Specific Aims

Ras family proteins, important in the control of cell growth, are commonly mutated in human cancer [1]. Furthermore, mutations in Ras are also a leading cause of resistance to modern targeted therapy, and patients who harbor Ras mutations have considerably poorer prognoses than those with wild type Ras [2]. Targeting Ras has proven difficult, however, as its oncogenic mutants activate Ras primarily by inactivating its enzymatic activity, leaving inhibition of the enzyme an unworkable strategy. Further, the high affinity of GTP for Ras, resulting in the active Ras-GTP complex, combined with its high intracellular concentration renders outcompeting the bound nucleotide extremely difficult [3]. While recent experimental approaches have seen some success in identifying allosteric modulators [3], these inhibitors are of limited application as they are currently of low affinity and therefore require the presence of a cysteine for covalent attachment [3], present in only a small fraction of cancers. Worse still, this need for a cysteine results in a straightforward path to drug resistance. Here, we propose a novel strategy to design noncovalent ligands that allosterically inhibit Ras, using computational strategies where structural biology cannot elucidate hidden conformers.

Aim 1: Computationally map conformations accessible to Ras along with their corresponding energetics to identify potential opportunities for allosteric modulation

We will first **identify metastable conformations and the associated energetics of oncogenic K-Ras using Markov State Model approaches**. Then, we intend to **mine the metastable conformations identified previously for potential ligand binding sites**. Following that analysis, we will **analyze the metastable conformations for putatively inactive conformations by comparison to other structures**. Finally, we will **develop fluorescence-based experimental assays to validate the conformational populations that we observed**.

Aim 2: Identify new small molecule ligands of Ras

We will **perform virtual screening on a library of commercially-available compounds against putative allosteric binding sites**. Following that, we intend to **filter hits using free energy calculations**. Finally, we will **develop and validate a fluorescence-based assay for measuring weak allosteric binding and experimentally measure quantitative binding affinities of promising hits**. Finally, **we will computationally optimize derivatives of weak-binding scaffolds**, including existing scaffolds and those discovered in Aim 2 **using a new chemical Monte Carlo algorithm based on the method of expanded ensembles to search over large spaces of possible derivative molecules.** We will then **experimentally validate our derivatized ligands** to more tightly bind K-Ras.

Conclusion

Having completed the above aims, our results will consist of novel chemical structures for targeting Ras, a novel computational pipeline for the discovery of allosteric sites and inhibitors, as well as insight into the conformational populations of a key signaling protein.

Innovation

In this proposal, we develop a novel strategy for the design of allosteric ligands, and apply this strategy to the "undruggable" oncoprotein K-Ras. Taking advantage of recent advances in massively-parallel molecular simulation, we propose to use the developing field of Markov State Modeling to enumerate metastable conformations of K-Ras, giving unprecedented insight into both its functional dynamics and opportunities for its inhibition. We then develop a novel technique for exploring the relevant chemical space in ligand design for identified binding sites, allowing us to use physical principles to examine the structure-activity relationship of ligand binding to these previously undiscovered sites. We also propose to experimentally validate these techniques, providing a guide for the future examination of Ras allosteric ligands experimentally.

Expanded-ensemble simulation to explore chemical space. We will use the technique of expanded-ensemble simulation [4] over chemical space, with an appropriate bias, to guide the simulation toward a tighter-binding derivative. This unprecedented use of molecular simulation to guide ligand enhancement is not only a technique to enhance new screening hits, but also provides insight into the quantitative structure-activity relationship (QSAR) of the ligands and the binding pocket, a key quantity in medicinal chemistry [5]. Importantly, this also allows us to ask questions about how different chemotypes affect binding of ligands, and whether traditional chemical motifs can be applied in the design of allosteric binders at sites where no natural ligand is known to bind.

Generation of Markov State Model (MSM) of K-Ras: While there exist a number of structures of K-Ras, experimental techniques cannot access the many metastable conformations that proteins naturally adopt in low populations. Molecular simulation can provide insight into this, though it is difficult to parallelize. Therefore, we will use the developing field of Markov State Modeling to aggregate many independent trajectories of Ras into a model of its equilibrium conformational populations [6]. This allows for the ability to discover binding sites otherwise hidden from experimental view, with the important feature of containing equilibrium weights. The MSM also allows us to better understand the way that Ras functions, and its transitions between active and inactive states.

Experimentally validate findings: We plan to use modern techniques such as Hydrogen-Deuterium Exchange Mass Spectrometry (HDX-MS) [7] to probe the conformational dynamics of Ras, and compare these experimental results with the calculated values from the MSM. The adaptation of this technique for K-Ras has several advantages. Among those are its high sensitivity, ability to detect multiple conformers, and very low requirement for protein [7].

Abstract

Ras, a family of small GTPases, is critical in pro-growth signaling and cancer survival. Ras proteins are activated after exchanging their nucleotide diphosphate with a nucleoside triphosphate, and become inactive after enzymatic activity hydrolyzes the gamma phosphate bond. In many cancers, Ras is mutated such that this hydrolysis is slow or nonexistent, rendering its pro-growth signal permanently on. It has proven difficult to target, however, as its mutation is a loss-of-function, and its bound nucleotide is difficult to outcompete. Here, we propose a computational strategy for the development of small-molecule allosteric modulators of Ras. First, we propose to use Markov State Models to generate a map of the metastable conformations of Ras and their populations, identifying those with potential ligand binding sites. Then, we propose to use virtual screening to identify potential scaffolds to bind and stabilize this site. Finally, we propose a novel application of expanded ensemble molecular simulation to computationally explore chemical space near the scaffold to improve its ligand binding affinity. Our computational method will be coupled to experimental techniques to ensure that our computation is accurately representing the system. Once complete, our proposal will result in a new avenue for the potential development of therapeutics targeting Ras family proteins.

