Collision-Free Poisson Motion Planning in Ultra High-Dimensional Molecular Conformation Spaces

Rasmus Fonseca[*], Dominik Budday[†], Henry van den Bedem[‡]

November 15, 2017

Abstract

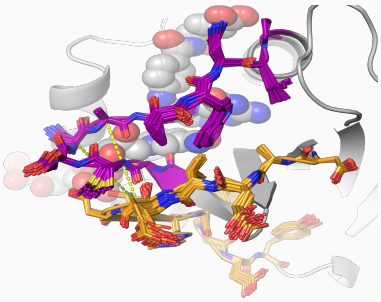
The function of protein, RNA, and DNA is modulated by fast, dynamic exchanges between three-dimensional conformations. Conformational sampling of biomolecules with exact and nullspace inverse kinematics, using rotatable bonds as revolute joints and non-covalent interactions as holonomic constraints, can accurately characterize these native ensembles. However, sampling biomolecules remains challenging owing to their ultra-high dimensional configuration spaces, and the requirement to avoid (self-) collisions, which results in low acceptance rates. Here, we present two novel mechanisms to overcome these limitations. First, we introduce temporary constraints between near-colliding links. The resulting constraint varieties instantaneously redirect the search for collision-free conformations, and couple motions between distant parts of the linkage. Second, we adapt a randomized Poisson-disk motion planner, which prevents local oversampling and widens the search, to ultra-high dimensions. Tests on several model systems show that the sampling acceptance rate can increase from 16% to 70%, and that the conformational coverage in loop modeling measured as average closeness to existing loop conformations doubled. Correlated protein motions identified with our algorithm agree with those from MD simulations.

Keywords: High Dimensional Motion and Path planning, Collision-avoidance, Computational Biology, Inverse Kinematics, Molecular Rigidity ■

[*]Stanford University, Molecular and Cellular Physiology, Stanford, CA, USA, rfon@stanford.edu

[†]University of Erlangen-Nuremberg, Chair of Applied Dynamics, 91058 Erlangen, Germany

[‡]SLAC National Accelerator Laboratory, Bioscience Division, Stanford University, Menlo Park, CA, USA, vdbedem@stanford.edu



The function of protein, RNA, and DNA is intimately associated with exchanges between three-dimensional conformations. Computationally modeling these motions is challenging due to their ultra-high dimensions (> 50) and rugged, sterically determined energy landscapes. To overcome these limitations, we first introduce temporary constraints between near-colliding atoms, instantaneously redirecting the search for collision-free conformations. Coupled to an adapted randomized Poisson-disk motion planner that ensures uniform sampling, we are able to expand the search to ultra-high dimensions, resulting in broad conformational ensembles closely related to functional mechanisms.

INTRODUCTION

Many proteins interact with their partners and perform their cellular functions by rapidly interchanging between three-dimensional substates[1]. Computational methods to readily characterize the conformational landscape of folded proteins or their complexes would help us interpret ensemble-averaged, experimental data, and can provide insight into how they function. Molecular dynamics simulations can reveal time-resolved, atomically detailed trajectories, but require sophisticated resources to overcome spatiotemporal barriers separating functional substates[2]. By contrast, non-deterministic conformational sampling procedures, such as Monte Carlo sampling, can rapidly collect a representative set of conformations but disregard time-scales of molecular change.

Fast, robotics-inspired algorithms for randomized exploration of molecular conformation spaces are quickly emerging as an alternative to these more traditional molecular simulations[3,4]. They are frequently designed around geometric motion planners, such as probabilistic roadmaps (PRMs,[5–7]), rapidly-exploring random trees (RRTs,[8–10]) or their variants[11,12].

While these planners efficiently handle high-dimensional configuration spaces, they require adaptations to sample broadly and uniformly when the dimensionality is ultra-high ($>$ 50), for example in molecular simulations[9,12–14].

These approaches often represent a protein as a kinematic linkage, with groups of atoms as rigid bodies and rotatable bonds as revolute joints (degrees-of-freedom DOF, Fig. 1.a). Randomly perturbing the rotatable bonds would quickly lead to unfolding the protein. Instead, the rotatable bonds require coordinated changes to maintain non-covalent interactions, such as hydrogen bonds, in the protein. Nullspace inverse kinematics[15] is an efficient technique to deform a protein while observing constraints[12,16]. In nullspace inverse kinematics, we express non-covalent interactions in the protein as holonomic constraints[17], which define a lower-dimensional variety in conformational space[15]. Deformations of the protein structure on the constraint variety therefore maintain the folded state.

Alternative, non-native contacts such as steric interactions by repulsive van der Waals forces (clashes) also play a critical role in protein conformational dynamics by stabilizing native states and redirecting the motion of coordinated degrees-of-freedom[18]. However, clashes

also severely hinder fast exploration of conformational space. This is particularly the case for proteins, whose cores are densely packed. While several techniques have been developed to address steric hindrance and increase the efficiency of conformational sampling, most neglect the role of non-native contacts by avoiding, retrospectively relieving, or simply ignoring clashes. For example, a common solution is to accept or reject a trial conformation after checking for clashes[9]. Alternatively, highly reduced structural representations of the protein[19,20] or prior information constraints and filtering were proposed to avoid steric clashes[21]. Another, highly efficient, approach is to combine a reduced structural representation with normal modes as the search space for an RRT[10]. Alternatively, the PCA-EA algorithm[13] takes a population of known conformations as input, computes PCA on their C_α coordinates, and performs conformational perturbations only on the principal components. In the latter cases, the side chains are only adjusted to relieve clashes once a new conformation was found. Recently, RRT was combined with basin hopping to explore low-energy regions of the landscape of the BLN69 protein[14].

These procedures either increase the efficiency of the algorithms at the expense of atomically detailed collision avoidance, or employ expensive collision checking and energy evaluations. However, most importantly, these solutions fail to account for functionally important, sterically coupled motions in proteins that govern their conformational dynamics.

Here, we address both limitations simultaneously by planning collision-avoiding motions. We capitalize on the observation that protein kinematic linkages are highly redundant. Sugiura *et al.*[22] earlier proposed velocity-based nullspace techniques to avoid self-collisions by introducing virtual, repulsive forces between near-colliding links. Petric *et al.*[23] recently proposed to project desired Cartesian velocities away from fixed obstacles onto nullspace joint velocities. In these solutions, collisions are locally resolved by prescribing desired velocities on the degrees-of-freedom.

Our Contributions. Here, we introduce a different approach that avoids explicitly prescribing velocities, but instead lets the global deformation determine how local collisions are resolved. We create a temporary one-dimensional constraint between near-colliding links, thereby altering the constraint variety of the linkage. The one-dimensional constraint prevents the two near-clashing links from moving closer only in the direction of their shortest

distance, but allows all other motions to let the atoms slide past each other (Fig. 1.b,c). Velocities on this new constraint variety are collision-avoiding, corresponding to a new search direction in conformational space. By adding desired constraints in successive sampling steps our algorithm hops between collision-free constraint varieties.

We combined our clash-avoiding procedure with a new motion planner we call Poisson-Explore, inspired by Poisson disk sampling[24]. Poisson Disk sampling generates random samples that 1) are separated by at least the radius of the Poisson disk, and 2) satisfy the *maximal sampling property*, which requires that the disks completely cover conformation space. These properties ensure broad and uniform sampling, but are difficult to meet in practice when the geometry of collision-free conformation space $\mathcal{C}_{\text{free}}$ is unknown *a priori.* Recently, Manocha and coworkers introduced Poisson-RRT: a Poisson Disk-based motion planner[25]. Despite favorable sampling properties, the complexity of Poisson planners scales exponentially with the dimension of conformation space. We therefore designed an efficient multi-query randomized Poisson disk-based planner adapted to ultra-high dimensional molecular conformational spaces for which $\mathcal{C}_{\text{free}}$ is unknown. To reduce the number of neighborhood queries, we introduced a bounding volume hierarchy (BVH) that was inspired by collision detection algorithms[26,27].

