bioRxiv preprint doi: https://doi.org/10.1101/158592; this version posted July 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Transferable tICA-Metadynamics: Efficient sampling of protein mutants by transferring information from the wild type's Markov state model

Mohammad M. Sultan[1], & Vijay S. Pande[1†]
[1]Department of Chemistry, Stanford University, 318 Campus Drive, Stanford, California 94305, USA.
[†]Pande@stanford.edu

Abstract:

We recently showed that the time-structure based independent component analysis method from Markov state model literature provided a set of variationally optimal slow collective variables for Metadynamics (tICA-Metadynamics). In this paper, we extend the methodology towards efficient sampling of protein mutants by borrowing ideas from transfer learning methods in machine learning. Our method explicitly assumes that a similar set of slow modes and states are found in both the wild type and its mutants. Under this assumption, we describe a few simple techniques using sequence mapping for transferring the slow modes and structural information contained in the wild type simulation to a mutant model for performing enhanced sampling. The resulting simulations can then be reweighted onto the full-phase space using MBAR, allowing for thermodynamic comparison against the wild type. We first benchmark our methodology by re-capturing alanine dipeptide dynamics across a range of different atomistic force fields after learning a set of slow modes using Amber ff99sb-ILDN. We next extend the method by including structural data from the wild type simulation and apply the technique to recapturing the affects of the GTT mutation on the FIP35 WW domain.

Introduction:

Efficient sampling of protein configuration space remains an unsolved problem in computational biophysics. While algorithmic advances in molecular dynamics (MD) code bases[1] combined with distributed computing hardware[2], specialized chips[3], and large-scale increasingly faster GPU clusters have provided routine access to microsecond timescale dynamics, there is still room for significant improvements. One such potential avenue is predicting the affects of mutations onto the protein's free energy landscape. Under the current scheme, one would have to re-run our entire simulation in order to ascertain the affects of a mutation onto a protein's free energy landscape. Due to the vast amount of computational resources required for even one simulation, most current MD papers run one simulation in a single force field for a single protein. However, considering the important role of mutagenesis experimentally as biophysical probes, the

bioRxiv preprint doi: https://doi.org/10.1101/158592; this version posted July 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

biological role of SNPs in medicine and disease, as well as phylogenic and evolutionary questions connecting mutations, often there are hundreds to thousands of mutations (or more) that would be relevant for simulation. Instead, the predictions from these simulations are extrapolated to other conditions but those changes/mutations are often not explicitly tested in-silico. While such hypothesis generation is useful for guiding future work, the gap between extrapolated predictions and experimental realization is large..

There is an obvious scaling problem between the computational and time cost of unbiased MD and the number of interesting mutants that could be investigated using simulation. For example, there are several hundred known protein kinases[4,5] with each having tens to hundreds of known mutants. These kinases have critical protonation and phosphorylation sites that significantly affect their free energy landscapes[6,7]. To predict these mutations' effects, do we need to re-run an entirely new simulation on the mutated protein? Are modern force fields even capable of elucidating such effects? Even if we assume an accurate enough force-field[8,9], how do we efficiently sample these mutants or perhaps even propose new novel variants to be probed via experimental assays. Arguably, for MD to decrease the gap between theoretical hypothesis and experimental realization, an ability to efficiently sample the effects of mutations is required. Since unbiased MD is too slow, we turn to enhanced sampling.

While enhanced sampling methods such as Metadynamics or Umbrella sampling offer promise, they require identification of a set of collective variables (CVs)[10] to sample along. Metadynamics[10–13] can be thought of as computational sand filling along CV of interest to enhance sampling between kinetically separate regions. Therefore, these CVs should correlate with the slowest structural degrees of freedom within the system, and exclusion of slow modes leads to hysteresis and convergence issue[12,14]. For example, even for the simplest test cases such as capped Alanine dipeptide, hysteresis can arise if we choose the faster ψ coordinate for enhanced sampling.

Given all of these problems with enhanced sampling algorithms, we instead aim to solve a simpler problem. What if we are given unbiased MD simulations for the wild type (WT) and we wish to learn the dynamics for a closely related mutant? The mutant could correspond to a change in force field, an amino acid substitution, post-translational modifications, or even an alternative drug in the case of drug-binding simulations. We expect that these mutants likely sample a very similar free energy landscape, albeit with different thermodynamics and kinetics. Could we design a better sampling scheme by transferring knowledge from the WT simulation to the mutant?

Transfer learning[15,16] is a method from the machine learning literature where knowledge learnt from modeling one task is transferred to the model for the purpose of learning another task. We wish to replicate a similar effect in molecular modeling where we transfer the knowledge learnt from a protein's wild type to a simulation of its mutant. Ultimately, we aim to efficiently sample the mutant to predict affects of force field changes, post translation modifications, and/or amino-acid substitutions etc.

bioRxiv preprint doi: https://doi.org/10.1101/158592; this version posted July 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The idea of knowledge transfer is not new in computational biophysics. Researchers constantly use homology modeling[17] to create models for systems which have not been crystallized or select CVs for enhanced sampling simulations[10] based upon an intuition learnt from failed runs, literature search, or previously published modeling work on homologous systems. However, this is often done in an ad-hoc or heuristic fashion. For example, it might be difficult to find the "right" template for homology modeling when a large set of similar sequence identity structures are available.

