Adaptive Ensemble Refinement of Protein Structures in High Resolution Electron Microscopy Density Maps with Flexible Fitting via Ensemble Toolkit

Daipayan Sarkar**,*,†,‖ Hyungro Lee**,‡ John W. Vant,¶ Matteo Turilli,§,⊥ Josh V. Vermaas,† Shantenu Jha,*,§,⊥ and Abhishek Singharoy*,¶

†*MSU-DOE Plant Research Laboratory, East Lansing, MI, USA*

‡*Pacific Northwest National Laboratory, Richland, WA, USA*

¶*School of Molecular Sciences, Arizona State University, Tempe, AZ, USA*

§*Rutgers University, Electrical & Computer Engineering, New Brunswick, NJ, USA*

‖*School of Molecular Sciences, Arizona State University, Tempe, AZ, USA*

⊥*Brookhaven National Laboratory, Computational Science Initiative, Upton, NY, USA*

E-mail: sarkarda@msu.edu; shantenu.jha@rutgers.edu; asinghar@asu.edu

***D.S. and H.L. contributed equally towards this manuscript. *Contact authors.*

Abstract

Recent advances in cryo-electron microscopy (cryo-EM) have enabled modeling macromolecular complexes that are essential components of the cellular machinery. The density maps derived from cryo-EM experiments are often integrated with manual, knowledge-driven, and physics-guided computational methods to build, fit, and refine molecular structures. Going beyond a single stationary-structure determination

scheme, it is becoming more common to interpret the experimental data with an ensemble of models, which contributes to an average observation. Hence, there is a need to decide on the quality of an ensemble of protein structures on-the-fly, while refining them against the density maps. We introduce such an adaptive decision making scheme during the molecular dynamics flexible fitting (MDFF) of biomolecules. Using RADICAL Cybertools, and the new RADICAL augmented MDFF implementation (R-MDFF) is examined in a high-performance computing environment for refinement of two protein systems, Adenylate Kinase and Carbon Monoxide Dehydrogenase. The use of multiple replicas in flexible fitting with adaptive decision making in R-MDFF improves the overall quality of the fit and model by 40 % relative to the refinements of the brute-force MDFF. The improvements are particularly significant at high, 2 - 3 Å map resolutions. More importantly, the ensemble model captures key features of biologically relevant molecular dynamics that is inaccessible to a single-model interpretation. Finally, this pipeline is applicable to systems of growing sizes, with the overhead for decision making remaining and robust to computing environments.

1 Introduction

Integrative modeling is an area of rapid methodological developments, wherein, atom-resolved structure(s) of biological systems are determined by merging data from multiple experimental sources with physics [1–3] and informatics-guided approaches. [4] These elegant fitting, [1–3,5–10] learning [11] sand inferencing [12–16] methodologies have been successful in resolving a range of structures, starting with soluble and membrane proteins up to sub-cellular complex architectures. [17,18] Integrative models routinely make it to top positions at the EMDB and PDB competitions, serving a diverse cross-section of the Biophysics community. [19] Advances in protein structure modeling, [20–23] evolutionary covariance or multi-sequence alignments offer excellent constraints for initiating such hybrid pipelines. [24]

A key issue in integrating structural or biochemical information with simulations stems from the heterogeneity of the data. The data quality can be spatially variant, spanning anywhere between coarse-grained to near-atomistic level of details. As a natural consequence of this heterogeneity, a single-model interpretation of the experimental data becomes implausible, opening the door to an ensemble treatment of the data. [25] The ensemble models capture on one hand, the most probable interpretation of the data, while on the

other, pinpoints rare-events and hidden conformations. Biology often employs such conformational diversity in problems of allostery and recognition, motivating the refinement of experimental knowledge against molecular ensembles.[14]

Another advantage of the ensemble interpretation is that, the generation of multiple independent atomic models using an EM density and statistical analysis of their map-model agreement offers metrics of global as well as local EM map quality.[26] This ensemble approach offers essentially both a quantitative and qualitative assessment of the precision of the models and their representation of the density.

The size of the ensembles that collectively describes the diversity in single-particle images (reflecting in the quality of the maps), however, grows non-linearly with system-size.[27] For proteins of molecular mass 500 kDa or bigger, composed of 5000 residues or more, a single CPU is expected to take 5000 years of wall-clock time for sampling the conformational ensembles using either brute force molecular dynamics (MD) or Monte Carlo simulations;[28] even the fastest GPUs of the day will not rescue this situation. Alternatively, data-guided enhanced sampling methodologies, such as MELD[13] (integrated with NAMD via the recently completed CryoFold plugin[14]) or backbone tracing methodologies such as MAINMAST[29] and analogous methods,[11] by themselves, either remain system-size limited, generating ensembles for only local regions within a map, or require further refinements using conjugate gradient minimization or free-energy schemes[30] to determine thermodynamic ensemble. As a step towards addressing this issue, by leveraging classical force fields (so-called CHARMM[31] energy functions) we have developed a range of molecular dynamics flexible fitting (MDFF) methodologies for integrating x-ray and cryo-EM data with MD simulations.[1–3,32] The simulations are biased towards conforming molecular models into forms consistent with the experimental density maps. These protocols are available through MD simulation engine NAMD,[33] recently to GROMACS c[6] and are also expanded as plugins, such as ISOLDE[34] in ChimeraX.[35,36] As a natural outcome of this fitting procedure, the most probable data-guided models are derived, e.g. for complex systems like the ribosomal machinery, virus particles and membrane proteins.[18] However, the conformational heterogeneity that contributes to the uncertainty of the experimental data is lost.

In this article we explore whether, it is possible to recover portions of the conformations lost in brute-force MDFF by running multiple replicas of MDFF in parallel with 'adaptive decision making'. Rather than physically enforcing a model into a map, this approach skews the probability of an ensemble of models towards maximizing their consistency with the map. This way, there remains a finite probability of visiting several uncertain structures, while still emphasizing determination of the most probable molecular models.

Traditional high-performance computing (HPC) approaches fail to make data-driven decisions within a multi-replica ensemble modeling workflow. We employ the RADICAL-Cybertools, [37] and in particular Ensemble Toolkit (EnTK),[38] to overcome this challenge of developing multi-replica MDFF as a workflow appli-

cation. Herein, EnTK deploys an application programming interface (API) for casting the MDFF simulation and analysis workflow as a hierarchy of pipelines, stages and tasks. Simultaneously, the RADICAL-Pilot (RP)[39,40] is employed as the high-performance and dynamic resource management layer. This workflow identifies all the flexible fitting tasks within a pipeline, acquires heterogeneously distributed resources to complete multiple parallel pipelines, and manages the overall execution of the stages iteratively.

A classical approach involves long brute-force MD simulations that are often stuck in local minima, requiring additional steps to find interesting regions of the conformational search space. In contrast, adaptive sampling implements an iterative loop that concurrently executes multiple simulations, each with short simulation time.[41–43] Map-model metrics are analyzed at every iteration - this analysis increases the probability of finding models that are consistent with the data, reducing the possibilities of getting trapped in any local energy minimum. The decision to enhance the sampling of specific models can be based a number of map-model metrics,[19] such as TM scores,[44] MolProbity,[45] EMRinger,[46] Q-score.[47] In our first proof-of-concept adaptive MDFF workflow, a simple global correlation coefficient (CC) is employed as a criteria to guide the choice of refinement models; Molprobilty statistics are employed for cross-validation.

Outlined in **Figure 1**, Workflow application using R-MDFF composes individual simulations and supports analysis calculation on intermediate results to perform adaptive sampling. The scheme iteratively screens model populations based on their CCs with the map, and improves efficiency of computing resource consumption over longer simulations. We find that, powered by EnTK's data-staging capabilities and checkpointing of the parallel MD simulations available on NAMD, MDFF trajectories intermittently screened by CC values offer an ensemble of refined models. We have tested the adaptive MDFF workflow with up to 100 replicas, each encompassing 16 iterations across resolutions of 2 Å to 5 Å, and achieved around 40 % improvement in map-model fitting over a long brute-force simulation. Similar efficiencies are noted while comparing the adaptive workflow with multi-replica, yet non-adaptive implementations of MDFF. The pipeline is further tested for using up to 400 replica with (1 node/replica). In all these cases, we find that an ensemble approach with adaptive decision making offers more diverse ensembles than brute force MDFF. Thus, going beyond traditional MDFF, these ensembles capture on one hand, the 'best' model, while simultaneously the uncertainty in the assignments on the other. Remarkably, the performance of the workflow improves with system-size (3341 atoms in Adenylate Kinase and 11452 atoms in Carbon Monoxide Dehydrogenase atoms), and remains robust to computing platforms. Taken together, our implementation breaks free of the traditional high-performance computing execution model that assumes singular jobs and static execution of tasks and data, to one that is fundamentally designed for data-integration and assimilation across different scales, quality and sparsity. The cryo-EM community has actively sought ways of extracting not just stationary structures, but ensembles and more importantly, molecular dynamics information from electron density data

are ever-increasing.[27,48–50]