Our contributions allow us to efficiently explore the conformational landscape of biomolecules, from large loop motions with > 50 degrees of freedom to whole molecules with hundreds or even thousands of degrees of freedom. It identifies collective motions, propagated by native and non-native contacts through constraints, which give insight into the mechanisms of molecular function.

METHODOLOGY

We designed our algorithms around the Kino-Geometric Sampling (KGS) software framework (`https://simtk.org/projects/kgs/`,[12,28]). The purpose of KGS is to efficiently explore feasible regions of a molecule's conformational space. To represent a conformation we first construct a molecular graph, $G_m = (V_m, E_m)$, such that V_m contains all atoms and E_m contains all covalent bonds (Fig. 1.a). Next, a rigid-body graph $G_r = (V_r, E_r)$ is constructed

from G_m by edge-contracting edges that correspond to double bonds or are part of pentameric rings (RNA ribose or the proline amino acid). Vertices in V_r are sets of atoms that form rigid bodies and edges in E_r are rotatable covalent bonds that form revolute joints. For linearly branched multi-chain molecules like proteins and RNA, this results in a set of acyclic trees where each tree represents one chain. For generally branched molecules, we can identify the minimum spanning tree and add left-out edges as constraints. Finally, we connect each tree in G_r to a super-root v_s via six global DOFs resulting in a single rooted tree, $T_k = (V_k, E_k)$. Forward kinematics are defined as propagation of atom coordinate transformations from $v_s \in V_k$, along the direction of edges in E_k. A conformation is represented as a vector $\mathbf{q} \in \mathbb{R}^n$ of all rotatable covalent bonds (DOFs) in T_k, where $n = |E_k|$.

The constraints are defined by sets of atomic pairs, C_k, in the molecule for which the local geometry (pair distance and angles to surrounding covalent bonds) must be maintained (see Fig.1.b). For example, in KGS, hydrogen bonds with energies below -1 kcal/mol are automatically detected using the hydrogen bond potential[29]):

$$E_{HB} = D_0 \left[5 \left(\frac{R_0}{R} \right)^{12} - 6 \left(\frac{R_0}{R} \right)^{10} \right] F(\theta, \phi, \psi) \tag{1}$$

where $R_0 = 2.8$ Å and $D_0 = 8$ kcal/mol are hydrogen-bond ideal distance and well-depth, R is donor-acceptor distance, and F is an angle term that depends on the hybridization state of donor and acceptor. Before starting sampling, the acceptor/donor atom-pairs are added to C_k as constraints. Only the rotation around a hydrogen bond axis is permitted, which leads to five constraints per hydrogen bond. Even if rotation around a hydrogen bond axis increases its energy above -1 kcal/mol the bond is maintained all through the simulation. For hydrogen bonds in secondary structures this is often a reasonable assumption, but in systems where functional mechanisms involve breaking and forming new bonds one would have to extend the simulations strategy to fit the specific problem.

Additionally, any covalent bonds that induce cycles in T_k are added to C_k as well. Optionally, the user can add additional constraints, for example disulfide bridges. The constraints enforce coordinated motion of DOFs across the protein or RNA, and can even rigidify DOFs or rotations of hydrogen bonds, which lead to larger rigid clusters. We previously showed that our geometric approach decomposes macromolecules into identical rigid clusters as those

found by topological rigidity-analysis based procedures[17].

A conformation is considered feasible if covalent bond lengths, constraint lengths, and bond angles remain constant in linear approximation and if no atoms are clashing. The geometry of covalent bonds is maintained implicitly by using revolute joints for the internal edges. Clashing atoms are detected by hashing atom coordinates into a 3D grid of $1 \times 1 \times 1$ (Å^3) cells and for each atom performing an expected constant-time query[30] in neighboring cells. The grid is rebuilt after each iteration. As protein cores are extremely tightly packed we multiply their van der Waals radii with 0.75 before checking if they overlap. Taking steps in conformational space while maintaining the geometry of constraints is the responsibility of the *conformational perturbations* and coordinating the use of these operators is the responsibility of the *planners* as described in the following sections.

Conformational Perturbations

To comprehensively explore conformational space we employ two distinct operators termed analytical inverse kinematics (AIK) and nullspace inverse kinematics (NIK) perturbation.

The AIK perturbation[31] takes a sub-chain spanned by three *pivot atoms* with a total of six adjacent rotational DOFs and analytically computes all possible closed conformations of the sub-chain. It is a requirement that each pivot atom is the end-point of exactly two distinct rotational axes and that there are no hydrogen bonds or other constraints in the sub-chains interior. This perturbation allows jumps between unconnected regions of the feasible conformational manifold and serves to generate distinct seed conformations for other, "local" perturbations.

The $5m$ holonomic constraints $\mathbf{\Phi} = \mathbf{\Phi}(\mathbf{q})$ from m hydrogen bonds define a constraint variety

$$\mathcal{V}_{\text{hb}} = \{\mathbf{q} \in \mathbb{R}^n \mid \mathbf{\Phi}(\mathbf{q}) = \mathbf{0}\}. \tag{2}$$

of dimension at least $n - 5m$. NIK perturbations are local perturbations on $\mathcal{V}_{\text{hb}}$, which are taken from the tangent space $\mathcal{T}_{\mathbf{q}}(\mathcal{V}_{\text{hb}})$ to $\mathcal{V}_{\text{hb}}$ at $\mathbf{q}$. NIK perturbations maintain the holonomic constraints in linear approximation by projecting a trial perturbation $\boldsymbol{\delta}_q$ onto the nullspace of the constraint Jacobian $\mathbf{J}$ evaluated at a seed conformation $\mathbf{q}$[17]. Right-singular vectors

corresponding to vanishing singular values in a singular value decomposition (SVD,[32]) form an orthonormal basis $\mathbf{N}$ for the nullspace of $\mathbf{J}$. We obtain an admissible perturbation $\boldsymbol{\Delta}_q$ through projection of a trial perturbation $\boldsymbol{\delta}_q$ via

$$\boldsymbol{\Delta}_q = \mathbf{N}\mathbf{N}^T\boldsymbol{\delta}_q. \tag{3}$$

Despite the linear nature of this equation, we have previously demonstrated how hydrogen bond-length deformations are limited to about 5% a majority of the time[33].

In addition to constraints defined by native contacts, such as hydrogen bonds, we designed a new procedure to modulate non-native contacts by adding 1-D dynamic clash-avoiding constraints (dCC). Whenever a perturbation in $\mathcal{V}$ leads to a prohibitive steric clash between atoms, the conformation can not be accepted. However, instead of discarding the intended search direction $\boldsymbol{\delta}_\mathbf{q}$ to find a new conformation, we redirect the perturbation onto a new variety $\mathcal{V}_{\text{dCC}}$ by introducing c new dCCs, which prevents each of the c clashing atom-pairs from approaching each other. Given two clashing atoms centered at $\mathbf{p}_i$ and $\mathbf{p}_j$, the dCC

$$\mathbf{n}_c^T\left(\frac{\partial\mathbf{p}_j}{\partial\mathbf{q}} - \frac{\partial\mathbf{p}_i}{\partial\mathbf{q}}\right)\boldsymbol{\delta}_\mathbf{q} = 0 \tag{4}$$

