calculation protocol, size of candidate ligands and number of ensemble simulations. As such, HTBAC advances the state of the art of binding affinity calculations and protocols.

Keywords: binding free energy, binding affinity, drug binding, middleware

1 Introduction

Recent developments in both algorithms and architectures (in particular GPGPUs) have led to increasing interest in the use of molecular simulation to estimate the strength of macromolecular binding free energies [1]. However, for molecular simulations to truly influence decison making in industrial and clinical settings, the dual challenges of scale (thousands of concurrent multi-stage pipelines) and sophistication (adaptive sampling or selection of binding affinity protocols based upon system behaviour and statistical uncertainty) will need to be tackled [**?**].

Tools that facilitate the scalable, automated and sophisticated computation of varied binding free energy calculations on high- performance computing resources are necessary. Towards that goal we recently introduced the High-Throughput Binding Affinity Calculator (HTBAC) [2], which brings advances

in the design of domain-specific workflow systems using building blocks to facilitate the building and automation of rapid and accurate calculation of binding affinities. In Ref. [2] we demonstrated how HTBAC scales almost perfectly to hundreds of concurrent binding affinity calculations on a leadership class machine. This permits the rapid time-to-solution that is essentially invariant of the size of candidate ligands as well as the type and number of protocols concurrently employed.

In this case we look to reproduce a collaboration run between UCL and GlaxoSmithKline to study a congeneric series of drug candidates binding to the BRD4 protein (inhibitors of which have shown promising preclinical efficacy in pathologies ranging from cancer to inflammation) [3]. This study compared two different protocols, known as TIES and ESMACS, both based on an ensemble simulation philosophy. In this approach multiple simulations are executed based on the same input system description, providing enhanced sampling and reduced time to completion. TIES is based on rigourous, but computationally expensive, calculations of relative free energies (i.e. results provide a comparison between two drugs). ESMACS, in contrast, provides absolute binding free energies at low computational cost, but to achieve this coarse grains many of the details of the system being studied. The results of the simplifications employed by ESMACS is that its performance is heavily system dependent. In the real world application of these technologies, drug design projects have limited resources and must make trade-offs between the needs for rigour and coverage of a wide range of chemical space. Initially large numbers of compounds must be screened to eliminate poor binders (using ESMACS), later more accurate methods (such as TIES) are needed as good binders are refined and improved. This means that many projects will combine the use of both protocols.

We explore the use of HTBAC for aforementioned sophisticated exploration of novel compounds binding to a target protein. In order to support such investigations HTBAC must be enhanced to support flexible resource reallocations schemes where resources can be moved between simulations run using different protocols or systems, for example, when one calculation has converged whilst another has not. This adaptability makes it easier to manage complex programmes where efficient use of resources is required in order to achieve a time to completion of studies comparable to those of high throughput chemistry. This functionality provides the foundations upon which we develop adaptable simulation schemes that automatically handle different system characteristics.

In this work we demonstrate the use of HTBAC to adaptively run both protocols, including mixed protocol runs. Both protocols can involve the running sets of simulations which require very different levels of computational power. Further to the use of a common framework to improve the ease of deployment and efficiency of execution of existing protocols we show how HTBAC can aid the development of enhanced approaches. The TIES protocol is highly sensitive to the chemical details of the compounds being studied which means that different runs may require different sampling strategies to achieve optimal time to convergence. We have developed an adaptive variant of TIES which automati-

cally increases the sampling in areas where it is needed for each system, allowing more rapid convergence of calculations. Furthermore, this approach facilitates a data driven approach to future protocol refinement.

This approach not only allows us to tailor the protocol to the particular system but to provide meaningful statistical uncertainties for all of our models. These developments fit into a wider vision in which the use of flexible and responsive computational protocols allow us to produce accurate, precise and reproducible estimates of the free energy of binding. Not only would this allow for wider uptake of computational techniques in industrial settings but opens up possibilities of using these technologies in clinical decison support scenarios. By creating a 'digital twin', where the target protein is based on the real patients genetic sequence, a specific individuals response to different treatements could be predicted. This approach would be applicable even in the case of rare variants where insufficient data is available for statistical methods to be informative.

In the next section, we review previous work using ensemble molecular dynamics and outline the challenges faced in bringing this approach up to extreme scale. Section 3 provides details of the TIES and ESMACS protocols and the BRD4 system we will applly them to. In Section 4, we discuss how HTBAC has been designed and implemented in order to meet the computational challenges associated with the scalable execution of multiple, and possibly concurrently executing protocols. In Sections 5 and 6 we describe the design and then results of a series of experiments characterizing the performance and scalability of HTBAC on the Blue Waters and Titan supercomputers. We conclude with a discussion of the impact of HTBAC, implications for binding affinity calculations and near-term future work.

2 Related Work

The strength of ligand binding is determined by a thermodynamic property known as the binding free energy (or binding affinity). One promising technology for estimating binding free energies and the influence of protein and ligand composition upon them is molecular dynamics (MD) [4]. A diversity of methodologies have been developed to calculate binding affinities MD sampling [5] and blind tests show that many have considerable predictive potential [6, 7]. The development of commercial approaches that claim accuracy of below 1 kcal mol^{-1} [8] has led to increased interest from the pharmaceutical industry [9]. The same technologies have also shown promise in analysing the influence of protein mutations on drug binding [10, 11]. These developments have also spurred renewed interest in the factors affecting the performance of different free energy methods [12–14]. As these developments combine with increased computational power it is becoming possible to integrate the information gained from large numbers of simulations to provide more predictive models using machine learning approaches [15]. It is this approach of combining predictive modelling with integrative technologies which provides one of the most promising avenues for developing decision support tools for drug discovery and clinical support.

As binding free energy approaches based on MD become more widely used, particularly by non-specialists, automated tools for system and simulation preparation have been developed [16–18], including our own tool the binding affinity calculator (BAC) [19]. Using BAC, we have designed two protocols with the demands of clinical decision support and drug design applications in mind: ESMACS (enhanced sampling of molecular dynamics with approximation of continuum solvent) [3] and TIES (thermodynamic integration with enhanced sampling) [20]. The former protocol is based on variants of the molecular mechanics Poisson-Boltzmann surface area (MMPBSA), which is an "approximate" endpoint method [21]. The latter on the rigorous "alchemical" thermodynamic integration (TI) approach [22].

