RAPTOR: Ravenous Throughput Computing

Andre Merzky[1], Matteo Turilli[1,2], Shantenu Jha[1,2]
[1] *Rutgers, the State University of New Jersey, Piscataway, NJ 08854, USA*
[2] *Brookhaven National Laboratory, Upton, NY 11973, USA*

2022 22nd International Symposium on Cluster, Cloud and Internet Computing (CCGrid) | 978-1-6654-9956-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CCGRID54584.2022.00069

***Abstract*—We describe the design, implementation and performance of the RADICAL-Pilot task overlay (RAPTOR). RAPTOR enables the execution of heterogeneous tasks—i.e., functions and executables with arbitrary duration—on HPC platforms, providing high throughput and high resource utilization. RAPTOR supports the high throughput virtual screening requirements of DOE's National Virtual Biotechnology Laboratory effort to find therapeutic solutions for COVID-19. RAPTOR has been used on 8300 compute nodes to sustain 144M/hour docking hits, and to screen 10^{11} ligands. To the best of our knowledge, both the throughput rate and aggregated number of executed tasks are a factor of two greater than previously reported in literature. RAPTOR represents important progress towards improvement of computational drug discovery, in terms of size of libraries screened, and for the possibility of generating training data fast enough to serve the last generation of docking surrogate models.**

I. Introduction

In response to COVID19, many researchers are using high-performance computing (HPC) to support epidemiological studies and to design anti-viral therapeutics. Significant effort has been invested in designing drug-discovery pipelines that can screen many more ligands than traditional *in-silico* drug-design approaches.

Nearly all high throughput virtual screening (HTVS) pipelines involve docking, i.e., the process of scoring a putative drug candidate (ligand) with a potential protein target. Docking algorithms are significantly cheaper but less accurate than full physics-based simulations to compute binding affinities between ligand and protease. The relative inexpensive scoring allows many more ligands to be investigated, which is necessary given the possible 10^{60} drug candidates.

One noteworthy COVID19 HTVS pipeline is the IMPECCABLE campaign [1], [2]—DoE's National Virtual Biotechnology Laboratory (NVBL) [3] effort to develop therapeutics for COVID-19. We discuss design, implementation and performance of RADICAL-Pilot task overlay (RAPTOR), which serves as the workhorse of the campaign's docking effort.

RAPTOR is a general purpose, portable, coordinator/worker framework for the execution of function and executable tasks on HPC platforms. RAPTOR is a subsystem of RADICAL-Pilot (RP) [4], and relies on it to acquire and manage resources, and to schedule and launch its coordinator and workers on those resources. RAPTOR extends RP's capabilities, providing: (1) steady utilization above 90% of the available resources with task executing for 1 second or longer; (2) partitioning of CPU and GPU resources across an arbitrary number of concurrently and/or sequentially executing batch jobs; and (3) partitioning of tasks across multiple, independent coordinators and workers.

To get a sense of the scale and impact: RAPTOR has been used on up to 466,816 concurrent CPU cores to sustain 144×10^6/hour docking hits [5] and to screen approximately 10^{11} ligands. To the best of our knowledge, both throughput rate and aggregated number of executed tasks are a factor of two greater than previously reported in literature [6]. RAPTOR was used to generate consensus and ensemble scoring against protein targets, and to generate training data for docking surrogate models [7], [8] that are up to 3–4 orders of magnitude faster than traditional docking programs [2]. Finally, RAPTOR has been used for more than 2M node-hours on primarily TACC Frontera and ORNL Summit, to support the identification of over 40 hits on COVID19 drug targets that are progressing to advanced testing [5], [9].

The main contributions of this paper are a description of the design and implementation of RAPTOR, and a performance evaluation of RAPTOR when used to perform computational docking at scale, as part of an HTVS pipeline. §II describes the HTVS use case and the state of the art infrastructure for it. §III discusses the design and implementation of RAPTOR, showing how it extends RP to support HTVS. §IV discusses experimental insight into the performance of RAPTOR for different workloads and configurations used to run HTVS.

II. High-Throughput Virtual Screening

HTVS is used in a variety of disciplines, from materials design [10] to drug discovery [11]. HTVS analyzes libraries of molecules, reducing them to a set of promising leads for experimental evaluation. Typically, HTVS follows a computational funnel (Fig. 1) to focus computational effort on promising molecules. Specifically, HTVS intelligently samples the large space of possible candidates and narrows the number of candidates down to an experimentally tractable set. HTVS is necessary for problems where exhaustive exploration is not an option, or for time critical options where random search is not acceptable.

In drug-discovery, HTVS enables rapid, low-cost screening of significantly larger and curated compound libraries, than feasible in experimental studies [11]. HTVS can now outperform equivalent experimental high throughput screening, and has been shown to rapidly identify tightly binding compounds. However, virtual libraries used in molecular discovery are often still too large to exhaustively evaluate, warranting the use of algorithms to help with exploration.

978-1-6654-9956-9/22/$31.00 ©2022 IEEE
DOI 10.1109/CCGrid54584.2022.00069

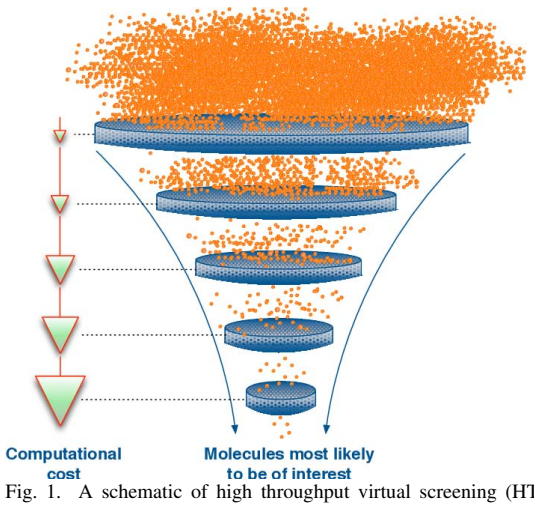


Fig. 1. A schematic of high throughput virtual screening (HTVS): Downstream stages are progressively computationally more expensive, but are focussed on increasingly promising candidates *Image source Ref. [10].*

Computational docking often forms the first stage in an HTVS pipeline. Docking fits trial drug-like compounds into (protein) binding sites in three-dimensional models of the protein targets characterized by a score. Docking is useful in early stages of molecular discovery to identify initial hits to be prioritized for experimental validation. This is true for both "regular" drug discovery pipelines, as well as for customized and AI-based ones.

A. HTVS Docking Infrastructure: State-of-the-Art

Several efforts created an open HTVS infrastructure, taking advantage of cloud platforms [6], [12] or HPC resources [13] to support large-scale ligand docking across various protein targets. Here we discuss four recent publications that represent the spectrum of performance considerations and design properties. Impressive results notwithstanding, given the diverse computing platforms and docking programs employed, as well as different measures of performance, it is difficult to provide a head-to-head performance comparison.

1) VirtualFlow [12] submits multiple "jobs" to different "clusters". VirtualFlow exhibited linear scalability with a peak at 160,000 CPUs on Google Cloud. VirtualFlow can dock 1B compounds in approximately 2 weeks, using 10,000 CPU cores simultaneously. In the context of COVID19, Ref. [12] investigate 17 virus-related targets, 45 screens, and $\approx 50 \times 10^9$ docking instances. However, they do not report how effectively the resources are utilized and focus on application performance measures (docks/time).

