# Enhanced Sampling of Protein Conformational Transitions via Dynamically Optimized Collective Variables

Z. Faidon Brotzakis[a,b] and Michele Parrinello[a,b,c*]

[a] *Department of Chemistry and Applied Bioscience, ETH Zürich, c/o USI Campus, Via Giuseppe Buffi 13 Lugano, CH-6900, Lugano, Ticino, Switzerland*

[b] *Institute of Computational Science, Universita della Svizzera Italiana (USI), Via Giuseppe Buffi 13 Lugano, CH-6900, Lugano, Ticino, Switzerland. and*

[c] *Instituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy* *

## Abstract

Protein conformational transitions often involve many slow degrees of freedom. Their knowledge would give distinctive advantages since it provides chemical and mechanistic insight and accelerates the convergence of enhanced sampling techniques that rely on collective variables. In this study, we implemented a recently developed variational approach to conformational dynamics metadynamics to the conformational transition of the moderate size protein, L99A T4 Lysozyme. In order to find the slow modes of the system we combined data coming from NMR experiments as well as short MD simulations. A Metadynamics simulation based on these information reveals the presence of two intermediate states, at an affordable computational cost.

* parrinello@phys.chem.ethz.ch

## I. INTRODUCTION

The L99A variant of the T4 Lysozyme (L99A T4L) protein exhibits two metastable states, G and E (see Fig. 1a,b). The former is the high populated ground state while the latter is a low populated excited state. The interconversion between the two takes place on the timescale of miliseconds [1, 2]. In the G state a large hydrophobic cavity (150 $Å^3$) is formed, in contrast with state E in which the hydrophobic cavity is filled by the F114 residue [1–3]. Even more intriguingly L99A T4L is druggable and several crystal structures of the holo form have been reported [3, 4]. It is therefore a prototypical example of protein plasticity.

Since the G to E transformation takes place on a milisecond timescale a direct Molecular Dynamics (MD) simulation is outside our present computational capability, especially if one wants to gather sufficient statistics. One way to tackle this timescale problem is to use enhanced sampling methods. One such method is Metadynamics (MetaD) that belongs to the class of enhanced sampling methods that use Collective Variables (CVs) to describe the slow modes involved in the transition of interest. In a remarkable application of MetaD, Wang et al made use of Path Collective Variables Metadynamics (Path CV-MetaD) and were able to discover an intermediate metastable state that is capable of accommodating a ligand.

In this paper we want to complement the effort of Wang et al by approaching the same problem from a different angle. Our purpose is to reveal what are the slow modes that are to be associated with the G to E transition. This will also assess the ability of recently developed methods of obtaining such modes in a relatively more complex scenario such as that of a real protein. Our approach will be based on Variational Approach to Conformational dynamics Metadynamics (VAC-MetaD). In VAC-MetaD one first performs a short MetaD run with non optimal CVs that later uses to derive more efficient CVs. Since CV efficiency is strongly related to how well they describe the physical process one gains not only efficiency but also insight. Using these CVs we not only rediscover the on pathway intermediate state of Ref. [5], but also we find a new one.

