bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Novel, provable algorithms for efficient ensemble-based computational protein design and their application to the redesign of the c-Raf-RBD:KRas protein-protein interface

Anna U. Lowegard[1,2☯], Marcel S. Frenkel[3☯], Jonathan D. Jou[2], Adegoke A. Ojewole[1,2], Graham T. Holt[1,2], Bruce R. Donald[2,3*],

1 Program in Computational Biology and Bioinformatics, Duke University Medical Center, Durham, NC, USA

2 Department of Computer Science, Duke University, Durham, NC, USA

3 Department of Biochemistry, Duke University Medical Center, Durham, NC, USA

☯These authors contributed equally to this work.

* brd+pcb19@cs.duke.edu

Abstract

The K^* algorithm provably approximates partition functions for a set of states (e.g., protein, ligand, and protein-ligand complex) to a user-specified accuracy ε. Often, reaching an ε-approximation for a particular set of partition functions takes a prohibitive amount of time and space. To alleviate some of this cost, we introduce two algorithms into the OSPREY suite for protein design: FRIES, a Fast Removal of Inadequately Energied Sequences, and *EWAK**, an Energy Window Approximation to K^*. In combination, these algorithms provably retain calculational accuracy while limiting the input sequence space and the conformations included in each partition function calculation to only the most energetically favorable. This combined approach leads to significant speed-ups compared to the previous state-of-the-art multi-sequence algorithm, *BBK**. As a proof of concept, we used these new algorithms to redesign the protein-protein interface (PPI) of the c-Raf-RBD:KRas complex. The Ras-binding domain of the protein kinase c-Raf (c-Raf-RBD) is the tightest known binder of KRas, a

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

historically "undruggable" protein implicated in difficult-to-treat cancers including pancreatic ductal adenocarcinoma (PDAC). FRIES/$EWAK^*$ accurately retrospectively predicted the effect of 38 out of 41 different sets of mutations in the PPI of the c-Raf-RBD:KRas complex. Notably, these mutations include mutations whose effect had previously been incorrectly predicted using other computational methods. Next, we used FRIES/$EWAK^*$ for prospective design and discovered a novel point mutation that improves binding of c-Raf-RBD to KRas in its active, GTP-bound state (KRas$^{\mathrm{GTP}}$). We combined this new mutation with two previously reported mutations (which were also highly-ranked by OSPREY) to create a new variant of c-Raf-RBD, c-Raf-RBD(RKY). FRIES/$EWAK^*$ in OSPREY computationally predicted that this new variant would bind even more tightly than the previous best-binding variant, c-Raf-RBD(RK). We measured the binding affinity of c-Raf-RBD(RKY) using a bio-layer interferometry (BLI) assay and found that this new variant exhibits single-digit nanomolar affinity for KRas$^{\mathrm{GTP}}$, confirming the computational predictions made with FRIES/$EWAK^*$. This study steps through the advancement and development of computational protein design by presenting theory, new algorithms, accurate retrospective designs, new prospective designs, and biochemical validation.

Author summary

Computational structure-based protein design is an innovative tool for redesigning proteins to introduce a particular or novel function. One such possible function is improving the binding of one protein to another, which can increase our understanding of biomedically important protein systems toward the improvement or development of novel therapeutics. Herein we introduce two novel, provable algorithms, FRIES and $EWAK^*$, for more efficient computational structure-based protein design as well as their application to the redesign of the c-Raf-RBD:KRas protein-protein interface. These new algorithms speed up computational structure-based protein design while maintaining accurate calculations, allowing for larger, previously infeasible protein designs. Using FRIES and $EWAK^*$ within the OSPREY suite, we designed the tightest known binder of KRas, an "undruggable" cancer target. This new variant of a KRas-binding domain, c-Raf-RBD, should serve as an important tool to probe the protein-protein interface

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

between KRas and its effectors as work continues toward an effective therapeutic targeting KRas.

Introduction

Computational structure-based protein design (CSPD) is an innovative tool that enables the prediction of protein sequences with desired biochemical properties (such as improved binding affinity). OSPREY (Open Source Protein Redesign for You) [1] is an open-source, state-of-the-art software package used for CSPD and is available at http://www.cs.duke.edu/donaldlab/osprey.php for free. OSPREY's algorithms focus on *provably* returning the optimal sequences and conformations for a given input model. In contrast, as argued in [2–7], stochastic, non-deterministic approaches [8–10] provide no guarantees on the quality of conformations, or sequences, and make determining sources of error in predicted designs very difficult.

When using OSPREY, the input model generally consists of a protein structure, a flexibility model (e.g., choice of sidechain or backbone flexibility, allowed mutable residues, etc.), and an all-atom pairwise-decomposable energy function that is used to evaluate conformations. OSPREY models amino acid sidechains using frequently observed rotational isomers or "rotamers" [11]. Additionally, OSPREY can also model continuous sidechain flexibility [12–15] along with discrete and continuous backbone flexibility [16–19], which allow for a more accurate approximation of protein behavior [13, 16, 20–23]. The output produced by CSPD generally consists of a set of candidate sequences and conformations. Many protein design methods have focused on computing a global minimum energy conformation (GMEC) [14, 18, 24–28]. However, a protein in solution exists not as a single, low-energy structure, but as a thermodynamic ensemble of conformations. Models that only consider the GMEC may incorrectly predict biophysical properties such as binding [12, 20–23, 29–31] because GMEC-based algorithms underestimate potentially significant entropic contributions. In contrast to GMEC-based approaches, the K^* algorithm [12, 29, 30] in OSPREY models thermodynamic ensembles to provably and efficiently approximate the K^* score. The K^* score is a ratio of the Boltzmann-weighted partition functions for a protein-ligand complex that estimates the association constant, K_a (further detailed in the Section

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

entitled "Computational materials and methods"). *BBK** [32] is the most recent improvement on the traditional K^* algorithm that allows for multi-sequence design. Previous algorithms [12, 27, 29, 30, 33–35] that design for binding affinity using ensembles are linear in the size of the sequence space N, where N is exponential in the number of simultaneously mutable residue positions. *BBK** is the first provable ensemble-based algorithm to run in time sublinear in N, making it possible not only to perform K^* designs over large sequence spaces, but also to enumerate a gap-free list of sequences in order of decreasing K^* score.

OSPREY has been used successfully on several empirical, prospective designs including designing enzymes [12, 16, 22, 29, 36], resistance mutations [2, 37, 38], protein-protein interaction inhibitors [30, 39], epitope-specific antibody probes [40], and broadly-neutralizing antibodies [41, 42]. These successes have been validated experimentally *in vitro* and *in vivo* and are now being tested in several clinical trials [43–45]. However, while OSPREY has been successful in the past, as the size of protein design problems grows (e.g., when considering a large protein-protein interface), enumerating and minimizing the necessary number of conformations and sequences to satisfy the provable halting criteria in previous K^*-based algorithms [12, 29, 30] becomes prohibitive (despite recent algorithmic improvements [32]). The entire conformation space can be monumental in size and heavily populated with energetically unfavorable sequences and conformations. *EWAK**, an Energy Window Approximation to K^*, seeks to alleviate some of this difficulty by restricting the conformations included in each sequence's thermodynamic ensemble. *EWAK** *guarantees* that each conformational ensemble contains *all* of the lowest energy conformations within an energy window of the GMEC for each design sequence. FRIES, a Fast Removal of Inadequately Energied Sequences, also mitigates this complexity problem by limiting the input sequence space to only the most favorable, low energy sequences. Previous algorithms have focused on optimizing for sequences whose conformations are similar in energy to that of the GMEC. In contrast, FRIES focuses on optimizing for sequences with energies better-than or comparable-to the wild-type sequence. FRIES *guarantees* that the restricted input sequence space includes all of the sequences within an energy window of the wild-type sequence, but excludes any potentially unstable sequences with significantly worse partition function values. Wild-type sequences are generally expected to be

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig 1. Design example using the structure of the LecB lectin *Pseudomonas aeruginosa* strain PA14 (PDB ID: 5A6Y [47]) and the OSPREY workflow for FRIES/*EWAK**. In the top panel, the full, 4 domain structure of lectin is shown on the left-hand side. (A) Zooming in on the region where domains A (green) and D (yellow) interact, showing the two mutable residues (Q80 and I82) along with the surrounding flexible shell of residues as lines. There were 11 flexible residues included in this design with Q80 and I82 allowed to mutate to all other amino acids except for proline. This design consisted of 8.102×10^{11} conformations and 441 sequences. FRIES limited this space to 5.704×10^{11} conformations and 206 sequences. FRIES/*EWAK** in combination reduced the amount of time taken by about 75% compared to *BBK**. FRIES alone was responsible for roughly 50% of this speed-up. (B) 10 low-energy conformations included in the thermodynamic ensemble of the design sequence with mutations Q80I and I82F. For this particular sequence, *BBK** minimized 10,664 conformations while *EWAK** minimized only 4,104 conformations. The bottom panel shows the general workflow for FRIES/*EWAK**. The workflow begins with the input model (as described in the Section entitled "Introduction"), which defines the design space for the first algorithm, FRIES. FRIES proceeds to prune the sequence space as described in the Section entitled "Fast Removal of Inadequately Energied Sequences (FRIES)" and as illustrated in the Venn diagram with the unpruned space shown as a yellow disk. Next, the remaining FRIES sequence space defines the conformation space (which contains multiple sequences as well as conformations) searched with *EWAK**. *EWAK** limits the conformations included in each partition function as described in the Section entitled "Energy Window Approximation to K* (*EWAK**)." *EWAK** generally searches over only a subset of the conformations (green area) that previous K^*-based algorithms like *BBK** [32] search (orange area). *EWAK** then returns the top sequences based on decreasing K^* score.

near-optimal for their corresponding folds [46]. Therefore, limiting the sequence space to sequences energetically similar to or better than the wild-type sequence is reasonable. A simplified workflow for FRIES/*EWAK** is presented in Fig 1. Compared to the previous state-of-the-art algorithm *BBK**, FRIES and *EWAK** improve runtimes by up to 2 orders of magnitude, FRIES decreases the size of the sequence space by up to more than 2 orders of magnitude, and *EWAK** decreases the number of conformations included in partition function calculations by up to almost 2 orders of magnitude.

As a proof of concept to test these algorithms and our design approach, we used FRIES and *EWAK** to study the protein-protein interface (PPI) of KRas$^{\mathrm{GTP}}$ in complex with its tightest-binding effector, c-Raf. As described in the Section entitled "Computational redesign of the c-Raf-RBD:KRas protein-protein interface," KRas is an important cancer target that has historically been considered "undruggable" [48]. Deepening the understanding of the PPI between KRas and its effectors is an important step toward developing effective new therapeutics. For this study, we focused on the redesign of the c-Raf Ras-binding domain (c-Raf-RBD) in complex with KRas$^{\mathrm{GTP}}$ (c-Raf-RBD:KRas$^{\mathrm{GTP}}$). First, our new algorithms successfully retrospectively predicted

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the effect on binding of mutations in the c-Raf-RBD:KRas$^{\mathrm{GTP}}$ PPI even where other computational methods previously failed [49]. Next, we used FRIES/*EWAK** prospectively to predict the effect of novel, previously unreported mutations in the PPI of the c-Raf-RBD:KRas$^{\mathrm{GTP}}$ complex. We then measured the binding of top OSPREY-predicted c-Raf-RBD variants to KRas using a bio-layer interferometry (BLI) assay with a single-concentration screen. This screen suggested that one of our new computationally-predicted c-Raf-RBD variants – c-Raf-RBD(Y), a c-Raf-RBD variant that includes the mutation V88Y – exhibits improved binding to KRas$^{\mathrm{GTP}}$. Next, we created a c-Raf-RBD variant, c-Raf-RBD(RKY), that included this new mutation, V88Y, together with two previously reported mutations [49], N71R and A85K. FRIES/*EWAK** computationally predicted that c-Raf-RBD(RKY) would bind more tightly to KRas$^{\mathrm{GTP}}$ than any other variant. The single-concentration screen using BLI also suggested that c-Raf-RBD(RKY) binds more tightly to KRas$^{\mathrm{GTP}}$ than the previously reported best variant [49]. The K_d values for the most promising variants were measured using a BLI assay with titration which confirmed our computational predictions and that, to the best of our knowledge, the novel construct c-Raf-RBD(RKY) is the highest affinity variant ever designed, with single-digit nanomolar affinity for KRas$^{\mathrm{GTP}}$.

Computational materials and methods

The K^* algorithm's [12, 29, 30] K^* score serves as an estimate of the binding constant, K_{a}, and is calculated by first approximating the Boltzmann-weighted partition function of each state: unbound protein (P), unbound ligand (L), and the bound protein-ligand complex (C). Each Boltzmann-weighted partition function $Z_x(\mathbf{s})$, $x \in \{P, L, C\}$, is defined as:

$$Z_x(\mathbf{s}) = \sum_{d \in \mathbf{Q}(\mathbf{s})} \exp(-E_x(d)/RT). \tag{1}$$

If $\mathbf{s}$ is any – generally amino acid – sequence of n residues, then $\mathbf{Q}(\mathbf{s})$ is the set of conformations defined by $\mathbf{s}$, $E_x(d)$ is the minimized energy of a conformation d in state x, and R and T are the gas constant and temperature, respectively. Many protein

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

design algorithms approximate these partition functions for each state using either stochastic [50–53] or provable [2,12,29–31,33,53] methods.

OSPREY's K^* algorithm provably approximates these partition functions to within a user-specified ε of the full partition function as defined in Eq (1). The binding affinity for sequence $\mathbf{s}$ is defined as:

$$K_a(\mathbf{s}) = \frac{Z_C(\mathbf{s})}{Z_P(\mathbf{s})Z_L(\mathbf{s})}. \tag{2}$$

The K^* algorithm provably approximates this binding affinity. This is enabled by the use of A^* [4,12,26,54], which allows for the gap-free enumeration of conformations in order of increasing lower bounds on energy [26]. However, enumerating a sufficient number of these conformations to obtain a guaranteed ε-approximation can be very time consuming because the set of all conformations $\mathbf{Q}(\mathbf{s})$ grows exponentially with the number of residues n. Also, the K^* algorithm was originally [12,29,30] limited to computing a K^* score for every sequence in the sequence space as defined by the input model for a particular design. However, BBK^* [32] builds on K^* and provably returns the top m sequences along with their ε-approximate K^* scores and runs in time sublinear in the number of sequences. That is, BBK^* does not require calculating ε-approximate K^* scores for (or even examining) every sequence in the sequence space before it returns the top sequences. Nevertheless, BBK^* may spend unnecessary time and resources evaluating unfavorable sequences before deciding to prune them.

