Focused conformational sampling in proteins

Marco Bacci, Cassiano Langini, Jiří Vymětal, Amedeo Caflisch, and Andreas Vitalis

Citation: The Journal of Chemical Physics **147**, 195102 (2017);
View online: https://doi.org/10.1063/1.4996879
View Table of Contents: http://aip.scitation.org/toc/jcp/147/19
Published by the American Institute of Physics

Focused conformational sampling in proteins

Marco Bacci,[a] Cassiano Langini, Jiří Vymětal,[b] Amedeo Caflisch,[a] and Andreas Vitalis[a]
Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

(Received 19 July 2017; accepted 25 October 2017; published online 20 November 2017)

A detailed understanding of the conformational dynamics of biological molecules is difficult to obtain by experimental techniques due to resolution limitations in both time and space. Computer simulations avoid these in theory but are often too short to sample rare events reliably. Here we show that the *progress index-guided sampling* (PIGS) protocol can be used to enhance the sampling of rare events in selected parts of biomolecules without perturbing the remainder of the system. The method is very easy to use as it only requires as essential input a set of several features representing the parts of interest sufficiently. In this feature space, new states are discovered by spontaneous fluctuations alone and in unsupervised fashion. Because there are no energetic biases acting on phase space variables or projections thereof, the trajectories PIGS generates can be analyzed directly in the framework of transition networks. We demonstrate the possibility and usefulness of such focused explorations of biomolecules with two loops that are part of the binding sites of bromodomains, a family of epigenetic "reader" modules. This real-life application uncovers states that are structurally and kinetically far away from the initial crystallographic structures and are also metastable. Representative conformations are intended to be used in future high-throughput virtual screening campaigns. *Published by AIP Publishing.* https://doi.org/10.1063/1.4996879

I. INTRODUCTION

Wet-lab experiments are the primary vehicle of discovery in the life sciences. Yet, at the molecular level, limitations to spatial and temporal resolution persist, and computer simulations are now employed routinely to complement insights from these experiments.[1–3] However, simulations of biomolecules, such as the ones carried out by integrating Newton's equations of motion,[4] are impaired by low scalability on general purpose hardware and by the rugged free energy landscape of atomistic models.[3,5] Interconversion rates between metastable states can be prohibitively low in conventional sampling (CS). Unfortunately, it is precisely these slow structural transitions involving many correlated degrees of freedom that are usually of interest.[6] The inability to sample the underlying free energy landscape exhaustively limits the power of brute-force molecular dynamics (MD) approaches as vast areas of phase space will remain undiscovered.[7,8] Techniques capable of enhancing interesting conformational transitions of complex systems are therefore desirable. The recently developed progress index-guided sampling (PIGS) method, briefly outlined in this section and explained in detail in Sec. II A and in the reference publication, is one of those.[9]

A comprehensive review of enhanced sampling methods is beyond the scope of this introduction and may be difficult in general.[10,11] We thus mention only those methods that, to the best of our knowledge, are appropriate for focusing explorations on specific degrees of freedom, which is what we achieve with PIGS. Force-[12–16] and energy-biased methods such as umbrella sampling and metadynamics[17,18] are well-established tools designed to enhance the sampling along (few) collective variables. Extensions to more dimensions are not trivial, however, especially from a practical perspective, even though improvements in this direction have been made.[19–22] Accelerated molecular dynamics is another popular approach used to boost the sampling of specific degrees of freedom.[23] It works by raising the minima of specific contributions to the potential energy in a threshold-dependent manner. As a result, it is difficult to enhance sampling in arbitrary collective variables or to recover the correct thermodynamics. The latter problem is due to the wide underlying energy spectra and the heterogeneous nature of the barriers.[24] For all methods utilizing an altered potential energy surface, it is not an easy task to retrieve correct kinetics and transition paths from simulations. This is because the connectivities between states and the microscopic rates are themselves biased. To make inferences in this regard, it may be necessary to impose plausible transition rates based on equilibrium distributions, geometrical proximity, and diffusivity.[25]

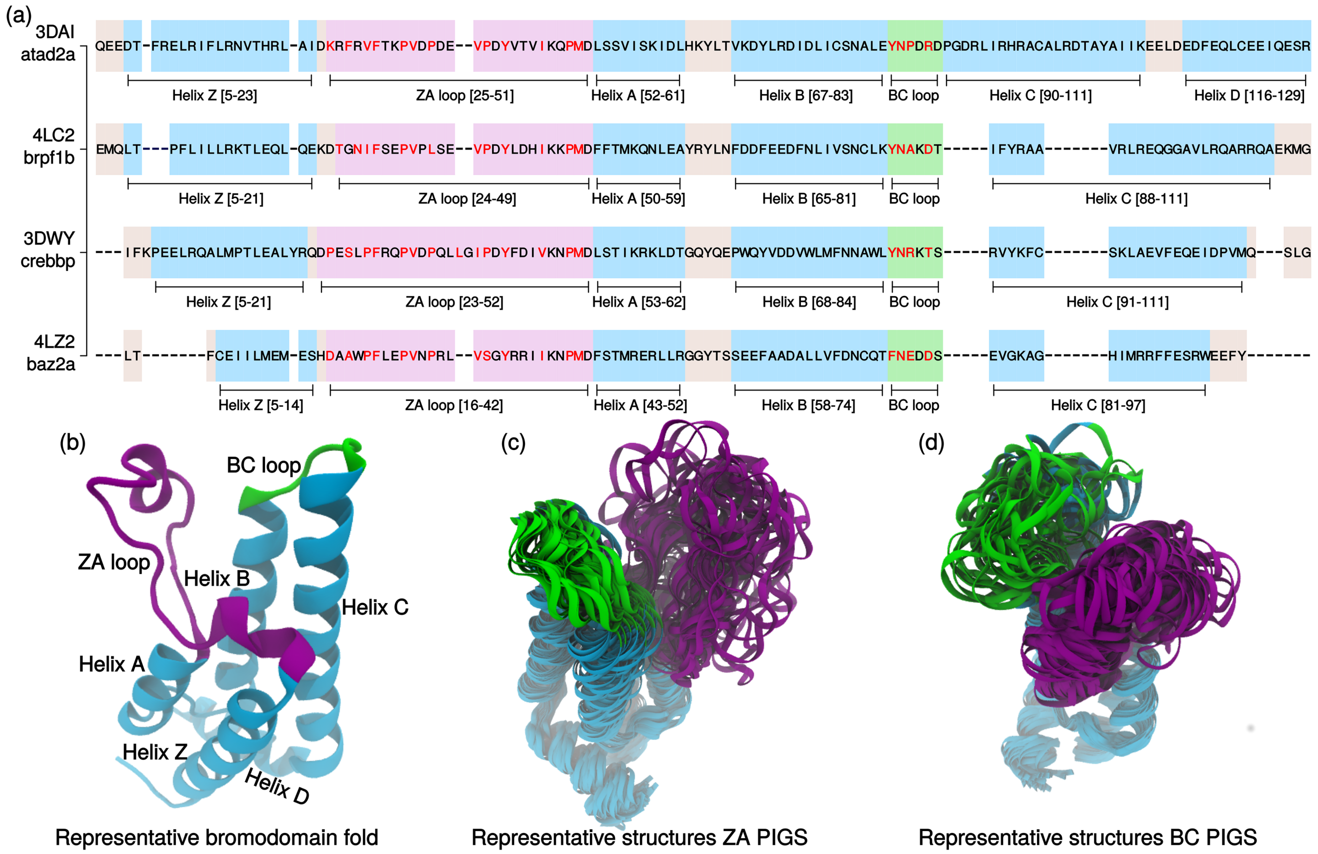
Adaptive sampling schemes are methods of a different type. We do not discuss further those approaches aiming to sample the pathways between a few main basins and/or along progress variables.[26–34] In general, adaptive schemes speed up the sampling by guiding the dynamics of the system according to the information on its evolution collected on-the-fly. A seminal and well-known example in protein folding is found in the work of Pande *et al.*[35] where energy variance is used as an indicator of large-scale transitions. The fundamental logic of adaptive sampling approaches can be described as follows: (1) simulations run, preferably in parallel, with an identical

[a] Authors to whom correspondence should be addressed: marco.bacci@uzh.ch; caflisch@bioc.uzh.ch; and a.vitalis@bioc.uzh.ch. Tel./Fax: +41 44 635 5568/+41 44 635 6862.
[b] Present address: Institute of Organic Chemistry and Biochemistry AS CR, v.v.i., Flemingovo náměstí 542/2, CZ-166 10 Prague 6, Czech Republic.

propagator, which is usually conventional MD; (2) information from the trajectories is collected and analyzed to determine which instantaneous conformations are most promising; (3) further generations of simulations take as starting points the most promising candidates from prior generations. The strength of these methods, aside from parallelizability, is precisely that the only bias they introduce to increase sampling is from a judicious and non-Boltzmann choice of starting conditions for each generation. In many cases, the notion of "promising" can be tuned easily toward a specific goal.

The adaptive sampling of Markov state models,[36] free energy-guided sampling,[37] diffusion map-directed MD,[38] WExplore,[39] and the recently developed fluctuation amplification of specific traits (FAST)[40] can be gathered along with PIGS[9] in this class. WExplore is an elegant method that builds upon the weighted-ensemble framework by dividing the conformational space of a system into Voronoi polyhedra.[31,32,39] A (possibly increasing) number of copies of the system are evolved in parallel and a hierarchy is used to inform cloning and merging operations to keep the sampling as uniform across phase space as possible. The representation for the space discretization can in principle be chosen to exclusively enhance specific degrees of freedom. Similarly, FAST uses phase space discretization to construct Markov state models (MSMs) at regular intervals.[41,42] In addition, it rewards low sampling weights and changes in a selected geometric transform such as total energy or solvent accessible surface area.[40] Both the discretization and the reward function can thus be tuned to focus the sampling enhancement without having to bias the potential energy surface. Consequently, FAST, like the other methods in its class, allows the data to be analyzed as locally equilibrated trajectories.[43,44]

PIGS also returns a set of trajectories that can be analyzed naturally in the framework of MSMs. It is designed to fit modern HPC resources as it evolves a constant and possibly large number of replicas of a system in parallel and relies on state-of-the-art scalable analysis algorithms to enhance the sampling.[9,45,46] Unlike in FAST[40] or adaptive sampling,[36] this does not imply building MSMs on-the-fly. In short, PIGS works as follows. From a starting condition, a selected number of replicas are evolved in parallel. Features are extracted from them at regular intervals. PIGS mandates the selection of features as a set of degrees of freedom, e.g., specific dihedral angles, to represent the system and compute distances between snapshots. It is these selections that enable PIGS to focus the sampling enhancements in response to specific questions. To accomplish this, all collected snapshots from all replicas are arranged jointly in the so-called progress index (PI),[46] which is analyzed to derive a ranking of the current end points of the simulation stretches in a way that rewards sampling


