Article

Direct Calculation of Protein Fitness Landscapes through Computational Protein Design

Loretta Au[1,*] and David F. Green[2]

[1]Department of Statistics, The University of Chicago, Chicago, Illinois; and [2]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York

ABSTRACT Naturally selected amino-acid sequences or experimentally derived ones are often the basis for understanding how protein three-dimensional conformation and function are determined by primary structure. Such sequences for a protein family comprise only a small fraction of all possible variants, however, representing the fitness landscape with limited scope. Explicitly sampling and characterizing alternative, unexplored protein sequences would directly identify fundamental reasons for sequence robustness (or variability), and we demonstrate that computational methods offer an efficient mechanism toward this end, on a large scale. The dead-end elimination and A* search algorithms were used here to find all low-energy single mutant variants, and corresponding structures of a G-protein heterotrimer, to measure changes in structural stability and binding interactions to define a protein fitness landscape. We established consistency between these algorithms with known biophysical and evolutionary trends for amino-acid substitutions, and could thus recapitulate known protein side-chain interactions and predict novel ones.

INTRODUCTION

Protein mutagenesis studies can disentangle how native interactions in wild-type are functionally important, but incrementing the number of mutations for a variant results in a combinatorial expansion of the possible protein sequence space. Single mutant variants of a 350-amino-acid protein, for instance, would yield 6650 sequences, while changes as pairs or triplets would allow $>2.4 \times 10^7$ and $>5.7 \times 10^{10}$ unique sequences, respectively. The sheer magnitude of protein sequences raises many challenges for interpreting the role of primary structure in dictating protein structure and function, and although progress continues to be made toward this understanding, it remains incomplete. Existing methods offer a range of analytical results, varying in the type and number of sequences that are evaluated (Fig. 1 *a*). Comparative sequence analysis methods can measure sequence conservation, identify motifs, and evaluate evolutionary relationships of known, sequenced proteins (1–7), while primary structures that deviate away from biases of natural evolution can be created via mutagenesis protocols. As examples, alanine scanning replaces original amino-acid side chains with alanine (8–10), and even larger protein libraries are possible via directed evolution experiments (11–13), which can scale up to 10^{12} or more sequences for sampling; both approaches require additional resources for functional characterization. High sequence similarity by itself cannot guarantee that structural motifs or protein folds are shared (14–16), and this can affect how results derived solely from sequence analysis should be interpreted. In contrast, mutagenesis studies may be more costly than comparative sequence analysis, but the protein expression and functional assays that accompany these methods provide a more comprehensive understanding of the biophysical requirements that are essential to sequence-function relationships. High-throughput and deep-sequencing methods for directed evolution have been improving (17–19), and continue to elucidate the functional requirements of protein fitness. However, financial and temporal costs may still impose some constraints, depending on problem size, which motivated our development of a resourceful computational approach that can still provide high-resolution data for analysis. In particular, our protocol methodically simulates mutant variants for computing the protein fitness requirements of a chosen protein system without selection bias, offering additional perspective to how the energetic landscape of sequence space is shaped.

The dead-end elimination and A* search algorithms (DEE/A*) were adapted here for large-scale in silico mutagenesis, and thus enabled us to explore protein sequences that would be inaccessible otherwise (Fig. 1 *b*; see Fig. S1 in the Supporting Material). By assessing all low-energy sequences and their corresponding structures, we could deconvolve the multiple contributions of wild-type amino acids to protein fitness, defined here as structure stabilizing and binding interactions. Our computational approach, demonstrated here for a G-protein heterotrimer, is applicable to any system. However, it requires a reliable structural template to define the wild-type sequence. Enhanced sampling of backbone conformations is also needed, to account for

Submitted June 25, 2015, and accepted for publication November 16, 2015.

*Correspondence: lorettaau@uchicago.edu

Editor: Amedeo Caflisch.

0006-3495/16/01/0075/10

http://dx.doi.org/10.1016/j.bpj.2015.11.029

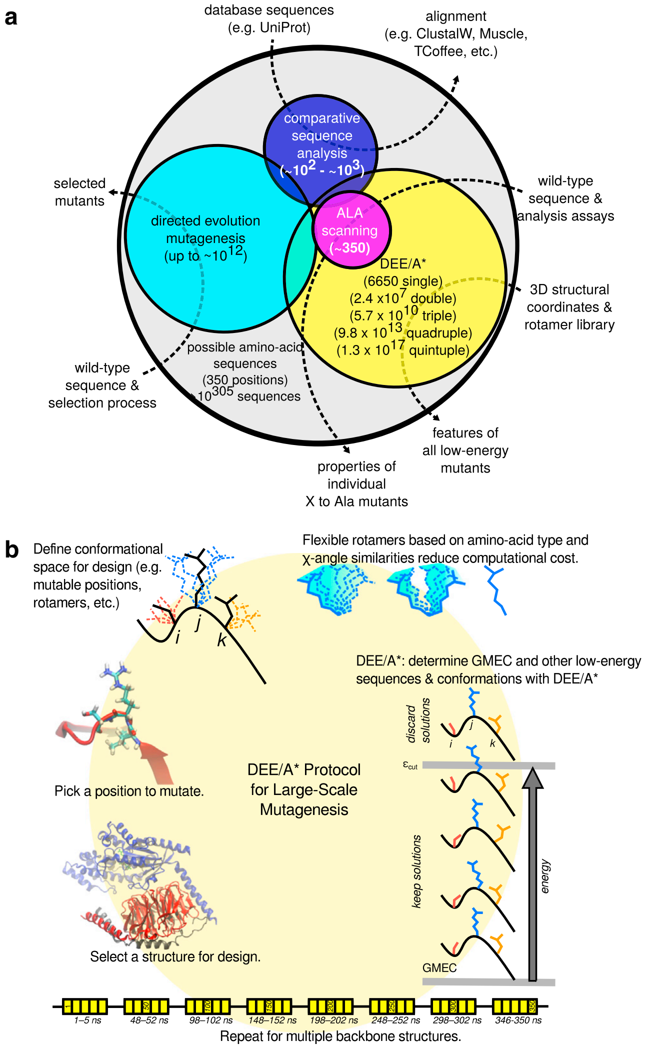
a
database sequences
(e.g. UniProt)
alignment
(e.g. ClustalW, Muscle, TCoffee, etc.)
comparative sequence analysis
(~10^2 - ~10^3)
selected mutants
directed evolution mutagenesis
(up to ~10^{12})
ALA scanning
(~350)
wild-type sequence & analysis assays
DEE/A*
(6650 single)
(2.4 x10^7 double)
(5.7 x 10^{10} triple)
(9.8 x 10^{13} quadruple)
(1.3 x 10^{17} quintuple)
3D structural coordinates & rotamer library
wild-type sequence & selection process
possible amino-acid sequences
(350 positions)
10^{305} sequences
features of all low-energy mutants
properties of individual X to Ala mutants
b
Define conformational space for design (e.g. mutable positions, rotamers, etc.)
i
j
k
Flexible rotamers based on amino-acid type and χ-angle similarities reduce computational cost.
DEE/A*: determine GMEC and other low-energy sequences & conformations with DEE/A*
discard solutions
ε_{cut}
keep solutions
energy
GMEC
Pick a position to mutate.
DEE/A* Protocol for Large-Scale Mutagenesis
Select a structure for design.
1
50
100
150
200
250
300
350
1–5 ns
48–52 ns
98–102 ns
148–152 ns
198–202 ns
248–252 ns
298–302 ns
346-350 ns
Repeat for multiple backbone structures.

FIGURE 1 Systematic mutagenesis using computational protein design algorithms. (*a*) The mutant sequence space for a protein with 350 amino acids can become combinatorially large, and methods for exploring this space are shown here. Data for input and analytical results are indicated (*arched arrows* passing through the corresponding technique; note: *circles* are not drawn to scale). (*b*) From a molecular dynamics simulation of a given protein system, multiple snapshots are taken to create an ensemble of representative backbone structures. Each position in the protein may be substituted to any amino acid, and nearby side chains are flexible while others remain fixed. Dead-end elimination will discard rotamers that are incompatible with a low-energy structure for a given sequence, and the A* search will evaluate the combination of rotamers at all flexible positions that will yield the global minimum energy conformation (GMEC) and additional solutions below a designated energy cutoff (ε_{cut}), within the given constraints. To see this figure in color, go online.

slight variations in the protein microenvironment and measure the consistency of a mutational effect. This sampling was established using multiple conformations from a molecular dynamics simulation, but as an alternative, backbone flexibility could be accounted for using methods that introduce ϕ- and ψ-angle perturbations to the backbone or by including multiple crystal structures of the protein (20–24).

We chose these algorithms because of their proven success in redesigning proteins to improve existing function or introduce novel ones (25–30). DEE evaluates amino acids and side-chain configurations that are incompatible with a low-energy protein conformation, based on the target protein structure, a rotamer library, and the energetic model used (31). Consequently, the number of possible structures is reduced; the lowest-energy protein conformation and additional ones within a designated energetic cutoff can then be identified from the remaining rotamers using the A* search algorithm, a heuristic, best first search that estimates the energetic cost of different rotamer combinations (Fig. 1 *b*) (32,33). Unlike a stochastic algorithm, the deterministic nature of DEE/A* guarantees the same solution every time, although it could be prone to completing an exhaustive search before doing so (34,35). A hierarchy of energetic models with increasing accuracy can be used to refine the solutions from DEE/A*: beginning with coarser pairwise decomposable approximations, high-energy sequences can be discarded early so that more intensive implicit or explicit solvent computations are performed on fewer molecules, reducing computational expense (36).

MATERIALS AND METHODS

Molecular dynamics setup

An all-atom molecular dynamics simulation was performed on the wild-type G-protein heterotrimer $G_i\alpha_1\beta_1\gamma_2$ (PDB: 1GP2) (37,38). CHARMM and NAMD were used for the simulation (39,40), with periodic boundary and NPT ensemble conditions (P = 1 atm, T = 300 K) using PARAM22/27 parameters (41,42), the TIP3P model for solvation (43), and a 2-fs time step. The structure was prepared using the REDUCE program to define initial protonation states (44), and hydrogen atom coordinates were determined using the HBUILD module in CHARMM (45). Randomly selected water molecules were replaced with sodium and chloride ions to establish a relevant physiological ionic strength (145 mM), with a minimum 10 Å distance between solute and the box edge. The structure was minimized after 240 steps and 200 ps of equilibration using Langevin dynamics in NAMD. A 12 Å cutoff was used for short-range interactions, while long-range electrostatic interactions were accounted for using the particle-mesh Ewald method. Snapshots were saved at every time step to ensure correlation with the same Boltzmann distribution throughout the simulation.

DEE/A* parameters

Forty wild-type conformations from the molecular dynamics trajectory were selected for analysis: the first 5 ns, the last 5 ns, and six additional intervals between them, each spanning 5 ns with a midpoint that was a multiple of 50 ns (Fig. S1). Side-chain orientations were defined using the original Dunbrack-Karplus rotamer library, augmented before use by adding ±10° to each χ_1- and χ_2-angle for enhanced sampling (46).

We applied the generalized-Born implicit solvent model with switching from CHARMM (47), after preliminary pruning using a distance-dependent dielectric ($\varepsilon = 4r$) (48). A flexible rotamer model was also used to discard unfavorable orientations quickly, by averaging together rotamers with similar χ_1- and χ_2-angles, reducing the size of the conformational space searched by DEE/A* (49). Each position in the wild-type sequence was mutated, and all sequences within 30 kcal/mol from the global minimum energy conformation were kept and referenced to the corresponding wild-type energy. Wild-type side chains within 5 Å of any GDP atom were included in the analysis of Gα-GDP interactions; not all side chains will interact with this ligand throughout the simulation due to backbone fluctuations, and free energy data were normalized accordingly.

Computing protein fitness of each mutant sequence

Protein fitness was measured as a combination of structural stability and binding interactions, either between Gα with GDP or Gα with the $\beta\gamma$-heterodimer. Structural stability is defined here as the energetic difference between an amino-acid side chain in the context of a folded structure and the reference state, in which the amino acid is isolated, then N-acetylated and N-methylamidated at the N- and C-termini, respectively; binding is defined energetically as the difference between the folded protein bound to its interaction partner and the same folded protein, unbound. A 500-Å rigid-body translation was used to separate binding partners to compute unbound-state energies. A Boltzmann-weighted average at an effective temperature of 4500 K was computed to represent the overall effect of each mutation (see the Supporting Material). The use of a discrete rotamer library and discretely sampled protein backbones leads to an exaggeration of unfavorable energies and the neglect of conformational entropy terms, which often partially compensate enthalpic terms, can further overstate the energetics. The use of an elevated effective temperature accounts for some of this exaggeration, albeit in an ad hoc manner.

Energy minimization of each DEE/A* result was performed using a Newton-Rhapson algorithm in CHARMM for 4000 steps each to identify shortcomings in the rotamer library. Mutational free energy was computed by decomposing amino acids into the amino-, carboxyl-, and variable side chain (starting from Cβ) groups, and the energetic difference was computed against a hydrophobic isostere of the wild-type amino acid. Similarity matrices were constructed using theoretical amino-acid probabilities found in the wild-type $G_i\alpha_1\beta_1\gamma_2$, and counting the number of sequences that survive an energetically defined evolutionary pressure. Frequency of substitution from amino acid i to j, e_{ij}, was computed using these counts, and compared to the expected frequency found in PAM120 and BLOSUM62 (see the Supporting Material).

Measuring amino-acid substitution rates

Entries in any PAM or BLOSUM matrix is a score, on a half-bit scale, that indicates the probability of observing substitutions to wild-type amino acid i with amino-acid j, S_{ij}. In a given set of protein sequences or within an aligned region of sequences (depending on the type of substitution matrix computed), the observed frequency of finding i substituted by j, e_{ij}, is compared to a corresponding theoretical probability that the amino-acid exchange may happen, p_i, p_j (where p_i and p_j are the natural, independent frequencies of occurrence for amino acids i and j, respectively) and thus $S_{ij} = 2\log_2(e_{ij}/p_ip_j)$. For comparison against these evolution-based observations, we defined e_{ij} as the number of DEE/A* sequences that simultaneously satisfied the 1.5 kcal/mol cutoff for structural stability and binding interactions after mutation of amino acid i into j (sequences that survived DEE/A* fitness pressures). Algebraically, scores from PAM and BLOSUM matrices can be converted to e_{ij} for comparison, because $e_{ij} = p_ip_j2^{(S_{ij}/2)}$; the values for S_{ij} were provided by PAM120, BLOSUM62, or a randomly generated matrix, and wild-type amino-acid distributions of the entire heterotrimer were used to define p_i and p_j accordingly (see the Supporting Material).

Statistical analysis for predictions

The Mann-Whitney-Wilcoxon statistical test was implemented using the *exactRankTests* library from the R statistical package. Neutral mutations were defined as changes from wild-type within a -1.5 and 1.5 kcal/mol range, and thus these values were set to zero before this analysis. An exact test was chosen to account for ties, and the null hypothesis (a zero vector, indicating no changes due to substitution) was compared to the empirical data collected for each position, a 20-dimensional vector representing the 20 possible amino acids underlying the cumulative distribution function (see the Supporting Material). A low p-value in these calculations suggests strong evidence that the position is mutationally sensitive for the aspect of fitness evaluated. Side-chain positions with known binding interactions were separated from all others to measure the true-positive rate of DEE/A*, based on a cutoff value, $p = 0.05$; this cutoff was also used as the premise for identifying additional side chains involved in binding interactions.

Computational resources

All DEE/A* mutations for an individual mutation were performed on a single 3.4 GHz Intel Pentium IV Xeon processor; most positions required ~4–5 h of computing time. There are 685 mutable positions in 1GP2, and using a cluster with 235 processing nodes, an average of 48 h was required to perform mutagenesis at all positions. Mutation free energy calculations required ~30 min of computing time on the same cluster.

RESULTS

Heterotrimeric G-proteins are ubiquitous in eukaryotic cell-signaling pathways, and we have chosen this as a model system for our approach, with $G_i\alpha_1\beta_1\gamma_2$ (PDB: 1GP2) as our wild-type reference (37,50,51). This family of proteins has been shown to have unique patterns in interaction specificity between subunits that enable complex formation and biological function (52–55). As determinants of broad-spectrum biological function, we have focused on 1) the structural stability of the complete protein complex, 2) the binding interactions between the $\beta\gamma$-heterodimer and Gα, and 3) the binding of Gα to GDP (see Materials and Methods).