To overcome the above limitations of BBK^* and K^*, we introduce FRIES, a Fast Removal of Inadequately Energied Sequences, and $EWAK^*$, an Energy Window Approximation to K^*. These two algorithms focus on limiting the input sequence space and the number of conformations included in each partition function estimate when approximating a sequence's K^* score to provably only the most energetically favorable options. The FRIES/$EWAK^*$ approach limits the number of conformations that must be enumerated (see the Section entitled "$EWAK^*$ limits the number of minimized conformations when approximating partition functions while maintaining accurate K* scores"), which leads to significant speed-ups (see the Section entitled "FRIES/$EWAK^*$ is up to 2 orders of magnitude faster than BBK*") because each enumerated conformation must undergo an energy minimization step. This minimization step is

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

relatively expensive, therefore, anything that reduces the number of minimized conformations while not sacrificing provable accuracy is desirable. For the importance of this minimization step to biological accuracy, see the discussions of continuous flexibility and its comparison to discrete flexibility in [4, 5, 7, 13, 14, 19]. $EWAK^*$ also maintains the advances made by BBK^* including running in time sublinear in the number of sequences N and returning sequences in order of decreasing K^* score. FRIES and $EWAK^*$ are described in further detail in the Section entitled "Algorithms" below.

Algorithms

Fast Removal of Inadequately Energied Sequences (FRIES). Generally in protein design when optimizing a protein-protein interface (PPI) for affinity, the designer aims to improve the K^* score of a variant sequence relative to the wild-type sequence, and, when performing a design targeting a similar fold, to minimally perturb the native structure. To accomplish this, FRIES guarantees to only keep sequences whose partition function values are not markedly worse than the wild-type sequence's partition function values for all of the design states (e.g. protein, ligand, and complex). How many orders of magnitude worse a particular sequence's partition function values are allowed to be is determined by a user-specified value m. The FRIES algorithm prunes sequences that exhibit massive decreases in partition function values that signal an increased risk of disturbing the native structure of the states in a given system. However, sequences with markedly worse, lower partition function values may be required when searching for, for example, resistance mutations, where positive and negative design are necessary [2, 37, 38]. Importantly, FRIES does still allow for sequences that may have lower, worse partition function values by allowing the user to specify how many orders of magnitude lower a candidate sequence's partition function is allowed to be relative to the wild-type sequence's partition function.

To prune the input sequence space, FRIES exploits A^* over a *multi-sequence tree* (as is described and used in COMETS [55]), which enjoys a fast sequence enumeration in order of lower bound on minimized energy. Each sequence v in this *multi-sequence tree* [55] has a corresponding *single-sequence conformation tree*, viz., a tree that can be searched for the lowest energy conformations for a sequence v. FRIES first enumerates

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sequences (in order of energy lower bounds) in the *multi-sequence tree* until the wild-type sequence is found. Then, FRIES searches the wild-type's corresponding *single-sequence conformation tree* using A^*. The first conformation enumerated according to monotonic lower bound on pairwise minimized energy is then subjected to a full-atom minimization [30] to calculate the minimized energy of one of the wild-type sequence's conformations E_{WT}. FRIES then continues enumerating sequences in the *multi-sequence tree* in order of increasing lower bound on minimized energy until the lower-bound on the energy of a sequence v, $E_v^{\ominus}$, is greater than $E_{WT} + w$ where E_{WT} is as described above and w is a user-specified energy window value (Fig 2). Any variant sequence v with a lower bound on minimized energy $E_v^{\ominus}$ not satisfying the following criterion is pruned:

$$E_v^{\ominus} \leq E_{WT} + w. \tag{3}$$

This criterion guarantees that the remaining, unpruned sequence space includes all sequences within an energy window of the wild-type sequence's energy. FRIES enumerates sequences in order of increasing lower bound on minimized energy. Therefore, it calculates an upper bound $q_v^{\oplus}$ on the partition function for each sequence v by Boltzmann-weighting the lower bound on its energy $E_v^{\ominus}$ and multiplying it by the size of the conformation space for that particular sequence $|\mathbf{Q}(v)|$:

$$q_v^{\oplus} = |\mathbf{Q}(v)| \exp(-E_v^{\ominus}/RT). \tag{4}$$

The lower bound for the wild-type sequence $q_{WT}^{\ominus}$ is calculated by Boltzmann-weighting the minimized energy of the single conformation found during the sequence search for the wild-type sequence E_{WT}:

$$q_{WT}^{\ominus} = \exp(-E_{WT}/RT). \tag{5}$$

$q_{WT}^{\ominus}$ is a lower bound because, in the worst case, at least this one conformation will contribute to the partition function for the wild-type sequence. FRIES then uses these bounds to remove all of the sequences whose partition function value is not within some user-specified m orders of magnitude of the lower bound on the wild-type partition

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

function $q^{\ominus}_{WT}$. If the following criterion is not met, the sequence v is pruned from the space:

$$\ln q^{\oplus}_{v} \leq \ln q^{\ominus}_{WT} + m. \tag{6}$$

FRIES prunes sequences for the protein, the ligand, and the protein-ligand complex independently, limiting the input sequence space to exclude unfavorable sequences for all of the states. The resulting smaller sequence space is subsequently used as input for *EWAK**. The set of sequences remaining is guaranteed to include all of the sequences within a user-specified energy window w of the wild-type sequence that also satisfy the partition function criterion given in Eq (4). Importantly, FRIES can be used to limit the size of the input sequence space in this fashion for any of the protein design algorithms available within OSPREY.

Fig 2. How FRIES chooses which sequences to keep and which sequences to prune. The solid curve represents the energy landscape of the conformation space that spans across, in this example, 7 different sequences (separated by dotted lines). Each sequence is labeled on the x-axis with an index indicating the order with which it is (or would be) enumerated with FRIES in order of increasing lower bound on minimized energy (red dotted curve). FRIES continues to enumerate in this way until it encounters the wild-type sequence (green), at which point FRIES calculates the minimized energy E_{WT} of the conformation with the lowest lower bound on minimized energy for the wild-type sequence (marked with a green dot). E_{WT} then becomes the baseline from which FRIES can provably enumerate all remaining sequences within some user-specified energy window w (yellow lines). Finally, FRIES prunes the sequences with energies provably higher than $E_{WT} + w$ (black) and keeps the sequences that occur within the shaded yellow region (colored in blue and green). More sequences are also pruned according to their partition function values as described in the Section entitled "Fast Removal of Inadequately Energied Sequences (FRIES)" and as defined by Eq (4).

Energy Window Approximation to K* (*EWAK).** After reducing the size of the input sequence space using FRIES, as described in the Section entitled "Fast Removal of Inadequately Energied Sequences (FRIES)," *EWAK** proceeds by using a variation on an existing algorithm: *BBK** (described in [32]). The crucial difference between *BBK** and *EWAK** is that with *EWAK** the ensemble of conformations used to approximate each K^* score is limited to those within a user-specified energy window of the GMEC for each sequence. This guarantees to populate the partition function for a particular sequence and state with all of the provably lowest, most-favorable conformations (that fall within the user-specified energy window). These conformations often account for the

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

majority of the full ε-approximate partition function (see the Section entitled "Computational materials and methods") in traditional K^* calculations [12]. Hence, $EWAK^*$ also empirically enjoys negligible loss in accuracy of K^* scores (see the Sections entitled "$EWAK^*$ limits the number of minimized conformations when approximating partition functions while maintaining accurate K* scores" and "FRIES/$EWAK^*$ retrospectively predicted the effect mutations in c-Raf-RBD have on binding to KRas"). $EWAK^*$ retains the beneficial aspects of BBK^*, including returning sequences in order of decreasing predicted binding affinity and running in time sublinear in the number of sequences.

Computational experiments

We implemented FRIES/$EWAK^*$ in the OSPREY suite of open source protein design algorithms [1]. FRIES was tested on 2,662 designs that range from an input sequence space size of 441 to 10,164 total sequences. The size of the reduced input sequence space produced by FRIES was compared to the size of the full input sequence space size for each design. For these tests, FRIES returned every sequence within 8 kcal/mol of the wild-type sequence and was set to include only those sequences that are at most 2 orders of magnitude worse in partition function value than the wild-type. The results for these tests are described in the Section entitled "FRIES can reduce the size of the input sequence space by more than 2 orders of magnitude while retaining the most favorable sequences." Computational experiments were also run comparing FRIES/$EWAK^*$ with the previous state-of-the-art algorithm in OSPREY: BBK^* [32]. Using BBK^* and FRIES/$EWAK^*$, we computed the top 5 best binding sequences for 167 different designs to compare the running time of BBK^* vs. FRIES/$EWAK^*$. FRIES was limited to sequences within 4 kcal/mol of the wild-type sequence that are at most 2 orders of magnitude worse in partition function values than the wild-type. The $EWAK^*$ partition function approximations were limited to conformations within an energy window of 1 kcal/mol of the GMEC for each sequence. BBK^* was set to return the top 5 sequences with an accuracy of $\varepsilon = 0.68$ (as was described in [32]). Using these same $EWAK^*$ and BBK^* parameters, we also compared the change in the size of the conformation space necessary to compute an accurate K^* score for BBK^* vs. $EWAK^*$

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for 661 partition functions from 161 design examples. The results for these tests are described in Sections entitled "FRIES/$EWAK^*$ is up to 2 orders of magnitude faster than BBK*" and "FRIES can reduce the size of the input sequence space by more than 2 orders of magnitude while retaining the most favorable sequences." The number of conformations that undergo minimization (as described in [12–15]) for each partition function calculation with $EWAK^*$ was also compared across different energy window sizes for 350 partition function calculations from 87 design examples. These partition function calculations were compared to BBK^*'s partition function calculations with a demanded accuracy of $\varepsilon = 0.10$. This smaller ε allowed for more accurate approximations of the K^* scores. The results for these tests are described in the Section entitled "FRIES can reduce the size of the input sequence space by more than 2 orders of magnitude while retaining the most favorable sequences."

Every design included a set of mutable residues along with a set of surrounding flexible residues (Fig 1 for an example). All of these residues were allowed to be continuously flexible [12–15]. The designs were selected from 40 different protein structures (listed in S1 Table and also used in [32, 56]), and were run on 40-48 core Intel Xeon nodes with up to 200 GB of memory.

Computational results

FRIES can reduce the size of the input sequence space by more than 2 orders of magnitude while retaining the most favorable sequences

The number of remaining sequences after FRIES was compared to the size of the complete input sequence space. In the best case, when using FRIES, the sequence space was decreased by more than 2 orders of magnitude and the conformation space was decreased by just over 4 orders of magnitude. The sequence space was reduced an average of 49% and the conformation space was reduced an average of 40%. These results are broken down further in Fig 3.

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig 3. Reduction in input sequence space size using FRIES. (A) A pie chart representing the reduction in the sequence space in percentages across all 2,662 designs. 7% of the designs had a reduction in sequence space over 95%, 24% of the designs had a reduction in sequence space between 66-95%, 31% of the designs had a reduction in sequence space between 36-65%, 32% of the designs had a reduction in sequence space between 6-35%, and 6% of the designs had a reduction in sequence space under 5%. (B) and (C) plot the number of sequences remaining after using FRIES starting with 441 and 9,261 sequences total, respectively. The number of sequences remaining for each design are sorted in order of decreasing size of the remaining conformation space after FRIES.

FRIES/*EWAK** is up to 2 orders of magnitude faster than BBK*

The overall runtime was compared between *BBK** and FRIES/*EWAK**. FRIES/*EWAK** was an average of 62% faster than *BBK** on 167 example design problems. FRIES removed unfavorable sequences (as described in the Section entitled "Fast Removal of Inadequately Energied Sequences (FRIES)") from the search space for 156 out of the 167 design problems. For the cases described in the Section entitled "Computational experiments," FRIES/*EWAK** performed consistently faster than *BBK** (in 92% of the design examples) as shown in Fig 4, Panel A. The longest running *BBK** design problem took nearly 8 days, whereas FRIES/*EWAK** completed the same example in just under 2 hours. In contrast, the design problem that took the longest for FRIES/*EWAK** out of the 167 tested only required about 22 hours (the same design took *BBK** over 178 hours).

*EWAK** limits the number of minimized conformations when approximating partition functions while maintaining accurate K* scores

We examined 661 K^* score calculations, and concluded that the total number of conformations minimized to approximate the K^* score was decreased by an average of 27%. In the best case the number of conformations minimized to approximate the K^* score was decreased by 93%. These results are plotted in Fig 4, Panel B. Even though the partition function approximations were limited to a smaller conformation space with *EWAK**, the K^* scores did not differ by more than 0.2 orders of magnitude between *EWAK** and *BBK** for these 661 example K^* score calculations.

A total of 350 of these 661 partition functions were subsequently re-estimated using

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig 4. Comparing runtimes and the number of minimized conformations between FRIES/*EWAK and *BBK** for a variety of designs.** (A) A plot of the runtime in seconds (the y-axis is on a log scale) for FRIES/*EWAK** (blue dots) and *BBK** (yellow dots) for 167 design examples. Each point represents one design and is plotted in increasing order of *BBK** running time. FRIES/*EWAK** was faster than *BBK** 92% of the time with an average improvement of 62% over *BBK** and a maximum improvement of 2.2 orders of magnitude. This improvement was evident in (A) since the blue dots (FRIES/*EWAK** times) fall mostly below the yellow dots (*BBK** times). (B) A plot of the number of conformations minimized (y-axis is on a log scale) for 661 partition function calculations from 161 design examples. The number of conformations minimized by *EWAK** (blue dots) was less than the number of conformations minimized by *BBK** (yellow dots) in 68% of these cases, as is evidenced by the blue dots landing mostly below the yellow dots. In the best case, *EWAK** decreased the number of conformations by 1.1 orders of magnitude. The average percent reduction in the number of minimized conformations was 27%. (C) Each dot represents a calculated partition function. Yellow dots are partition functions limited to within a 1.0 kcal/mol window of the GMEC, red dots are partition functions limited to a 3.0 kcal/mol window of the GMEC, and green dots are partition functions limited to within a 5.0 kcal/mol window of the GMEC. These dots are plotted according to the number of minimized conformations required for each corresponding *BBK** partition function calculation. The solid black line represents the number of *BBK** minimized conformations, so dots that fall below the black line represent examples that required fewer minimized conformations than with *BBK**. As they approach the 5.0 kcal/mol window, the dots begin to converge with the *BBK** line. However, as the number of *BBK** minimized conformations rises beyond $\sim 10^4$, even the green dots drop below the *BBK** line.