Introduction and Background

The role of Ras family proteins in cancer

Ras, a family of small GTPases, occupies a central role in cell growth signaling, and, unsurprisingly, plays a central role in the progression of serious cancers [2]. In Figure 1, a Kaplan-Meier plot of time to recurrence in K-Ras mutant vs. wild-type or unknown genotype in non-small cell lung cancer, one can see that tumors harboring a mutant K-Ras genotype carries a considerably worse prognosis; by about 25 months from remission, nearly all patients have had a recurrence [2]. In Figure 1, one can see that while death due to other causes dominates for patients with wild-type or unknown K-Ras status, death due to disease dominates for those with mutant K-Ras. In addition to the lack of clinically-available K-Ras inhibitors, there is evidence that oncogenic K-Ras may assist in conferring the deadly metastatic phenotype to cancer cells via interference with cell-cell interactions and induction of a migratory phenotype [1]. Ras is able to achieve these ends in cancer via its numerous interactions with downstream effectors, such as the interactions shown below in Figure 2.
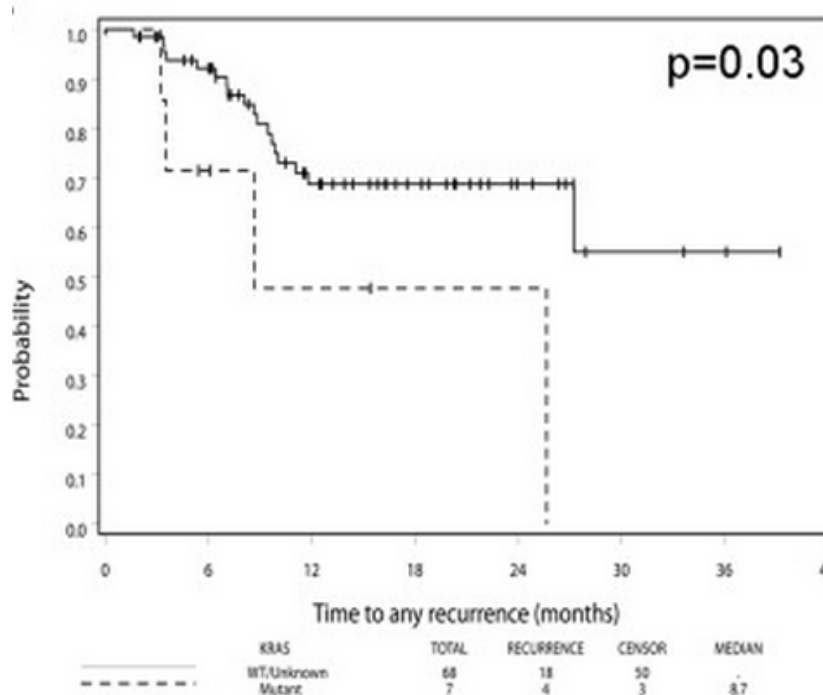

Figure 1: **Time to recurrence**. Dotted line, K-Ras mutant, Solid line, K-Ras unknown or wild-type. [2]

Ras functional biophysics

The Ras family proteins are small GTPases; that is, they are enzymes which can bind GTP and hydrolyze it to GDP, which remains bound until an exchange [1]. In its GTP-bound form, Ras is active, and engages downstream pro-growth signaling factors as seen in Figure 2. Ras is then inactivated by hydrolyzing its bound GTP; however, this process is greatly assisted by a family of proteins known as GTPase Activating Proteins, or GAPs [1].
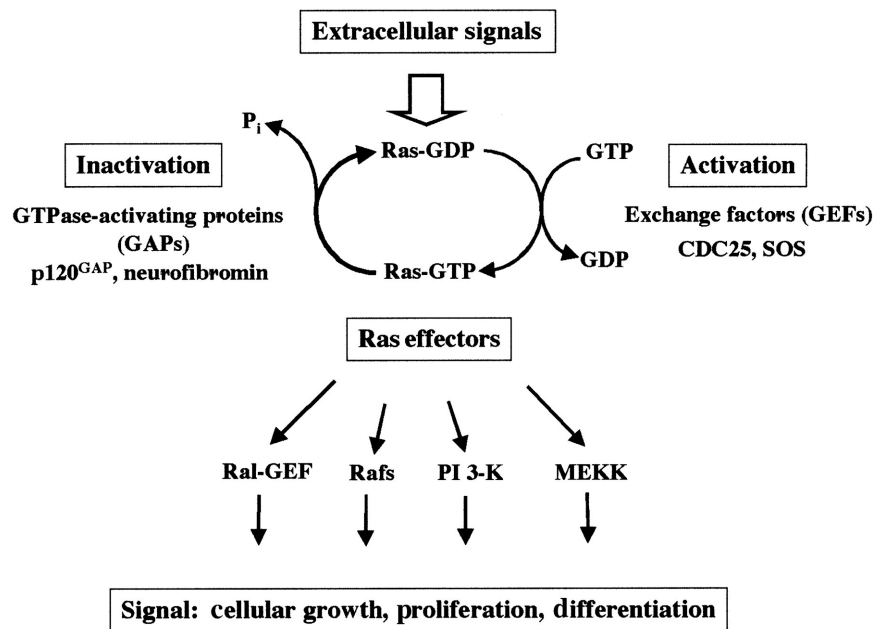
Figure 2: **A diagram of major Ras growth signaling partners, indicating its connectedness in growth signaling pathways** [8]

The oncogenic potential of K-Ras mutations relates to its central role in cellular growth signaling [1], as indicated by Figure 2. This position in relevant pathways makes Ras both a critical means of cancer proliferation and an attractive drug target. However, its biophysical nature causes great difficulty in designing an inhibitor; first, oncogenic mutations generally ablate the ability for Ras to hydrolyze GTP, as well as its ability to interact with partners that assist in hydrolysis and thus become inactive [1]. This loss-of-function mutation means that inhibition of the enzymatic activity is not a feasible avenue. Furthermore, outcompeting bound GTP is generally regarded as infeasible due to its picomolar affinity for GTP [3] , and the lack of any other apparent binding sites in crystal structures [3].

Approach

Aim 1: Computationally map conformations accessible to Ras along with their corresponding energetics to identify potential opportunities for allosteric modulation

Hypothesis

Ras family proteins can adopt multiple kinetically metastable states containing binding sites not present in the ground state, which can be observed using advanced molecular simulation techniques with the apo-protein.

Overview:

As discussed above, Ras is a key protein in the ability for cancer cells to grow and divide, but the loss-of-function nature of its mutations and lack of apparent druggable sites render it a difficult target. While no apparent viable sites for ligand binding are present in crystal structures, there is evidence that there exist lower-populated conformations [9] [3] which contain ligand binding sites, and may be signaling-inactive. However, the question remains of whether these sites, and potentially others, can be discovered in the absence of ligand. In order to answer this question, we will use parallel high-throughput molecular dynamics to generate trajectory data, and Markov State Modeling (MSM) [6] to combine the trajectory data into a conformational map, as conceptually illustrated in Figure 3.
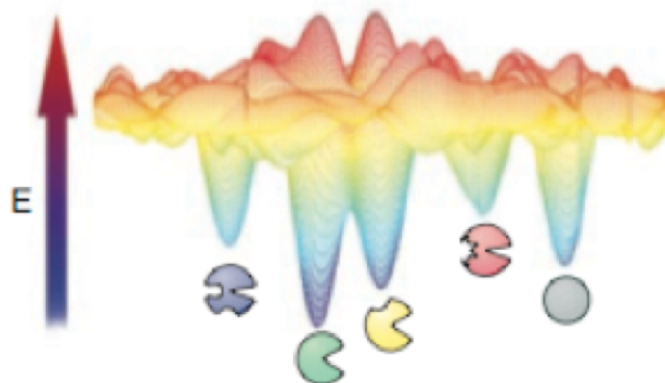
Figure 3: **A schematic of an energy landscape with multiple wells for metastable states** [10]

Aim 1.1: Identify metastable conformations and the associated energetics of oncogenic K-Ras using Markov State Model approaches

Here we propose to use parallel molecular dynamics simulations with OpenMM [11], and to combine the independent trajectories using Markov State Model approaches [6] to elucidate the equilibrium conformational populations of oncogenic K-Ras and their interconversion kinetics. We also propose to use techniques developed in our laboratory to accelerate sampling, described below. The completion of this aim will not only render useful conformations for cancer drug discovery, but will also demonstrate that molecular dynamics can discover allosteric sites not visible in experimental structures, even in the absence of a ligand to stabilize those structures.