on the desired perturbation $\boldsymbol{\delta}_\mathbf{q}$ allows free motions of $\mathbf{p}_1$ and $\mathbf{p}_2$ within the contact plane (Fig. 1.c)), but only a joint motion in the clash direction $\mathbf{n}_c$. In essence, the two atoms can slide past each other but maintain their distance with respect to $\mathbf{n}_c$. The constraints are formulated individually for each pair of clashing atoms and added as an additional row to the constraint Jacobian matrix $\mathbf{J}$. An SVD leads to a basis $\mathbf{N}_{\text{dCC}} \in \mathbb{R}^{d\times(d-r')}$ for the equal or lower-dimensional nullspace of the Jacobian with rank $r' \geq r$. Accordingly, we obtain the corresponding, clash-avoiding nullspace perturbation using the previously introduced projection $\boldsymbol{\Delta}_\mathbf{q} = \mathbf{N}_{\text{dCC}}\mathbf{N}_{\text{dCC}}^T\boldsymbol{\delta}_\mathbf{q}$. The dCCs increase the probability of finding a new, clash-free conformation close to the desired search direction $\boldsymbol{\delta}_\mathbf{q}$ but give no guarantee as it is a linearized procedure. Note that a proposed perturbation can introduce additional clashes at different sites. We therefore perform up to k iterations , each with the same dsired gradient, but each time introducing additional clash-avoiding constraints where necessary. This perturbation is denoted NIK_k, and the default setting of k is 5. The ultra-high dimensions of molecular conformation spaces allow for a large number of dCCs while avoiding complete

rigidification of the molecule. Traditional, lower-dimensional linkages would suffer almost immediate immobility. Finally, note that dCC cannot be used in combination with AIK moves. Perturbing AIK analytical solutions by dCC would no longer guarantee closure of the sub-chain.

Planner

The original planner for KGS[12] resembles the RRT planner[8] but has been adapted to high-dimensional conformation spaces where it is not possible to generate a random conformation in the region of interest (Algorithm 1). The input is an initial conformation $\mathbf{q}_{\text{init}}$, an exploration radius R (in Å RMSD), and a number of desired iterations I. The exploration radius is subdivided into 100 spherical bins. Like an RRT, a seed is selected by first generating a random conformation, $\mathbf{q}_{\text{rand}}$. However, for molecular linkages $\mathbf{q}_{\text{rand}}$ is not necessarily a feasible conformation, and is typically located far outside the exploration region. To prevent the planner from only selecting seeds near the border of the exploration region, we randomly select a spherical bin centered on $\mathbf{q}_{\text{init}}$, and let $\mathbf{q}_{\text{seed}}$ be the conformation RMSD-closest to $\mathbf{q}_{\text{rand}}$. Furthermore, to avoid only stepping away from the initial, perturbations do not go toward $\mathbf{q}_{\text{rand}}$ but rather in a random direction. This procedure is denoted *binned RRT* in the following.

The binned RRT planner tends to oversample regions near the initial conformation, while simultaneously limiting fast exploration of unknown territory. We therefore introduce a multi-query *Poisson planner* (Algorithm 2), which is based on Poisson disk sampling[34]. The planner is initialized with a minimum Poisson disk radius r and an initial conformation $\mathbf{q}_{\text{init}}$ which is added to the open set. At each iteration we randomly select a seed conformation $\mathbf{q}_{\text{seed}}$ from the set of open conformations (Fig. 2.a) and attempt P perturbations such that the new conformation $\mathbf{q}_{\text{new}}$ lies within the Poisson disk with inner radius r and outer radius $2r$, i.e., the distance from $\mathbf{q}_{\text{new}}$ to $\mathbf{q}_{\text{seed}}$ is between r and $2r$. A perturbation attempt is successful if the resulting conformation is non-clashing and at least a distance r from any existing conformation. Finally, $\mathbf{q}_{\text{seed}}$ is moved to the set of closed conformations. If there are no more conformations in the open set the procedure ends.

The critical step is fast and efficient determination of all nearby conformations. However,

analyses of existing Poisson sampling algorithms[35,36] tend to ignore the dimensionality, n, of conformations, so their asymptotic behavior hides an exponential growth in n. The dimension of molecular conformation spaces easily exceeds $n = 50$, which, as yet, has limited applicability of Poisson sampling to ultra-high dimensional problems. For example, spatial hashing algorithms to identify neighboring samples[35] would require checking $3^n \approx 7 \cdot 10^{23}$ adjacent bins. This is clearly infeasible.

To address the extremely high dimensionality we introduce a bounding volume hierarchy (BVH) algorithm, called BVHCollect,inspired by collision detection in protein structures[27,37]. Each conformation $\mathbf{q}$ is associated with a reference to its parent seed, all conformations it serves as seed (its descendants), and the radius $R^B(\mathbf{q})$ of the sphere containing all descendants of $\mathbf{q}$ (Fig. 2.c). When a seed is selected, nearby nodes are located by traversing the BVH in a depth-first-manner starting at the root ($\mathbf{q}_{\text{init}}$). A visited node, $\mathbf{q}_{\text{c}}$, and all its descendants can be pruned from the traversal if they are sufficiently far from $\mathbf{q}_{\text{seed}}$ such that a collision with a new conformation is impossible. Considering all existing descendants of $\mathbf{q}_{\text{c}}$ within $R^B(\mathbf{q}_{\text{c}})$, the maximum allowed distance of $2r$ between $\mathbf{q}_{\text{new}}$ and $\mathbf{q}_{\text{seed}}$ plus the required empty inner disk with radius r around $\mathbf{q}_{\text{new}}$ requires

$$|\mathbf{q}_{\text{c}}\mathbf{q}_{\text{seed}}| > 2r + r + R^B(\mathbf{q}_{\text{c}}) \tag{5}$$

to exclude $\mathbf{q}_{\text{c}}$ and all its descendants from collision detection with $\mathbf{q}_{\text{new}}$. After perturbations have been generated, bounding volumes with associated R^B-values on the path from $\mathbf{q}_{\text{seed}}$ up to $\mathbf{q}_{\text{init}}$ need to be updated to reflect the new descendants in their subtrees.

Depending on the shape of conformational space the tree can be arbitrarily unbalanced, so BVHCollect takes linear time worst-case and consequently PoissonExplore takes $\mathcal{O}(N^2)$-time worst-case, where N is the total number of conformations generated. For real conformational spaces, however, a balanced tree is expected, which yields $\mathcal{O}(N \log_b N)$ complexity, where b is the branching factor of nodes in the tree. The traditional Poisson planner with spatial hashing takes $\mathcal{O}(N \cdot 3^n)$ time where n is the dimensionality of the problem. This approach becomes infeasible for $n \geq 6$ Clearly, the additional book-keeping via BVH pays off especially for large n, where existing algorithms using standard spatial hashing become prohibitively expensive.

RESULTS

We evaluated the performance of our algorithms on two proteins and an RNA with different sizes and characteristics. *Escherichia coli* dihydrofolate reductase (PDB ID 3QL3, DHFR in the remainder) is a 755 degree of freedom enzyme, which is essential to nucleic acid synthesis[38]. In some experiments we separately consider DHFR's *M20* (residues 14-25) and *FG* (residues 116-128) loops. The loops share the same underlying molecular graph and initial structure as the full enzyme, but while all torsional DOFs are active in full DHFR, only loop DOFs are activated when we consider loops. The *Pseudoknot* is the topologically knotted T-arm non-coding RNA of *turnip yellow mosaic virus* (PDB ID 1A60)[39]. Finally, protein $G\alpha s$ represents the alpha subunit of heterotrimeric G-protein (PDB ID 1AZT) which is an intracellular switch activated by extracellular stimuli through G-protein coupled receptors (GPCRs)[40] Unless otherwise stated, PoissonExplore uses only the NIK_5 perturbation and BVHCollect to query for neighboring conformations. In the following we evaluate the improvements for each of our contributions separately and the last subsection demonstrates their practical applicability.