*BBK** with a more accurate, stringent ε value of 0.1 and using *EWAK** with varied energy windows: 1.0 kcal/mol, 3.0 kcal/mol, and 5.0 kcal/mol. We examined the number of conformations minimized for each complex partition function calculation across the examples. When using 1.0 kcal/mol, *EWAK** minimized up to 1.7 orders of magnitude fewer conformations (Fig 4, Panel C for more details). Despite this decrease in the number of included conformations, *EWAK** reported accurate K^* scores. The largest difference in scores between *BBK** and *EWAK** was 0.3 orders of magnitude. The accuracy of *EWAK** is explored further in the Section entitled "FRIES/*EWAK** retrospectively predicted the effect mutations in c-Raf-RBD have on binding to KRas."

Computational redesign of the c-Raf-RBD:KRas protein-protein interface

We previously showed, by investigating 58 mutations across 4 protein systems, that OSPREY can accurately predict the effect of mutations on PPI binding [1]. Herein, we

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

tested the biological accuracy of the new modules FRIES and *EWAK** after adding them to OSPREY by applying them to a particular system of significant biomedical and pharmacological importance: c-Raf-RBD in complex with KRas. The c-Raf Ras-binding domain (c-Raf-RBD) is a small self-folding domain that does not include the kinase signaling domains normally present in c-Raf. The c-Raf-RBD normally binds to KRas when KRas is GTP-bound (KRas$^{\mathrm{GTP}}$). A c-Raf-RBD variant that has high affinity for KRas$^{\mathrm{GTP}}$ could be an important first step toward discovering a tool that disrupts the KRas:effector interaction. Despite the recent successes with inhibitors targeting mutant KRas(G12C) by trapping it in the inactive GDP-bound state [57–62] and their recent move to clinical trials [63], these inhibitors are susceptible to resistance in the form of up-regulation of guanine nucleotide exchange factors (GEFs) and nucleotide exchange [60] which both push KRas to remain in its GTP-bound state. An inhibitor of the interaction between KRas$^{\mathrm{GTP}}$ and its effectors is hypothesized to have the advantage of not being susceptible to these mechanisms of resistance because it would directly interrupt KRas signaling. Hence, to further verify the accuracy and utility of FRIES/*EWAK**, we focused on this important PPI between KRas$^{\mathrm{GTP}}$ and one of its many effectors, c-Raf. First, in the Section entitled "FRIES/*EWAK** retrospectively predicted the effect mutations in c-Raf-RBD have on binding to KRas," we retrospectively investigated previously reported mutations in the c-Raf-RBD [49, 64, 65] and how they affect the binding of c-Raf-RBD to KRas. This retrospective study lays the groundwork for the prospective study we present that investigates novel mutations. So, following the retrospective study, we computationally redesigned the PPI using FRIES/*EWAK** in search of new c-Raf-RBD variants with improved affinity for KRas$^{\mathrm{GTP}}$ (see the Section entitled "Prospective redesign of the c-Raf-RBD:KRas protein-protein interface toward improved binding" for details). To perform these computational designs, we first made a homology model of c-Raf-RBD bound to KRas$^{\mathrm{GTP}}$ (see S1 Text for details).

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

FRIES/$EWAK^*$ retrospectively predicted the effect mutations in c-Raf-RBD have on binding to KRas

Each previously reported c-Raf-RBD variant [49, 64, 65] was tested computationally using FRIES/$EWAK^*$ by calculating a K^* score, a computational approximation of K_a, for each variant along with its corresponding wild-type sequence. A percent change in binding was then calculated by comparing the variant's K^* score to the corresponding wild-type sequence's K^* score. The $\log_{10}$ of this value was then calculated and normalized to the wild-type by subtracting 2. A similar procedure was completed using the reported experimental data in order to easily compare the computationally predicted effect with the experimentally measured effect. The resulting value, called Δb, represents the change in binding. If a variant has a Δb less than 0, it is predicted to decrease binding. If a variant has a Δb greater than 0, it is predicted to increase binding. Δb values that are roughly equivalent to 0 indicate variants that have little to no effect on binding since the wild-type sequence was normalized to 0. The Δb values for the 41 computationally tested variants were plotted and compared to experimental values in Fig 5.

Fig 5. Predicting the effect of mutations in c-Raf-RBD on binding with KRas. Each bar represents either the experimental (red) or computationally predicted (blue) effect each variant has on binding. The bars are sorted in increasing order of Δb value (see the Section entitled "FRIES/$EWAK^*$ retrospectively predicted the effect mutations in c-Raf-RBD have on binding to KRas") of the experimental (red) bars. If the Δb value is less than 0, binding decreases. If the Δb value is greater than 0, binding increases. If the Δb value is close to 0, the effect is neutral. Quantitative values of K^* tend to overestimate the biological effects of mutations (leading to the much larger blue bars) due to the limited nature of the input model compared to a biologically accurate representation. However, K^* in general does a good job ranking variants, as can be seen here in Fig 6, in [1], and in [38]. Out of the 41 variants listed on the x-axis, only 3 were predicted incorrectly (marked with black asterisks) by $EWAK^*$. In terms of accuracy, BBK^* performed very similarly to $EWAK^*$ (data not shown), however, in 2 cases (marked with green boxes), BBK^* ran out of memory and was unable to calculate a score. BBK^* also did not return values for the 2 variants marked with orange boxes. The variants marked with purple dots were tested in [49] experimentally – not computationally – and decreased binding of c-Raf-RBD to KRas^{GTP} was observed, which $EWAK^*$ was able to predict correctly. The two variants marked with yellow triangles were computationally predicted in [49] to improve binding of c-Raf-RBD to KRas^{GTP}. However, the experimental validation in [49] showed that these variants exhibit decreased binding, which $EWAK^*$ accurately predicted.

Out of the 41 variants tested (see S2 Table), $EWAK^*$ predicted the

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

experimentally-reported effect (increased vs. decreased binding) correctly in 38 cases. The three designs where the effect was predicted incorrectly are marked with a star in Fig 5. To make these predictions, the corresponding computational designs ranged in size from single point mutations up to 6 simultaneous mutations. Results are outlined in Fig 5. Furthermore, the Spearman's ρ value – a measure of the correlation between two sets of rankings – when comparing the experimental data to the computational predictions is 0.81. This ρ value indicates that not only can $EWAK^*$ correctly predict the effect of a particular set of mutations, but that $EWAK^*$ also does a good job ranking the variants in order according to change in binding upon mutation (Fig 6). This value is very similar to Spearman's ρ values for other PPI systems when using OSPREY [1].

Fig 6. Comparing the computational $\boldsymbol{EWAK^*}$ ranking with the experimental ranking for 41 c-Raf-RBD variants binding to KRas. Each green dot represents a variant of c-Raf-RBD and is plotted according to the experimental ranking along with the corresponding computational ranking of its binding to KRas. A least squares fit line is shown in gray. Calculating the Pearson correlation coefficient between the two sets of rankings yields a Spearman's ρ of 0.81.

BBK^* produced similarly accurate results, but took up to 10 times longer and failed to produce results in 4 cases. In particular, in 2 cases (marked in green in Fig 5), BBK^* ran out of memory. These cases serve as examples of large designs where $EWAK^*$ outperforms BBK^*. In the 2 other cases (marked in orange in Fig 5), BBK^* failed to return a result for the requested sequence in the top 5 reported sequences. This illustrates how $EWAK^*$ and FRIES are particularly helpful when performing larger designs that contain more simultaneous mutations and more flexible residues.

Finally, we compared our predictions to the interesting biological predictions in [49]. It is unclear how many mutants were computationally evaluated, but the authors do report computational predictions for 6 point mutations. Of those, point mutants R67L, N71R, and V88I were predicted to improve the intermolecular interactions between c-Raf-RBD and KRas^{GTP}. However, experiments found that R67L and V88I actually reduced the binding of c-Raf-RBD to KRas^{GTP} [49, 64]. In contrast to [49], $EWAK^*$ accurately predicted that these mutations decrease binding of c-Raf-RBD to KRas^{GTP}. For a more detailed view of one of these designs, V88I, see Fig 7. Additionally, a number of mutations were combined and experimentally tested in [49]. Unfortunately, none of these variants improved binding to either KRas^{GTP} or KRas^{GDP}, which

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

FRIES/$EWAK^*$ correctly predicted computationally (Fig 5). In [49], the authors do not present any computational predictions for these combined variants, but our results show that a computational prediction using OSPREY's $EWAK^*$ would have saved the time and resources taken to experimentally test these variants.

Fig 7. Redesign of c-Raf-RBD residue position 88 from valine to isoleucine. The left-hand side shows c-Raf-RBD (yellow) in complex with KRas (pink). Panels (A-D) zoom in on one particular design at residue position 88 and are rotated 180°. Residue position 88 has a valine in the native, wild-type sequence (panels A & C) which was redesigned to an isoleucine (panels B & D). A mutation to isoleucine at this position was computationally predicted by $EWAK^*$ to decrease the binding of c-Raf-RBD to KRas^{GTP}. This was experimentally validated in [49], where the authors incorrectly computationally predicted the effect of this particular mutation on the binding of c-Raf-RBD to KRas^{GTP}. (A) The wild-type residue (valine) is shown in green with dots that indicate molecular interactions [66] with the surrounding residues (residues allowed to be flexible in the design are shown as lines). (B) The mutant residue (isoleucine) is shown in blue with dots that indicate molecular interactions [66] with the surrounding residues (residues allowed to be flexible in the design are shown as lines). Contacts made by the wild-type valine residue (circled dots in (A)) were lost upon mutation to isoleucine (circled space in (B)). (C & D) A set of 10 low-energy conformations that were included in the corresponding partition function calculation are shown for the wild-type (green) and the variant (blue).

Prospective redesign of the c-Raf-RBD:KRas protein-protein interface toward improved binding

The ability to accurately predict the effect mutations have on the binding of c-Raf-RBD to KRas^{GTP} (see the Section entitled "FRIES/$EWAK^*$ retrospectively predicted the effect mutations in c-Raf-RBD have on binding to KRas") gave us confidence in the $EWAK^*$ algorithm's ability to predict new mutations in this interface toward a c-Raf-RBD variant that exhibits an even higher affinity for KRas^{GTP} than previously reported variants which focused on targeting KRas^{GDP} [49]. Therefore, to do a prospective study, we computationally redesigned 14 positions in c-Raf-RBD in the c-Raf-RBD:KRas PPI to identify promising mutations. After extending OSPREY to include FRIES and $EWAK^*$, 14 different designs were completed where each design included 1 mutable position that was allowed to mutate to all amino acid types except for proline. Each design also included a set of surrounding flexible residues within roughly 4 Å of the mutable residue. These designs were run using FRIES and $EWAK^*$ and included continuous flexibility [12–15]. FRIES was first used to limit each design to

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

only the most favorable sequences (as described in the Section entitled "Fast Removal of Inadequately Energied Sequences (FRIES)") and then $EWAK^{*}$ was used to estimate the K^{*} scores (as described in the Section entitled "Energy Window Approximation to K* ($EWAK^{*}$)"). We report the upper and lower bounds on the $EWAK^{*}$ score for each design in Table 1 (also see S3 Table), where the listed sequences are those that were not pruned during the FRIES step. From these results, the predicted binding effect (increased vs. decreased) was determined based on comparing each variant's K^{*} score to its corresponding wild-type K^{*} score. We then selected 5 novel point mutations – that to our knowledge are not reported in any existing literature – for experimental validation (Table 1). It is worth noting that these 5 point mutations were selected out of an initial 294 possible mutations. We limited our experimental validation to only these 5 new mutations and 2 previously reported mutations. This greatly reduced the amount of resources necessary for experimental validation compared to testing all 294 possibilities. These mutations were selected based on having a promising K^{*} score and through examining structures calculated by $EWAK^{*}$. Of the mutations selected, T57M was selected to act as a variant that was computationally predicted to be comparable to wild-type. This variant was included to further verify the accuracy of OSPREY's predictions. On the other hand, some of OSPREY's top predictions were excluded, for instance, T57R (included in S3 Table) was not selected for experimental testing because it has an unsatisfied hydrogen bond as evidenced in the structures calculated by OSPREY. Therefore, we do not believe that the score accurately represents the effect the mutation will have. Other excluded top predictions (see S3 Table) displayed similar characteristics or have been reported and tested previously [49, 64, 65].

Experimental validation of mutations in the c-Raf-RBD:KRas protein-protein interface

The mutations selected (highlighted in Table 1) from computational design were experimentally validated using a bio-layer interferometry (BLI) assay. Results from an initial single-concentration BLI screen (Fig 8) suggested that, contrary to the computational predictions, the T57K and V88F variants decrease binding, whereas the T57M and K87Y mutations both have a roughly neutral effect on binding, which is

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1. Computational predictions by OSPREY/FRIES/$EWAK^*$ that were selected for experimental validation. Each row of the table shows the results of the redesign of a residue position in c-Raf-RBD in the c-Raf-RBD:KRas PPI that were also selected for experimental validation (all of the computational results are listed in S3 Table). The table contains the values for upper and lower bounds on $\log(K^*)$ values (the calculation of these bounds is described in detail in [32]). Mutations highlighted in yellow, blue, and pink were selected for experimental testing and validation. The two residues highlighted in blue are the best previously discovered [49] mutations that improve binding (independently and additively) and are included in our tightest binding variant, c-Raf-RBD(RKY) (Figs 9, 8, and 10). The variants highlighted in yellow are, to the best of our knowledge, never-before-tested variants that are predicted to increase the binding of c-Raf-RBD to KRas^{GTP}. The variant highlighted in pink was selected for experimental testing to act as a mutation predicted to be comparable to wild-type to test how accurately OSPREY predicted the effects of these mutations.