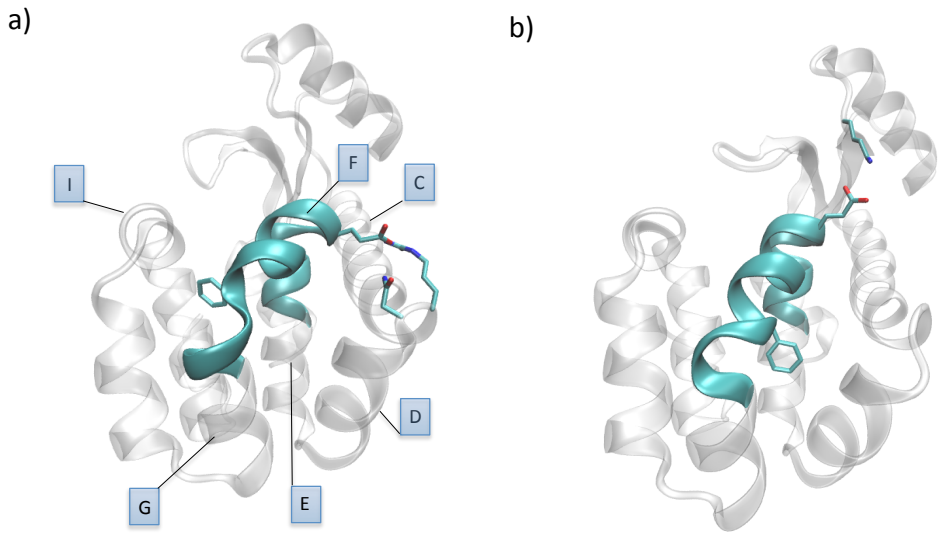


FIG. 1: Representation of the L99A T4 Lysozyme in a) state G and, b) in state E. Note that in cyan is highlighted the region of aminoacids showing an increased chemical shift signal [1]. In licorice representation we show the important salt bridge formed between residues E108 and R80,N81 in state G and between residues E108 and K35 in state E. Surrounding helices are labelled in blue.

## II. METHODS

### A. Metadynamics

Metadynamics (MetaD) is a well known enhanced sampling method for studying rare events [6–8]. In MetaD one speeds up sampling by adding a history-dependent bias constructed by depositing Gaussian kernels on selected degrees of freedom, known as collective variables (CVs).

In well-tempered Metadynamics [9] this goal is achieved by periodically adding a bias that is built according to the iterative procedure

$$V_n(\mathbf{s}) = V_{n-1}(\mathbf{s}) + G(\mathbf{s}, \mathbf{s}_n) \exp\left(-\frac{1}{\gamma - 1} V_{n-1}(\mathbf{s}_n)\right) \tag{1}$$

where $\mathrm{V}_n(\mathbf{s})$ is the total bias deposited at iteration $n$ and is obtained by adding to $\mathrm{V}_{n-1}$ a Gaussian term $G(\mathbf{s}, \mathbf{s}_n)$ centered on the present value of the CV and scaled by the factor $\exp\left(-\frac{1}{\gamma-1} V_{n-1}(\mathbf{s}_n)\right)$. This factor makes the height of the added Gaussian decay with time.

The bias factor $\gamma$ determines the rate with which the added bias decreases and regulates the $\mathbf{s}$ fluctuations. In this way, the MetaD potential allows the system to explore the FES, escaping the free-energy minima and crossing large energy barriers at an affordable computational cost.

An advantage of well-tempered MetaD is that, after a transient period, the simulation enters a quasi-stationary limit in which the expectation value of any observable $\langle O(\mathbf{R})\rangle$ [10], can be estimated as running average according to:

$$\langle O(\mathbf{R})\rangle = \lim_{T\to\infty} \frac{\int_0^T O(\mathbf{R_t}) e^{\beta(V(\mathbf{s}(\mathbf{R_t}))-c(t))} dt}{\int_0^T e^{\beta(V(\mathbf{s}(\mathbf{R_t}))-c(t))} dt} \tag{2}$$

where $\mathbf{R}_t$ are atomic positions at time $t$, $\beta$ is the reciprocal of the temperature multiplied by the Boltzmann constant $\mathrm{k}_B$ and c(t) is a time dependent offset defined as:

$$c(t) = -\frac{1}{\beta} \log \frac{\int e^{-\beta(F(\mathbf{s})+V(\mathbf{s},t)} ds}{\int e^{-\beta(F(\mathbf{s})} ds} \tag{3}$$

where F($\mathbf{s}$) is the free energy as a function of the CV. In this way, one can reweight to Boltzmann averages of any variables even if they are not included in the biased CV.

### B. Variational Approach to Conformational dynamics Metadynamics

In MetaD the cost of reconstructing the free energy grows exponentially with the number of CVs, therefore, the number of CVs should be kept as small as possible. This condition prevent us from describing all relevant degrees of freedom of the system. This limitation is particularly restrictive in proteins where a slow process often involves crossing multiple small barriers, involving several degrees of freedom. In 2017 McCarty and Parrinello developed the Variational Approach to Conformational dynamics in Metadynamics (VAC-MetaD) [11] that allows combining several degrees of freedom into a number of CVs that are linear expansions of several descriptors. So far this approach has shed light into chemical reactions and the folding of small peptides [12, 13].