2) Ref. [13] outlines a supercomputer-driven pipeline for *in silico* drug discovery, using enhanced sampling molecular dynamics (MD) and ensemble docking. Ensemble docking makes use of MD results by docking compound databases into representative protein binding-site conformations, thus taking into account the dynamic properties of the binding sites. On Summit, that pipeline docked 1B compounds in 24 hours, using Autodock-GPU [14]. In Fig. 6 of Ref. [13], the authors outline the variation in docking time, of both GPU-based and "regular" CPU-based docking programs, observing fluctuations of 2x and 10x respectively.

3) Ref. [15], an ACM 2020 Gordon Bell Finalist, reported a peak performance of 20,000 docks/second on Summit, i.e., an aggregated performance of 50M docks/hour, for up to 20 poses per dock. This was primarily achieved through adaptation of AutoDock-GPU: GPU offloading feature calculations in re-scoring and database queries. A further 10x performance improvement was achieved by using parallel database methods. So far, 70M docks/hour is the highest reported throughput in literature.

4) The COVID-Moonshot project [16] used the folding@home approach in conjunction with high-throughput experimental screening, MD simulations and ML to identify covalent and non-covalent inhibitors against main protease (MPro) which demonstrated viral inhibition. Although folding@home is an "exascale" platform, the peak or sustained throughput is not precisely reported, nor is it easy to discern from available data. Based upon published science results, our best-effort estimate is that the COVID-Moonshot project screened 10×10^9 ligands over a period of several months, using steady state resources, thus about 100×10^6 docks/day over approximately 1000 CPUs.

B. RAPTOR for HTVS Docking

RAPTOR is used for the docking phase of DoE NVBL's IMPECCABLE campaign [1]–[3], [5], which integrates algorithmic and methodological innovations with advanced infrastructure to dock a large number of ligands with protein targets. A protein target represents a well-defined binding site, expressed as PDB file. For each target, we iterate through ligands from certain molecule libraries and compute a docking score for that ligand-protein pair. To increase the reliability of results, we used OpenEye and Autodock-GPU for the same ligand set and targets, which also allowed us to leverage HPC resource heterogeneity. We executed OpenEye on Frontera's x86 architectures and Autodock-GPU on Summit's GPUs. A docking call is executed either as a task function of the OpenEye Python library, or as an executable task launching AutoDock-GPU. In both cases, RAPTOR is used for orchestration.

Docking was used for consensus and ensemble scoring of large libraries [9], and to generate training data for docking surrogate models [7]. Several libraries, the largest of which (`mcule-ultimate-200204-VJL`) has 126M drug candidates, were used to dock against more than 100 targets. Other libraries include `Orderable-zinc-db-enaHLL` with 6.6M candidates, details of which can be found in Ref. [5]. Based upon library sizes, we estimate RAPTOR has been used to screen close to 100×10^9 molecules against over a dozen drug targets in SARS-CoV-2.

RAPTOR is a general-purpose, high-throughput task execution system that is not limited to a specific docking program or computing platform. This is in contrast to Refs. [12], [16] and Refs. [13], [15], [17], respectively. Since the docking programs are different, a direct comparison of scientific docking results are not meaningful, however RAPTOR reached a throughput

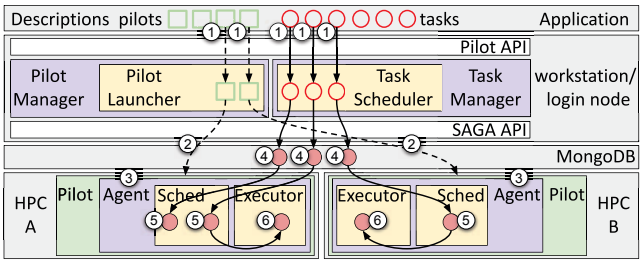


Fig. 2. RADIAL-Pilot (RP) architecture and execution model. RP is a distributed system that execute its components between the user's workstation or the HPC platform's login node, and the HPC platform's compute nodes. RAPTOR executes on the target HPC platform's compute nodes (see Fig. 3).

of 144M docks/h and performed a total of 100×10^9 docking tasks. These are both factor of two greater than previously reported in literature. RAPTOR achieved this on Frontera, using 8300 nodes at peak and with resource utilization above 90%. Further, different from VirtualFlow [6], [12], RAPTOR can manage the entire workload within a single large job, or spread across multiple jobs.

III. Design and Implementation

The RADICAL-Pilot Task OveRlay (RAPTOR) is a coordinator/worker framework for the execution of function and executable tasks on HPC platforms. It is designed to enable high-rate task execution at scale, e.g., up to $144 \times 10^6 tasks/h$ on TACC Frontera's 8300 nodes. RAPTOR is a component of RADICAL-Pilot (RP) [4] and relies on RP to acquire and manage resources, and to schedule and launch its coordinators and workers on those resources. RAPTOR is implemented in Python as is RP.

Fig. 3 shows RP's architecture and execution model. RP exposes an API to describe pilots and tasks [18], and uses four modules to manage them: PilotManager, TaskManager, Agent and DB [4]. Once described, pilots and tasks are passed to RP's runtime system ①. The PilotManager submits pilots on one or more resources via the SAGA API ②. The SAGA API implements an adapter for each supported resource type, exposing uniform methods for job and data management [19]. Once a pilot becomes active on a resource, it bootstraps the Agent module ③.

The TaskManager schedules each task to an Agent ④ via a queue on a MongoDB instance. This instance is used as the RP DB module to communicate task descriptions between the TaskManager(s) and the Agent(s). Each Agent pulls tasks from the DB module ⑤ and schedules ⑥ each task on an Executor upon resource availability (e.g., number of cores or GPUs). The Executor sets up the task's execution environment and then spawns the task for execution.

RP's tasks are fully-decoupled, i.e., they have no data dependences. Task dependences can be resolved before submission to RP via workflow engines, e.g., EnTK [20] and [21]. Each task is assumed to be self-contained, executed by RP as a black-box that returns success or failure codes. RP has no control or knowledge about the code each task executes, enabling the separation of concern among resource management, execution management and task executables. At application level, RP implements a 'batch-like' programming model to describe groups of tasks (i.e., workloads) and submit them for execution. Concurrency is implicit as users do not control it: RP executes tasks with the maximum concurrency allowed by the available resources.

RP is designed to schedule and launch executable tasks and not function tasks, i.e., tasks coded as functions in a specific programming language. Executable tasks are comprised of a self-contained program that can execute on the compute nodes' operating system. RP handles such tasks as an object containing its requirements, e.g., number of processes, type of process communication, type and number of CPU cores and/or GPUs, and so on. RP's tasks are relatively 'heavy' and require a certain time to be scheduled and launched. That limits RP's throughput and, ultimately, the efficiency at which RP can use resources with tasks shorter than 1 minute [4], [22].

Scheduling in RP is global: all the tasks that are submitted to RP's Agent are managed by a single scheduler. While the scheduling algorithm is tweaked to reach peaks of 350 tasks/s, its performance degrades for short running tasks on large resources (less than $\sim 60s$ for ~ 1000 nodes, $\sim 120s$ for ~ 2000 nodes, etc.).

RAPTOR extends RP to support the execution of tasks like those required by the COVID19 campaign described in §II. RAPTOR can: (1) execute both function and executable tasks; (2) achieve high throughput with arbitrary short running tasks; (3) arbitrarily partition resources and tasks; (4) use a multilevel scheduling in which workloads are partitioned and then subsets of tasks are scheduled to subset of resources; and (5) partition tasks across multiple, independent executors.