We hypothesize an efficient use of transfer learning would maximally leverage the *reaction coordinates, thermodynamic, and structural* information contained in the WT simulation. Our *key* results stem from recognizing that protein mutants sample a similar set of free-energy minima connected via similar slow modes. Our model assumes that these slow modes involve the same set of residues across the WT and mutant sequences and all that remains are identifying those slow modes (Figure 1) in the WT simulation[12] and transferring them on to a mutant simulation.

We propose transferring information from the WT's tICA (time-structure based independent component analysis) model and MSM (Markov state model) to the mutant Metadynamics or Umbrella sampling simulations (Figure 1). tICA is a dimensionality reduction technique[18–21] capable of finding reaction coordinates(tICs) within the dataset. are kinetic models of protein dynamics that model the dynamics as memory-less jump processes. tICA was initially used as a dimensionality reduction process[21] for defining the Markov models' state space though it was later shown that both tICA and MSM solve the same problem[22] of approximating the underlying transfer operator, albeit with a differing choice of basis. The tICA[19–21] method has non-linear[18] extensions available which significantly improve its descriptive abilities. Furthermore, a variational principle[22] for tICA and MSMs allows a researcher to systematically validate[23] modeling parameters to potentially integrate out subjective modeling decisions. We recently showed that these tICs[21] provided a set of excellent CVs for enhanced sampling via Metdaynamics[12] or other schemes. Therefore, we hypothesize the answer lies in transferring these tICs over from one simulation to another.

But how do we transfer these slow tICA coordinates? At this point it is worth recalling that tICA is a linear combination of input features[12,18,19,21,24]. These input features are a set of real numbers encoding the protein's conformational state and concretely might be dihedrals or contacts or RMSD to a set of landmark points. Furthermore, these features might be the result of a non-linear transform such as a Guassian kernel[12,24]. Therefore, what we wish to compute are these protein strucutral features for a new closely related sequence (Figure 1). For this, we will need to determine a set of features that can be applied to both the WT and mutant system after performing a structural or sequence alignment (Figure 1). For example, this might involve figuring out the equivalent atom indices for backbone dihedrals/contact distances/rmsds etc that make up the set of features used to construct the WT's tICA model. Once such a mapping has been established, it is straightforward to transfer the linear combinations that make up the slowest modes for enhanced sampling simulations. In practice, we find we only have to modify small parts of input scripts that are fed into Plumed[25] for performing the enhanced sampling simulations.

bioRxiv preprint doi: https://doi.org/10.1101/158592; this version posted July 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Our method explicitly makes the following set of assumptions:
1). The wild type and mutant proteins occupy similar set of configurations in phase space, are connected via similar pathways, and have a similar set of slow modes.
2). The wild type simulation captures a large portion this accessible phase space, and tICA and MSMs correctly enumerate these slowest modes.
3). We can calculate equivalent features for the mutant and WT proteins.

There has been some previous work in using MSMs for efficient sampling of protein mutants. In particular, Voelz et al.[26] used an information theoretic approach to find maximally surprising changes to a mutant MSM for performing new rounds of iterative sampling. However, their approach requires at least partial convergence of a rudimentary mutant MSM before such comparisons can be made. The amount of sampling required to make this rudimentary MSM could easily exceed the sampling of the WT, e.g. if the mutation slows down the dominant kinetics by an order of magnitude. Furthermore, at least initially, the rudimentary mutant MSM is likely to have large statistical uncertainties, potentially leading to false positives for the suprisal/self-information distance metric proposed in the paper[26]. Here, we are approaching the mutant problem from a fundamentally different perspective that aims to cannibalize all available data in the WT MSM.