2 Adaptive Integrative Modeling using R-MDFF

Adaptive decision making for the refinement of multiple protein structures coupled to 3-dimensional (3D) electron density maps is implemented as an iterative simulation-analysis workflow enabled by EnTK (**Figure 1** and **Algorithm 1**). Within R-MDFF, a workflow is defined as an ensemble of *simulation + analysis* pipelines that synchronously execute on HPC resources. Each constituent pipeline enable seven serial tasks: **(1)** load an empirically determined density map or generate a simulated map. Then convert this map to an MDFF potential, Eq. 1 and 2. Independently, examine the quality of an initial search model in terms of stereochemical properties, and perform rigid body docking to place this search model inside the EM density map, e.g. using Chimera [36] or Situs.[51] **(2)** Define the secondary structure restraints. Visual Molecular Dynamics (VMD)[52] then prepares the input files required by NAMD[33,53] to deploy MDFF. Multiple replicas of the system are prepared, under the same R-MDFF pipeline, as shown in **Figure 1**. Then, the ensemble of MDFF simulations are performed in parallel. **(3)** VMD's scripting interface is re-used to calculate the interim cross-correlation value between the atomic models from each replica and the EM density map. The CC values are extracted from VMD log files for different replicas are then combined to construct a matrix. EnTK uses a data staging area to move this matrix from flexible fitting to the adaptive decision-making block across multiple replicas. **(4)** Here, a decision is made on whether the flexible fitting simulations will continue or terminate, based on whether the computed CC is greater than or equal to the user-defined threshold. This on-the-fly map-model analysis enables an adaptive flexible fitting algorithm to run recursively inside EnTK, without user intervention. When the threshold CC is not met at the end of tasks 1-4, subsequent iterations are performed, wherein **(5)** all replicas are reseeded with the atom coordinates, velocities, and periodic system information corresponding to MDFF model with the best CC from the previous iteration, and the next round of multi-replica MDFF proceeds. If the map correlation of the best-fitted model decreases along the forward iteration, the new poorly fitted starting conformation is accepted with a weight $min\ (1, e^{\Delta E_{CC}/KT})$. Here, we follow $\Delta E_{CC} = k(CC_{N+1} - CC_N)$ for iteration no. N and $k = 5 \times 10^5$ kJ/mol.[6] For a failed move, the fitting restarts with the initial conformations of the last iteration, and the criterion is reused to find a new starting structure. **(6)** Again, EnTK uses data staging area to store these information in files and provide them to the replicas. This feature not only makes the algorithm adaptive, but offers future scope of improvement for applications requiring advanced decision-making, either based on inferencing[12–14,54,55] or neural network based machine learning algorithms.[56–60] Finally, the application converges to yield a refined ensemble **(7)**, which exit the R-MDFF workflow and downloads results to the end-user's working directory.

Algorithm 1: R-MDFF scheme with adaptive decision making based on CC.

```
begin
    perform rigid body docking of protein atomic structure in EM density map
    generate N replicas with this initial coordinate
    while CC replica resolution ≤ CC threshold do
        generate N replicas with initial coordinates from highest CC ensemble to EM density map
        repeat simulation stage (selected coordinates, density map)
        repeat analysis stage (best cc coordinates, replica index)
        increase iteration by 1
        end
    refined protein ensemble in end-user's working directory
```

The R-MDFF API is implemented as a Python module, loaded into the workflow application's code.[61] The API exposes classes for pipeline, stage, and task, allowing one to directly map the workflow description to the logical representation of an ensemble of simulations. Each task object exposes a set of variables with which to configure input, output files, executable, resource requirements, and pre/post execution directives. Finally, an appmanager object is used to contain the workflow description and execute it with a single `AppManager.run()` method. The iteration logic to change the workflow description and issue another `AppManager.run()` is written in pure Python as part of the workflow application. The entire R-MDFF workflow application of this paper required only 500 lines of Python code.[61]

As already described in Balasubramanian et al.[38], EnTK complements the ensemble simulation paradigm with decision making through real-time workflow and parameter changes, based on the results of the analysis stages. In the present context, this feature enables iterative workflow executions with a single HPC batch-job submission, avoiding costly manual evaluation of cross-correlation coefficient, workflow editing, and re-submission, as demonstrated here for flexibly fitting biomolecules in cryo-EM density maps. EnTK also abstracts from the users the need to explicitly manage data flow and task execution. It manages data staging so that each task of each stage has either a copy or a link to all the NAMD input files it requires, allowing the users to focus on the MDFF simulation and VMD analysis methods, without having to explicitly handle data sourcing, saving, and exchange. Furthermore, EnTK schedules and executes the workflow's tasks, managing the mapping of tasks to available resources on each compute node allocated to the workflow execution. Users only have to specify the amount of CPU cores/GPUs needed by each task and whether the task is (Open)MPI.

3 Results and Discussion

We conducted a series of experiments using R-MDFF by varying the number of replicas and iterations, and length of flexible fitting simulations. With these parameters, we compared the quality of the refined models for two example systems, namely, adenylate kinase (ADK) and carbon monoxide dehydrogenase (CODH) proteins. The robustness of the protocol is demonstrated on two HPC facilities, namely on Oak Ridge Leadership Computing's Summit, each node on which has two IBM Power9 processors and six NVIDIA V100 GPU accelerators, and on Pittsburgh Supercomputer Center's Bridges2 having two AMD EPYC 7742 processors per node.

3.1 Adaptive decision making using R-MDFF provides variance in ensemble refinement of high-resolution density maps

We start by fitting the 'closed' conformation of ADK (PDB: 1AKE) into synthetic density maps derived from its 'open' form (PDB: 4AKE). In **Figure 2** the CC changes are presented as a function of iterations for ensemble members with 4, 8, 16 and 32 replicas, and at map resolutions of 1.8, 3 and 5 Å. Using the utility `Phenix.maps` and the reported structure factors, ADK density maps were constructed at the native 1.8 Å resolution, and were further truncated at intermediate (3 Å) and low (5 Å) map resolutions. The product of number of replicas ($M_{replicas}$) × number of iterations ($N_{iterations}$) × the flexible fitting MD steps per iteration and per replica ($t_{steps/iteration/replica}$) which results in 16 nanoseconds (ns) of sampling.

For intermediate and low resolutions, the CC peak is ≈ 0.9, while at high resolution this value shifts to ≈ 0.8. While the former took one to two iterations only, the later took up to ten iterations to reach the highest CC values. The determination of the best-fitted model is almost independent of the number of replicas used, barring at 1.8 Å where a small increase in peak CC from 0.78 to 0.80 is observed when $M_{replicas}$ doubled from 4 to 8. As expected, the change in the quality of fit from 0.9 (against the 3.0 Å and 5.0 Å maps) to 0.8 (against the 1.8 Å maps) stems from the many conformations of a model that can match intermediate to low resolution density, the number of which decreases at higher resolutions.

In **Figure 2**, as the replica count further increases to 16 and 32, the distribution of CC traces across iterations becomes wider for a high-resolution density map. Interestingly, the best CC values at the end of every iteration does not linearly increase with iterations, but rather show fluctuations with an overall increasing trend. For example, in iterations 3 to 5 of the 32-replica refinement the CC's decrease, suggesting that the E_{CC}-selection criterion does not uniquely bias the distribution towards higher CC values, offering a probability to sample also the poorly fitted structures. This wider distribution implies conformational

diversity in the ensemble of refined protein structures. Conventionally, MDFF generates protein structures with low variance and high bias towards maximizing correlation with the target density. At higher resolutions, the population of the structures is further skewed. Even with the use of a relatively small number of R-MDFF replicas ($M_{replicas} = 32$) encoded through EnTK, the workflow generates a range of structures with CCs between 0.78 to 0.82. On the contrary, a single long MDFF trajectory with identical cumulative time of 16 ns produces final models strongly peaked at CC ~ 0.67. Therefore, unlike conventional flexible fitting, the R-MDFF workflow generates models that represent different propensities for large-scale protein conformations, still including the most probable model with an improvement of 22% over a single long MDFF.

3.2 Statistics of map-model fits improve with larger replica simulations

An improvement in ADK fit quality on a single long MDFF trajectory, particularly at 1.8 Å resolution, motivated further exploration of fit quality as a function of $M_{replicas}$. First, we have repeated the $M_{replicas} = 64$ computations without the adaptive decision block in R-MDFF. The most probable CC values decrease to 0.52 compared to 0.82 determined by executing the entire workflow, see **Figure 3**. This significant difference of 40% in the fit of the model to map implies that it is not just the sampling of an increased number of replicas, rather the adaptive re-initialization strategy employed every iteration successfully improves the simple MDFF results.

Next, it is examined whether increasing the simulation length per iteration for every replica, i.e. $t_{steps/iteration/replica}$ has an effect on improving the quality of fit for larger number of replicas. From **Figure 4**, one notice that the CC value increases and then drops and forks as the $M_{replicas}$ increases beyond 32. This is expected, since for higher $M_{replicas}$ the $t_{steps/iteration/replica}$ is decreased to conserve the cumulative simulation length as that of a single MDFF trajectory. To address this issue, for replicas 64, 100, 200, 400 we increase the simulation length per iteration for each replica to match that of the 16-replica workflow. We chose the $t_{steps/iteration/replica} = 16$ ns from the $M_{replicas}$ =16 setup, as that provides some rare yet high CC peaks.

As $t_{steps/iteration/replica}$ increased from 20 ps/iteration/replica to 80 ps/iteration/replica, for $M_{replicas}$ = 64, to match the simulation length for $M_{replicas} = 16$ from **Figure 4**, the CCs improved systematically by 11% from 0.72 to 0.82 (**Figure 5**). More importantly, at the higher values of $M_{replicas}$, e.g. 100, 200, and 400, a broad, and in fact, bimodal distribution of refined models is derived, while maintaining the same $t_{steps/iteration/replica} = 80$ ps/iteration/replica.

