Specific Aims

Ras family proteins, important in the control of cell growth, is a common site for oncogenic mutants 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 [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 require the presence of a cysteine in the active site [3], present in only a small fraction of cases, or are weak [4] Worse still, the need for a cysteine results in a straightforward path to drug resistance mutants lacking the cysteine. Here, we propose a novel computational pipeline to identify targetable conformations and discover noncovalent ligands to allosterically inhibit Ras, starting from a computational analysis of Ras conformational populations, allowing rational design of allosteric inhibitors without beginning with an inhibitor.

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**.

Aim 3: Computationally explore chemical space near identified ligands

We will **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.

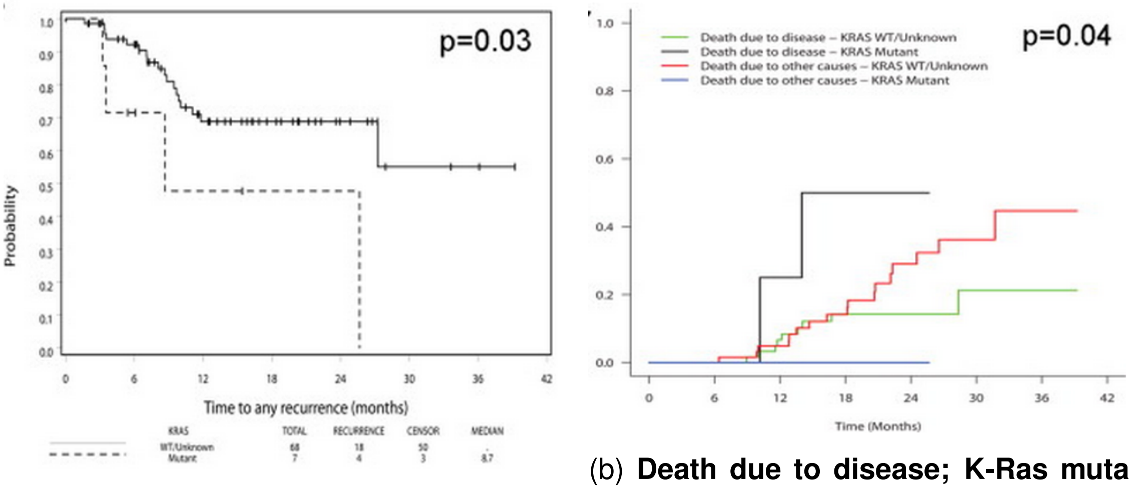
Abstract

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 mutates 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 pipeline 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, dysregulated mutants of it play a central role in the progression of serious cancers [2]. In Figure 1a, 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 1b, 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.



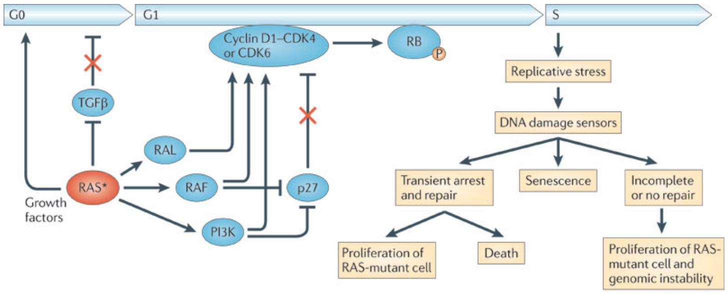
(a) **Time to recurrence**

(b) **Death due to disease; K-Ras mutant, black**

Figure 1: Kaplan-Meier plots of K-Ras mutant vs. wild-type show that prognosis is considerably worse for K-Ras mutant genotypes. [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].



Nature Reviews | Cancer

Figure 2: **A diagram of major Ras growth signaling partners, indicating its connectedness in growth signaling pathways** [1]

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 conformations, some of which are signaling-inactive and expose druggable binding sites for stabilizing ligands

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 [4] [3] which contain ligand binding sites, and may be signaling-inactive. Discovery of such sites allows us to rationally design ligands that can stabilize these inactive conformations, thus potentially inhibiting oncogenic growth signaling and leading to a new path to cancer therapeutics. In order to discover these conformations, we will use parallel high-throughput molecular dynamics to generate trajectory data, and Markov State Modeling (MSM) [5] to combine the trajectory data into a conformational map, as conceptually illustrated in Figure 3.