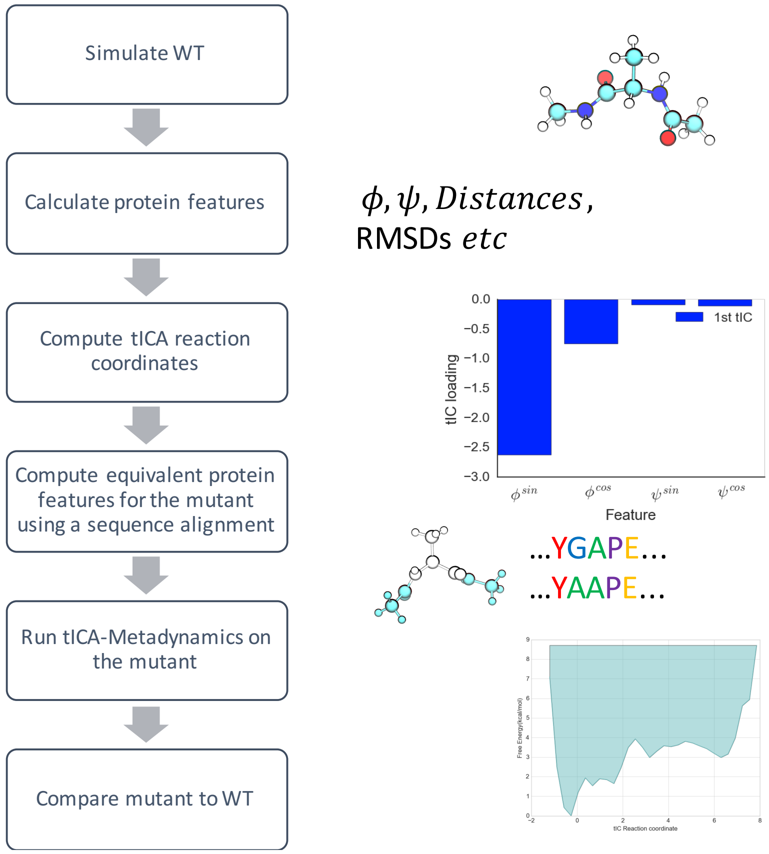

Figure 1: Pictorial representation for the presented method. Starting from the WT simulation, we first calculate wild type protein features. We then reduce the dimensionality keeping only the slowest modes in the system. We then transfer the tIC loadings to the mutant for faster sampling of the mutant.

bioRxiv preprint doi: https://doi.org/10.1101/158592; this version posted July 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Transferable tICA-Metadynamics is an efficient way to sample mutations

We begin by showing as a simple proof of concept that the dynamics of alanine dipeptide can be re-captured across several FFs after learning the slowest modes in the "WT" model (Amber99sb-ildn[8]). We downloaded a previously generated dataset[27] that contained 4μs of capped Alanine dipeptide run using the Amber99sb-ildn force field (FF)[8]. We then trained a tICA model on the backbone dihedrals at a lagtime of 1ns. As shown in Figure 2a, the tICA model captures the slowest mode as corresponding to movement in and out of the α_L basin while the next mode is flux in and out of the α_R basin. We next ran bias-exchange[10,11] tICA-Metadynamics simulations in 3 different FFs (Amber99sbiln, Charmm27, and Amber03). The exact parameters for the well-tempered Metadynamics runs are given in SI table 1, though we empirically found that a range of parameters worked. All MD trajectories were run in the NPT ensemble with a MonteCarlo Barostat (1 atm), a Langevin integrator (300 K), and a 2 fs timestep. We used the PME method[28] to handle long range electrostatics using a 1nm cutoff. The simulations were performed on GPUs using OpenMM[29,1] and Plumed[25]. After running the Metadynamics simulation, we combined the data across the two tICs using Multi-state Bennett Acceptance Ratio (MBAR)[30,31] algorithm. For each simulated frame, we used the last reported bias across the tIC CVs as an estimate for input into the MBAR algorithm.

The results are given in Figure 2b and 2c. We explicitly projected the Charmm27 and Amber03 datasets[27] using Amber99sb-ildn's state decomposition, allowing us to compare the models across force fields without having to worry about state equivalence. It can be seen that our sampling scheme efficiently learns the differences between the dynamics upon mutating the force field from Amber99sb-ildn to Charmm27 or Amber03 (Figure 2b). For example, the α_L basin in Amber03 is significantly higher in free-energy (Figure 2c) compared to Amber99sb-ildn and Charmm27.

bioRxiv preprint doi: https://doi.org/10.1101/158592; this version posted July 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