Mutation	Lower Bound $\log(K^*)$	Upper Bound $\log(K^*)$
T57M	3.43	3.46
T57	3.82	3.92
T57K	5.01	5.07
N71	7.25	7.49
N71R	9.66	10.10
A85	26.3	26.9
A85K	30.7	32.3
K87	13.4	14.1
K87Y	14.1	14.2
V88	16.5	16.6
V88Y	17.3	17.6
V88F	18.0	18.2

consistent with the computational predictions. The final computationally predicted point mutant, V88Y, improves binding a comparable amount to the improvement seen with A85K or N71R, two previously reported variants also predicted by OSPREY and experimentally tested herein that improve binding. With the discovery of this new variant containing the point mutant V88Y (referred to herein as c-Raf-RBD(Y)) the next natural step was to combine it with the mutations found in the best reported variant, N71R and A85K (referred to herein as c-Raf-RBD(RK)). Therefore, we also included the double-mutant, c-Raf-RBD(RK), and the new triple-mutant – which contains N71R, A85K, and V88Y and is referred to herein as c-Raf-RBD(RKY) – in our initial BLI screen. Additionally, the c-Raf-RBD(RKY) variant was computationally predicted by FRIES/$EWAK^*$ to bind to KRas^{GTP} more tightly than the previous best known binder, c-Raf-RBD(RK) (results are detailed in Fig 9). Given the promising screening and computational results for the c-Raf-RBD(Y) and c-Raf-RBD(RKY) variants, we measured K_d values for each variant by titrating the analyte over the

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig 8. Single-concentration experimental screening of c-Raf-RBD variants binding to KRas using BLI. (A) Binding curves are shown for each variant (labeled on the plot) tested at a concentration of 250 nM. The colors and labels in panel (A) correspond to those in panel (B). (B) Plot of estimated K_d values for each tested variant from a single-concentration screen (plotted in panel (A)). The c-Raf-RBD(RKY) variant (in green on the far left) is a novel, newly discovered variant of c-Raf-RBD. Top variants were further validated and had their K_d values calculated more accurately using BLI titration experiments (Fig 10).

Fig 9. Computational predictions in the protein-protein interface of the c-Raf-RBD:KRas complex for c-Raf-RBD(RK) and the novel variant c-Raf-RBD(RKY). Shown on the left is only the relevant protein-protein interface between c-Raf-RBD and KRas. Each panel zooms in on this interface and details a different c-Raf-RBD variant and its corresponding computational predictions. The upper and lower bounds on the $\log(K^*)$ score for each design variant (wild-type, c-Raf-RBD(RK), and c-Raf-RBD(RKY)) are given in the bottom table. These computational predictions correspond with and are supported by the experimental results presented in the Section entitled "Experimental validation of mutations in the c-Raf-RBD:KRas protein-protein interface." Panels (A) and (B) show the wild-type sequence, panels (C) and (D) show the variant c-Raf-RBD(RK), and panels (E) and (F) show the novel computationally predicted variant c-Raf-RBD(RKY). Panels (A), (C), and (E) show the wild-type, c-Raf-RBD(RK), and c-Raf-RBD(RKY), respectively, along with probe dots [66] that represent the molecular interactions within each structure calculated by OSPREY. These probe dots were selected to only show interactions between the residues included in the computational designs (shown as green and blue lines) with their surrounding residues. Panels (B), (D), and (F) show 10 low-energy structures from each conformational ensemble calculated by OSPREY/*EWAK**. Panel (G) shows a zoomed-in overlay of the wild-type variant with the c-Raf-RBD variant that includes only the V88Y mutation. Purple arrows indicate the change in positioning of the lysine at residue position 84 upon mutation of residue position 88 from valine to tyrosine. When valine is present at position 88, the lysine residue (shown in green) primarily hydrogen bonds with an aspartate (labeled) in KRas. When valine is mutated to tyrosine (shown in cyan), the lysine at position 84 moves to make room for the tyrosine and positions itself to hydrogen bond with both the aspartate and the glutamate (labeled) in KRas.

ligand in a BLI-based assay (Fig 10). Excitingly, c-Raf-RBD(RKY) is calculated by the data from the BLI assay (Figs 8 and 10) to bind KRas$^{\mathrm{GTP}}$ roughly 5 times better than the previous best known binder, c-Raf-RBD(RK), and approximately 36 times better than wild-type c-Raf-RBD. Given how heavily studied the KRas system is, with several reported mutational and structural studies [49, 64, 64, 65, 65, 67, 67–73, 73, 74, 74–79], this is a discovery of some significance.

Bio-layer interferometry (BLI) assay Binding of wild-type and variants of c-Raf-RBD were experimentally measured using a bio-layer interferometry (BLI) assay. Each variant of c-Raf-RBD was expressed and purified (S2 Text) with cysteine residues at positions 81 and 96 substituted for isoleucine and methionine, respectively. These

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig 10. BLI titration experiments to calculate K_d values for select c-Raf-RBD variants. The plots shown here are representative and the data from replicate experiments are presented in S4 Table. Each plot shows the data collected from a titration BLI experiment where the concentration of the c-Raf-RBD variant is incrementally increased. The concentrations for the wild-type variant were 10, 50, 150, 200, and 300 nM. The concentrations for all of the other variants were 10, 25, 25, 75, 75, 125, and 200 nM. Repeat intermediate concentrations were used as loading controls. These curves were then fit using a mass transport model within the Octet Data Analysis HT software provided by FortéBio in order to calculate the K_d value for each variant's binding to KRas. The values in the table here (bottom right) are average K_d values shown with 2 standard deviations calculated from replicate experiments (S4 Table). The values presented here for Wild-Type, A85K, and c-Raf-RBD(RK) agree well with previously reported K_d values [49]. The best binding variant, c-Raf-RBD(RKY), binds to KRas about 5 times better than the previous tightest-known binder, c-Raf-RBD(RK), and about 36 times better than wild-type c-Raf-RBD.

mutations were previously reported to minimally affect on the stability of c-Raf-RBD [73] and their substitution allows for the use of the c-Raf-RBD constructs in other assays (not mentioned herein). Additionally, we do not believe these residue substitutions have a large effect since the K_d values determined herein align with previously reported K_d values [49] (Fig 10). KRas was expressed and purified (S3 Text) with a poly-histidine protein tag (His-tag) and loaded with a non-hydrolyzable GTP analog, GppNHp. KRas was also made to include a substitution at position 118 from cysteine to serine in order to increase expression and stability [80]. Ni-NTA tips were then used to perform the BLI experiments to determine binding of the c-Raf-RBD variants to KRas$^{\mathrm{GppNHp}}$ (results are shown in Figs 8 and 10 and S4 Table). All experiments were carried out in 30 mM phosphate pH 7.4, 327 mM NaCl, 2.7 mM KCl, 5 mM $MgCl_2$, 1.5 mM TCEP, 0.1% BSA, and 0.02% Tween-20 + Kathon at 25°C with 1000 RPM shaking and a KRas loading concentration of 20 µg/ml. Each curve presented (Figs 8 and 10) was fit using the built-in mass transport model within the Octet Data Analysis HT software provided by FortéBio. We only accepted fits with a sum of square deviations χ^2 less than 1 (FortéBio recommends a value less than 3) and a coefficient of determination R^2 greater than 0.98.

Discussion

FRIES and $EWAK^*$ are new, provable algorithms for more efficient ensemble-based computational protein design. Efficiency and efficacy were tested and shown across a

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

total of 2,826 different design problems. An implementation of FRIES/*EWAK** is available in the open-source protein design software OSPREY [1] and all of the data has been made available (see "Data availability statement"). FRIES/*EWAK** in combination achieved a significant runtime improvement over the previous state-of-the-art, *BBK**, with runtimes up to 2 orders of magnitude faster. *EWAK** also limits the number of minimized conformations used in each K^* score approximation by up to about 2 orders of magnitude while maintaining provable guarantees (see the Section entitled "Energy Window Approximation to K* (*EWAK**)"). FRIES alone is capable of reducing the input sequence space while provably keeping all of the most energetically favorable sequences (see the Section entitled "Fast Removal of Inadequately Energied Sequences (FRIES)"), decreasing the size of the sequence space by more than 2 orders of magnitude, and leading to more efficient design given the smaller search space.

To further validate OSPREY with FRIES/*EWAK**, we applied these algorithms to a biomedically significant design problem: the c-Raf-RBD:KRas PPI. First, we performed a series of retrospective designs where FRIES/*EWAK** accurately predicted how a variety of mutations affect the binding of c-Raf-RBD to KRas$^{\mathrm{GTP}}$ that previous computational methods had failed to accurately predict [49]. This success supports the use of OSPREY and FRIES/*EWAK** to evaluate the effect mutations in the protein-protein interface of c-Raf-RBD:KRas have on binding (more, similar successes of the K^* algorithm are presented and discussed in [1]). FRIES/*EWAK** also prospectively predicted the effect of new mutations in the c-Raf-RBD:KRas PPI and discovered a novel c-Raf-RBD mutation V88Y with improved affinity for KRas. We went on to combine this new mutation with two previously reported mutations, N71R and A85K [49], to create c-Raf-RBD(RKY), an even stronger binding c-Raf-RBD variant, which FRIES/*EWAK** accurately predicted. We biochemically screened top predicted variants using an initial bio-layer interferometry (BLI) single-concentration assay. Only a promising subset of the computationally predicted and initially screened variants were then evaluated using a BLI titration assay to calculate K_d values for individual c-Raf-RBD variants. We determined that c-Raf-RBD(RKY) binds to KRas$^{\mathrm{GTP}}$ roughly 36 times more tightly than wild-type c-Raf-RBD, making it the tightest known c-Raf-RBD variant binding partner of KRas$^{\mathrm{GTP}}$.

Given that numerous groups have explored this protein-protein

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

interaction [64, 65, 67–77] and performed mutagenesis on c-Raf-RBD either, through rational means [64, 67, 74, 78], computational methods [49, 65] or high-throughput evolutionary methods [73, 79] and that none identified V88Y, this discovery validates our computational approach and the use of computational algorithms such as FRIES and *EWAK** to redesign protein-protein interfaces toward improved binding. Finally, previous mutations that enhanced the affinity of c-Raf-RBD binding to KRas did so by supercharging c-Raf-RBD [49, 64, 65]. In contrast, our mutation V88Y introduces a novel, aromatic residue. The discovery that such a mutation can improve the binding of c-Raf-RBD to KRas$^{\mathrm{GTP}}$ is of considerable significance. These new c-Raf-RBD variants serve as an important step toward better understanding the KRas:effector interface and eventually developing successful therapeutics to directly target and block the aberrant behavior of mutant KRas.

Supporting information

S1 Text. Homology model of c-Raf-RBD in complex with KRas.

S2 Text. Details of the expression and purification of c-Raf-RBD variants.

S3 Text. Details of the expression and purification of KRas.

S1 Table. Protein structures used in computational experiments as described in the Section entitled "Computational materials and methods." Each protein structure has its PDB ID listed along with its molecule names as presented in the Protein Database entry for each structure. Individual designs are not listed or described here, but the necessary code and data is provided for the interested reader (see Data availability statement).

S2 Table. Experimental and computational percent change in binding and rankings. For each listed variant, we give the experimental percent change in binding relative to wild-type as reported in [64] and as calculated from reported binding values in [65] and [49], the *EWAK** computationally predicted percent change in binding (as described in the Section entitled "FRIES/*EWAK** retrospectively predicted the effect

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mutations in c-Raf-RBD have on binding to KRas") and the rankings that correspond to these values. The rankings have a Pearson correlation of 0.81.

S3 Table. Table of computational predictions for point mutants in c-Raf-RBD. Each section of the table shows the results of the redesign of a residue position in c-Raf-RBD in the c-Raf-RBD:KRas PPI in order of increasing upper bound on $\log(K^*)$. The table contains the values for upper and lower bounds on $\log(K^*)$ values (these bounds are described in detail in [32]). *Design results for the wild-type amino acid identity for each position. †Mutations that were selected for experimental testing and validation.

S4 Table. K_d values for each tested variant for all replicates of BLI titration experiments. For each listed variant, we give the dissociation constant K_d for each BLI titration experiment calculated from the fit done using the built-in mass transport model within the Octet Data Analysis HT software provided by ForteBio. We only accepted fits with a sum of square deviations χ^2 less than 1 (ForteBio recommends a value less than 3) and a coefficient of determination R^2 greater than 0.98. Presented in the table in Fig 10 are averages of these K_d values.

Acknowledgments

We thank Rachel Kimbrough, Chanelle Simmons, Catherine Ehrhart, and Kelly Huynh for assistance with experiments and biochemical validation, Ben Fenton and Michael Kennedy for the initial KRas plasmids, and Paul Modrich for the Rosetta 2(DE3) cell lines. We also thank Terrence Oas and all members of the lab for helpful discussions.

References

1. Hallen MA, Martin JW, Ojewole A, Jou JD, Lowegard AU, Frenkel MS, et al. OSPREY 3.0: Open-source protein redesign for you, with powerful new features. Journal of Computational Chemistry. 2018;39(30):2494–2507. doi:10.1002/jcc.25522.

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2. Ojewole A, Lowegard A, Gainza P, Reeve SM, Georgiev I, Anderson AC, et al. OSPREY predicts resistance mutations using positive and negative computational protein design. In: Computational Protein Design. Springer; 2017. p. 291–306.

3. Gainza P, Roberts KE, Georgiev I, Lilien RH, Keedy DA, Chen CY, et al. OSPREY: protein design with ensembles, flexibility, and provable algorithms. Methods Enzymol. 2013;523:87–107. doi:10.1016/B978-0-12-394292-0.00005-9.

4. Donald BR. Algorithms in Structural Molecular Biology. Cambridge, MA: MIT Press; 2011.

5. Gainza P, Nisonoff HM, Donald BR. Algorithms for protein design. Current Opinion in Structural Biology. 2016;39:16–26.