Models with high CC around 0.8 are expected given the inherent bias of the density data within MDFF simulations. However, the distribution of models isolated with statistical significance, and lower yet still

decent CC values between 0.7 to 0.8, were obscured by single, smaller-replica or non R-MDFF jobs. The quality of these structures was determined employing a MolProbity[45] analysis of all members of the generated protein ensemble. Despite a broad distribution in the quality of fit, the quality of model remains universally high as seen through MolProbity scores peaked between 0.75 - 1.75, see SI figures, **Figures S3 - S7**. Thus, the multi-model description inferred from the 1.8 Å ADK density map remain energetically viable conformations of the protein that remain in the vicinity of the best-fitted model, but with variations that can reflect the dynamics of ADK.

In the open state of ADK, its so-called "lid" domain undergoes a hinge-like movement[30,62–64] to maintain a conformational pre-equilibrium with the closed state. Such movements are confirmed by transition path sampling simulations and FRET experiments.[65,66] Illustrated in **Figure 5**, principal components from the ensemble of converged R-MDFF models collated across 400 replicas clearly captures this hinge movement of "lid" opening and closing essential towards the biological function catalyzes the interconversion of the various adenosine phosphates (ATP, ADP, and AMP). However, a skewed distribution of the just the best-fitted models is obtained with $M_{replicas} = 16$, indicating only the "lid" open state. Thus, by using a probabilistic selection criterion within R-MDFF, rather than a deterministic one used conventionally with data-guide simulations such as MDFF, the space of CCs is more exhaustively sampled during flexible fitting, and the evolution of an ensemble can be monitored to gain insights on the structure-function relationship for biomolecules.

3.3 Refinement protocol is robust to system size

Robustness of the newly implemented R-MDFF parameters estimated from our multi-replica ADK simulations is tested using a second example of larger protein, namely carbon monoxide dehydrogenase. The closed conformation (PDB - 1OAO: chain C) was used as the search model, while the open state (1OAO: chain D) was the target. Similar to ADK, the fitting was performed with maps of the reported 1.9 Å and synthetically reduced 3.0 Å resolutions. After reconstruction of the density for the entire protein, again using *phenix.maps*, the target density for the open state was extracted by masking the map about chain D using the *volutil* module of VMD.

Figure 6 suggests improvement in the refinement of CODH both at the high and intermediate resolutions. Similar to the ADK example, a higher number of replica improved the distribution of models across the range of CCs. But now, the best-fitted MDFF model improved from CC= 0.75 to 0.8 between $M_{replicas}$=16 and 100, which an improvement of 6.7%, higher than the improvement of 3% seen in ADK over a similar range of replicas. Thus, the workflow on one hand scales with system size, while on the other benefits from

the deployment of multiple replica simulations as the system size grows.

Our procedure of performing flexible fitting with on-the-fly adaptive decision making to transition from closed to open state, results in an ensemble within 2.0 Å RMSD to the experimentally determined "open" state comparing backbone atoms. This outcome is also comparable to our past refinement of CODH using a so-called cascade or simulated-annealing protocol, where the refined CODH model also reached within 2 - 2.5 Å of the open target. The larger number of replicas offer a search model greater number of opportunities to conform to density features in high resolution EM maps, and the min (1, $e^{\Delta E_{CC}/KT}$) based selection-rule installed in R-MDFF enables avoiding of local minima in the CC space. Since the larger systems are prone to degeneracy of density features, we expect R-MDFF workflow enabled via EnTK to be more useful in overcoming the local minima and exhaustively sampling the conformational space as the size grows.

3.4 R-MDFF: Performance characterization

This section characterizes the computing performance of R-MDFF on HPC resources. We provide evidence that R-MDFF manages computing resources efficiently, with comparatively small overhead when running multiple replicas.

3.4.1 Experiment Configuration

We designed 11 experiments to evaluate the efficiency of R-MDFF enabled via EnTK and we summarized their setup and results in Table 1. We utilized two biological systems—adenylate kinase (ADK) and carbon monoxide dehydrogenase (CODH)—running between 2 and 100 replicas (Rep), with varying simulation length (Sim. Len.) and resolution (Res.). Each experiment executed between 256 and 12800 tasks on PSC Bridges and ORNL Summit.

We characterized the performance of EnTK by measuring its overhead (OVH), i.e., the amount of time in which compute nodes are available but not used to execute tasks. Specifically, we separate between then time taken by the middleware (EnTK and its runtime system) to acquire resources, bootstrap the components and schedule the tasks; from the time taken by all the tasks to perform their scientific computation. Thus, OVH gives a simple but effective way to evaluate the cost of executing MDFF with EnTK and its components in terms of time spent to do everything else but science.

Experiment's runs utilize up to 4 compute nodes on Bridges2 and 100, and execute each replica on a full compute node. On Bridges2, the NAMD MD engine uses 128 cores (AMD EPYC 7742 with of 256GB DDR4 memory), without GPU acceleration. Note that Bridges2 offers 24 compute nodes, each with 8 GPU V100 accelerators but we decided to use only CPU resources due to their limited availability. On Summit,

Table 1: Experiments to characterize R-MDFF performance. System: biological system name; Rep. (M): Total number of replicas between 2 and 100; Sim. Len. (ps): timescale per iteration in picoseconds; Res. (Å): resolutions in Angstrom (high 1.8Å and intermediate 3.0Å); Resource: GPUs and CPU cores on OLCF Summit and CPU cores only on PSC Bridges2; Tasks: number of tasks for each experiment; OVH(s): Overhead of R-MDFF enabled via EnTK in seconds.

Exp. ID	**System**	**Rep. (M)**	**Sim. Len. (ps)**	**Res. (Å)**	**Resource**	**Tasks**	**OVH (s)**
1	ADK	2	64	1.8	Bridges (CPU)	256	81.0 ± 10
2	ADK	4	32	1.8	Bridges (CPU)	512	126.0 ± 10
3	ADK	4	250	1.8	Summit (GPU&CPU)	512	92.0
4	ADK	8	160	1.8	Summit (GPU&CPU)	1024	105.27 ± 18
5	ADK	16	80	1.8	Summit (GPU&CPU)	2048	114.06 ± 16
6	ADK	32	40	1.8	Summit (GPU&CPU)	4096	109.33 ± 10
7	ADK	64	20	1.8	Summit (GPU&CPU)	8092	158.87 ± 57
8	ADK	100	10	1.8	Summit (GPU&CPU)	12800	266.98 ± 245
9	CODH	16	80	1.8	Summit (GPU&CPU)	2048	93.34 ± 17
10	CODH	16	80	3.0	Summit (GPU&CPU)	2048	99.44 ± 20
11	CODH (long)	100	80	1.8	Summit (GPU&CPU)	12800	113.61 ± 20

we run the CUDA-enabled NAMD MD engine on 6 NVIDIA V100 GPU accelerators per node. Different hardware platforms show wide performance gaps in time to solution but the cross-correlation is similar when using the same configurations.

We provide templates to allow users to replicate the experiments presented in this paper or as a starting point to create a run new experiments. The templates, written in YAML, store user-defined attributes for experiments and HPC resources separately, ensuring flexible analysis on diverse computing platforms. The source code and configuration parameters of the experiments are published on the R-MDFF Github repository.[61]

3.4.2 EnTK Overhead is Steady Across HPC Platforms

We measured the time spent by EnTK to bootstrap and clean up the execution environment. Those are overheads as they measure the time spent before and after the execution of the workflow's tasks, when computing resources are already available. We measured the overheads on both Bridges and Summit, and at different scales.

Both bootstrap and clean up overheads are independent of the workflow scale as the time taken to manage the execution environment does not depend on the number of tasks executed in it. However, the bootstrap overhead can vary, depending on the filesystem performance and network latency, when serving packages and files during the bootstrapping process. We used a pre-configured environment to reduce the

bootstrapping overhead by limiting the number of downloaded packages and the I/O operations required to build the execution environment of EnTK and the other RADICAL-Cybertools.

Figure 7 shows that OVH is between 3% and 5% of the total execution time of the workflow presented in §5, across all our experiments. As summarized in Table 1, OVH is invariant of the number of replicas executed on Summit (150.90 $\pm$ 115 seconds) and on Bridges (103.5 $\pm$ 22.5 seconds) when running from 2 replicas to 100 replicas.

Bridges2 shows three times larger overhead compared to Summit, mainly due to the different performance of the parallel filesystems: Lustre on Bridges2, GPFS on Summit. On Lustre, the initial access to files takes longer than continuous access because Lustre has to retrieve the location of the actual storage device over the network. The additional results from both platforms are reported in SI figures, **Figures S8 and S9**.

4 Conclusions

Cryo-EM data of a protein represents an average of many two-dimensional images transformed to a three-dimensional density map. Classical methods in statistical mechanics such as MD fail to determine such an ensemble in finite length simulations, as structures remain trapped in deep potential wells corresponding to local dense points in the density maps. To circumvent this algorithmic bottleneck of importance sampling and to decide the quality of an ensemble of protein structures on-the-fly, we present a framework for ensemble refinement of protein structures with adaptive decision making that improves both the quality of model and fit. We call this method R-MDFF. A refined protein ensemble offers, on one hand, the most probable structural representation based on available density information, while offering insights on protein conformational dynamics that are often ignored in traditional single-model interpretation derived from single-particle experiments.