Acceptance Rate Increases with Dynamic Clash Constraints

To illustrate that the dCCs used in NIK_5 result in higher acceptance rates, we carried out simulations with PoissonExplore with NIK_0 (no dCC) and NIK_5 perturbations on the M20 loop and our three full systems. Since acceptance rates could be affected by the choice of planner we performed the same tests on a random walk strategy that rejects clashing structures. This strategy resembles frequently used Monte Carlo sampling methods and is therefore labeled MCl.

With NIK_0, the clash rejection rate of PoissonExplore is independent of system size and extremely high, around 80-90% (Table 1). For all test systems, using NIK_5 dramatically reduces conformation rejection rates due to clashes. For the M20 loop, the rejection rate is reduced by an order of magnitude. The reductions are substantial for larger molecular structures, reducing rejections by a factor of two in the case of DHFR, but diminish as the size increases owing to cascading collisions. For larger systems, resolving a collision

with NIK_5 is more likely to introduce new clashes owing to large, densely packed protein cores. Introducing arbitrarily many collision constraints would rigidify large portions of the molecule, and is not likely to result in higher acceptance rates.

The computational cost of NIK_5 is higher than NIK_0 as there are more SVD computations to perform. The number of accepted conformations per time-unit for NIK_5 is half that of NIK_0. However, computational cost can be further reduced by making use of iterative SVD computations for successively added clash constraints.

A major strength of the dCC procedure is that it opens up regions of conformational space that are otherwise difficult to access. Unlocking such sterically constrained regions often requires coordinated motion of the linkage. In turn, such long-range collective motions can give insight into molecular mechanisms of conformational change (see Section "Correlated Motions in DHFR Active Site").

Finally, we observe that naive sampling using a random walk strategy (MCl) results in extremely high rejection rates regardless of dCC, making a strong case for motion planning-based strategies.

Bounding Volume Hierarchy Speeds Up Neighbor Search

To test the efficiency of CollectBVH we recorded the number of distance computations as the planner explored the native state of the pseudoknot molecule.

Fig. 4.a shows the number of distance computations performed by BVHCollect and the number performed by a linear search through all existing conformations. As expected, the linear search subroutine makes PoissonExplore perform a quadratic number of total distance computations. For the first few hundred conformations, BVHCollect results in about half the number of distance computations, but after about 400 conformations the evolution of distance computations takes on a nearly linear trend. This linear-time behavior of PoissonExplore is a best-case performance and can not be expected for general problems. One explanation might be that the pseudoknot structure is particularly elongated and flexible around the middle (see Fig. 4.b). PoissonExplore will close conformations near the initial conformation and then explore in separate directions which permits BVHCollect to efficiently prune large branches in distant regions.

Comparison With RCD+

To test the sampling quality of PoissonExplore, we generated conformations of the M20 loop in DHFR. This functionally important loop is well-characterized, and adopts three distinct conformations during the catalytic cycle: closed, occluded, and open[41]. We generated 1000 conformations, starting at the 'closed' conformation, and evaluated the distribution of all-atom RMSD distances to each of these three conformations. We then compared our conformational ensembles to 1000 structures generated using the state-of-the-art kinematics-based sampler, random coordinate descent (RCD+,[42]). RCD+ mimics cyclic coordinate descent but selects bonds for optimization randomly, and updates loop conformations by spinor-matrices and geometric filters.

When we examined the effect of the Poisson disk radius on exhaustive exploration of ultra-high dimensional conformation spaces, i.e., close all open conformations, we found that PoissonExplore is sensitive to finely calibrating the Poisson disk size. If the radius is too large, $\mathbf{q}_{\text{init}}$ is immediately closed without opening up any new conformations and if it is too small the search will proceed indefinitely. To ensure broad sampling, we therefore adjusted r until it was just below the point where the search gets exhausted,

Both methods sample a broad ensemble of states (Fig. 5). For all three cases PoissonExplore samples closer to the experimentally observed conformation than RCD+, but for the closed and occluded states the middle 50th percentile of PoissonExplore conformations extend to a broader RMSD range than the comparable RCD+ ensemble. We anticipate this behavior as the Poisson disks result in a dense set of approximately equidistant conformations. RCD+ conformations extend to an *overall* larger RMSD range for all three states, but at the expense of structural quality of the samples. Among the 1,000 RCD+ loop conformations, 120 have collisions even when van der Waals radii are scaled by 75%, i.e., they would have been rejected by our method. Nonetheless, we expect PoissonExplore to diffuse more slowly through conformation space than a random sampler like RCD+. A larger number of conformations could possibly mitigate this drawback.

Correlated Motions in DHFR Active Site

Next, we examine if the kinematic cycles defined by native and non-native contacts propagate collective motions in proteins. The functionally important FG loop (residues 116-128) in DHFR connects the F and G β-strands. A 'distal' amino acid mutation G121V in this loop reduces catalytic activity of the enzyme 200-fold, despite the fact that the residue is nearly 15Å from the active site. Nuclear Magnetic Resonance (NMR) spectroscopy data furthermore suggests that the FG loop and the M20 loop are dynamically linked[43,44]. There is only one hydrogen bond between the loops (D122/H to G15/O). We generated 1,000 conformations for the FG and M20 loops simultaneously, by proposing trial moves $\mathbf{\Delta}_q$ for the rotatable bonds in both loops and projecting $\mathbf{\Delta}_q$ onto the corresponding nullspace.

We analyzed collective motions of the two loops by computing the correlations between the positions of their C_α atoms over all conformations. The correlation between atoms i and j is characterized by the quantity[45]

$$C_{i,j} = \frac{\langle \Delta\mathbf{p}_i(q) \cdot \Delta\mathbf{p}_j(q) \rangle_q}{\sqrt{\langle |\Delta\mathbf{p}_i(q)|^2 \rangle_q \cdot \langle |\Delta\mathbf{p}_j(q)|^2 \rangle_q}} \tag{6}$$

where $\langle \rangle_q$ denotes the average over all conformations, $\mathbf{p}_i(q)$ is the position of atom i in conformation q, and $\Delta\mathbf{p}_i(q) = \mathbf{p}_i(q) - \langle \mathbf{p}_i(q) \rangle_q$. Note that the normalization term in the denominator results in downscaling of $C_{i,j}$ if there is little motion.

We observed a higher degree of self-correlation within the M20 loop than in the FG loop (Fig. 6). The first half of the M20 loop backbone shows slightly reduced correlations with the second half, suggesting somewhat independent motions. Proline P21 and tryptophan W22, a more rigid and a bulky side-chain, likely reduce the feasible region of conformational space for the second half. Correlations between the two halves of the FG loop are highly reduced, with the N-terminal end of the loop showing more motion. The reduced self-correlations of residues reflect smaller motions overall.

To verify these observations we carried out a 100ns molecular dynamics simulation using the CHARMM force field and explicit solvent. In this case, the entire molecule is free to move. After aligning the entire trajectory to the first frame, the correlations shown in the lower triangle matrix of Fig. 6 were computed. Unlike the KGS samples, these trajectories

show motion of the loop terminals, but the correlated motion between loops is maintained very accurately.

Interestingly, in both simulations the highest inter-loop correlations are seen between residue pairs that are not in direct contact, but communicate through steric collisions and the hydrogen bond network. The region around G121 of the FG loop clearly stands out, collectively moving with N18 and A19 of the M20 loop. A19 is hydrogen bonded to M16, which sterically interacts with G121. Thus, both native and non-native contacts propagate these collective motions. Chemical shifts from NMR experiments of a related DHFR^{G121V}:NADPH:MTX complex indicated that these FG loop residues respond dynamically to the G121V mutation[46].

Residues D127 and E17 stand out by their anti-correlated motions. We propose that these residues are the result of a somewhat asymmetric hinge motion around the last half of the FG loop driven by interactions with the M20 loop. The lack of correlations for residues 124-126 supports this claim; when E17 on top of the M20 loop moves upwards, the hinge turns and D127 moves downwards, and vice versa.