6. Simoncini D, Allouche D, de Givry S, Delmas C, Barbe S, Schiex T. Guaranteed Discrete Energy Optimization on Large Protein Design Problems. J Chem Theory Comput. 2015;11(12):5980–9. doi:10.1021/acs.jctc.5b00594.

7. Hallen MA, Donald BR. Protein Design by Algorithm. arXiv preprint arXiv:180606064. 2018;.

8. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A. 2000;97(19):10383–8.

9. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545–74. doi:10.1016/B978-0-12-381270-4.00019-6.

10. Lee C, Subbiah S. Prediction of protein side-chain conformation by packing optimization. Journal of Molecular Biology. 1991;217(2):373–388.

11. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins. 2000;40(3):389–408.

12. Georgiev I, Lilien RH, Donald BR. The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for computing partition functions over molecular ensembles. J Comput Chem. 2008;29(10):1527–42. doi:10.1002/jcc.20909.

13. Gainza P, Roberts KE, Donald BR. Protein design using continuous rotamers. PLoS Comput Biol. 2012;8(1):e1002335. doi:10.1371/journal.pcbi.1002335.

14. Hallen MA, Gainza P, Donald BR. Compact Representation of Continuous Energy Surfaces for More Efficient Protein Design. J Chem Theory Comput. 2015;11(5):2292–306. doi:10.1021/ct501031m.

15. Hallen MA, Jou JD, Donald BR. LUTE (Local Unpruned Tuple Expansion): Accurate Continuously Flexible Protein Design with General Energy Functions and Rigid-rotamer-like Efficiency. Research in Computational Molecular Biology (RECOMB). 2016;9649:122–136.

16. Georgiev I, Donald BR. Dead-end elimination with backbone flexibility. Bioinformatics. 2007;23(13):i185–94. doi:10.1093/bioinformatics/btm197.

17. Georgiev I, Keedy D, Richardson JS, Richardson DC, Donald BR. Algorithm for backrub motions in protein design. Bioinformatics. 2008;24(13):i196–204. doi:10.1093/bioinformatics/btn169.

18. Hallen MA, Donald BR. CATS (Coordinates of Atoms by Taylor Series): protein design with backbone flexibility in all locally feasible directions. Bioinformatics. 2017;33(14):i5–i12. doi:10.1093/bioinformatics/btx277.

19. Hallen MA, Keedy DA, Donald BR. Dead-end elimination with perturbations (DEEPer): a provable protein design algorithm with continuous sidechain and backbone flexibility. Proteins. 2013;81(1):18–39. doi:10.1002/prot.24150.

20. Tzeng SR, Kalodimos CG. Protein activity regulation by conformational entropy. Nature. 2012;488(7410):236.

21. Gilson MK, Given JA, Bush BL, McCammon JA. The statistical-thermodynamic basis for computation of binding affinities: a critical review. Biophys J. 1997;72(3):1047–69. doi:10.1016/S0006-3495(97)78756-3.

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

22. Chen CY, Georgiev I, Anderson AC, Donald BR. Computational structure-based redesign of enzyme activity. Proc Natl Acad Sci U S A. 2009;106(10):3764–9. doi:10.1073/pnas.0900266106.

23. Sciretti D, Bruscolini P, Pelizzola A, Pretti M, Jaramillo A. Computational protein design with side-chain conformational entropy. Proteins. 2009;74(1):176–91. doi:10.1002/prot.22145.

24. Georgiev I, Lilien RH, Donald BR. Improved Pruning algorithms and Divide-and-Conquer strategies for Dead-End Elimination, with application to protein design. Bioinformatics. 2006;22(14):e174–83. doi:10.1093/bioinformatics/btl220.

25. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science. 1997;278(5335):82–7.

26. Leach AR, Lemon AP. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. Proteins. 1998;33(2):227–39.

27. Traoré S, Allouche D, André I, de Givry S, Katsirelos G, Schiex T, et al. A new framework for computational protein design through cost function network optimization. Bioinformatics. 2013;29(17):2129–36. doi:10.1093/bioinformatics/btt374.

28. Bernard Chazelle MS Carl Kingsford. A semidefinite programming approach to side chain positioning with new rounding strategies. NFORMS JOURNAL ON COMPUTING. 2004;16(4):380–392.

29. Lilien RH, Stevens BW, Anderson AC, Donald BR. A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the substrate specificity of the gramicidin synthetase a phenylalanine adenylation enzyme. J Comput Biol. 2005;12(6):740–61. doi:10.1089/cmb.2005.12.740.

30. Roberts KE, Cushing PR, Boisguerin P, Madden DR, Donald BR. Computational design of a PDZ domain peptide inhibitor that rescues CFTR activity. PLoS Comput Biol. 2012;8(4):e1002477. doi:10.1371/journal.pcbi.1002477.

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

31. Silver NW, King BM, Nalam MNL, Cao H, Ali A, Kiran Kumar Reddy GS, et al. Efficient Computation of Small-Molecule Configurational Binding Entropy and Free Energy Changes by Ensemble Enumeration. J Chem Theory Comput. 2013;9(11):5098–5115. doi:10.1021/ct400383v.

32. Ojewole AA, Jou JD, Fowler VG, Donald BR. In: BBK* (Branch and Bound over K*): A Provable and Efficient Ensemble-Based Algorithm to Optimize Stability and Binding Affinity over Large Sequence Spaces. Springer International Publishing; 2017. p. 157–172.

33. Viricel C, Simoncini D, Barbe S, Schiex T. Guaranteed weighted counting for affinity computation: Beyond determinism and structure. In: International Conference on Principles and Practice of Constraint Programming. Springer; 2016. p. 733–750.

34. Traoré S, Allouche D, André I, Schiex T, Barbe S. Deterministic Search Methods for Computational Protein Design. Methods Mol Biol. 2017;1529:107–123.

35. Traoré S, Roberts KE, Allouche D, Donald BR, André I, Schiex T, et al. Fast search algorithms for computational protein design. J Comput Chem. 2016;37(12):1048–58. doi:10.1002/jcc.24290.

36. Stevens BW, Lilien RH, Georgiev I, Donald BR, Anderson AC. Redesigning the PheA domain of gramicidin synthetase leads to a new understanding of the enzyme's mechanism and selectivity. Biochemistry. 2006;45(51):15495–504. doi:10.1021/bi061788m.

37. Frey KM, Georgiev I, Donald BR, Anderson AC. Predicting resistance mutations using protein design algorithms. Proc Natl Acad Sci U S A. 2010;107(31):13707–12. doi:10.1073/pnas.1002162107.

38. Reeve SM, Gainza P, Frey KM, Georgiev I, Donald BR, Anderson AC. Protein design algorithms predict viable resistance to an experimental antifolate. Proc Natl Acad Sci U S A. 2015;112(3):749–54. doi:10.1073/pnas.1411548112.

39. Gorczynski MJ, Grembecka J, Zhou Y, Kong Y, Roudaia L, Douvas MG, et al. Allosteric inhibition of the protein-protein interaction between the

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

leukemia-associated proteins Runx1 and CBFbeta. Chem Biol. 2007;14(10):1186–97. doi:10.1016/j.chembiol.2007.09.006.

40. Georgiev I, Schmidt S, Li Y, Wycuff D, Ofek G, Doria-Rose N, et al. Design of epitope-specific probes for sera analysis and antibody isolation. Retrovirology. 2012;9.

41. Georgiev IS, Rudicell RS, Saunders KO, Shi W, Kirys T, McKee K, et al. Antibodies VRC01 and 10E8 neutralize HIV-1 with high breadth and potency even with IG-framework regions substantially reverted to germline. J Immunol. 2014;192(3):1100–6. doi:10.4049/jimmunol.1302515.

42. Rudicell RS, Kwon YD, Ko SY, Pegu A, Louder MK, Georgiev IS, et al. Enhanced potency of a broadly neutralizing HIV-1 antibody in vitro improves protection against lentiviral infection in vivo. J Virol. 2014;88(21):12669–82. doi:10.1128/JVI.02213-14.

43. A Phase 1, Single Dose Study of the Safety and Virologic Effect of an HIV-1 Specific Broadly Neutralizing Human Monoclonal Antibody, VRC-HIVMAB080-00-AB (VRC01LS) or VRC-HIVMAB075-00-AB (VRC07-523LS), Administered Intravenously to HIV-Infected Adults. ClinicalTrials.gov Identifier: NCT02840474. NIAID And National Institutes of Health Clinical Center. September (2018). `https://clinicaltrials.gov/ct2/show/NCT02840474`;.

44. Evaluating the Safety and Serum Concentrations of a Human Monoclonal Antibody, VRC-HIVMAB075-00-AB (VRC07-523LS), Administered in Multiple Doses and Routes to Healthy, HIV-uninfected Adults. ClinicalTrials.gov Identifier: NCT03387150. NIAID And National Institutes of Health Clinical Center. September (2018). `https://clinicaltrials.gov/ct2/show/NCT03387150`;.

45. VRC 610: Phase I Safety and Pharmacokinetics Study to Evaluate a Human Monoclonal Antibody (MAB) VRC-HIVMAB095-00-AB (10E8VLS) Administered Alone or Concurrently With MAB VRC-HIVMAB075-00-AB (VRC07-523LS) Via Subcutaneous Injection in Healthy Adults. ClinicalTrials.gov Identifier: NCT03565315. NIAID And National Institutes of Health Clinical

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Center. September (2018). https://clinicaltrials.gov/ct2/show/NCT03565315;.

46. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proceedings of the National Academy of Sciences. 2000;97(19):10383–10388.

47. Sommer R, Wagner S, Varrot A, Nycholat CM, Khaledi A, Häussler S, et al. The virulence factor LecB varies in clinical isolates: consequences for ligand binding and drug discovery. Chemical Science. 2016;7(8):4990–5001.

48. Papke B, Der CJ. Drugging RAS: Know the enemy. Science. 2017;355(6330):1158–1163.

49. Filchtinski D, Sharabi O, Rüppel A, Vetter IR, Herrmann C, Shifman JM. What makes Ras an efficient molecular switch: a computational, biophysical, and structural study of Ras-GDP interactions with mutants of Raf. Journal of molecular biology. 2010;399(3):422–435.

50. Lee J. New Monte Carlo algorithm: entropic sampling. Physical Review Letters. 1993;71(2):211.

51. Nosé S. A molecular dynamics method for simulations in the canonical ensemble. Molecular physics. 1984;52(2):255–268.

52. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;.

53. Lou Q, Dechter R, Ihler AT. Dynamic Importance Sampling for Anytime Bounds of the Partition Function. In: Advances in Neural Information Processing Systems; 2017. p. 3196–3204.

54. Roberts KE, Gainza P, Hallen MA, Donald BR. Fast gap-free enumeration of conformations and sequences for protein design. Proteins. 2015;83(10):1859–77. doi:10.1002/prot.24870.

55. Hallen MA, Donald BR. COMETS (Constrained Optimization of Multistate Energies by Tree Search): A provable and efficient protein design algorithm to

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

optimize binding affinity and specificity with respect to sequence. Journal of Computational Biology. 2016;23(5):311–321.

56. Jou JD, Holt GT, Lowegard AU, Donald BR. Minimization-Aware Recursive K^* ($MARK^*$): A Novel, Provable Algorithm that Accelerates Ensemble-Based Protein Design and Provably Approximates the Energy Landscape. In: International Conference on Research in Computational Molecular Biology. Springer; 2019. p. 101–119.

57. Ostrem JM, Peters U, Sos ML, Wells JA, Shokat KM. K-Ras (G12C) inhibitors allosterically control GTP affinity and effector interactions. Nature. 2013;503(7477):548.

58. Hunter JC, Gurbani D, Ficarro SB, Carrasco MA, Lim SM, Choi HG, et al. 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. 2014;111(24):8895–8900.

59. Lito P, Solomon M, Li LS, Hansen R, Rosen N. Allele-specific inhibitors inactivate mutant KRAS G12C by a trapping mechanism. Science. 2016;351(6273):604–608.

60. Patricelli MP, Janes MR, Li LS, Hansen R, Peters U, Kessler LV, et al. Selective inhibition of oncogenic KRAS output with small molecules targeting the inactive state. Cancer discovery. 2016;6(3):316–329.

61. Zeng M, Lu J, Li L, Feru F, Quan C, Gero TW, et al. Potent and selective covalent quinazoline inhibitors of KRAS G12C. Cell chemical biology. 2017;24(8):1005–1016.

62. Janes MR, Zhang J, Li LS, Hansen R, Peters U, Guo X, et al. Targeting KRAS mutant cancers with a covalent G12C-specific inhibitor. Cell. 2018;172(3):578–589.

63. Fakih M, O'Neil B, Price TJ, Falchook GS, Desai J, Kuo J, et al.. Phase 1 study evaluating the safety, tolerability, pharmacokinetics (PK), and efficacy of AMG

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

510, a novel small molecule KRASG12C inhibitor, in advanced solid tumors.; 2019.

64. Fridman M, Maruta H, Gonez J, Walker F, Treutlein H, Zeng J, et al. Point mutants of c-raf-1 RBD with elevated binding to v-Ha-Ras. Journal of Biological Chemistry. 2000;275(39):30363–30371.

65. Kiel C, Filchtinski D, Spoerner M, Schreiber G, Kalbitzer HR, Herrmann C. Improved binding of Raf to Ras·GDP is correlated with biological activity. Journal of Biological Chemistry. 2009;284(46):31893–31902.

66. Roberts KE. http://www.cs.duke.edu/donaldlab/software/proteinInteractionViewer/. Protein Interaction Viewer. 2012;.

67. Nassar N, Horn G, Herrmann C, Block C, Janknecht R, Wittinghofer A. Ras/Rap effector specificity determined by charge reversal. Nature Structural and Molecular Biology. 1996;3(8):723.

68. Sydor JR, Seidel RP, Goody RS, Engelhard M. Cell-free synthesis of the Ras-binding domain of c-Raf-1: binding studies to fluorescently labelled H-Ras. FEBS letters. 1999;452(3):375–378.

69. Herrmann C, Horn G, Spaargaren M, Wittinghofer A. Differential interaction of the ras family GTP-binding proteins H-Ras, Rap1A, and R-Ras with the putative effector molecules Raf kinase and Ral-guanine nucleotide exchange factor. Journal of Biological Chemistry. 1996;271(12):6794–6800.