Complete mutagenesis profiles calculated from using DEE/A*

Many mutations have a neutral effect on the protein (Figs. S2–S4; Tables S1–S3), but there is a tendency for mutations to be less favorable than wild-type. Approximately two-thirds of the sequences explored by DEE/A* are destabilizing to the wild-type structure, and greater energetic variance is seen in these sequences than those measured for changes in binding interactions (Fig. 2). This is due to both having fewer amino acids involved in binding (compared to stabilization), and having a broad range of microenvironments, from hydrophobic to highly solvent-exposed, available in the folded protein. A complete sequence profile for every position was established for our model system, identifying specific regions of unfavorable amino-acid substitution and highlighting those that are less sensitive to mutation (Figs. S5–S8, S10, and S11). Positions with several allowable and favorable substitutions usually have fewer geometric or electrostatic constraints; when very diverse functional groups cannot be accommodated at a position, it suggests

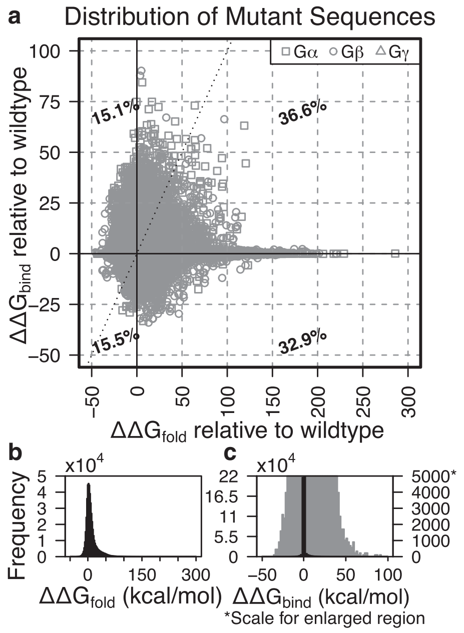

FIGURE 2 Protein fitness landscape for mutant sequences. Sequences are mapped according to energy relative to the wild-type sequence for structural stability ($\Delta\Delta G_{\text{fold}}$) and binding interactions ($\Delta\Delta G_{\text{bind}}$) in (*a*), and the relative proportions in each quadrant of this landscape are shown as percentages. The distribution of mutant sequences according to each aspect of protein fitness is shown in the bottom row. (*b*) For stability, there is a heavy tail in the distribution of $\Delta\Delta G_{\text{fold}}$, which indicates that most mutants are less stable than wild-type. (*c*) As for binding interactions, the value of $\Delta\Delta G_{\text{bind}}$ is often ~0 kcal/mol (shown in *black*, *y* axis on *left*), but a closer look at this histogram reveals that the distribution is skewed (shown in *gray*, *y* axis on *right*), with more sequences having a positive $\Delta\Delta G_{\text{bind}}$ value.

that unique side-chain interactions exist in the region and are required to maintain protein fitness.

Conformational space adequately modeled with rotamer library and protein structures

The movement of protein side chains is most accurately simulated using small χ-angle perturbations, but amino-acid orientations have been examined and statistically determined to favor specific combinations of χ-angles, establishing the basis for rotamer libraries (56,57). Without this discretization, DEE/A* simply cannot work—the algorithm evaluates unique combinations of side-chain placement on a given backbone structure, and rotamer libraries provide clear definitions of these possibilities in a limited number. As an alternative, energy minimization algorithms can be used to find favorable side-chain orientations that may be unlisted in such libraries, and can work well when the number of structures needed for analysis is not overwhelming. A comparison was made between all DEE/A* sequences sharing the same backbone orientation and their corresponding minimized structures (each starting as a DEE/A* solution) to assess the influence of a discretized rotamer space; a positive linear correlation was found between them (Fig. S15), suggesting energetic similarity despite slight differences in side-chain positioning. Approximately 60% of the data is found to be energetically unfavorable using DEE/A* and remains unfavorable after applying energy minimization, while roughly 20% of the data is favorable in both calculations. As the reference structure is derived from the wild-type sequence, a bias is found (and expected) toward mutant structures becoming more energetically favorable from using the minimization algorithm. However, as both wild-type and mutant sequences were minimized, relative energies from DEE/A* could be better than those from minimization (Fig. S15). Large discrepancies between the two methods of calculation (>20 kcal/mol) tended to involve substitutions to charged side chains or involve geometric constraints: in one particular β-sheet (GβAsp^{247}, GβThr^{249}, and GβArg^{251}), nonaliphatic amino acids were disfavored to preserve directionality, hydrogen-bond interactions, and the size of $(i, i+2)$ side chains (58), suggesting that finer sampling could be beneficial in specific side-chain packing contexts (Fig. S17). Even so, ~80% of the sequences were within ±5 kcal/mol of the alternative energy calculation, indicating a satisfactory evaluation of most sequences using DEE/A* without additional energy minimization (Fig. S16). The wild-type protein structures used were representative of major changes or fluctuations that the complete heterotrimer may undergo during simulation. The energetic variance of each mutant sequence was measured as the number of states in the structural ensemble increased (Fig. S12), and consistency in free energy was established first for sequences from densely packed, hydrophobic regions of the protein. Structural constraints within these regions were further reflected in the number of unsuitable mutations at these positions (Figs. S13 and S14; Table S4). In contrast, a greater number of backbone conformations was necessary to capture structural features of loops and other flexible regions of the heterotrimer, due to greater degrees of freedom.

Amino-acid functional roles can be disentangled

Energetic profiles were created separately for stability and binding interactions using the free energy of all mutant sequences (see Materials and Methods). By measuring these two aspects of fitness independently, functional trade-offs in the protein could be identified, as demonstrated by the GDP-binding pocket of Gα (Fig. 3). If either requirement for stability or binding was not satisfied, the overall fitness of the protein was worse than wild-type—the maximum energy of either stability or binding, max($\Delta\Delta G_{\text{fold}}$, $\Delta\Delta G_{\text{bind}}$), could distinguish this for a given mutant. Asp^{150}, for example, could be easily replaced by most amino acids and remain functional, because the native orientation points the carboxyl group away from GDP, despite being near a guanine nitrogen (Fig. 3 *a*). Nearly all substitutions could

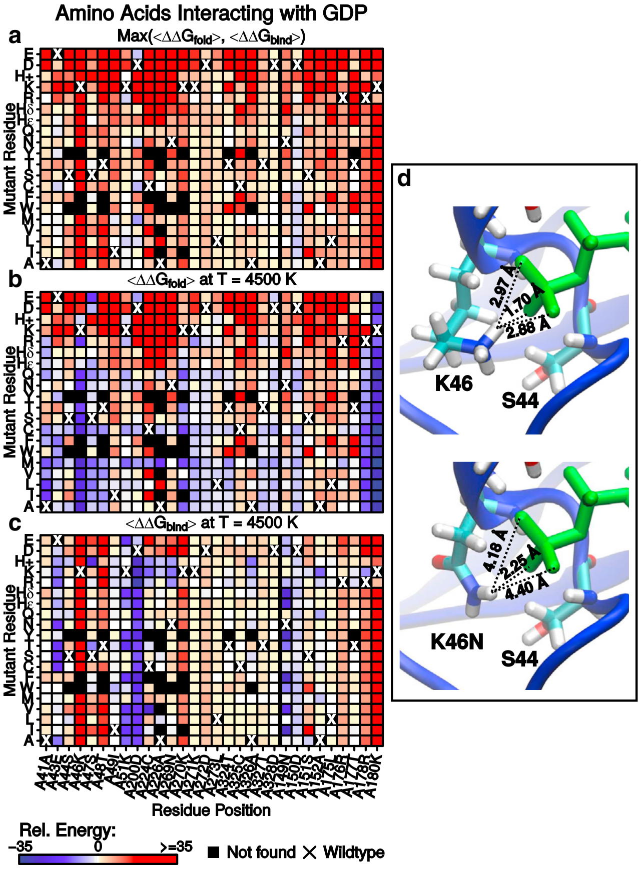

FIGURE 3 Gα interactions with GDP. Side chains within 5 Å of GDP can have functional trade-offs as substitutions are made to wild-type. Energy referenced to wild-type is shown for each mutant. A black box for a sequence indicates that the corresponding structure had steric clashes, and was thus not found by DEE/A*. Black boxes with a large *X* highlight the wild-type amino acid. (*a*) Evaluating max($\langle \Delta\Delta G_{\mathrm{fold}} \rangle$, $\langle \Delta\Delta G_{\mathrm{bind}} \rangle$) reveals difficulty in simultaneously satisfying both fitness criteria. (*b*) Average stability, $\langle \Delta\Delta G_{\mathrm{fold}} \rangle$, and (*c*) average binding interactions, $\langle \Delta\Delta G_{\mathrm{bind}} \rangle$, for Gα-GDP indicate varying degrees of mutational sensitivity at different positions. (*d*) Mutations often alter the proximity of important interactions, as seen in K46, in which hydrogen bonds are lost in K46N. To see this figure in color, go online.

improve protein stabilization for Ala^{41}, Lys^{46}, Lys^{270}, and Lys^{180} in Gα, but the same mutations were poor candidates for binding GDP (Fig. 3 *b*). Similarly, side-chain substitutions in the same subunit could improve binding interactions relative to wild-type at positions Ile^{49}, Asp^{200}, and Asn^{149}, but doing so would generally destabilize the α-subunit (Fig. 3 *c*). Bulkier, aromatic amino acids did not fit well in this region, and charged side chains were also poor candidates because of their electrostatic requirements. These general trends were exhibited throughout the heterotrimer, and functional trade-offs were only a concern for the small number of positions present at the protein-binding interface or involved in protein-ligand interactions (Figs. S7 and S8). Most positions were sensitive to substitution, and this is expected for a highly evolved protein family. Gα positions at the amino terminus or in switch II (residues 202–209) have the greatest energetic variation after mutation than other positions in the subunit, and these regions were known to interact with the β-subunit when inactive (50,59,60). Gβ, an example of a WD40 β-propeller protein, has positions at the binding interface that also show a similar trend, and where stability is lost after mutation is consistent with our expectations of the WD40 protein family (Fig. S9) (50,61).

Mutant structures from DEE/A* provide practical computational models

Alternative methods for studying wild-type contributions to protein stability, some of which require significantly fewer computational resources than DEE/A*, do exist. Amino acids may be decomposed according to functional groups, for instance, so that the energy required to convert a side chain into its hydrophobic isostere can be measured, and this mutational free energy elucidates any underlying electrostatic interactions (62–66). Such calculations can be completed in ~30 min for a single heterotrimer, while DEE/A* would require ~48 h for the same system using the same computing cluster. The expense of using DEE/A* is well compensated for, however—all 20 amino-acid choices are evaluated when finding low-energy sequences and simultaneously modeling tertiary structures. Having up to 19 mutant variants thus provides multiple frames of reference for assessing how tolerant a wild-type side chain can be to different kinds of mutation. The outcomes also include visual examples of less intuitive substitutions and energetic data that can help rank mutational effects or quantify the mutational robustness of wild-type. Energetic comparisons made with hydrophobic isosteres has its advantages in efficiency, but relies on an artificial construction that is not found in biology; DEE/A* offers practical models in its representations of actual amino acids.

To illustrate the compatibility between these two kinds of calculations, and their differences, the mutational free energy of all positions involved in Gα-GDP interactions were compared to the sequences from DEE/A* (Fig. 4). Each aspect of fitness was treated independently for assessment; the number of stable states found (defined by energetic cutoff) and the mutant sequence energies distributed were compared to mutational free energies computed using hydrophobic isosteres; positions could be separated easily according to mutational robustness in this way. Lys^{46} had negative mutational free energy, an indication that important interactions were made by this side chain to bind GDP, but from DEE/A*, we could understand that only wild-type would ever make these contributions—no other substitutions are allowed here. Conversely, we found that electrostatic contributions of Glu^{43} were also important in the wild-type, but all mutations were allowed and tended to

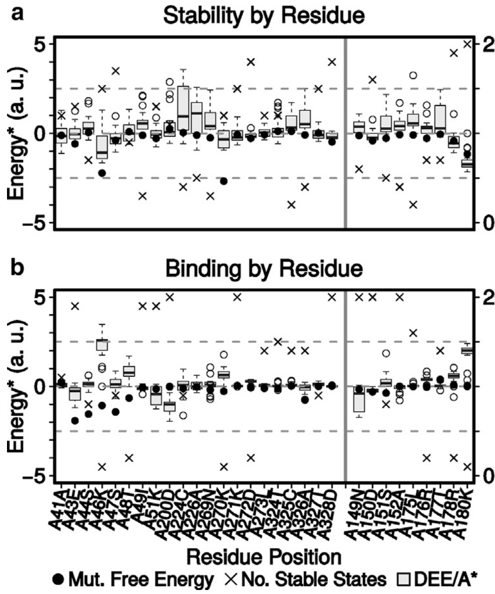

FIGURE 4 Energetic contributions of amino acids. Electrostatic calculations were performed for Gα amino acids that interact with GDP. Box-and-whisker plots represent the energetic distribution of all mutant sequences from DEE/A* at the specified position, and are overlaid onto mutational free energy data for (*a*) structural stability and (*b*) binding interactions. Arbitrary units (a.u.) were used for the *y* axes on the left at a scales: (*i*) 6 kcal/mol for stability mutation free energy, (*ii*) 1.6 kcal/mol for binding mutation free energy, and (*iii*) 16 kcal/mol for DEE/A* results in both contexts; the respective energetic ranges are thus (*i*) [−30,30] kcal/mol, (*ii*) [−8,8] kcal/mol, and (*iii*) [−80,80] kcal/mol. The number of sequences found at each position from DEE/A* are all those ≤1.5 kcal/mol from the wild-type energy; these quantities were marked with an *X* and follow the *y* axes on the right.

be more favorable than wild-type. Lys[180] could be substituted by anything to improve stability, and also have important native interactions; however, substitutions adversely affected binding interactions with GDP. All mutations were disallowed, even though the wild-type amino acid had little impact on binding, again demonstrating that geometry or interactions with neighboring residues play an important secondary role.

DEE/A* substitutions are strongly correlated with known amino-acid exchanges

Finally, the overall DEE/A* substitution rates were compared with the PAM120 and BLOSUM62 similarity matrices to measure how well DEE/A* (and our choice of folding and binding as measures of fitness) can reflect protein evolutionary pressures. Each PAM and BLOSUM matrix accounts for a broad range of sequence evolution, and are standard matrices for use in sequence alignments (67–69). Energetically favorable DEE/A* sequences were used to derive the expected frequency of substituting amino acid *i* with *j*, e_{ij}, for comparison to analogous values of e_{ij} using PAM or BLOSUM (see the Supporting Material). We defined protein fitness to depend on a combination of

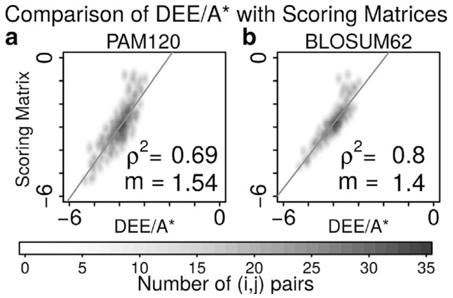

FIGURE 5 Comparison of DEE/A* and evolutionary substitution frequencies. Expected frequencies of substitution for any (*i*, *j*) amino-acid pair were computed based on the number of mutant sequences satisfying the 1.5-kcal/mol cutoff. Amino-acid probabilities from wild-type provide a basis for deriving substitution rates for comparison to (*a*) PAM120 and (*b*) BLOSUM62.

protein stabilization and binding ability, at unknown proportions (Figs. S18–S20); a balanced weighting of both fitness criteria optimized the correlation between our DEE/A* data and a chosen similarity matrix, with $\rho^2 \approx 0.7$ and $\rho^2 \approx 0.8$ for PAM and BLOSUM, respectively (Fig. 5). To determine whether these correlations were meaningful, our DEE/A* data were also compared to random matrices constructed 1) from a uniform distribution bounded by the maximum and minimum scores of both PAM120 and BLOSUM62, and 2) by permuting the entries of each similarity matrix. Correlation was generally poor ($\rho^2 \approx 0.1$) between DEE/A* and a completely arbitrary matrix; correlation between PAM120 (or BLOSUM62) with a permuted version of itself was first established ($\rho^2 \approx 0.3$ for PAM120, and $\rho^2 \approx 0.5$ for BLOSUM62), then DEE/A* data were compared to the randomized version of each matrix and showed slight improvement ($\rho^2 \approx 0.4$ for PAM120 and $\rho^2 \approx 0.6$ for BLOSUM62) (Fig. S22). These data suggest that the general distribution of values in the DEE/A*-derived matrices is similar to that in the PAM and BLOSUM matrices (it is this that leads to nonzero correlation between randomized matrices, but there are deeper similarities in the detailed structure of the matrices; Figs. S21 and S22; Tables S5 and S6). Comparison to alternative PAM and BLOSUM did not yield any statistically significant differences, due to low variance between different versions of PAM and BLOSUM scores overall (not shown).