Here we review shortly the necessary steps for applying this procedure. Initially one identifies a set of descriptors $d_k(\mathbf{R})$ that are important for the process under study. Then one searches for the best CV $\mathbf{s}_i(\mathbf{R})$ that can be expressed as a linear combination of $\mathrm{d}_k(\mathbf{R})$:

$$\mathbf{s}_i(\mathbf{R}) = \sum_k^N b_{ik} d_k(\mathbf{R}) \tag{4}$$

This is achieved by a variant of Time-lagged Independent Conformational Analysis (TICA) developed in Noe's group [14, 15]. In this framework, one computes the time-lagged covariance matrix C($\tau$) defined as:

$$C_{mn} = \langle r_m(0) r_n(\tau) \rangle \tag{5}$$

where $\mathrm{r}_k(\tau)$=$\mathrm{d}_k(\tau)$-$\langle d_k \rangle$. The expansion coefficients are determined by solving the eigenvalue equation:

$$\overline{\mathbf{C}}(\tau)\mathbf{b}_i = \overline{\mathbf{C}}(0)\lambda_i \mathbf{b}_i \tag{6}$$

where $\overline{\mathbf{C}(\tau)}$ is the time lagged covariance matrix, $\overline{\mathbf{C}(\mathbf{0})}$ is the same matrix at time zero, $\lambda_i$ is the $i$th eigenvalues and $\mathbf{b}_i$ is the $i$th vector of coefficients of the CV $\mathbf{s}_i$. In Ref. [14, 15], one starts from the assumptuon that a trajectory that already crosses the barriers several times is available. McCarty and Parrinello [11] showed that an analogous procedure can be applied to a MetaD trajectory, if the MetaD timescale was properly rescaled:

$$d\tau = e^{\beta(V(\mathbf{s}(\mathbf{R_t}))-c(t))} dt \tag{7}$$

For the purpose of studying a protein conformational transitions, the eigenvectors of the eigenvalue equation obtained from a MetaD trajectory, correspond to the slowest decaying modes, and can be used as CVs in the production MetaD simulation [11, 16].

### C. Selecting the set of descriptors

In the application of VAC it is essential to choose the right set of descriptors. In this choice we shall be guided mostly by experimental facts. NMR chemical shifts indicate that residues 101 to 118, whose position is highlighted in Fig. 1 are mobile on the milisecond timescale. We choose to express their motion in terms of descriptors related to these residues such as dihedral angles and contacts, as detailed in Table S1 and Fig. S1. The rest of the descriptors come from an analysis of two short MD simulations in state G and E. These simulations showed that a salt bridge away from the hydrophobic cavity between E108 with R80, N81 and E108 with K35 did change in the two states. For this reason we expressed these two possibilities in terms of two coordination number descriptors.

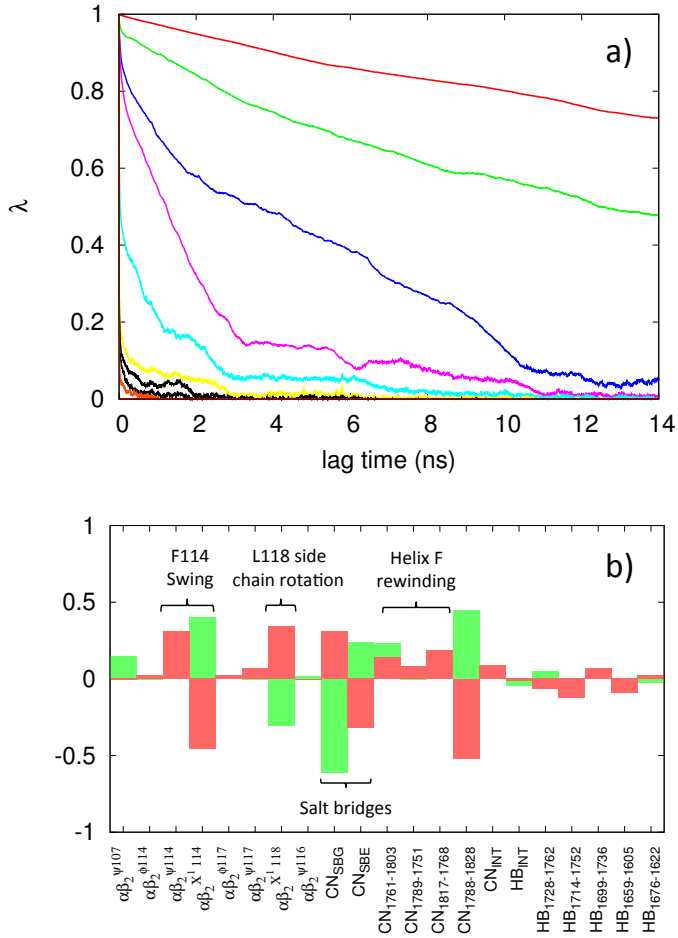


FIG. 2

FIG. 3: a) Eigenvalues decay as a function of lag time and b) coefficients of the set of descriptors for the two eigenvectors associated with the two first slowly decaying eigenvalues (red and green respectively).