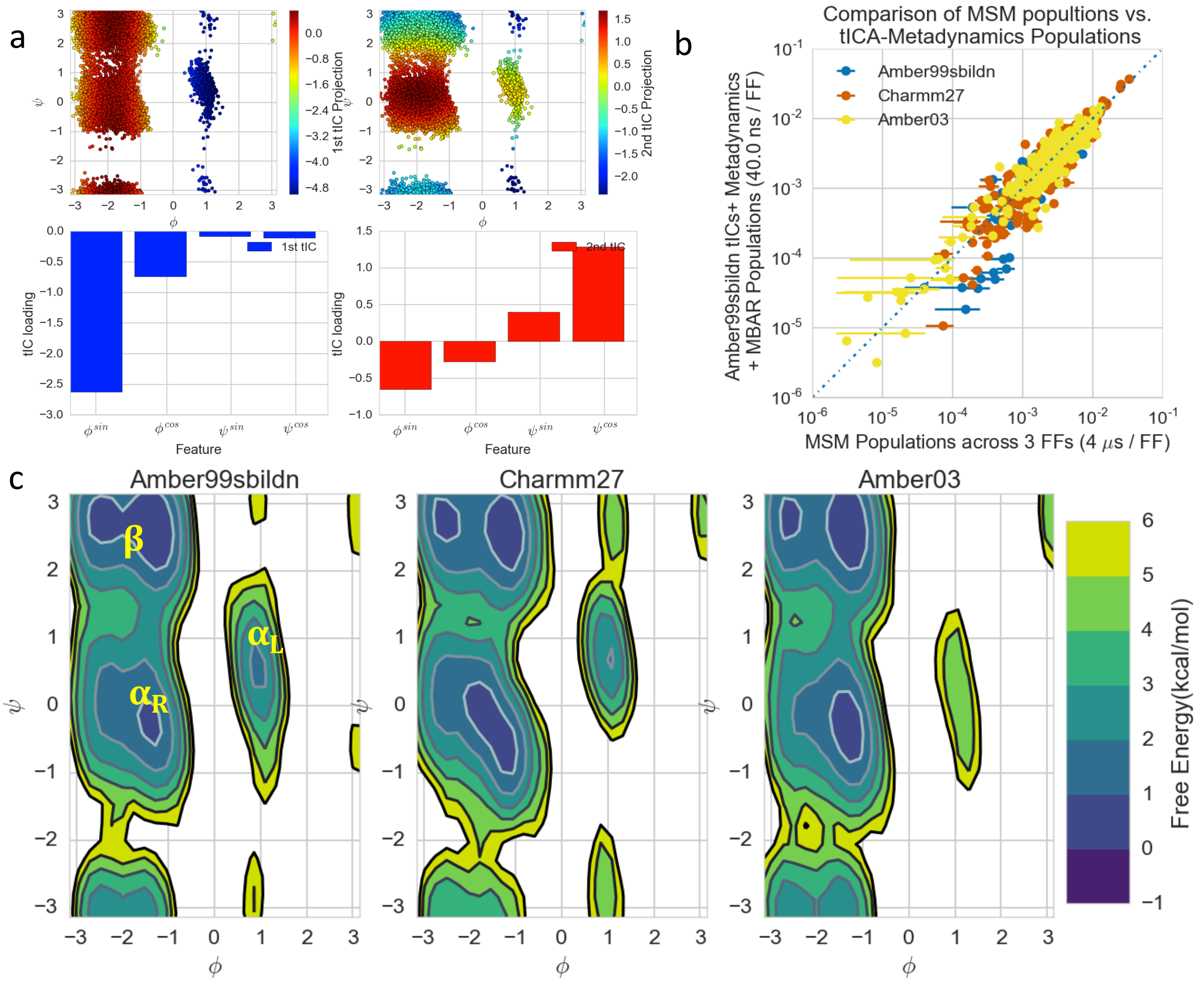

Figure 2: Transferable tICA-Metadynamics is an efficient method to understand the effect of mutations. a). Top two slowest tIC modes for capped Alanine learnt from unbiased MD using the Amber99sbildn model. b). Scatter plot of MSM microstate populations versus the tICA-Metadynamics inferred microstate populations. The error bars represent two standard deviations and are obtained by 10 rounds of bootstrapping. We used Amber99sbildn's state definition across all 3 FFs for a systematic comparison. c). Reweighting the starting Ambe99sb-ildn MD dataset using tICA-Metadynamics population across 3 FFs. The 0kcal/mol is defined using Amber99sb-ildn's most heavily populated state.

Transferable tICA-Metadynamics can use Wild type simulation's structural data by coupling to a MSM structural reservoir

Up to this point, our modeling efforts have only focused on using the slow tICs within the WT simulation for efficiently sampling the mutant. This might be sufficient for small peptides systems but is unlikely to work for large systems due to for example missing structural features in the construction of our tICA coordinates. While we could systematically improve the quality of our tICA model via the variational analysis[22], there is always a finite chance of missing structural degrees of freedom. To overcome this, we recommend coupling the Metadynamics simulations to a structural reservoir containing structures sampled from the WT MSM simulation (Figure 3a). Then, all that remains is creating a proposal distribution and an acceptance criterion for inserting the WT MSM state into the mutant Metadynamics simulation (Figure 3a). Ordinary Bias-Exchange[10,32] swaps protein coordinates according the following criterion:

bioRxiv preprint doi: https://doi.org/10.1101/158592; this version posted July 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$P_{accept} = \min\left(1, \exp\left[\beta\left(V^a(x^a,t) + V^b(x^b,t) - V^a(x^b,t) - V^b(x^a,t)\right)\right]\right)$$

where $V^a(x^a,t)$ is the Metadynamics bias potential acting on coordinates, x^a, of replica a at time t. However, since a MSM structural reservoir has no external bias acting on it, we change the swap probability to an insertion (from MSM to Metadynamics) probability:

$$P_{insert\ MSM \rightarrow tICA\ Simulation} = \min\left(1,\ \exp\left[\beta\left(V^a(x^a,t) - V^a(x^{MSM},t)\right)\right]\right)$$

where x^{MSM} are the coordinates for the MSM state under consideration. If accepted, the MSM state is put into the mutant Metadynamics simulation. Given enough sampling, this scheme resembles a Metropolis step. To improve the acceptance probability, we used the WT Markovian transition model to *propose* a transition state after figuring out the mutant's current MSM state within the simulation. Using the WT transition matrix provides an excellent proposal distribution since we hypothesize that the mutant only minimally perturbs certain elements of the matrix. Our reservoir approach is similar to the high-temperature reservoir introduced by Okur et al[33], though in this instance, the ensemble of structures is obtained via a regular MD run, and the proposal is dealt using the WT transition matrix. While the WT MSM transition matrix serves as an excellent proposal distribution it is also possible to use other proposal distributions such as the uniform distribution. Furthermore, several sampling techniques from the MonteCarlo literature such as the Wang-Landau scheme can be employed as well. We note that for mutant simulations, generating this MSM state reservoir would require additional steps of homology[17] modeling, minimization and equilibration, though this is a pleasantly parallelized problem.