Compatibility between alanine mutations from DEE/A* and thermal stability experiments

A full alanine scan was performed by Sun et al. (70) to understand the role of native interactions in stabilizing Gα, and we used these data for comparison with mutant alanine sequences from our DEE/A* calculations. In their experiment, each wild-type Gα residue was systematically mutated in the α-subunit to alanine (and wild-type alanine to glycine) for GDP- and GTP-bound states. The change in thermal stability

($\Delta T_m = T_{\text{mut}} - T_{wt}$) was measured for each single mutant, and a threshold of $\Delta T_m \leq 2^\circ$C was proposed by Sun et al. (70) for defining a destabilizing mutation. The analogous description from our DEE/A* calculations would be a mutation in which $\Delta\Delta G_{\text{fold}} > +1.5$ kcal/mol. We directly compared alanine substitutions from DEE/A* with the data provided by Sun et al. (70) using these two interpretations for destabilizing mutations (Fig. S23; see Table S7). We considered an unfavorable alanine mutation to be a positive outcome, and by these measurements, the sensitivity of DEE/A* was computed to be 0.52, while its specificity was 0.78. Energetic data for alanine mutants were then randomized and reassessed to establish a quantitative reference for these values, and we measured sensitivity and specificity to be 0.32 ± 0.04 and 0.67 ± 0.02, respectively, after 5000 independent trials (Table S8). Compared to all random trials performed, our DEE/A* calculations correctly classified important side-chain interactions (true-positives) and positions that were insensitive to mutation (true-negatives) at a consistently higher rate than any of the randomized cases (Fig. S24). (Consequently, the number of false-positives and false-negatives were both much lower than the randomized trials.) For all Gα side chains, 75% (162 positions) were predicted to be mutationally insensitive by both our DEE/A* calculations and the data provided by Sun et al. (70) (Table S7). Based only on the structural data from computational simulations, the role of native interactions was correctly determined in ~68% of Gα (the total number of true-positives and true-negatives) using only alanine substitutions. Although DEE/A* cannot perfectly replicate in cyto conditions and related assays, these statistics were strong indicators that DEE/A* can reasonably predict the importance of wild-type interactions.

Consistency between DEE/A* and known point mutations

Oncogenic point mutations are available in public databases, such as COSMIC and cBioPortal (71–73), and several were found for the GNB1 gene, which encodes Gβ_1. In addition to compiling a complete list of these single mutants, Yoda et al. (74) discovered a few additional ones in their experiment that explored cancerous mutations affecting the β-subunit. These results from Yoda et al. (74) were used for comparison with corresponding DEE/A* mutants (Table S9). Either gain-of-function or loss-of-function mutations could be oncogenic, and both of these possibilities were considered in our comparison with DEE/A*. Furthermore, lethal mutations could affect function by altering heterotrimer stability or by disrupting proper association between Gα and the $\beta\gamma$-heterodimer, but the distinction between these two mechanisms is not always known from the available data. Thus, an energetic definition that accounted for both gain-of-function and loss-of-function mutations, and the possible contexts for which mutations may affect heterotrimer function, would be the maximum magnitude of either protein stability or binding interactions: max($|\Delta\Delta G_{\text{fold}}|$, $|\Delta\Delta G_{\text{bind}}|$). A value >1.5 kcal/mol in our DEE/A* calculations would indicate either an activating or deactivating mutation. We found a positive correlation between our computational results and the set of known point mutations: of the 36 single mutants available for comparison, only three of them had a neutral change after mutation (both $|\Delta\Delta G_{\text{fold}}| \leq 1.5$ kcal/mol and $|\Delta\Delta G_{\text{bind}}| \leq 1.5$ kcal/mol (see the Supporting Material). The remaining 33 mutations (92%) were either activating or deactivating mutations in at least one of the fitness contexts. These results further demonstrated that our computational approach can capture important trends in mutational effects found in biological systems.

High predictive ability of computational results

The Mann-Whitney-Wilcoxon statistical test was used for evaluating quantitative differences between mutational effects, and to assess mutational sensitivity of a wild-type side chain (see the Materials and Methods.) Binding interactions between subunits are documented for 49 positions (in total) of Gα and Gβ (50), and 38 (~78%) of these were detected by DEE/A* (Fig. 6 *a*), based on a threshold of $p = 0.05$. Five additional positions could be included in this count, if the threshold were adjusted to 0.10 instead, accounting for ~87% of known positions. Discrepancies for false-negatives from the remaining 11 positions (6 at threshold $p = 0.10$) were likely due to differences in conformational sampling. These positions tended to be in highly flexible protein regions: near the switch region of Gα and near the N-terminal helix of Gα (Fig. 6 *b*). The original observed interactions were established using x-ray crystallographic data, while our computational results were based on an ensemble of structures for the heterotrimer. Having this distinction for our data made it possible to computationally determine a broader regime of side chains involved in protein binding: 30 additional side chains were predicted by DEE/A* using the same statistical analysis (see Materials and Methods). The same metrics could be extrapolated for understanding structurally stabilizing interactions, and nearly all positions were found to have some significant contribution (Fig. S25; Tables S10 and S11). The molecular requirements for structural stability are not necessarily interchangeable with binding requirements, however, and further experimentation would be needed to verify the predictive ability of DEE/A* for this aspect of fitness and to separate side chains that are fundamental to stabilizing tertiary structure from less influential ones.

DISCUSSION

Proteins must satisfy a number of conditions, including the ability to stably form an appropriate fold and associate with

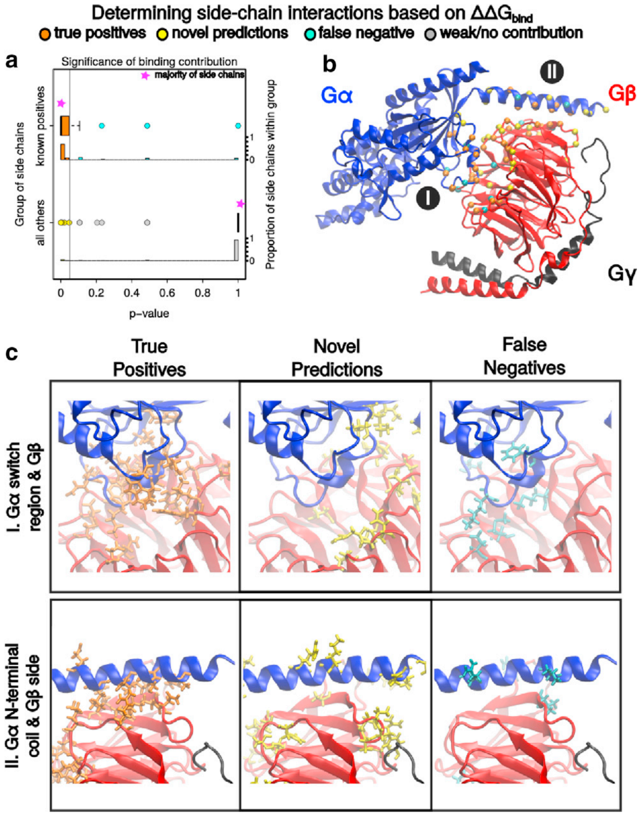

FIGURE 6 Recapitulation and prediction of side-chain interactions. Statistical differences based on all mutant $\Delta\Delta G_{bind}$ values were measured with the Mann-Whitney-Wilcoxon statistical test. (*a*) The distribution of *p* values for positions known to make binding interactions and all other positions are shown. (*Vertical line*) $p = 0.05$ for visual purposes. (*Magenta stars*) Nearly all positions of a subgroup that are found in the *p*-value distribution. (*b*) Positions for true-positives (*orange*, $p \leq 0.05$), novel predictions (*yellow*, $p \leq 0.05$), and false-negatives (*cyan*, $p \leq 0.05$) are mapped as spheres onto the heterotrimer for reference. (*Spheres*) The α-carbon positions of each residue. (*c*) Structural examples of where true-positive, novel predictions, and false-negatives are typically found. To see this figure in color, go online.

various cognate binding targets, to be biologically functional and overcome different selection pressures. Many advances in high-throughput methods have considerably improved how protein sequence-function relationships can be studied, but the volume of possible sequence space remains inevitably greater. Our DEE/A* protocol provides a mechanism for studying mutations on a very large scale to help mend a part of this disparity, and provide a perspective that is different, although complementary, to these approaches. By analyzing novel sequence variants systematically, the energetic landscape of a protein was computed, and the functional role of each amino acid could be deconvolved. The performance of DEE/A* also relies on different resources than existing methods: when the number of sequences for alignment is inadequate or the evolutionary history of a given protein is not well understood, a thorough analysis of protein structural stability and binding interactions is still possible. Whether alone or combined with existing methods, this level of detail can provide a better understanding of how protein design or engineering goals can be met.

Depending on the problem being considered, variations to our approach could be made to improve modeling details and computational accuracy. Longer molecular dynamics simulations and/or a greater number of representative structures would enhance sampling in highly flexible proteins or regions, for instance. Additional solvation models could also be used, e.g., by passing important sequences found using implicit-solvent models onto explicit-solvent simulations, to provide a more detailed explanation of electrostatic interactions. Furthermore, epistatic relationships from pairwise interactions and amino-acid covariation could be studied in depth using DEE/A* by introducing multiple mutations into the sequence at once. While these modifications may provide a better picture of how mutagenesis affects the wild-type protein, improvements in capturing amino-acid substitutions that completely reflect evolutionary biology might never be possible—evolutionary fitness pressures extend far beyond what can be measured by energetic change. Despite this, modeling side-chain substitutions with DEE/A* has shown consistency with structural studies and binding assays. DEE/A* thus complements comparative sequence analysis methods very well: sequence conservation can be directly linked to measurable aspects of fitness, and regions of allowable sequence variation can be explained.

CONCLUSIONS

The dead-end elimination and A* search algorithms simultaneously search over protein sequence and conformational spaces, and we have leveraged these to elucidate many sequence-function relationships in a heterotrimeric G-protein. By adapting these two algorithms to find all low-energy single mutants, the multiple roles amino acids play in overall protein fitness could be deconvolved as a function of mutational robustness. Large-scale mutagenesis using this computational approach is able to capture many biophysical features of side-chain substitutions, and these changes in the initial wild-type structure satisfy expectations based on preexisting structural and experimental studies. DEE/A* reveals several relationships among primary structure, structural stability, and protein function, enhancing the utility of techniques in comparative sequence analysis and extending the boundaries of accessible protein sequence space.

SUPPORTING MATERIAL

Supporting Materials and Methods, twenty-five figures, and eleven tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(15)01212-6.

AUTHOR CONTRIBUTIONS

L.A. and D.F.G. designed the experiments; L.A. performed the experiments; L.A. and D.F.G. contributed analytical tools; L.A. analyzed the data; and L.A. and D.F.G. wrote the article.

ACKNOWLEDGMENTS

We thank the National Institutes of Health grant No. GM086199 to D.F.G. for partial funding, Dr. Bruce Tidor for access to the DEE/A* software, and the SEAWULF cluster at Stony Brook University for computing resources.

REFERENCES

1. Thompson, J. D., T. J. Gibson, …, D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25:4876–4882.
2. Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302:205–217.
3. Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
4. Lockless, S. W., and R. Ranganathan. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science.* 286:295–299.
5. Süel, G. M., S. W. Lockless, …, R. Ranganathan. 2003. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* 10:59–69.
6. Magliery, T. J., and L. Regan. 2005. Sequence variation in ligand binding sites in proteins. *BMC Bioinformatics.* 6:240.
7. McLaughlin, R. N., Jr., F. J. Poelwijk, …, R. Ranganathan. 2012. The spatial architecture of protein function and adaptation. *Nature.* 491:138–142.
8. Cunningham, B. C., and J. A. Wells. 1989. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science.* 244:1081–1085.
9. Massova, I., and P. A. Kollman. 1999. Computational alanine scanning to probe protein-protein interactions: a novel approach to evaluate binding free energies. *J. Am. Chem. Soc.* 121:8133–8143.
10. Weiss, G. A., C. K. Watanabe, …, S. S. Sidhu. 2000. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc. Natl. Acad. Sci. USA.* 97:8950–8954.
11. Sidhu, S. S., and S. Koide. 2007. Phage display for engineering and analyzing protein interaction interfaces. *Curr. Opin. Struct. Biol.* 17:481–487.
12. Ernst, A., D. Gfeller, …, S. S. Sidhu. 2010. Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Mol. Biosyst.* 6:1782–1790.
13. Araya, C. L., D. M. Fowler, …, S. Fields. 2012. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. USA.* 109:16858–16863.
14. Cordes, M. H. J., R. E. Burton, …, R. T. Sauer. 2000. An evolutionary bridge to a new protein fold. *Nat. Struct. Biol.* 7:1129–1132.
15. Newlove, T., J. H. Konieczka, and M. H. J. Cordes. 2004. Secondary structure switching in Cro protein evolution. *Structure.* 12:569–581.
16. van Dorn, L. O., T. Newlove, …, M. H. J. Cordes. 2006. Relationship between sequence determinants of stability for two natural homologous proteins with different folds. *Biochemistry.* 45:10542–10553.
17. Fowler, D. M., C. L. Araya, …, S. Fields. 2010. High-resolution mapping of protein sequence-function relationships. *Nat. Methods.* 7:741–746.
18. Fowler, D. M., J. J. Stephany, and S. Fields. 2014. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* 9:2267–2284.
19. Stiffler, M. A., D. R. Hekstra, and R. Ranganathan. 2015. Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell.* 160:882–892.
20. Harbury, P. B., J. J. Plecs, …, P. S. Kim. 1998. High-resolution protein design with backbone freedom. *Science.* 282:1462–1467.
21. Davis, I. W., W. B. Arendall, 3rd, …, J. S. Richardson. 2006. The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure.* 14:265–274.
22. Georgiev, I., and B. R. Donald. 2007. Dead-end elimination with backbone flexibility. *Bioinformatics.* 23:i185–i194.
23. Smith, C. A., and T. Kortemme. 2008. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.* 380:742–756.
24. Smith, C. A., and T. Kortemme. 2011. Predicting the tolerated sequences for proteins and protein interfaces using RosettaBackrub flexible backbone design. *PLoS One.* 6:e20451.
25. Dahiyat, B. I., and S. L. Mayo. 1997. De novo protein design: fully automated sequence selection. *Science.* 278:82–87.
26. Shimaoka, M., J. M. Shifman, …, T. A. Springer. 2000. Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nat. Struct. Biol.* 7:674–678.
27. Bolon, D. N., and S. L. Mayo. 2001. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. USA.* 98:14274–14279.
28. Sarkar, C. A., K. Lowenhaupt, …, D. A. Lauffenburger. 2002. Rational cytokine design for increased lifetime and enhanced potency using pH-activated "histidine switching". *Nat. Biotechnol.* 20:908–913.
29. Looger, L. L., M. A. Dwyer, …, H. W. Hellinga. 2003. Computational design of receptor and sensor proteins with novel functions. *Nature.* 423:185–190.
30. Bolon, D. N., R. A. Grant, …, R. T. Sauer. 2005. Specificity versus stability in computational protein design. *Proc. Natl. Acad. Sci. USA.* 102:12724–12729.
31. Desmet, J., M. De Maeyer, …, I. Lasters. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature.* 356:539–542.
32. Leach, A. R., and A. P. Lemon. 1998. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins.* 33:227–239.
33. Pierce, N. A., J. A. Spriet, and S. L. Mayo. 2000. Conformational splitting: a more powerful criterion for dead-end elimination. *J. Comput. Chem.* 21:999–1009.
34. Desjarlais, J. R., and N. D. Clarke. 1998. Computer search algorithms in protein modification and design. *Curr. Opin. Struct. Biol.* 8:471–475.
35. Voigt, C. A., D. B. Gordon, and S. L. Mayo. 2000. Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* 299:789–803.
36. Green, D. F. 2010. A statistical framework for hierarchical methods in molecular simulation and design. *J. Chem. Theory Comput.* 6:1682–1697.
37. Wall, M. A., D. E. Coleman, …, S. R. Sprang. 1995. The structure of the G protein heterotrimer $G_i\alpha1\beta1\gamma2$. *Cell.* 83:1047–1058.
38. Carrascal, N., and D. F. Green. 2010. Energetic decomposition with the generalized-Born and Poisson-Boltzmann solvent models: lessons from association of G-protein components. *J. Phys. Chem. B.* 114:5096–5116.
39. Brooks, B. R., C. L. I. Brooks, 3rd, …, M. Karplus. 2009. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* 30:1545–1614.
40. Phillips, J. C., R. Braun, …, K. Schulten. 2005. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26:1781–1802.