### D. Protocol

We make use of the CHARMM27 and TIP3P forcefields for the protein and water respectively [17, 18]. All our MetaD simulations initiate from state G of L999A T4L, obtained from X-ray crystal structure (PDB ID: 3DMV). For more information about the system preparation we refer the reader to the SI. Briefly, we performed a 350 ns NPT MetaD simulation at 310 K and atmospheric pressure, starting from state G and biasing a suboptimal CV constructed as an equal weights linear combination of the descriptors mentioned in the subsection IIB. This led to the Free Energy Surface (FES) depicted in Fig. 4a where only

the G and E states are explored. After that we analysed this trajectory as described in subsection IIB and obtained the eigenvalues shown in Fig. 3a. It can be seen that the two top most eigenvalues decay more slowly than the others. We thus chose two corresponding eigenvectors as CVs, henceforth abbreviated $\mathbf{s}_1$ and $\mathbf{s}_2$. An analysis of these eigenvectors shows that not all 21 descriptors are relevant. As expected, the descriptors $\alpha\beta_2^{\psi_{114}}$,$\alpha\beta_2^{\chi^1_{114}}$ related to the swing of phenylalanine F114 from states G to E are relevant. Then, we find that the H-bond contacts descriptors ($\mathrm{CN}_{1761-1803}$,$\mathrm{CN}_{1789-1751}$,$\mathrm{CN}_{1817-1768}$,$\mathrm{CN}_{1788-1828}$) related to the rewinding of the F helix in going from G to E, play a role. Also relevant is the descriptor $\alpha\beta_2^{\chi^1_{118}}$ depicting the side-chain rotation of the hydrophobic residue L118. Finally, the relevance of the salt bridges between residues E108 and R80, N81 and residues E108 and K35 ($\mathrm{CN}_{SBG}$ and $\mathrm{CN}_{SBE}$) is apparent. It is also to be noted a tendency of the $\mathrm{CN}_{SBG}$ and $\mathrm{CN}_{SBE}$ to be in anti-phase. For a detailed explanation of the descriptors the reader can go to Table S1 and Fig. S1.

## III. RESULTS

Using the VAC-MetaD derived CVs $\mathbf{s}_1$ and $\mathbf{s}_2$ we performed a new 800 ns MetaD simulation, henceforth abbreviated as VAC-MetaD. Not only does the system go reversibly directly from G to E but remarkably it also follows a different route by passing through two other metastable intermediates, shown in Fig. 4b. One is the previously found $\mathrm{I}_{036}$, the other is a newly found metastable intermediate, abbreviated $\mathrm{I}_{misf}$. Both metastable intermediates were verified to be stable after running unbiased simulations of 20 ns for each of the states.

### A. Conformational free energies

In order to give quantitative insight on the population of the for states, we compute the free energy differences between the states in both MetaD and VAC-MetaD simulations. Regarding the technical aspects of this procedure we refer the interested reader to the SI.

The MetaD performed with suboptimal CVs gives a free energy difference between states G and E of 0.2 $\pm$ 2kJ/mol. This value is not close to the experimentally known value of 7.35 $\pm$ 0.8 kJ/mol[2] ns which we attribute to the fact that the CV is suboptimal and the simulation is not fully converged. However, the free energy difference between G and