(a)
3DAI
atad2a
QEEDT-FRELRIFLRNVTHRL-AIDKRFRVFTKPVDPDE--VPDYVTVIKQPMDLSSVISKIDLHKYLTVKDYLRDIDLICSNALEYNPDRDPGDRLIRHRACALRDTAYAIIKEELDEDFEQLCEEIQESR
Helix Z [5-23]
ZA loop [25-51]
Helix A [52-61]
Helix B [67-83]
BC loop
Helix C [90-111]
Helix D [116-129]
4LC2
brpf1b
EMQLT---PFLILLRKTLEQL-QEKDTGNIFSEPVPLSE--VPDYLDHIKKPMDFFTMKQNLEAYRYLNFDDFEEDFNLIVSNCLKYNAKDT-----IFYRAA-------VRLREQGGAVLRQARRQAEKMG
Helix Z [5-21]
ZA loop [24-49]
Helix A [50-59]
Helix B [65-81]
BC loop
Helix C [88-111]
3DWY
crebbp
---IFKPEELRQALMPTLEALYRQDPESLPFRQPVDPQLLGIPDYFDIVKNPMDLSTIKRKLDTGQYQEPWQYVDDVWLMFNNAWLYNRKTS-----RVYKFC-------SKLAEVFEQEIDPVMQ---SLG
Helix Z [5-21]
ZA loop [23-52]
Helix A [53-62]
Helix B [68-84]
BC loop
Helix C [91-111]
4LZ2
baz2a
---LT-------FCEIILMEM-ESHDAAWPFLEPVNPRL--VSGYRRIIKNPMDFSTMRERLLRGGYTSSEEFAADALLVFDNCQTFNEDDS-----EVGKAG-------HIMRRFFESRWEEFY-------
Helix Z [5-14]
ZA loop [16-42]
Helix A [43-52]
Helix B [58-74]
BC loop
Helix C [81-97]
(b)
BC loop
ZA loop
Helix B
Helix C
Helix A
Helix Z
Helix D
(c)
(d)

Representative bromodomain fold

Representative structures ZA PIGS

Representative structures BC PIGS

FIG. 1. Sequences and segments of the bromodomains used in this study and cartoons of the atad2a domain. Cartoons are rendered with VMD[71] and Tachyon (http://jedi.ks.uiuc.edu/~johns/raytracer). (a) Amino acid sequence alignment[72] of the bromodomains. Background colors and text annotations distinguish the different segments (helices in blue, ZA loop in purple, and BC loop in green). The residues that are part of the PIGS representations (Table I) are highlighted in red. (b) The bromodomain of atad2a in cartoon representation exemplifies the common fold of bromodomains. Helix D is present only for atad2a. (c) 50 snapshots representative of different basins characteristic of ZA PIGS simulations of the atad2a domain are aligned and displayed as ribbons in the same color code as (b). (d) Same as (c) for BC PIGS.

uniqueness. The ranking is used to reseed (stochastically) those copies sampling overlapping regions of phase space with more interesting ones, which have arisen by spontaneous fluctuations alone. After reseeding, the cycle starts anew. Notably, PIGS can deal with a broad set of features, making it a flexible tool suited to tackle different problems. It is also scalable, unsupervised, and synergistic, viz., sampling benefits scale with the number of copies used in parallel.

The ability to focus sampling enhancements with high precision is of importance because many biomolecules, in particular proteins, have functionally distinct regions such as loops, surface patches, catalytic sites, allosteric binding sites, structured cores, or disordered linkers. Epigenetic regulators are members of different protein families that participate in modulating DNA accessibility through covalent modifications of chromatin. These include small "reader" modules called bromodomains.[47,48] Their conserved fold consists of four α-helices (termed αZ, αA, αB, and αC), which are connected by loops of different lengths, see Figs. 1(a) and 1(b).[49] Bromodomains bind acetylated lysine side chains on histone tails with well-defined hydrophobic pockets that are framed by two nonadjacent loops, viz., the ZA and BC loops.[50] Numerous mutations in epigenetic regulators, including bromodomains, have been identified in several types of cancers.[51–55] In general, epigenetic regulation is under dynamic control and appears to be a feasible target for cancer therapies.[56–60] However, in finding small molecule effectors, bromodomains challenge standard *in silico* docking protocols in that the binding site shows considerable plasticity.[61] It may thus be useful to virtual screening approaches to have access to additional conformations of the protein with differences first and foremost in the binding site.[62,63] In particular, a metastable state unique to a given bromodomain could enable the identification of selective ligands.

Here, we use the ZA and BC loops of four different bromodomains as prototypical examples to demonstrate that PIGS is not only able to reach time scales and conformations that are difficult to access with CS but that its design allows enhancing the rates of phase space exploration for specific parts of complex molecules without perturbing the remaining parts directly. In the remainder of the text, we first review the PIGS protocol followed by a sufficient description of the simulation settings and analysis methods. We then summarize the results, which show that it is a straightforward task to obtain focused enhanced sampling with PIGS by selecting the appropriate degrees of freedom. Importantly, the newly discovered areas of phase space are meaningful in both a thermodynamic and a kinetic sense. We conclude by discussing the gist of our results in the context of the general applicability of PIGS.

II. METHODS

A. PIGS protocol

In addition to the description below, all technical details can be found in the original paper.[9] In PIGS, a set of N_r replicas of a system are propagated in parallel under the same conditions by a stochastic sampler, e.g., Langevin dynamics, Monte Carlo, or MD with a stochastic thermostat. Here, we used an independent MD engine (GROMACS) for system propagation (see Sec. II B for settings). All copies were run for a stretch of a given and fixed length (here, 100 ps), and snapshots sufficient to allow the extraction of the required features (Table I) were saved at constant frequency. At the end of each

TABLE I. Dihedral angles used to represent the four bromodomains in the PIGS simulations, sampling times (per copy and cumulative), and saving frequency of all runs. The residue numbering is congruent with the sequences in Fig. 1.

	atad2a (3DAI)	baz2a (4LZ2)	brpf1b (4LC2)	crebbp (3DWY)
ZA PIGS torsional angles	Ψ: L25, F27, V29, F30, P33, V34, P36, V39, P40, Y42, I46, P49, M50 χ_1: Y42	Ψ: D16, A18, P20, F21, P24, V25, P27, V30, S31, Y33, I37, P40, M41 χ_1: Y33	Ψ: T24, N26, I27, F28, P31, V32, L34, V37, P38, Y40, I44, P47, M48 χ_1: Y40	Ψ: P24, S26, P28, F29, P32, V33, P35, L38, I40, P41, Y43, V47, P50, M51 χ_1: Y43
ZA PIGS sampling	90 ns per copy; 5.76 μs cumulative	90 ns per copy; 5.76 μs cumulative	90 ns per copy; 5.76 μs cumulative	88.8 ns per copy; 5.68 μs cumulative
BC PIGS torsional angles	Ψ: N85, Y86, R88 χ_1: Y84, N85 χ_2: N85	Ψ: N76, E77, D79 χ_1: F75, N76 χ_2: N76	Ψ: N83, A84, D86 χ_1: Y82, N83 χ_2: N83	Ψ: N86, R87, T89 χ_1: Y85, N86 χ_2: N86
BC PIGS sampling	~97 ns per copy; 6.22 μs cumulative	~90 ns per copy; 5.76 μs cumulative	~90 ns per copy; 5.76 μs cumulative	~75 ns per copy; 4.84 μs cumulative
CS sampling	127.5 ns per copy per bromodomain; 8.16 μs cumulative per bromodomain			
No. of atoms	60 779	60 703	60 736	60 725
Saving frequency; no. of copies	4 ps (all cases); 64 copies (all cases)			

stretch, all saved snapshots from all replicas were analyzed concurrently by state-of-the-art analysis algorithms[45,46] implemented in CAMPARI (http://campari.sourceforge.net), which are central to the PIGS protocol. Then, new stretches were restarted according to the reseeding decisions made by PIGS. Here, an entire PIGS simulation consisted of nearly 1000 such cycles.

PIGS aims to reseed copies that are sampling overlapping regions of phase space with putatively more interesting ones. The progress index (PI) is at the core of PIGS. It is an unsupervised analysis method able to reveal different free energy basins and barrier regions explored by a stochastic dynamical system such as a biomolecule in solution.[46] PIGS exploits this information to make reseeding decisions as detailed below. The PI has as its only essential input parameter the choice of features and metric function defining the conformational (geometric) distance between two snapshots, and the choice of features is what we vary in this contribution to achieve a focused exploration.

In an approximate but scalable implementation, the reseeding process happens in five main steps. **First**, the data are preorganized into a multi-resolution clustering tree,[45] which is a data structure created by clustering all the snapshots from all replicas into mutually similar groups at a number of resolutions in a numerically efficient manner. Because the construction of the PI relies on short distances only, this information is sufficient to produce a very good approximation of the exact PI, which implies that the preliminary clustering has a marginal effect on the final outcome of the protocol. **Second**, we construct the PI. Starting from the snapshot that is the centroid representative of the largest cluster at the finest resolution, all snapshots are arranged such that the next one ideally is the snapshot closest to *any of the ones* already accounted for. The approximation we use is to exploit the preorganization of the data in the multi-resolution clustering tree to find closest neighbors heuristically (rather than exhaustively) while maintaining the scalability of the algorithm. This gives rise to the approximate PI.[46] **Third**, the N_r final frames of all copies, which are the only states we consider for reseeding, are ranked according to a consensus ranking from the following three criteria defining sampling uniqueness and interestingness: (i) their position along the PI (right is better); (ii) the distance to the snapshot on the left by which they were added to the PI (larger is better); (iii) the smallest distance, as relative position along the PI, to any other final conformation (larger is better). Criteria (i) and (ii) evaluate favorably if a final frame resides in an area of low sampling density whereas criterion (iii) indicates how likely a final frame is to be dissimilar from the other final frames. The consensus ranking of a replica $\xi(R)$ is simply obtained as the sum of the three individual components (smaller sum is better). **Fourth**, the algorithm attempts to reseed the N_r-N_t lower ranked conformations with the top N_t ones. Specifically, for every low-ranked copy, $\mathrm{R_L}$, a high-ranked one, $\mathrm{R_T}$, is drawn uniformly, and the reseeding probability is computed as

$$p\left(R_L \rightarrow R_T\right) = \left[\xi\left(R_L\right) - \xi\left(R_T\right)\right] / \left[\xi\left(R_{Worst}\right) - \xi\left(R_{Best}\right)\right] . \tag{1}$$

This probability is compared with a random number drawn from the [0:1] interval. If the probability in Eq. (1) is larger than this number, the reseeding is putatively accepted. **Fifth**, a heuristic is used to cancel reseeding events for those replicas that are deemed to have been sampling a relatively unique region of phase space during the last stretch. This check is required because the final conformation of any replica being reseeded is lost irrevocably for trajectory continuation. This is the only heuristic to explicitly consider the position of all snapshots in the PI, and it basically corresponds to a locality criterion per replica. More precisely, if the difference between the 3rd and 1st quartiles of the snapshots of R_L within the PI is less than the number of analyzed snapshots per replica, then the reseeding is cancelled and the low-ranked copy continues to be propagated in the next stretch. Accepted reseedings involve the replacement of all phase space variables (here, positions and velocities) with those from another copy. Trajectory divergence is achieved by the stochastic component of the MD engine. After each reseeding cycle, the sampling history is forgotten completely, and this memorylessness ensures the scalability of the algorithm. As described above, the construction of the PI mandates as input a choice of how to define conformational distances between snapshots, and this involves selecting a set of geometric coordinates (e.g., a set of dihedral angles or interatomic distances) as features. This selection of specific coordinates, which the metric (here, Euclidean) is based on, allows directing the protocol to focus the enhancement of phase space exploration toward parts and questions of interest, which is the key contribution in this manuscript. The only sampling bias incurred by PIGS is due to the killing and restarting of simulations in an unsupervised but non-Boltzmann way. Initial condition bias of this type is inherent to both CS and other methods using MSMs to guide the sampling, and the MSMs themselves are commonly used to remove this bias.[40,43,44]

B. PIGS and CS simulations

To enhance the sampling of the ZA and BC loops with PIGS, we used two independent sets of segment-specific coordinates, viz., the dihedral angles listed in Table I. PIGS simulations with these two sets gave rise to the **ZA PIGS** and **BC PIGS** data sets for a given bromodomain. Table I provides further information about the simulations. Other possible representations and applications of PIGS are discussed in Sec. IV D.