Using these protocols, we have demonstrated the lack of reproducibility of individual simulations for a variety of protein systems, with calculations for the same protein-ligand combination using almost identical initial conditions producing widely varying results (binding affinities varying by up to 12 kcal mol $^{-1}$ for small molecules, whilst flexible ligands can vary even more) [23–25]. Indeed, our work has revealed how completely unreliable single simulation based approaches are. However, we have shown that averaging across multiple runs can reliably produce results in agreement with previously published experimental findings [3, 20, 24–27], and correctly predicted the results of experimental studies performed by colleagues in collaboration [11]. We term this approach ensemble molecular dynamics, "ensemble" here referring to the set of individual (replica) simulations conducted for the same physical system.

Both TIES and ESMACS protocols are ensemble based. Basing these computations on the direct calculation of ensemble averages facilitates the determination of statistically meaningful results along with complete control of errors. In addition to accuracy and precision, in order for MD simulations to attain widespread use in industrial (and potentially clincal) settings, it is necessary that results can be obtained soon enough in order that the results can feed into decision making processes. These motivations provided the impetuse behind our automated model building and simulation input creation package, the binding affinity calculator (BAC) [19]. As we scale up to larger datasets the use of the ensemble approaches will increasingly necessitate the use of middleware to provide reliable coordination and distribution mechanisms with low performance overheads. This has provided the motivation behind the development of HTBAC.

3 Science Drivers

The basic use case for HTBAC is to enable large scale free energy s  es of protein-ligand binding using ensemble simulations. We have already briefly introduced two free energy calculation protocols, ESMACS and TIES [3, 20], with reference to the comparative rigour of the former relative to the later. Here we provide more details of their practical specifications and of the protein system we used to benchmark and refine them in this work.

3.1 ESMACS and TIES

In both protocols, ensembles of MD simulations are employed to perform averaging and to obtain tight control of error bounds in our estimates. In addition, the ability to run replica simulations concurrently means that, as long as sufficient compute resources are available, turn around times can be significantly reduced compared to the generation of single long trajectories. Due to their shared philosophical underpinning both protocols share similar middleware requirements.

The current implementation of both TIES and ESMACS uses the the NAMD package [28] to conduct the simulations. Conceptually, each replica simulation consists of three stages: minimization, equilibration and production MD. In practice the equilibration phase is broken into multiple steps to ensure that the size of the simulation box does not alter too much over the simulation. During these steps positional constraints are gradually released from the structure and the system is heated to a physiologically realistic temperature. Whilst both protocols share a common workflow for individual replicas, the make up of the ensemble is different.

In the case of ESMACS, an ensemble consists of a set of identical simulations differing only in the initial velocities assigned to each atom. Upon completion of the MD simulation, free energy computation (via MMPBSA and potentially normal mode analysis) is performed using AmberTools [29–31].

TIES is a so called "alchemical" method in which the fact that the potential used to decribe the system is user definable to transform one system into another. This allows for the calculation of free energy differences between the two systems, such as those induced by an alteration to a candidate drug. Typically a transformation parameter (called λ) is defined such that as it increases from zero to one the system description transforms from containing an initial drug to a target compound via a series of hybrid states. Sampling along λ is then required in order to compute the difference in binding free energy. Consequently, whereas in ESMACS all replicas sample using the same system description, in TIES sub-ensembles are executed at diffent points along λ. The change in free energy associated with the transformation is calculated by numerically integrating (via adaptive quadrature) the values of the $dU/d\lambda$ across the full set of λ windows simulated. The TIES protocol, for example, relies on the numerical integration of $dU/d\lambda$.

Obtaining accurate and precise results from TIES requires that the λ windows correctly capture the changes of $dU/d\lambda$ over the transformation. This behaviour may vary considerably between systems. Typically, windows are evenly spaced between 0 and 1 with the spacing between them set before execution at a distance determined by the simulator to be sufficient for a wide range of systems. Typically, a TIES ensemble 65 replicas evenly distributed between 13 λ windows In order to obtain a meaningful TIES result it is necessary to not only simulate the drug pair in the protein but also in an aqueous environemnt, adding a further 65 replicas albeit it using a smaller system at lower computational cost.

Following simulation (and in the case of ESMACS free energy analysis steps) both protocols require the execution of short serial steps to provide summary

statistics. Both protocols can are highly customizable, for example the number of simulation replicas in the ensemble and the lengths of their runs can be varied. More generally though, there may be cases where it is important to increase the sampling of phase space possibly through expanding the ensemble or in TIES changing the distributon of λ windows. In addition, the ESMACS protocol can also be extended to account for adaptation energies involved in altering the conformation of the protein or ligand during binding through the use of separate component simulations.

3.2 The Value of Adaptivity

Both ESMACS and TIES have been sucessfully used to predict binding affinities quickly and accurately. Nonetheless, they are very expensive computationally, and optimizing the execution time while still improving the accuracy is desirable. Given the very large number of drug candidates, it is imperative to gain maximum insight into potential candidate compounds using time and resources efficiently. This provides one clear motivation for the use of adaptive methods which minimize the compute time used whilst producing binding free energy estimates meeting pre-defined quality criteria (such as convergence or statistical uncertainty below a given theshold).

A second driver for adaptivity is that such algorithmic methods will typically involve compounds with a wide range of chemical properties which can impact not only the time to convergence, but the type of sampling required to gain accurate results. In general, there is no way to know before running calculations exactly which setup of calculation is required for a particular system. For example in TIES, the number (or the exact location) of the λ windows that will most impact the calculation are not known *a priori*, and change between physical systems (drugs). As multiple simulations must be run for each window, sampling with a very high frequency is expensive and impractical. Furthermore, adaptive placement of λ windows is likely to better capture the shape of the $dU/d\lambda$ curve, leading to more accurate and precise results for a given computational cost.

On occasion, alchemical methods may be very slow to converge, in such circumstances use of another method, such as ESMACS, may be the best option. This means that the most effective way to gain accurate and precise free energy results on industrially or clinically relavant timescales is to be able to adapt both sampling (intra-protocol) and even adapt protocol (inter- protocol) used at run time. With potentially thousands of simulations, often employing multiple analysis methodologies, this is indeed the only way to effectively utilize these techniques and resources at scale.