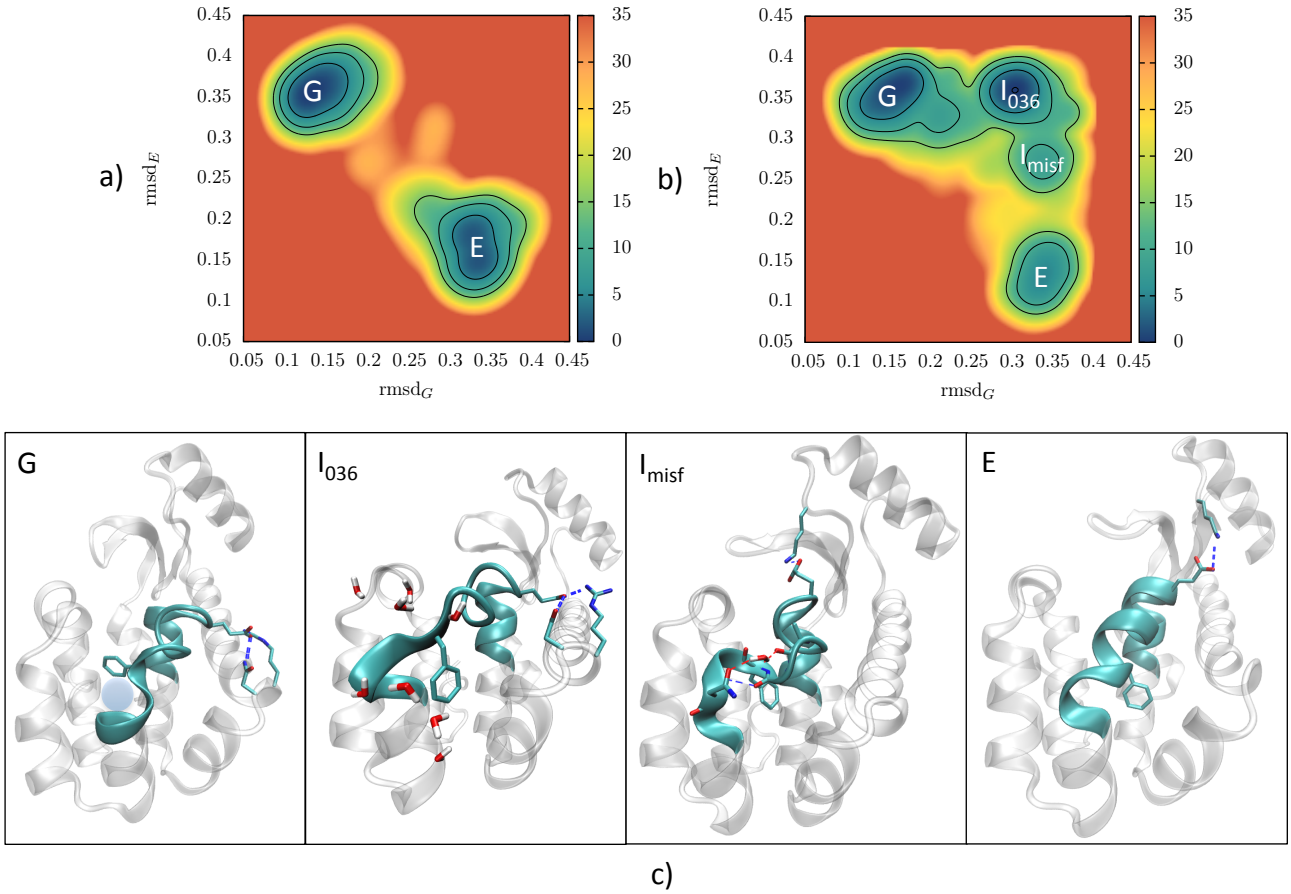


FIG. 4: FES in kJ/mol as a function of the $RMSD_G$ and $RMSD_E$ for a) the MetaD and b) the VAC-MetaD simulation. c) States G, $I_{036}$, $I_{misf}$ and E explored during the VAC-MetaD simulation. In shaded purple is depicted schematically the position of the hydrophobic cavity of state G.

E obtained from the VAC-MetaD simulations is 5.2 ± 2.0 kJ/mol, which is in excellent agreement with the experiment, thus proving the quality of these CVs. The free energy difference between G and $I_{036}$ is found to be 5.6 ± 4.4 kJ/mol and the one between G and $I_{misf}$ is 14.8 ± 4.6 kJ/mol. Notably, state $I_{misf}$ is found to be much higher in energy compated to state G. We attribute this to the fact that in $I_{misf}$ the protein assumes a misfolded conformation.