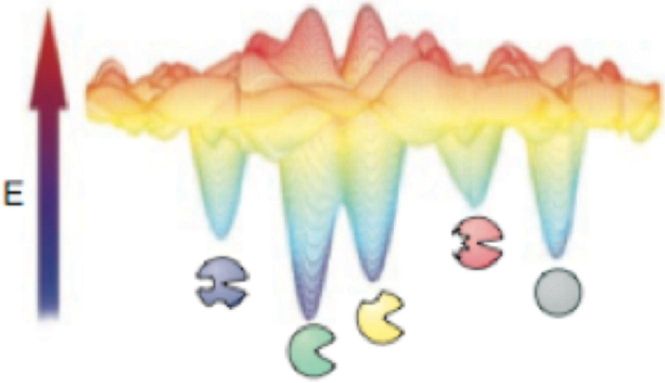


Figure 3: **A schematic of an energy landscape with multiple wells for metastable states** [6]

Objective 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 [7], and to combine the independent trajectories using Markov State Model approaches [5] 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 will give us unprecedented insight into opportunities for drugging this target.

Markov State Modeling: This process involves several steps. First, as the trajectory data is very high-dimensional, we will use a dimensionality-reduction technique, such as time-lagged independent component analysis (tICA) [8]. This technique transforms input coordinates, such as distances or torsions, into a new set of coordinates, highlighting the slowest degrees of freedom [5]. Next, a technique such as Voronoi tessellation is used to assign the reduced-dimension trajectories to discrete states [5]. These now-discretized trajectories are used to estimate the transition matrix between states [5]. 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 [9].

Initialization with starting structures: Despite the use of parallel, high-performance computing hardware, it remains difficult to fully sample the conformation space of a protein. In order to more rapidly and effectively sample the conformational states of K-Ras, we will use a pipeline developed in our laboratory that identifies protein structures in the protein data bank (PDB) by BLAST [10] sequence similarity, then constructs homology models of the target sequence using MODELLER [11], followed by a brief molecular dynamics refinement as shown in Figure 4. By starting from diverse conformations, we expect to more rapidly cover conformational space in the MD trajectories.

Massively Parallel Molecular Dynamics (MD): After generating the initial conformations, we will run explicit-solvent molecular dynamics from this diverse set of starting structures, creating several copies of each with randomized starting velocities. This work will be performed primarily using Oak Ridge National Lab's Titan, the second-fastest civilian supercomputer in the world, for which we have been allocated compute time. We will use OpenMM [7], an open-source, GPU-accelerated high-performance molecular simulation package, to obtain trajectories for analysis. Initially, we will simulate the G12C mutant, as it is both clinically relevant and the subject of recent investigation [3], as well as the apo form of the K-Ras protein. We intend to repeat this protocol for GDP- and GTP- bound forms, as well as other clinically-relevant mutants.

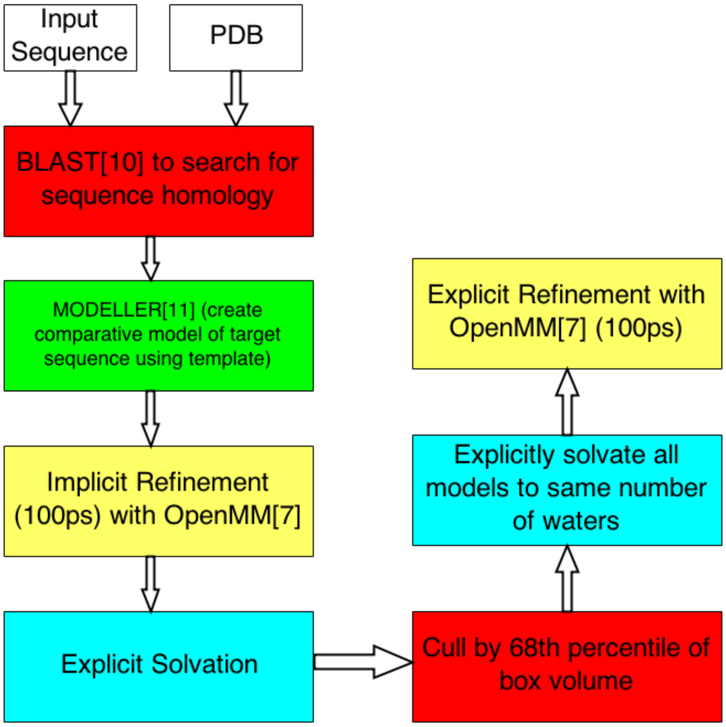

Input Sequence
PDB
BLAST[10] to search for sequence homology
MODELLER[11] (create comparative model of target sequence using template)
Implicit Refinement (100ps) with OpenMM[7]
Explicit Solvation
Cull by 68th percentile of box volume
Explicitly solvate all models to same number of waters
Explicit Refinement with OpenMM[7] (100ps)

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

Analyzing trajectory data: As trajectory data is received from Titan, we will proceed by creating an MSM of the conformations of K-Ras. The output of the MSM analysis will render several important pieces of information. First, it will enumerate a set of metastable conformations adopted by K-Ras. Moreover, the model will also yield the kinetics and energetics of these conformations, allowing us to identify near-ground state conformations in K-Ras.