We also generated 1000 structures permitting full-chain flexibility which is compared with loop-only sampling in the supplementary Figure 1. The high inter-loop correlations are preserved, but some of the anti-correlated motions are reduced or disappear. The correlation equation depends on performing an alignment of all structures which might account for a higher degree of correlation.

CONCLUSIONS

The ultra-high dimensions of molecular conformation spaces require efficient methods to broadly and uniformly sample conformations and relate these conformational sets to biological function. We combine different molecular perturbations from analytical inverse kinematics solutions, nullspace projections and our new, dynamic clash-avoiding constraint strategy, to approach a comprehensive representation of the underlying constraint variety. Our four examples consistently illustrate that this strategy increases acceptance rates in highly rugged search spaces. Coupled to an adapted Poisson disk inspired motion plan-

ner with bounding volume hierarchy (PoissonExplore), our algorithm efficiently samples distinct conformations in the ultra-high conformation space close to experimentally known functional states. Compared to RCD+ on DHFR's M20 loop, we obtain broad conformational ensembles around the known closed, occluded, and open state.

Finally, we analyzed how functional motions in DHFR are coupled through native and non-native contacts between the M20 and the FG-loop. Surprisingly, correlation analysis reveals coordinated motions in both loops linked to residues that are not direct neighbors. This reveals how motions encoded by our clash-avoiding constraints are propagated over long distances, revealing underlying functional mechanisms in the molecule. Our method is generally applicable to ultra-high dimensional problems in robotics and related research fields, subject to holonomic constraints and (self-) collisions. Software, documentation, and examples are available at `https://github.com/ExcitedStates/KGS`.

ACKNOWLEDGMENTS

D.B. is funded by the Deutsche Telekom Stiftung. R.F. is funded by grant NNF15OC0015268 from the Novo Nordisk Foundation and the Stanford Bio-X Program.

References

[1] H. van den Bedem and J. S. Fraser, Nat Meth **12**, 307 (2015), ISSN 1548-7091.

[2] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw, Annu Rev Biophys **41**, 429 (2012).

[3] I. Al-Bluwi, T. Siméon, and J. Cortés, Comput Sci Rev **6**, 125 (2012), ISSN 15740137.

[4] B. Gipson, D. Hsu, L. E. Kavraki, and J.-C. Latombe, Annual review of analytical chemistry **5**, 273 (2012).

[5] L. E. Kavraki, P. Švestka, J.-C. Latombe, and M. H. Overmars, IEEE trans on Rob and Autom **12**, 566 (1996).

[6] S. Thomas, X. Tang, L. Tapia, and N. M. Amato, J Comput Biol **14**, 839 (2007).

[7] K. Molloy, R. Clausen, and A. Shehu, Robotica pp. 1–29 (2014).

[8] S. M. Lavalle, J. J. Kuffner, and Jr., pp. 293–308 (2000).

[9] J. Cortés, T. Siméon, V. R. De Angulo, D. Guieysse, M. Remaud-Siméon, and V. Tran, Bioinf **21**, i116 (2005).

[10] S. Kirillova, J. Cortés, A. Stefaniu, and T. Siméon, Proteins: Struct, Funct, Bioinf **70**, 131 (2008).

[11] F. Noé, D. Krachtus, J. C. Smith, and S. Fischer, J Chem Theory Comput **2**, 840 (2006).

[12] P. Yao, L. Zhang, and J.-C. Latombe, Proteins: Struct, Funct, Bioinf **80**, 25 (2012).

[13] R. Clausen and A. Shehu, J Comp Biol **22**, 844 (2015).

[14] C.-A. Roth, T. Dreyfus, C. H. Robert, and F. Cazals, J Comp Chem **37**, 739 (2016).

[15] J. W. Burdick, in *Proc IEEE Int Conf Robot Autom* (IEEE Comput Soc Press, 1989), pp. 264–270.

[16] H. van den Bedem, I. Lotan, J. C. Latombe, and A. M. Deacon, Acta Cryst **D61**, 2 (2005), ISSN 0907-4449.

[17] D. Budday, S. Leyendecker, and H. van den Bedem, J Mech Phys Solids **83**, 36 (2015).

[18] M. Oliveberg and P. G. Wolynes, Q Rev Biophys **38**, 245 (2005).

[19] N. Haspel, M. Moll, M. L. Baker, W. Chiu, and L. E. Kavraki, BMC Struct Biol **10**, S1 (2010).

[20] I. Al-Bluwi, M. Vaisset, T. Siméon, and J. Cortés, BMC Struct Biol **13**, 1 (2013).

[21] B. Raveh, A. Enosh, O. Schueler-Furman, and D. Halperin, PLoS Comput Biol **5**, e1000295 (2009).

[22] H. Sugiura, M. Gienger, H. Janssen, and C. Goerick, in *IEEE Int Conf Intell Rob Syst* (2007), pp. 2053–2058, ISBN 1424409128.

[23] T. Petrič and L. Žlajpah, Rob Aut Syst **61**, 948 (2013), ISSN 09218890.

[24] A. Lagae and P. Dutré, in *Comp Graph Forum* (Wiley Online Library, 2008), vol. 27, pp. 114–129.

[25] C. Park, J. Pan, and D. Manocha, in *IEEE Int Conf Rob and Aut* (IEEE, 2014), pp. 4667–4673.

[26] V. de Angulo, T. Siméon, and J. Cortés, in *Proc of Rob: Science and Systems* (2005).

[27] I. Lotan, F. Schwarzer, D. Halperin, and J.-C. Latombe, J Comput Biol **11**, 902 (2004).

[28] R. Fonseca, D. V. Pachov, J. Bernauer, and H. van den Bedem, Nucl Acids Res **42**, 9562 (2014).

[29] B. I. Dahiyat, D. Benjamin Gordon, and S. L. Mayo, Protein Sci **6**, 1333 (1997).

[30] D. Halperin and M. H. Overmars, in *Proc Tenth Ann Symp Comp Geom* (ACM, 1994), pp. 113–122.

[31] E. A. Coutsias, C. Seok, M. J. Wester, and K. A. Dill, Int J Quant Chem **106**, 176 (2006).

[32] G. H. Golub and C. F. Van Loan, *Matrix computations*, vol. 3 (JHU Press, 2012).

[33] R. Fonseca, H. van den Bedem, and J. Bernauer, J Comput Biol **23**, 362 (2016).

[34] R. L. Cook, ACM Trans on Graph **5**, 51 (1986).

[35] R. Bridson, in *SIGGRAPH sketches* (2007), p. 22.

[36] D. Dunbar and G. Humphreys, ACM Trans on Graph **25**, 503 (2006).

[37] R. Fonseca and P. Winter, J Comp Biol **19**, 1203 (2012).

[38] G. Bhabha, J. Lee, D. C. Ekiert, J. Gam, I. A. Wilson, H. J. Dyson, S. J. Benkovic, and P. E. Wright, Science **332**, 234 (2011).

[39] M. H. Kolk, M. van der Graaf, S. S. Wijmenga, C. W. Pleij, H. A. Heus, and C. W. Hilbers, Science **280**, 434 (1998).

[40] D. V. Pachov, R. Fonseca, D. Arnol, J. Bernauer, and H. van den Bedem, J Chem Theory Comp **12**, 946 (2016).

[41] M. R. Sawaya and J. Kraut, Biochem **36**, 586 (1997).

[42] P. Chys and P. Chacón, J Chem Theory Comput **9**, 1821 (2013).

[43] D. D. Boehr, J. R. Schnell, D. McElheny, S.-H. Bae, B. M. Duggan, S. J. Benkovic, H. J. Dyson, and P. E. Wright, Biochem **52**, 4605 (2013).