The R-MDFF based workflow application allows adaptive decision-making for flexible fitting simulations by the the integration of correlation analysis with MD simulations. This workflow is implemented on two distinct heterogeneous high-performance national supercomputers facilities, Bridges2 and Summit. The workflow performs an user-defined number of iterative fitting and analysis tasks. This multi-replica scheme improves statistical significance, the quality of models over those derived from the traditional scheme of performing a single long MDFF simulation. Consequently, the new scheme arrives not just at the best-fit but a population of models with varied ranges of data-consistency. In addition, we show that R-MDFF enabled via EnTK, is well suited for heterogeneous extreme-scale high-performance computing environments[67] by managing resource utilization of GPU and CPU computing units and the workflow overhead for increased ensemble members. We also show that our approach would have a similar computational cost as the tradi-

tional single long MDFF simulation, but with a quick turnaround time (shorter wall time of workload), while exploring interesting regions in the density map. Larger system sizes that are more akin to cryo-EM structure determination offer further performance advantages. We continue to extend the capability of R-MDFF in complex applications in exascale high-performance computing environments.

5 Methods

Modern adaptive sampling frameworks are dynamic, extensible, scalable and robust to facilitate hundreds or thousands of experiments for searching different structures, and specialized features can be added to solve existing problems through the framework. We developed a workflow application using RADICAL-Cybertools[37] that provides a scalable workflow framework for implementing ensemble refinement with cross correlation calculation on HPC computing resources. R-MDFF (RADICAL-Cybertools enhanced Molecular Dynamics Flexible Fitting), depicted in **Figure ??**, supports adaptive decision making algorithms to iterate between molecular dynamics flexible fitting simulation and cross-correlation analysis. Our workflow application is portable to explore the space of experimental configurations and support various use cases, so that the ensemble refinement produces results on different dimensions of a physical system; resolution density, simulation length, replica count, and HPC resource. The complete integration is explained in the following sections: (a) MDFF simulation, (b) CC analysis, (c) RADICAL cybertools and (d) validation approaches.

5.1 Molecular Dynamics Flexible Fitting simulation:

In the pipeline simulation stage, R-MDFF uses the conventional MDFF algorithm, as described in.[3] Briefly, MDFF requires, as input data, an initial structure and a cryo-EM density map. A potential map is generated from the density and subsequently used to bias a MD simulation of the initial structure. The structure is subject to the EM-derived potential while simultaneously undergoing structural dynamics as described by the MD force field.

Let the Coulomb potential associated with the EM map be $\Phi(\mathbf{r})$. Then the MDFF potential map is given by,

$$V_{EM}(\mathbf{r}) = \begin{cases} \zeta \left[\frac{\phi((r)) - \phi_{th}}{\phi_{max} - \phi_{th}} \right], & \text{if } \phi(r) \geq \phi_{th} \\ \zeta, & \text{if } \phi(r) < \phi_{th} \end{cases} \tag{1}$$

where ζ is a scaling factor that controls the strength of the coupling of atoms to the MDFF potential, ϕ_{th} is a threshold for disregarding noise, and $\phi_{max} = max(\phi(r))$. The potential energy contribution from the MDFF forces is then

$$U_{EM}(\mathbf{r} = \sum_i w_i V_{EM}(r_i) \tag{2}$$

where i labels the atoms in the structure and w_i is an atom-dependent weight, usually the atomic mass. During the simulation, the total potential acting on the system is given by,

$$U_{total} = U_{MD} + U_{EM} + U_{SS} \tag{3}$$

where U_{MD} is the MD potential energy as provided by MD force fields (e.g. CHARMM) and U_{SS} is a secondary structure restraint potential that prevents warping of the secondary structure by the potentially strong forces due to U_{EM}. A detailed description of the MDFF methodology is presented in.[1,2] Specific simulation parameters for the example cases of ADK and CODH are provided on the GitHub page.[61]

5.2 Cross-correlation analysis

The analysis and decision-making part of the ensemble refinement involves calculating the map-model cross correlation (CC) value for all replicas at every iteration of the R-MDFF workflow. For t MD steps and M replicas per iteration, the total simulation time is equal to $t_{steps/iteration} \times N_{iteration} \times M_{replicas}$. At the end of each iteration, the CC for $M_{replicas}$ number of resulting structures is computed against the target map to examine the quality of fitting. Atomic coordinates corresponding to the Monte Carlo-like selection rule described in Sect. 2 are used to restart all the replicas for the next iteration.

5.3 RADICAL Cybertools

In order to implement the pipeline, we have extended an open-source, Python framework – RADICAL Ensemble Toolkit (EnTK) – that facilitates adaptive ensemble biomolecular simulations at scale. The first step of writing the EnTK workflow code is to construct a task parallel execution of MDFF simulation using NAMD 2.14,[33,53] and to connect the analysis stage to find highest CC values among replicas. While all the necessary information such as NAMD checkpoints and CC values are kept under EnTK's data staging area, distributed computing resources are coordinated to ensure the workflow performance over CPUs and GPUs from heterogeneous HPC platforms. In addition, several features have been added to the application by utilizing existing capabilities of RADICAL-Cybertools. Tcl scripting is interfaced with EnTK APIs to interact with VMD software directly and the partitioned scheduling is introduced to assign a single node per replica exclusively for the best performance of NAMD simulations. Usability and productivity have been addressed to automate resource configurations and experiment settings as well as ensuring reproducibility of

scientific data. With the R-MDFF application, replacing MD engines or analysis methods needs to change a few lines of settings in a workflow management file without source code modifications. The application, R-MDFF is available on GitHub (https://github.com/radical-collaboration/MDFF-EnTK) and implemented to support adaptive decision making for ensemble-based simulations and to enable the novel analysis method, MDFF or others on HPC resources.

5.4 Other analyses - MolProbity and PCA

MolProbity[45] scores are calculated to determine the quality of the ensemble of structures for protein ADK after adaptive decision based flexible fitting. The distribution of MolProbity scores indicate that as the ensemble members increase, ranging from 16 to 400, a population of high quality structural ensemble (MolProbity score ~ 0.75) is observed.

For the analysis of protein molecular dynamics simulations, principal component analysis (PCA)[68–70] approach can monitor the individual modes, thereby allowing one to filter the major modes of collective motion from local fluctuations. Often these principal modes of motion is correlated with protein function, the reduced dimensional subspace spanned by these modes was termed essential dynamics,[71] reflecting the modes which correspond to essential biological function. Also, using PCA we can distinguish converged structures (fitted to EM density map) from outlier structures (outside the EM density map).[69]

We performed PCA using ProDY[72] and represent the essential dynamics using the Normal Mode Wizard[72] plugin in VMD.[73] The principal components are represented in **Figures 5 B and C**, corresponding to 16 and 400 ensemble members respectively. PCA results of evaluating the structural ensemble of protein ADK across multiple iterations in R-MDFF, suggests while fitting to a high resolution EM density map with low ensemble members (16 replicas), the essential dynamics capture mostly the biologically relevant "lid-closed" state. However, increasing the number of ensemble members (400 replicas), results indicate that normal modes can probe both the the biologically relevant "lid-open" and "lid-closed" states.

Acknowledgement

Singharoy acknowledges start-up funds from SMS and CASD at Arizona State University, CAREER award from NSF (MCB-1942763), and an NDEP grant from the Department of Defense. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. The RADICAL Lab acknowledges NSF Awards 1835449 and 1931512. The benchmarks were also carried out using the resources of the OLCF at Oak Ridge National

Laboratory, which is supported by the Office of Science at DOE under Contract No. DE-AC05-00OR22725, made available via the INCITE program.

References

(1) Trabuco, L. G.; Villa, E.; Mitra, K.; Frank, J.; Schulten, K. Flexible Fitting of Atomic Structures into Electron Microscopy Maps Using Molecular Dynamics. *Structure* **2008**, *16*, 673–683, DOI: 10.1016/j.str.2008.03.005.

(2) Trabuco, L. G.; Villa, E.; Schreiner, E.; Harrison, C. B.; Schulten, K. Molecular dynamics flexible fitting: A practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods* **2009**, *49*, 174–180, DOI: 10.1016/j.ymeth.2009.04.005.

(3) Singharoy, A.; Teo, I.; McGreevy, R.; Stone, J. E.; Zhao, J.; Schulten, K. Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. *eLife* **2016**, *5*, 1–33, DOI: 10.7554/eLife.16105.

(4) Leaver-Fay, A. et al. *Methods in Enzymology*; Academic Press, 2011; Vol. 487; pp 545–574, DOI: 10.1016/B978-0-12-381270-4.00019-6.

(5) McGreevy, R.; Teo, I.; Singharoy, A.; Schulten, K. Advances in the molecular dynamics flexible fitting method for cryo-EM modeling. *Methods* **2016**, *100*, 50–60, DOI: 10.1016/j.ymeth.2016.01.009.

(6) Igaev, M.; Kutzner, C.; Bock, L. V.; Vaiana, A. C.; Grubmüller, H. Automated cryo-EM structure refinement using correlation-driven molecular dynamics. *eLife* **2019**, *8*, 1–33, DOI: 10.7554/eLife.43542.

(7) Kim, D. N.; Moriarty, N. W.; Kirmizialtin, S.; Afonine, P. V.; Poon, B.; Sobolev, O. V.; Adams, P. D.; Sanbonmatsu, K. Cryo_fit: Democratization of flexible fitting for cryo-EM. *Journal of Structural Biology* **2019**, *208*, 1–6, DOI: 10.1016/j.jsb.2019.05.012.

(8) Costa, M. G.; Fagnen, C.; Vénien-Bryan, C.; Perahia, D. A New Strategy for Atomic Flexible Fitting in Cryo-EM Maps by Molecular Dynamics with Excited Normal Modes (MDeNM-EMfit). 2020; https://pubs.acs.org/doi/abs/10.1021/acs.jcim.9b01148.