PIGS simulations were run using GROMACS[64] coupled to a custom Python script to perform the PIGS reseeding process with CAMPARI v3b every 100 ps. Each of the 64 replicas provided the features in Table I every 200 fs and the PI was constructed from the combined data. Scalable calculations of the PI require an approximation to the exact solution, which relies on data preorganization by scalable clustering.[45] Here, it was feasible to employ an OpenMP parallelization of this step on a single node (relative time cost below 10%). As parameters, this required a number of guesses (500) and clustering settings (tree height of 12, fine and coarse thresholds of 10° and 55°). The number of top-ranked replicas (N_t), which are protected from being reseeded at the end of a given cycle, was 32. Overall, these settings resulted in an average trajectory length between successful reseedings of 1-3 ns for the different cases.

TABLE II. Sets of atoms selected for computing RMSDs based on Cartesian coordinates. "N" and "O" refer to backbone atoms. Side chain atoms were added only for the two loops. By visual inspection of crystal structures, at most one atom per side chain, which appeared both informative and structurally constrained by the remainder of the molecule, was picked.

	atad2a (3DAI)	baz2a (4LZ2)	brpf1b (4LC2)	crebbp (3DWY)
ZA loop rep.	**N, O**: I23-S53; **Cβ**: V29, T31, V34, V39, V43, T44, V45; **Cδ_1**: I46; **Cζ**: F27, F30; **OH**: Y42	**N, O**: S14-S44; **Cβ**: A17, A18, V25, V30; **Cδ_1**: I36, I37; **Cγ**: L22, N26, L29; **Cζ**: F21; **Nε_1**: W19; **OH**: Y33; **Oγ**: S31	**N, O**: L22-F51; **Cβ**: T24, V32, V37; **Cδ_1**: I27, I44; **Cγ**: N26, L34, L41; **Cζ**: F28	**N, O**: R21-S54; **Cβ**: V33, V47; **Cδ_1**: Q31, I40, I46; **Cγ**: L27, L37, L38; **Cζ**: F29, F44; **OH**: Y43; **Oγ**: S26
BC loop rep.	**N, O**: L82–G91; **Cγ**: N85; **OH**: Y84	**N, O**: Q73-V82; **Cζ**: N76; **Oγ**: S80	**N, O**: L80-F89; **Cβ**: A84, T87; **Cγ**: N83; **OH**: Y83	**N, O**: W83-V92; **Cβ**: T89; **Cγ**: N86; **Oγ**: S90
Helix bundle	**N, O**: T6-L21, S54-I59, D69-A81, D92-I110	**N, O**: E6-M12, T45-L50, E60-C72, G83-R96	**N, O**: T6-L19, T52-L57, D67-C79, Y90-Q110	**N, O**: E6-L19, T55-L60, Q70-A92, T93-V110

CS simulations simply ran GROMACS continuously with identical numbers of replicas (these are referred to simply as **CS**). We simulated the different bromodomains [Fig. 1(a)] as described by the CHARMM36[65] force field with modified TIP3P water and an ionic background of ~150 mM KCl in the NVT ensemble at 310 K. All simulations were run on the GPU nodes of the supercomputer Piz Daint. Temperature was maintained by the velocity rescaling thermostat.[66] All non-bonded interactions employed a cutoff of 1.2 nm with the help of Verlet neighbor lists, and we treated electrostatic interactions with the generalized reaction field approach.[67] Aside from water molecules, which were held rigid with the SETTLE algorithm,[68] constraints were applied to all covalent bonds and enforced by LINCS[69] with default settings.

C. Root mean square deviation (RMSD)

As an independent measure of conformational distance from reference states (see Sec. III B), we defined the sets of degrees of freedom used in RMSDs with prior alignment reported in Table II.

D. Principal component analysis (PCA)

PCA is a dimensionality reduction technique relying on variance. Here, we computed the PC transformation and projected the raw data onto the first two principal components, which are orthogonal linear combinations of input features capturing the largest variance in the data set (see Sec. III C). As input features, we picked the sine and cosine values of the dihedral angles listed in Table III. These representations combine both loops per domain and differ slightly from the ones used in PIGS. This is intended as they are meant to also be able to report on changes in the directly adjacent residues. Note that all the available snapshots for a bromodomain were analyzed jointly (ZA PIGS, BC PIGS, and CS).

E. Mean first passage times (MFPTs)

To understand the kinetic distances of states discovered by PIGS from the crystallographic reference states, we first constructed mesostate networks by grouping conformations into clusters with a tree-based clustering algorithm that is well-known to perform well in preserving kinetic information.[45] Subsequently, we derived transition networks to infer MFPTs, for all clusters, to the cluster containing the structure closest to the reference PDB (identified by the RMSD across all Cα atoms). Here, unlike for PCA, we employed two distinct representations (Table IV) and treated the individual simulation groups (ZA PIGS, BC PIGS, and CS) separately (see Sec. III E).

This means that 6 clusterings and derived transition networks were obtained for each bromodomain (3 simulation groups times 2 representations, see Table V). Detailed balance of mesostate transitions was imposed by symmetrization

TABLE III. Dihedral angles used to compute principal components. For the PC analysis, all angles were included in a joint representation, regardless the loop (ZA or BC) they were part of.

	atad2a (3DAI)	baz2a (4LZ2)	brpf1b (4LC2)	crebbp (3DWY)
PCA torsional angles	Ψ: I23-S53, L82-P90 χ_1: Y42, N85	Ψ: S16-S31, Y33-S44, Q73-V82 χ_1: Y33, N76	Ψ: L22-T24, N26-F51, L80-F89 χ_1: Y40, N83	Ψ: R21-L38, I40-S54, W83-V92 χ_1: Y43, N86

TABLE IV. Dihedral angles used to group snapshots to derive transition networks. This is the same selection as that in Table III only split into two sets corresponding to the ZA and BC loops.

	atad2a (3DAI)	baz2a (4LZ2)	brpf1b (4LC2)	crebbp (3DWY)
ZA loop torsional angles	Ψ: I23-S53 χ_1: Y42	Ψ: S16-S31, Y33-S44 χ_1: Y33	Ψ: L22-T24, N26-F51 χ_1: Y40	Ψ: R21-L38, I40-S54 χ_1: Y43
BC loop torsional angles	Ψ: L82 - P90 χ_1: N85	Ψ: Q73 - V82 χ_1: N76	Ψ: L80 - F89 χ_1: N83	Ψ: W83 - V92 χ_1: N86

of the count matrix to the maximum count per pair of states at a lag time of 1.0 ns and a clustering resolution of 17°. Transitions between clusters were accumulated with a sliding window approach. These settings were the same as those for the global MSMs, which include all data (BC PIGS, ZA PIGS, and CS) and are described in Sec. II F.

F. Ensemble reweighting based on Markov state models (MSMs)

Transition networks can not only be used to infer MFPTs (see Sec. II E) but also to calculate the probability distribution at equilibrium across clusters. This assumes that the trajectory at the chosen lag time is a Markov process. If the MSM is constructed from a set of short trajectories, the equilibrium distribution will generally differ from the raw, count-based sampling weights. We thus used MSMs to infer the equilibrium (steady state) weights of the clusters in networks which, for a given bromodomain, included all snapshots from ZA PIGS, BC PIGS, and CS. The MSM steady state probabilities give rise to snapshot-based weights for subsequent analyses calculated as $w_i = p_c^{ss}/p_c^{raw}$ where c denotes the cluster that snapshot i is part of, and p^{SS} and p^{raw} are the steady-state and raw sampling weights of clusters, respectively. We constructed a set of MSMs at different lag times and resolutions with the representations in Table III. As in Sec. II E, we imposed detailed balance and used the sliding window approach. We wanted a consistent choice across bromodomains, and visual inspection of the implied time scales lets us choose a lag time of 1.0 ns at a clustering resolution of 17° (see Table VI and Figs. S7–S10 of the supplementary material). MSMs-derived weights were used to compute reweighted histograms and averages wherever noted.

G. Metastability analysis

Identical to the CS simulations described in Sec. II B, we ran 64 copies of simulations from 2 additional starting states per bromodomain: a representative of a newly discovered ZA and BC loop conformation each (see Secs. III E and III F). The positive controls are in fact the CS simulations which were run from the crystal structures. The other simulations were run for 85 ns (instead of 127.5 ns) per replica, and, for the metastability analysis alone (Sec. III F), we truncated the CS simulations at 85 ns to achieve exact comparability.

To estimate metastability, we relied on RMSDs as a function of time. RMSDs were calculated based on the union of sets of atoms listed in Table II, which means that these RMSDs are sensitive to changes in any of the parts of the bromodomain. Thus, the approach is prone to underestimate local metastability due to different processes contributing to conformational drift. This is a particular issue for the BC loop due to the size and disorder of the ZA loop. In general, however, we deemed the approach acceptable since we are primarily looking to derive a lower bound. Metastability was described by a summary statistic, which is the characteristic (life) time of an exponential fit, τ, calculated as $\langle t_i \rangle$ where the time-dependent probability for being in the initial state is $p(t) = \exp(-t/\tau)$. The individual t_i escape times were counted with the help of two RMSD thresholds spaced 1 Å apart. Initially, the RMSD values are small and the system is in its starting state. A leaving event and associated escape time were registered when the RMSD exceeded the larger threshold. Reentries were allowed and counted whenever the RMSD fell below the lower threshold after a prior escape. Using this strategy, most trajectories gave rise to only 0 or 1 leaving events for a range of thresholds. We note that the estimate of τ as $\langle t_i \rangle$ is the maximum

TABLE V. Numbers of clusters in the transition networks used for MFPT analysis. Because of the imposition of detailed balance, the net statistical weight of the largest reversibly connected subset of clusters (the largest strongly connected component) was always >99%. The parameters were the same throughout and inherited from the global MSMs (Sec. II F): a resolution of 17° and a lag time of 1.0 ns. The tree-based clustering always utilized a tree height of 16.

	atad2a (3DAI)		baz2a (4LZ2)		brpf1b (4LC2)		crebbp (3DWY)	
	Representation		Representation		Representation		Representation	
	ZA	BC	ZA	BC	ZA	BC	ZA	BC
BC PIGS	9 767	28 632	11 148	17 777	11 627	8839	9 412	26 332
ZA PIGS	53 683	985	58 688	689	69 636	1002	46 596	425
CS	12 649	875	14 336	813	17 532	853	20 114	631

TABLE VI. Characteristic quantities of the MSMs used for deriving snapshot weights. Because of the imposition of detailed balance, the net statistical weight of the largest strongly connected component was always >99%.

	atad2a (3DAI)	baz2a (4LZ2)	brpf1b (4LC2)	crebbp (3DWY)
No. of clusters in network	65 238	70 769	73 547	60 855
No. of edges in network	1 280 781	1 910 162	2 263 452	1 506 314
Network resolution	17°			
Lag time	1.0 ns			

likelihood estimate for this parameter, which is biased toward smaller values for small sample sizes. Again, this is acceptable if we are interested in a lower bound.

III. RESULTS

A. Common bromodomain fold and representative structures discovered by PIGS

As summarized in Fig. 1(a), we used PIGS to diversify the dihedral angles of the ZA loop of four bromodomains (ZA PIGS, Table I), the crystal structures of which have PDB codes 3DAI (the bromodomain of the protein atad2a), 3DWY (crebbp), 4LC2 (brpf1b), and 4LZ2 (baz2a).[48,70] ZA PIGS consists of 64 copies per domain, each spanning ∼90 ns. Analogously and independently, we ran 64 copies per domain focusing on the torsional angles of the BC loop (giving rise to the BC PIGS data sets). For comparison, we also obtained data from CS simulations starting from the same initial structures for each bromodomain. The CS simulations consist of 64 replicas per bromodomain, each covering ∼127 ns (Table I). All simulations were, for the propagation stretches, standard MD calculations in an explicit solvent (water and 150 mM of monovalent salt) and periodic boundary conditions run with GROMACS 5 (see Sec. II B for details).[64]

Figures 1(c) and 1(d) provide a graphical manifestation of the efficacy of the PIGS simulations. Despite the qualitative and exemplary character of this analysis, it is readily appreciated that the simulations succeeded in focusing the explorations on the ZA and BC loops: the ZA loop of atad2a clearly adopts more diverse conformations in Fig. 1(c) than in Fig. 1(d), and the opposite is true for the BC loop.