Objective 2: Mine the Metastable Conformations for potential ligand binding sites

The creation of a Markov State Model for Ras will provide us with unparalleled insight into this protein's conformational populations. With this data, we can begin searching for opportunities for allosteric modulation; to do so, we will look for conformations that contain potential ligand binding sites, as well as conformations which display similarity to the GDP-bound (inactive) conformation, especially in the Switch I and II effector regions displayed in Figure 8, as described below. We expect that this will lead to new avenues for therapeutic lead discovery.

Elucidating excited state conformations: Since our aim is to allosterically target Ras by stabilizing an inactive conformation with a small molecule, we will confine our search to metastable conformational states that are within $10k_BT$ of the ground state. Conformations beyond this energy level would likely be difficult to stabilize with a small molecule [12].

Discovering Potential Small Molecule Binding Sites: With this 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 [13]. This technique searches for points nearly-surrounded by solvent-inaccessible regions [13]. 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 [14], as illustrated in Figure 5.

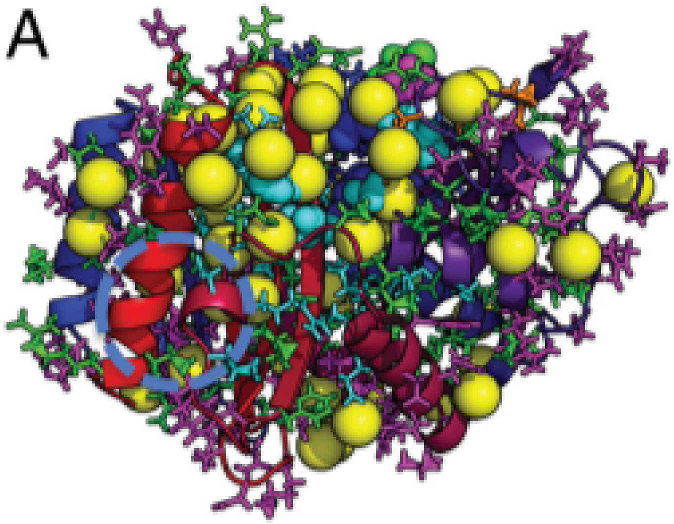


Figure 5: **An example of the identification of cryptic binding sites from MSM data on β-lactamase**. The known allosteric binding site is circled in blue, identified via MSM techniques without ligand. [14]

Second, the existence of weakly-binding molecules to allosteric sites on K-Ras, such as the covalent inhibitors in [3] depicted in Figure 6, whose bound structure is illustrated in Figure 7. However, the molecules in [3] was discovered by a screening technique [15] 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.

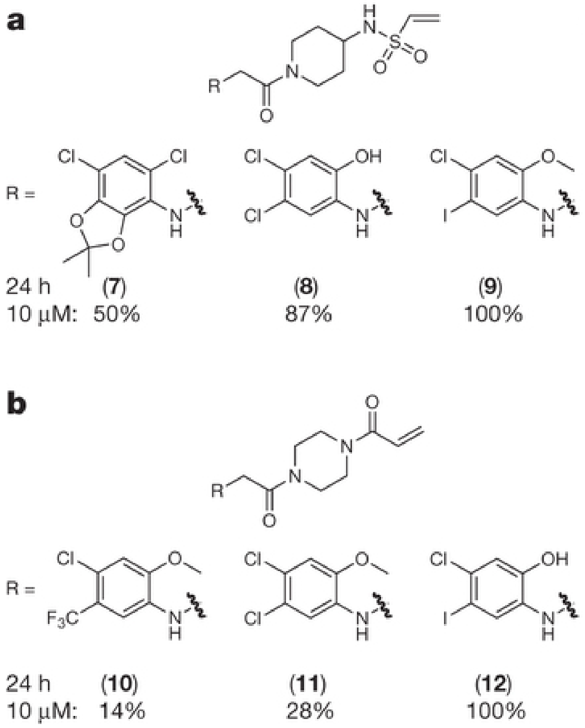


Figure 6: **Several molecules identified in [3] to inhibit K-Ras G12C.** Percentages denote covalent attachment after 24h in 10μM compound. Note the presence of an electrophile for covalent attachment to the required cysteine. Image from [3]