Rationale: As we would like to determine whether the conformational populations of K-Ras in detail we will use the computational technique of atomistic molecular dynamics. In using atomistic molecular dynamics to test our hypothesis, we are presented with several options. It may seem straightforward to simply run dynamics near an interesting conformation until binding sites appear. However, this suffers from the serious limitation that it does not provide equilibrium population data for an identified conformation. A more rigorous approach would involve running a single trajectory of a Ras protein for a very long time; however, this approach suffers from the practical limitation that it is difficult to take advantage of parallel computing resources. Instead, we propose to utilize the developing field of Markov State Modeling, which enables us to run many trajectories in parallel and distill trajectories into a set of kinetically metastable states with [6], enabling us to achieve vastly greater sampling and take advantage of modern hardware such as Titan and Folding@Home.

Specific Experiments: In order to generate our conformational model via Markov State Modeling (MSM), we will first run many independent simulations of G12C-mutant apo K-Ras in explicit solvent, as this mutation is both clinically relevant and the subject of recent investigation [3]. We intend to repeat this protocol on other mutations and with GTP- and GDP-bound forms of K-Ras, allowing us to explore differences in equilibrium populations. To run the simulations, we will use Titan, a supercomputer at Oak Ridge National Laboratory and the second-fastest in the world. To enhance sampling of conformational space, we will use a pipeline developed in our laboratory that identifies protein structures in the protein data bank (PDB) by BLAST [12] sequence similarity, then constructs homology models of the target sequence using MODELLER [13], followed by a brief molecular dynamics refinement as shown in Figure 4.

Data Analysis and Anticipated Results: After trajectory data is collected, our first step in MSM creation will involve a dimensionality-reduction technique. This transforms input coordinates, such as distances or torsions, into a new, lower-dimensional, set of coordinates representing important (such as slowly-changing) features. Next, a technique such as Voronoi tessellation is used to assign the reduced-dimension trajectories to discrete states [6]. These now-discretized trajectories are used to

estimate the transition matrix between states [6]. By using a maximally-connected subset of states in this final step, we can avoid the inclusion of potentially problematic simulations of irrelevant conformational states [14]. The result of this process is a set of metastable conformations, along with transition probabilities between them, allowing us to examine the equilibrium conformational populations of K-Ras for ligand binding sites. We will verify that similar results are obtained using different subsets of the trajectory data, providing an indication that the conformational space of K-Ras has been sufficiently sampled. We expect that the conformations recovered in this process will include those identified in both inactive and active crystal structures, as well as a variety of intermediate states.
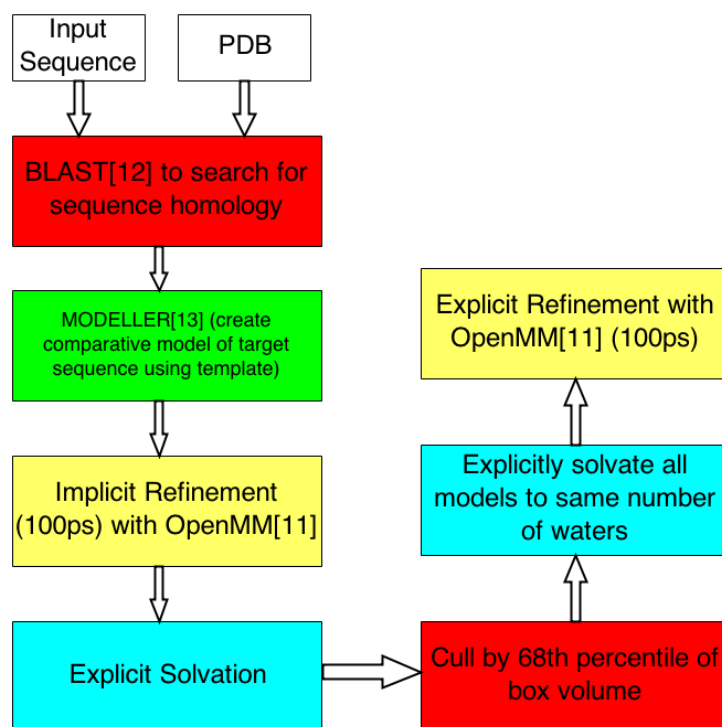

Figure 4: **A flowchart demonstrating the generation of starting conformations.**

Potential Problems and Alternative Strategies: While it is unlikely, it is possible that even given extensive sampling via supercomputers and from a diverse set of conformations, transitions between known relevant states of K-Ras will not be observed. In this event, we could further pursue techniques such as adaptive sampling [6] to generate an MSM including transitions between known relevant states.

Aim 1.2: Mine the Metastable Conformations for potential ligand binding sites

Rationale: By generating an MSM of K-Ras, we are now afforded unprecedented insight into its equilibrium conformational populations. To test our hypothesis that this population, discovered without the presence of ligand, contains putative ligand binding sites, we intend to use an algorithm such as LIGSITE [15]. Furthermore, in order to identify potentially useful binding sites for drug discovery, we intend to use equilibrium population data to limit our selection of useful conformations to those which have an energy difference from ground that can be stabilized with a small molecule. [16]

Specific Experiments and Data Analysis: With our computational approach, we are afforded the opportunity to discover novel ligand binding sites that were not visible in experimental structures. We intend to screen the identified conformations with an algorithm known as LIGSITE [15]. This technique searches for points nearly-surrounded by solvent-inaccessible regions [15]. We believe that this will be successful for two reasons. First, this, approach has been used before to discover known cryptic binding sites in simulations without ligand [17]. Second, the existence of weakly-binding molecules to allosteric

sites on K-Ras, such as the covalent inhibitors in [3] depicted in Figure 5, whose bound structure is illustrated in Figure 6. However, the molecules in [3] was discovered by a screening technique [18] that requires covalent attachment, and still relies on serendipity. Our approach can discover allosteric sites without first requiring an allosteric ligand, allowing for rational allosteric drug design. Furthermore, in order to confine our identification of conformations with ligand binding sites to those which can conceivably be stabilized by a ligand, we will choose metastable conformations within $10k_BT$ of the ground conformation [16].

Anticipated Results: We expect that the analysis of metastable conformations will result in the identification of conformations with ligand binding sites, including those already discovered, despite the absence of ligand in the simulation. This provides us with several important pieces of information. First, it would confirm our hypothesis that the protein adopts these allosteric conformations even in the absence of ligand, albeit at a lower rate. Second, the result yields a structure to use for the discovery of novel ligands, which is not afforded by current experimental techniques.