3.3 Target protein: BRD4

Bromodomain-containing proteins, and in particular the four members of the BET (bromodomain and extra terminal domain) family, are currently a major focus of research in the pharmaceutical industry. Small molecule inhibitors

of these proteins have shown promising preclinical efficacy in pathologies ranging from cancer to inflammation. Indeed, several compounds are progressing through early stage clinical trials and are showing exciting early results [32]. One of the most extensively studied targets in this family is the first bromodomain of bromodomain-containing protein 4 (BRD4-BD1) for which extensive crystallographic and ligand binding data are available [33–35].

We have previously investigated a congeneric series of ligands binding to BRD4-BD1 (we shall from now on refer to this are simply BRD4) using both ESMACS and TIES. This was performed in the context of a blind test of the protocols in collaboration with GlaxoSmithKline [3]. The goal was to benchmark the ESMACS and TIES protocols in a realistic drug discovery scenario. In the original study, we investigated chemical structures of 16 ligands based on a single tetrahydroquinoline (THQ) scaffold [36]. Here we focus on the first seven of these ligands to test and refine the protocols used and the way in which they were executed. The results of our previous work provide a benchmark of both accuracy and statistical uncertainty to which we can compare our new results.

Initial coordinates for the protein-ligand system where taken from the X-ray crystal structure PDB ID: 4BJX [37]. This structure contains a ligand based on the THQ template and other ligands were alligned with this common scaffold. Preparation and setup of the simulations were implemented using our automated tool, BAC [19]. This process including parametrization of the compounds, solvation of the complexes, electrostatic neutralization of the systems by adding counterions and generation of configurations files for the simulations. The AMBER ff99SB-ILDN [38] force field was used for the protein, and TIP3P was used for water molecules. Compound parameters were produced using the general AMBER force field (GAFF) [39] with Gaussian 03 [40] to optimize compound geometries and to determine electrostatic potentials at the Hartree–Fock level with 6-31G** basis functions. The restrained electrostatic potential (RESP) module in the AMBER package [30] was used to calculate the partial atomic charges for the compounds. All systems were solvated in orthorhombic water boxes with a minimum extension from the protein of 14 Å (the resulting systems contain approximately 40 thousand atoms).

4 HTBAC: Design and Implementation

Most advances in high-performance computing have focused on the scale, performance and optimization of single long simulations. However, due to the end of Dennard scaling and methodological advances, many applications are now formulated using multiple, shorter ensemble-based simulations. There are limited software solutions that can support the scalable execution of heterogeneous computational tasks. HTBAC builds upon established middleware solutions [], validated runtime abstractions [] for scalable execution of heterogeneous computational tasks, and customizes them for the computation of binding free energies.

In principle, many workflow management systems (WMS) can be used to express ensemble-based workflows to compute binding free energies. General-

purpose WMS however, are limited by a lack of specificity: workflow system with feature rich, but complex interface models impose substantial overhead when integrating application workflows in the runtime system, preventing users from quick and flexible applications prototyping.

Furthermore, statistically meaningful results derived from simulations are of critical interest, as they can leverage additional sampling of simulations which could lead to an improvement in precision of free energy calculations. Such decisions typically, cannot be made *a priori* and require postmortem user intervention in order to analyze and spawn additional simulations. Specific features such as steering and adaptivity (e.g., spawning additional simulations) during runtime often require a redistribution of resources. Current WMS do not provide adequate support for dynamic resource utilization and require special extensions for adaptive formulations thus limiting the possibility of adaptive workflows.

This suggests three design requirements for HTBAC: (1) enable the scalable execution of concurrent multiple protocols, (2) abstract the complexity of building protocols, execution and resource management from the user; and (3) provide adaptive features (i.e., modifiying the task-graph during runtime, based on previous tasks), to enable more effective protocols, without explicit resource management requirements.

HTBAC derives many of the advantages of a lightweight, flexible domain specific workflow layer from its use of RADICAL-Cybertools (RCT) which are functionally well-defined and delineated middleware building blocks. RCT [41] are engineered to support extensible and scalable workflows across diverse computing platforms. Although enscapulated from end- users of HTBAC, the two primary RCT components that HTBAC depends upon are the Ensemble Toolkit (EnTK) and RADICAL-Pilot (RP).

HTBAC is implemented in Python, and uses RCT to provide ensemble-execution capabilities and a runtime system to execute tasks. Ensemble Toolkit (EnTK), the topmost layer of RCT simplifies the process of creating ensemble-based applications with complex coordination and communication requirements. The EnTK API exposes the PST model which consists of three components: **Pipeline**, **Stage**, and **Task**.

Consistent with EnTK's programming model, HTBAC also uses the PST model to to express **Protocols**. HTBAC promotes binding affinity protocols as the user-facing constructs, and provides programming model defined by the Pipeline, Stage and Tasks model to express a variety of protocols. Each protocol contains multiple stages with simulations and analysis tasks interspersed, in the most general case. We define each stage as a computational **task** and the ordered aggregation of these stages alongside their dependencies as a **pipeline**. Figure **??** provides the visual implementation of the TIES protocols into the PST model. HTBAC allows the simple expression and concurrent execution of multiple distinct protocols, thereby enabling concurrent screening of drug candidates. HTBAC not only simplifies the expression of complex binding affinity protocols, but also provides hitherto unavailable capabilities, viz., the adaptive

formation/expression and execution of these protocols, without any additional programming burden.