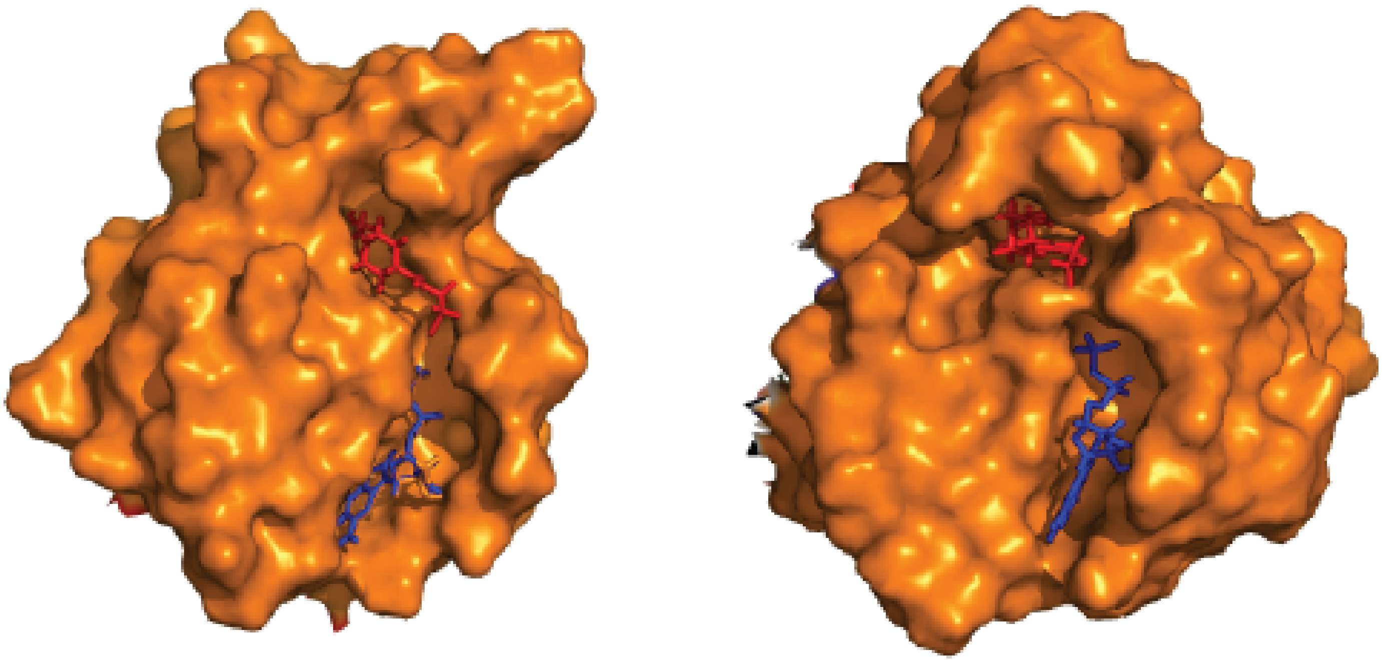


Figure 7: **K-Ras G12C with bound inhibitor, in red, aligned to inactive conformation on left.** [3]. The presence of the inhibitor (red) reveals an allosteric pocket above the nucleotide (blue) binding site in bound PDB entry 4LV6, unbound 4OBE.

Objective 3: Identify putative druggable signaling-inactive conformations

Once we have identified metastable conformations, it is not only important to locate opportunities for ligand binding, but especially those states that may represent signalling-inactive states that could be preferentially stabilized by an allosteric ligand.

Favoring the GDP-bound form: The molecules in Figure 6 were shown to cause K-Ras to favor its GDP-bound form instead of its active GTP bound form. [3]. Thus, we will search for conformations which are more similar to the GDP bound conformations such as those found in PDB codes 4LPK and 4LRW using metrics such as RMSD, as well as placement of key residues involved in nucleotide binding.

Disrupting interaction with Raf or PI3K: Key interaction partners of Ras family proteins include SOS, an upstream effector that encourages nucleotide exchange, Raf, a group of protein kinases, and 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 [16], such as those indicated in Figure 8.

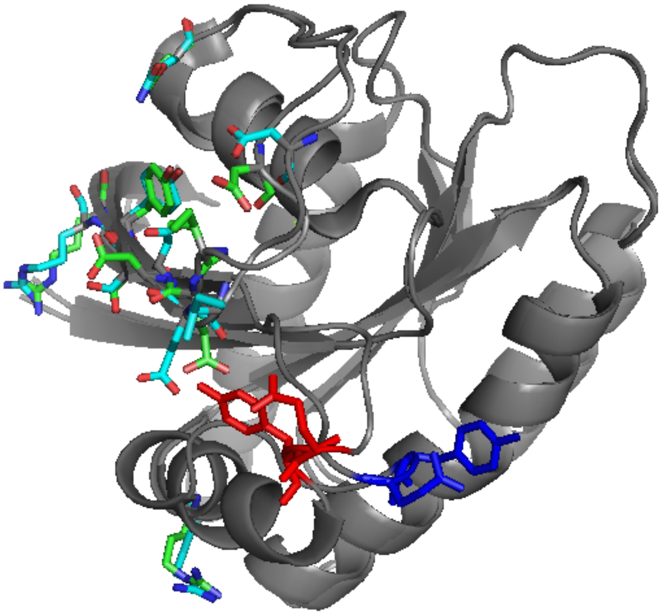


Figure 8: **Alignment of GDP and GTP bound conformations of H-Ras, showing positioning of PI3K-interacting residues [16].** PDB codes: inactive 1Q21 [17], green side chains, active 5P21 [18], teal side chains. Red and blue side chains denote dramatically different positioning between active and inactive forms, respectively.