Potential Problems and Alternative Strategies: It is possible that the MSM that we generated fails to generate novel allosteric binding sites. In this case, we will proceed with the proposal using already-identified sites. It is also possible that the MSM fails to identify pre-existing sites, though this seems highly unlikely. In that event, we will explore means to increase sampling, as well as run simulations including known ligands to test if they are required for appearance of the ligand binding site.
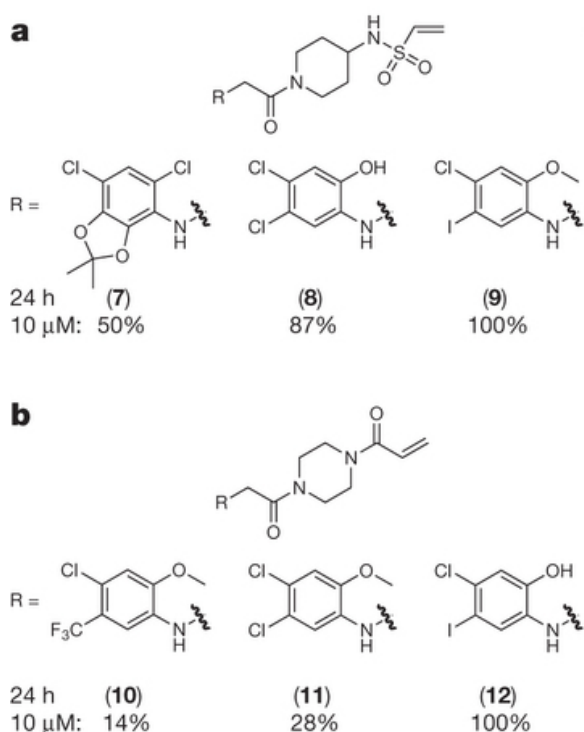

Figure 5: **Several molecules identified in [3] to inhibit K-Ras G12C.** Percentages denote covalent attachment after 24h in $10\mu M$ compound. Note the presence of an electrophile for covalent attachment to the required cysteine. Image from [3]
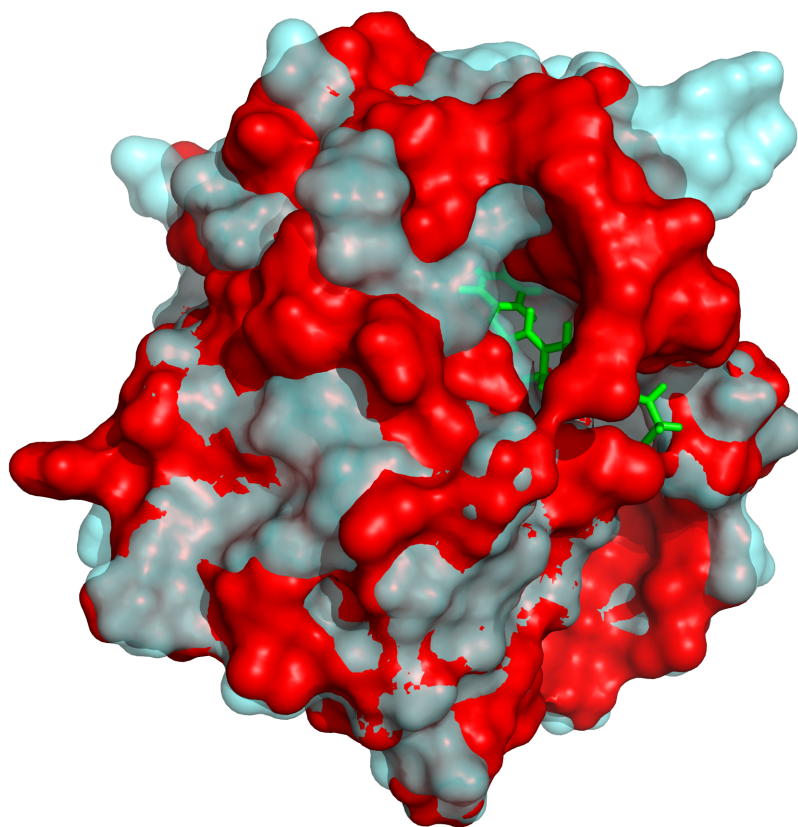
Figure 6: **K-Ras G12C (red surface) with bound inhibitor (green molecule), aligned to unbound structure (aquamarine surface)**. The ligand's presence reveals the presence of a binding pocket not visible in the unbound structure. [3]. PDB codes 4LV6 (bound) [3], 4OBE (unbound) [19].

Aim 1.3: Identify putative druggable signaling-inactive conformations

Rationale: Once we have identified metastable conformations, it is not only important to locate opportunities for ligand binding, but especially those states that may represent signaling-inactive states that could be preferentially stabilized by an allosteric ligand. The wealth of Ras structures bound to effectors as well as inactive will aid us in identifying conformations that may be unable to signal to downstream partners.

Specific Experiments and Data Analysis: Since the molecules in Figure 5 were shown to cause K-Ras to favor its guanosine diphosphate (GDP)-bound form instead of its active guanosine triphosphate (GTP)-bound form. [3], it is plausible for us to search the metastable conformations for those with minimum RMSD to GDP-bound inactive structures. We will also examine the placement of nucleotide-interacting residues in identified metastable conformations and compare them to those of known GDP- and GTP-bound Ras structures. Additionally, we will examine structures involving key interaction partners of Ras. These proteins include SOS, an upstream effector that encourages nucleotide exchange, Raf, a group of protein kinases, and phosphoinositide 3-kinase (PI3K) [1]. Comparison of available crystal structures of Ras-family proteins bound to these effectors with structures of inactive Ras can also lend insight into which conformational features are required for Ras activity [20], such as those indicated in Figure 7.
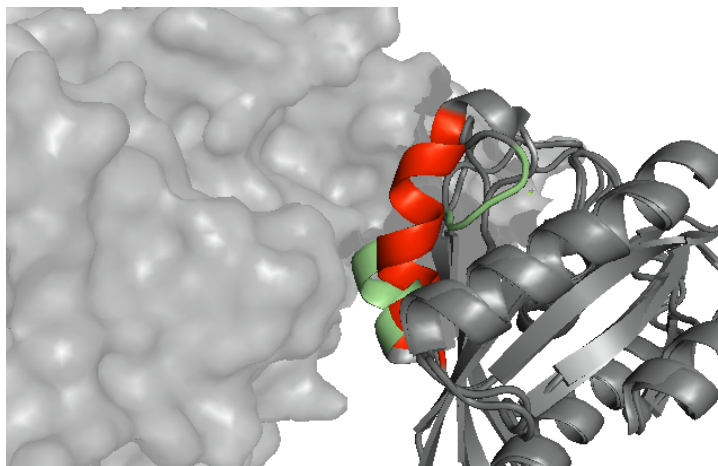
Figure 7: **Alignment of GDP and GTP bound conformations of H-Ras, showing Switch II effector region interacting with Phosphoinositide 3-kinase γ (PI3Kγ) [20].** Red and aquamarine show active and inactive conformations, respectively, with PI3Kγ surface. PDB codes: inactive 1Q21 [21], active 1HE8 [20]

Anticipated Results: We expect that our analysis of the metastable conformations of Ras will include conformations in both active and inactive states. Furthermore, we expect that we will identify conformations with potential ligand binding sites that have key residues, as shown in Figure 7, in the inactive positions. This discovery of previously unknown inactive states would represent a number of important steps forward. First, it would further the understanding of how Ras functions, revealing novel metastable states and generating new hypotheses about the interaction of Ras with its effectors. Second, the discovery of metastable, inactive states with ligand binding sites would enable novel opportunities in the development of therapeutics targeted at Ras.