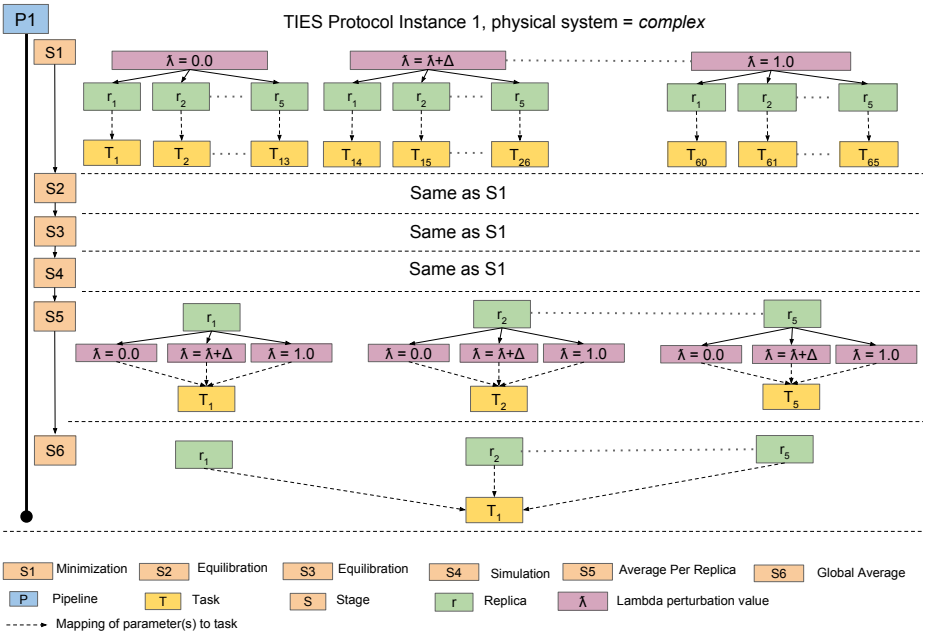

Fig. 1. TIES protocol expressed using the EnTK PST model. Each protocol instance maps to a single Pipeline, comprised of Stage(s) which maintain temporal order. Each Stage executes n tasks, where n represents the number of lambda-replica combinations.

A workflow is comprised of N_P instances of the Pth protocol. Once the workflow is described, it is submitted to the Application Manager which sets up multiple processes, threads and a RabbitMQ message queue for communication. EnTK identifies tasks which have dependencies satisfied and can be executed concurrently. EnTK's **Execution Manager** uses the underlying runtime system, RADICAL-Pilot to execute the tasks on specific target resources.

HTBAC exposes a **Protocol** component as the main user facing component. The **Protocol** class can be instantiated as one of two protocols. Currently the HTBAC provides the two protocol objects: ESMACS and TIES. For TIES, each protocol typically corresponds to a single physical system, and multiple instances can be supported. The TIES protocol object enables the user to select a physical system, number of replicas, core allocation per replica, and lambda windows (for alchemical pertubations). Similarly, the ESMAC protocol enables the user to select ...

In Section 2 we highlighted how an ensemble simulation approach can both aid sampling and improve uncertainty quantification for free energy calculations.

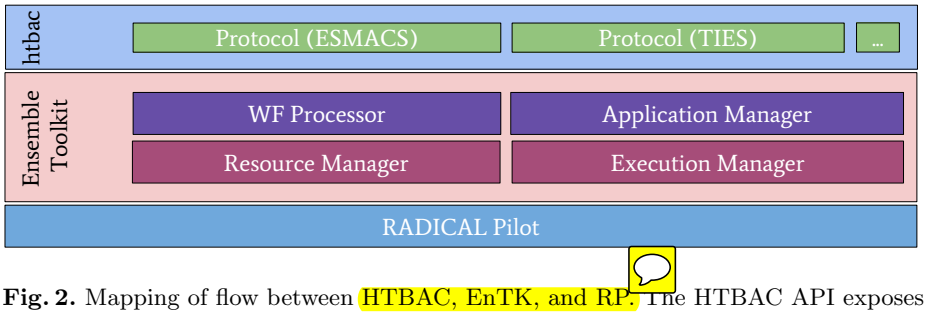

Fig. 2. Mapping of flow between HTBAC, EnTK, and RP. The HTBAC API exposes the Protocol component. EnTK serves as the workflow execution system and by managing the workflow and workload. RADICAl-Pilot serves as the runtime system

Despite these crucial advantages, it remains non-trivial for field researchers to write biosimulation applications that involve individual protocols supporting multiple replicas, and by extension multiple protocols. With HTBAC, this burden is minimized by specifying the number of concurrent instances of the **Protocol** object. Moreover, the ability to generate multiple protocol instances enables the user to investigate a range of physical systems (i.e., drug candidates) concurrently.

In our earlier work [2] we demonstrated how we applied the PST model to implement the ESMACS protocol into an EnTK application. In this work we validate the generality of HTBAC by implementing the TIES protocol and investigating its performance with multiple TIES protocol.

A protocol instance

We compare overheads of HTBAC and demonstrate that its overheads and performance are invariant of the specific protocol..... ***shantenu: connective tissue for Experiments Section...

5 Experiments

Design and Methods

For the TIES protocol, each pipeline consists of six stages. Each of the simulation stages contains a task for every unique λ-replica combination.

In the non-adaptive workflow scenario, the first 11 λ windows consist of the following values: L is a vector with

$$L = \{x_i : x_i \in [0,1] \; and \; x_{i+1} = x_i + \delta\}, where \; \delta \; is \; 0.1. \tag{1}$$

We append two additional values on both ends of L, completing 13 λ windows. Each λ window consists of five replicas. Therefore there are a total of 65 tasks for every simulation stage. The production simulations stage, $s4$ as described in 1 executes a four ns simulation duration. The analysis stages of the protocol reduce the number of tasks. The first analysis task consists of five tasks where each task performs an aggregate analysis over all λ windows for each replica. The

second analysis stage consists of one task that aggregates the previous results and computes a single average across all replicas.

In the adaptive workflow, over the course of a protocol instance we alter the number of λ windows being simulated. The number of λ windows and their respective values depend on the difference between the $dU/d\lambda$ measured between adjacent windows. Increasing the number of λ windows in regions of rapid change will increase the accuracy of the overall integral to a greater degree than an arbitrarily placed window. In order to access the $dU/d\lambda$ values during runtime, we break down the single production simulation stage (S4) from the nonadaptive workflow into multiple stages for the adaptive workflow. Each stage in the adaptive workflow produces only 1 ns of simulation. Once each stage is complete a decision is made about whether more λ windows are required, and if so where they should be placed. We start the workflow with only five λ parameters that consist of the following values:

$$L = \{x_i : x_i \in [0,1] \; and \; x_{i+1} = x_i + \delta\}, where \; \delta \; is \; 0.33. \tag{2}$$

For every λ window we initialize with three replicas therefore yielding a total of 15 tasks. We run 15 tasks for stages $S1$ through $S4.1$. Between stages $S4.1$ and $S4.3$ the number of λ windows doubles for every stage, which doubles the total number of tasks. The last production simulation stage, $S4.4$, runs for the remaining 2 ns durations.