Fig. 3 shows how RAPTOR integrates within RP to enable the setup of the coordinator/worker infrastructure and how it launches and executes tasks on it. Due to RP's task model, scheduling and launching RAPTOR's coordinators and workers do not require additional capabilities: once bootstrapped, ① and ②, RP manages coordinators and workers as any other task ③. Once running, a coordinator schedules one or more workers on RP's Scheduler ④. Each worker is then launched on a compute node by RP Executor ⑤. Finally, the coordinator schedules function calls on the available workers for execution ⑥, load-balancing across workers as to obtain maximal resource utilization.

As a subsystem of RP, RAPTOR integrates with overall RP's capabilities. RP's TaskManager schedules and launches RAPTOR's coordinators and workers but also other executable tasks. In turn, RAPTOR can execute both function and executable tasks on the resources acquired by RP's PilotManager and allocated to RAPTOR's workers, independent of whether RP executes tasks on other resources. This is illustrated in Fig. 3 by showing RP executing a 32 CPU MPI task (green) on two compute nodes, while RAPTOR executes Python function calls of diverse sizes (yellow) on its workers (red).

RAPTOR's coordinators and workers manage the execution of tasks via several queues, depending on configuration pa-

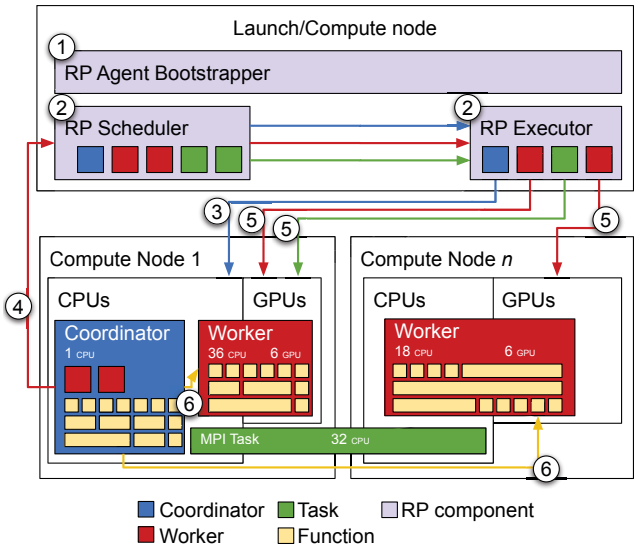


Fig. 3. RAPTOR: architecture and execution model.

rameters and performance requirements. A coordinator pushes tasks to a queue and N workers concurrently pull that queue for tasks to execute. The number of coordinators, queues and workers can be tuned so that the rate of (de)queuing does not exceed the capabilities of the queue implementation and of the used network. That keeps resources busy and avoid worker starvation. Both coordinator and worker are implemented in Python, using ZeroMQ to implement their queues.

RAPTOR's performance mainly depends on set up and management times of coordinators and workers. Five design choices improve performance: (1) launch coordinators and workers via MPI so to reduce latency, given the available technologies; (2) use a dedicated communication channel between each coordinator, its workers, and RP's scheduler; (3) partition resources across a user-defined number of coordinators and workers; (4) limit each worker to use at most one compute node; and (5) submit function tasks in bulk from a coordinator to its workers.

RAPTOR's design enables the implementation of workload control at coordinator level. RAPTOR coordinator's API consists of the `rp.raptor.coordinator` class and four main methods: `submit`, `start`, `join` and `stop`. Users inherit the class and initialize their coordinator with parameters like the workers' description (`dscr`), and the number of workers (`n_worker`), CPUs (`cpn`) and/or GPUs (`gpn`). Users then specify the workers' payload—either a Python function or an arbitrary executable—and, in case, callbacks to receive updates about the workers' status. Class methods are then used to submit the payload to the workers (e.g., `coordinator.submit(descr=descr, ...)`).

RAPTOR limits workers to a single node, making impossible to execute multi-node MPI Python function tasks. This is acceptable considering that RP would still be able to execute multi-node MPI executable tasks alongside the coordinators/-workers tasks. As discussed in the next section, the maximum number of coordinators and workers that RP can manage depends on the HPC platform's MPI performance. RAPTOR's throughput depends on the workload executed and on system capabilities such as shared file system performance, availability of caches, etc. RAPTOR enables resource-specific optimizations (e.g., using nodes' SSD storage) that are not necessarily portable due to hardware and system constraints. Currently, RAPTOR provides workers for Python functions and code snippets, and for arbitrary non-MPI executables. Prototypes of workers to execute multi-node MPI functions and executables are being tested.

IV. Performance Characterization

We characterize RAPTOR's performance and overheads, showing how it supports the execution of the workload described in §II at scale. We perform experiments with: production runs for NVBL-Medical Therapeutics campaigns on Frontera; runs for largest achievable size on both Frontera and Summit; and runs with both function and executable tasks. Tab. I summarizes the parameterization and results of each experiment. We measure: docking time in seconds (s); docking rate in docks/h; resource utilization; and RAPTOR set-up time in seconds.

Resource utilization measures the percentage of available CPU and/or GPUs used for docking operations. Resources become available as soon as the HPC platform's batch system schedules the job(s) submitted by RP. As such, resource utilization is a measure of how efficiently RP and RAPTOR use the available resources to execute the given workload. Extending the results presented in Ref. [1], tab. I provides two values for resource utilization: `avg` for the average utilization over the pilot runtime, and `steady` for the steady-state utilization. For the latter, we remove the contributions of startup and cool-down. We define startup as the time where the concurrency of tasks rises, and cool-down where the concurrency decreases.

We assign one pilot for each protein to which a set of ligands will be docked. Within each pilot, each coordinator will manage the workers defined via its interface (see §III). Each coordinator iterates at different strides through the ligands database, using pre-computed data offsets for faster access, and generating the docking requests to be distributed to the workers. Each worker runs on one node, executing docking requests across the CPU cores or the GPUs of that node.

A. Experiment 1

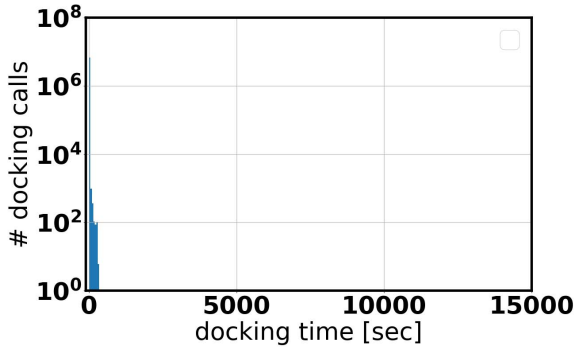
Experiment 1 performed the docking of 6.6×10^6 ligands—from the `Orderable-zinc-db-enaHLL` database—on 31 proteins. We used RP to acquire resources via 31 independent pilots, i.e., 31 jobs submitted to Frontera's batch system. We used 1 pilot for each protein and, for each pilot, we used RP to initialize RAPTOR's coordinators and workers. Each coordinator then managed the execution of the docking function tasks on its workers. Due to the different queue waiting times, at most 13 pilots executed concurrently in experiment 1, with a peak throughput of $\sim 17.4 \times 10^6$ docks/h.