B. Distance from reference PDB

To quantify the visual hints provided by Figs. 1(c) and 1(d), we first calculated the distances of the trajectory snapshots from the respective crystal structures (3DAI for atad2a, 4LZ2 for baz2a, 4LC2 for brpf1b, and 3DWY for crebbp). A coarse and simple way to do this is given by the root mean square deviations with alignment (RMSD) of apposite sets of atoms (listed in Sec. II C).

In detail, this means that an RMSD value for a specific segment in Fig. 2 is obtained by aligning a given snapshot to the reference PDB according solely to the coordinates of N and O backbone atoms and a few side chain atoms (Table II) of the segment in question. The actual RMSD is calculated across this restricted set as well. Aside from ZA and BC loops, helices αA, αZ, αB, and αC define a joint set ("Helices rep." in Fig. 2). For a given bromodomain, all snapshots derive from one of the three simulation groups: ZA PIGS, BC PIGS, or CS. We thus report three results per segment, i.e., we can measure the RMSD of the ZA loop in ZA PIGS, BC PIGS, and CS, and we can do the same for the BC loop and the helix bundle. If focused

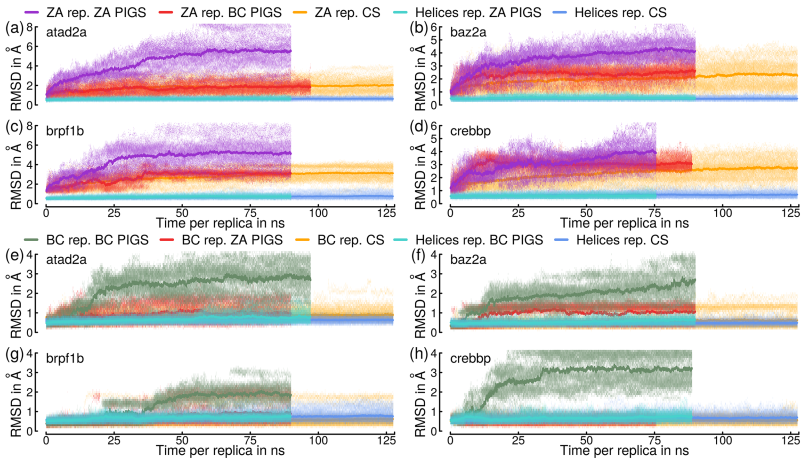


FIG. 2. RMSD time traces of various segments (ZA loop, BC loop, and helix bundle) comprising the typical fold of bromodomains. Each RMSD is computed from the corresponding segment of the reference crystal structure after alignment to the same segment. Each multi-replica run (ZA PIGS, BC PIGS, and CS) provides one RMSD trace per segment. In the top panels, (a)-(d), RMSD traces for the ZA loop and helix bundle are shown. In the bottom panels, (e)-(h), data for the BC loop and helix bundle are plotted. Lines mark the average RMSD across all copies of a multi-replica run at a given time. The 2D histograms in time and RMSD are in logarithmic scale, delineate the envelope across all replicas, and respect MSM steady state weights (see Sec. II F). (a) ZA loop and helix bundle RMSD traces for the atad2a domain (ref. PDB 3DAI). (b) Same as (a) for baz2a (ref. PDB 4LZ2). (c) Same as (a) for brpf1b (ref. PDB 4LC2). (d) Same as (a) for crebbp (ref. PDB 3DWY). (e) BC loop and helix bundle RMSDs for atad2a. (f) Same as (e) for baz2a. (g) Same as (e) for brpf1b. (h) Same as (e) for crebbp.

explorations are successful, we expect to detect greater values when the RMSD representation echoes the dihedral one used in PIGS (e.g., "ZA rep. ZA PIGS" in Fig. 2) and to find smaller values (comparable to CS) in the other cases (e.g., "ZA rep. BC PIGS"). We also expect that CS results are comparable with PIGS ones for all the segments that do not benefit from sampling enhancement, helices included. Figure 2 confirms these predictions: the ZA loop is not perturbed in BC PIGS beyond its intrinsic flexibility but reaches much larger distances in ZA PIGS runs [Figs. 2(a)–2(d)]. The analogous result holds for the BC loop as evidenced by Figs. 2(e)–2(h). The helix bundle in PIGS simulations never undergoes significant rearrangements indicated by the relevant RMSDs barely exceeding ∼1 Å irrespective of simulation group or bromodomain. Figure 2 shows histograms across replicas that account for MSM weights but the impact, relative to using raw sampling weights, is minor here (compare Fig. S1 of the supplementary material).

As a second result, Fig. 2 confirms the large flexibility of the ZA loop observed in prior studies.[61] The example of crebbp [Fig. 2(d)] appears to suggest that an enhancement of sampling is not needed in this case. However, some considerations are in order. First, the ZA RMSD in ZA PIGS does in fact reach larger values than in BC PIGS or CS. Second, the ZA PIGS group features a wider spectrum of RMSD values than the other two groups. This implies that ZA PIGS samples structures both further away and closer to the crystal than BC PIGS or CS. It is generally expected that PIGS runs contain a wider spectrum of basins also in terms of kinetic distance from the initial condition, and this spread is likely beneficial to achieve locally equilibrated trajectories. In Fig. 2(d), we also note a rapid increase in ZA loop RMSD in BC PIGS, which eventually converges onto the same plateau level as the CS data, albeit with a narrower distribution. It is important to understand that the PIGS protocol reseeds system snapshots globally. This means that reseeding decisions, which are more frequent at the beginning of a run, can have a stochastic impact on the distribution of degrees of freedom that are not part of the PIGS set. This is the likely reason for the result in Fig. 2(d) and for the generally narrower envelopes of ZA loop RMSDs in BC PIGS runs.

Figure 2 demonstrates quantitatively that unsupervised focused explorations with PIGS are possible and successful, and that structural changes do not automatically propagate spatially or along the sequence. This is likely the result of the choices made for the sets of degrees of freedom and, more importantly, of bromodomain architecture. Bromodomains are not expected to propagate signals allosterically as they are primarily competitive binders and help in the recruitment and assembly of complexes that regulate transcription and/or modify the histone code.[49,73–75] In general, however, spatial couplings can and will be exposed by PIGS between different parts of a system, e.g., two adjacent monomers in an aggregate.[76] This suggests to us that PIGS can be used precisely to discover the presence (or lack) of allosteric effects, which are known to be difficult to predict and/or simulate.[77,78]

C. Conformational envelopes

RMSDs become dramatically degenerate as the actual value increases. Therefore, as a complement, Fig. 3 presents a

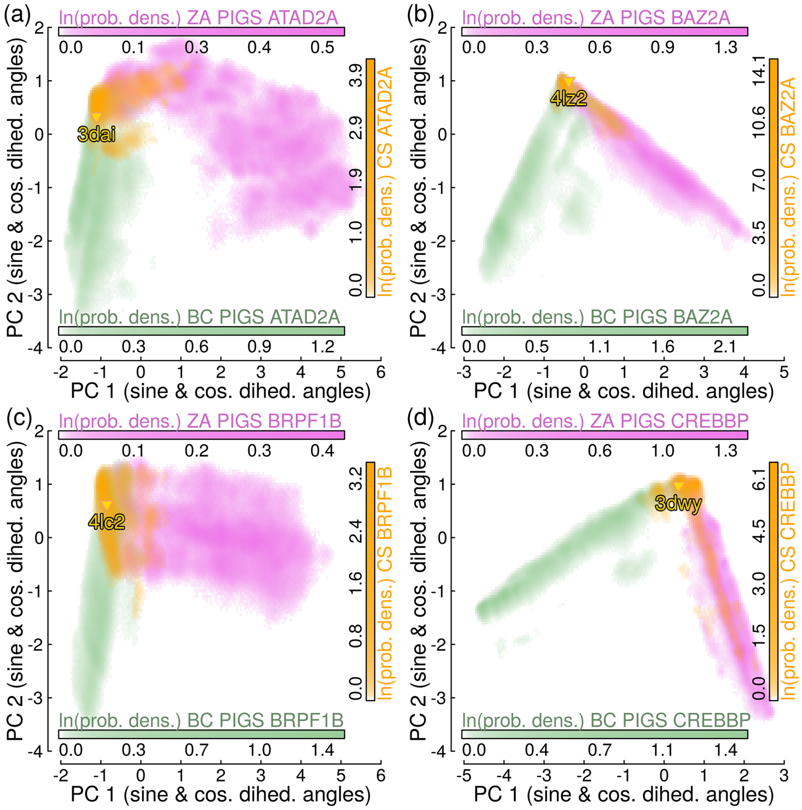


FIG. 3. PC projections colored by simulation set of origin (ZA PIGS, BC PIGS, or CS). We computed PC transformations based on sine and cosine values of the dihedral angles in both loops for the combined data sets for each bromodomain (Table III). The plotted probability densities (logarithmic scale) reflect the steady state of underlying MSMs (see Sec. II F), which here have only a small effect (compare Fig. S2 of the supplementary material). The positions of the corresponding reference PDB structures are highlighted in each panel by their closest representatives (distance was calculated as the RMSD across all Cα atoms). (a) PC projection for the atad2a bromodomain (ref. PDB 3DAI). (b) Same as (a) for baz2a (ref. PDB 4LZ2). (c) Same as (a) for brpf1b (ref. PDB 4LC2). (d) Same as (a) for crebbp (ref. PDB 3DWY).

two-dimensional projection onto principal components (PCs, see Sec. II D).[79] The components are derived by using the sine and cosine values of the dihedral angles of ZA and BC loop residues in a joint representation (Table III) for all the snapshots of a given bromodomain combined. In the histograms, we can then partition densities by their simulation condition of origin: ZA PIGS (magenta), BC PIGS (green), and CS (orange). Projections utilize the first two components, which capture between 30% and 40% of the total variance. As seen in Fig. 3, the low-dimensional PC projections highlight conformational envelopes and their overlap regions much more clearly than Fig. 2. In particular, Fig. 3 establishes unequivocally that the overlap between ZA and BC PIGS is restricted to the phase space area near the snapshot closest to the reference crystal structure. The lack of overlap elsewhere demonstrates that ZA and BC PIGS explore different areas of phase space with enhanced rates. Since PCs are based on variance, it is reasonable to expect that the envelopes are larger for ZA PIGS than for BC PIGS given that the ZA loop is both longer and more flexible.

Importantly, Fig. 3(d) addresses the concern regarding the ZA loop RMSD trace in BC PIGS discussed in the context of Fig. 2(d). Clearly, the observed rapid increase is not due to overlapping coverage of the ZA loop phase space by BC PIGS. The overlap between BC and ZA PIGS is indeed small also for crebbp and lower than the one between ZA PIGS and CS. In general, CS simulations tend to overlap more with ZA PIGS than BC PIGS runs. This is owed to the structural characteristics of the ZA loop and particularly evident in Fig. 3(d). The ZA loop is reasonably described as intrinsically disordered and conformations diversify spontaneously in CS, albeit at a much lower effective rate. Figure S3 of the supplementary material shows the same data as Fig. 3 in raw probability scale and confirms that the phase space discovered is not reachable by CS on the 100 ns time scale. The fact that BC PIGS runs have a lower coverage of the ZA loop phase space than CS was addressed above already; it is almost certainly related to the BC loop-based reseeding decisions. Because these decisions terminate replicas and duplicate others, there may be, relative to a CS simulation with the same number of replicas, a loss of information in degrees of freedom not covered by the PIGS representation.