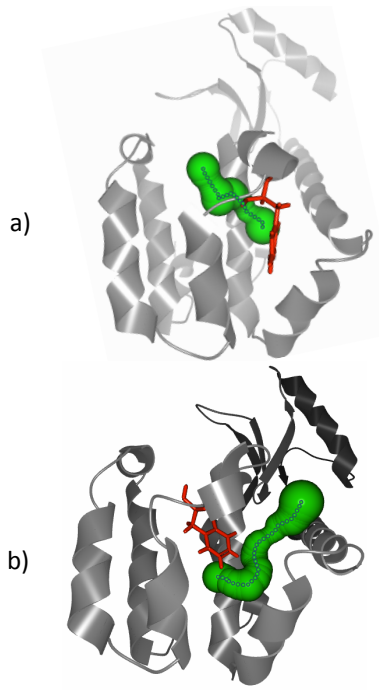


FIG. 5: The most probable tunnel formed in states a) $I_{036}$, formed between helices F and I and b) $I_{misf}$, formed between helices F,E and D.

### B. Description of the states

Here, we give a brief description of the four states found and we refer to the interested reader to the Supporting Information.

**State G.** Our VAC-MetaD simulation captures well the minimum of state G as shown in Fig. 4b and Fig. 4c,G . In agreement with the crystal structure of state G, we verify that F114 residue is turned with an angle $\psi$ at 49°, and buried inside the cavity, while making hydrophobic contacts with aminoacids of helices G, E, I [1, 3]. Another characteristic feature of this state is the type of winding of the first part of helix G (residues 116-121) with a kink at the position of residues 112-117. Note also the presence of the salt bridge between aminoacids E108 with R80 and N81 that helps stabilization of this state although distant from hydrophobic cavity. The hydrophobic cavity highlighted in Fig. 4c,G. is made inaccessible by the presence of phenylalanine F114.

**State $I_{036}$** Some of the most important features of this state are the rotation of F114 that is now exposed to the solvent, the partial unfolding of the last part of helix F (residues 113-115), and the permanence of the salt bridge between E108 with R80 and N81. The new F114

position is now stabilized by hydrophobic contacts with G and D helices. The formation of a cavity is able to accommodate a ligand, its shape and size can be seen in Fig. 5a.

**State E** Before describing $I_{misf}$ it is more illuminating to first describe the E state. Since $I_{misf}$ can be looked at as a modification of the E state in the same way as $I_{036}$ is best understood moving reference to G. In the E state the kink between helices G and F moves to the position of residues 115 to 120, and at the end of helix F hydrogen bonds between residues 111 and 116 are formed. Also, here there is a stabilizing salt bridge but now between E108 and K35.

**State $I_{misf}$** In this state a some H-bonds that stabilized helices G and F in state E are broken and now are replaced by new ones, a typical characteristic of misfolding. More precisely, direct H-bonds between F114 and S117 and a water mediated hydrogen bond between G113 and N116 are formed. Now F114 assumes an orientation intermediate between those that this residue has in G and E. The E108-K35 salt bridge present in state E is maintained. An elongated cavity that sits between the F, E and D helices is formed (see Fig. 5b). Contrary to state $I_{036}$ the diameter of this cavity is too small to allow a ligand insertion. The F114 is now in an intermediate position in the hydrophobic cavity between states G and E. Like in state E, there is a stabilizing salt bridge between E108 and K35.

## IV. DISCUSSION

In this study we implemented for the first time the VAC-MetaD framework to a conformational transition of the moderate size protein, L99A T4 Lysozyme. Our aim was to obtain insight in the slow degrees of freedom, by looking at the slowly decaying eigenvectors of the VAC-MetaD analysis, and accelerate convergence using these dynamically optimized CVs and unravel the complex free energy landscape of this conformational transition.

Combining experimental knowledge from NMR and short MD simulations in states G and E and applying the VAC-MetaD framework to an initial MetaD trajectory we unravelled the slow degrees of freedom. In particular these involve, the swing motion of phenylalanine F114 in going from G to E, the H-bond contacts related to the rewinding of helix F in going from G to E, a side chain rotation of Leucine L118 in the hydrophobic cavity and finally, a salt bridge formation away from the hydrophobic cavity between residues E108 and K35 or residues E108 and R80 and N81. Convergence was much improved using the VAC-MetaD

optimized CVs, compared to suboptimal CVs. The free energy difference between state G and E was found in very good agreement with experiments.