The number of pilots used depends on the policies that govern: (1) the number of jobs that can be concurrently submitted

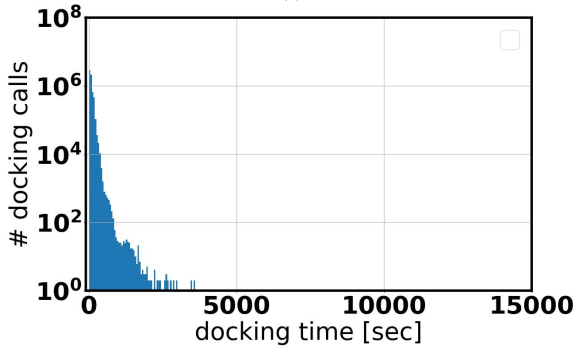
TABLE I

EXPERIMENTS. RAPTOR USES ONE PILOT FOR EACH PROTEIN, COMPUTING THE DOCKING SCORE OF A VARIABLE NUMBER OF LIGANDS TO THAT PROTEIN. OPENEYE AND AUTODOCK-GPU IMPLEMENT DIFFERENT DOCKING ALGORITHMS AND DOCKING SCORES, RESULTING IN DIFFERENT TASK TIMES AND RATES. RESOURCE UTILIZATION IS OFTEN IMPEDED BY THE LONG TAIL TASK TIME DISTRIBUTIONS WHICH CAUSE AN EXPENSIVE COOLDOWN PERIOD. HOWEVER, THE STEADY STATE RESOURCE UTILIZATION IS >=90% FOR ALL EXPERIMENTS.

ID	Platform	Application	Nodes	Pilots	Tasks [$\times 10^6$]	Startup [sec]	1st Task [sec]	Utilization	Task Time [sec]		Rate [$\times 10^6/h$]	
								avg / steady	max	mean	max	mean
1	Frontera	OpenEye	128	31	205	129	125	90% / 93%	3582.6	28.8	17.4	5.0
2	Frontera	OpenEye	7600	1	126	81	140	90% / 98%	14958.8	10.1	144.0	126.0
3	Frontera	OpenEye	8336	1	13	451	142	63% / 98%	219.0	25.3	91.8	11.0
4	Summit	AutoDock	1000	1	57	107	220	95% / 95%	263.9	36.2	11.3	11.1



(a)



(b)

Fig. 4. Experiment 1: Distribution of docking times with the (a) shortest and (b) longest average time out of the 31 proteins analyzed. The distributions of the docking times for all 31 proteins have a long tail.

to the batch job system; (2) the amount of resources a batch job can request; and (3) the maximal walltime allowed for each job. We used Frontera's normal queue for experiment 1 that allowed us up to 100 concurrent jobs, each with a maximum of 1280 nodes and 48h of walltime. For experiments 2 and 3, we used instead a single pilot as we had access to a special queue that spanned all the machine for 24 and 3 hours respectively. Importantly, RP and RAPTOR required no further coding to support those diverse running modalities but only the setting of some of their configuration parameters.

Earlier runs with a setup similar to experiments 1 encountered performance bottlenecks related to Frontera's shared file system which stalled the simulation progress. To keep an acceptable load on Frontera's shared filesystem, only 34 of the 56 cores available were used on each node in experiment 1. Also in this case, no coding was required but just changing a configuration parameter.

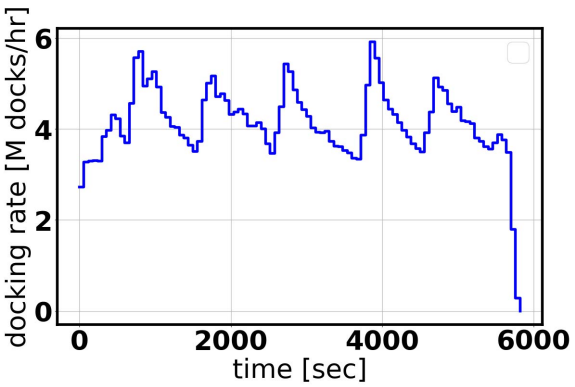
Figs. 4a and 4b show the distribution of docking times for proteins with the shortest and longest average docking time, using the `Orderable-zinc-db-enaHLL` ligand database. All proteins are characterized by long-tailed docking time distributions. Across the 31 proteins, the max/mean docking times are 3582.6/28.8s (Tab. I).

The large number of resulting docking requests ($31 \times 6.6 \times 10^6 = 205 \times 10^6$) poses a challenge to scalability due to the communication and coordination overheads. The long tail distributions necessitate load balancing across available workers to maximize resource utilization and minimize overall execution time. Thus, docking requests cannot be assigned statically to workers, but need to be dispatched dynamically.

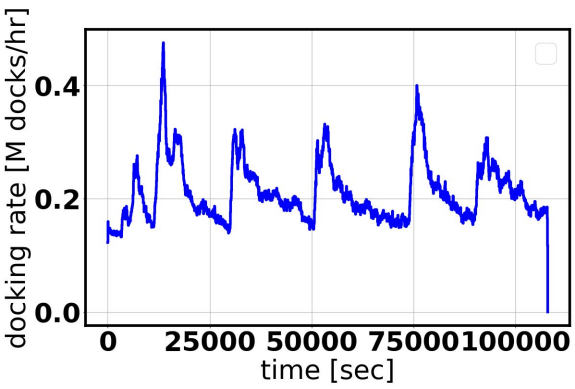
Consistent with RAPTOR's design, we addressed load balancing by: (i) communicating tasks in bulk so as to limit the communication frequency and therefore overhead; (ii) using multiple coordinator processes to limit the number of workers served by each coordinator, avoiding bottlenecks; (iii) using multiple concurrent pilots to partition the docking computations.

Figs. 5a and 5b show the docking rates for the pilots depicted in Figs. 4a and 4b, respectively. As with docking time distributions, the docking rate is similar across proteins. It seems likely that rate fluctuations depend on the interplay of machine performance, pilot size, and specific properties of the ligands being docked, and the protein. We measured a mean docking rate of 5×10^6 docks/h, with a max rate 17.4×10^6 docks/h when 13 pilots where executing concurrently, using about 20% of Frontera's resources (Tab. I).

Note that each pilot needs some time to launch coordinators and workers, and to begin distributing data and docking requests. Further, each pilot also needs some time to terminate and collect trailing results. That behavior is visible in all experiment plots. The respective overheads depend on the pilot size, and we will discuss them in more details for a larger setup in experiment 3. On Frontera, the time between when the pilot started and the first worker started to execute the first task was ~120s on average (Tab. I). That includes ~55s taken by the workers' coordinator to setup the execution and the time needed to prepare some of the input parameters of the docking functions. As the total execution time was between 1h:40m and 27h:46m (Fig. 5), setup time overhead was not relevant. Tab. I also provides resource utilization (as defined above) for the steady state between startup and cooldown, and as average for the full pilot lifetime.



(a)



(b)

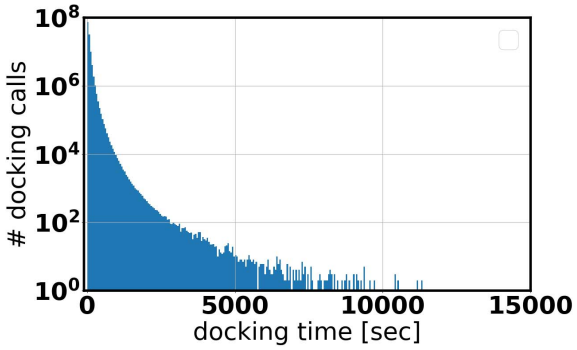
Fig. 5. Experiment 1: Docking rates for the protein with (a) shortest and (b) longest average docking time. The docking rates given in Tab. I are aggregated over concurrent pilots and thus larger than shown here for individual pilots.