Objective 4: Experimentally validate predicted conformational populations

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. To that end, I will develop several techniques to biophysically examine the effect of allosteric ligand binding on the conformational populations of Ras in an experimentally- and computationally-observable manner.

Free energy calculations with known ligands: We will perform free energy calculations on known Ras ligands such as those in [4], to the MSM-derived conformations, combining metastable-state-specific binding affinities with apo equilibrium populations to compute the overall effective binding affinity and ligand concentration-dependent state populations. If the conformational populations are predicted correctly, the free energy calculations should give a similar value to measured binding affinity.

Fluorescent labeling of Ras: We will also use an assay involving the introduction of multiple minimally-perturbative tryptophan mutants as described in [19]. Depending on their location, these mutants can be used to confirm binding of a ligand (if the tryptophan is near the binding site) or a conformational change upon ligand binding (if the tryptophan is distant from the binding site). We intend to use known ligands to perturb tryptophan-mutant Ras, and compare fluorescence results to the computed fluorescence results using computed conformation-dependent binding affinities, apo-populations, and a convenient surrogate for the spectroscopic signal (such as the average solvent-exposed surface area of tryptophan residues for each conformational state).

Anticipated Results

We anticipate that this aim will generate a valuable model of the conformational landscape of K-Ras, including potentially inactive conformations that expose putative ligand binding sites. We have high confidence that this will be possible, as there are several allosteric ligands known already [3] [4], albeit with serious shortcomings—those that bind tightly enough to be useful are covalent inhibitors, and therefore require the presence of a cysteine at the binding site. Where our method provides considerable improvements over currently available data and techniques is in creating a map of the conformations of K-Ras, with associated kinetics and energetics—these are quantities that are difficult, if not impossible, to determine from bound structures of serendipitously discovered allosteric modulators.

Potential Problems and Alternative Approaches

The forcefield may be inaccurate: It is conceivable that the classical forcefields used in molecular dynamics are either incorrectly parameterized or insufficiently detailed to recapitulate conformational populations. To that end, our laboratory is engaged in collaborative projects to benchmark and improve forcefields and their parameters. However, recent studies of forcefields show continuing improvement, as indicated in Figure 9, where lower scores are better [20] and suggest that the problem may not be so dire.

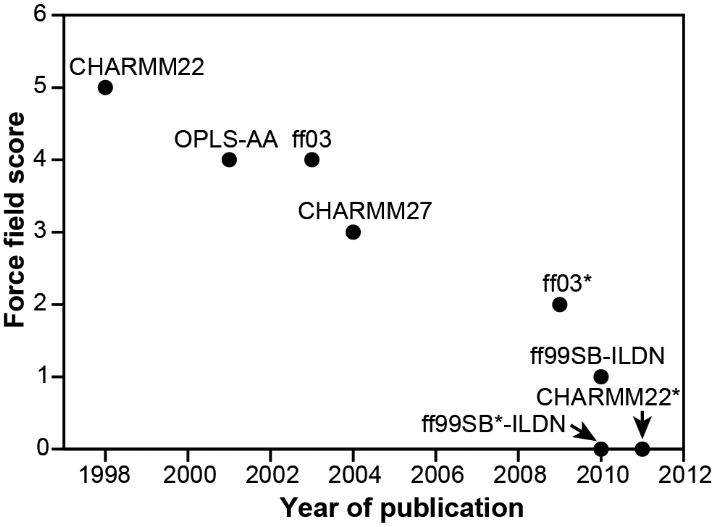


Figure 9: **A comparison of forcefields over time** [20]. Lower scores indicate better agreement with experiment, demonstrating consistent improvement.

There may be no identified conformations with binding pockets: This seems highly unlikely, given the existence of demonstrated binding pockets [3] [4], but in this event we can screen against the binding pockets identified in [3] and [4].

Minimally-perturbative tryptophan mutants may not be possible at the required sites: As an alternative to labeling via tryptophan mutants, we can use fluorescent labeling on surface cysteines in the region of interest to probe conformational populations.

Compare MSM conformational population to NMR results: If the above methods prove inconclusive, we intend to compare the conformational populations obtained via MSM to NMR data such as ligand concentration-dependent NOEs.

Aim 2: Identify new small molecule ligands of Ras

Hypothesis

Virtual screening can identify fragment-like compounds that bind to putative allosteric sites from Aim 1.