Potential Problems and Alternative Approaches: While we expect to recapitulate the inactive conformations of Ras, it is possible that we do not identify a conformation that is both inactive and exposes a ligand binding site. We find this to be highly unlikely, due to recent work [3] [9] which give us confidence that such sites exist. However, a potentially more interesting problem is that a conformation identified as inactive is actually active, potentially signaling through a non-canonical effector. While our work could continue even in this case (as we use the previously-discovered structure in [3]), this finding would lead to an interesting avenue of research into a novel Ras signaling partner.

Aim 1.4: Experimentally validate predicted conformational populations

Rationale: Once we have constructed a computationally useful MSM of Ras, it is important to perform experiments to validate that we have actually modeled the system in a relevant manner. Therefore, we propose to perform experiments where we can measure observable quantities, such as a hydrogen-exchange protection factor, that we can also predict within our MSM.

Specific Experiments and Data Analysis: Given a set of metastable conformations and their relative equilibrium populations, we first propose to calculate a quantity known as the "protection factor" for each backbone amide hydrogen, such as is done in [22]. This quantity reflects the likelihood of a proton to exchange with protons in the solvent, and is directly related to solvent-accessibility as well as lack of intramolecular hydrogen bonding at that point. To measure the quantity experimentally, we propose an experiment known as hydrogen-deuterium exchange mass spectrometry (HDX-MS). In this experiment, the protein is exposed to heavy water, allowing protons in the protein to exchange with deuterons in the solvent. Then, the protein is analyzed by mass spectrometry for sites which have exchanged their protons with deuterons. Due to the high sensitivity of mass spectrometry, even relatively low populations can be detected via this method, using very little protein [7]. We will then

compare the results of this experiment with the results of our calculations to validate that the MSM is accurately representing conformational populations of Ras.

Anticipated Results: We expect that the protection factors calculated from the MSM of K-Ras will match well with those measured by HDX-MS or NMR. Furthermore, we expect that the experiment will confirm the presence of existing and putative metastable binding sites, as a protection factor different from that expected from the crystal structure should be measured at those sites.

Potential Problems and Alternative Approaches: While these experiments are demonstrated in other studies [7], it is possible that we will find it difficult to obtain single-residue resolution using HDX-MS. In this case, we will turn to an established NMR technique for measurement of the same quantity. Alternatively, we could perform the NOESY experiment, which generates data that is related to the equilibrium distance between nuclei.

Aim 2: Identify new small molecule ligands of Ras

While screening can often identify weakly binding ligands, it is not typically clear where to proceed from there. We propose to develop a novel technique in which we can use the principles of statistical mechanics to derivatize ligands in a systematic way. We will develop an expanded-ensemble simulation that explores chemical space, allowing us to design ligands using principles of physics.

Hypothesis

Certain chemotypes are more likely to be found at a particular binding site than others. Here, we ask the question of which chemotypes are necessary to bind to the allosteric sites identified in Aim 1 using a novel theoretical framework, presented below.

Aim 2.1: Perform virtual screening on a library of commercially-available compounds against putative allosteric binding sites

Rationale: Once we have identified suitable metastable conformations, we intend to begin a virtual screen with a large library of purchasable drug-like fragments. This will allow us to discover new scaffolds for allosteric Ras ligands, and is an important step in the direction of novel Ras inhibitor design, serving as a launching pad for our method for rational optimization.

Specific Experiments: Databases such as ZINC [23] can provide collections of molecules that are readily available for purchase and meet certain criteria. We intend to begin by virtual screening compounds in the "Frags now" database, which is a database of over 1 million readily-available "fragment-like" (predicted logP$\leq$3.5, molecular weight$\leq$250, and the number of rotatable bonds$\leq$5). We will also filter by solubility, as the initial assays will take place at high concentration. The choice of the fragment database will allow us to screen a broad area of drug-like chemical space. We then intend to perform a docking-based virtual screen of our library against the conformations identified in Aim 1s, targeted at the potential binding sites for a small molecule. We will use a variety of docking tools to perform this screen after testing for known ligand binding over decoys. Molecular docking is a technique that uses a scoring function based on select features of the protein and ligand. While it often omits important features of the interaction, docking can be used to enrich a large library of compounds for potential hits. [24].

Data Analysis and Validation: Because docking is not typically able to produce an accurate binding affinity [25] we will use the more rigorous alchemical free energy calculations on the enriched set from docking. These methods create multiple systems which are alchemically modified; that is, the ligand and protein are decoupled to varying degrees, from completely interacting to completely non-interacting. Free energy differences, along with error estimates, are then calculated between each state [26]. We also intend to experimentally test the affinity of the compounds identified after the filter.
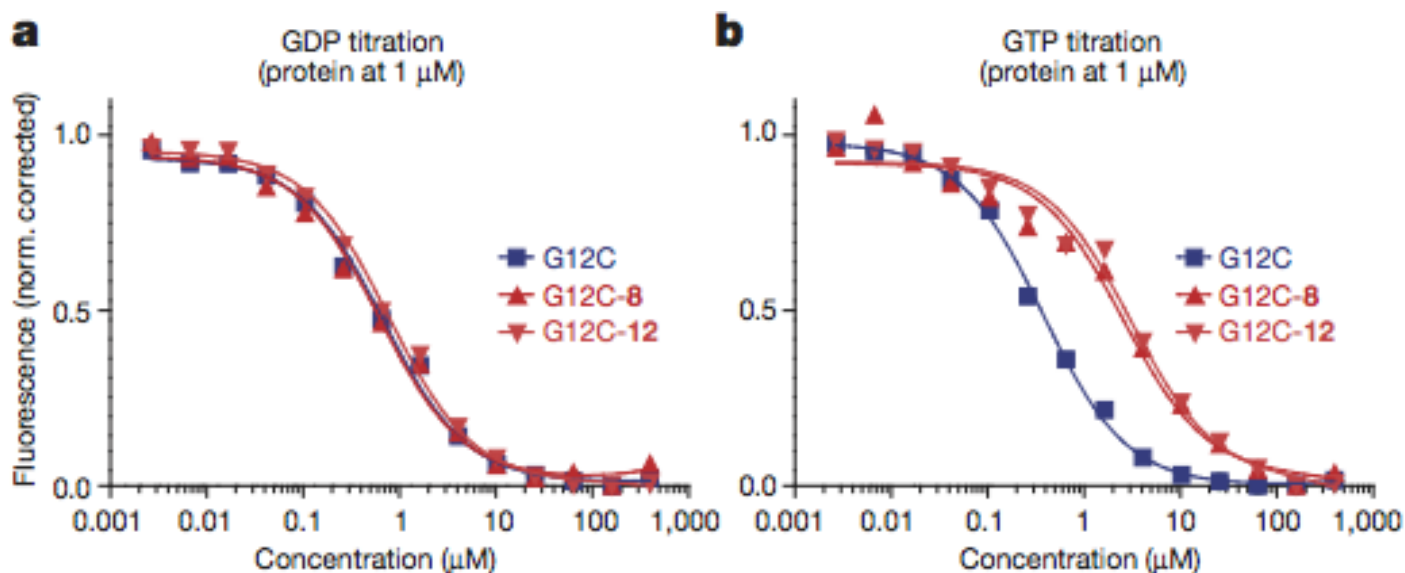
Figure 8: **As demonstrated here, perturbation with a hit should result in a shift in the affinity for GTP** [3]

If the compound is predicted to affect GTP affinity, we will use the assay depicted in Figure 8. We will also create multiple tryptophan mutants of Ras. These mutants will each contain a single tryptophan that will provide a fluorescent reporter of the conformational state off K-Ras; when a ligand predicted to stabilize an alternate conformation is added, we would expect to see a shift in the fluorescence of tryptophan residues exposed to a different solvent environment.