41. MacKerell, A. D. J., J. Wiórkiewicz-Kuczera, and M. Karplus. 1995. An all-atom empirical energy function for the simulation of nucleic acids. *J. Am. Chem. Soc.* 117:11946–11975.

42. MacKerell, A. D., D. Bashford, …, M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102:3586–3616.

43. Jorgensen, W. L., J. Chandrasekhar, …, M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926.

44. Word, J. M., S. C. Lovell, …, D. C. Richardson. 1999. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* 285:1735–1747.

45. Brünger, A. T., and M. Karplus. 1988. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. *Proteins.* 4:148–156.

46. Dunbrack, R. L., Jr., and M. Karplus. 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* 230:543–574.

47. Im, W., M. S. Lee, and C. L. Brooks, 3rd. 2003. Generalized Born model with a simple smoothing function. *J. Comput. Chem.* 24:1691–1702.

48. Lippow, S. M., K. D. Wittrup, and B. Tidor. 2007. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.* 25:1171–1176.

49. Mendes, J., A. M. Baptista, …, C. M. Soares. 1999. Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. *Proteins.* 37:530–543.

50. Wall, M. A., B. A. Posner, and S. R. Sprang. 1998. Structural basis of activity and subunit recognition in G protein heterotrimers. *Structure.* 6:1169–1183.

51. Neves, S. R., P. T. Ram, and R. Iyengar. 2002. G protein pathways. *Science.* 296:1636–1639.

52. Fawzi, A. B., D. S. Fay, …, J. K. Northup. 1991. Rhodopsin and the retinal G-protein distinguish among G-protein $\beta\gamma$ subunit forms. *J. Biol. Chem.* 266:12194–12200.

53. Schmidt, C. J., T. C. Thomas, …, E. J. Neer. 1992. Specificity of G protein β and γ subunit interactions. *J. Biol. Chem.* 267:13807–13810.

54. Rens-Domiano, S., and H. E. Hamm. 1995. Structural and functional relationships of heterotrimeric G-proteins. *FASEB J.* 9:1059–1066.

55. Yan, K., V. Kalyanaraman, and N. Gautam. 1996. Differential ability to form the G protein $\beta\gamma$ complex among members of the β and γ subunit families. *J. Biol. Chem.* 271:7141–7146.

56. Dunbrack, R. L., Jr. 2002. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* 12:431–440.

57. Petrella, R. J., and M. Karplus. 2001. The energetics of off-rotamer protein side-chain conformations. *J. Mol. Biol.* 312:1161–1175.

58. Murzin, A. G. 1992. Structural principles for the propeller assembly of β-sheets: the preference for seven-fold symmetry. *Proteins.* 14:191–201.

59. Conklin, B. R., and H. R. Bourne. 1993. Structural elements of G α subunits that interact with G β γ, receptors, and effectors. *Cell.* 73:631–641.

60. Neer, E. J. 1995. Heterotrimeric G proteins: organizers of transmembrane signals. *Cell.* 80:249–257.

61. Wu, X.-H., Y. Wang, …, Y.-D. Wu. 2012. Identifying the hotspots on the top faces of WD40-repeat proteins from their primary sequences by β-bulges and DHSW tetrads. *PLoS One.* 7:e43005.

62. Hendsch, Z. S., and B. Tidor. 1994. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.* 3:211–226.

63. Archontis, G., T. Simonson, and M. Karplus. 2001. Binding free energies and free energy components from molecular dynamics and Poisson-Boltzmann calculations. Application to amino acid recognition by aspartyl-tRNA synthetase. *J. Mol. Biol.* 306:307–327.

64. Elcock, A. H. 2001. Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.* 312:885–896.

65. Hendsch, Z. S., M. J. Nohaile, …, B. Tidor. 2001. Preferential heterodimer formation via undercompensated electrostatic interactions. *J. Am. Chem. Soc.* 123:1264–1265.

66. Green, D. F., and B. Tidor. 2005. Design of improved protein inhibitors of HIV-1 cell entry: optimization of electrostatic interactions at the binding interface. *Proteins.* 60:644–657.

67. Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. Chapter 22: A model of evolutionary change in proteins. *In* Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, DC, pp. 345–352.

68. Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA.* 89:10915–10919.

69. Altschul, S. F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219:555–565.

70. Sun, D., T. Flock, …, D. B. Veprintsev. 2015. Probing Gα_i1 protein activation at single-amino acid resolution. *Nat. Struct. Mol. Biol.* 22:149–170.

71. Forbes, S. A., D. Beare, …, P. J. Campbell. 2015. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* 43:D805–D811.

72. Cerami, E., J. Gao, …, N. Schultz. 2012. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2:401–404.

73. Gao, J., B. A. Aksoy, …, N. Schultz. 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal complementary data sources and analysis options. *Sci. Signal.* 6:1–20.

74. Yoda, A., G. Adelmant, …, A. A. Lane. 2015. Mutations in G protein β subunits promote transformation and kinase inhibitor resistance. *Nat. Med.* 21:71–75.

Supporting Material

Direct Calculation of Protein Fitness Landscapes
through Computational Protein Design

L. Au and D. F. Green, 2015

Note: Residue positions in all figures are listed alpha-numerically with the protein chain (A, B, and G for the α-, β- and γ-subunits, respectively), followed by the position number according to the PDB file, and a single-letter code for the wild-type amino acid. Histidine states may be singly protonated at δ- or ε-nitrogens and are listed with lower-case 'd' (or Hδ) and 'e' (or Hε), respectively; when it is doubly protonated, we indicate this with lower-case 'p' or with H+.

List of Figures

List of Tables

1 Overview of computational protocol

A molecular dynamics simulation was performed on $G\alpha_{i1}\beta_1\gamma_2$, and protein conformations were taken from 50-ns intervals over a 350-ns trajectory; a total of 40 snapshots were used in our analysis (5 conformations from each interval,) as shown in **Fig. S1**. Every position in a given protein conformation is mutated to each of the naturally occurring amino acids, except proline and glycine, and the rotamers that are incompatible with a low-energy conformation for a given sequence are discarded. From the remaining rotamer choices, the global minimum energy conformation and additional structures within a designated energy cutoff from it are identified. We have chosen a 30-kcal/mol cutoff for this, and have found that over 90% of all possible single-site mutations satisfy this constraint.

2 Neutral mutations defined by energetic landscape

Neutral mutations, those that neither worsen nor improve fitness, were defined based on the distribution of $\Delta\Delta G_{fold}$ and $\Delta\Delta G_{bind}$ on the energy landscape, after referencing to the wild-type sequence; the proportion of mutant sequences was highest around the origin, suggesting that most mutations have a neutral effect. Sequences simultaneously sharing $\Delta\Delta G_{fold}$ and $\Delta\Delta G_{bind}$ within a designated energy cutoff, ε_{cut}, were removed from the data set (e.g. $|\Delta\Delta G_{fold}| \leq \varepsilon_{cut}$ and $|\Delta\Delta G_{bind}| \leq \varepsilon_{cut}$) (**Fig. S2**). The proportion of remaining sequences was calculated for various energy cutoffs, and compared. At the 1.5-kcal/mol cutoff, a significant reduction in sequence space is observed, and thus taken as the energetic cutoff for defining neutrality (**Fig. S3**).

3 Effective temperature for DEE/A*

A Boltzmann-weighted average was computed for each mutation over all 40 snapshots, and an effective temperature for the distribution is necessary to rescale computed results so that the values could better reflect energetic changes due to conformational entropy. A number of temperatures were tested to gauge the compatibility between structures found and their computed energies; we focused on energetic changes after mutation at salt-bridges and hydrogen bonds to set a baseline (**Fig. S4**).

Most hydrogen bonds, involving nitrogen and oxygen, are expected to have ∼2–7 kcal/mol each, and we used this as a guideline to compare our calculations for specific mutations at different temperatures (**Tables S1–S3**). Our goal was to identify a temperature at which the loss or gain of hydrogen bonds would fall within the 2–7 kcal/mol interval, and the lowest energy that achieves this is at 4500K.

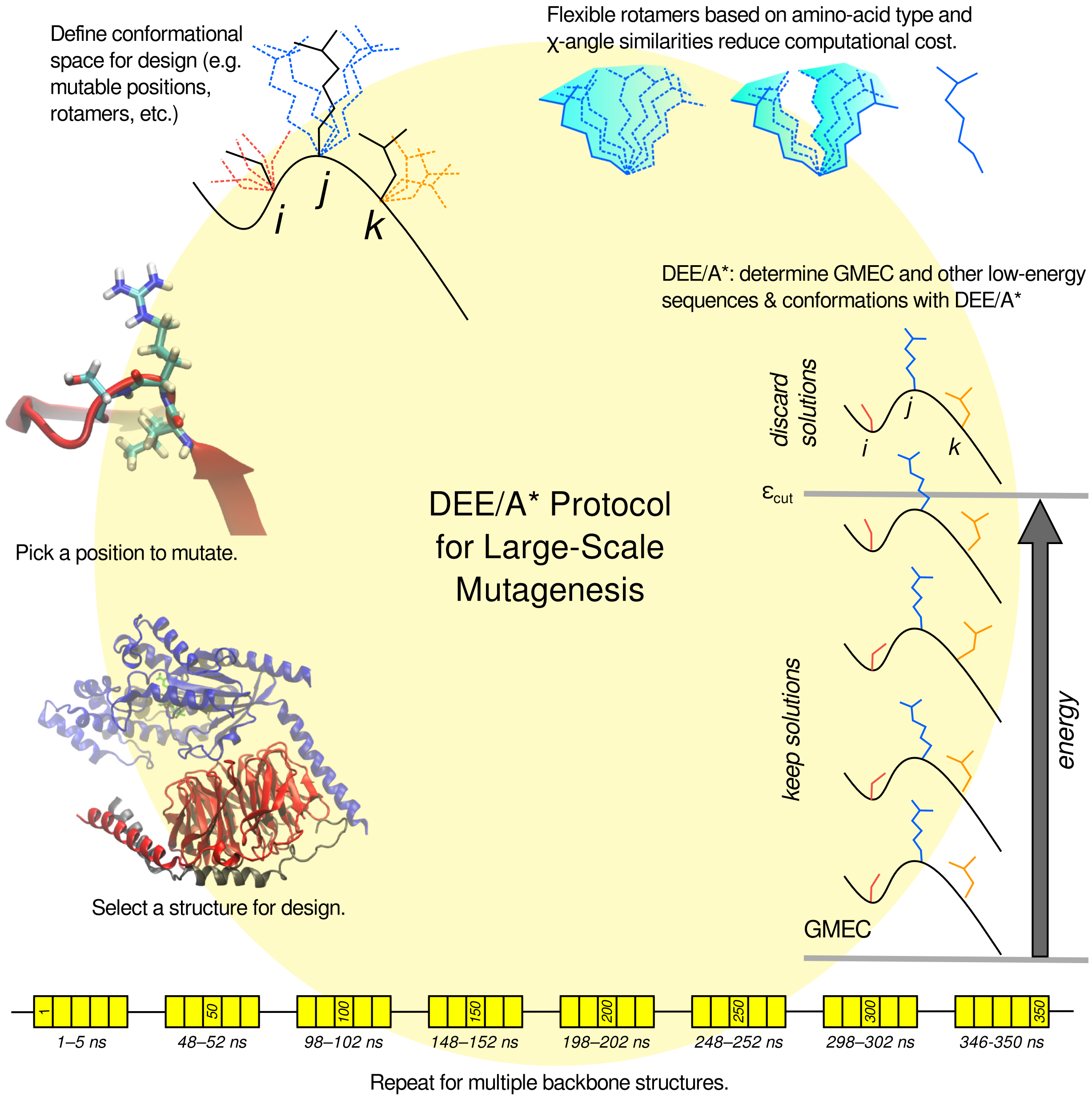
Define conformational space for design (e.g. mutable positions, rotamers, etc.)
i
j
k
Flexible rotamers based on amino-acid type and χ-angle similarities reduce computational cost.
DEE/A*: determine GMEC and other low-energy sequences & conformations with DEE/A*
discard solutions
i
j
k
ε_{cut}
keep solutions
energy
GMEC
DEE/A* Protocol for Large-Scale Mutagenesis
Pick a position to mutate.
Select a structure for design.
1
50
100
150
200
250
300
350
1–5 ns
48–52 ns
98–102 ns
148–152 ns
198–202 ns
248–252 ns
298–302 ns
346-350 ns
Repeat for multiple backbone structures.

Figure S1: (Main-text figure 1b.) Representative protein backbones are mutated systematically until low-energy sequences within ε_{cut} and their corresponding conformations are found.

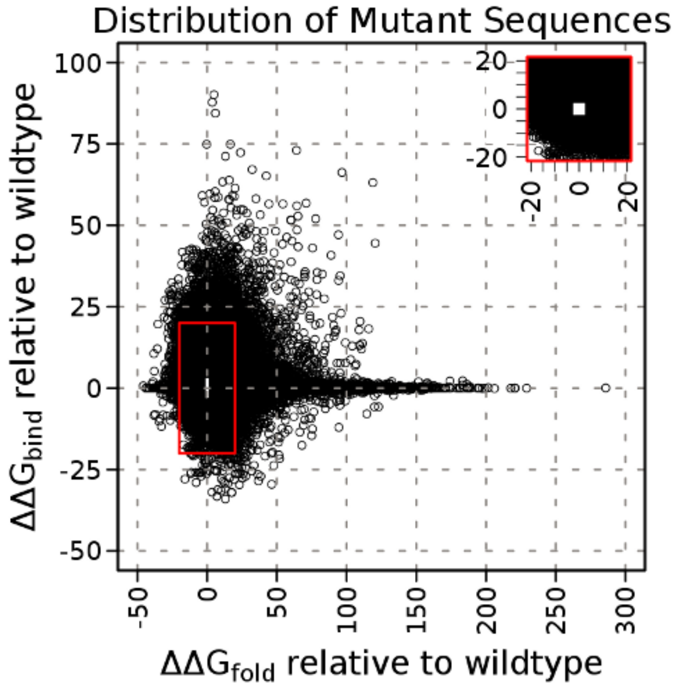

Figure S2: Sequences within 1.5 kcal/mol of $\Delta\Delta G_{fold}$ and $\Delta\Delta G_{bind}$, based on absolute value, removed.