Strikingly, while the MetaD simulation only contained direct transitions between states G and E, the VAC-MetaD optimized CVs contained more information, and when biased where able to drive the system transitioning through an unprecedented route, passing through two other metastable intermediate states $I_{036}$ and $I_{misf}$. The first is a previously found metastable intermediate state able to accommodate a ligand, and the second a newly found state with a misfolded F helix, higher in energy and unable to accommodate a ligand.

This work, proves the power of VAC-MetaD in accelerating convergence of the free energy, exploring the underlying free energy landscape and providing chemical insight through the VAC-MetaD eigenvectors. We see VAC-MetaD analysis as a tool that achieves all three goals and is up to the user whether he pursuits all of them. For example, a VAC-MetaD analysis of an initial MetaD or MD trajectory, is already insightful giving acess to the slow degrees of freedom.

A downside of the VAC-MetaD is that of the initial trajectory necessary to to optimize the descriptor set CVs. Hence, we anticipate that faster methods such as Variational Enhanced Sampling and/or multiple walkers would help obtaining the initial trajectory faster.

**ACKNOWLEDGEMENTS**

This research was supported by the VARMET European Union Grant ERC-2014-ADG-670227. Computational resources were provided by the Swiss National Supercomputing Centre (CSCS).

---

[1] G. Bouvignies, P. Vallurupalli, D. F. Hansen, B. E. Correia, O. Lange, A. Bah, R. M. Vernon, F. W. Dahlquist, D. Baker, and L. E. Kay, Nature **477**, 111 (2011).

[2] F. A. A. Mulder, A. Mittermaier, B. Hon, F. W. Dahlquist, and L. E. Kay, Nat. Struct. Biol. **8**, 932 (2001).

[3] A. E. Eriksson, W. A. Baase, J. A. Wozniak, and B. W. Matthews, Nature **355**, 371 (1992).

[4] M. Merski, M. Fischer, T. E. Balius, O. Eidam, and B. K. Shoichet, Proc. Natl. Acad. Sci. **112**, 5039 (2015).

[5] Y. Wang, E. Papaleo, and K. Lindorff-Larsen, Elife **5**, 1 (2016).

[6] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).

[7] A. Barducci, G. Bussi, and M. Parrinello, Phys. Rev. Lett. **100**, 1 (2008).

[8] A. Barducci, M. Bonomi, and M. Parrinello, Wiley Interdiscip. Rev. Comput. Mol. Sci. **1**, 826 (2011).

[9] M. Bonomi and M. Parrinello, Phys. Rev. Lett. **104**, 1 (2010).

[10] P. Tiwary, J. F. Dama, and M. Parrinello, J. Chem. Phys. **143**, 1 (2015).

[11] J. Mccarty, M. Parrinello, J. Mccarty, and M. Parrinello, J. Chem. Phys. **147**, 204109 (2017).

[12] G. Piccini, D. Polino, and M. Parrinello, J. Phys. Chem. Lett. **8**, 4197 (2017).

[13] Y. I. Yang and M. Parrinello, J. Chem. Theory Comput. **14**, 2889 (2018).

[14] G. Pérez-hernández, F. Paul, T. Giorgino, G. D. Fabritiis, and F. Noé, **015102-13** (2013).

[15] F. Noé and F. Nüske, Multiscale Model. Simul. **11**, 635 (2013).

[16] M. M. Sultan and V. S. Pande, J. Chem. Theory Comput. Theory Comput. **13**, 2440 (2017).

[17] W. L. Jorgensen and J. D. Madura, J. Am. Chem. Soc. **105**, 1407 (1983).

[18] C. Blanchet, M. Pasi, K. Zakrzewska, R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, K. Zakrzewska, D. B. Nikolov, H. Chen, E. D. Halay, A. Hoffmann, R. G. Roeder, S. K. Burley, E. Y. D. Chua, D. Vasudevan, G. E. Davey, B. Wu, C. A. Davey, N. Foloppe, A. D. Mackerell, M. Feig, C. L. Brooks, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, M. Karplus, N. K. Banavali, P. Bjelkmar, P. Larsson, M. A. Cuendet, B. Hess, and E. Lindahl, J. Chem. Theory Comput. **6**, 459 (2010).