D. Discovery of states

In Fig. 4, we consider an explicit measure of the number of states discovered by PIGS and CS runs along with the per-residue average α-helical content.[80]

The number of discovered states is a direct indicator of the exploration rate and of interest to any enhanced sampling method. We define a 3-dimensional conformational space based on the values of backbone dihedral angles of 3 consecutive residues. Each residue's instantaneous ϕ- and ψ-values are mapped to 1 of 8 different coarse states, e.g., the PP_{II}-basin (see Fig. S4 of the supplementary material).[81,82] Thus, there are 8^3 possible states for a given stretch along the sequence

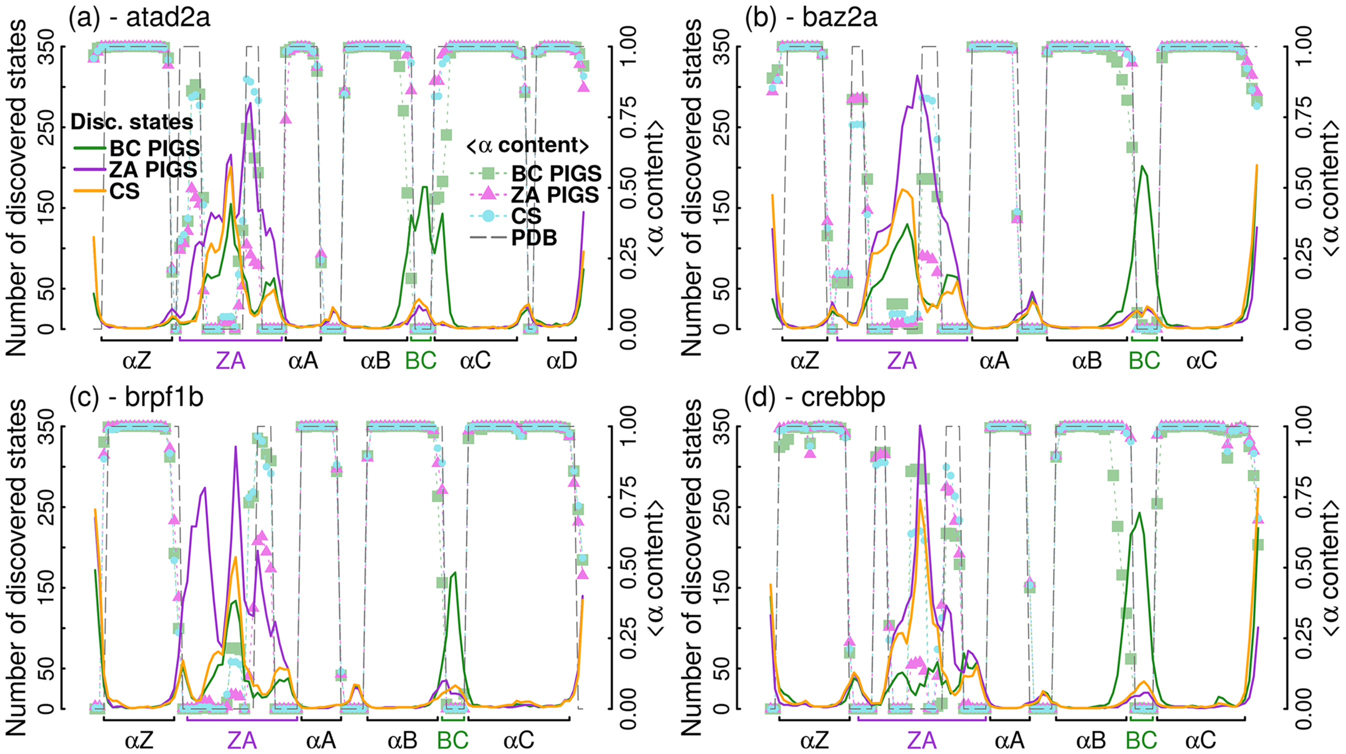


FIG. 4. Number of states discovered by PIGS and CS runs and average α-helical content per residue. States are defined based on per-residue torsional assignments along 3-residue long stretches (see text and Fig. S4 of the supplementary material). Helical content is based on the DSSP algorithm and respects MSMs weights (see Sec. II F and compare Fig. S5 of the supplementary material). Values for the crystal structures are displayed as well. The legend in panel (a) applies to all panels. (a) Data for the atad2a domain (ref. PDB 3DAI). (b) Same as (a) for baz2a (ref. PDB 4LZ2). (c) Same as (a) for brpf1b (ref. PDB 4LC2). (d) Same as (a) for crebbp (ref. PDB 3DWY).

of a bromodomain. By scanning the per-residue assignments along the sequence (excepting the residues closest to the termini), we increment the count of discovered states any time that a stretch is found in a state not previously sampled by it. The final count is assigned to the central residue of the stretch in question. In general, we expect to find larger counts for loop residues, in particular the ZA loop ones and negligible counts for residues that are part of the helix bundle. We also expect that for ZA PIGS and BC PIGS, respectively, the counts clearly exceed the ones encountered in CS for the ZA and BC loops, respectively (but not the other way around). Figure 4 demonstrates that this is the case. It is striking how the PIGS enhancement leads to the discovery of more states for precisely the residues in the respective sets, and how PIGS appears to have no influence elsewhere, i.e., how it accomplished focused exploration. As mentioned above, if the two loops were allosterically coupled, we would have expected a propagation of discovered states between them. This is clearly not the case here since for residues not in the respective PIGS set, the curves are similar across all domains and compare very well with CS. The only exception is the ZA loop of crebbp. In BC PIGS, considerably less states are discovered than in CS, which is consistent with the data presented in Figs. 2(d) and 3(d) and corroborates again the arguments made above in this respect. In addition, while ZA PIGS discovers more states than CS for the ZA loop, the numbers are more similar than for the other domains. This could indicate that the implied time scales of the ZA loop of crebbp are shorter (as confirmed in Sec. III E below). Figure 4 also plots the average per-residue α content, ⟨α⟩, which is always close to 1 for the helical residues as it is for CS. The only consistent exception seems to be the C-terminal cap of the αB helix in BC PIGS. This is an intuitive near-neighbor effect given that the cap is directly adjacent in sequence to residues that are part of the PIGS representation.

We want to point out that a memoryless and time-normalized version of the number of discovered states as plotted in Fig. 4 would provide for a measure of the speed of interconversion between dihedral states at the single residue level in terms of per-replica sampling time. As we will see more clearly below, this masks an underlying separation of time scales for the enhanced segment, i.e., states discovered by PIGS are not only more numerous and more diversified but also kinetically more distant from the crystal structures. This hypothesis is testable by first grouping the snapshots with metrics that describe the torsional states of the BC or ZA loop residues wholly. The second step is then to extract the slow time scales from derived transition networks as shown next.

E. Mean first passage times (MFPTs)

By using grouping (clustering) strategies based on either BC loop or ZA loop torsional angles (Table IV), we construct specific transition networks (Table V) from the sets of trajectories. A PIGS simulation is essentially an ensemble of short trajectories with well-defined start and end points, and the implied cluster connectivity is fully accounted for in our analysis. The transition networks are used to compute steady state weights and MFPTs (see Sec. II E) to the cluster containing the snapshot closest to the reference crystal structure. The groupings are performed on BC PIGS, ZA PIGS, and CS data sets separately. This way, we obtain 3 MFPT curves for a given representation, e.g., BC loop MFPTs for BC PIGS, ZA PIGS, and CS.

In Fig. 5, the reference cluster is always leftmost on the x-axis, and all other clusters are ordered according to their MFPTs to it. It is clear that the MFPTs are much larger whenever there is a match between the enhanced and the analyzed segment; BC MFPTs reach between 0.2 μs and 0.9 μs in BC PIGS and ZA MFPTs reach between 3 μs and 14 μs in ZA PIGS, depending on the bromodomain. ZA MFPTs in ZA PIGS have to be compared to the hundreds of ns reached in the absence of an enhancement of sampling, and this includes the case of crebbp [Fig. 5(d)]. It is a necessary result that the time scales in CS cannot dramatically exceed the simulation length of an individual replica, and this rule extends to PIGS

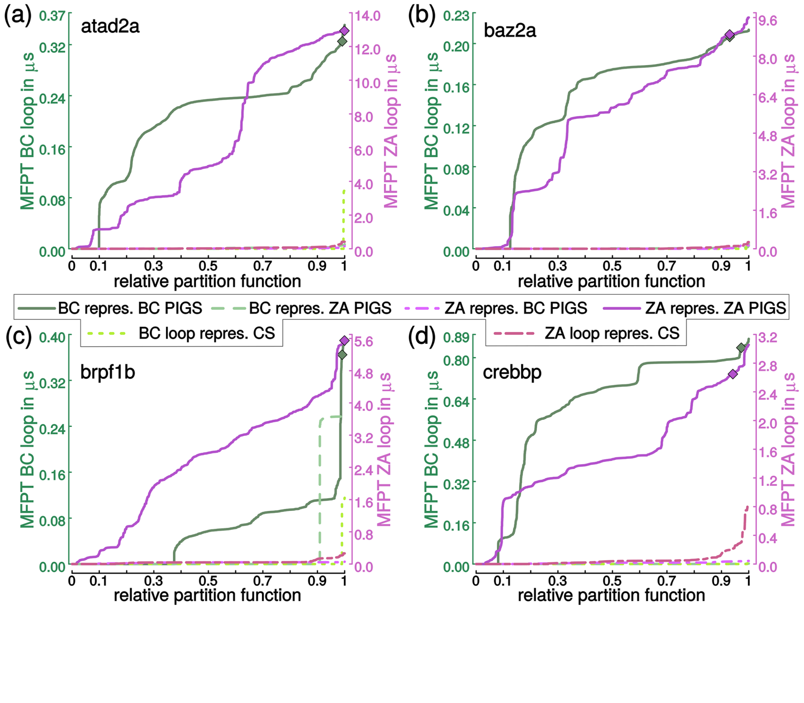


FIG. 5. Mean first passage times to the cluster that contains the structure closest to the reference PDB (distance measured by RMSD across all Cα atoms). Snapshots were clustered with torsional metrics. Clusters are ordered by MFPT and spaced by their MSM steady state weights (see Sec. II E). The cumulative sum of these weights constitutes the relative partition function. Each curve refers to a specific subset of the data (see legend in the middle, which applies to all panels). Diamonds highlight those clusters we selected for investigating their metastabilities (see Secs. III F and II G) (a) Data for the atad2a domain (ref. PDB 3DAI). (b) Same as (a) for baz2a (ref. PDB 4LZ2). (c) Same as (a) for brpf1b (ref. PDB 4LC2). (d) Same as (a) for crebbp (ref. PDB 3DWY).