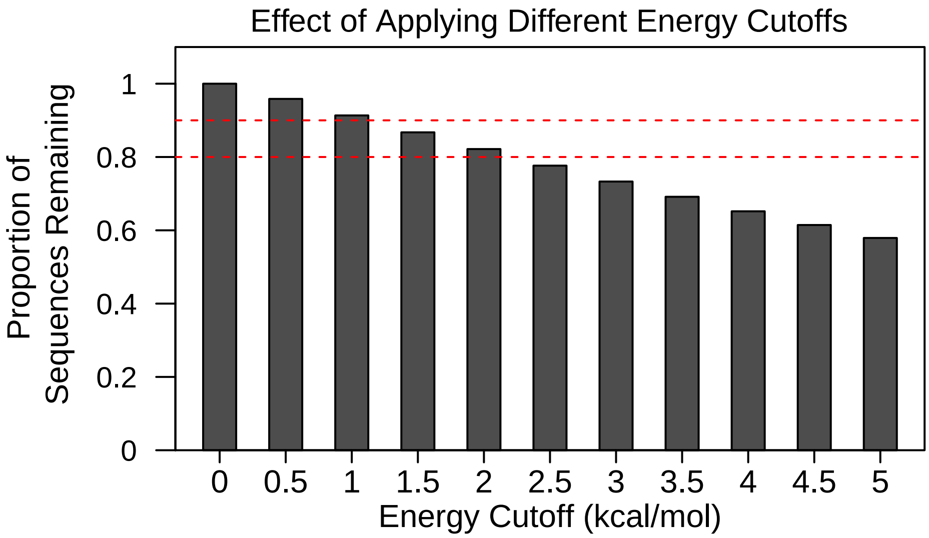

Figure S3: The proportion of sequences remaining after removing neutral sequences, defined by different ε_{cut}, is shown.

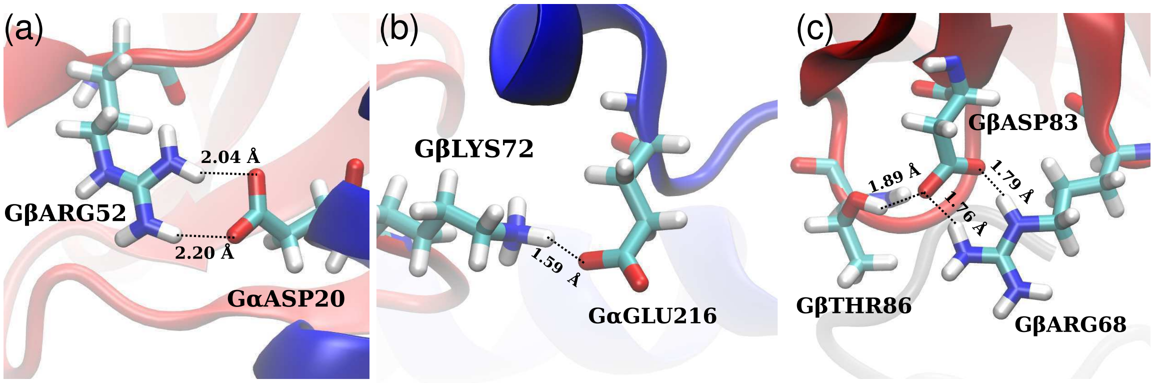

Figure S4: Temperature for effective energy was based on well-established intermolecular interactions: (a) doubly-bonded salt bridge, GαD20–GβR52, (b) singly-bonded salt bridge, GαE216–GβK57, and (c) a hydrogen-bond network between GβR68–GβE83–GβT86

Table S1: $\langle \Delta\Delta G_{bind} \rangle$ in kcal/mol for select mutations at GαD20–GβR52 salt bridge.

Temp. (K)	GαD20A	GαD20E	GαD20N	GαD20Q	GβR52A	GβR52K
300	-1.5± 0.2	-13.1 ± 2.0	-4.8 ± 0.7	-7.2 ± 0.9	5.2 ± 1.0	0.0 ± 0.0
1500	-0.2 ± 0.1	-10.4 ± 1.0	-2.9 ± 0.4	-5.6 ± 0.5	1.8 ± 0.4	1.3 ± 0.1
3000	2.1 ± 0.1	-6.7 ± 0.5	0.4 ± 0.2	-2.6 ± 0.3	2.2 ± 0.2	1.8 ± 0.1
4500	3.3 ± 0.1	-4.6 ± 0.3	1.9 ± 0.2	-1.1 ± 0.2	2.4 ± 0.1	2.0 ± 0.1
6000	4.0 ± 0.2	-3.5 ± 0.3	2.7 ± 0.2	-0.3 ± 0.2	2.6 ± 0.1	2.2 ± 0.1
9000	4.9 ± 0.2	-2.4 ± 0.2	3.6 ± 0.2	0.7 ± 0.2	2.7 ± 0.1	2.5 ± 0.1
300,000	7.7 ± 0.3	0.3 ± 0.3	6.3 ± 0.3	3.3 ± 0.3	3.0 ± 0.1	3.2 ± 0.1

Table S2: $\langle \Delta\Delta G_{bind} \rangle$ in kcal/mol for select mutations at GαE216–GβK57 salt bridge.

Temp. (K)	Gα E216A	Gα E216D	Gα E216N	Gα E216Q	Gβ K57A	Gβ K57R
300	2.0 ± 0.3	-2.6 ± 0.4	-8.7 ± 1.4	-7.6 ± 1.1	-8.0 ± 1.3	-10.2 ± 1.6
1500	1.2 ± 0.3	-0.7 ± 0.1	-4.5 ± 0.7	-7.0 ± 0.5	-8.0 ± 1.3	-10.2 ± 1.6
3000	1.1 ± 0.2	1.3 ± 0.1	-1.5 ± 0.3	-5.0 ± 0.2	-7.9 ± 1.3	-10.0 ± 1.6
4500	1.9 ± 0.1	2.3 ± 0.1	-0.2 ± 0.2	-4.2 ± 0.2	-6.4 ± 1.1	-7.8 ± 1.3
6000	2.4 ± 0.1	2.8 ± 0.1	0.5 ± 0.2	-3.8 ± 0.1	-3.1 ± 0.8	-4.4 ± 0.9
9000	2.9 ± 0.1	3.3 ± 0.1	1.2 ± 0.1	-3.3 ± 0.1	2.3 ± 0.4	-0.3 ± 0.4
300,000	3.9 ± 0.1	4.4 ± 0.1	2.4 ± 0.1	-2.5 ± 0.1	8.0 ± 0.1	4.6 ± 0.2

Table S3: $\langle \Delta\Delta G_{bind} \rangle$ for select mutations in the hydrogen-bond network

Temp. (K)	Gβ R68A	Gβ R68K		
300	9.3 ± 1.3	-2.9 ± 0.4		
1500	3.7 ± 0.4	-1.4 ± 0.3		
3000	3.1 ± 0.2	-0.1 ± 0.2		
4500	3.2 ± 0.2	0.6 ± 0.1		
6000	3.3 ± 0.2	1.0 ± 0.1		
9000	3.4 ± 0.2	1.4 ± 0.1		
300,000	3.8 ± 0.1	2.3 ± 0.1		
	Gβ D83A	Gβ D83E	Gβ D83N	Gβ D83Q
300	-5.3 ± 0.9	-1.5 ± 0.2	-4.3 ± 0.7	-5.3 ± 0.8
1500	-4.3 ± 0.6	2.8 ± 0.3	-3.2 ± 0.4	-3.1 ± 0.6
3000	-1.8 ± 0.3	4.2 ± 0.2	-1.4 ± 0.3	-1.0 ± 0.3
4500	-0.1 ± 0.2	4.8 ± 0.2	0.0 ± 0.2	0.3 ± 0.2
6000	0.8 ± 0.2	5.1 ± 0.2	0.9 ± 0.1	1.0 ± 0.2
9000	1.8 ± 0.1	5.7 ± 0.2	1.9 ± 0.1	1.9 ± 0.2
300,000	3.5 ± 0.2	7.1 ± 0.2	4.0 ± 0.2	3.8 ± 0.2
	Gβ T86A	Gβ T86S	Gβ T86C	
300	-3.1 ± 0.5	-2.1 ± 0.3	-3.0 ± 0.3	
1500	-1.1 ± 0.2	-0.5 ± 0.1	-2.2 ± 0.2	
3000	0.3 ± 0.1	0.1 ± 0.1	-0.9 ± 0.1	
4500	0.9 ± 0.1	0.3 ± 0.1	-0.3 ± 0.1	
6000	1.2 ± 0.1	0.5 ± 0.1	0.0 ± 0.1	
9000	1.5 ± 0.1	0.8 ± 0.1	0.4 ± 0.1	
300,000	2.5 ± 0.1	1.9 ± 0.1	1.4 ± 0.1	

4 Complete mutation profiles for G$\alpha_{i1}\beta_1\gamma_2$

For each mutant sequence, the energetic difference relative to wild type is computed over an ensemble of states using backbone structures from the 40 chosen conformations, and an effective temperature of 4500 K was used to compute a Boltzmann-weighted average over them for each sequence. Energy profiles for each subunit of the heterotrimer were compiled to identify regions of high and low mutational sensitivity. These energetic changes due to mutation are shown with secondary structure for:

- Stability ($\langle \Delta\Delta G_{fold} \rangle$) of G$\alpha$ in context of a complete heterotrimer (**Fig. S5**).
- Stability ($\langle \Delta\Delta G_{bind} \rangle$) of $\beta\gamma$-heterodimer in context of a complete heterotrimer (**Fig. S6**).
- Binding interactions ($\langle \Delta\Delta G_{bind} \rangle$) of G$\alpha$ to $\beta\gamma$-heterodimer (**Fig. S7**).
- Binding interactions ($\langle \Delta\Delta G_{bind} \rangle$) of $\beta\gamma$ to the α subunit (**Fig. S8**).
- Residues involved in binding that show significant energetic variation (**Fig. S9**).
- Maximum of either stability or binding in each Gα mutant (**Fig. S10**).
- Maximum of either stability or binding in each $\beta\gamma$-heterodimer mutant (**Fig. S11**).

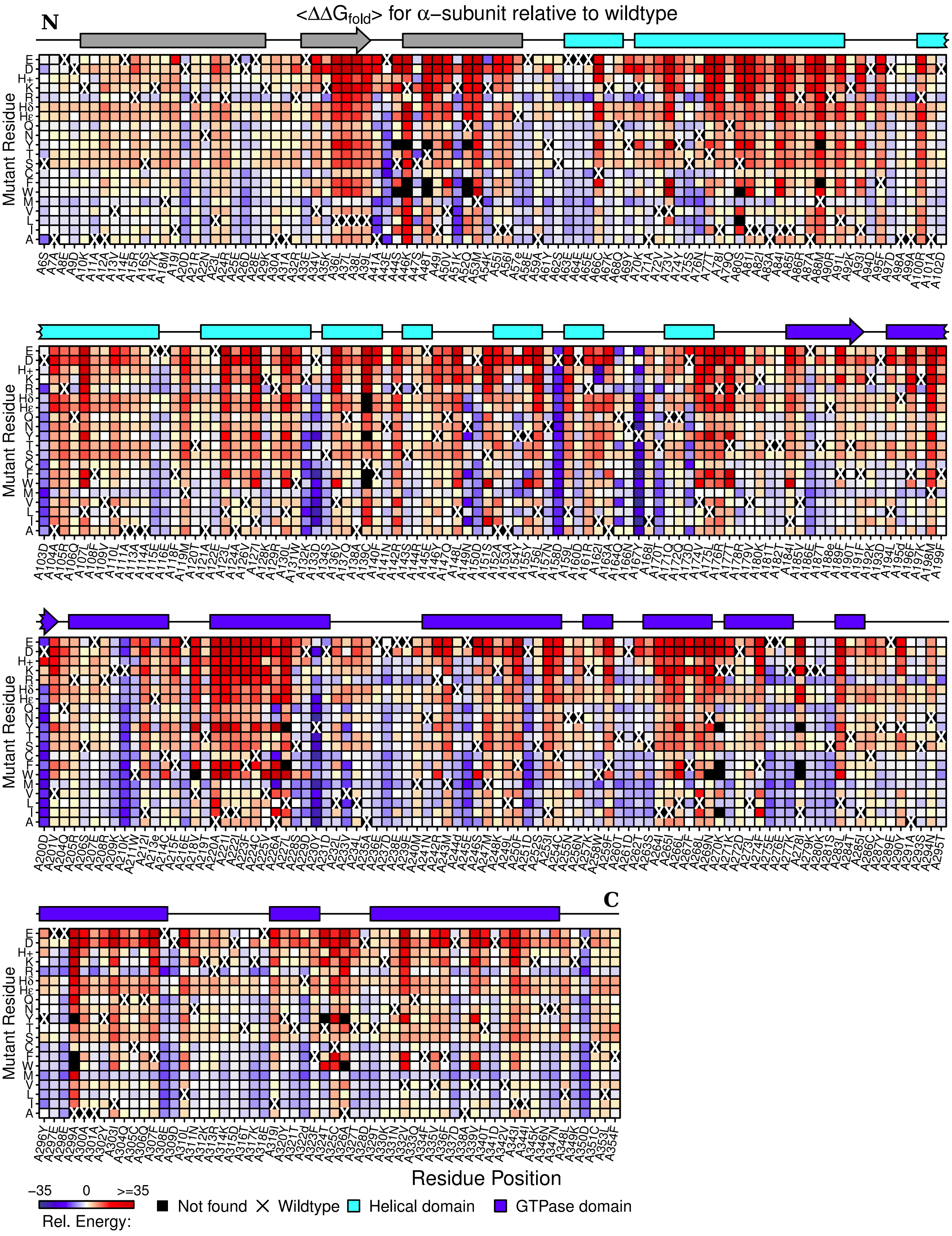

Figure S5: Stability ($\langle \Delta\Delta G_{fold} \rangle$) of G$\alpha$ sequences are organized according to position in the subunit and mutations are arranged according to amino-acid properties. The energy of each mutant (in kcal/mol) is referenced to the wild-type structure prior to averaging over the ensemble of states.

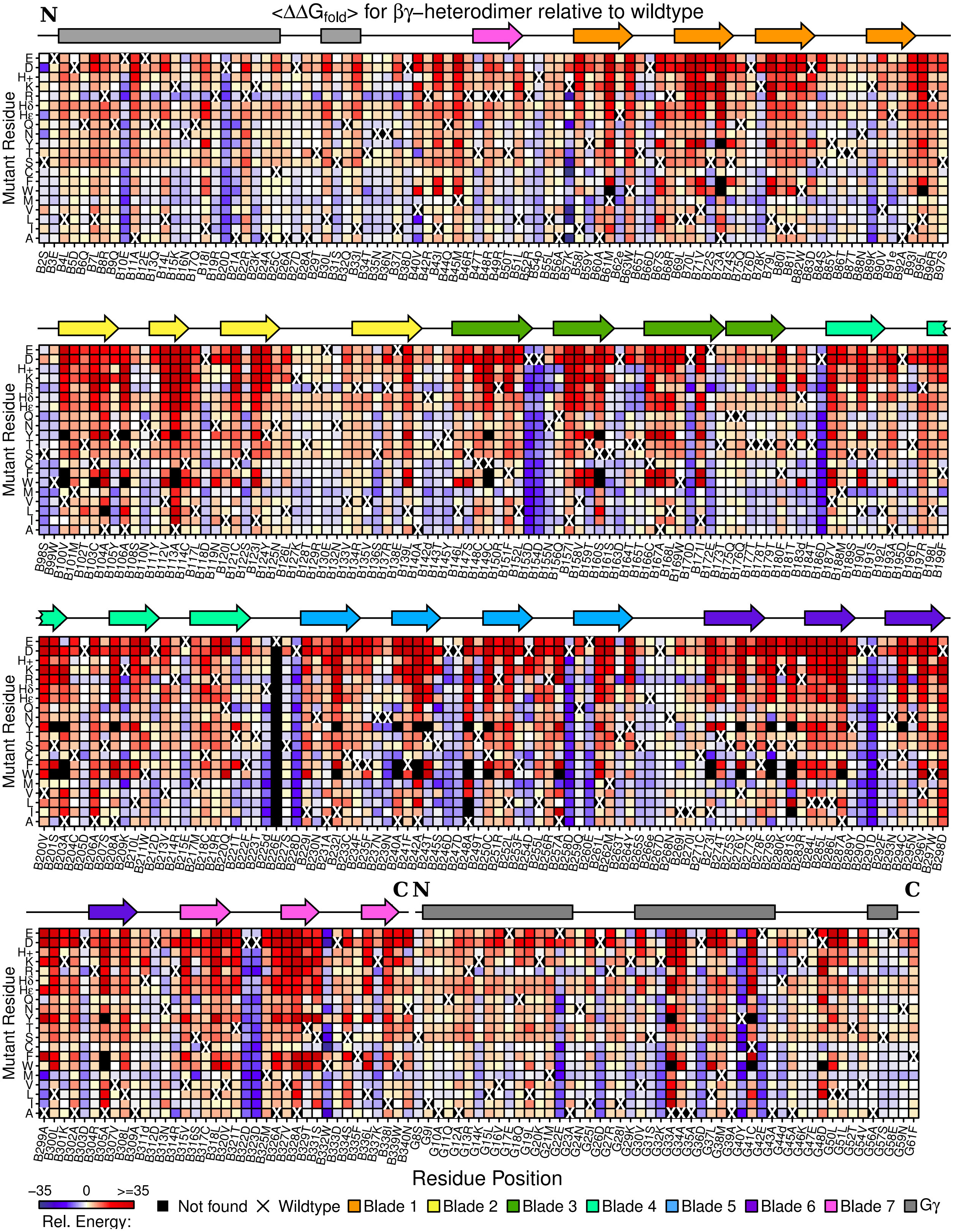

Figure S6: Stability ($\langle \Delta\Delta G_{fold} \rangle$) for the complete heterodimer is separated according to protein chain: Gβ is organized in repeating propeller blades, indicated by secondary structure illustrations, and Gγ follows. Wild-type amino acids are distinguished from mutations for reference, and mutant amino acids are ordered according to side-chain properties; favorable (blue) and unfavorable substitutions (red) can be identified quickly from this subset of protein sequence space.