bioRxiv preprint doi: https://doi.org/10.1101/158592; this version posted July 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

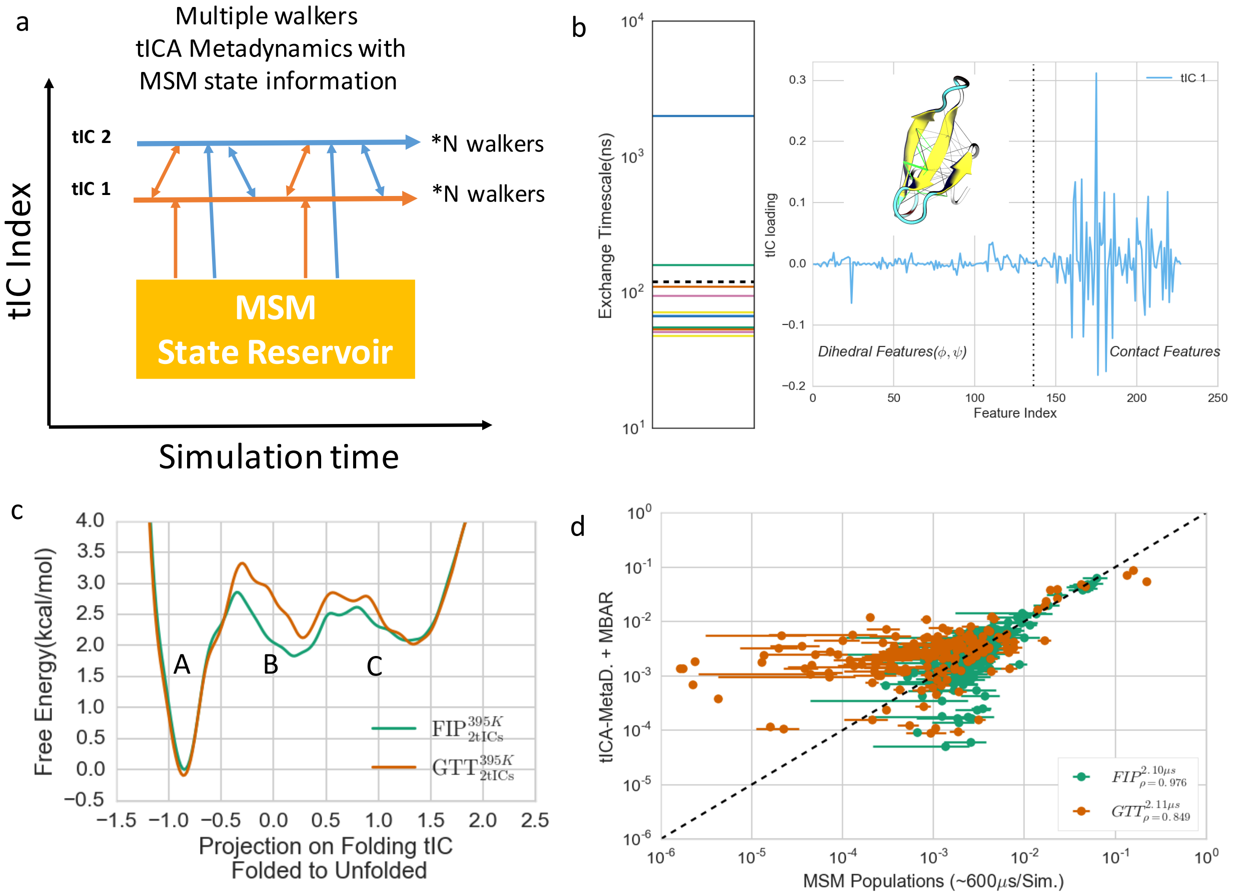

Figure 3: tICA-Metadynamics can be used to understand the effects of mutations upon folding simulations . a). Model of the multiple walkers' swap scheme used within this example. The double sided arrows indicate swap attempts while the single sided arrows indicate insertion attempts from the MSM state reservoir. b). Left Panel: Exchange timescales for the backbone dihedral and contact tICA model built using the Anton WW-FIP mutant simulation data. We decided to sample every coordinate up to the second one (dashed line) because all other modes had exchange timescales <= 100ns. b). Right Panel: tIC loading for the dominant folding tIC shows that it is a complex combination of both the WW's backbone dihedrals and contacts. The inset shows the contact distances used to build the model with the most important (highest tIC loading) distances shown in green. c). PMFs obtained from the Metadynamics simulations using the scheme highlighted in a). The GTT mutants' unfolded state is de-stabilized relative to the folded state. In both cases, we used the WW-FIP mutant to define the 0kcal/mol. The total amount of sampling across all the walkers was about 1-2µs per mutant. d). Comparison of the Anton-MSM state-populations vs. the MBAR reweighted tICA-Metadynamics results. The MSM error bars represent two standard deviations obtained via 10 rounds of bootstrapping. The legend shows both the sampling per mutant, and the correlation coefficients between the MSM thermodynamics vs the Metadynamics populations. Protein images were generated using VMD[34], while the graphs were generated using IPython Notebook[35] and MSMExplorer[36].