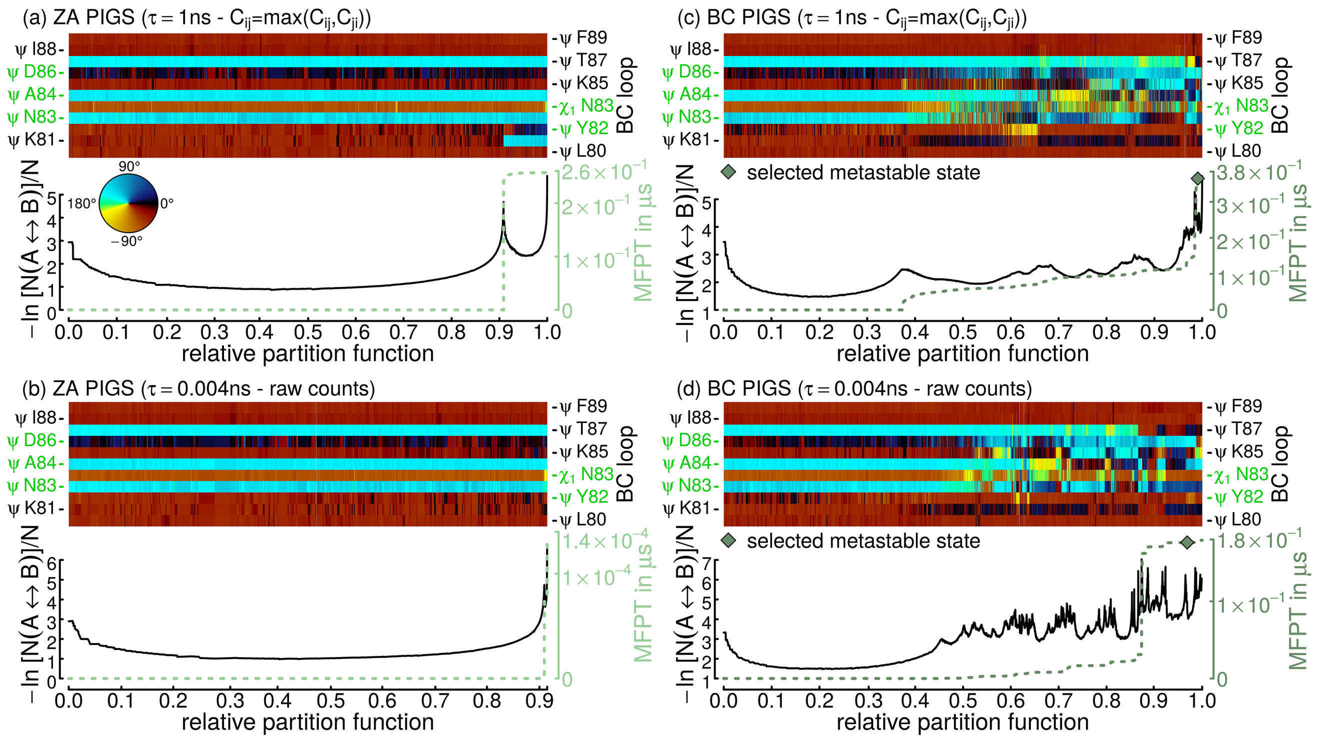


FIG. 6. BC MFPTs, cut-profiles,[83] and torsional state annotations for the brpf1b domain in ZA and BC PIGS. The MFTPs plotted in panels (a) and (c) are the same ones as in Fig. 5(c) for ZA and BC PIGS, respectively. MFPTs are also used to order the clusters along the x-axis, which are spaced by their steady state weights. The cumulative weight defines the relative partition function. Cut profiles (black lines) report on the barriers between all clusters to the left and all clusters to the right at a given point. Torsional annotations are at the top, and the axis labels are found both left and right to improve legibility. The angles included in the BC PIGS representation (Table I) are written in green. For each angle, the value of the centroid of each cluster is taken as a consensus value (vertical white lines are due to resolution limitations of raster images). The color wheel in (a) for the torsion angle values applies to all panels. (a) ZA PIGS, τ = 1.0 ns, and detailed balance is imposed with naïve symmetrization of the count matrix (see Sec. II E). (b) ZA PIGS, τ = 4 ps, and detailed balance is not imposed. (c) Same as (a) for BC PIGS. (d) Same as (b) for BC PIGS.

data sets for segments that were not enhanced. It is important to note that MFPTs exceed the CS background level for 70%-90% of the MSM-weighted data, i.e., PIGS does not just discover a small number of low likelihood states that are kinetically distant from the reference state. In general, the BC loop time scales are intrinsically smaller than those for the ZA loop possibly because the small space allows only few events for structural diversification.

It is evident from Figs. 5(c) and 5(d) that low likelihood events can and do of course occur in non-enhanced sampling, e.g., in BC MFPTs for ZA PIGS on brpf1b or ZA MFPTs for CS data in crebbp. In the brpf1b example, the jump in MFPT is caused by a transition in a single slow coordinate. This is revealed by comparing the MFPTs, cut profiles,[83] and torsional annotations of the brpf1b BC loop [Fig. 6(a)]. We can see that the (pseudo-)free energy barrier and resultant jump in MFPT coincide with the isomerization of the Ψ angle of residue K81 in ZA PIGS simulations. This transition is undersampled, and the actual interconversion rate cannot be estimated with high precision. In fact, without detailed balance imposition (see Sec. II E), the network becomes fractured and the isomerized state is cut off [Fig. 6(b) shows the largest strongly connected component only]. Furthermore, this transition is not sampled by BC PIGS [see Figs. 6(c) and 6(d)] as the Ψ angle of K81 was not part of the BC PIGS representation (Table I). This observation highlights how accurately the protocol can focus sampling enhancements.

Figure 6 demonstrates that the BC MFPTs and cut profiles of ZA PIGS are much more featureless than the BC PIGS ones, and that each dihedral coordinate visits at most two states. Conversely, when the sampling enhancement is on the BC loop, barriers are crossed repeatedly, and this achieves better connectivity and the discovery of more torsional states. Even when detailed balance is not imposed [Fig. 6(d)], the time scales by BC PIGS do not change dramatically [relative to Fig. 6(b)], and the largest strongly connected component still encompasses almost 100% of the data. This suggests that the BC loop, constrained by helices αB and αC, is sampled almost exhaustively in BC PIGS. The main effect of detailed balance imposition at large lag time for BC PIGS is a smoothing of the barriers and a moderate increase of the MFPTs. The second example mentioned above, viz., the ZA MFPTs for CS data in crebbp [Fig. 5(d)], is analyzed in Fig. S6 of the supplementary material and allows the following conclusions: the increase in CS MFPTs is associated with a combination of events [Fig. S6(a) of the supplementary material]; ZA PIGS robustly samples slow transitions beyond the CS time scale [mainly associated with residue Y43, Fig. S6(b) of the supplementary material]; both networks remain well-connected when detailed balance is not imposed [Figs. S6(c) and S6(d) of the supplementary material]; while the MFPTs drop more considerably than in Fig. 6 without detailed balance, the desired time scale gap between CS and ZA PIGS persists.

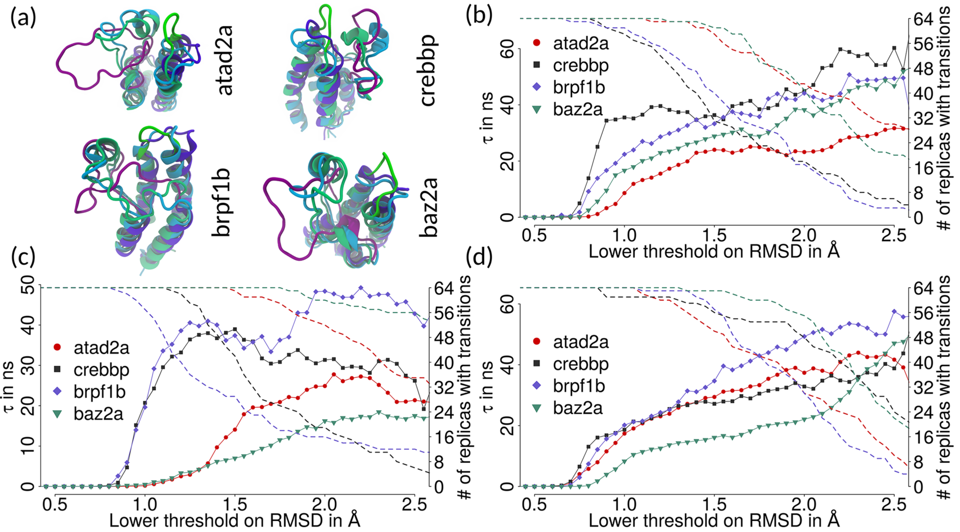


FIG. 7. Metastability of the crystal structure reference states and one kinetically distant state per bromodomain selected from ZA PIGS and BC PIGS each. Metastability was assessed explicitly in 64 independent copies run from the same initial structure (solvent and protein) for 85 ns. A 1 Å-band in RMSD from the initial structure was used to measure escape with the position of the band varied systematically (see Sec. II G). The data were smoothed by the *supsmu* filter[84] in R[85] with a span of 425 ps. (a) Cartoon views of the initial backbone conformations of the proteins in cyan (reference), purple (ZA PIGS-selected state), and green (BC PIGS-selected state) after alignment on the entire domains. (b) Metastability of BC PIGS-selected states. We plot life times from exponential fits (symbols and solid lines) along with the number of replicas contributing to this estimate (dashed lines). (c) The same as (b) for ZA PIGS-selected states. (d) The same as (b) for crystallographic reference states. This analysis is based on the CS data described in Sec. II B and Table I truncated at 85 ns to ensure 1:1 comparability.

The reported MFPTs are important as they point out clearly that the states discovered by PIGS are not just fluctuations around the initial basin but span time scales inaccessible by the CS data. We next confirm that selected discovered states are not just kinetically distant but also metastable.

F. Steady state probabilities and kinetics of metastable states

Figure 6 and Fig. S6 of the supplementary material suggest, in two examples, that PIGS discovers a rich free energy landscape with many states separated by well-defined free energy barriers. This implies that these states should be metastable, i.e., remain self-similar on a time scale of at least nanoseconds. Because PIGS does not rely on energetic biases, it is practically impossible to sample regions of phase space that are enthalpically very unfavorable. In fact, the spontaneous fluctuations promoted by PIGS are extremely unlikely to visit high-energy states, due to the exponential decay in probability. This is often an advantage, as energy-biased approaches may spend a considerable amount of time in states with very low sampling weight. In Fig. 7, we demonstrate that kinetically distant states picked in an *ad hoc* fashion (we required only that snapshots were available with full coordinates, including solvent, which were saved at a much lower frequency) are indeed metastable to a similar extent as crystal structure representatives.

To measure metastability, we performed additional independent simulations for the selected states and relied on the CS data for the crystallographic references, which serve as positive controls. Relaxation clearly proceeded in two stages, a very fast local relaxation on the ps-time scale followed by slower escape processes. To illustrate both, we measured escape from the initial structure with a moving RMSD-threshold band as described in Sec. II G and the caption of Fig. 7. It is clear that a band extending from 1 to 2 Å is the smallest possible choice to describe the slow modes. For lower thresholds, all the way up to 2.5 Å, the life time estimates do not change by more than a factor of ∼2 in most cases. These life times are comparable in their absolute values (20-50 ns). Notable exceptions are the ZA PIGS-selected states for atad2a and baz2a [Fig. 7(c)]. As seen in Fig. 7(a), here the ZA loops extend far into the solvent, and it seems reasonable that the inherent amplitude of conformational fluctuations is larger, thus explaining the generally lower values at all thresholds. It is interesting to note that these two proteins tend to appear the least metastable also in Fig. 7(b) and, partially, Fig. 7(d), suggesting that the conformational rigidity differs in general. As pointed out in Sec. II G, the life time estimates are likely to be severe underestimates. This is highlighted in Fig. 7 by the often small number of trajectories contributing to a given life time.

It is important that residence times of 20-50 ns explain why degrees of freedom in CS and also those in PIGS that are not part of the PIGS representation [e.g., BC loop residues in ZA PIGS as in Fig. 6(a)] cannot converge on the chosen time scale for individual replicas (∼100 ns). Thus, any predictions regarding the thermodynamics of sampled states are heavily biased by the chosen initial condition(s), and the resultant free energy landscape can be a dramatic oversimplification [as in Fig. 6(a)].