(9) Vant, J. W.; Lahey, S. L. J.; Jana, K.; Shekhar, M.; Sarkar, D.; Munk, B. H.; Kleinekathöfer, U.; Mittal, S.; Rowley, C.; Singharoy, A. Flexible Fitting of Small Molecules into Electron Microscopy

Maps Using Molecular Dynamics Simulations with Neural Network Potentials. *Journal of Chemical Information and Modeling* **2020**, *60*, 2591–2604, DOI: 10.1021/acs.jcim.9b01167.

(10) Vant, J. W.; Sarkar, D.; Gupta, C.; Shekhar, M. S.; Mittal, S.; Singharoy, A. In *Protein Structure Prediction*; Kihara, D., Ed.; Methods in Molecular Biology; Springer US: New York, NY, 2020; pp 301–315, DOI: 10.1007/978-1-0716-0708-4_18.

(11) Pfab, J.; Phan, N. M.; Si, D. DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on cov-related complexes. *Proceedings of the National Academy of Sciences of the United States of America* **2021**, *118*, DOI: 10.1073/pnas.2017525118.

(12) Perez, A.; MacCallum, J. L.; Dill, K. A. Accelerating molecular simulations of proteins using Bayesian inference on weak information. *Proceedings of the National Academy of Sciences* **2015**, *112*, 11846–11851, DOI: 10.1073/pnas.1515561112.

(13) Perez, A.; Morrone, J. A.; Brini, E.; MacCallum, J. L.; Dill, K. A. Blind protein structure prediction using accelerated free-energy simulations. *Science Advances* **2016**, *2*, DOI: 10.1126/sciadv.1601274.

(14) Shekhar, M. et al. CryoFold: Determining protein structures and data-guided ensembles from cryo-EM density maps. *Matter* **2021**, *4*, 3195–3216, DOI: 10.1016/j.matt.2021.09.004.

(15) Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. Metainference: A Bayesian inference method for heterogeneous systems. *Science Advances* **2016**, *2*, 1501177, DOI: 10.1126/sciadv.1501177.

(16) Bonomi, M.; Hanot, S.; Greenberg, C. H.; Sali, A.; Nilges, M.; Vendruscolo, M.; Pellarin, R. Bayesian Weighing of Electron Cryo-Microscopy Data for Integrative Structural Modeling. *Structure* **2019**, *27*, 175–188.e6, DOI: 10.1016/j.str.2018.09.011.

(17) Gupta, C.; Sarkar, D.; Tieleman, D. P.; Singharoy, A. The Ugly, Bad, and Good Stories of Large-Scale Biomolecular Simulations. *Current Opinion in Structural Biology* **2022**, *73*, 102338, DOI: 10.1016/j.sbi.2022.102338.

(18) Vant, J. W.; Sarkar, D.; Nguyen, J.; Baker, A. T.; Vermaas, J. V.; Singharoy, A. Exploring Cryo-Electron Microscopy with Molecular Dynamics. *Biochemical Society Transactions* **2022**, *50*, 569–581, DOI: 10.1042/BST20210485.

(19) Lawson, C. L. et al. Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge. *Nature Methods* **2021**, *18*, 156–164, DOI: 10.1038/s41592-020-01051-w.

(20) Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, DOI: 10.1038/s41586-021-03819-2.

(21) Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876, DOI: 10.1126/science.abj8754.

(22) Kryshtafovych, A.; Moult, J.; Billings, W. M.; Della Corte, D.; Fidelis, K.; Kwon, S.; Olechnovič, K.; Seok, C.; Venclovas, Č.; Won, J.; Participants, C.-C. Modeling SARS-CoV-2 Proteins in the CASP-commons Experiment. *Proteins: Structure, Function, and Bioinformatics* **2021**, *89*, 1987–1996, DOI: 10.1002/prot.26231.

(23) Lensink, M. F. et al. Prediction of Protein Assemblies, the next Frontier: The CASP14-CAPRI Experiment. *Proteins: Structure, Function, and Bioinformatics* **2021**, *89*, 1800–1823, DOI: 10.1002/prot.26222.

(24) Lee, H.; Turilli, M.; Jha, S.; Bhowmik, D.; Ma, H.; Ramanathan, A. DeepDriveMD: Deep-Learning Driven Adaptive Molecular Simulations for Protein Folding. 2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS). 2019; pp 12–19, DOI: 10.1109/DLS49591.2019.00007.

(25) Balasubramanian, V.; Jensen, T.; Turilli, M.; Kasson, P.; Shirts, M.; Jha, S. Adaptive Ensemble Biomolecular Applications at Scale. *SN Computer Science* **2020**, *1*, 1–15, https://doi.org/10.1007/s42979-020-0081-1.

(26) Herzik, M. A.; Fraser, J. S.; Lander, G. C. A Multi-model Approach to Assessing Local and Global Cryo-EM Map Quality. *Structure* **2019**, *27*, 344–358.e3, DOI: 10.1016/j.str.2018.10.003.

(27) Frank, J.; Ourmazd, A. Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. *Methods* **2016**, *100*, 61–67, DOI: 10.1016/j.ymeth.2016.02.007.

(28) Netz, R. R.; Eaton, W. A. Estimating computational limits on theoretical descriptions of biological cells. *Proceedings of the National Academy of Sciences* **2021**, *118*, e2022753118, DOI: 10.1073/pnas.2022753118.

(29) Terashi, G.; Kihara, D. De novo main-chain modeling for em maps using MAINMAST. *Nature Communications* **2018**, *9*, 1–11, DOI: 10.1038/s41467-018-04053-7.

(30) Vant, J. W.; Sarkar, D.; Streitwieser, E.; Fiorin, G.; Skeel, R.; Vermaas, J. V.; Singharoy, A. Data-guided Multi-Map variables for ensemble refinement of molecular movies. *Journal of Chemical Physics* **2020**, *153*, DOI: 10.1063/5.0022433.

(31) Huang, J.; Mackerell, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry* **2013**, *34*, 2135–2145, DOI: 10.1002/jcc.23354.

(32) McGreevy, R.; Singharoy, A.; Li, Q.; Zhang, J.; Xu, D.; Perozo, E.; Schulten, K. xMDFF: Molecular Dynamics Flexible Fitting of Low-Resolution X-ray Structures. *Acta Crystallographica. Section D, Biological Crystallography* **2014**, *70*, 2344–2355, DOI: 10.1107/S1399004714013856.

(33) Phillips, J. C. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics* **2020**, *153*, 044130, DOI: 10.1063/5.0014475.

(34) Croll, T. I. ISOLDE: A physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallographica Section D: Structural Biology* **2018**, *74*, 519–530, DOI: 10.1107/S2059798318002425.

(35) Goddard, T. D.; Huang, C. C.; Meng, E. C.; Pettersen, E. F.; Couch, G. S.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Science* **2018**, *27*, 14–25, DOI: 10.1002/pro.3235.

(36) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **2004**, *25*, 1605–1612, DOI: 10.1002/jcc.20084.

(37) Turilli, M.; Balasubramanian, V.; Merzky, A.; Paraskevakos, I.; Jha, S. Middleware building blocks for workflow systems. *Computing in Science & Engineering* **2019**, *21*, 62–75.

(38) Balasubramanian, V.; Turilli, M.; Hu, W.; Lefebvre, M.; Lei, W.; Modrak, R.; Cervone, G.; Tromp, J.; Jha, S. Harnessing the power of many: Extensible toolkit for scalable ensemble applications. International Parallel and Distributed Processing Symposium. 2018; pp 536–545.

(39) Luckow, A.; Santcroos, M.; Zebrowski, A.; Jha, S. Pilot-data: an abstraction for distributed data. *Journal of Parallel and Distributed Computing* **2015**, *79*, 16–30.

(40) Luckow, A.; Rattan, K.; Jha, S. Pilot-Edge: Distributed Resource Management Along the Edge-to-Cloud Continuum. *arXiv preprint arXiv:2104.03374* **2021**, Accepted for PAISE'21 (IPDPS 21).

(41) Dakka, J.; Farkas-Pall, K.; Turilli, M.; Wright, D. W.; Coveney, P. V.; Jha, S. Concurrent and adaptive extreme scale binding free energy calculations. 2018 IEEE 14th International Conference on e-Science (e-Science). 2018; pp 189–200.

(42) Hruska, E.; Balasubramanian, V.; Lee, H.; Jha, S.; Clementi, C. Extensible and scalable adaptive sampling on supercomputers. *Journal of Chemical Theory and Computation* **2020**, *16*, 7915–7925.

(43) Zwier, M. C.; Adelman, J. L.; Kaus, J. W.; Pratt, A. J.; Wong, K. F.; Rego, N. B.; Suárez, E.; Lettieri, S.; Wang, D. W.; Grabe, M., et al. WESTPA: An interoperable, highly scalable software package for weighted ensemble simulation and analysis. *Journal of chemical theory and computation* **2015**, *11*, 800–809.

(44) Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **2004**, *57*, 702–710, DOI: `10.1002/prot.20264`.

(45) Williams, C. J. et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* **2018**, *27*, 293–315, DOI: `10.1002/pro.3330`.

(46) Barad, B. A.; Echols, N.; Wang, R. Y. R.; Cheng, Y.; Dimaio, F.; Adams, P. D.; Fraser, J. S. EMRinger: Side chain-directed model and map validation for 3D cryo-electron microscopy. *Nature Methods* **2015**, *12*, 943–946, DOI: `10.1038/nmeth.3541`.