Objective 1: Perform virtual screening on a library of commercially-available compounds against putative allosteric binding sites

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.

Retrieve database of readily available fragment-like compounds: Databases such as ZINC [21] 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.

Use small molecule docking to discover potential weakly-binding fragments to allosteric sites in conformati

Once potentially inactive conformations have been identified, we intend to perform a docking-based virtual screen of our library agains these conformations, 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. [22]

Objective 2: Filter hits using alchemical free energy calculations:

Highest scoring compounds are enriched for binders: Although docking makes many approximations, it has been shown that docking with a large library of molecules leads to an enrichment among the highest-scoring molecules for potential hits. Therefore, we will use the top 1% of the screening results, filtered by physical properties such as solubility, for this stage.

Free energy calculations: Because docking is not typically able to produce an accurate binding affinity [23] 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 [24].

Objective 3: Experimentally measure binding affinities of hits

Test differential affinity for GDP and GTP: If the identified putative inhibition of Ras involves the stabilization of a GDP-favored conformation as in [3], we intend to test the compound’s ability to bias Ras toward a GDP-bound form. We will follow the protocol outlined in [3]. This assay begins with mant-GDP-loaded Ras, which is fluorescent. The assay then titrates in unlabeled GTP, or a nonhydrolyzable analog, and measures the change in fluorescence as a result. As shown in Figure 10, if the ligand actually biases Ras toward a GDP-bound conformation, a shift in the affinity from untreated Ras should be seen, denoting a decrease in affinity for GTP. As a control, both ligand-bound and ligand-free Ras should be titrated with unlabeled GDP, and there should be no difference in the change in fluorescence vs. concentration GDP.

Test modulation of conformational populations with fluorescent labeling: We will use the assay developed in Aim 1, validated on known Ras ligands, to experimentally test the fragments discovered in our virtual screen.

Anticipated Results

Out of the large library of over 1,000,000 fragment-like molecules in the ”frags now” database of ZINC [21], 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 detectable by our fluorescence assay. Finally, we expect that the results of this

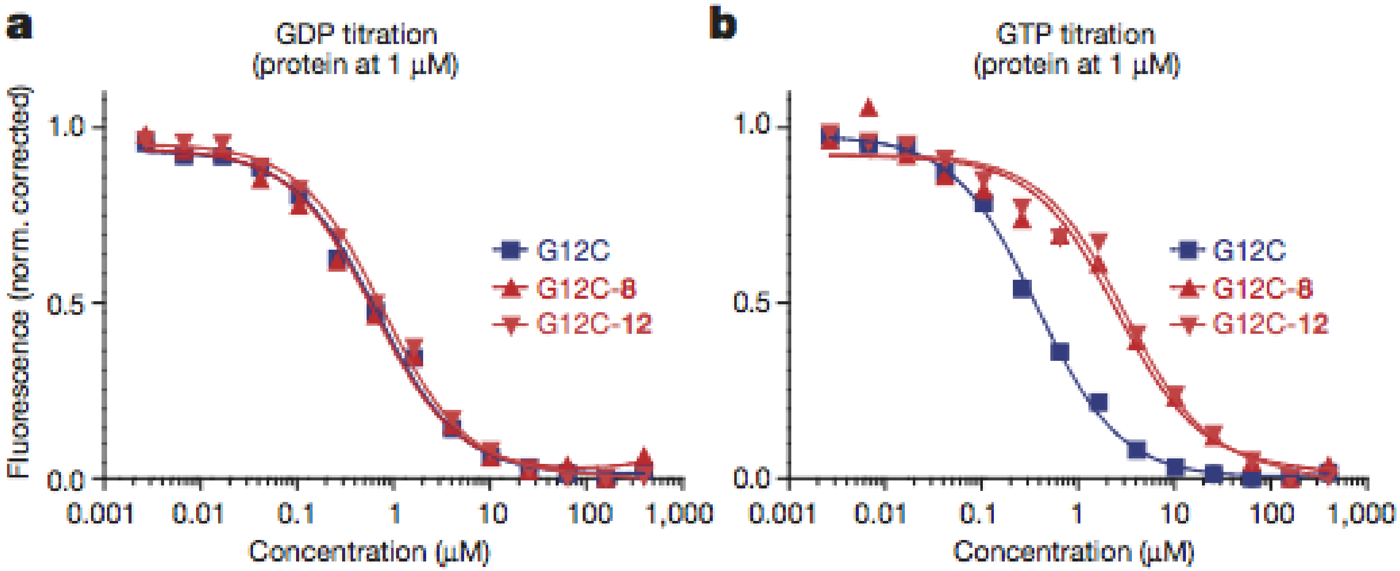


Figure 10: **As demonstrated here, perturbation with a hit should result in a shift in the affinity for GTP** [3]

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 the large library of available fragments to contain some hits, there are several unlikely issues that we may encounter:

The fragment database may not contain any binding ligands: It is possible that the database we will use does not contain any binding ligands, or contains only ligands that bind so weakly that the effect is undetectable even at the solubility limit. This might be evidenced by few high-scoring docking poses, or free energy calculations that suggest no binding. To resolve such an issue, we would seek other screening libraries, such as a library used internally at Memorial Sloan Kettering, or other libraries on ZINC.