Anticipated Results: Out of the large library of over 1,000,000 fragment-like molecules in the "frags now" database of ZINC [23], we expect docking to enrich a collection of molecules for fragments that will more likely bind to the predicted binding cavity. We further expect that that binding to sites exposed only in excited state conformations will perturb the equilibrium conformational populations of Ras in a ligand concentration-dependent manner. Finally, we expect that the results of this stage will be useful leads to pursue tighter-binding derivatives and a set of much-needed starting points for allosteric Ras modulators.

Potential Problems and Alternative Approaches: While we expect that a large and diverse library will result in weak hits, it is possible that no such hits are present in the library. In this event, we could expand the library, or use a database of chemical transformations [27] to expand the complexity of the libraries. Finally, if no compounds are discovered, we will proceed with the weak scaffolds discovered in [3] [9].

Aim 2.2: Perform Expanded Ensemble Simulations Over Chemical Space to Enhance Ligands

Rationale: Often, when hits are discovered, medicinal chemists use their chemical intuition and available chemistry to explore chemical space near the hit. However, this can be error-prone, time-consuming, and costly, especially as one realizes the combinatorial explosion of possible modifications of only a single hit. We propose instead to use a rigorous statistical mechanics framework to computationally explore the chemical space near derivatives of the scaffold to improve binding affinity. Using the theory of expanded ensemble simulation [4], we will allow the lead to change various chemical substituents during simulation, biased toward tighter-binding ligands.

Theory: As indicated in Figure 9, even a scaffold with four potential sites for modification and 10 modifications at each site can result in a search space of 10^4 molecules. This is difficult both for traditional medicinal chemistry and for free energy calculations. Therefore, to computationally explore this combinatorially large chemical space, we propose to use expanded ensemble simulation [4], as

described below. We intend to use a database of chemical transformations, curated by medicinal chemists, to identify potential transformations of our ligand. These databases, such as [27], can allow us to propose changes to the ligand that are likely to be chemically accessible.
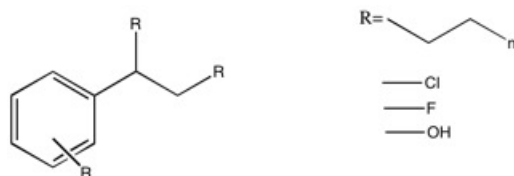

Figure 9: **An example of a Markush structure, containing combinatorially many possibilities.**

While simulation of biomolecules and ligands often proceeds via sampling (molecular dynamics or Markov Chain Monte Carlo) from the canonical ensemble, whose equilibrium distribution is given by

$$\pi(\mathbf{x}) = \frac{\exp[-u(\mathbf{x})]}{\int \exp[-u(\mathbf{x})]\mathrm{d}\mathbf{x}} = Z^{-1}\exp[-u(\mathbf{x})] \tag{1}$$

where $u(\mathbf{x})$ represents the reduced potential, given in its general form by

$$u(\mathbf{x}) = \beta[U(\mathbf{x}) + pV(\mathbf{x})] \tag{2}$$

where $\mathbf{x}$ is the configuration of the system, U($\mathbf{x}$) is the potential energy, V($\mathbf{x}$) is the volume (for a constant pressure simulation), p is the external pressure, and $\boldsymbol{\mu}$ is the chemical potential of each particle represented in $\mathbf{n}(\mathbf{x})$, which represents the number of particles in a (semi) grand canonical simulation [28]. Z represents the partition function, that is, the integral of $\exp[-u(\mathbf{x})]$ over its entire domain. However, we need not restrict ourselves to this; we can define an *expanded ensemble* such that we instead sample from a joint distribution given by

$$\pi(\mathbf{x}, k) = \frac{\exp[-u_k(\mathbf{x}) + g_k]}{\sum_{i=1}^{k} \int \exp[-u_i(\mathbf{x}) + g_i]\mathrm{d}\mathbf{x}} \tag{3}$$

where the new parameter *k* indexes the state of the simulation, and its corresponding u_k gives the the reduced potential for that state and configuration. In our case, the state k represents the current molecule, and a jump to a different state would represent a jump to a different ligand molecule. We also have the free parameter g_k, which enables us to weight each state with a biasing potential. Sampling from Equation (3) often proceeds by Gibbs sampling [29]; that is, we draw a correlated sample from the conditional of $\mathbf{x}$ given *k*,

$$\pi(\mathbf{x}|k) = \frac{\exp[-u_k(\mathbf{x})]}{\int \exp[-u_k(\mathbf{x})]\mathrm{d}\mathbf{x}} \tag{4}$$

followed by a sample from the conditional of *k* given $\mathbf{x}$:

$$\pi(k|\mathbf{x}) = \frac{\exp[-u_k(\mathbf{x}) + g_k]}{\sum_{i=1}^{k} \exp[-u_i(\mathbf{x}) + g_i]} \tag{5}$$

To address the question of how to set the weights g_k, we will show that, for a specific convenient choice, the marginal distribution $\pi(k)$ is proportional to the binding affinity, K_a, such that the expanded ensemble simulation will be driven to spend more time in chemical states that are tighter binders. This will enable expanded ensemble simulations to effectively hunt through large chemical spaces for good ligands. In order to derive this, we first marginalize out $\mathbf{x}$ in Equation (5):

$$\pi(k) = \frac{\exp(g_k) \int \exp[-u_k(\mathbf{x})]\mathrm{d}\mathbf{x}}{\sum_{i=1}^{k} \int \exp[-u_i(\mathbf{x}) + g_i]\mathrm{d}\mathbf{x}} \propto \exp(g_k)Z_{PL_k} \tag{6}$$
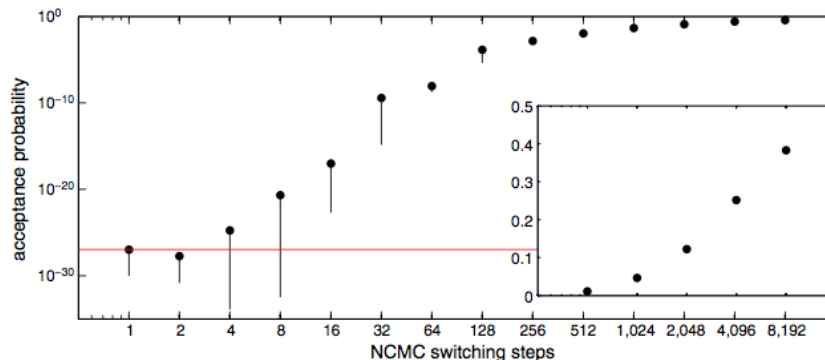
Figure 10: **NCMC can improve the acceptance rates of MC moves (such as those that will be used to propose changes in chemical identities) by orders of magnitude.** [31]

where Z_{PL_k} represents the partition function for the protein-ligand complex *k*. We also know from thermodynamics that binding affinity, excluding multiplicative constants, is calculated as [30]