(47) Pintilie, G.; Zhang, K.; Su, Z.; Li, S.; Schmid, M. F.; Chiu, W. Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nature Methods* **2020**, *17*, 328–334, DOI: `10.1038/s41592-020-0731-1`.

(48) Punjani, A.; Rubinstein, J. L.; Fleet, D. J.; Brubaker, M. A. CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods* **2017**, *14*, 290–296, DOI: `10.1038/nmeth.4169`.

(49) Ourmazd, A. Cryo-EM, XFELs and the structure conundrum in structural biology. *Nature Methods* **2019**, *16*, 941–944, DOI: `10.1038/s41592-019-0587-4`.

(50) Mashayekhi, G.; Vant, J.; Polavarapu, A.; Ourmazd, A.; Singharoy, A. Energy landscape of the SARS-CoV-2 reveals extensive conformational heterogeneity. *Current Research in Structural Biology* **2022**, *4*, 68–77, DOI: `10.1016/j.crstbi.2022.02.001`.

(51) Wriggers, W. Conventions and workflows for using Situs. *Acta crystallographica. Section D, Biological crystallography* **2012**, *68*, 344–351, DOI: `10.1107/S0907444911049791`, 22505255[pmid] PMC3322594[pmcid] S0907444911049791[PII] Acta Crystallogr D Biol Crystallogr.

(52) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38, DOI: `https://doi.org/10.1016/0263-7855(96)00018-5`.

(53) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **2005**, *26*, 1781–1802, DOI: 10.1002/jcc.20289.

(54) MacCallum, J. L.; Perez, A.; Dill, K. A. Determining Protein Structures by Combining Semireliable Data with Atomistic Physical Models by Bayesian Inference. *Proceedings of the National Academy of Sciences* **2015**, *112*, 6985–6990, DOI: 10.1073/pnas.1506788112.

(55) Tang, W. S.; Silva-Sánchez, D.; Giraldo-Barreto, J.; Carpenter, B.; Hanson, S.; Barnett, A. H.; Thiede, E. H.; Cossio, P. Ensemble Reweighting Using Cryo-EM Particles. 2022.

(56) Ho, N.; Cava, J. K.; Vant, J.; Shukla, A.; Miratsky, J.; Turaga, P.; Maciejewski, R.; Singharoy, A. Learning Free Energy Pathways through Reinforcement Learning of Adaptive Steered Molecular Dynamics. 2022.

(57) Tsai, S.-T.; Fields, E.; Xu, Y.; Kuo, E.-J.; Tiwary, P. Path Sampling of Recurrent Neural Networks by Incorporating Known Physics. *Nature Communications* **2022**, *13*, 7231, DOI: 10.1038/s41467-022-34780-x.

(58) Wang, Y.; Herron, L.; Tiwary, P. From Data to Noise to Data for Mixing Physics across Temperatures with Generative Artificial Intelligence. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2203656119, DOI: 10.1073/pnas.2203656119.

(59) Evans, L.; Cameron, M. K.; Tiwary, P. Computing Committors via Mahalanobis Diffusion Maps with Enhanced Sampling Data. *The Journal of Chemical Physics* **2022**, *157*, 214107, DOI: 10.1063/5.0122990.

(60) Evans, L.; Cameron, M. K.; Tiwary, P. Computing Committors in Collective Variables via Mahalanobis Diffusion Maps. *Applied and Computational Harmonic Analysis* **2023**, DOI: 10.1016/j.acha.2023.01.001.

(61) MDFF Integration with EnTK. https://github.com/radical-collaboration/MDFF-EnTK, 2019.

(62) Arora, K.; Brooks, C. L. Large-Scale Allosteric Conformational Transitions of Adenylate Kinase Appear to Involve a Population-Shift Mechanism. *Proceedings of the National Academy of Sciences* **2007**, *104*, 18496–18501, DOI: 10.1073/pnas.0706443104.

(63) Matsunaga, Y.; Fujisaki, H.; Terada, T.; Furuta, T.; Moritsugu, K.; Kidera, A. Minimum Free Energy Path of Ligand-Induced Transition in Adenylate Kinase. *PLOS Computational Biology* **2012**, *8*, e1002555, DOI: 10.1371/journal.pcbi.1002555.

(64) Li, D.; Liu, M. S.; Ji, B. Mapping the Dynamics Landscape of Conformational Transitions in Enzyme: The Adenylate Kinase Case. *Biophysical Journal* **2015**, *109*, 647–660, DOI: 10.1016/j.bpj.2015.06.059.

(65) Olsson, U.; Wolf-Watz, M. Overlap between Folding and Functional Energy Landscapes for Adenylate Kinase Conformational Change. *Nature Communications* **2010**, *1*, 111, DOI: 10.1038/ncomms1106.

(66) Aviram, H. Y.; Pirchi, M.; Mazal, H.; Barak, Y.; Riven, I.; Haran, G. Direct Observation of Ultrafast Large-Scale Dynamics of an Enzyme under Turnover Conditions. *Proceedings of the National Academy of Sciences* **2018**, *115*, 3243–3248, DOI: 10.1073/pnas.1720448115.

(67) Merzky, A.; Turilli, M.; Titov, M.; Al-Saadi, A.; Jha, S. Design and Performance Characterization of RADICAL-Pilot on Leadership-Class Platforms. *IEEE Transactions on Parallel & Distributed Systems* **2022**, *33*, 818–829, DOI: 10.1109/TPDS.2021.3105994.

(68) Stein, S. A. M.; Loccisano, A. E.; Firestine, S. M.; Evanseck, J. D. In *Annual Reports in Computational Chemistry*; Spellmeyer, D. C., Ed.; Elsevier, 2006; Vol. 2; pp 233–261, DOI: 10.1016/S1574-1400(06)02013-5.

(69) Sittel, F.; Jain, A.; Stock, G. Principal Component Analysis of Molecular Dynamics: On the Use of Cartesian vs. Internal Coordinates. *The Journal of Chemical Physics* **2014**, *141*, 014111, DOI: 10.1063/1.4885338.

(70) Schultze, S.; Grubmüller, H. Time-Lagged Independent Component Analysis of Random Walks and Protein Dynamics. *Journal of Chemical Theory and Computation* **2021**, *17*, 5766–5776, DOI: 10.1021/acs.jctc.1c00273.

(71) David, C. C.; Jacobs, D. J. In *Protein Dynamics: Methods and Protocols*; Livesay, D. R., Ed.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2014; pp 193–226, DOI: 10.1007/978-1-62703-658-0_11.

(72) Bakan, A.; Meireles, L. M.; Bahar, I. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **2011**, *27*, 1575–1577, DOI: 10.1093/bioinformatics/btr168.

(73) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38, DOI: https://doi.org/10.1016/0263-7855(96)00018-5.

Supporting Information Available

The Supporting Information (SI) includes all MolProbity scores for MDFF with adaptive decision making trajectories and the performance of the R-MDFF algorithm when applied to larger biomolecular systems such as CODH.

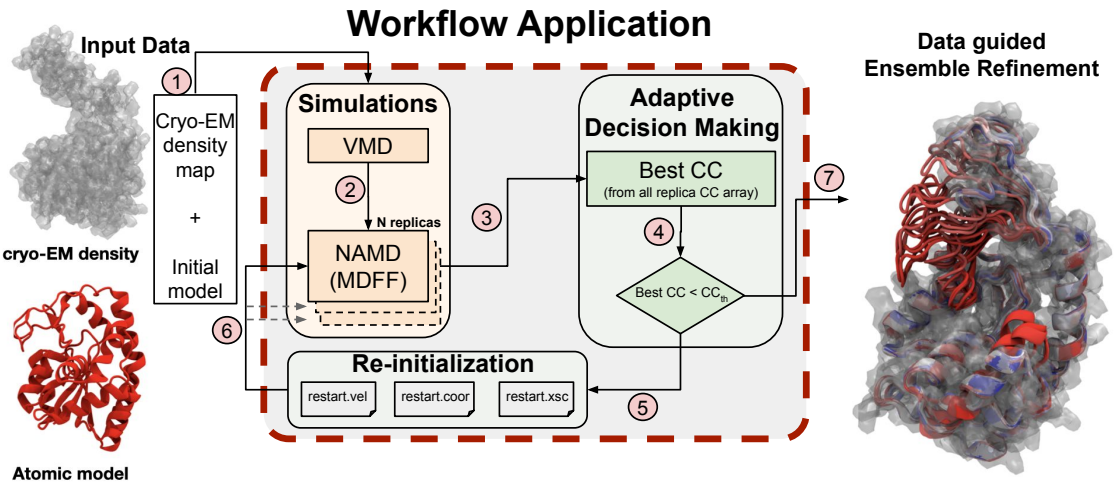

Figure 1: **Overview of the workflow application**. The schematic shows how NAMD/VMD is used to perform flexible fitting iteratively. Internal boxes with annotation numbers indicate the sequence of the workflow: (1) Input data (2) Simulation preparation and execution using VMD and NAMD respectively (3) Building CC matrix and sort through CC matrix to select best CC (4) Check if best CC is lower than threshold CC (5) Use the current state of the molecular sytem corresponding to the best CC using the restart files (6) Re-seed all replicas with the restart files and perform the next iteration of flexible fitting (7) Data guided ensemble refined models.