Lastly, it is possible to use a neutral replica within this setup. The neutral replica has no external bias acting on it and approximately samples from the canonical distribution. However, if a neutral replica is used, we recommend only allowing the neutral replica to swap with the biased replicas since an appropriate asymptotically correct swapping criterion for swapping between the neutral replica and the MSM state reservoir doesn't exist.

We tested our methodology by predicting the effects of the GTT mutation upon the folding of the WW domain[3,37,38]. We began by learning a tICA model (50ns lagtime) on the backbone

bioRxiv preprint doi: https://doi.org/10.1101/158592; this version posted July 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

dihedrals and selected contacts for the WT mutant (WW-FIP). We kept the top 15 tICs, and made a MSM at a 50ns lag time on a 200 state model. Our tICA model indicated that the slowest mode (Exchange time scale > 1 μs) corresponded to the folding while the second slowest mode (Exchange time scale > 100 ns) corresponded to formation of an off-pathway register shifted state (Figure 3b-c). Since every subsequent slow tIC mode has exchange timescales of less than ~ 100ns, we chose to focus our sampling on these two tICs. We ran the simulations for *both* the FIP35 WT protein and the GTT triple mutant for a more systematic comparison. Similar to previous work[37], all the simulations were performed in the NVT ensemble with a 2fs time step at 395K. We used the PME method[28] to handle long range electrostatics using a 1nm cutoff. The simulations were performed on GPUs using OpenMM[29,1] and Plumed[25]. After running the Metadynamics simulations, we used MBAR to re-weight to the MSM state space and obtained the PMFs along the dominant tIC. All relevant simulations parameters are shown in SI Table 2.

The results for both of our enhanced sampling simulations is given in Figure 3d. Two different insights emerge from our enhanced sampling scheme relative to the Anton results (SI Figure 1). Similar to the Anton simulations, our FIP unfolded state (Figure 3d, tiC value >-0.25) has a distinct two state behavior. Basin ‘C’ corresponds to the unfolded and collapsed state. This basin also includes an off-pathway register shifted state. The second high free-energy basin (Figure 3d, B) is an on-pathway intermediate state where two of the three beta-strands have formed. The unfolded state in our ensemble is more populated than in the Anton simulations. These on and off-pathway intermediate states were not detected in the original two-state folding reaction coordinate for the WW domain[37,38] though it was later found from the simulations using a variety of techniques[39]. We note that our tICA analysis was able to identify the on-pathway folding intermediate and the off-pathway state as the top two slowest modes (tICs) within our model.

As can be seen in Figure 3c-d, our simulations indicate that the GTT mutant de-stabilizes the unfolded state and the on-path intermediate state, leading to increased folded population and faster folding timescales. These results are in line with the previous computational and experimental work[37] though our simulations required about 200-300x less aggregate sampling(~1-3μs vs 600μs). More importantly, the current sampling was performed in parallel so that no single walker had to be run for more than 50-200ns (~3-7 days on K40 GPUs using OpenMM[1] and Plumed[25]). We also believe it might be possible to optimize this further by modifying the Metadynamics parameters and Metropolis swap schemes/rate.

Transferable tICA-Metadynamics can use Wild type simulation’s thermodynamic data as a prior for the underlying free energy landscape

Lastly, we turn to efficiently using the thermodynamic information contained in the WT simulation. To that end, we recommend using the WT simulation to identify minimum values along each tIC coordinate, aka the thermodynamic minima, to plug into a variant of Metadynamics, namely Transition-Tempered Metadynamics (TTMetaD)[40]. In TTMetaD, the Gaussians heights are scaled according to the number of trips between basins. We also believe that it is possible use to the WT free energy surface as a Bayesian[41] prior for the mutant Metadynamics simulation , though that is beyond the scope of this work. The latter might involve

bioRxiv preprint doi: https://doi.org/10.1101/158592; this version posted July 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

starting off with a 'partially' constructed free energy-landscape such that the Metadynamics engine only has to fill in the regions that are different between the WT and the mutant.

Our current results open up several interesting avenues for future work. For example, up to this point, we have only focused on enumerating the thermodynamic differences between the mutants. However, the recent work in kinetic reweighting either via Maximum caliber[42] , TRAM[43], or plain transition state theory could potentially be used to obtain the an estimate for the mutants' perturbed kinetics. This raises the intriguing possibility of getting estimates for both the kinetic and thermodynamics of a mutant simulation for a miniscule fraction of the WT's compute cost. An excellent application for this would be the ability to predict changes in a drug's binding and unbinding kinetics. Our approach explicitly includes all of the protein's slow conformational modes, in addition to the drug binding mode—making it more accurate.