[44] H. van den Bedem, G. Bhabha, K. Yang, P. E. Wright, and J. S. Fraser, Nat Meth **10**, 896 (2013).

[45] H. Kamberaj and A. van der Vaart, Biophys J **96**, 1307 (2009).

[46] M. R. V, S. P. J, P. C. M, and L. A. L, PLoS ONE **7** (2012).

[47] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, Acta Crystallographica Section D: Biological Crystallography **66**, 12 (2010).

Figure 1: Kinematic representation of molecules and constraints. a) A molecular graph with double bonds and partial double bonds highlighted. After edge-contractions these atoms become a single rigid body vertex. The remaining bonds can rotate around the bond axis and are described by dihedral angles. b) Small section of a protein molecule with hydrogen bonds marked as purple lines and a nitrogen-hydrogen clash marked with van der Waals spheres and the bisecting plane. c) A dynamic clash-avoiding constraint (dCC) allows individual motions of the near-clashing atoms $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ along directions $\boldsymbol{t}_c$ in the plane with normal vector $\boldsymbol{n}_c$, but only a joint move along $\boldsymbol{n}_c$. This allows atoms to slide past each other, but prevents them from getting closer.

Figure 2: A two-dimensional representation of a high-dimensional Poisson planner. Closed circles denote previously obtained conformations, open circles are proposed conformations. Pink areas are sterically excluded regions of conformation space. a) A random unexplored seed conformation is selected and a Poisson disk with inner (outer) radius r ($2r$) placed around it. b) A zoomed view of panel a for a subsequent sampling step. An admissible perturbation within the Poisson disk, $\Delta_q(\mathbf{N}_q)$, is generated by projecting a random perturbation onto the nullspace of $\mathbf{q}_{\text{seed}}$. If a collision occurs we add a dynamic clash constraint, resulting in a new, reduced nullspace $\mathbf{N}_q^{\text{dCC}}$ with a projection in which the clash is resolved. c) To avoid oversampling, perturbations are only accepted if they are further than r from any existing conformation. Using a BVH, we maintain the largest distance from $\mathbf{q}_{\text{c}}$ to any of its descendants, $R^B(\mathbf{q}_c)$. Distance computations between a perturbation of $\mathbf{q}_{\text{seed}}$ to all descendants of $\mathbf{q}_{\text{c}}$ can be ignored if $|\mathbf{q}_{\text{c}}\mathbf{q}_{\text{seed}}| > R^B(\mathbf{q}_c) + (2r + r)$, i.e. the purple sphere does not collide with the blue sphere.

Figure 3: Pseudocodes for the RRT-like planner and the Poisson planner. Both algorithms take an initial conformation from which the search is started. Algorithm 1 additionally takes the argument R as the exploration radius around the initial, σ the step size of the perturbation, and I the number of iterations. Algorithm 2 takes the argument r as the inner radius of the Poisson disk and P the number of random perturbations that are attempted before closing a seed conformation. The PERTURB function performs either an AIK or NIK move with equal probability.

Figure 4: Computational complexity of BVH. a) A subset of generated DHFR and pseudoknot structures. b) Number of distance computations as a function of number of samples of DHFR and the pseudoknot using the BVH versus linearly checking all conformations. A sub-quadratic behavior is observed for both tested systems when the BVH is enabled.

Figure 5: Accuracy of loop sampling. Using PoissonExplore with AIK and NIK_5 as well as RCD+ we sampled 1,000 conformations of the M20 loop (residue 14-25) and measured their backbone heavy-atom RMSD distances to 3 loop conformations known to represent a closed, an occluded, and an open conformation. PoissonExplore was initialized near the closed conformation while RCD+ completely rebuilds loops in each iteration. The middle 50th percentile of each distribution is represented using green and blue box-plots. The three different states are shown below with the lowest backbone heavy-atom RMSD structure from KGS (green) and RCD+ (blue) respectively.

Figure 6: Correlated loop motions in DHFR. Both the M20 loop and the FG loop of DHFR were sampled using the Poisson planner with dCC (upper right triangle) and a molecular dynamics simulation (lower left triangle). The correlation between C_α atoms in each loop is plotted. Red colors indicate high correlation ($C_{i,j}$) and blue colors anti-correlation (negative $C_{i,j}$). Inter-loop pairs with particularly high correlation are highlighted in the structural models on the right. Several of the inter-loop correlations are spatially non-local. For example, the crystal structure distance is 9.6Å between CA-atoms in residues 19 and 121 (high inter-loop correlation) and 20.6Å between residues 17 and 127 (highest anti-correlation).

Figure 2: Supp Mat. Correlated loop motions in full chain DHFR. Comparison of correlated motions in the M20 loop and the FG loop of DHFR when sampling loops only (upper triangle) versus the full-chain (lower triangle). The correlation between C_α atoms in each loop is plotted. Red colors indicate high correlation ($C_{i,j}$) and blue colors anti-correlation (negative $C_{i,j}$).

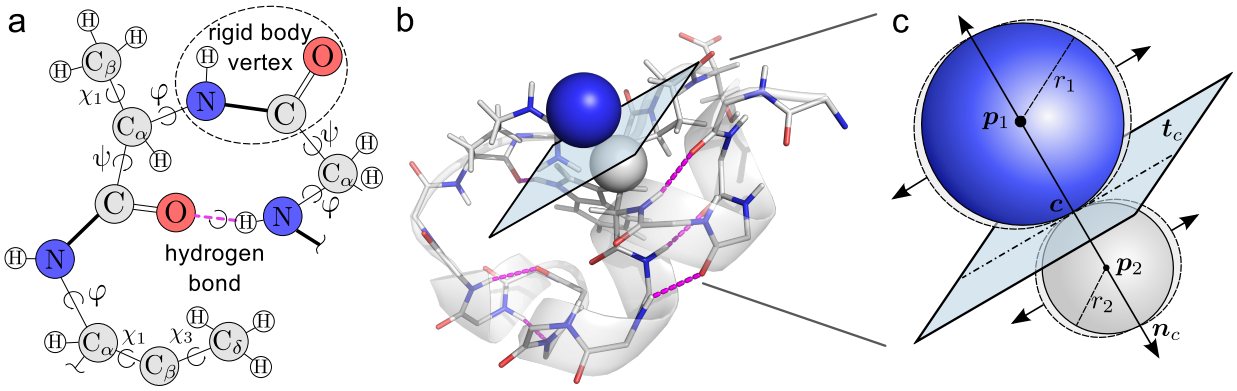


Figure 1
Rasmus Fonseca, Dominik Budday, Henry van den Bedem
J. Comput. Chem.

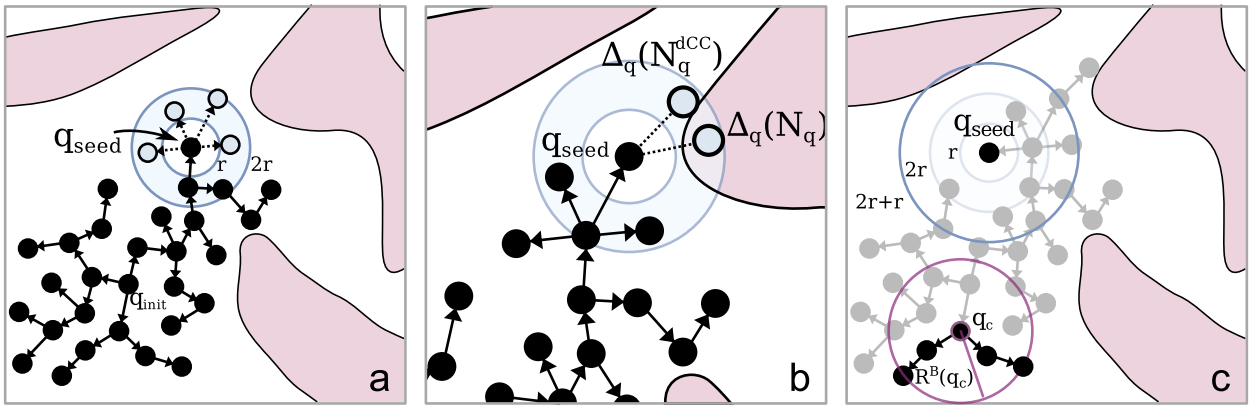


Figure 2
Rasmus Fonseca, Dominik Budday, Henry van den Bedem
J. Comput. Chem.