B. Experiment 2

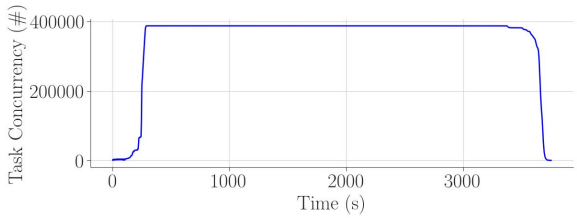
Experiment 2 characterizes the scalability of RAPTOR with a single pilot, spanning all available nodes on Frontera (about 1000 nodes were reserved for system work at the time of this run). To minimize the overhead caused by repeated loading of the receptor data into memory, the data were loaded once per node and then reused for all docking runs assigned to that specific node. The individual cores hosting the docking computations received cloned copies of the receptor data so as to isolate the individual docking computations.

To reduce the overhead of loading compound data from disk, the storage offsets in the dataset were precomputed at startup and staged to the compute nodes. Intermediate data were stored on node-local SSDs, further reducing the load on the shared file system. For the same reason, during startup we stored a static Python virtual environment with the OpenEye docking modules on the local storage of the nodes. Together, those improvements enabled the use of all the 56 cores of each compute node and required minor programming at application level. RP and RAPTOR enable that kind of performance tuning for every application written against their APIs and every HPC platform with local storage on the compute nodes.

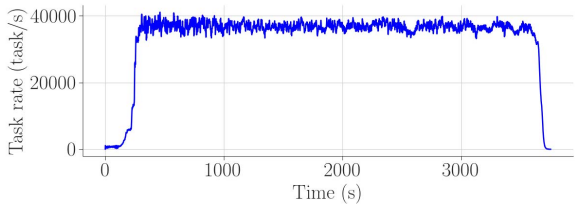
Fig. 6a shows the distribution of docking times of approximately 126×10^6 ligands from the `mcule-ultimate--200204-VJL` library to a single protein, using OpenEye on Frontera. Note that the distribution is highly dependent on



(a)



(b)



(c)

Fig. 6. Experiment 2: (a) docking time distribution; (b) docking concurrency; and (c) docking rate for a single protein and 126×10^6 ligands. Executed with 158 coordinators, each using ∼50 nodes/2800 cores on Frontera.

the protein being used: for the specific protein used in this run, we measured a max of 14985.8 seconds and a mean of 61.5 seconds (Tab. I). The set of proteins available to us varied in mean docking time from ∼3 to ∼70 seconds.

Fig. 6c shows the docking rate for a single pilot with 7650 compute nodes. Compared to experiment 1, the rate does not fluctuate over time and is consistently near $\sim 40 \times 10^3$ docks/s ($\sim 144 \times 10^6$ docks/h—see Tab. I). Note that the long tail distribution of runtimes results in a long tail of docking calls and causes the "cooldown" phase. That phase and the startup time ultimately lower utilization from 98.3% in the steady-state (before cooldown starts) to a total average of 90.0%. The time taken to create the task was reduced to 35s from the 55s required in experiment 1, mainly due to using the more efficient local storage. The time taken to execute the first task from when resources become available marginally increased to 140s compared to the 120s of experiment 1.

As discussed, the docking times depend on the proteins used. Thus, the docking rate inversely depends on that protein choice and, ultimately, not on RAPTOR's design and capabilities. The range of rates is very wide: for the proteins available to us, we observed a mean docking rate between $\sim 14 \times 10^6$ and $\sim 300 \times 10^6$, for runs of the same size.

C. Experiment 3

One of RAPTOR's distinguishing capability is executing function and executable tasks concurrently. In experiment 3, we launched executable tasks alongside the OpenEye docking function tasks, thus emulating an heterogeneous workload. Each executable task run the `stress` command.

As with experiment 2, we used only one pilot but with 8336 compute nodes, for a total of 466,816 cores. On that pilot, we launched 8 coordinators and each coordinator launched 1041 workers. As we use MPI to launch the workers, we created a total of 8328 ranks, utilizing all the available compute nodes (reserving 8 nodes for the coordinator's tasks). We tested larger numbers of coordinators/workers but without measurable improvements. That implies that the communication system is not a bottleneck in this setup. More experiments are needed but, currently, we believe we reached the limit of the platform's performance.

Executing experiments on whole machines of the size of Frontera requires special agreements and support. We worked with Frontera staff at TACC to use the whole machine for 3 hours after a maintenance period. Our runs uncovered issues with a switch, triggered two faults in the shared file system and ultimately overwhelmed the telemetry system of the machine. That reduced the time we had for each run to 1200s and reduced the number of runs we were able to perform.

With short runs, the startup and cooldown periods have a relatively large impact on the performance and utilization numbers presented here. Note that, in production, users run RAPTOR for up to the maximal walltime allowed—usually between 24 and 48 hours. RAPTOR startup time happens only once per run so, as far as startup costs few minutes, it will be negligible from a RAPTOR's efficiency point of view. On the contrary, filesystem slow downs and bottlenecks recur across the whole run, greatly affecting overall resource utilization. Thus, trading off slower startup time for better filesystem performance benefits the overall run efficiency.

Despite the optimizations done to reduce the load on the shared file system in experiments 2 and 3, most workers' task collection stalled for ∼150 seconds after running as expected for ∼800 seconds. That stall led to longer task runtimes visible in Fig. 7b, where several tasks run significantly longer beyond their nominal 60s cutoff time. The average utilization listed in Tab. I was lowered significantly due to the stalls. Our traces show no errors, delays or overloads and TACC telemetry service did not provide conclusive data.

The startup time in experiment 3 is non-negligible: 451s. That amount of time can be separated into 6 contributions: (1) pilot bootstrapping and (2) staging to node storage overlap, contributing 78s; (3) coordinator startup that contributes 1s; (4) input data pre-processing in the coordinators contributes 42s; (5) worker startup (all ranks) and (6) bootstrapping of communication system overlap, contributing 330s.

Fig. 7a shows a histogram of the startup times for all ranks, and one for the communication channel setup which the worker ranks can only initiate once the ranks are up. Interestingly, the first worker rank for each coordinator took only

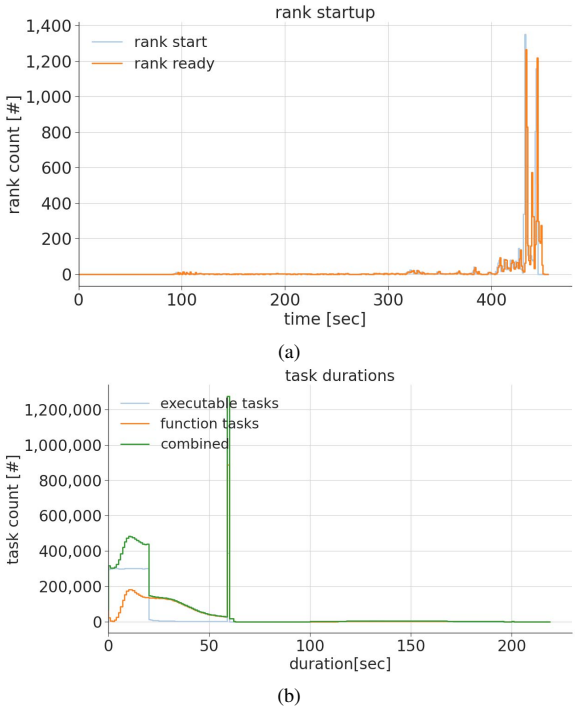


Fig. 7. Experiment 3: worker rank startup times.