$$K_a \equiv \frac{[PL]}{[P][L]} \propto \frac{Z_{PL}}{Z_{P\emptyset}} / \frac{Z_L}{Z_\emptyset} \tag{7}$$

and since the partition functions of the protein alone will not change during our expanded ensemble calculations, Equation (7) is proportional to

$$Z_{PL} / \frac{Z_L}{Z_\emptyset} \tag{8}$$

Thus, setting the weight $g_k = -ln(Z_{L_k}/Z_\emptyset)$ yields an unnormalized probability for being in state *k* of

$$\pi(k) \propto Z_{PL_k} / \frac{Z_{L_k}}{Z_\emptyset} \tag{9}$$

which is equivalent to the expression in Equation (8). Therefore, by using a specific $g_k = -ln(Z_{L_k}/Z_\emptyset)$, we can steer the simulation toward molecules with better K_a naturally. Though the above framework is theoretically sound, the presence of explicit waters near the site of transformation may create very low acceptance rates for state-change proposals, as there will likely be a clash if the ligand is instantaneously modified. We intend to take advantage of recent work [31] that has shown that one can make nonequilibrium moves and preserve the equilibrium distribution. Using this scheme, we would not simply propose to instantaneously change the state, but rather propose a protocol to transition between the current state and the next proposed state, interleaving partial perturbation steps with propagation steps. The final move is then accepted or rejected based on the nonequilibrium work performed, and, to preserve the equilibrium distribution, momenta are reversed if the move is rejected. As shown in Figure 10, for the test case of a WCA dimer, the use of NCMC dramatically increased acceptance rates and dramatically decreased correlation times.

Specific Experiments and Data Analysis: Although the above is tantalizing, calculating the normalizing constant of the ligand (equivalent to its hydration free energy) is nontrivial. Therefore, we propose to use single point energies, given by a simple evaluation of the forcefield at a minimized conformation in implicit solvent to provide an estimate for the weight to use in expanded ensemble simulation. This choice is considerably faster to evaluate, and, as we suggest below, may be a reasonable choice. We plan to run an expanded-ensemble simulation, starting from a weakly-binding ligand identified in Aim 2.1, using the aforementioned choice for the biasing potential. This choice is motivated by recent work [32] indicating that though point energies of ligands may be insufficient for calculating hydration free energy estimates, they are relatively close (RMSE=0.39±0.05 kcal/mol for minimized ligands in

solvent) [32]. After running simulations, and binning the number of samples from each ligand, we will perform free energy calculations on the ligands that the simulation visited, and compare the rankings from the expanded ensemble to those of the free energy calculation.

Anticipated Results: We expect that this theoretical result, and its practical application, will serve as a major innovation in drug design. Not only does it provide a method for enhancing ligand affinity that takes into account the dynamics of the protein and ligand, but it also allows us to ask questions about the nature of the relevant chemical space for binding a particular protein at a particular site, even in the absence of a ligand. We will use this information to interrogate identified Ras binding pockets for the relationship between chemotype and binding affinity. Additionally, because the single point estimates are relatively close to the hydration free energy values, we would expect them to still bias our expanded ensemble in the correct direction to discover tighter-binding ligands. Therefore, our free energy calculations should reflect a similar ranking as the expanded ensemble calculations.

Potential Problems and Alternative Approaches: Though previous results indicate that the point energy of minimized ligands is a good approximate to the hydration free energy, it is possible that, especially for larger and more flexible ligands, this approximation will not be appropriate. In this event, we could use hydration free energies to bias the simulation, as indicated in the theoretical result. Another potential issue is that the large number of low-affinity ligands that could be sampled will overwhelm the simulation, causing it to spend inordinate time sampling poorly-binding ligands. To remedy this, we will explore techniques to amplify the bias toward tightly-binding ligands.

Aim 2.3 Test whether compounds identified via the expanded-ensemble method will bind more tightly to oncogenic K-Ras and potentially disrupt its phenotype

Rationale: Once we have obtained simulation statistics in the expanded-ensemble technique described above, we will want to confirm that the identified molecules actually bind.

Specific Experiments and Data Analysis: The first step in confirming that these ligands bind is to run alchemical free energy calculations on the top-scorers, confirming that they bind to a considerable enough degree to warrant synthesis and wet lab experimentation. This is done because free energies of binding can be computed much more precisely this way than looking at $-kTln[\pi(k)]$ from the expanded ensemble simulation. If the ligands appear to be tightly-binding in calculation, we will have them synthesized by collaborators at the organic synthesis core for *in vitro* testing. Then, we will use assays as described above to assay for GTP/GDP presence with the ligand, as well as solvent-exposed regions. We will also measure the affinity of the ligand using isothermal titration calorimetry (ITC), very popular and important in binding affinity measurement [33], for which our lab is developing novel techniques. If the ligand is experimentally confirmed to bind, we will then collaborate with cancer biologists to assay cell growth in the presence of the inhibitor, as shown in [3]. Finally, having shown that the ligand inhibits cancer cell growth, we will contact a structural biology collaborator at Stony Brook University to assist us in generating X-ray crystal or NMR evidence of the bound pose of the ligand and K-Ras.

Anticipated Results: We expect that subjecting our lead compounds identified in the previous screen to our technique of expanded-ensemble simulation will identify derivatives that are both synthetically realistic and tighter binders. We anticipate that this result will also indicate novel chemical motifs that may be relevant in targeting Ras, and may suggest a general approach for identifying relevant chemical space for targeting a particular protein/binding pocket. We furthermore expect that the ligand will stabilize the expected conformation and cause measurable changes in the activity of Ras. This will help pave the way for the development of novel anticancer drugs targeting this heretofore very difficult-to-target protein.

Potential Problems and Alternative Approaches: While we expect this aim to result in a tighter-binding ligand, there are some potential pitfalls. One issue is that the proposed changes to the ligand are very difficult or infeasible to make in the lab; while we expect our synthetic feasibility model to avoid this, if

it does occur, we will restrict the modifications available to a particular lead based on a more detailed curation process.

Conclusion

Once completed, our technique will have accomplished several innovations. First, we will have produced novel ligands that target and stabilize Ras family proteins, providing a new avenue for development of therapeutics targeting this "undruggable" target, including producing data that can produce a quantitative structure-activity relationship for Ras ligands. We will also have produced a map of the conformational populations of Ras, enabling further studies of its functional biophysics, critical to understanding the oncogenic potential of Ras. Furthermore, as our technique is a systematic one, it will serve as a template for the discovery of allosteric sites and ligands, as well as a rigorous technique for the improvement of scaffold binding in the face of combinatorially-large chemical space.

References

[1] Yuliya Pylayeva-Gupta, Elda Grabocka, and Dafna Bar-Sagi. RAS oncogenes: weaving a tumorigenic web. *Nat Rev Cancer*, 11(11):761–774, oct 2011.

[2] Raymond H. Mak, Gretchen Hermann, John H. Lewis, Hugo J.W.L. Aerts, Elizabeth H. Baldini, Aileen B. Chen, Yolonda L. Colson, Fred H. Hacker, David Kozono, Jon O. Wee, Yu-Hui Chen, Paul J. Catalano, Kwok-Kin Wong, and David J. Sher. Outcomes by tumor histology and KRAS mutation status after lung stereotactic body radiation therapy for early stage non-small cell lung cancer. *Clinical Lung Cancer*, sep 2014.