Alg. 1 BINNEDRRT($\mathbf{q}_{\text{init}}, R, \sigma, I$)

$B \leftarrow$ array of 101 empty bins
$B[0]$.add($\mathbf{q}_{\text{init}}$)
for $i = 0$ **to** I **do**
  $\mathbf{q}_{\text{rand}} \leftarrow$ random conformation
  repeat
    $b_{\text{rand}} \leftarrow \text{RAND}(0, 100)$
  until $B[b_{\text{rand}}] \neq \emptyset$
  $\mathbf{q}_{\text{seed}} \leftarrow \arg\min_{\mathbf{q} \in B[b_{\text{rand}}]} |\mathbf{q}_{\text{rand}}\mathbf{q}|$
  $\mathbf{q}_{\text{new}} \leftarrow \text{PERTURB}(\mathbf{q}_{\text{seed}}, \sigma)$
  if CLASHFREE($\mathbf{q}_{\text{new}}$) **then**
    $b_{\text{new}} \leftarrow \lfloor |\mathbf{q}_{\text{init}}\mathbf{q}_{\text{new}}| \cdot \frac{100}{R} \rfloor$
    $B[b_{\text{new}}]$.add($\mathbf{q}_{\text{new}}$)
  end if
end for
return $\bigcup_{b=0}^{100} B[b]$

Alg. 2 POISSONEXPLORE($\mathbf{q}_{\text{init}}, r, P$)

$S_{\text{open}} \leftarrow \{\mathbf{q}_{\text{init}}\}$
$S_{\text{closed}} \leftarrow \{\}$
while $S_{\text{open}} \neq \emptyset$ **do**
  $\mathbf{q}_{\text{seed}} \leftarrow S_{\text{open}}$.Pop()
  $S' \leftarrow$ BVHCOLLECT($\mathbf{q}_{\text{seed}}$)
  for $p = 0$ **to** P **do**
    $\mathbf{q}_{\text{new}} \leftarrow \text{PERTURB}(\mathbf{q}_{\text{seed}}, \frac{r+2r}{2})$
    if CLASHFREE($\mathbf{q}_{\text{new}}$) $\wedge$ $\forall \mathbf{q} \in S'. |\mathbf{q}_{\text{new}}\mathbf{q}| > r$ **then**
      S_{open}.add($\mathbf{q}_{\text{new}}$)
    end if
  end for
  S_{closed}.add($\mathbf{q}_{\text{seed}}$)
end while
return S_{closed}

Figure 3
Rasmus Fonseca, Dominik Budday, Henry van den Bedem
J. Comput. Chem.

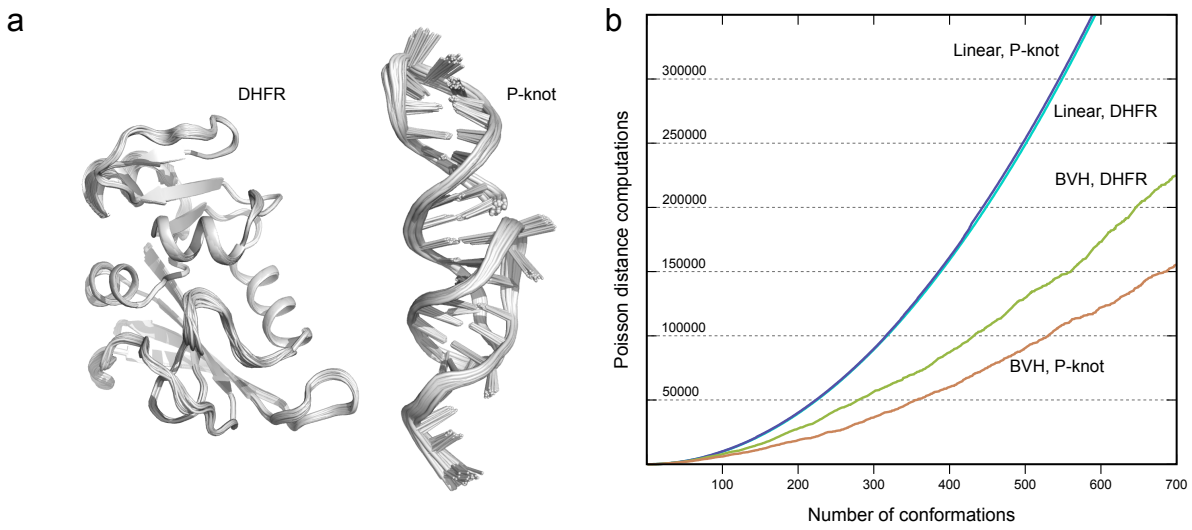


Figure 4
Rasmus Fonseca, Dominik Budday, Henry van den Bedem
J. Comput. Chem.

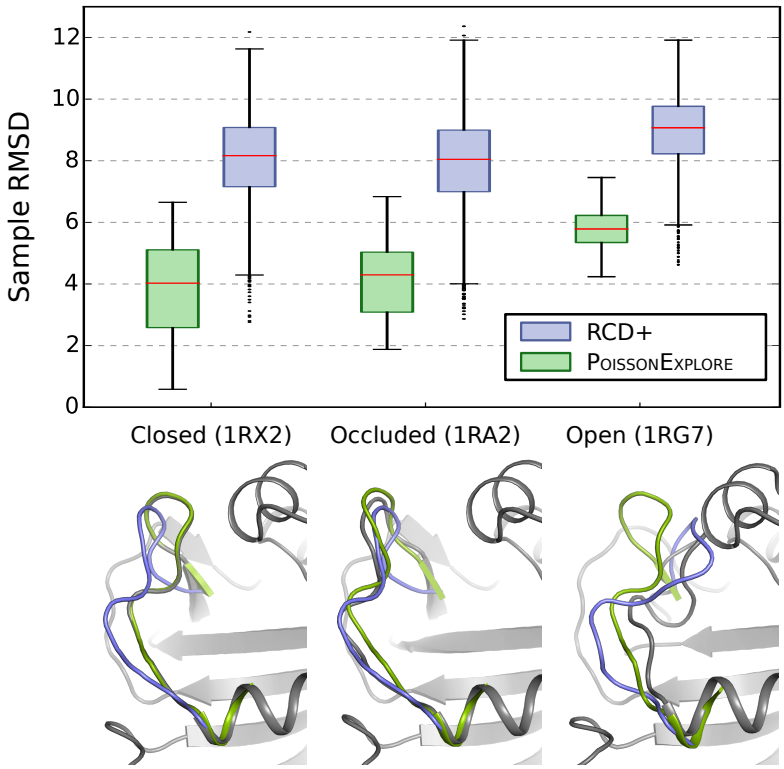


Figure 5
Rasmus Fonseca, Dominik Budday, Henry van den Bedem
J. Comput. Chem.