IV. DISCUSSION AND CONCLUSIONS

In this contribution, we have addressed a problem of interest, namely, how to focus sampling enhancements to specific parts of proteins to drive the discovery of relevant conformational states. These parts are described by sets of degrees of freedom of dimensionalities as high as 15 (Table I). Complex systems often exhibit emergent behavior that is not known beforehand and difficult to predict, and this difficulty can arise precisely because there are delicate and nonlinear correlations between degrees of freedom, for example, as seen in protein allostery. PIGS ultimately achieves a diversification in a target space, and if this space is chosen appropriately, both the sampling enhancement and the focusing thereof can be achieved with ease.

Specifically, the data in Figs. 2–4 demonstrate that PIGS widens the envelope of discovered states dramatically relative to a CS run even when slightly more resources are used for CS (Table I). Importantly, it does so for a real world application, viz., a protein domain of >100 residues in explicit solvent. The biggest obstacle to brute-force MD is the large time scale of processes of interest. CS offers only limited scaling when deployed to HPC resources, and this problem is thus a fundamental one.[3] Here, PIGS routinely increased the covered time scales (Fig. 5) by up to two orders of magnitude in a focused manner. Given that PIGS returns a set of locally equilibrated trajectories (if the simulation stretches are of reasonable length), we employed a reweighting approach based on MSMs (see Sec. II F) to establish the thermodynamic stability of the discovered states. In all cases, we found that MSM-weights are similar to the raw sampling weights indicating that these states would in fact be sampled spontaneously in much longer MD simulations. To corroborate this result, we analyzed also the kinetic stability of a small selection of states (Fig. 7) and determined residence times comparable to that of the crystallographic reference states.

Below, we summarize some basic recommendations on how to use PIGS efficiently and discuss advantages and limitations in relation to alternative protocols that modify the energy landscape. This is followed by a brief discussion on what emerged regarding bromodomain architecture and an outlook commenting on future developments and the applicability to other systems.

A. What makes an appropriate selection of features for PIGS?

For the present system, a general knowledge of the bromodomain function and architecture was sufficient to make the choices in Table I. This of course implies knowledge of experimental structures, which were a prerequisite to begin with. We suggest the following three guidelines for selecting the PIGS representation to practitioners:

1. Degrees of freedom that diversify rapidly in CS given a target simulation time should be excluded from the PIGS representation. These are usually weakly coupled variables, e.g., the rotamer states of solvent-exposed side chains in a protein. This suggestion applies to any PIGS simulation. The presence of many weakly coupled degrees of freedom evolving quickly creates a combinatorial explosion of states in this high-dimensional space. This combinatorial increase means that all replicas rapidly appear diversified, which deteriorates the reseeding rate. This ultimately results in all "slower" degrees of freedom failing to receive sampling enhancements. This issue is in theory addressed by increasing the number of replicas. While in our experience this is indeed helpful, the combinatorial growth in the number of states means that resource limitations come into play extremely quickly.
2. It is not necessary to include all in a set of closely coupled degrees of freedom. In many cases, couplings are directly apparent, e.g., sets of interatomic distances involving the same pair of protein side chains will be highly correlated. Similarly, next-neighbor couplings in backbone dihedral angles result directly from steric considerations rather than system-specific issues (compare Table I). While the presence of additional but tightly coupled variables generally should have little influence on the PIGS reseeding decisions, their inclusion decreases computational efficiency. This is again a general rule.
3. For focused explorations, there is an additional concern: it is almost certain that not all slow "modes" a system offers are of interest, and some may even be detrimental to a given study. In the concrete example here, observing the onset of the unfolding transition by including helix residues, albeit interesting *per se*, would have created an undesirable overlap of long time scales, which would have prevented clear statements about the states accessible to the two loops and their intrinsic time scales *in the folded state.* Of course, the system must allow for this separation. If loop conformations were invariably linked to the folding equilibrium for bromodomains, this overlap would have been both the inevitable and the biologically relevant result.

B. A comparative assessment of virtues and limitations

As mentioned in Sec. IV A, complex systems can pose the difficulty that several slow processes overlap in time scale. If one is interested in only one or a few of these processes, the ability to focus the sampling enhancement is critical. This is precisely one of the appealing properties of low-dimensional collective variables used as reaction coordinates in methods like umbrella sampling. However, systems of as high a dimensionality as the bromodomains investigated here challenge many advanced sampling methods because low-dimensional reaction coordinates become difficult to define, and because the energy spectrum is unfeasibly wide and the number of relevant states can be so large that approaches requiring explicit human supervision become intractable.

Unlike umbrella sampling, PIGS, like the many other adaptive methods, uses no energetic biases along collective variables. Instead sampling enhancements are achieved by detecting and rewarding spontaneous fluctuations within a (relatively) high-dimensional feature space, which can be tailored to accommodate different needs. The lack of an energetic bias prevents PIGS from sampling states that have very low statistical weight purely on account of their enthalpies. In turn, it is thus reasonable to expect that newly discovered states are indeed metastable rather than being located outside of local free energy minima (Fig. 7). In contrast, methods biasing the potential energy always run the risk of spending significant resources on exploring practically irrelevant states. As a downside, the lack of an energetic bias also means that PIGS does not provide clear benefits when a barrier is fully enthalpic, e.g., in the *cis*/*trans* isomerization of a single polypeptide ω bond. In practice, interesting free energy barriers are probably not generally of this type but rather tend to involve many degrees of freedom. As we showed in Ref. 9, globular states of an α-helix-forming peptide found at low temperatures, which seemed enthalpically trapped, were readily diversified with PIGS.

It is useful to recall a general limitation of focused approaches (regardless of methodology): it is never possible to rigorously ensure the equilibration of degrees of freedom in parts of the system that the sampling enhancements were not focused on. This means that the system's intrinsic exploration rates are no longer uniform (as they are in CS), which implies that initial conditions persist to different extents for different parts of the system. In umbrella sampling, this limitation is known as the difficulty to achieve equilibration of orthogonal degrees of freedom.[86] In focused PIGS, it is manifest clearly, for example, in Figs. 3 and 4. The view of the state space is partial when focusing on one loop at a time, and states with joint kinetically distant states of both loops are absent in this data set. Importantly, the likelihood of such states can be predicted *a posteriori* by taking the products of MSM probabilities from the two ensembles of PIGS simulations. There is also no fundamental limitation to enlarge the representation to include both loops in the PIGS representation simultaneously (see Secs. IV A and IV C).

An additional advantage of PIGS is the absence of any supervision once a representation has been chosen. Bearing the caveat in mind that weakly coupled and "fast" degrees of freedom should be avoided (see Sec. IV A), PIGS offers the favorable property that those degrees of freedom most amenable to spontaneous change can drive the diversification, and that these coordinates can and do emerge on-the-fly in an unsupervised manner, i.e., they are effectively learned by the algorithm.

C. Insights into bromodomain architecture

Bromodomains have the conserved helical fold shown in Fig. 1(b). Our results show clearly that it is possible to specifically enhance the time scales of exploration and the explored conformational space for either the BC or the ZA loop without significantly altering the properties of the other loop or the helix bundle relative to CS. Because PIGS provides a relatively gentle way of enhancing the sampling in a focused manner, the diversification is unlikely to propagate to parts that are not directly enhanced unless their conformational fluctuations are tightly coupled. In this study, we always kept the two loops separated, and our results demonstrate implicitly that there is little allosteric cross talk between the BC and ZA loops for any of the domains. We have suggested ZA loop disorder as a recruitment vehicle in recent work,[87] and the lack of coupling would consequently imply that the recruitment step cannot predispose the BC loop toward conformations compatible with binding. However, this may indeed be unnecessary: Fig. 5 suggests that the BC loop of bromodomains is less likely to be the kinetic bottleneck in rearrangements upon the (un)binding of natural and pharmaceutical ligands than the ZA loop, which changes conformations on time scales that reach well into the μs-regime.

Based on Figs. 3 and 4 in particular, it is clear that our state space is truncated, i.e., kinetically distant states for both loops are not populated jointly. While it would require using a larger number of replicas to maintain a comparable reseeding rate, it is both feasible and meaningful to include the two loops in a joint PIGS representation. We have chosen not to do so here because we wanted to highlight focused sampling enhancements. From a modeling point of view, the construction of chimeric structures as receptors for virtual screening campaigns on bromodomains should be possible. Such a piecewise reconstruction may obviously be more useful for larger systems with multiple independent components.

Clearly, both the ZA and BC loops are directly adjacent to the helix bundle both in sequence and in space. However, aside from a minor cap effect visible in Fig. 4, we detected no influence of loop diversification on the helix bundle. This suggests that the rearrangement of the bromodomain tertiary structure and the onset of unfolding transitions are either much slower in time scale or at most weakly coupled to loop conformation (or both). This is different from the critical role turn sequences are known to play in the formation of β-sheets.[88]

D. Applicability to other problems and outlook

While for the bromodomains, no significant coupling between the two loops emerged, focused conformational explorations with PIGS do not mandate that this be the case. Allosteric effects caused by strong couplings have been revealed by PIGS in other types of systems, e.g., amyloid fibrils.[76] This means that degrees of freedom not part of the PIGS representation will automatically respond to the sampling enhancement because the spontaneous fluctuations are necessarily coupled themselves. In general, the strength of allosteric effects depends on the physical and chemical properties of the system, and they are difficult to assess *a priori*. PIGS could be used precisely to answer questions regarding their strength.

We thus predict that the possibility to focus phase space explorations on a broad set of degrees of freedom (from one to many, from dihedral angles to collective variables, etc.) in a simple and yet precise manner can be useful in many applications. In computer simulations of biomolecules, these applications could include the diversification of receptor conformations or the discovery of allosteric binding sites for drug design, studies of the propagation of signals through structural changes and binding/unbinding processes, or the disorder-to-order transition of molecular recognition features.[89] It is important to realize that PIGS works as long as there is sampling redundancy. It is therefore not the right choice for systems that diversify spontaneously in CS on the affordable time scale or that already offer multiple and heterogeneous initial conditions, e.g., an expanded intrinsically disordered protein in solution. Even then, it may still be possible to tailor the representation of the system toward a feasible goal, e.g., by using coarse collective variables as features such as the radius of gyration or secondary structure content.

In our experience with biomacromolecules, representations based on torsional angles or interatomic distances are both well-suited to drive conformational changes in ordered systems, and we have used anywhere from 2 to about 150 of these "atomistic" features at once. RMSD with alignment can be used as well even though the alignment operation slows down the data analysis steps. We have also explored other types of features such as contact patterns or solvent accessibility measures. Interested readers are referred to the available publications[9,76] and to the documentation of CAMPARI (http://campari.sourceforge.net). As a more technical conclusion, we demonstrated here that the analysis algorithms in CAMPARI can in principle be applied to implement PIGS with any propagation code. The PIGS reseeding decisions will work as intended so long as the parts of interest of the system undergo stochastic evolution with the potential for sampling recurrence and overlap. Thus, outside the life science community, PIGS can be potentially useful in applications as diverse as numerical optimization using Monte Carlo algorithms, agent-based simulations of financial markets,[90] or, of course, any type of particle-based molecular simulations. By splitting the reseeding heuristic and the propagation engine, PIGS is a versatile tool deployed easily on most HPC architectures.

Ongoing work explores a number of avenues. First, we continue our efforts to parallelize the data mining steps efficiently across the entire set of resources allocated to a given PIGS run. Second, we want to fine-tune the representation according to the evolution of the system. Specifically, the idea is to use feature weights capable of dynamically emphasizing degrees of freedom for which no or little diversification has been detected. This would allow these features to benefit maximally from the sampling enhancements and avoid that fast motions mask redundancy in the replicas, especially in high-dimensional feature spaces (as described in Sec. IV A). Another approach would be to use dimensionality reduction techniques.[91] Third, we are trying to understand any potential fault lines in the MSM-based thermodynamic reweighting process employed here. The removal of initial state bias is a tricky problem, and an error estimation and reweighting strategy specific to PIGS would be of particular interest. Forth, we are using PIGS to tackle challenging problems, including the (un)binding of disordered peptides from/to bromodomains or amyloid (proto)fibrils.[76] Given the success demonstrated here, we anticipate that PIGS can be useful in a large number of specific research questions involving molecular systems.