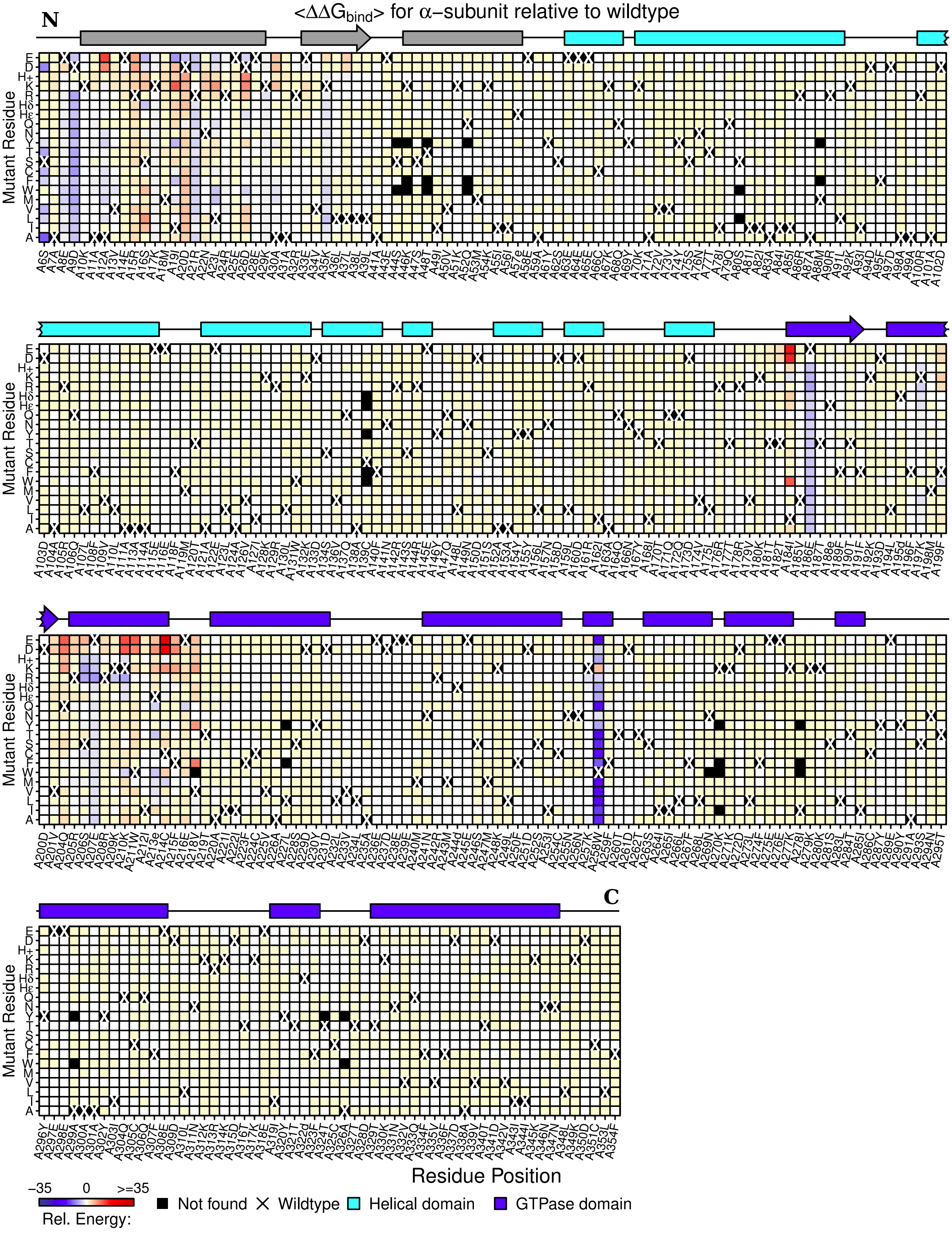

Figure S7: Average energy referenced to wild type is shown here for $\Delta\Delta G_{bind}$ for the α-subunit, based on how it binds to the $\beta\gamma$-heterodimer. Most mutations have a neutral effect, hence the profile is largely dominated by yellow tones.

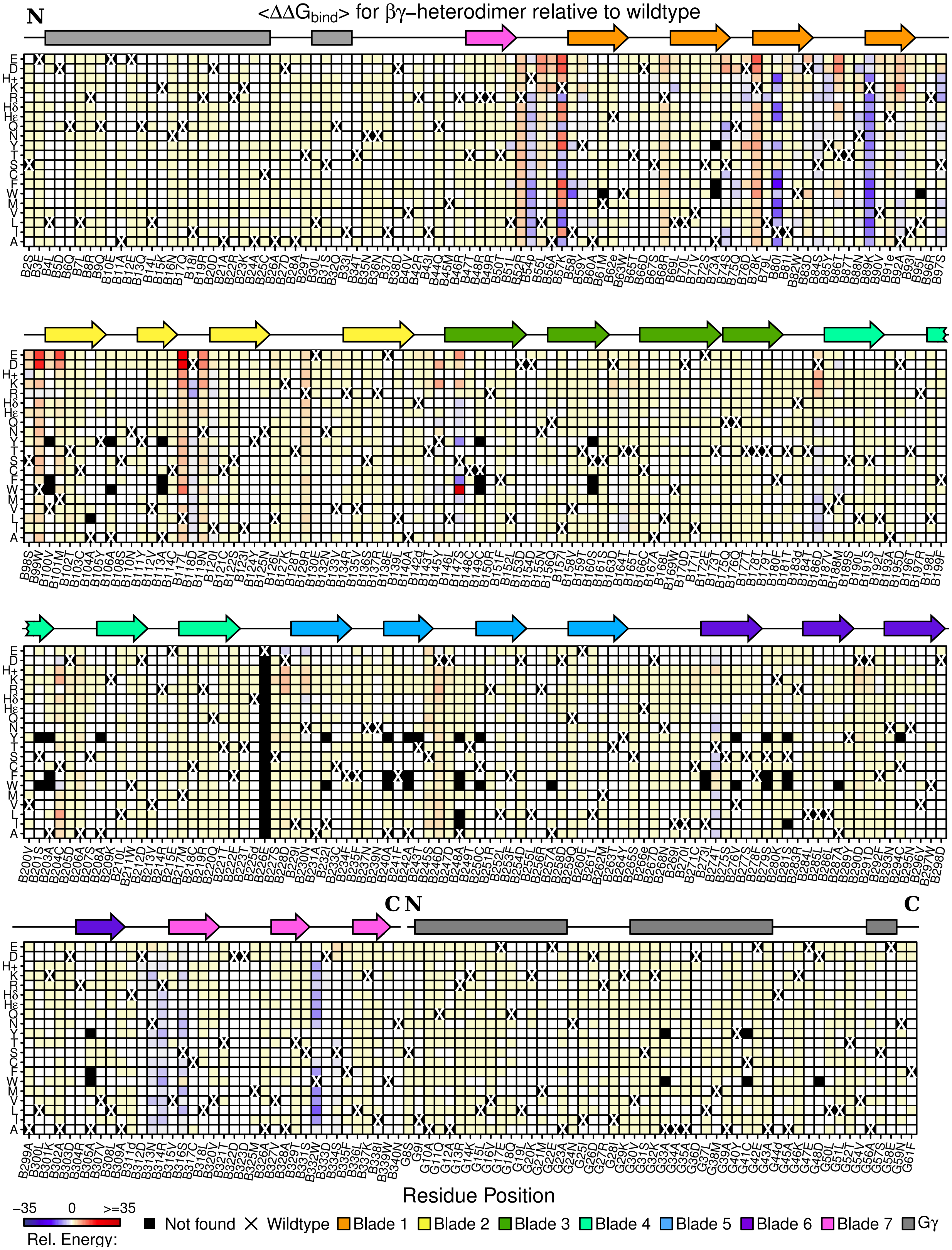

Figure S8: Average energy referenced to wild type is shown here for $\Delta\Delta G_{bind}$ for the $\beta\gamma$-heterodimer, based on its binding interactions with Gα. Most mutations have a neutral effect, as indicated by the yellow tones.

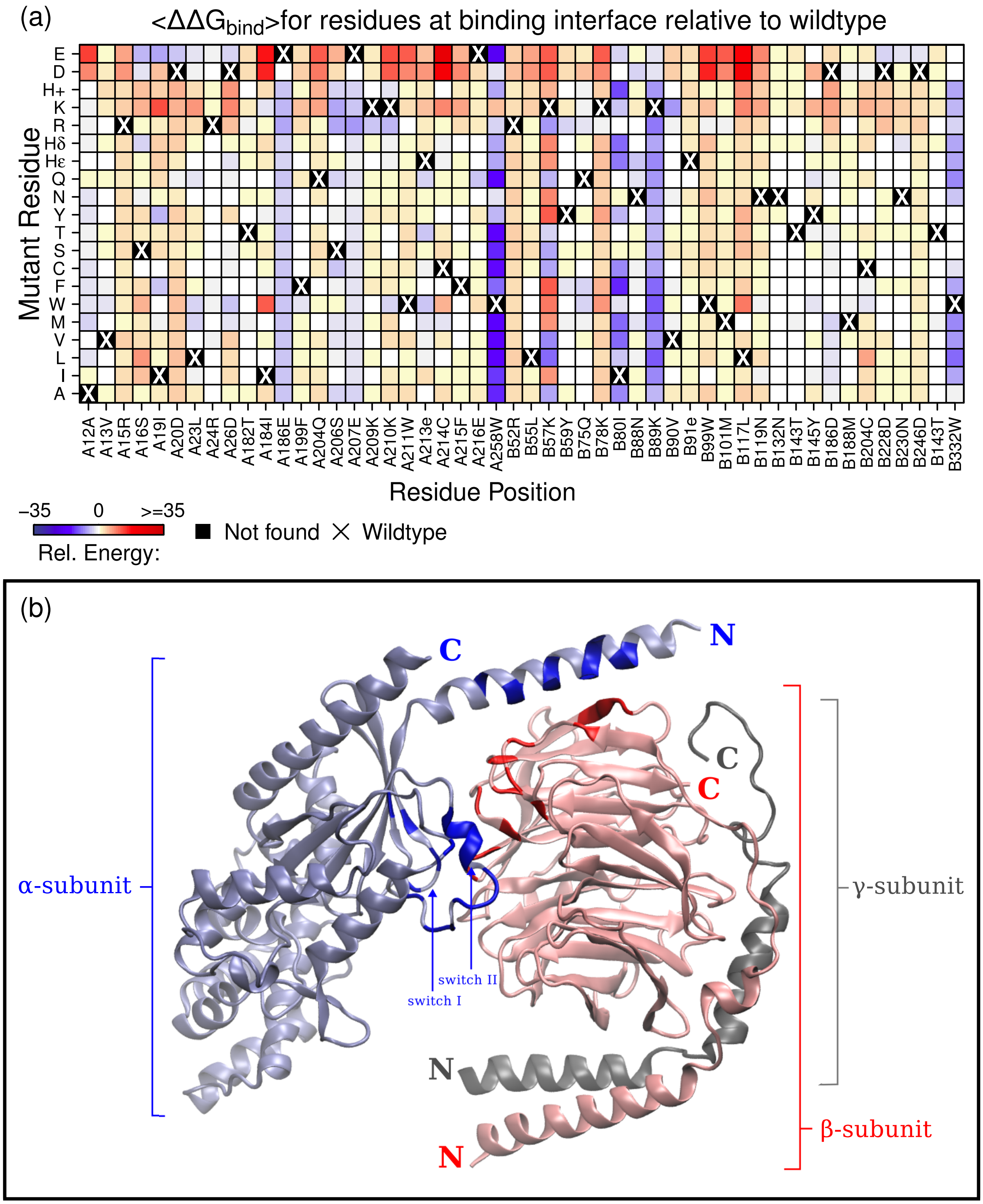

Figure S9: Most positions show no change in binding energy after mutation. (a) Here, the subset of positions with noticeable energetic variation are shown. (b) Structurally, these positions correspond well with the switch II region and amino terminus of Gα, both of which are known to bind with the $\beta\gamma$-heterodimer. Residues involved in binding according to (a) are shown in blue or red, superimposed onto the light blue-gray and light red subunits of the heterotrimer.

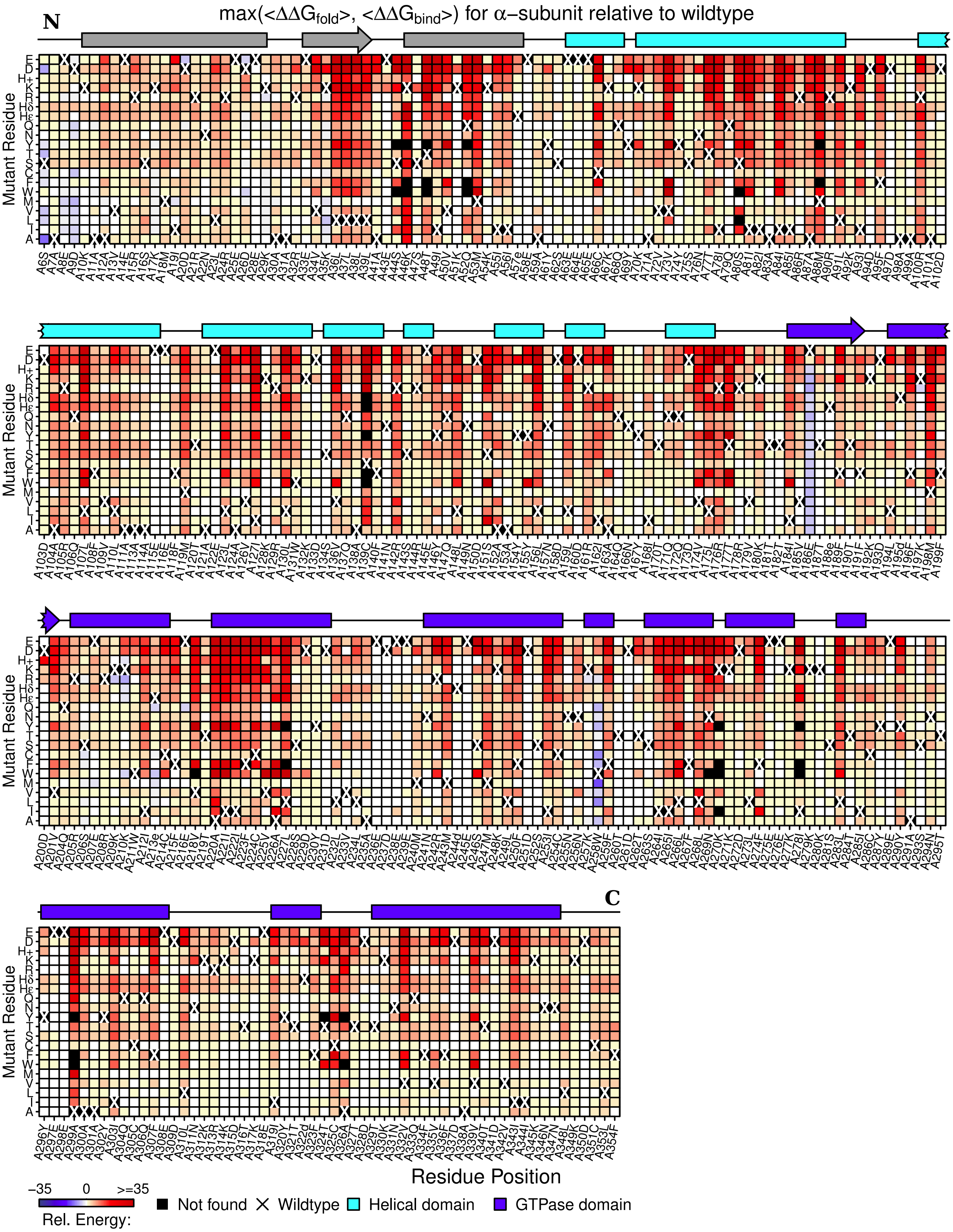

Figure S10: Proteins must stabily fold and also bind to interaction partners. Here, the maximum energy of either term is shown; most worst-case scenario mutations are either neutral or unfavorable, as seen in yellow and red, respectively.