[3] Jonathan M. Ostrem, Ulf Peters, Martin L. Sos, James A. Wells, and Kevan M. Shokat. K-ras(g12c) inhibitors allosterically control GTP affinity and effector interactions. *Nature*, 503(7477):548–551, nov 2013.

[4] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov. New approach to monte carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys.*, 96(3):1776, 1992.

[5] Richard A. Lewis. A general method for exploiting QSAR models in lead optimization. *Journal of Medicinal Chemistry*, 48(5):1638–1648, mar 2005.

[6] John D Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Current Opinion in Structural Biology*, 25:135–144, apr 2014.

[7] Lars Konermann, Jingxi Pan, and Yu-Hong Liu. Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem. Soc. Rev.*, 40(3):1224, 2011.

[8] C. W. Reuter, M. A. Morgan, and L. Bergmann. Targeting the Ras signaling pathway: a rational, mechanism-based treatment for hematologic malignancies? *Blood*, 96(5):1655–1669, Sep 2000.

[9] Qi Sun, Jason P. Burke, Jason Phan, Michael C. Burns, Edward T. Olejniczak, Alex G. Waterson, Taekyu Lee, Olivia W. Rossanese, and Stephen W. Fesik. Discovery of small molecules that bind to k-ras and inhibit sos-mediated activation. *Angew. Chem. Int. Ed.*, 51(25):6140–6143, may 2012.

[10] G. M. Lee and C. S. Craik. Trapping moving targets with small molecules. *Science*, 324(5924):213–215, apr 2009.

[11] Peter Eastman, Mark S. Friedrichs, John D. Chodera, Randall J. Radmer, Christopher M. Bruns, Joy P. Ku, Kyle A. Beauchamp, Thomas J. Lane, Lee-Ping Wang, Diwakar Shukla, Tony Tye, Mike Houston, Timo Stich, Christoph Klein, Michael R. Shirts, and Vijay S. Pande. OpenMM 4: A reusable, extensible, hardware independent library for high performance molecular simulation. *J. Chem. Theory Comput.*, 9(1):461–469, jan 2013.

[12] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, oct 1990.

[13] Andrej Šali and Tom L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, dec 1993.

[14] Kyle A. Beauchamp, Gregory R. Bowman, Thomas J. Lane, Lutz Maibaum, Imran S. Haque, and Vijay S. Pande. MSMBuilder2: Modeling conformational dynamics on the picosecond to millisecond scale. *J. Chem. Theory Comput.*, 7(10):3412–3419, oct 2011.

[15] Manfred Hendlich, Friedrich Rippmann, and Gerhard Barnickel. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling*, 15(6):359–363, dec 1997.

[16] I. D. Kuntz, K. Chen, K. A. Sharp, and P. A. Kollman. The maximal affinity of ligands. *Proceedings of the National Academy of Sciences*, 96(18):9997–10002, aug 1999.

[17] G. R. Bowman and P. L. Geissler. Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proceedings of the National Academy of Sciences*, 109(29):11681–11686, jul 2012.

[18] D. A. Erlanson, A. C. Braisted, D. R. Raphael, M. Randal, R. M. Stroud, E. M. Gordon, and J. A. Wells. Site-directed ligand discovery. *Proceedings of the National Academy of Sciences*, 97(17):9367–9372, aug 2000.

[19] J. C. Hunter, D. Gurbani, S. B. Ficarro, M. A. Carrasco, S. M. Lim, H. G. Choi, T. Xie, J. A. Marto, Z. Chen, N. S. Gray, and K. D. Westover. In situ selectivity profiling and crystal structure of SML-8-73-1, an active site inhibitor of oncogenic k-ras g12c. *Proceedings of the National Academy of Sciences*, 111(24):8895–8900, jun 2014.

[20] Michael E. Pacold, Sabine Suire, Olga Perisic, Samuel Lara-Gonzalez, Colin T. Davis, Edward H. Walker, Phillip T. Hawkins, Len Stephens, John F. Eccleston, and Roger L. Williams. Crystal structure and functional analysis of ras binding to its effector phosphoinositide 3-kinase . *Cell*, 103(6):931–944, dec 2000.

[21] Liang Tong, Abraham M. de Vos, Michael V. Milburn, and Sung-Hou Kim. Crystal structures at 2.2 {aa resolution of the catalytic domains of normal ras protein and an oncogenic mutant complexed with GDP. *Journal of Molecular Biology*, 217(3):503–516, feb 1991.

[22] J. J. Skinner, W. Yu, E. K. Gichana, M. C. Baxa, J. R. Hinshaw, K. F. Freed, and T. R. Sosnick. Benchmarking all-atom simulations using hydrogen exchange. *Proceedings of the National Academy of Sciences*, 111(45):15975–15980, oct 2014.

[23] John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. ZINC: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, jul 2012.

[24] Brian K Shoichet, Susan L McGovern, Binqing Wei, and John J Irwin. Lead discovery using molecular docking. *Current Opinion in Chemical Biology*, 6(4):439–446, aug 2002.

[25] Gregory L. Warren, C. Webster Andrews, Anna-Maria Capelli, Brian Clarke, Judith LaLonde, Millard H. Lambert, Mika Lindvall, Neysa Nevins, Simon F. Semus, Stefan Senger, Giovanna Tedesco, Ian D. Wall, James M. Woolven, Catherine E. Peishoff, and Martha S. Head. A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry*, 49(20):5912–5931, oct 2006.

[26] Stefan Boresch, Franz Tettinger, Martin Leitgeb, and Martin Karplus. Absolute binding free energies: a quantitative approach for their calculation. *J. Phys. Chem. B*, 107(35):9535–9551, sep 2003.

[27] Markus Hartenfeller, Martin Eberle, Peter Meier, Cristina Nieto-Oberhuber, Karl-Heinz Altmann, Gisbert Schneider, Edgar Jacoby, and Steffen Renner. A collection of robust organic synthesis reactions for in silico molecule design. *Journal of Chemical Information and Modeling*, 51(12):3093–3098, dec 2011.

[28] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129(12):124105, 2008.

[29] Jun S. Liu. *Monte Carlo strategies in scientific computing*. Springer, 2001.

[30] M.K. Gilson, J.A. Given, B.L. Bush, and J.A. McCammon. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical Journal*, 72(3):1047–1069, mar 1997.

[31] J. P. Nilmeier, G. E. Crooks, D. D. L. Minh, and J. D. Chodera. Nonequilibrium candidate monte carlo is an efficient tool for equilibrium simulation. *Proceedings of the National Academy of Sciences*, 108(45):E1009–E1018, oct 2011.

[32] David L. Mobley, Ken A. Dill, and John D. Chodera. Treating entropy and conformational changes in implicit solvent simulations of small molecules. *J. Phys. Chem. B*, 112(3):938–946, jan 2008.

[33] Stephanie Leavitt and Ernesto Freire. Direct measurement of protein binding energetics by isothermal titration calorimetry. *Current Opinion in Structural Biology*, 11(5):560–566, sep 2001.