HTBAC provides the functional capability to adaptively deterine the time at which the λ window are determined. However in this paper, we eschew a discussion of this type of adaptivity, as the objective is to determine the feasibility of adaptive execution and demonstrate the scientific merit of implementing adaptive decision making. In order to provide "control" baseline experiments, our experiments implement "adaptive change" not "when" i.e., we introduce only a single degree of freedom relative to baseline "non-adaptive" experiments.

Weak Scaling Experiments ***shantenu: 1 sentence about how pipelines in ESMACS were simpler than TIES? else the reader will say: you did 128 pipelines already... why should I be impressed by a small number of pipelines. Important to highlight that the scalability is over PROTOCOL INSTANCES and not pipeline instances. ***jumana: added blurb in design & implementation, removed it from here

Here we show weak scalability for the TIES protocol by growing the number of protocol instances while adhering to the required number of pipelines. By design of each protocol, an increase in the number of instances simply means an increase in the number of pipelines. Therefore our previous ESMACS experiment referenced in [ref SC] already demonstrates the scalability of ESMACS as a growth in the number of pipelines.

The first weak scalability experiment demonstrates the behavior of HTBAC, EnTK and RP using the multiple instances of the TIES protocol. By design of weak scaling, the ratio between the number of pipelines and cores are kept

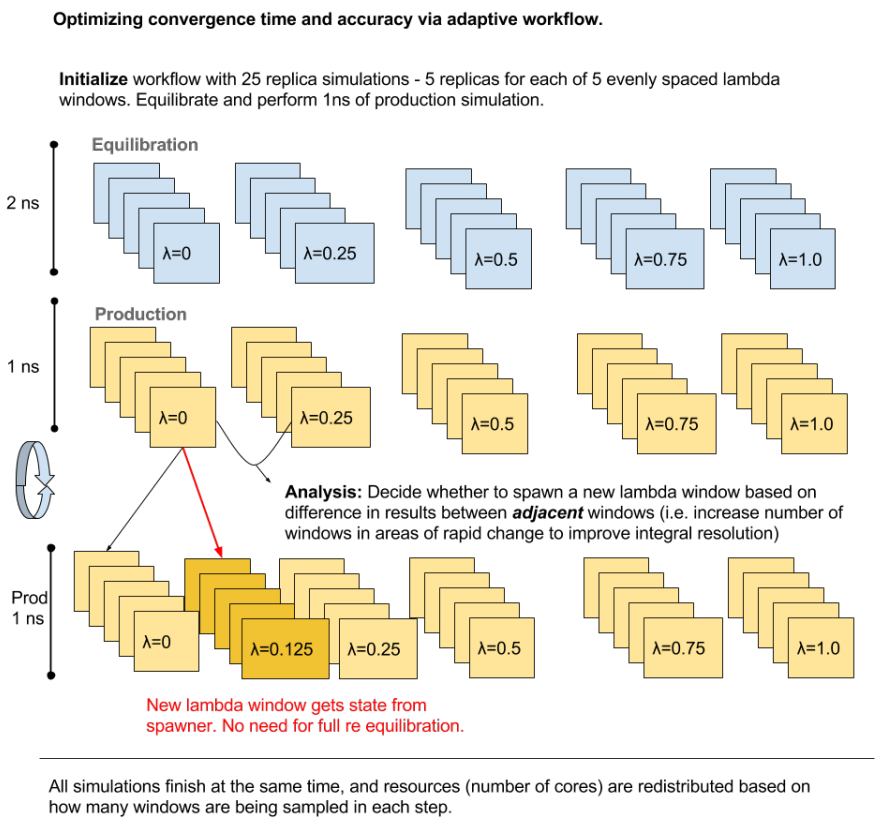

Fig. 3. ***shantenu: Please provide caption. Please reference and use in text.

constant. ***shantenu: This is some what inconsistent with Fig. 2, which shows constant ratio between protocol instances and resources used ***jumana: will modify once we get new data in the early a.m. As the number of cores (measure of resource) changes by a factor of 2, we introduce twice as many protocol instances. As designed, the weak scaling property provides insight into the size of the workload that can be investigated in a given amount of time.

Strong Scaling Experiments Next we repeat the same design of the weak scalability experiments but examine performance of strong scaling when fixing the number of pipelines and varying the resources. The comparison between weak and strong scalability demonstrates the overhead introduced by load balancing and scheduling tasks in multiple generations.

5.1 Experiment Setup

We perform weak (and strong) scalability experiments on NCSA Blue Waters–a 13.3. petaFLOPS Cray, with 32 Interlago cores/50 GB RAM per node, Cray Gemini, Lustre shared file system. We perform our experiments from a virtual machine hosted in Europe, as Blue Waters does not allow for executing applications directly on the login node.

All experiments use HTBAC version 0.1, EnTK version 0.6 and RP version 0.47. The MD engine used is NAMD-MPI for tasks pertaining to $S1$ - $S4$, while the analysis stage, $S5$ use AmberTools. For both adaptive and nonadaptive experiments, the minimization tasks of $S1$ are assigned 100(0) steps, while the equilibration tasks in $S2$ and $S3$ are assigned 5000 steps. In the nonadaptive experiments, the production simulation tasks in $S4$ are assigned 50000 steps. For the adaptive experiments, each substage of $S4$ i.e. $S4.1$ - $S4.4$ is assigned X steps.

HTBAC submits a resource request to EnTK, to which EnTK uses RP to acquire resources via a single pilot. The performance of the pilot is contingent upon characterization of performance, in this case, weak and strong scalability. Accordingly, we request the maximum number of cores required by the workload as the number of cores in a pilot. We use between 4160 and 33280 cores as indicated in figure 4 because the NAMD executable used in all tasks from $S1$-$S4$ require at least 32 cores per task. From our own scalability performance measurements of NAMD on Blue Waters, we observe the ideal cores per task to be 16, however Blue Waters does not permit running multiple MPI applications on the same node, hence each NAMD task requires a full node for execution.