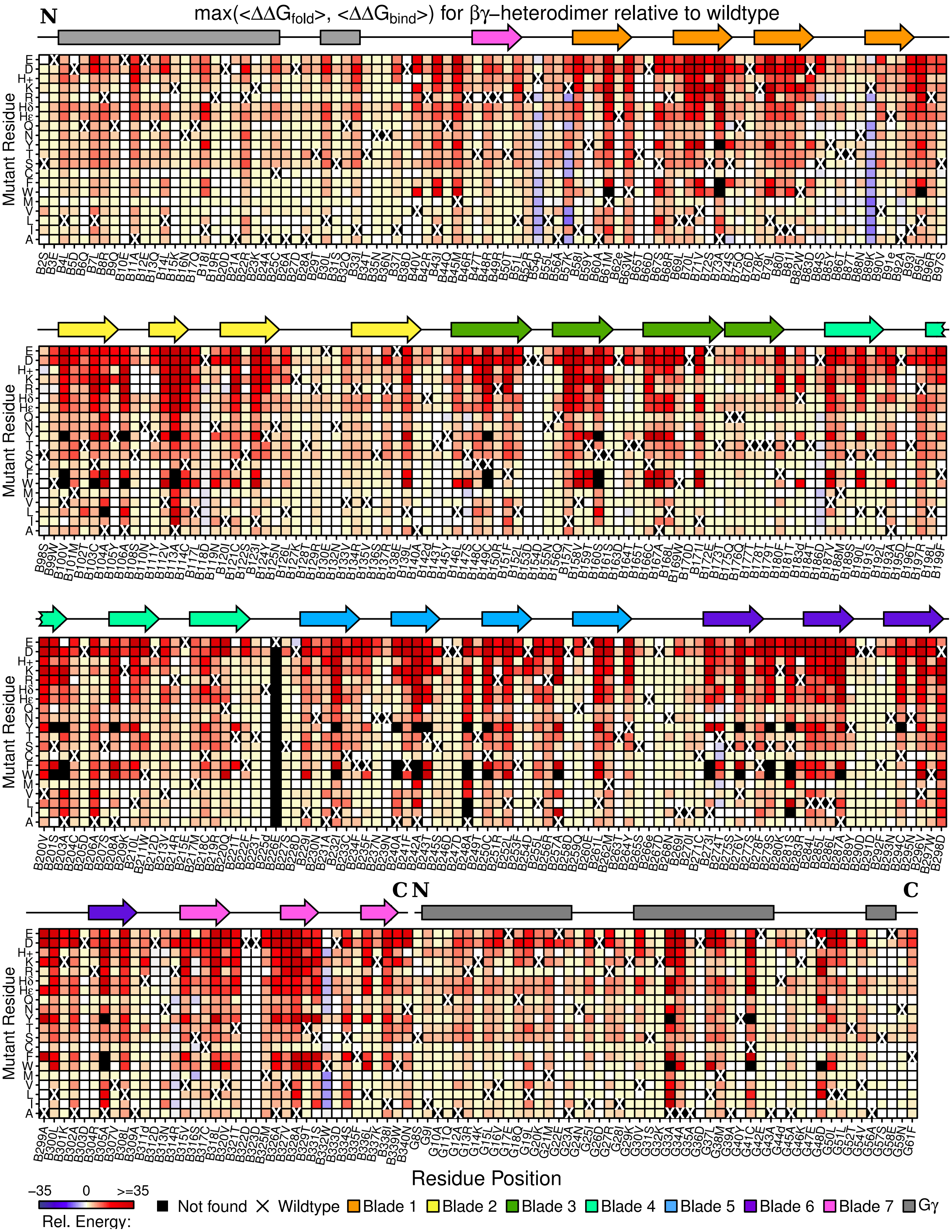

Figure S11: Each value shown is the worst of either stability or binding energy terms, since a functional protein must satisfy both requirements. The majority of mutations are either neutral or unfavorable relative to the wild-type sequence, as seen in yellow and red, respectively.

5 Completeness of computational sampling

Sufficiency of sample size. Including additional backbone conformations for performing DEE/A* is expected to improve the accuracy of our sampling, and these data can show how mutations behave consistently over an ensemble of structures. To determine whether or not a sufficient number of representations were used in our analysis, we considered the average energy difference between sequential subsets of intervals. Only the energy data for stability are considered in this analysis, since variance in binding energy is naturally low due to a small number of positions actually involved in binding. From our data, we defined the following subsets of conformations (**Table S4**):

Table S4: Structures are listed in intervals according to nanosecond in simulation.

A:	[1,5]
B:	[1,5] ∪ [48,52]
C:	[1,5] ∪ [48,52] ∪ [98,102]
D:	[1,5] ∪ [48,52] ∪ [98,102] ∪ [148,152]
E:	[1,5] ∪ [48,52] ∪ [98,102] ∪ [148,152] ∪ [198, 202]
F:	[1,5] ∪ [48,52] ∪ [98,102] ∪ [148,152] ∪ [198, 202] ∪ [248, 252]
G:	[1,5] ∪ [48,52] ∪ [98,102] ∪ [148,152] ∪ [198, 202] ∪ [248, 252] ∪ [298, 302]
H:	[1,5] ∪ [48,52] ∪ [98,102] ∪ [148,152] ∪ [198, 202] ∪ [248, 252] ∪ [298, 302] ∪ [346, 350]

The average energy of each sequence was computed for each subset, and we measured how this average energy will change from one subset to another, in alphabetical order by taking the absolute difference between corresponding sequences: $|\langle A\rangle - \langle B\rangle|$, $|\langle B\rangle - \langle C\rangle|$, $|\langle C\rangle - \langle D\rangle|$, and so forth. The difference in energy variation as the number of structural ensembles increased was partitioned into 1-kcal/mol bins, and the distribution of variance across them is measured to summarize how sampling improves the consistency of our energetic data (**Fig. S12**). We found that including additional snapshots could dramatically reduce the number of outliers in our protein sequence space. In practice, eliminating outliers entirely might not be possible, if flexible regions exist in the protein being studied; it is expected that highly flexible proteins will require more structural conformations for analysis.

Selection of rotamer library. We were interested in seeing how well the augmented Dunbrack–Karplus library (±10° to each χ_1- and χ_2-angle) performed by taking DEE/A* results from a single backbone conformation, and applying a Newton–Rhapson energy minimization algorithm to it (**Fig. S15**). Due to the large energetic calculations and number of residues involved in stability, our analysis only focused on changes in $\Delta\Delta G_{fold}$. As indicated in the main text, approximately 14% of the sequences were found to be unfavorable in one approach and favorable in the other. Unfavorable states remained unfavorable in about 60% of the sequences, while favorable sequences remained favorable in approximately 20% of the data. This indicates that only about 7% of over 6000 sequences could be improved in a meaningful way using energy minimization. Discrepancies in energy calculations tend to arise with the aromatic amino acids, or the charged ones (**Fig. S17**). Most of the energetic improvements that arise from off-rotamer sampling are very modest: the majority of energy differences between the two methods were within 5 kcal/mol of each other, prior to any adjustment with effective temperature (**Fig. S16**). Furthermore, minimization of the wild-type structure also contributed to energetic discrepancies between the two methods of calculation, by lowering the energy of the reference state.

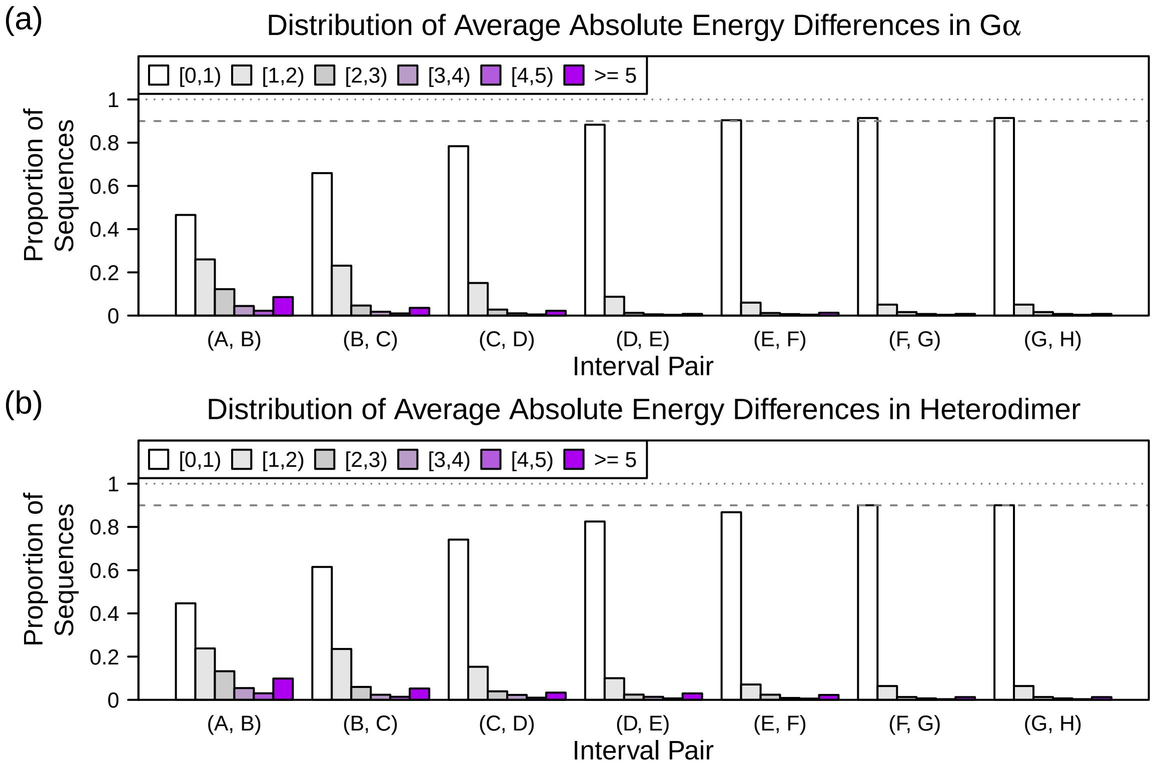

Figure S12: The proportion of data at [0,1)-, [1,2)-, [2,3)-, [3,4)-, [4,5)-kcal/mol or $\geq$ 5-kcal/mol as the number of conformations included in sampling increases (see **Table S4**) is shown. Sequences for Gα and the $\beta\gamma$-heterodimer show similar patterns in convergence, as the number of conformations used to represent an ensemble of states increases.

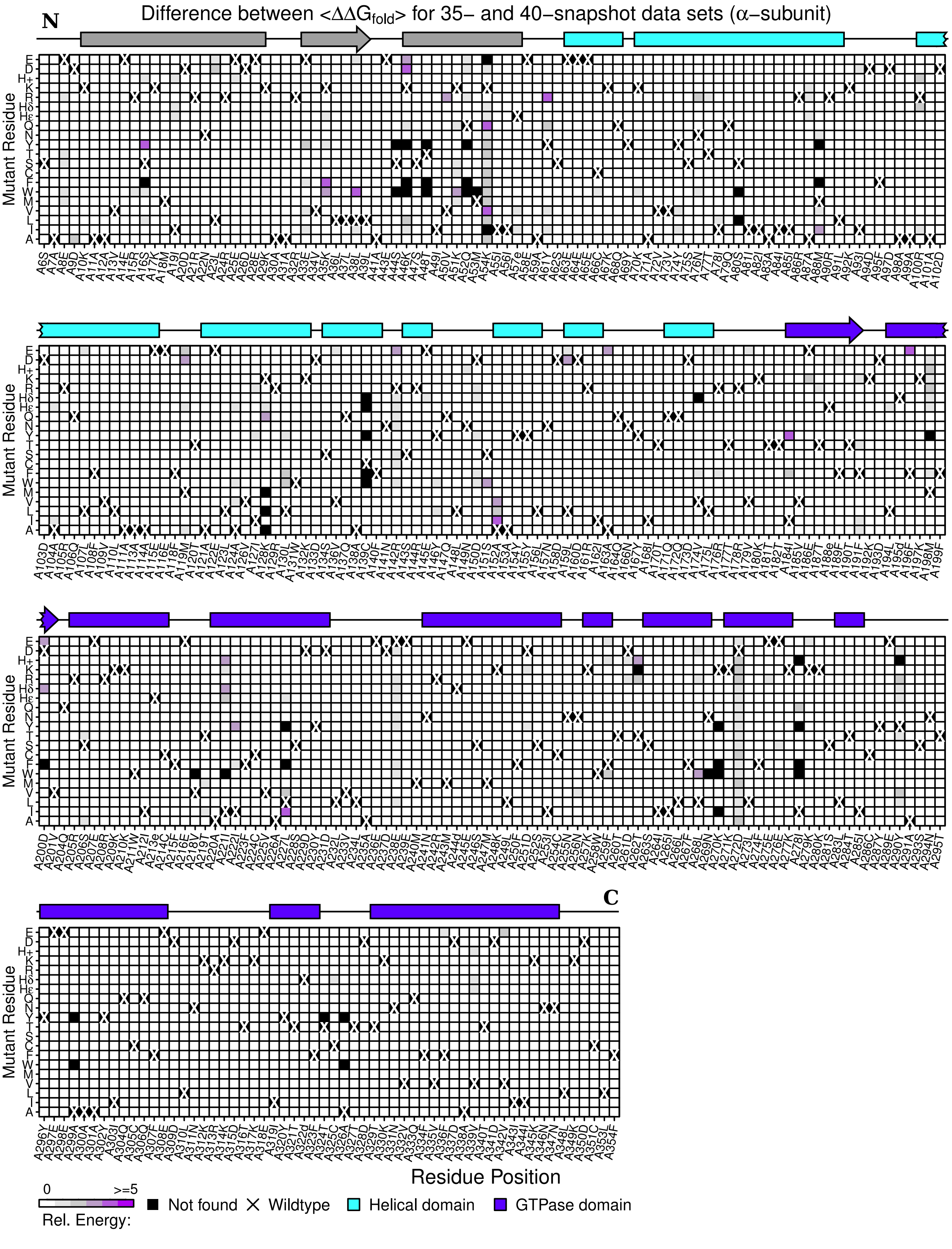

Figure S13: Difference between $\langle\Delta\Delta G_{fold}\rangle$ using a 35- and 40-snapshot data set for Gα is shown here for comparison. Most positions have converged (white and light purple), but there are a few outliers in regions that are harder to sample.