70. Herrmann C, Martin GA, Wittinghofer A. Quantitative analysis of the complex between p21 and the ras-binding domain of the human raf-1 protein kinase. Journal of Biological Chemistry. 1995;270(7):2901–2905.

71. Lakshman B, Messing S, Schmid EM, Clogston JD, Gillette WK, Esposito D, et al. Quantitative biophysical analysis defines key components modulating recruitment of the GTPase KRAS to the plasma membrane. Journal of Biological Chemistry. 2019;294(6):2193–2207.

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

72. Block C, Janknecht R, Herrmann C, Nassar N, Wittinghofer A. Quantitative structure-activity analysis correlating Ras/Raf interaction in vitro to Raf activation in vivo. Nature structural biology. 1996;3(3):244.

73. Campbell-Valois FX, Tarassov K, Michnick S. Massive sequence perturbation of the Raf Ras binding domain reveals relationships between sequence conservation, secondary structure propensity, hydrophobic core organization and stability. Journal of molecular biology. 2006;362(1):151–171.

74. Fridman M, Walker F, Catimel B, Domagala T, Nice E, Burgess A. c-Raf-1 RBD associates with a subset of active vH-Ras. Biochemistry. 2000;39(50):15603–15611.

75. Fetics SK, Guterres H, Kearney BM, Buhrman G, Ma B, Nussinov R, et al. Allosteric effects of the oncogenic RasQ61L mutant on Raf-RBD. Structure. 2015;23(3):505–516.

76. Gorman C, Skinner RH, Skelly JV, Neidle S, Lowe PN. Equilibrium and kinetic measurements reveal rapidly reversible binding of Ras to Raf. Journal of Biological Chemistry. 1996;271(12):6713–6719.

77. Hunter JC, Manandhar A, Carrasco MA, Gurbani D, Gondi S, Westover KD. Biochemical and structural analysis of common cancer-associated KRAS mutations. Molecular cancer research. 2015;13(9):1325–1335.

78. Fridman M, Tikoo A, Varga M, Murphy A, Nur-E-Kamal M, Maruta H. The minimal fragments of c-Raf-1 and NF1 that can suppress v-Ha-Ras-induced malignant phenotype. Journal of Biological Chemistry. 1994;269(48):30105–30108.

79. Campbell-Valois FX, Tarassov K, Michnick S. Massive sequence perturbation of a small protein. Proceedings of the National Academy of Sciences. 2005;102(42):14988–14993.

80. Sun Q, Burke JP, Phan J, Burns MC, Olejniczak ET, Waterson AG, et al. Discovery of small molecules that bind to K-Ras and inhibit Sos-mediated activation. Angewandte Chemie International Edition. 2012;51(25):6140–6143.

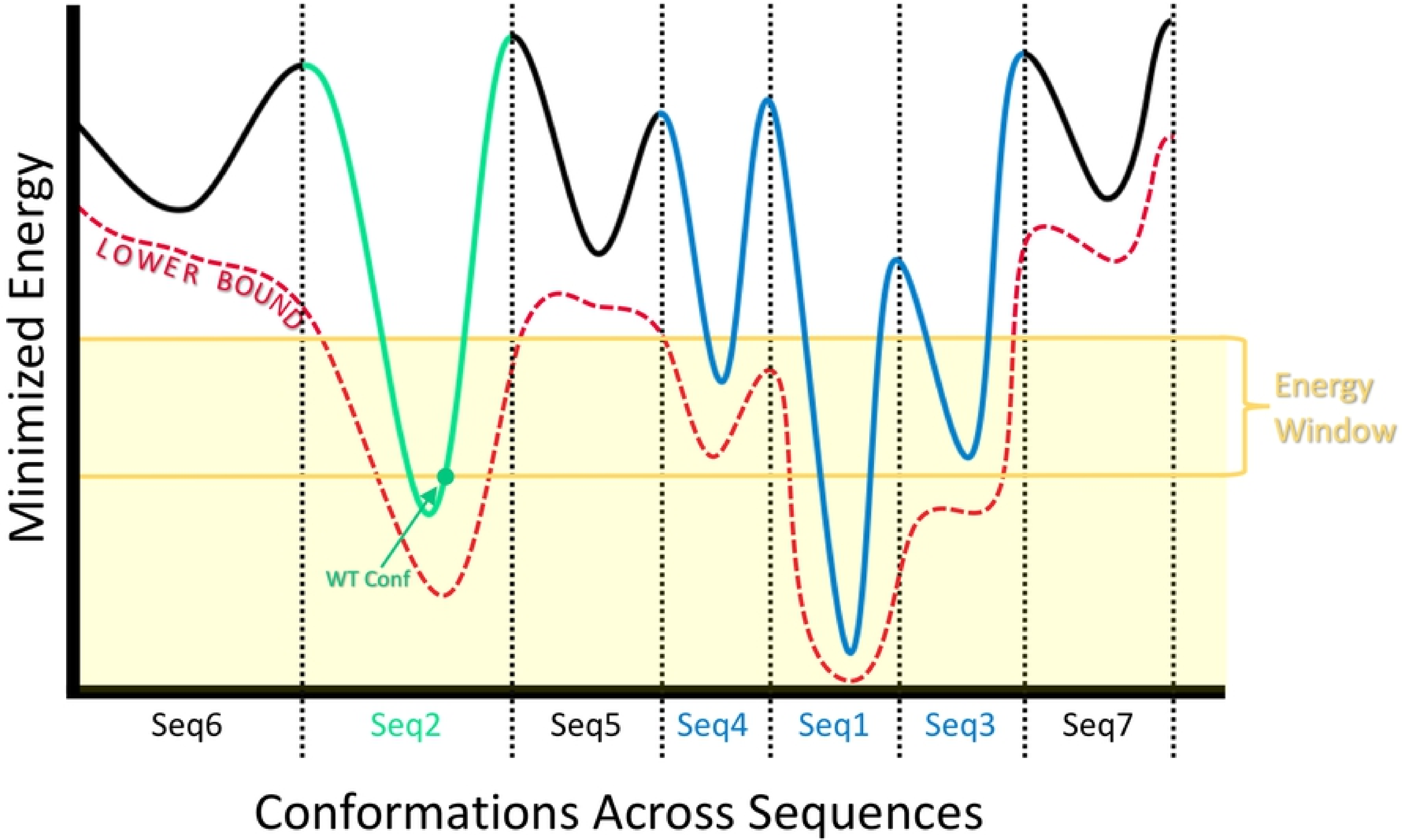

Minimized Energy
LOWER BOUND
WT Conf
Energy Window
Seq6
Seq2
Seq5
Seq4
Seq1
Seq3
Seq7
Conformations Across Sequences

Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/790949; this version posted October 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

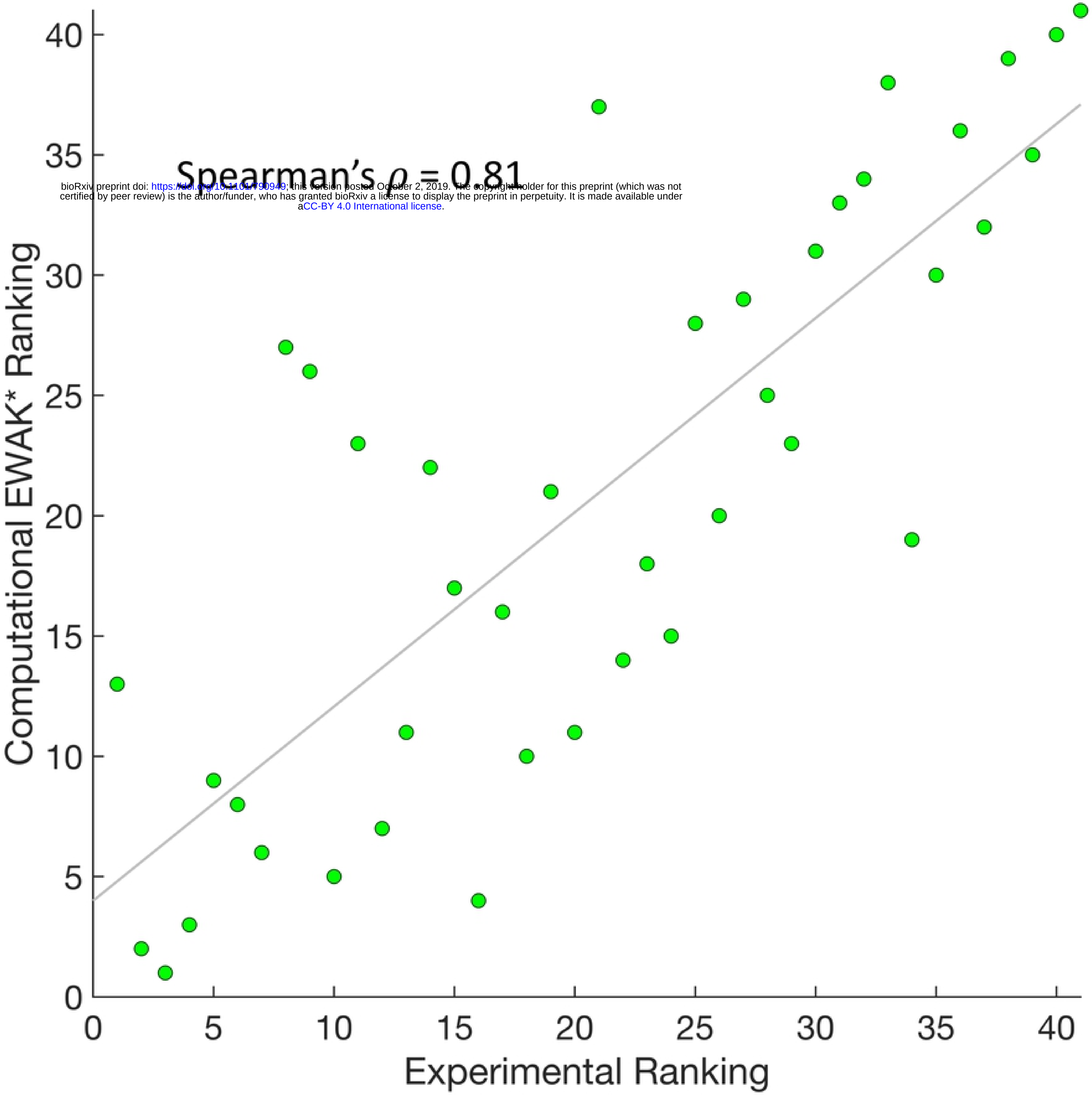


Figure 6

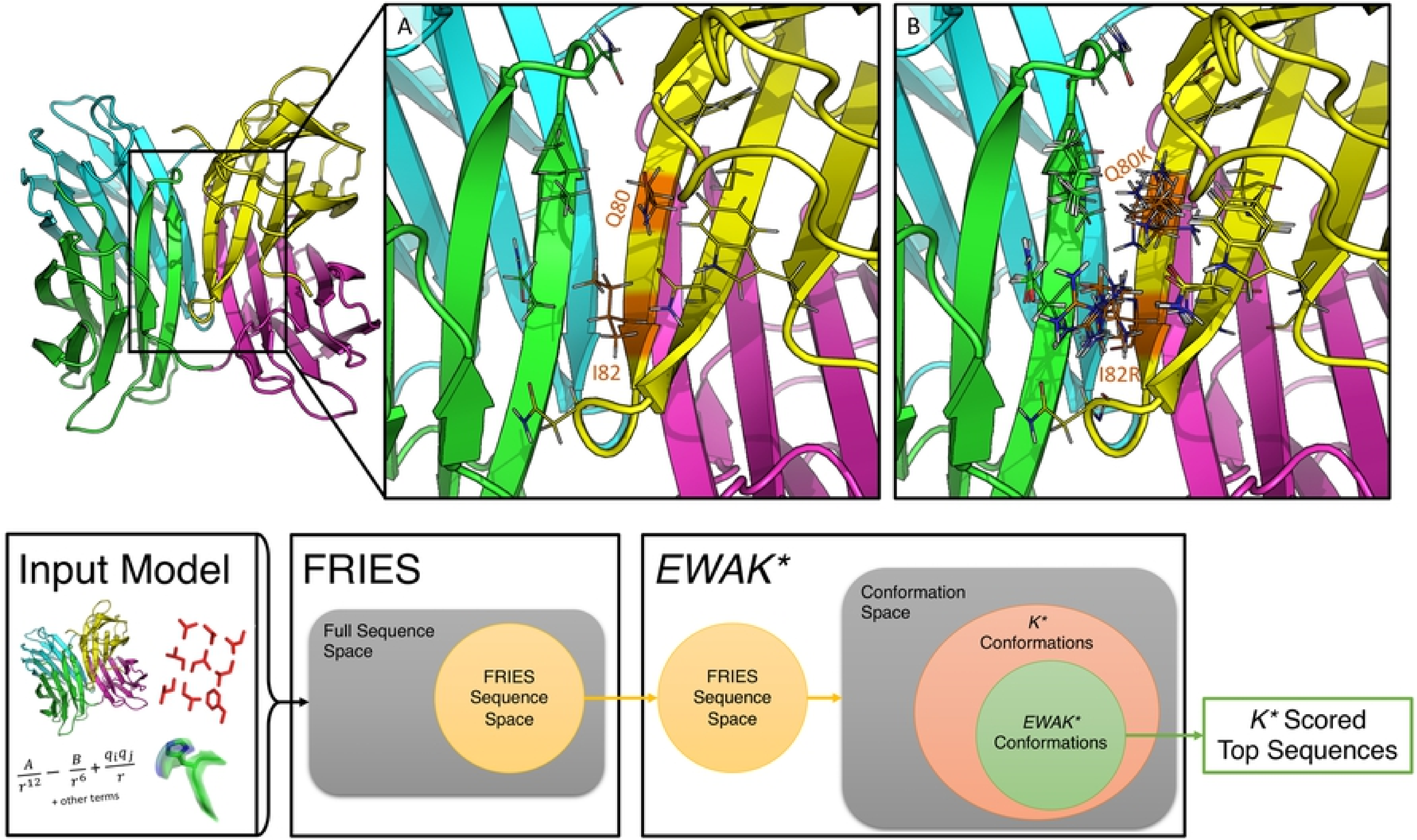

A
Q80
I82
B
Q80K
I82R
Input Model
$\frac{A}{r^{12}} - \frac{B}{r^{6}} + \frac{q_i q_j}{r}$
+ other terms
FRIES
Full Sequence Space
FRIES Sequence Space
EWAK*
FRIES Sequence Space
Conformation Space
K* Conformations
EWAK* Conformations
K* Scored Top Sequences