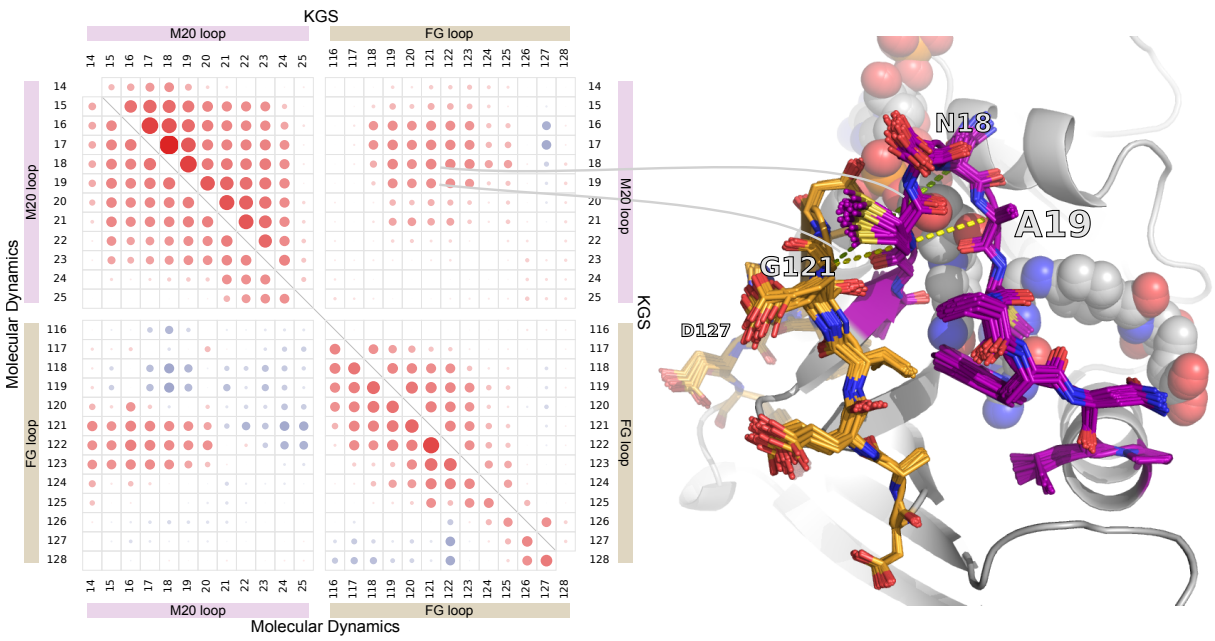
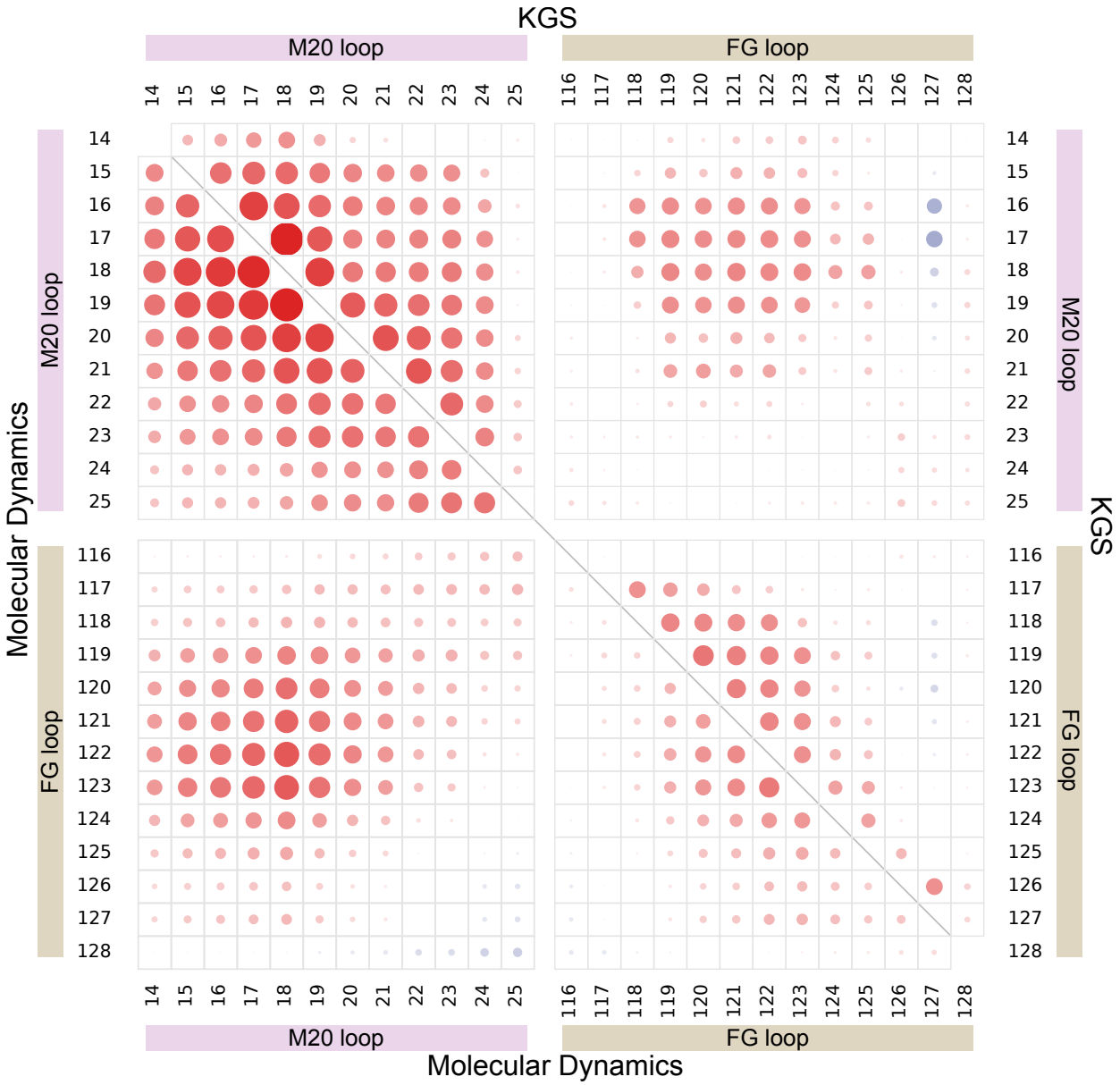


Figure 6
Rasmus Fonseca, Dominik Budday, Henry van den Bedem
J. Comput. Chem.



Supp. mat. Figure 1
Rasmus Fonseca, Dominik Budday, Henry van den Bedem
J. Comput. Chem.

		M20 loop	P-knot	DHFR	Gαs
Degrees of freedom		48	326	755	1713
PoissonExplore	Clash rate	4%	14%	44%	60%
	Disk reject rate	6%	1%	1%	1%
	Perturbation time	210ms	294ms	2479ms	
	Clash score	446 ± 1	68 ± 6	447 ± 2	
PoissonExplore (no dCC)	Clash rate	88%	81%	88%	79%
	Disk reject rate	4%	1%	2%	7%
	Perturbation time	166ms	58ms	85ms	
	Clash score	446 ± 1	73 ± 6	448 ± 2	
MCl	Clash rate	19%	23%	100%	100%
	Perturbation time	1194ms	230ms	15222ms	
	Clash score	449 ± 1	74 ± 9	449 ± 4	
MCl (no dCC)	Clash rate	91%	89%	100%	100%
	Perturbation time	1433ms	55ms	186ms	
	Clash score	452 ± 3	57 ± 7	450 ± 3	

Table 1: Comparison of conformation rejection rates using the PoissonExplore planner or Monte Carlo-like (MCl) sampling using NIK_5 and NIK_0 (no dCC) perturbations. For PoissonExplore we distinguish between rejections due to clashes (clash rate) and those rejected by failing to meet the Poisson disk criteria for new conformations (disk reject rate). Perturbation times indicate the average time it takes for each accepted conformation. The clash scores were computed using molprobity[47] and are generally different than those for the starting structures (DHFR: 0.8 and P-knot: 80.0).