5.2 Overheads of HTBAC

HTBAC enables concurrent execution of multiple protocol instances. With each new protocol instance generated for an application, the HTBAC overhead grows to match the additional requirement of generating and coordinating protocols. In order to understand the contribution of the various events in HTBAC, termed as

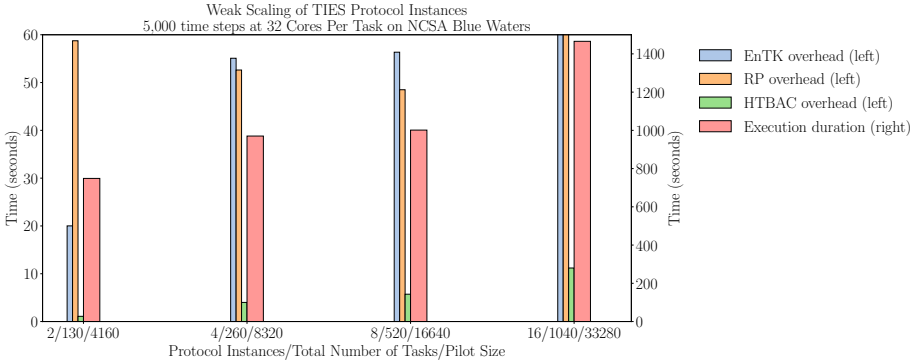

Fig. 4. Weak scaling properties of HTBAC (right side). We investigate the weak scaling of HTBAC as the ratio of the number of protocol instances to resources is kept constant. (Left) Overheads of HTBAC, EnTK and RP for experimental configurations investigating the weak scaling of TIES. We ran two trials for each protocol instance configuration. ***shantenu: (i) I would replace "Pilot Size" with "Number of Cores". (ii) How is HTBAC Overhead measured? Maybe extend Y-left range to 75 or so? (iii) Why do we not have "closer" to linear scaling for 5,000 timesteps?? (iv) Minor nitpick: Is it possible to plot the bars equal widths? ***jumana: will do

HTBAC overhead, to the total time to run. We construct the following metrics : $Time - to - run = T(overhead_{\mathrm{HTBAC}} + T(overhead_{\mathrm{entk}}) + T(overhead_{\mathrm{rp}}) + T(execution)$. We define $T(execution)$ as $TTX = TTC - T_q$ where TTC is time-to-completion and T_q is time spent queuing on the HPC machine.

6 Results

7 Conclusions

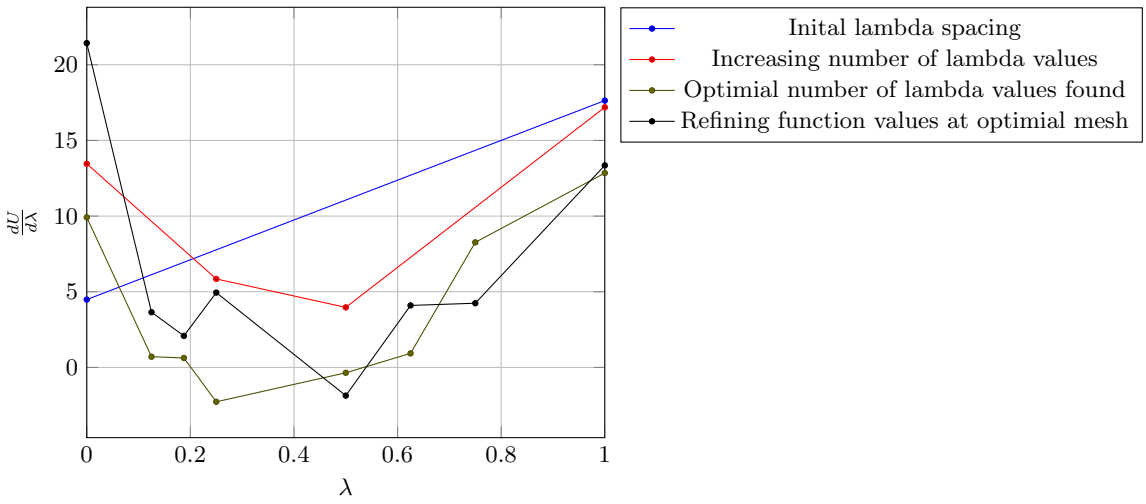

Fig. 5. Example figure of adaptive illustration. Caption.

Bibliography

[1] M. De Vivo, M. Masetti, G. Bottegoni, and A. Cavalli, "Role of molecular dynamics and related methods in drug discovery," *Journal of Medicinal Chemistry*, vol. 59, no. 9, pp. 4035–4061, 2016, pMID: 26807648. [Online]. Available: http://dx.doi.org/10.1021/acs.jmedchem.5b01684

[2] J. Dakka, M. Turilli, D. W. Wright, S. J. Zasada, V. Balasubramanian, S. Wan, and S. J. Peter V. Coveney, "High-throughput binding affinity calculations at extreme scales," *Computational Approaches for Cancer Workshop at SuperComputing17*, 2017. [Online]. Available: http://arxiv.org/abs/1712.09168

[3] S. Wan, A. P. Bhati, S. J. Zasada, I. Wall, D. Green, P. Bamborough, and P. V. Coveney, "Rapid and reliable binding affinity prediction of bromodomain inhibitors: a computational study," *J. Chem. Theory Comput.*, vol. 13, no. 2, pp. 784–795, 2017.

[4] M. Karplus and J. Kuriyan, "Molecular dynamics and protein function." *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 6679–6685, May 2005.