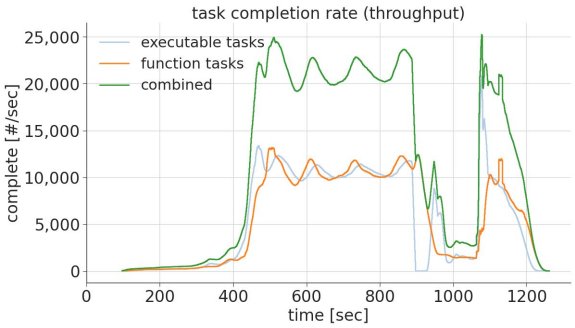
∼10 seconds to start, but the startup of the remaining ranks was significantly slower, with the last worker to come alive only after 330 seconds. These times depended on the performance of MPI on Frontera and it calls for further optimization.

As soon as each coordinator and its workers become active, they started executing bulks of 128 mixed function and executable tasks. The first worker began executing tasks 142s after the job started; the last worker however executed its first task at 368s, leading to a total ramp up time of 374s observed in Fig. 7a. Only then, RAPTOR could begin to utilize the full system.

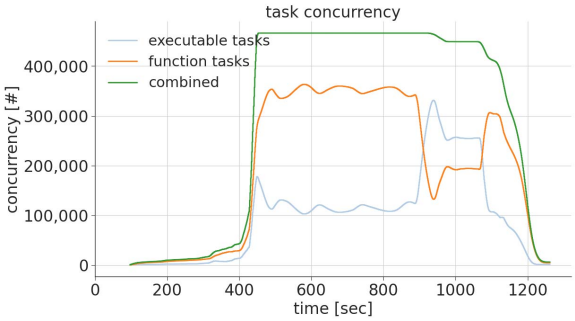
Fig. 7b shows the task runtime distribution for this run for 6,685,316 ligands from the `Orderable-zinc-db-enaHLL` library that are docked to the protein `3CLPro_6LU7_A_1_F` which is particularly relevant to study the binding of drugs to the spike of SARS-CoV-2. The figure shows durations between 3 and 60s and then a certain number of tasks that have been terminated at 60s. This is the threshold used by the scientists to determine when a ligand should be stopped to be computed, either because it would not be relevant or because, more commonly, the simulation stalled.

Fig. 7b also shows the distribution of the additional 6,685,316 **executable** tasks. We drew the tasks runtimes from a uniform distribution between 0s and 20s. Note that both distributions show several tasks running for longer than the 60s cutoff, up to a runtime of 360s. Those tasks were predominantly observed during the performance dip after 800s of runtime discussed above.

Fig. 8a shows the task completion rate for experiment 3. The rate shows a ramp up of ∼360s (see discussion above) followed by a peak of ∼25,000 tasks/s and an average of 22,000



(a)



(b)

Fig. 8. Experiment 3: task completion rate and task concurrency.

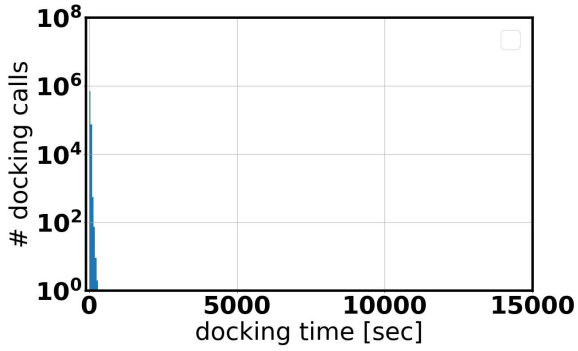
tasks/s for most of the remaining time of the run. This is equivalent to a peak $\sim 90 \times 10^6$ tasks/h and an average $\sim 79 \times 10^6$ tasks/h. The plot also shows the individual completion rate of the executable and function tasks. Both rate and behavior over time are comparable for function and executable tasks. After the ramp up phase, the execution rate peaks at ∼13,000 task/s to then stabilize at an average of ∼11,000 task/s.

As discussed, we observe some stalling at around 800s and then a rate peak of ∼25,000 task/s at ∼1000s that includes the tasks completed after the stalling period. We observe an cooldown phase in which the number of remaining tasks progressively decreases until no tasks are left to be executed. The consistency of behavior for function and executable tasks indicates that RAPTOR can concurrently execute both types of task in isolation, without affecting overall performance.

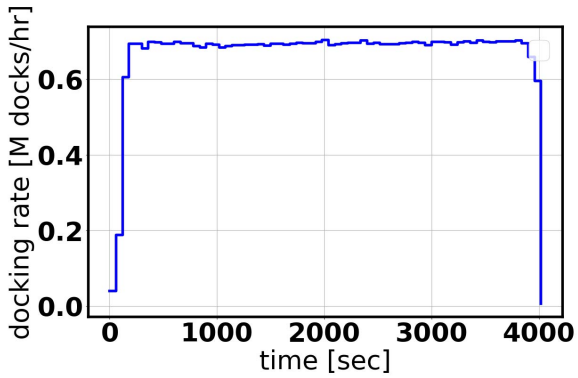
D. Experiment 4

Figure 9a shows the distribution of docking times of $\sim 57 \times 10^6$ ligands from the `mcule-ultimate-200-204--VJL` database, using AutoDock-GPU on Summit. The distribution has a max/mean of 263.9/36.2s (Tab. I). Compared to experiment 1, Fig. 4, max docking time is shorter, but the mean is longer. Compared to experiments 2 and 3, both max and mean are shorter. As observed, those differences are due to specific properties of the docked ligands and the protein.

Fig. 9b shows the docking rate for one pilot with 1000 nodes (6000 GPUs). Different from experiments 1, 2 and especially 3, the rate peaks very rapidly at $\sim 11 \times 10^6$ docks/h, showing a very short startup time. RAPTOR maintains that steady rate until the end of the execution, with a very rapid cooldown phase compared to the other experiments. We explain the observed sustained dock rate with an interplay between the scoring function and its implementation in AutoDock-GPU, and specific features of the 57×10^6 docked ligands.



(a)



(b)

Fig. 9. Experiment 3: (a) Distribution of docking time and (b) docking rate for a single protein and 57×10^6 ligands. A pilot is concurrently executed on Summit with 6000 GPUs.

Different from OpenEye on Frontera, AutoDock-GPU bundles 16 ligands into one GPU computation in order to efficiently use the GPU memory, reaching an average docking rate of 11.1×10^6 docks/h (Tab. I). Currently, our profiling capabilities allow us to measure GPU utilization with 5% relative error. Based on our profiling, we utilized between 93 and 98% of the available GPU resources.

Experiment 4 shows how RAPTOR can manage resource and executable tasks heterogeneities. Moving from Frontera to Summit required few changes in the pilot description and a new task description for AutoDock-GPU *executable* tasks instead of OpenEye *function* tasks. RAPTOR coordinator and worker also required minimal configuration changes to account for different number of resources available per node but, importantly, no change was required to manage executable tasks instead of function tasks. Both RP and RAPTOR are agnostic towards the type of code executed by each task.