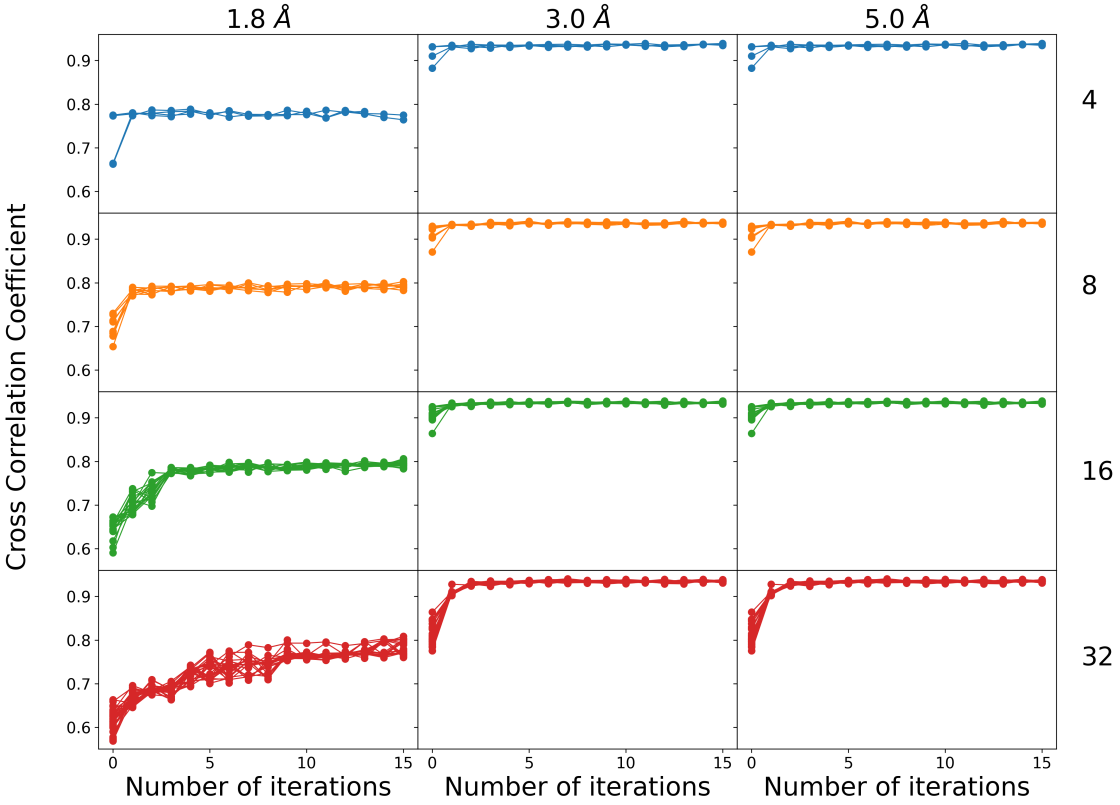

Figure 2: **CC vs. iteration comparison.** Data presented for high resolution (1.8 Å, intermediate resolution (3.0 Å) and low resolution (5.0 Å) cryo-electron microscopy density maps, for different ensemble members, 4 (blue), 8 (orange), 16 (green) and 32 (red) respectively. We find that replica number directly correlates with we have larger variabilty amongst ensemble members for high resolution density map.

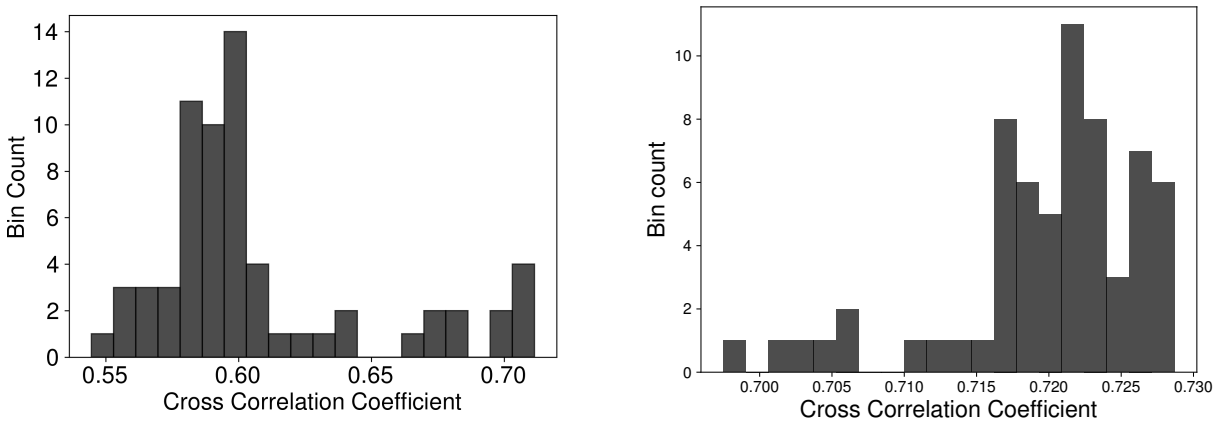

Figure 3: **Conventional vs. Adaptive MDFF** Comparing cross correlation between conventional (a) and adaptive (b) flexible when fitting ADK to high resolution EM map of 1.8 Å using 64 replicas. We find adaptive MDFF to perform better than a conventional MDFF, with a population of high cross correlation ensemble members.

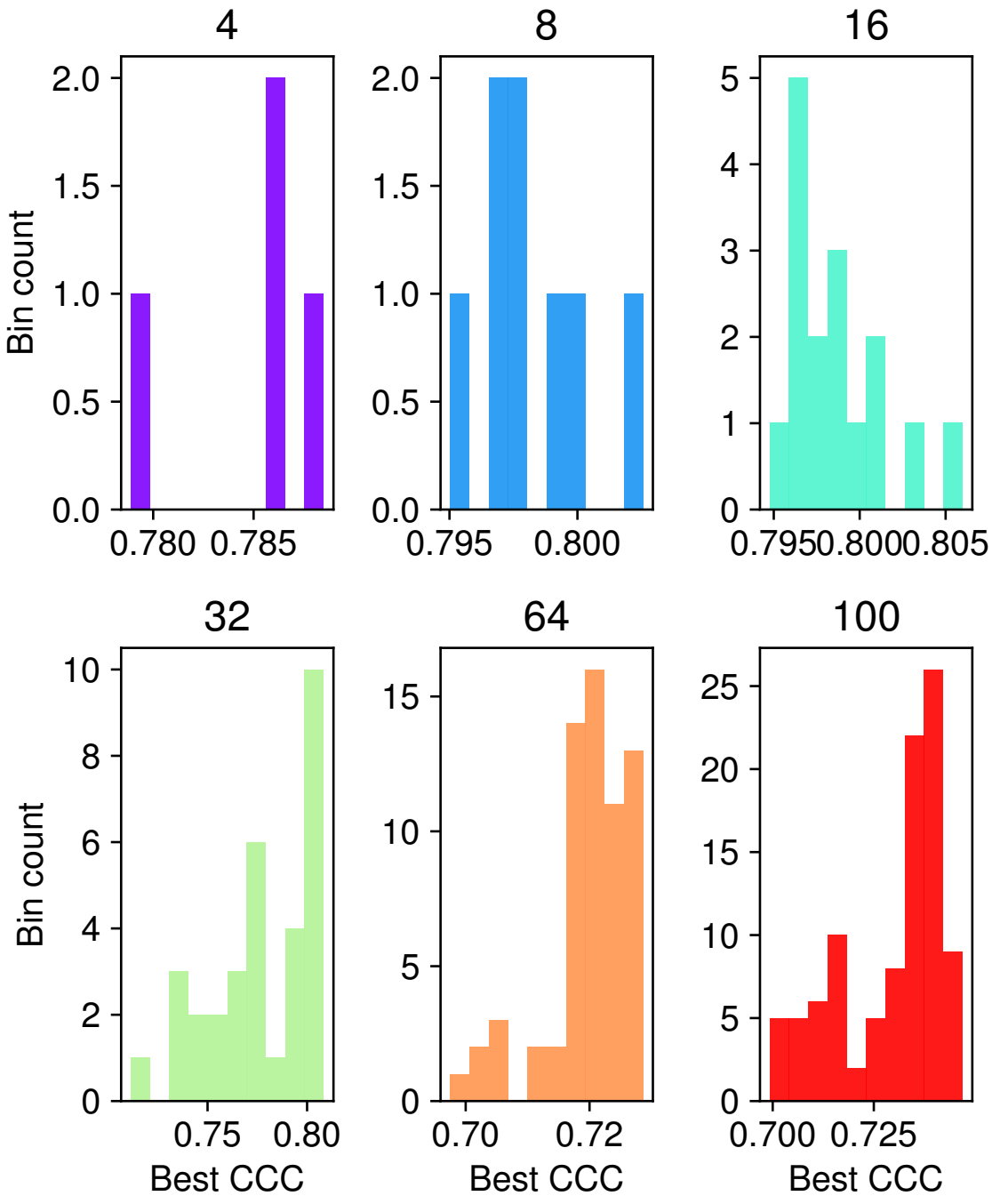

Figure 4: **Change in quality of fit with replica size** Best cross-correlation per iteration after adaptive flexible fitting of ADK to a high resolution 1.8 Å map, for different ensemble sizes with $M_{replicas}$ = 4 - 100. Monitoring the effect of replica size on the quality of fit. We observe cross-correlation to increase till 32 replicas, after which the value starts decreasing.