Figure 1

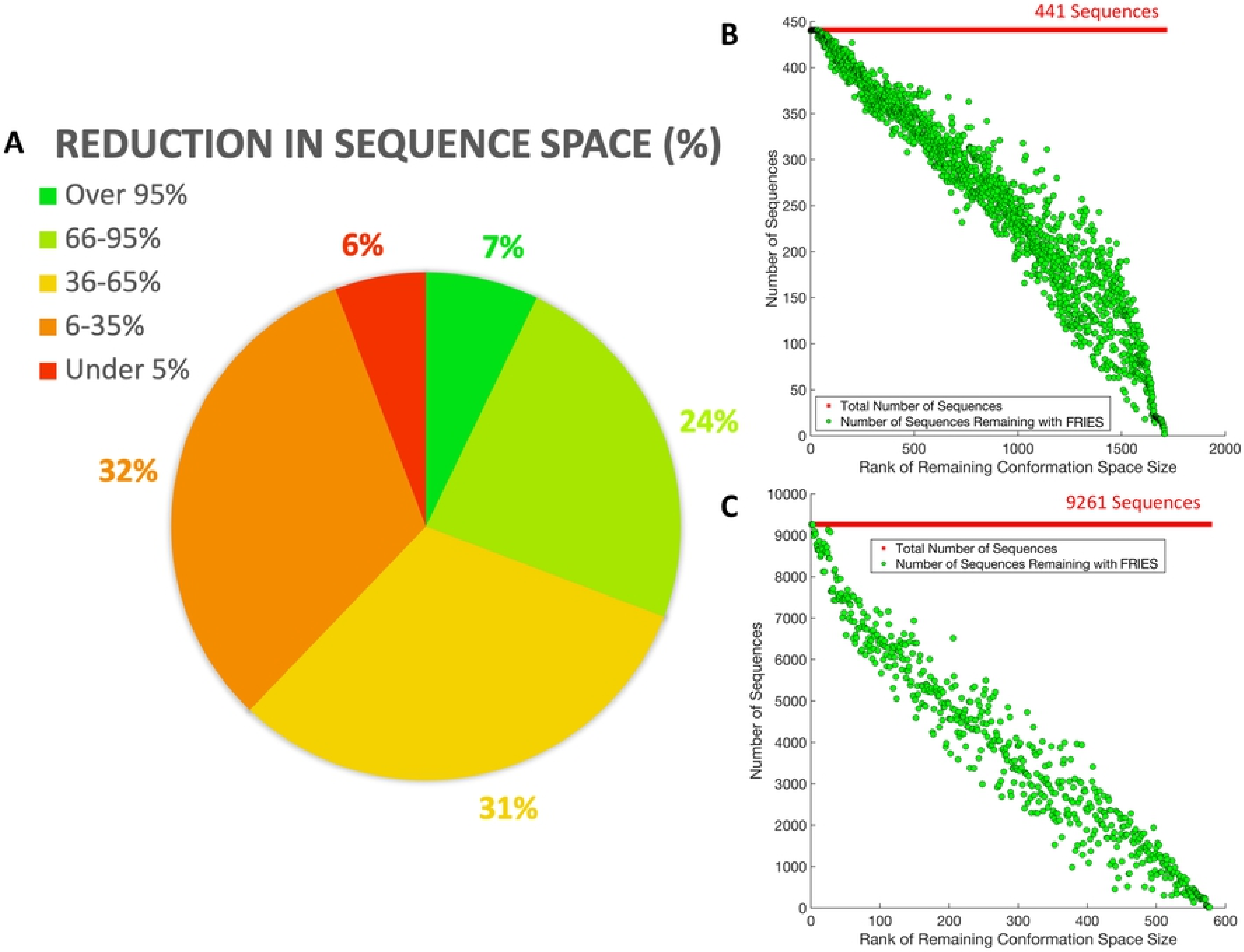

A
REDUCTION IN SEQUENCE SPACE (%)
Over 95%
66-95%
36-65%
6-35%
Under 5%
6%
7%
24%
32%
31%
B
441 Sequences
Number of Sequences
0
50
100
150
200
250
300
350
400
450
Total Number of Sequences
Number of Sequences Remaining with FRIES
0
500
1000
1500
2000
Rank of Remaining Conformation Space Size
C
9261 Sequences
Total Number of Sequences
Number of Sequences Remaining with FRIES
Number of Sequences
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
0
100
200
300
400
500
600
Rank of Remaining Conformation Space Size

Figure 3

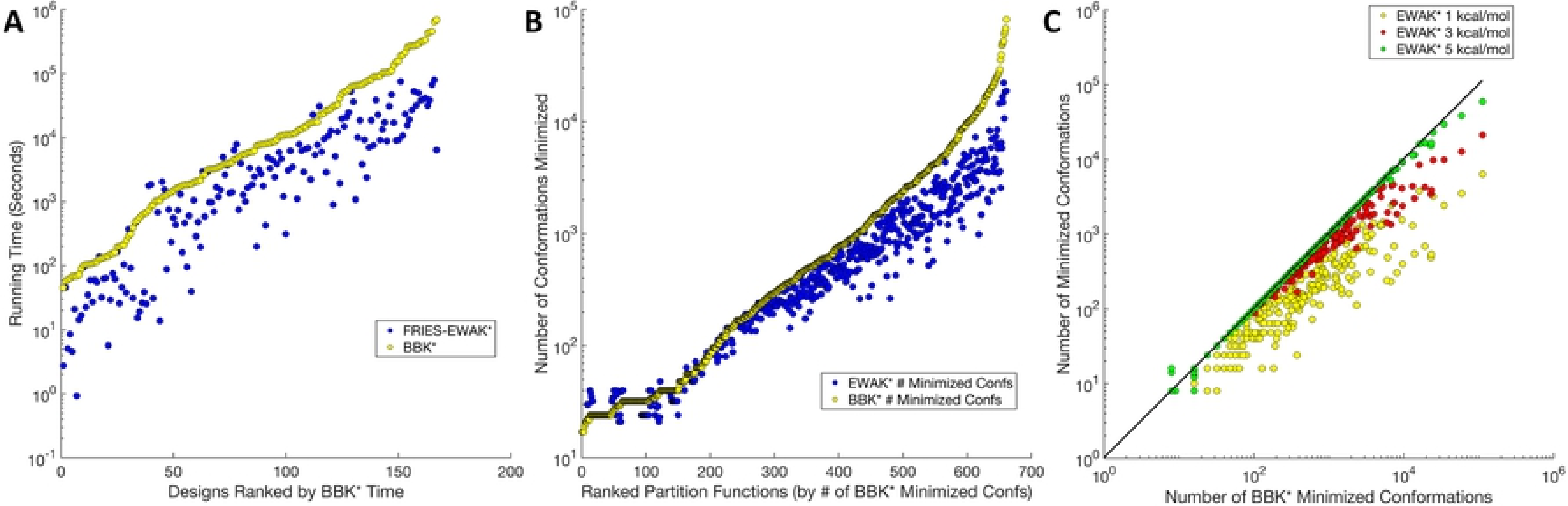

A
Running Time (Seconds)
Designs Ranked by BBK* Time
FRIES-EWAK*
BBK*
B
Number of Conformations Minimized
Ranked Partition Functions (by # of BBK* Minimized Confs)
EWAK* # Minimized Confs
BBK* # Minimized Confs
C
Number of Minimized Conformations
Number of BBK* Minimized Conformations
EWAK* 1 kcal/mol
EWAK* 3 kcal/mol
EWAK* 5 kcal/mol

Figure 4

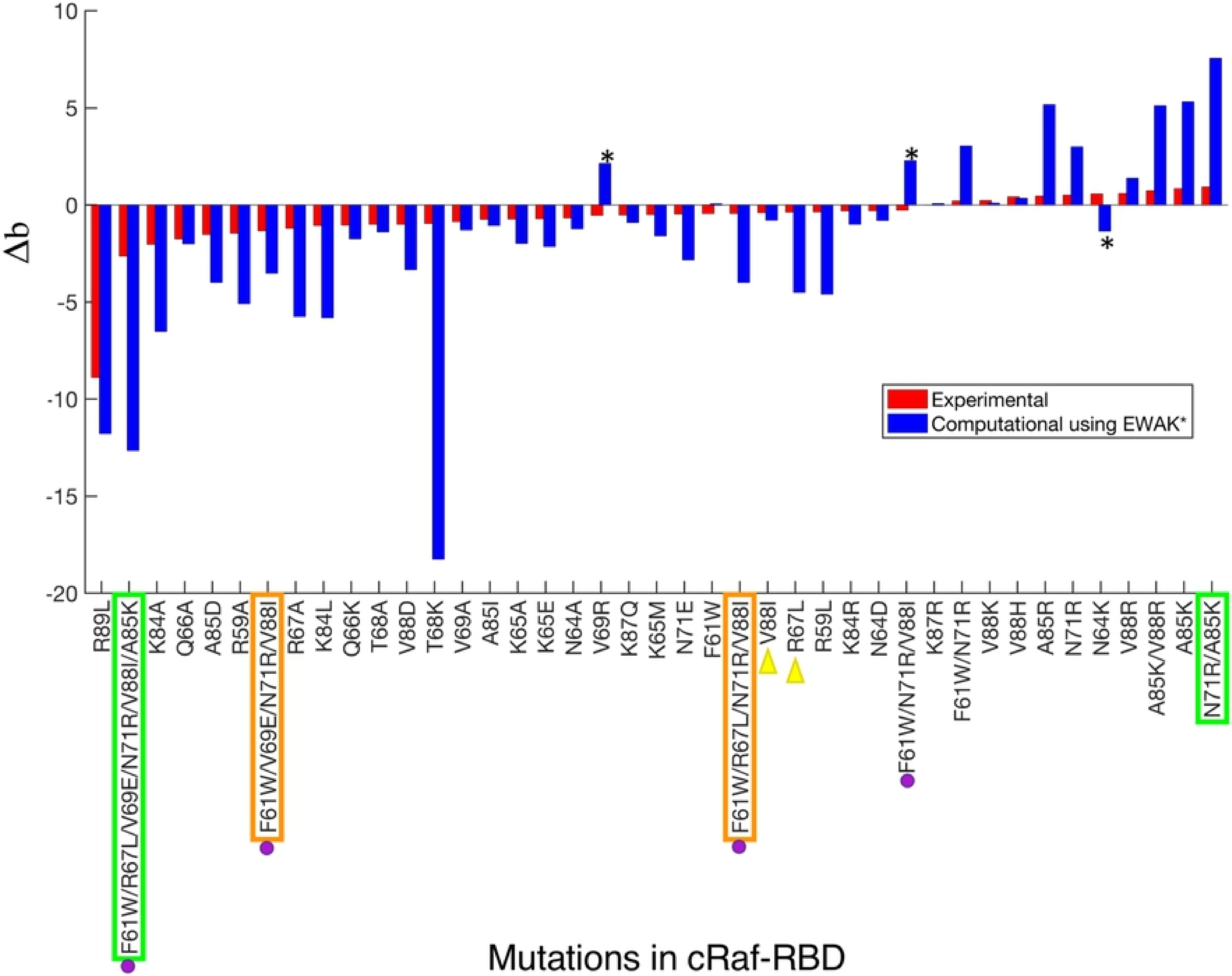

10
5
0
-5
-10
-15
-20
Δb
*
*
*
Experimental
Computational using EWAK*
R89L
F61W/R67L/V69E/N71R/V88I/A85K
K84A
Q66A
A85D
R59A
F61W/V69E/N71R/V88I
R67A
K84L
Q66K
T68A
V88D
T68K
V69A
A85I
K65A
K65E
N64A
V69R
K87Q
K65M
N71E
F61W
F61W/R67L/N71R/V88I
V88I
R67L
R59L
K84R
N64D
F61W/N71R/V88I
K87R
F61W/N71R
V88K
V88H
A85R
N71R
N64K
V88R
A85K/V88R
A85K
N71R/A85K
Mutations in cRaf-RBD