[5] D. L. Mobley and P. V. Klimovich, "Perspective: Alchemical free energy calculations for drug discovery," *The Journal of Chemical Physics*, vol. 137, no. 23, p. 230901, 2012. [Online]. Available: https://doi.org/10.1063/1.4769292

[6] A. S. J. S. Mey, J. J. Jiménez, and J. Michel, "Impact of domain knowledge on blinded predictions of binding energies by alchemical free energy calculations," *Journal of Computer-Aided Molecular Design*, Nov 2017. [Online]. Available: https://doi.org/10.1007/s10822-017-0083-9

[7] J. Yin, N. M. Henriksen, D. R. Slochower, M. R. Shirts, M. W. Chiu, D. L. Mobley, and M. K. Gilson, "Overview of the sampl5 host–guest challenge: Are we doing better?" *Journal of Computer-Aided Molecular Design*, vol. 31, no. 1, pp. 1–19, Jan 2017. [Online]. Available: https://doi.org/10.1007/s10822-016-9974-4

[8] L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, D. L. Romero, C. Masse, J. L. Knight, T. Steinbrecher, T. Beuming, W. Damm, E. Harder, W. Sherman, M. Brewer, R. Wester, M. Murcko, L. Frye, R. Farid, T. Lin, D. L. Mobley, W. L. Jorgensen, B. J. Berne, R. A. Friesner, and R. Abel, "Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field," *J. Am. Chem. Soc.*, vol. 137, no. 7, pp. 2695–2703, Feb. 2015. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ja512751q

[9] A. Ganesan, M. L. Coote, and K. Barakat, "Molecular dynamics-driven drug discovery: leaping forward with confidence," *Drug Discovery Today*, vol. 22, no. 2, pp. 249 – 269, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1359644616304147

[10] J. Mondal, P. Tiwary, and B. J. Berne, "How a kinase inhibitor withstands gatekeeper residue mutations," *J. Am. Chem. Soc.*, vol. 519, p. 102, 2016.

[11] T. D. Bunney, S. Wan, N. Thiyagarajan, L. Sutto, S. V. Williams, P. Ashford, H. Koss, M. A. Knowles, F. L. Gervasio, P. V. Coveney, and M. Katan, "The effect of mutations on drug sensitivity and kinase activity of fibroblast growth factor receptors: a combined experimental and theoretical study," *EBioMedicine*, vol. 2, no. 3, pp. 194–204, 2015.

[12] M. Aldeghi, M. J. Bodkin, S. Knapp, and P. C. Biggin, "Statistical analysis on the performance of molecular mechanics poisson–boltzmann surface area versus absolute binding free energy calculations: Bromodomains as a case study," *Journal of Chemical Information and Modeling*, vol. 57, no. 9, pp. 2203–2221, 2017, pMID: 28786670. [Online]. Available: http://dx.doi.org/10.1021/acs.jcim.7b00347

[13] D. Cappel, M. L. Hall, E. B. Lenselink, T. Beuming, J. Qi, J. Bradner, and W. Sherman, "Relative binding free energy calculations applied to protein homology models," *Journal of Chemical Information and Modeling*, vol. 56, no. 12, pp. 2388–2400, 2016, pMID: 28024402. [Online]. Available: http://dx.doi.org/10.1021/acs.jcim.6b00362

[14] A. d. Ruiter and C. Oostenbrink, "Extended thermodynamic integration: Efficient prediction of lambda derivatives at nonsimulated points," *Journal of Chemical Theory and Computation*, vol. 12, no. 9, pp. 4476–4486, 2016, pMID: 27494138. [Online]. Available: http://dx.doi.org/10.1021/acs.jctc.6b00458

[15] J. Ash and D. Fourches, "Characterizing the chemical space of ERK2 kinase inhibitors using descriptors computed from molecular dynamics trajectories," *J. Chem. Inf. Model.*, vol. 57, no. 6, pp. 1286–1299, 2017, pMID: 28471171. [Online]. Available: http://dx.doi.org/10.1021/acs.jcim.7b00048

[16] V. Gapsys, S. Michielssens, D. Seeliger, and B. L. de Groot, "pmx: Automated protein structure and topology generation for alchemical perturbations," *Journal of Computational Chemistry*, vol. 36, no. 5, pp. 348–354, 2015. [Online]. Available: http://dx.doi.org/10.1002/jcc.23804

[17] S. Doerr, M. J. Harvey, F. Noé, and G. De Fabritiis, "Htmd: High-throughput molecular dynamics for molecular discovery," *Journal of Chemical Theory and Computation*, vol. 12, no. 4, pp. 1845–1852, 2016, pMID: 26949976. [Online]. Available: http://dx.doi.org/10.1021/acs.jctc.6b00049

[18] A. Rizzi, P. B. Grinaway, D. L. Parton, M. R. Shirts, K. Wang, P. Eastman, M. Friedrichs, V. S. Pande, K. Branson, D. L. Mobley, and C. J. D, "YANK: A GPU-accelerated platform for alchemical free energy calculations. in preparation." *In preparation.*

[19] S. K. Sadiq, D. W. Wright, S. J. Watson, S. J. Zasada, I. Stoica, and P. Coveney, "Automated Molecular Simulation Based Binding Affinity Calculator for Ligand-Bound HIV-1 Proteases," *J. Chem. Inf. Model.*, vol. 48, no. 9, pp. 1909–1919, 2008.

[20] A. P. Bhati, S. Wan, D. W. Wright, and P. V. Coveney, "Rapid, accurate, precise and reliable relative free energy prediction using ensemble based thermodynamic integration," *J. Chem. Theory Comput.*, vol. 13, no. 1, pp. 210–222, 2017.

[21] I. Massova and P. Kollman, "Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies," *J. Am. Chem. Soc.*, vol. 121, no. 36, pp. 8133–8143, 1999.

[22] T. P. Straatsma and H. J. C. Berendsen, "Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations," *The Journal of Chemical Physics*, vol. 89, no. 9, pp. 5876–5886, 1988. [Online]. Available: https://doi.org/10.1063/1.455539

[23] S. Wan, B. Knapp, D. W. Wright, C. M. Deane, and P. V. Coveney, "Rapid, precise, and reproducible prediction of peptide–MHC binding affinities from molecular dynamics that correlate well with experiment," *J. Chem. Theory Comput.*, vol. 11, no. 7, pp. 3346–3356, 2015.

[24] S. K. Sadiq, D. W. Wright, O. A. Kenway, and P. V. Coveney, "Accurate ensemble molecular dynamics binding free energy ranking of multidrug-resistant HIV-1 proteases," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 890–905, 2010. [Online]. Available: http://dx.doi.org/10.1021/ci100007w

[25] D. W. Wright, B. A. Hall, O. A. Kenway, S. Jha, and P. V. Coveney, "Computing clinically relevant binding free energies of HIV-1 protease inhibitors," *J. Chem. Theory Comput.*, vol. 10, no. 3, pp. 1228–1241, 2014.