Docking may not accurately score ligands: This problem would be evidenced by free energy calculations showing no significant improvement of binding affinity among high-scoring docking results [23], or a failure to distinguish known ligands from decoys. This is a significant issue [23], which we would first attempt to resolve by adding various amounts of detail to the docking calculations. Absent improvement, we could potentially turn to previously-discovered scaffolds such as those in [4] or [3].

Aim 3: Computationally explore chemical space near identified ligands

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 [25], we will allow the lead to change various chemical substituents during simulation, biased toward tighter-binding ligands.

Hypothesis

Expanded-ensemble techniques that allow simulations to explore chemical space can identify novel chemical motifs necessary for binding to a specific protein.

Objective 1: Perform expanded-ensemble simulations on synthetically-feasible derivatives of the identified leads

Synthetic feasibility model: We intend to use a database of chemical transformations, curated by medicinal chemists, to identify potential transformations of our ligand. These databases, such as [26], can allow us to make proposed transformations.

Chemical space becomes combinatorially large: As indicated in Figure 11, 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 [25], as described below.

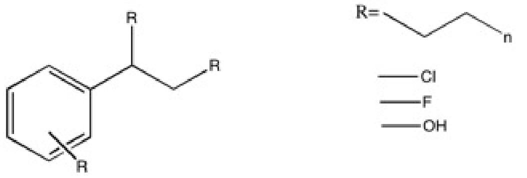


Figure 11: **An example of a Markush structure, containing combinatorially many possibilities.**

Expanded ensemble simulation: 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

$$u(\mathbf{x}) = \beta[U(\mathbf{x}) + pV(\mathbf{x}) + \boldsymbol{\mu}^T\mathbf{n}(\mathbf{x})] \tag{2}$$

where **x** is the configuration of the system, U(**x**) is the potential energy, V(**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 **n**(**x**), which represents the number of particles in a (semi) grand canonical simulation [27]. 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 [28]; that is, we draw a correlated sample from the conditional of **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 **x**:

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

Biasing the simulation toward tighter binders: We will show that, for a specific convenient choice of the weights g_k, 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 **x** in Equation (5):

$$\pi(k) = \frac{\exp(g_k) \int \exp[-u_k(\mathbf{x})] \mathrm{d}\mathbf{x}}{\sum_{i=1}^{k} \int \exp[-u_i(\mathbf{x}) + g_i] \mathrm{d}\mathbf{x}} \propto \exp(g_k) Z_{PL_k} \tag{6}$$

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 [29]

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

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

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

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

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

which is equivalent to the expression in Equation (8). Therefore, by using a specific $g_k = -lnZ_{L_k}$, which we can approximate with the hydration free energy in explicit or implicit solvent (discussed below), we can steer the simulation toward molecules with better K_a naturally.

Use Nonequilibrium Candidate Monte Carlo (NCMC) moves for explicit-solvent simulation: 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 [30] 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 12, for the test case of a WCA dimer, the use of NCMC dramatically increased acceptance rates and dramatically decreased correlation times.

Objective 2: Test whether single point energies of minimized ligands are sufficient to bias the simulation toward tight binders

Setting bias to a single point energy is a useful approximation to unbound partition function: 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

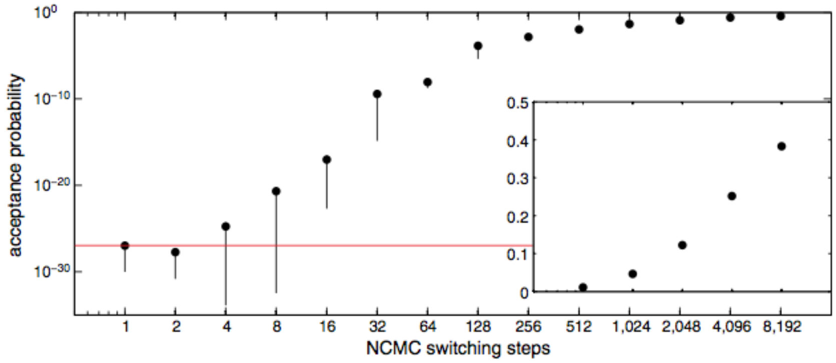


Figure 12: **NCMC can improve the acceptance rates of MC moves (such as those that will be used to propose changes in chemical identities) by orders of magnitude.** [30]

in expanded ensemble simulation. This choice is motivated by recent work [31] 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) [31]. 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.