Together, experiments 1–4 show how RAPTOR manages 6 types of heterogeneity: (1) different number of pilots per run; (2) different types of tasks concurrently executed on the same pilot, coordinator and worker; (3) different task runtimes; (4) different type of task executable; (5) different type of HPC platform; and (6) different type of computing resource, CPU cores and GPUs. Across these heterogeneities, RAPTOR manages to reach more than 90% peak resource utilization and

unprecedented scales for the docking calculations.

By supporting heterogeneity, flexible API and scale, RP and RAPTOR enable users to write more general-purpose and flexible applications, abstracting resource provisioning and management, and task scheduling. Ultimately, the goal is to avoid having to write special purpose applications that support a single use case on a specific platform and at a specific scale. This is what the COVID19 use case required: rapid deployment at scale on diverse platforms to efficiently and effectively leverage all the computational capacity across multiple institutions, to obtain results in the shortest possible amount of time.

V. Related Work

We discussed the performance and scale that RAPTOR achieves for the computational docking problem relative to similar efforts (Sec. §II) that represent the state-of-the-art. RAPTOR achieves at least a factor of two greater throughput on any platform than published results. We attribute this to the combination of the of the coordinator/worker and pilot paradigms [18] and their scalable implementations.

Coordinator/worker is one of the most common paradigms in distributed computing and programming [23]. Many classes of algorithms naturally fit the coordinator/worker paradigm, making it useful both for writing user-facing applications and scheduler components for middleware systems, especially in presence of heterogeneous resources and a dynamic runtime environment. Coordinator/worker is commonly adopted by middleware to enable the execution of many-task applications on distributed computing infrastructures [24].

In HPC, coordinator/worker is commonly used to coordinate the concurrent execution of processes and tasks across multiple resources and compute nodes. At programming level, coordinator/worker is used with MPI libraries [25], [26] and language extensions like Charm++ [27], [28] or COMPSs [29], to implement large-scale, single-executable applications. At task-level, diverse frameworks use workers coordinated by a coordinator to distribute and then execute tasks across HPC resources. For example, Dask [30], Parsl [21], Spark [31] and Arkouda [32] all use the coordinator/worker paradigm but many single-point solutions for domain-specific use cases also use coordinator/-worker [33]–[35].

While in this paper we present experiments specific to the problem described in Sec. §II, via the , RAPTOR supports the development of domain-independent applications with homogeneous and heterogeneous tasks. This is because the coordinator/worker paradigm is domain-independent and RAPTOR poses no constraints on the type of computation performed by the functions or executable of the workload. Further, RAPTOR supports extreme scale on HPC platforms with diverse architectures and resource usage policies.

VI. Discussion and Conclusions

HTVS pipelines are used in a variety of disciplines, ranging from materials design [10] to molecular design [11]. In particular, multi-scale biophysics-based HTVS pipelines are an important strategy for computational drug development. While HTVS can be considerably faster than experimental screening, until now, it has been too slow to explore libraries with billions of molecules, even on the fastest machines.

In this paper we describe the design and implementation of RAPTOR, and offer a performance evaluation of RAPTOR when used to perform computational docking at scale. Docking is often the first stage of multi-scale, biophysics-based HTVS pipelines and we report progress towards addressing the performance challenges of computational docking at scale. This, in turn, provides a path towards an overall improvement of throughput of computational drug discovery, in terms of size of libraries screened, and the possibility of integrating machine learning components with physics-based components.

RAPTOR extends the Pilot abstraction with the coordinator/worker paradigm, and offers a general-purpose, task-level application programming interface (API) to code distributed applications. RAPTOR supports: executing function and executable tasks; achieving high throughput and high resource utilization with arbitrary short running tasks; arbitrary partitioning of resources and tasks; multilevel scheduling in which workloads are partitioned and then subsets of tasks are locally scheduled to subset of resources; and partitioning of tasks across multiple, independent executors.

RAPTOR offers three main advantages compared to existing frameworks for multi-task applications: (1) users do not have to explicitly manage task concurrency when coding multi-task workloads; (2) user applications can execute up to 100×10^6 arbitrary Python functions on up to 500×10^3 cores and 24×10^3 GPUs; and (3) users can avoid coding resource management and task coordination. Additionally, RAPTOR abstracts away the notion of task concurrency from the users, while supporting unprecedented scale and managing 6 types of heterogeneity: (1) number of pilots per run; (2) types of tasks concurrently executed on the same pilot, coordinator and worker; (3) task runtimes; (4) type of task executable; (5) type of HPC platform; and (6) type of computing resource, CPU cores and GPUs. Overall, RAPTOR enables pilot-based execution of multi-task workloads on most DoE and NSF HPC platforms, independent of the type of executable launched by each task and, ultimately, of the use case supported.

RAPTOR allows users to implement rich control logic for their applications, expressed as an implementation of the coordinator/worker pattern. This will become increasingly important as emerging platforms offer progressively more heterogeneous resources and execution environments. Consistently, we plan to extend RAPTOR with workload management features such as: enacting failure management policies; making decisions based on state or output of tasks, both during runtime or after task completion; satisfy data dependencies, enabling both data and control flow management.

Acknowledgements We acknowledge Arvind Ramanathan, Rick Stevens, Austin Clyde (Argonne National Laboratory), and other members of the National Virtual Biotechnology Laboratory (NVBL) Medical Therapeutics group. We acknowledge computing time via the COVID19 HPC Con-

sortium. We thank TACC for the opportunity for scaling runs during the TexaScale days. This work is supported by NSF-1931512 (RADICAL-Cybertools), NSF-1835449 (SCALE-MS) and ECP CANDLE and ExaWorks. This research used resources at the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

REFERENCES

[1] H. Lee, A. Merzky, L. Tan, M. Titov, M. Turilli, D. Alfe *et al.*, "Scalable HPC and AI infrastructure for COVID-19 therapeutics," *Platform for Advanced Scientific Computing Conference (PASC '21), July 5–9, 2021, Geneva, Switzerland. ACM, New York, NY, USA*, 2021, https://arxiv.org/abs/2010.10517.

[2] A. A. Saadi, D. Alfe, Y. Babuji, A. Bhati, B. Blaiszik, T. Brettin *et al.*, "Impeccable: Integrated modeling pipeline for covid cure by assessing better leads," *50th International Conference on Parallel Processing (ICPP '21), August 9–12, 2021, Lemont, IL, USA. ACM, New York, NY, USA, 12 pages*, 2021.

[3] "National Virtual Biotechnology Laboratory (NVBL)," https://science.osti.gov/nvbl.

[4] A. Merzky, M. Turilli, M. Titov, A. Al-Saadi, and S. Jha, "Design and performance characterization of radical-pilot on leadership-class platforms," *IEEE Transactions on Parallel & Distributed Systems*, vol. 33, no. 04, pp. 818–829, apr 2022.

[5] A. Clyde, S. Galanie, D. W. Kneller, H. Ma, Y. Babuji, B. Blaiszik *et al.*, "High throughput virtual screening and validation of a sars-cov-2 main protease non-covalent inhibitor," *JCIM (in press)*, 2021.

[6] C. Gorgulla, A. Boeszoermenyi, Z.-F. Wang, P. D. Fischer, P. W. Coote, K. M. Das *et al.*, "An open-source drug discovery platform enables ultra-large virtual screens," *Nature*, vol. 580, no. 7805, pp. 663–668, 2020.