Figure 5

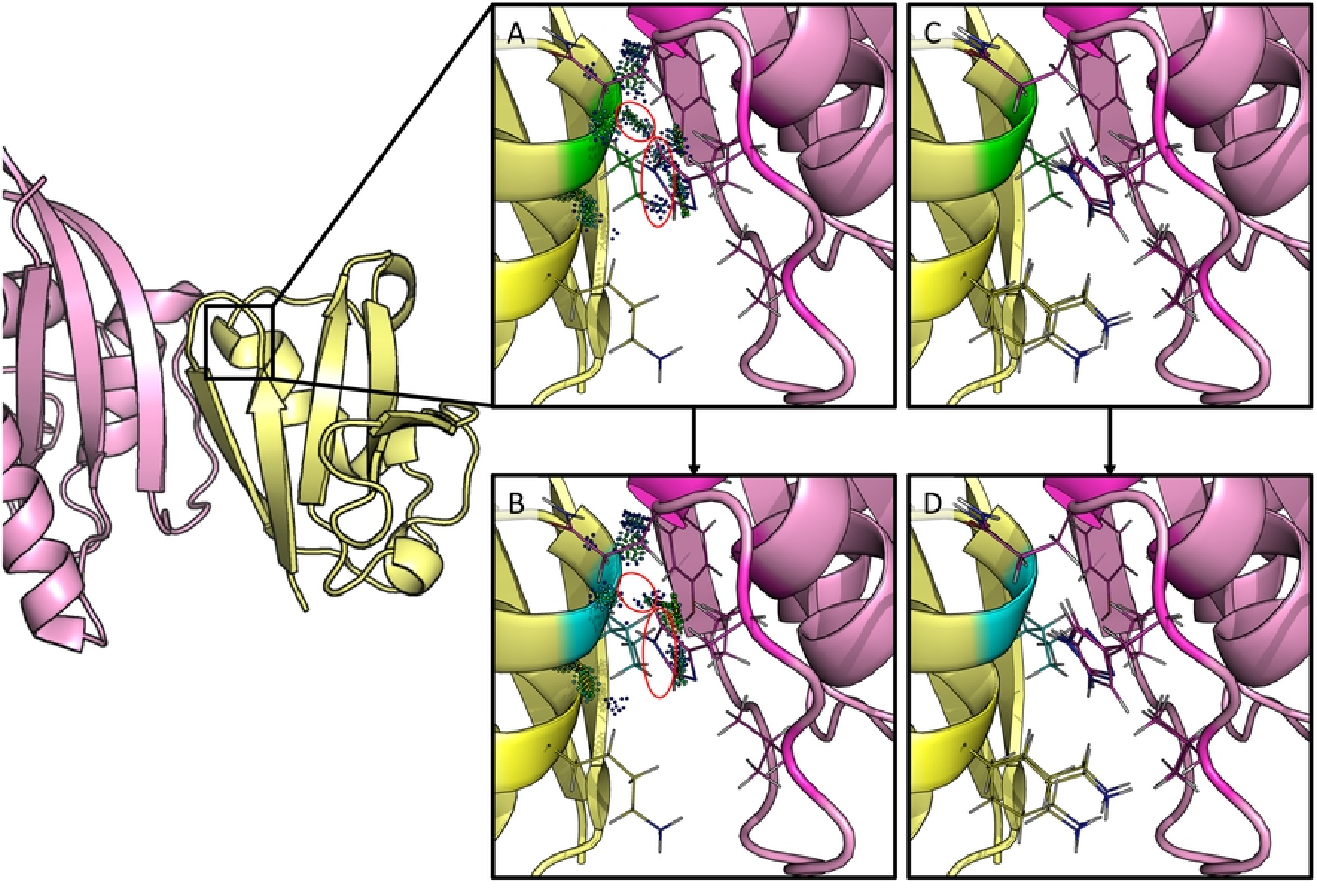


Figure 7

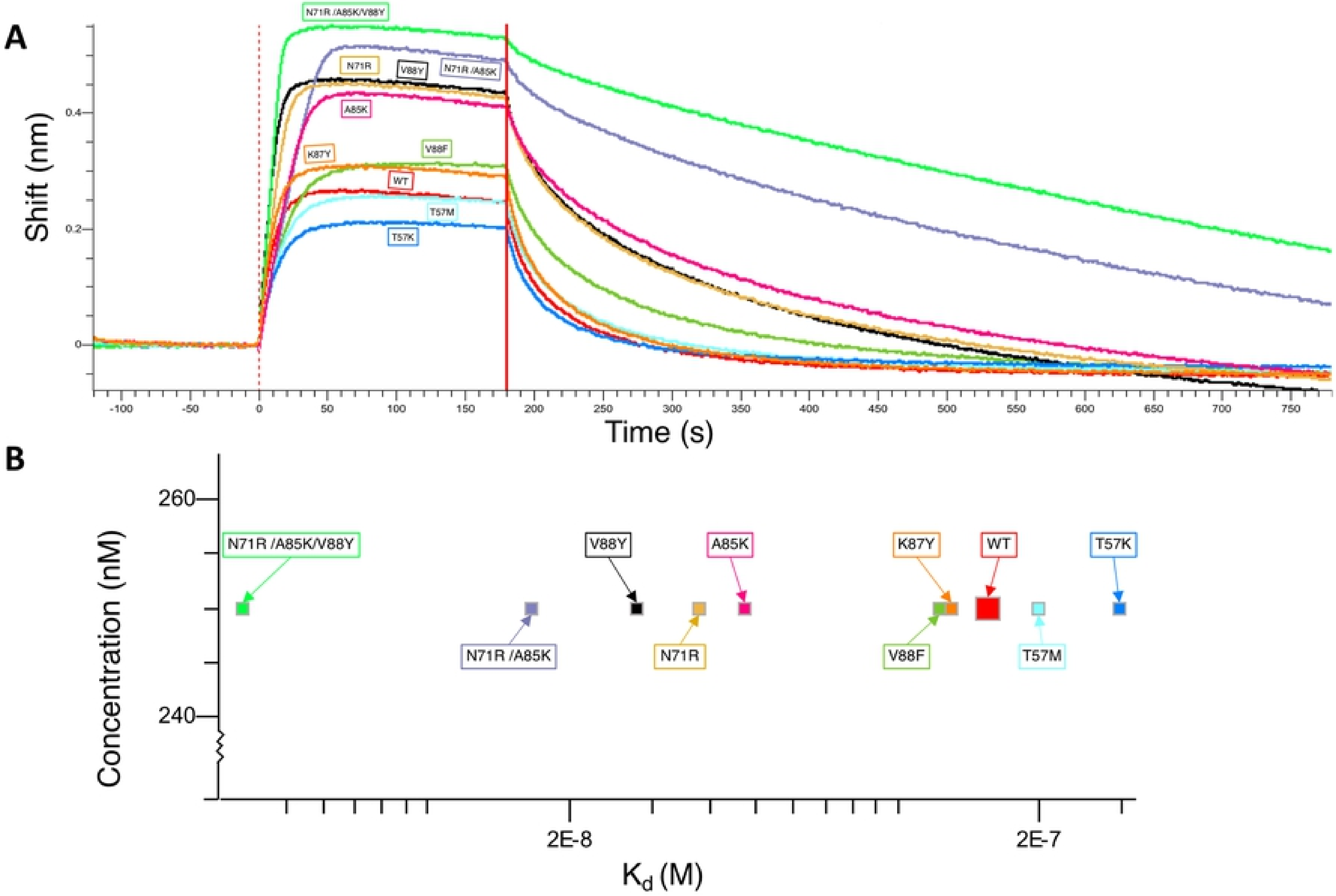
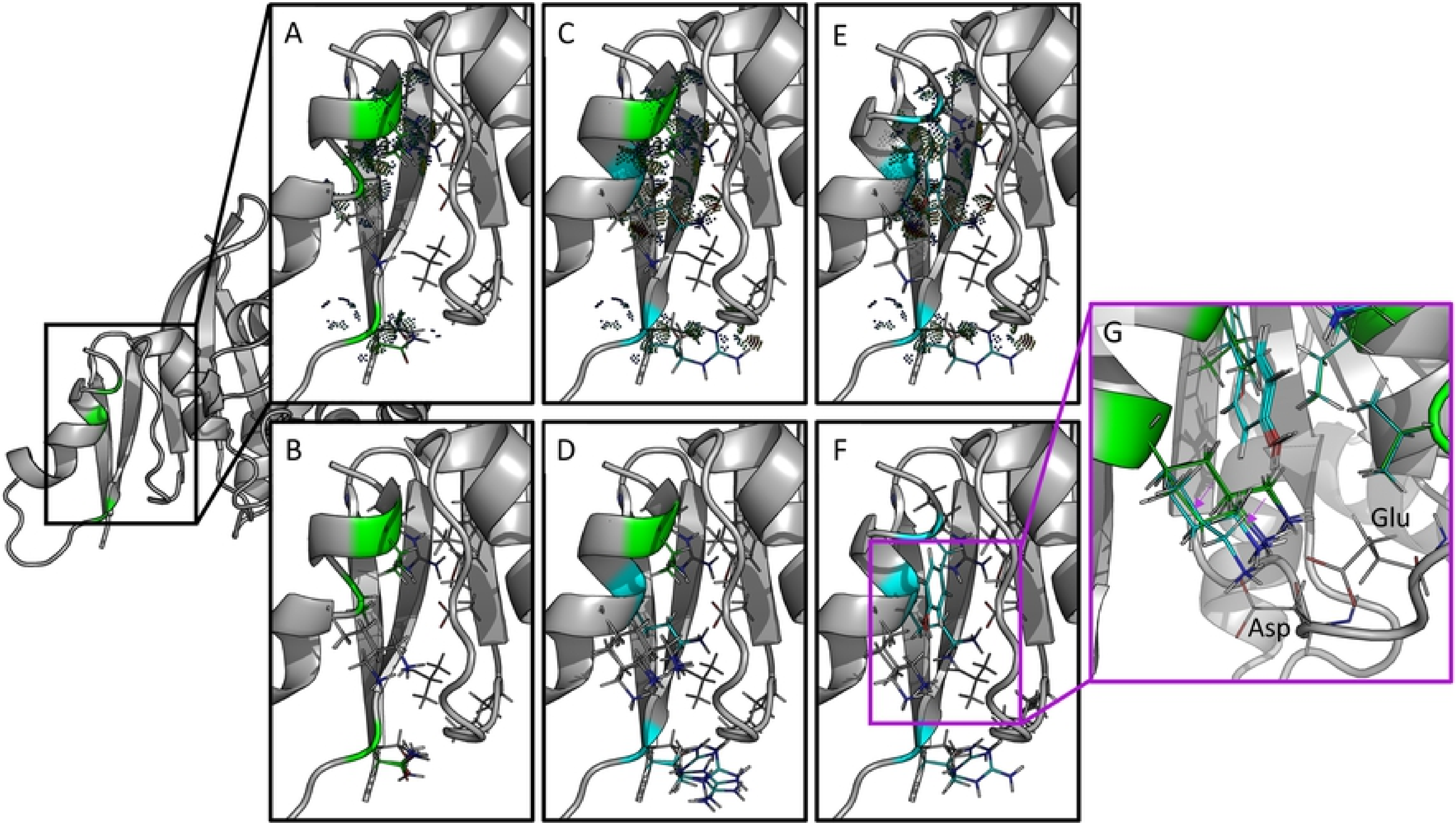
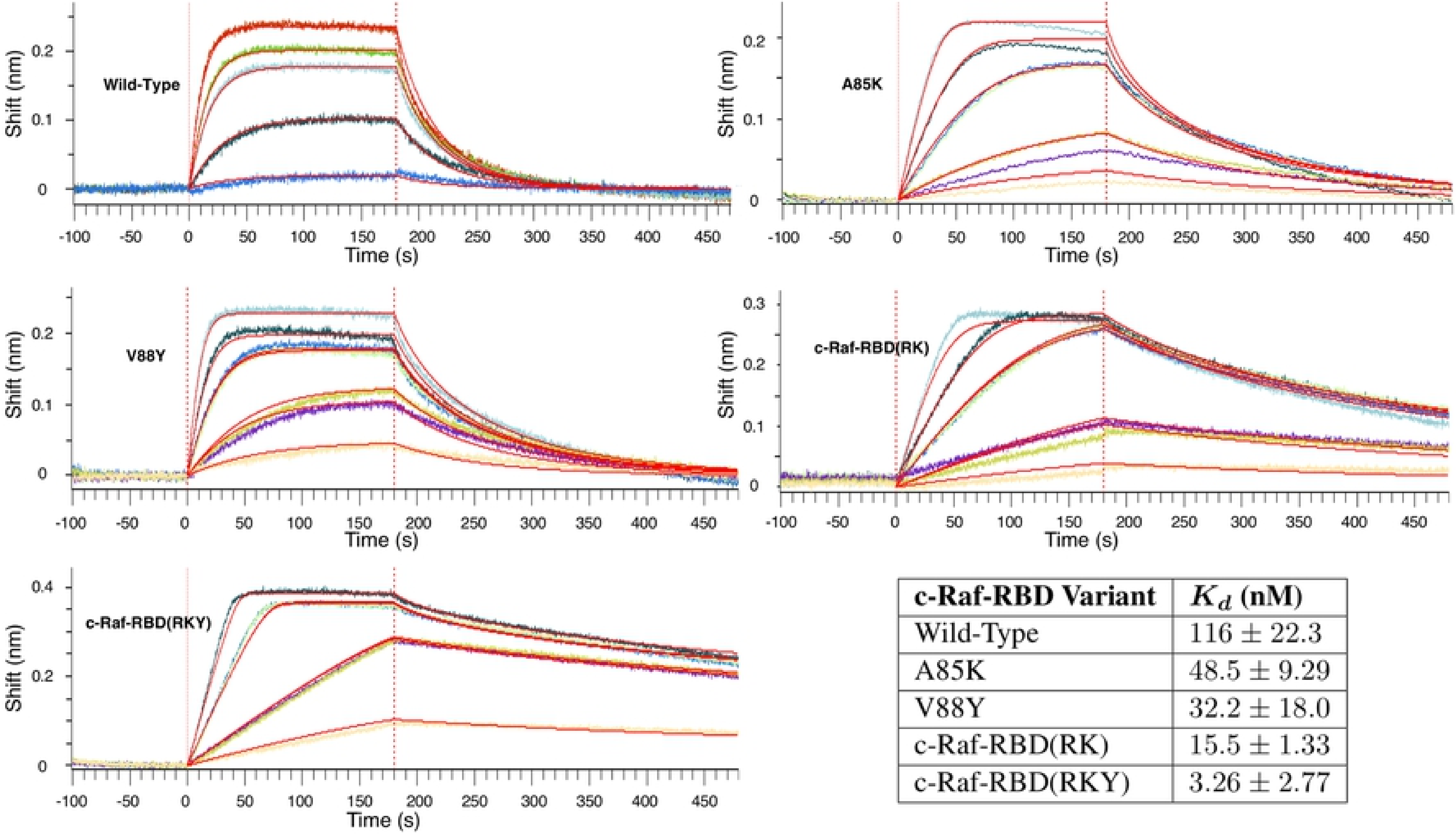

A
N71R /A85K/V88Y
N71R
V88Y
N71R /A85K
A85K
K87Y
V88F
WT
T57M
T57K
Shift (nm)
0.4
0.2
0
-100
-50
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
Time (s)
B
260
240
Concentration (nM)
N71R /A85K/V88Y
V88Y
A85K
K87Y
WT
T57K
N71R /A85K
N71R
V88F
T57M
2E-8
2E-7
K_d (M)

Figure 8



c-Raf-RBD Variant	Lower Bound log (K^*)	Upper Bound log (K^*)
Wild-Type	30.67	30.77
c-Raf-RBD(RK)	38.80	39.12
c-Raf-RBD(RKY)	39.29	39.80

Figure 9



c-Raf-RBD Variant	K_d **(nM)**
Wild-Type	116 ± 22.3
A85K	48.5 ± 9.29
V88Y	32.2 ± 18.0
c-Raf-RBD(RK)	15.5 ± 1.33
c-Raf-RBD(RKY)	3.26 ± 2.77

Figure 10