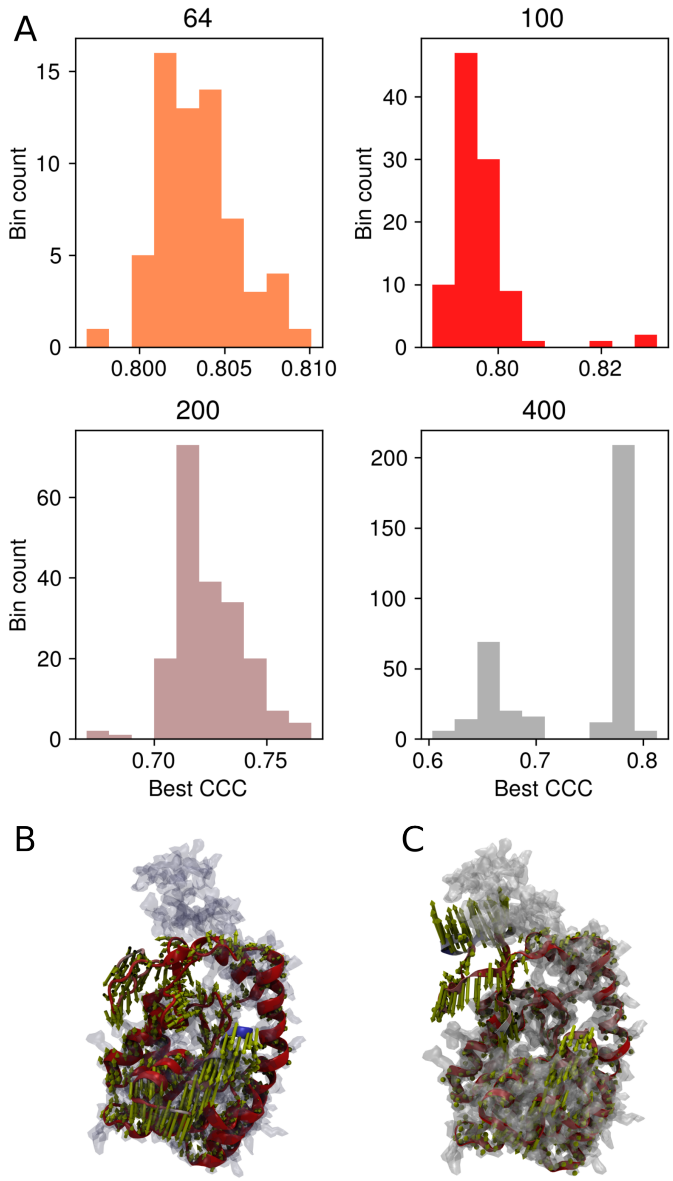

Figure 5: **Impact of total simulation time on the quality of fit.** (A) Best cross-correlation (CC) after adaptive flexible fitting of ADK to a high resolution (1.8 Å) map, for different ensemble members (64, 100, 200 and 400). (B) Principal component analysis (PCA) of ADK protein with 16 replicas flexibly fitting to 1.8 Å using adaptive protocol (C) PCA of ADK protein with 400 replicas flexibly fitting to 1.8 Å using adaptive protocol. PCA results demonstrate the experimentally observed hinge motion, can be found with a high replica number, but not for a low replica number.

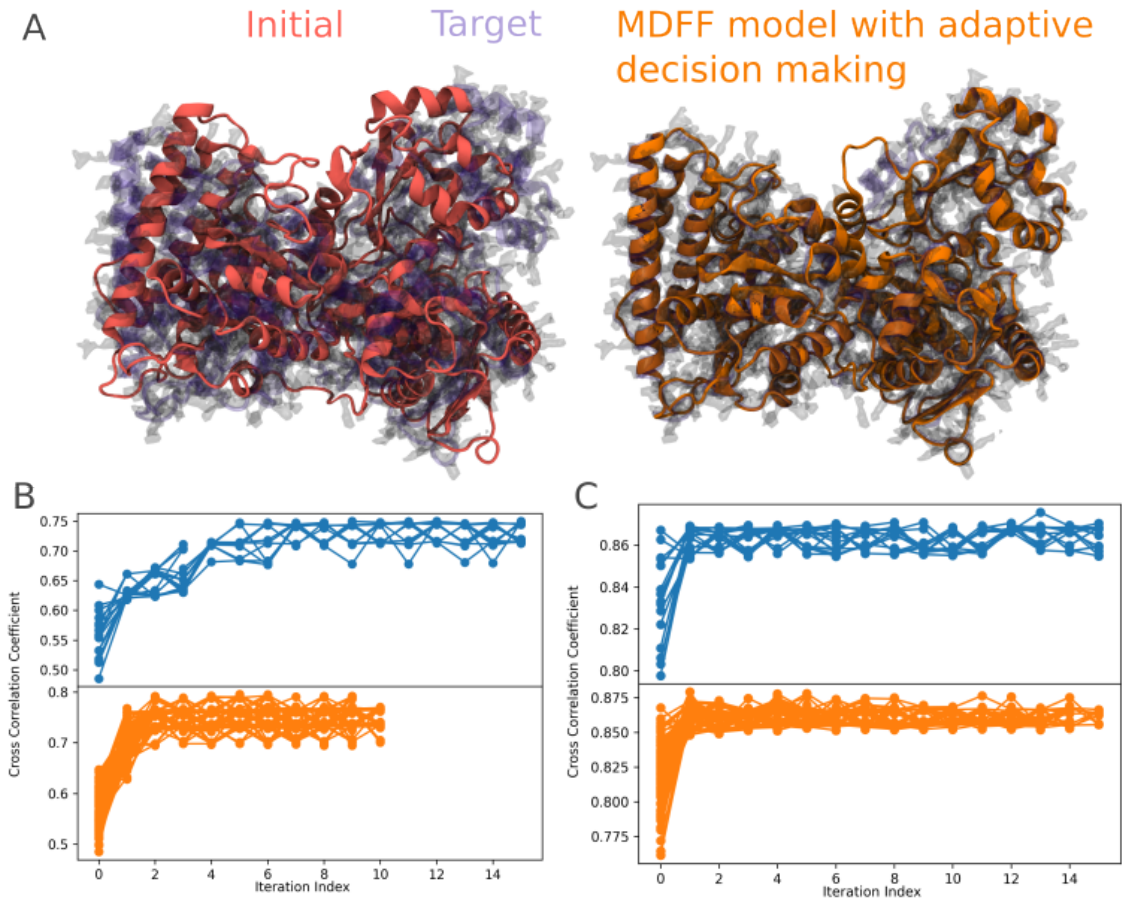

Figure 6: **Robustness of adaptive MDFF protocol to system size**. Closed (red) to open (blue) are the initial and final (target) states of the dynamics for protein CODH. Flexibly fitting with adaptive decision to high resolution 1.8 Å EM density map. CODH cross correlation vs iteration index for different ensemble members using the optimal parameters taken from ADK results at different resolutions - 1.8 Å and 3.0 Å.

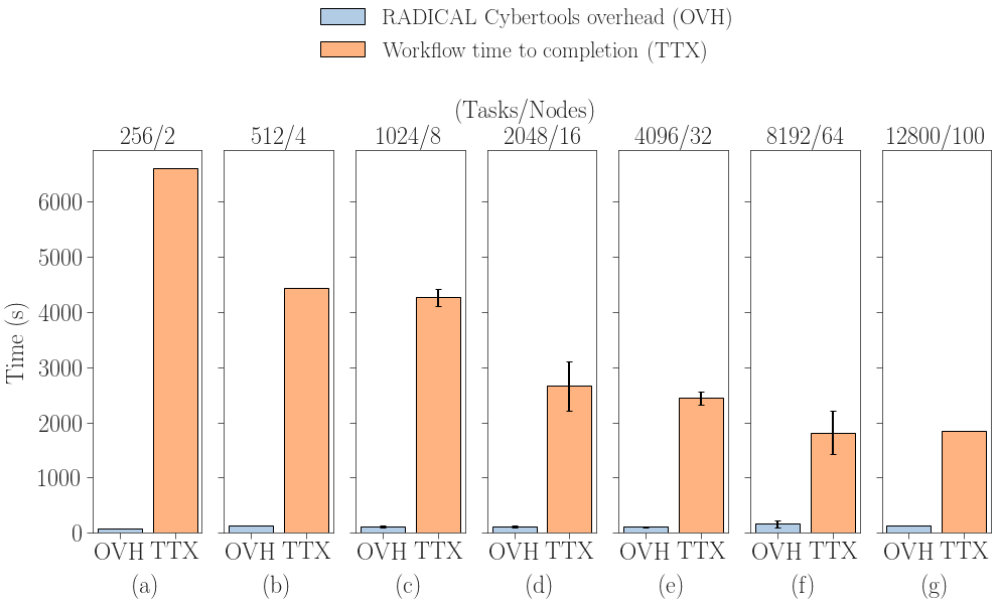

Figure 7: **Performance of EnTK during flexible fitting with adaptive decision making.** EnTK Overheads on PSC Bridges2 (a, b) and Summit (c - g) compute nodes. The total simulation time (equal to $t_{steps/iteration} \times N_{iteration} \times M_{replicas}$) is 16ns. TTX tends to decrease with the increasing of the number of compute nodes.

TOC Graphic

Some journals require a graphical entry for the Table of Contents. This should be laid out "print ready" so that the sizing of the text is correct.
Inside the `tocentry` environment, the font used is Helvetica 8 pt, as required by *Journal of the American Chemical Society*.
The surrounding frame is 9 cm by 3.5 cm, which is the maximum permitted for *Journal of the American Chemical Society* graphical table of content entries. The box will not resize if the content is too big: instead it will overflow the edge of the box.
This box and the associated title will always be printed on a separate page at the end of the document.