[26] S. Wan and P. V. Coveney, "Rapid and accurate ranking of binding affinities of epidermal growth factor receptor sequences with selected lung cancer drugs." *J. R. Soc. Interface*, vol. 8, no. 61, pp. 1114–1127, 2011. [Online]. Available: http://dx.doi.org/10.1098/rsif.2010.0609

[27] S. Wan, A. P. Bhati, S. Skerratt, K. Omoto, V. Shanmugasundaram, S. K. Bagal, and P. V. Coveney, "Evaluation and characterization of trk kinase inhibitors for the treatment of pain: Reliable binding affinity predictions from theory and computation," *J. Chem. Inf. Model*, vol. 57, no. 4, pp. 897–909, 2017.

[28] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable molecular dynamics with NAMD." *J. Comput. Chem.*, vol. 26, no. 16, pp. 1781–1802, 2005. [Online]. Available: http://dx.doi.org/10.1002/jcc.20289

[29] D. A. Case, V. Babin, J. T. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham, III, T. A. Darden, R. E. Duke, H. Gohlke, A. W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T. S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K. M. Merz, F. Paesani, D. R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, and P. A. Kollman, "Amber 14," University of California, San Francisco, 2014.

[30] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The Amber

biomolecular simulation programs." *J. Comput. Chem.*, vol. 26, no. 16, pp. 1668–1688, 2005. [Online]. Available: http://dx.doi.org/10.1002/jcc.20290

[31] B. R. Miller III, T. D. McGee Jr, J. M. Swails, N. Homeyer, H. Gohlke, and A. E. Roitberg, "MMPBSA. py: an efficient program for end-state free energy calculations," *J. Chem. Theory Comput.*, vol. 8, no. 9, pp. 3314–3321, 2012.

[32] N. H. Theodoulou, N. C. Tomkinson, R. K. Prinjha, and P. G. Humphreys, "Clinical progress and pharmacology of small molecule bromodomain inhibitors," *Current Opinion in Chemical Biology*, vol. 33, no. Supplement C, pp. 58 – 66, 2016, chemical genetics and epigenetics * Molecular imaging. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1367593116300758

[33] C.-w. Chung, H. Coste, J. H. White, O. Mirguet, J. Wilde, R. L. Gosmini, C. Delves, S. M. Magny, R. Woodward, S. A. Hughes, E. V. Boursier, H. Flynn, A. M. Bouillot, P. Bamborough, J.-M. G. Brusq, F. J. Gellibert, E. J. Jones, A. M. Riou, P. Homes, S. L. Martin, I. J. Uings, J. Toum, C. A. Clément, A.-B. Boullay, R. L. Grimley, F. M. Blandel, R. K. Prinjha, K. Lee, J. Kirilovsky, and E. Nicodeme, "Discovery and characterization of small molecule inhibitors of the bet family bromodomains," *Journal of Medicinal Chemistry*, vol. 54, no. 11, pp. 3827–3838, 2011, pMID: 21568322. [Online]. Available: http://dx.doi.org/10.1021/jm200108t

[34] P. Bamborough, H. Diallo, J. D. Goodacre, L. Gordon, A. Lewis, J. T. Seal, D. M. Wilson, M. D. Woodrow, and C.-w. Chung, "Fragment-based discovery of bromodomain inhibitors part 2: Optimization of phenylisoxazole sulfonamides," *Journal of Medicinal Chemistry*, vol. 55, no. 2, pp. 587–596, 2012, pMID: 22136469. [Online]. Available: http://dx.doi.org/10.1021/jm201283q

[35] C.-w. Chung, A. W. Dean, J. M. Woolven, and P. Bamborough, "Fragment-based discovery of bromodomain inhibitors part 1: Inhibitor binding modes and implications for lead discovery," *Journal of Medicinal Chemistry*, vol. 55, no. 2, pp. 576–586, 2012, pMID: 22136404. [Online]. Available: http://dx.doi.org/10.1021/jm201320w

[36] R. Gosmini, V. L. Nguyen, J. Toum, C. Simon, J.-M. G. Brusq, G. Krysa, O. Mirguet, A. M. Riou-Eymard, E. V. Boursier, L. Trottet, P. Bamborough, H. Clark, C.-w. Chung, L. Cutler, E. H. Demont, R. Kaur, A. J. Lewis, M. B. Schilling, P. E. Soden, S. Taylor, A. L. Walker, M. D. Walker, R. K. Prinjha, and E. Nicodème, "The discovery of i-bet726 (gsk1324726a), a potent tetrahydroquinoline apoa1 up-regulator and selective bet bromodomain inhibitor," *Journal of Medicinal Chemistry*, vol. 57, no. 19, pp. 8111–8131, 2014, pMID: 25249180. [Online]. Available: http://dx.doi.org/10.1021/jm5010539

[37] A. Wyce, G. Ganji, K. N. Smitheman, C.-w. Chung, S. Korenchuk, Y. Bai, O. Barbash, B. Le, P. D. Craggs, M. T. McCabe, K. M. Kennedy-Wilson, L. V. Sanchez, R. L. Gosmini, N. Parr, C. F. McHugh, D. Dhanak, R. K. Prinjha, K. R. Auger, and P. J. Tummino, "Bet inhibition silences expression of mycn and bcl2 and induces cytotoxicity in neuroblastoma

tumor models," *PLOS ONE*, vol. 8, no. 8, 08 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0072967

[38] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 8, pp. 1950–1958, 2010. [Online]. Available: http://dx.doi.org/10.1002/prot.22711

[39] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general Amber force field." *J. Comput. Chem.*, vol. 25, no. 9, pp. 1157–1174, 2004. [Online]. Available: http://dx.doi.org/10.1002/jcc.20035

[40] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, "Gaussian 03, Revision C.02," Gaussian, Inc., Wallingford, CT, 2004.

[41] S. Jha and M. Turilli, "A building blocks approach towards domain specific workflow systems?" 2017, (under review) http://arxiv.org/abs/1609.03484.