SUPPLEMENTARY MATERIAL

See supplementary material for additional Figs. S1-S10 mentioned in the main text.

ACKNOWLEDGMENTS

This work was supported by a grant from the Swiss National Science Foundation (SNF #149897) to A.C., by a grant from the Platform for Advanced Scientific Computing (PASC "Scalable Advanced Sampling in Molecular Dynamics") to A.V. and A.C., and by a CPU time grant from CSCS ("s636") to M.B., A.C., and A.V.

[1] T. Schlick, R. Collepardo-Guevara, L. A. Halvorsen, S. Jung, and X. Xiao, Q. Rev. Biophys. **44**, 191 (2011).
[2] A. Gray, O. G. Harlen, S. A. Harris, S. Khalid, Y. M. Leung, R. Lonsdale, A. J. Mulholland, A. R. Pearson, D. J. Read, and R. A. Richardson, Acta Crystallogr., Sect. D: Biol. Crystallogr. **71**, 162 (2015).
[3] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw, Annu. Rev. Biophys. **41**, 429 (2012).
[4] M. Karplus and J. A. McCammon, Nat. Struct. Biol. **9**, 646 (2002).
[5] W. F. Van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger, M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. A. Van Der Vegt, and H. B. Yu, Angew. Chem., Int. Ed. Engl. **45**, 4064 (2006).
[6] D. Shukla, Y. Meng, B. Roux, and V. S. Pande, Nat. Commun. **5**, 3397 (2014).
[7] S. Genheden and U. Ryde, Phys. Chem. Chem. Phys. **14**, 8662 (2012).
[8] S. D. Bond and B. J. Leimkuhler, Acta Numer. **16**, 1 (2007).
[9] M. Bacci, A. Vitalis, and A. Caflisch, Biochim. Biophys. Acta **1850**, 889 (2015).
[10] D. M. Zuckerman, Annu. Rev. Biophys. **40**, 41 (2011).
[11] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu, PLoS Comput. Biol. **12**, e1004619 (2016).
[12] S. Park and K. Schulten, J. Chem. Phys. **120**, 5946 (2004).
[13] X. Wu and B. R. Brooks, Chem. Phys. Lett. **381**, 512 (2003).
[14] X. Wu and S. Wang, J. Chem. Phys. **110**, 9401 (1999).
[15] I. Andricioaei, A. R. Dinner, and M. Karplus, J. Chem. Phys. **118**, 1074 (2003).
[16] X. Wu, M. Hodoscek, and B. R. Brooks, J. Chem. Phys. **137**, 044106 (2012).
[17] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[18] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).
[19] V. Leone, F. Marinelli, P. Carloni, and M. Parrinello, Curr. Opin. Struct. Biol. **20**, 148 (2010).
[20] L. Sutto, S. Marsili, and F. L. Gervasio, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **2**, 771 (2012).
[21] G. A. Tribello, M. Ceriotti, and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **107**, 17509 (2010).
[22] W. Wojtas-Niziurski, Y. Meng, B. Roux, and S. Bernèche, J. Chem. Theory Comput. **9**, 1885 (2013).
[23] D. Hamelberg, J. Mongan, and J. A. McCammon, J. Chem. Phys. **120**, 11919 (2004).
[24] W. Sinko, Y. Miao, C. A. F. de Oliveira, and J. A. McCammon, J. Phys. Chem. B **117**, 12759 (2013).
[25] W. Han and K. Schulten, J. Am. Chem. Soc. **136**, 12450 (2014).
[26] G. Henkelman, B. P. Uberuaga, and H. Jónsson, J. Chem. Phys. **113**, 9901 (2000).
[27] E. Weinan, W. Ren, and E. Vanden-Eijnden, J. Phys. Chem. B **109**, 6688 (2005).
[28] P. G. Bolhuis, C. Dellago, D. Chandler, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).
[29] L. T. Chong, A. S. Saglam, and D. M. Zuckerman, Curr. Opin. Struct. Biol. **43**, 88 (2017).
[30] R. J. Allen, D. Frenkel, and P. R. ten Wolde, J. Chem. Phys. **124**, 024102 (2006).
[31] G. A. Huber and S. Kim, Biophys. J. **70**, 97 (1996).
[32] B. W. Zhang, D. Jasnow, and D. M. Zuckerman, Proc. Natl. Acad. Sci. U. S. A. **104**, 18043 (2007).
[33] F. A. Escobedo, E. E. Borrero, and J. C. Araque, J. Phys.: Condens. Matter **21**, 333101 (2009).
[34] M. C. Zwier, J. L. Adelman, J. W. Kaus, A. J. Pratt, K. F. Wong, N. B. Rego, E. Suárez, S. Lettieri, D. W. Wang, and M. Grabe, J. Chem. Theory Comput. **11**, 800 (2015).
[35] V. S. Pande, I. Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C. D. Snow, E. J. Sorin, and B. Zagrovic, Biopolymers **68**, 91 (2002).
[36] G. R. Bowman, D. L. Ensign, and V. S. Pande, J. Chem. Theory Comput. **6**, 787 (2010).
[37] T. Zhou and A. Caflisch, J. Chem. Theory Comput. **8**, 2134 (2012).
[38] W. Zheng, M. A. Rohrdanz, and C. Clementi, J. Phys. Chem. B **117**, 12769 (2013).
[39] A. Dickson and C. L. Brooks III, J. Phys. Chem. B **118**, 3532 (2014).

[40] M. I. Zimmerman and G. R. Bowman, J. Chem. Theory Comput. **11**, 5747 (2015).

[41] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, J. Chem. Phys. **134**, 174105 (2011).

[42] V. S. Pande, K. Beauchamp, and G. R. Bowman, Methods **52**, 99 (2010).

[43] X. Huang, G. R. Bowman, S. Bacallado, and V. S. Pande, Proc. Natl. Acad. Sci. U. S. A. **106**, 19765 (2009).

[44] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. U. S. A. **106**, 19011 (2009).

[45] A. Vitalis and A. Caflisch, J. Chem. Theory Comput. **8**, 1108 (2012).

[46] N. Blöchliger, A. Vitalis, and A. Caflisch, Comput. Phys. Commun. **184**, 2446 (2013).

[47] X. de la Cruz, S. Lois, S. Sánchez-Molina, and M. A. Martínez-Balbás, BioEssays **27**, 164 (2005).

[48] P. Filippakopoulos, S. Picaud, M. Mangos, T. Keates, J. P. Lambert, D. Barsyte-Lovejoy, I. Felletar, R. Volkmer, S. Muller, T. Pawson, A. C. Gingras, C. H. Arrowsmith, and S. Knapp, Cell **149**, 214 (2012).

[49] P. Filippakopoulos and S. Knapp, FEBS Lett. **586**, 2692 (2012).

[50] J. R. Marchand and A. Caflisch, ChemMedChem **10**, 1327 (2015).

[51] A. H. Lund and M. van Lohuizen, Genes Dev. **18**, 2315 (2004).

[52] M. Esteller, N. Engl. J. Med. **358**, 1148 (2008).

[53] J. S. You and P. A. Jones, Cancer Cell **22**, 9 (2012).

[54] C. Plass, S. M. Pfister, A. M. Lindroth, O. Bogatyrova, R. Claus, and P. Lichter, Nat. Rev. Genet. **14**, 765 (2013).

[55] C. Koschmann, F. J. Nunez, F. Mendez, J. A. Brosnan-Cashman, A. K. Meeker, P. R. Lowenstein, and M. G. Castro, Cancer Res. **77**, 227 (2017).

[56] M. A. Dawson, T. Kouzarides, and B. J. P. Huntly, N. Engl. J. Med. **367**, 647 (2012).

[57] R. K. Prinjha, J. Witherington, and K. Lee, Trends Pharmacol. Sci. **33**, 146 (2012).

[58] O. Khan and N. B. La Thangue, Immunol. Cell Biol. **90**, 85 (2012).

[59] C. H. Arrowsmith, C. Bountra, P. V. Fish, K. Lee, and M. Schapira, Nat. Rev. Drug Discovery **11**, 384 (2012).

[60] P. Filippakopoulos and S. Knapp, Nat. Rev. Drug Discovery **13**, 337 (2014).

[61] S. Steiner, A. Magno, D. Huang, and A. Caflisch, FEBS Lett. **587**, 2158 (2013).

[62] G. Schneider, Nat. Rev. Drug Discovery **9**, 273 (2010).

[63] D. Spiliotopoulos and A. Caflisch, Isr. J. Chem. **54**, 1084 (2014).

[64] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, SoftwareX **1–2**, 19 (2015).

[65] J. Huang and A. D. MacKerell, J. Comput. Chem. **34**, 2135 (2013).

[66] G. Bussi, D. Donadio, and M. Parrinello, J. Chem. Phys. **126**, 014101 (2007).

[67] I. G. Tironi, R. Sperb, P. E. Smith, and W. F. van Gunsteren, J. Chem. Phys. **102**, 5451 (1995).

[68] S. Miyamoto and P. A. Kollman, J. Comput. Chem. **13**, 952 (1992).

[69] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).

[70] C. Tallant, E. Valentini, O. Fedorov, L. Overvoorde, F. M. Ferguson, P. Filippakopoulos, D. I. Svergun, S. Knapp, and A. Ciulli, Structure **23**, 80 (2015).

[71] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).

[72] R. C. Edgar, Nucleic Acids Res. **32**, 1792 (2004).

[73] R. Sanchez, J. Meslamani, and M.-M. Zhou, Biochim. Biophys. Acta, Gene Regul. Mech. **1839**, 676 (2014).

[74] Y. Xu, S. Zhang, S. Lin, Y. Guo, W. Deng, Y. Zhang, and Y. Xue, Nucleic Acids Res. **45**, D264 (2017).

[75] A. K. Jain and M. C. Barton, J. Mol. Biol. **429**, 2003 (2017).

[76] M. Bacci, J. Vymětal, M. Mihajlovic, A. Caflisch, and A. Vitalis, J. Chem. Theory Comput. **13**, 5117 (2017).

[77] J. A. Hardy and J. A. Wells, Curr. Opin. Struct. Biol. **14**, 706 (2004).

[78] S. Lu, W. Huang, and J. Zhang, Drug Discovery Today **19**, 1595 (2014).

[79] I. Jolliffe, *Principal Component Analysis* (Wiley Online Library, 2005).

[80] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).

[81] A. Vitalis and A. Caflisch, J. Chem. Theory Comput. **8**, 363 (2012).

[82] A. Vitalis and R. V. Pappu, J. Chem. Phys. **141**, 034105 (2014).

[83] S. V. Krivov and M. Karplus, J. Phys. Chem. B **110**, 12689 (2006).

[84] J. H. Friedman, Stanford University Technical Report No. 5, 1984.

[85] R Core Team, *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria, 2017).

[86] F. Zhu and G. Hummer, J. Comput. Chem. **33**, 453 (2012).

[87] C. Langini, A. Caflisch, and A. Vitalis, J. Biol. Chem. **292**, 16734 (2017).

[88] K. A. Olsen, R. M. Fesinmeyer, J. M. Stewart, and N. H. Andersen, Proc. Natl. Acad. Sci. U. S. A. **102**, 15483 (2005).

[89] A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker, and V. N. Uversky, J. Mol. Biol. **362**, 1043 (2006).

[90] T. Lux and M. Marchesi, Nature **397**, 498 (1999).

[91] Z. Tang and C. A. Chang, J. Chem. Theory Comput. **13**, 2230 (2017).