[7] A. Clyde, X. Duan, and R. Stevens, "Regression Enrichment Surfaces: a Simple Analysis Technique for Virtual Drug Screening Models," *arXiv e-prints*, p. arXiv:2006.01171, Jun. 2020.

[8] A. Clyde, T. Brettin, A. Partin, H. Yoo, Y. Babuji, B. Blaiszik *et al.*, "Protein-ligand docking surrogate models: A sars-cov-2 benchmark for deep learning accelerated virtual screening," *accepted for BMC Bioinformatics (2022)*, 2022. [Online]. Available: https://arxiv.org/abs/2106.07036

[9] Y. Babuji, B. Blaiszik, T. Brettin, K. Chard, R. Chard, A. Clyde *et al.*, "Targeting SARS-CoV-2 with AI-and HPC-enabled lead generation: A first data release," *arXiv preprint arXiv:2006.02431*, 2020.

[10] E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, and A. Aspuru-Guzik, "What is high-throughput virtual screening? a perspective from organic materials discovery," *Annual Review of Materials Research*, vol. 45, pp. 195–216, 2015.

[11] S. Zhang, K. Kumar, X. Jiang, A. Wallqvist, and J. Reifman, "DOVIS: An implementation for high-throughput virtual screening using AutoDock," *BMC Bioinformatics*, vol. 9, no. 1, pp. 1–4, 2008.

[12] C. Gorgulla, K. M. Das, K. E. Leigh, M. Cespugli, P. D. Fischer, Z. Wang *et al.*, "A multi-pronged approach targeting sars-cov-2 proteins using ultra-large virtual screening," *Iscience*, vol. 24, no. 2, p. 102021, 2021.

[13] A. Acharya, R. Agarwal, M. Baker, J. Baudry, D. Bhowmik, S. Boehm *et al.*, "Supercomputer-based ensemble docking drug discovery pipeline with application to covid-19," *Journal of chemical information and modeling*, 2020. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.jcim.0c01010

[14] D. Santos-Martins, L. Solis-Vasquez, A. F. Tillack, M. F. Sanner, A. Koch, and S. Forli, "Accelerating autodock4 with gpus and gradient-based local search," *Journal of Chemical Theory and Computation*, vol. 17, no. 2, pp. 1060–1073, 02 2021. [Online]. Available: https://doi.org/10.1021/acs.jctc.0c01006

[15] J. Glaser, J. V. Vermaas, D. M. Rogers, J. Larkin, S. LeGrand, S. Boehm *et al.*, "High-throughput virtual laboratory for drug discovery using massive datasets," *The International Journal of High Performance Computing Applications*, p. 10943420211001565.

[16] H. Achdout, A. Aimon, E. Bar-David, H. Barr, A. Ben-Shmuel, J. Bennett *et al.*, "Covid moonshot: open science discovery of sars-cov-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning," *bioRxiv*, 2020.

[17] J. V. Vermaas, A. Sedova, M. Baker, S. Boehm, D. Rogers, J. Larkin *et al.*, "Supercomputing pipelines search for therapeutics against covid-19," *Computing in Science & Engineering*, 2020.

[18] M. Turilli, M. Santcroos, and S. Jha, "A comprehensive perspective on pilot-job systems," *ACM Comput. Surv.*, vol. 51, no. 2, pp. 43:1–43:32, Apr. 2018. [Online]. Available: http://doi.acm.org/10.1145/3177851

[19] A. Merzky, O. Weidner, and S. Jha, "SAGA: A standardized access layer to heterogeneous distributed computing infrastructure," *Software-X*, 2015, dOI: 10.1016/j.softx.2015.03.001. [Online]. Available: http://dx.doi.org/10.1016/j.softx.2015.03.001

[20] V. Balasubramanian, M. Turilli, W. Hu, M. Lefebvre, W. Lei, R. Modrak *et al.*, "Harnessing the power of many: Extensible toolkit for scalable ensemble applications," in *International Parallel and Distributed Processing Symposium*. IEEE, 2018, pp. 536–545.

[21] Y. Babuji, A. Woodard, Z. Li, D. S. Katz, B. Clifford, R. Kumar *et al.*, "Parsl: Pervasive parallel programming in python," in *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, 2019, pp. 25–36.

[22] A. Merzky, M. Turilli, M. Maldonado, M. Santcroos, and S. Jha, "Using pilot systems to execute many task workloads on supercomputers," in *Workshop on Job Scheduling Strategies for Parallel Processing*. Springer, 2018, pp. 61–82.

[23] J.-P. Goux, J. Linderoth, and M. Yoder, "Metacomputing and the master-worker paradigm," in *Preprint MCS/ANL-P792-0200, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne*. Citeseer, 2000.

[24] J.-P. Goux, S. Kulkarni, J. Linderoth, and M. Yoder, "An enabling framework for master-worker applications on the computational grid," in *Proceedings of the 9th International Symposium on High-Performance Distributed Computing*. IEEE, 2000, pp. 43–50.

[25] M. Rynge, S. Callaghan, E. Deelman, G. Juve, G. Mehta, K. Vahi, and P. J. Maechling, "Enabling large-scale scientific workflows on petascale resources using mpi master/worker," in *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond*, 2012, pp. 1–8.

[26] R. Reussner, P. Sanders, L. Prechelt, and M. Müller, "Skampi: A detailed, accurate mpi benchmark," in *European Parallel Virtual Machine/Message Passing Interface Users' Group Meeting*. Springer, 1998, pp. 52–59.

[27] L. V. Kale and S. Krishnan, "Charm++ a portable concurrent object oriented system based on c++," in *Proceedings of the eighth annual conference on Object-oriented programming systems, languages, and applications*, 1993, pp. 91–108.

[28] A. Langer, R. Venkataraman, U. Palekar, L. Kale, and S. Baker, "Performance optimization of a parallel, two stage stochastic linear program," in *2012 IEEE 18th International Conference on Parallel and Distributed Systems*. IEEE, 2012, pp. 676–683.

[29] J. Conejero, S. Corella, R. M. Badia, and J. Labarta, "Task-based programming in compss to converge from hpc to big data," *The International Journal of High Performance Computing Applications*, vol. 32, no. 1, pp. 45–60, 2018.

[30] M. Rocklin, "Dask: Parallel computation with blocked algorithms and task scheduling," in *Proceedings of the 14th python in science conference*, vol. 126. Citeseer, 2015.

[31] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, I. Stoica *et al.*, "Spark: Cluster computing with working sets." *HotCloud*, vol. 10, no. 10-10, p. 95, 2010.

[32] M. Merrill, W. Reus, and T. Neumann, "Arkouda: interactive data exploration backed by chapel," in *Proceedings of the ACM SIGPLAN 6th on Chapel Implementers and Users Workshop*, 2019, pp. 28–28.

[33] Y. Guo, W. Bland, P. Balaji, and X. Zhou, "Fault tolerant mapreduce-mpi for hpc clusters," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2015, pp. 1–12.

[34] S. R. Young, D. C. Rose, T. Johnston, W. T. Heller, T. P. Karnowski, T. E. Potok *et al.*, "Evolving deep networks using hpc," in *Proceedings of the Machine Learning on HPC Environments*, 2017, pp. 1–7.

[35] I. Laguna, D. F. Richards, T. Gamblin, M. Schulz, B. R. de Supinski, K. Mohror, and H. Pritchard, "Evaluating and extending user-level fault tolerance in mpi applications," *The International Journal of High Performance Computing Applications*, vol. 30, no. 3, pp. 305–319, 2016.