One possible problem with our current approach is the determination of how far we can move away from the WT in sequence space before the transfer approach fails. Are the tICs learnt from a WT simulation applicable to a sequence with minimally sequence similarity? What is the distance metric and how do we define minimal? A similar problem is faced in homology modeling, where the quality of the model depends on the underlying sequence conservation. It is possible that the heuristic value of 40-50% sequence identity cutoff used in homology modeling might be applicable here too, but we concede that that value is simple conjecture at this point.

A more involved solution to this problem is to consider clustering the entire sequence super family. For example, there are 518 known human kinases[5]. One could potentially cluster the sequences using evolutionary distance metrics into *m* representative sequence clusters, where *m* is the number of possible unbiased simulations that can be performed. Those *m*- simulations are then run and analyzed via tICA and Markov models. It is worth noting that the simulations for the *m* sequences are perfectly parallelizable, allowing for synergistic collaborations between different research institutes. For all other sequences, we can then use the tICs from its closest cluster center or perhaps even combine the tICs from the k nearest neighbors.

To summarize, we present a new method Transferable tICA–Metadynamics for the efficient sampling of protein mutations by transferring the reaction coordinates, structural, and thermodynamic data from the WT simulation to the mutant. Our method explicitly assumes that the WT and the mutant share a similar set of slow modes. Under this assumption, we then show that the slow modes of the WT can be transferred to the mutant simulation by computing an equivalent set of protein structural features. This requires using a protein structural alignment to identify equivalent residues which is readily possible using modern software[44,45]. We benchmarked our method on two test cases showing how switching force field in alanine dipeptide causes shifts in the propensity and location of the α_L basin, and recapturing the previous results that the GTT mutant of WW domain stabilizes the active state.

Acknowledgements:

bioRxiv preprint doi: https://doi.org/10.1101/158592; this version posted July 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The authors would like to thank various members of the Pande lab for useful discussions and feedback on the manuscript. M.M.S would like to acknowledge support from the National Science Foundation grant NSF-MCB-0954714.

Code and data availability:

All the code needed to reproduce the main results of this paper is available at https://github.com/msultan/tica_metadynamics .

(1) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *bioRxiv* **2016**.

(2) Shirts, M.; Pande, V. S. Screen Savers of the World Unite! *Science* **2000**, *290*, 1903–1904.

(3) Shaw, D. E.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossváry, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Deneroff, M. M.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. In *Proceedings of the 34th annual international symposium on Computer architecture - ISCA '07*; ACM Press: New York, New York, USA, 2007; Vol. 35, p 1.

(4) Parsons, S. J.; Parsons, J. T. Src Family Kinases, Key Regulators of Signal Transduction. *Oncogene* **2004**, *23* (48), 7906–7909.

(5) Taylor, S. S.; Kornev, A. P. Protein Kinases: Evolution of Dynamic Regulatory Proteins. *Trends Biochem. Sci.* **2011**, *36* (2), 65–77.

(6) Shan, Y.; Seeliger, M. A.; Eastwood, M. P.; Frank, F.; Xu, H.; Jensen, M. Ø.; Dror, R. O.; Kuriyan, J.; Shaw, D. E. A Conserved Protonation-Dependent Switch Controls Drug Binding in the Abl Kinase. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (1), 139–144.

(7) Xiao, Y.; Lee, T.; Latham, M. P.; Warner, L. R.; Tanimoto, A.; Pardi, A.; Ahn, N. G. Phosphorylation Releases Constraints to Domain Motion in ERK2. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (7), 2506–2511.

(8) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins* **2010**, *78* (8), 1950–1958.

(9) Best, R. B.; Buchete, N.-V.; Hummer, G. *Are Current Molecular Dynamics Force Fields Too Helical?*; 2008; Vol. 95.

(10) Abrams, C.; Bussi, G. Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy* **2013**, *16* (1), 163–199.

(11) Pfaendtner, J.; Bonomi, M. Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics. *J. Chem. Theory Comput.* **2015**, *11* (11), 5062–5067.

(12) M. Sultan, M.; Pande, V. S. tICA-Metadynamics: Accelerating Metadynamics by Using Kinetically Selected Collective Variables. *J. Chem. Theory Comput.* **2017**, acs.jctc.7b00182.

(13) Sun, R.; Dama, J. F.; Tan, J. S.; Rose, J. P.; Voth, G. A. Transition-Tempered Metadynamics Is a Promising Tool for Studying the Permeation of Drug-like Molecules through Membranes. *J. Chem. Theory Comput.* **2016**, *12* (10), 5157–5169.

bioRxiv preprint doi: https://doi.org/10.1101/158592; this version posted July 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(14) Tiwary, P.; Berne, B. J. Spectral Gap Optimization of Order Parameters for Sampling Complex Molecular Systems. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (11), 2839–2844.

(15) Torrey, L.; Shavlik, J. Transfer Learning.

(16) Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How Transferable Are Features in Deep Neural Networks? **2014**.

(17) Eswar, N.; Webb, B.; Marti-Renom, M. a; Madhusudhan, M. S.; Eramian, D.; Shen, M.-Y.; Pieper, U.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Protein Sci.* **2007**, *Chapter 2* (November), Unit 2.9.