Objective 3: Test whether compounds identified via the expanded-ensemble method will bind more tightly to oncogenic K-Ras and potentially disrupt its phenotype

Once we have obtained simulation statistics in the expanded-ensemble technique described above, we will move to confirm that these ligands actually bind to oncogenic K-Ras.

Perform free energy calculations on derivatives: 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.

Collaborate with organic synthesis core to have derivatives synthesized: Since we have started with purchasable fragments and derivatized them *in silico*, we will collaborate with the organic synthesis core at Memorial Sloan Kettering to have the derivatives synthesized for *in vitro* testing.

Biophysical assay for GDP preference: If the conformation being stabilized is one that we predict to prefer GDP (for instance, by similarity to the GDP-bound conformation), we will use the GTP-GDP preference assay described above to compare the affinity of our optimized ligand to the original fragment.

Fluorescently labeling positions on the protein: We will also use fluorescence assays developed in Aim2 to compare the ability of the optimized ligands to affect the conformational populations of Ras with the corresponding unoptimized fragments.

Perform Isothermal Titration Calorimetry: This technique, very popular and important in binding affinity measurement [32] measures the heat released by ligand binding by monitoring the amount of heat required to keep a reference cell at the same temperature as a sample cell, into which ligand is titrated. ITC has been used for fragment-based studies [33], indicating that it is viable even for compounds with weak affinity. Additionally, our lab is developing new techniques to cope with low-affinity ITC. This technique is useful for measuring whether the ligand binds at all, and with what affinity. Due to the large amount of ligand required, however, its use may be limited for low-solubility compounds. s

Assay the effect of the ligand on cell growth: We will collaborate with cancer biologists to assay the ligand in inhibiting cancer cell growth, in both cells dependent on Ras and cells with Ras knocked out to control for nonspecific binding, following a protocol similar to that followed by Ostrem, et al. [3].

Engage structural biology collaborator for structural evidence of binding: After demonstrating that the ligand successfully binds and has an effect on cancer 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

Expanded ensemble simulation techniques will improve ligand affinity: 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:

The synthetic feasibility model may need refinement: Synthetic feasibility models curated by medicinal chemists may not always reflect what is synthetically feasible on every scaffold that matches the pattern. As an alternative, if we are generating very unrealistic proposals, we can limit the scope of the modifications, as well as restrict the modifications available to a particular lead based on curation in greater detail.

The single point-energy estimate may be a poor approximation for complex ligands: For ligands with high entropic contributions, the point-energy estimate may prove insufficient. As a replacement, the more computationally costly but more rigorous approach of using solvation free energies can be substituted.

The large number of weakly-binding derivatives can overwhelm the small number of tight binders: It is possible that the very large number of weakly-binding variants that can be visited in a simulation will overwhelm the small number of tightly-binding derivatives. To combat this issue, we will explore techniques to amplify the bias toward very tightly-binding ligands.

Fluorescence-based methods may not reveal relevant activity even if it exists: If the predicted conformational or nucleotide-preference changes do not occur, the fluorescence measurements may report no change. To combat this and confirm the negative, we will use techniques such as isothermal titration calorimetry (ITC), which are label-free and can confirm that no binding within detection limits is occurring.

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] 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.

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

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

[7] 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.

[8] Guillermo Perez-Hernandez, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noe. Identification of slow molecular order parameters for markov model construction. *J. Chem. Phys.*, 139(1):015102, 2013.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] E. F. Pai, U. Krengel, G. A. Petsko, R. S. Goody, W. Kabsch, and A. Wittinghofer. Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 A resolution: implications for the mechanism of GTP hydrolysis. *EMBO J.*, 9(8):2351–2359, Aug 1990.

[19] Bruno Antonny, Pierre Chardin, Michel Roux, and Marc Chabre. GTP hydrolysis mechanisms in ras p21 and in the ras-GAP complex studied by fluorescence measurements on tryptophan mutants. *Biochemistry*, 30(34):8287–8295, aug 1991.

[20] Kresten Lindorff-Larsen, Paul Maragakis, Stefano Piana, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. Systematic validation of protein force fields against experimental data. *PLoS ONE*, 7(2):e32131, feb 2012.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] Nyssa Drinkwater, Hoan Vu, Kimberly M. Lovell, Kevin R. Criscione, Brett M. Collins, Thomas E. Prisinzano, Sally?Ann Poulsen, Michael J. McLeish, Gary L. Grunewald, and Jennifer L. Martin. Fragment-based screening by x-ray crystallography, MS and isothermal titration calorimetry to identify PNMT (phenylethanolamine n-methyltransferase) inhibitors. *Biochem. J.*, 431(1):51–61, sep 2010.