(18) Schwantes, C. R.; Pande, V. S. Modeling Molecular Kinetics with tICA and the Kernel Trick. *J. Chem. Theory Comput.* **2015**, *11* (2), 600–608.

(19) Pérez-Hernández, G.; Noé, F. Hierarchical Time-Lagged Independent Component Analysis: Computing Slow Modes and Reaction Coordinates for Large Molecular Systems. *J. Chem. Theory Comput.* **2016**, acs.jctc.6b00738.

(20) Noé, F.; Clementi, C. Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2015**, *11* (10), 5002–5011.

(21) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9* (4), 2000–2009.

(22) Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. Variational Approach to Molecular Kinetics. *J. Chem. Theory Comput.* **2014**, *10* (4), 1739–1752.

(23) McGibbon, R. T.; Pande, V. S. Variational Cross-Validation of Slow Dynamical Modes in Molecular Kinetics. *J. Chem. phyics* **2015**, *142* (12).

(24) Harrigan, M. P.; Pande, V. S. Landmark Kernel tICA For Conformational Dynamics. *bioRxiv* **2017**.

(25) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185* (2), 604–613.

(26) Voelz, V. A.; Elman, B.; Razavi, A. M.; Zhou, G. Surprisal Metrics for Quantifying Perturbed Conformational Dynamics in Markov State Models. *J. Chem. Theory Comput.* **2014**, *10* (12), 5716–5728.

(27) Vitalini, F.; Noé, F.; Keller, B. G. Molecular Dynamics Simulations Data of the Twenty Encoded Amino Acids in Different Force Fields. *Data in Brief*. June 2016, pp 582–590.

(28) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N · log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089.

(29) Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L.-P.; Shukla, D.; Tye, T.; Houston, M.; Stich, T.; Klein, C.; Shirts, M. R.; Pande, V. S. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9* (1), 461–469.

(30) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129* (12), 124105.

(31) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J. H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11* (11), 5525–5542.

bioRxiv preprint doi: https://doi.org/10.1101/158592; this version posted July 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(32) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562.

(33) Okur, A.; Roe, D. R.; Cui, G.; Hornak, V.; Simmerling, C. Improving Convergence of Replica-Exchange Simulations through Coupling to a High-Temperature Structure Reservoir.

(34) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38, 27–28.

(35) Pérez, F.; Granger, B. E. IPython: A System for Interactive Scientific Computing. *Comput. Sci. Eng.* **2007**, *9* (3), 21–29.

(36) Hernández, C. X.; Harrigan, M. P.; Sultan, M. M.; Pande, V. S. MSMExplorer: Data Visualizations for Biomolecular Dynamics. *J. Open Source Softw.* **2017**, *2* (12).

(37) Piana, S.; Sarkar, K.; Lindorff-Larsen, K.; Guo, M.; Gruebele, M.; Shaw, D. E. Computational Design and Experimental Testing of the Fastest-Folding β-Sheet Protein. *J. Mol. Biol.* **2011**, *405* (1), 43–48.

(38) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science (80-.).* **2011**, *334* (6055).

(39) Krivov, S. V. The Free Energy Landscape Analysis of Protein (FIP35) Folding Dynamics. *J. Phys. Chem. B* **2011**, *115*, 12315–12324.

(40) Dama, J. F.; Rotskoff, G.; Parrinello, M.; Voth, G. A. Transition-Tempered Metadynamics: Robust, Convergent Metadynamics via on-the-Fly Transition Barrier Estimation. *J. Chem. Theory Comput.* **2014**, *10* (9), 3626–3633.

(41) Hines, K. E. A Primer on Bayesian Inference for Biophysical Systems. *Biophys. J.* **2015**, *108* (9), 2103–2113.

(42) Wan, H.; Zhou, G.; Voelz, V. A. A Maximum-Caliber Approach to Predicting Perturbed Folding Kinetics Due to Mutations. *J. Chem. Theory Comput.* **2016**, *12* (12), 5768–5776.

(43) Wu, H.; Mey, A. S. J. S.; Rosta, E.; Noé, F. Statistically Optimal Analysis of State-Discretized Trajectory Data from Multiple Thermodynamic States. **2014**.

(44) Harrigan, M. P.; Sultan, M. M.; Hernández, C. X.; Husic, B. E.; Eastman, P.; Schwantes, C. R.; Beauchamp, K. A.; Mcgibbon, R. T.; Pande, V. S. MSMBuilder: Statistical Models for Biomolecular Dynamics. *Biophys. J.* **2016**, *112* (1), 10–15.

(45) Mcgibbon, R. T.; Beauchamp, K. A.; Schwantes, C. R.; Wang, L.-P.; Hernández, C. X.; Harrigan, M. P.; Lane, T. J.; Swails, J. M.; Pande, V. S.; Hern, C. X.; Herrigan, M. P.; Lane, T. J.; Swails, J. M.; Pande, V. S. MDTraj: A Modern, Open Library for the Analysis of Molecular Dynamics Trajectories. *bioRxiv* **2014**, 9–10.