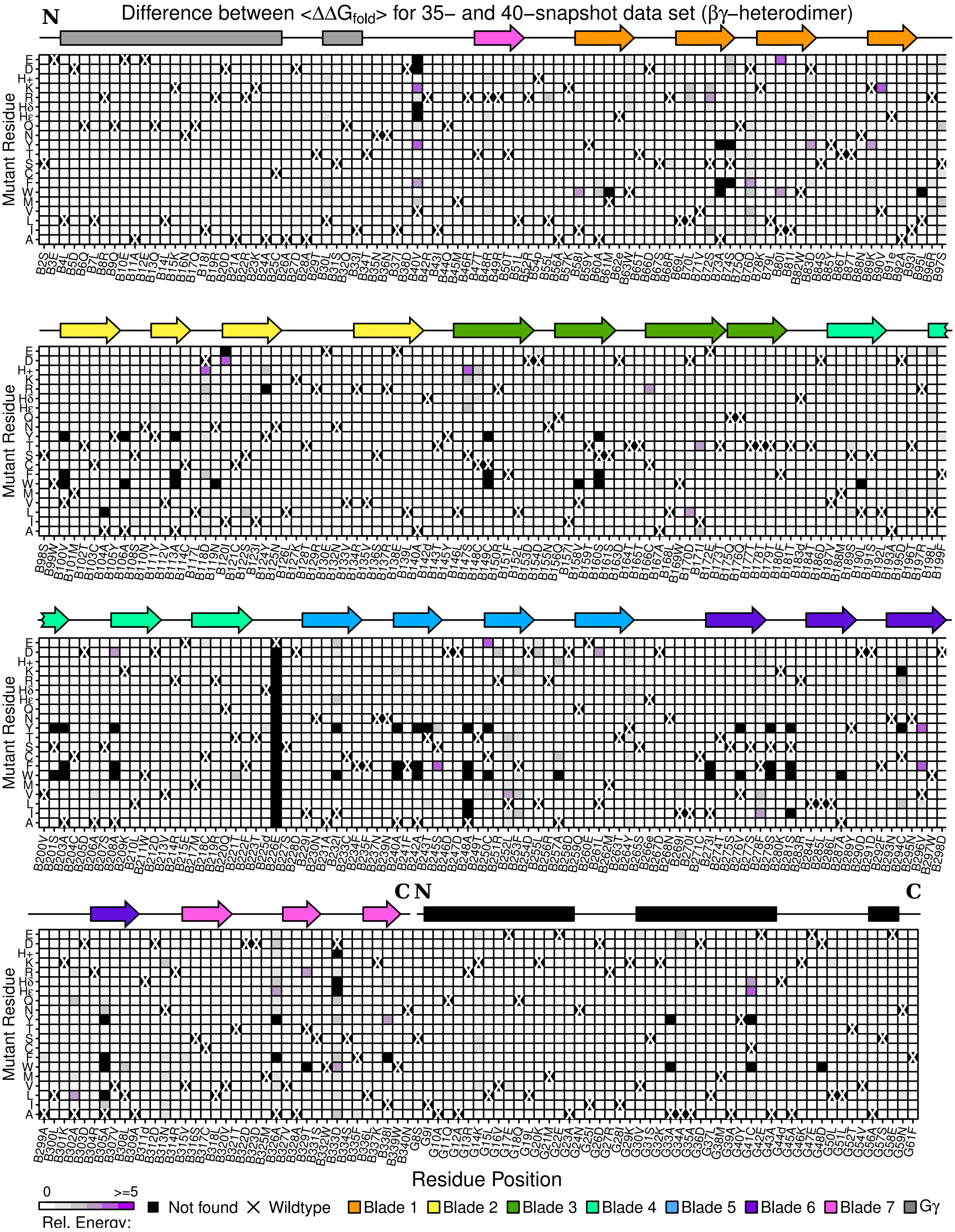

Figure S14: Convergence in the data is found in most of the $\beta\gamma$-heterodimer, shown here as the difference between $\langle \Delta\Delta G_{fold} \rangle$ between a 35- and 40-snapshot data set. Purple regions indicate larger energetic variance, while lighter areas suggest minimal energetic variation.

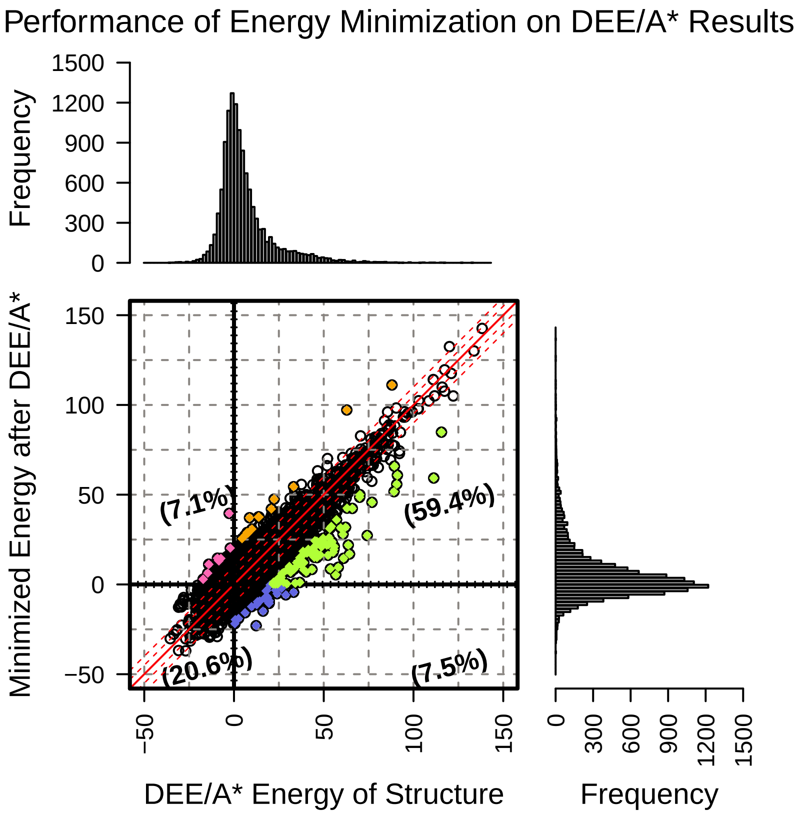

Figure S15: A Newton–Rhapson algorithm was used to perform energy minimization on all DEE/A* structures for one snapshot, to compare the two methods in searching rotamers. The energy landscape is divided into four quadrants to illustrate regions in which energetic improvements can and cannot be achieved when one approach is chosen instead of the other. Red lines indicate where $y = x$, and boundaries that are -10, -5, 5 and 10 kcal/mol from it are shown in dotted red lines.

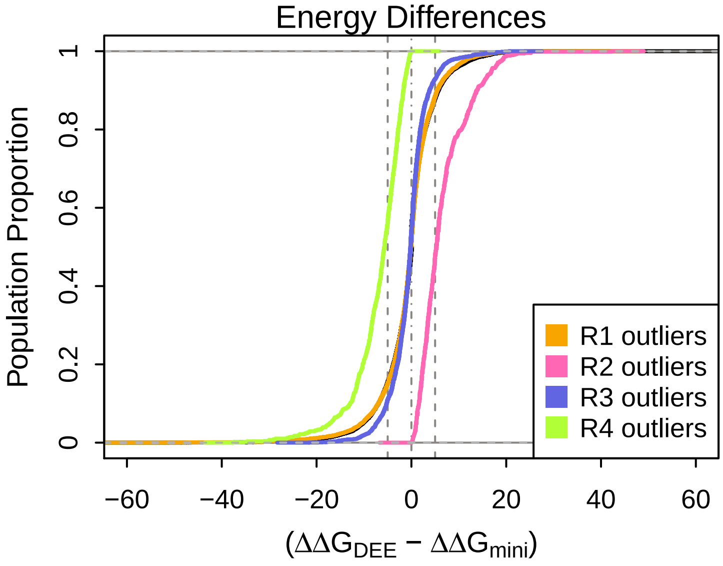

Figure S16: Energy difference between the two protocols was computed for each sequence, and a cumulative distribution of all mutant sequences from a single conformation is shown here, based on this computed difference. Colors correspond to the different quadrants defined previously in **Fig. S15**, and gray lines indicate a difference of -5, 0 or 5 kcal/mol for visual reference.

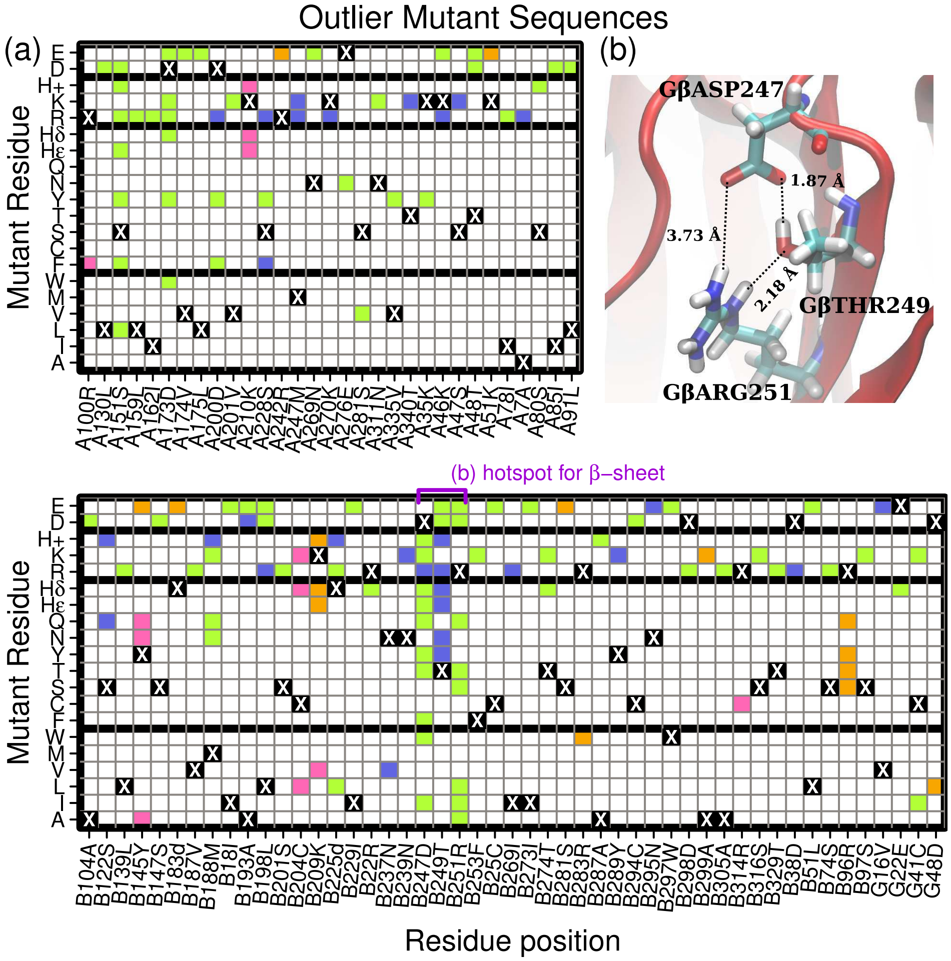

Figure S17: Outlier sequences, those having a 20-kcal/mol difference or greater between the two energy calculation methods, for the four quadrants (defined in **Fig. S15**) are shown and reveal underlying substitutions that yield the greatest discrepancy in (a). An example of a β-sheet with specific side-chain packing requirements that consistently favor wild type is shown in (b).

6 Reflecting evolutionary fitness pressures using DEE/A* approach

Determining frequency of substitution, e_{ij}. Similarity matrices traditionally compute scores as half-bit units. In general, the observed frequencies of amino acid i converted into amino acid j, e_{ij}, is computed, and compared with the expected probabilities of finding each amino acid naturally (p_i and p_j, respectively.) The score, S_{ij}, reflects how closely the observed and theoretical (expected) probability are to each other by taking their ratio (Eq. 1):

$$S_{ij} = 2log_2 \frac{e_{ij}}{p_i p_j} \qquad \text{(Eq. 1)}$$

For computing e_{ij}, the number of sequences found to satisfy the 1.5 kcal/mol energetic cutoff for both structural stability and binding interactions simultaneously was counted, for each (i, j)-pair of amino acids, and this number was normalized by the total number of sequences satisfying this evolutionary pressure. Meanwhile, the distribution of wild-type amino acids in G$\alpha_{i1}\beta_1\gamma_2$ was used to compute independent probabilities, p_i and p_j (Fig. **S18**). Unlike the standard similarity matrices, wild type and the mutant amino acid that it transitions to is clearly defined in DEE/A*, and thus (i, j)- and (j, i)-pairs are unique; these differences cannot be distinguished in PAM and BLOSUM, and so DEE/A* yields a non-symmetric matrix instead.

By taking wild type as the probability distribution of each amino acid found in G$_i\alpha_1\beta_1\gamma_2$ (e.g. the p_i and p_j terms) we convert PAM and BLOSUM scores into the appropriate e_{ij} terms with these as theoretical probabilities. By rearrangement, the expected frequency of substitution can be expressed as a function of these independent probabilities and the score given by the similarity matrix, (Eq. 2).

$$\begin{aligned} log_2 \frac{e_{ij}}{p_i p_j} &= \frac{1}{2} S_{ij} \\ e_{ij} &= p_i p_j 2^{(S_{ij}/2)} \qquad \text{(Eq. 2)} \end{aligned}$$

Correlation between DEE/A* with PAM and BLOSUM) A strong correlation exists between the expected values from DEE/A* and those derived starting from PAM or BLOSUM scores. To start, we looked at protein fitness as the sum of structural stability and binding interactions, at different proportions, and found that a uniform contribution from both aspects of fitness optimizes the correlation between the two different approaches of computing e_{ij}, regardless of the similarity matrix being used for comparison, **Fig. S19** and **Fig. S20**. The exact contribution of each term to overall fitness, of course, cannot be determined; regardless of the defined proportions, the comparison between DEE/A* with these similarities matrices outperforms randomly generated data, either from drawing random values within the boundaries of PAM and BLOSUM scores or by shuffling the entries of each respective similarity matrix. Computations are very quick for these random samples, and we have found from 250, 500 and 1000 trials that the Pearson's correlation coefficient remains unchanged (**Fig. S22**, and **Table S5**). The influence of permuting PAM120 or BLOSUM62

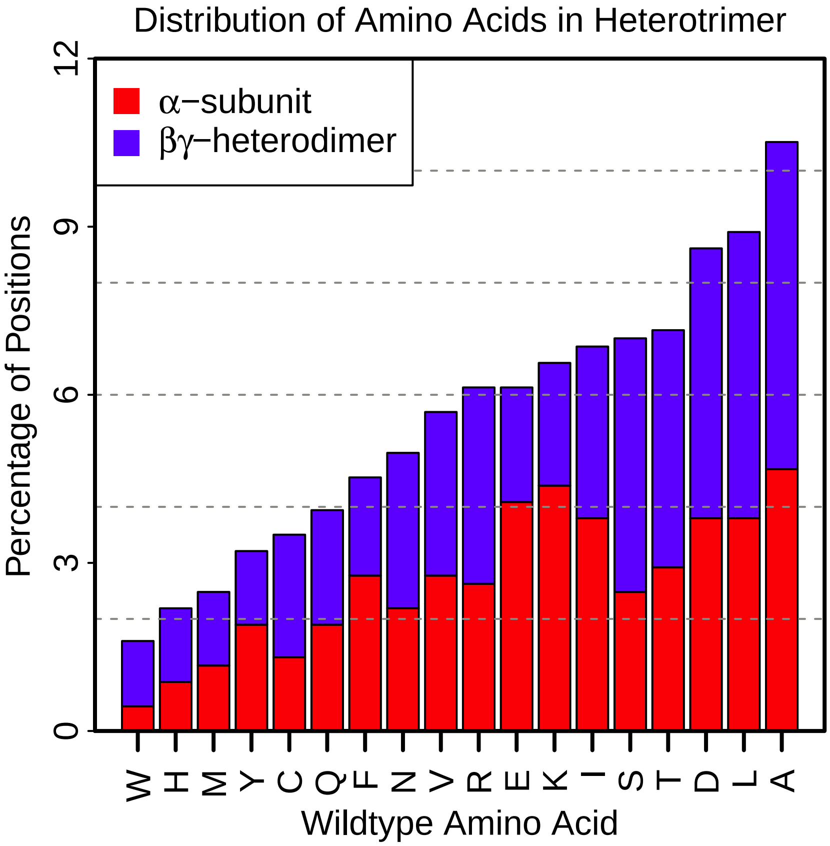

Figure S18: Distribution of wild type amino acids in G$\alpha_{i1}\beta_1\gamma_2$ are normalized and used to provide theoretical probabilities for assessing DEE/A* performance.

is shown by comparing the correlation between the chosen original matrix and the permuted version; a total of 250, 500 and 1000 trials were performed, and the correlation between these two sets of matrices is consistent ($\rho^2 \approx 0.7$ or 0.8 for PAM120 and BLOSUM62, respectively.) Furthermore, while some correlation between DEE/A* and the permuted matrices exist, this relationship is strongest when the (i, j) pairs are clearly identified, suggesting that DEE/A* can discriminate between amino acids well (**Fig. S21**).

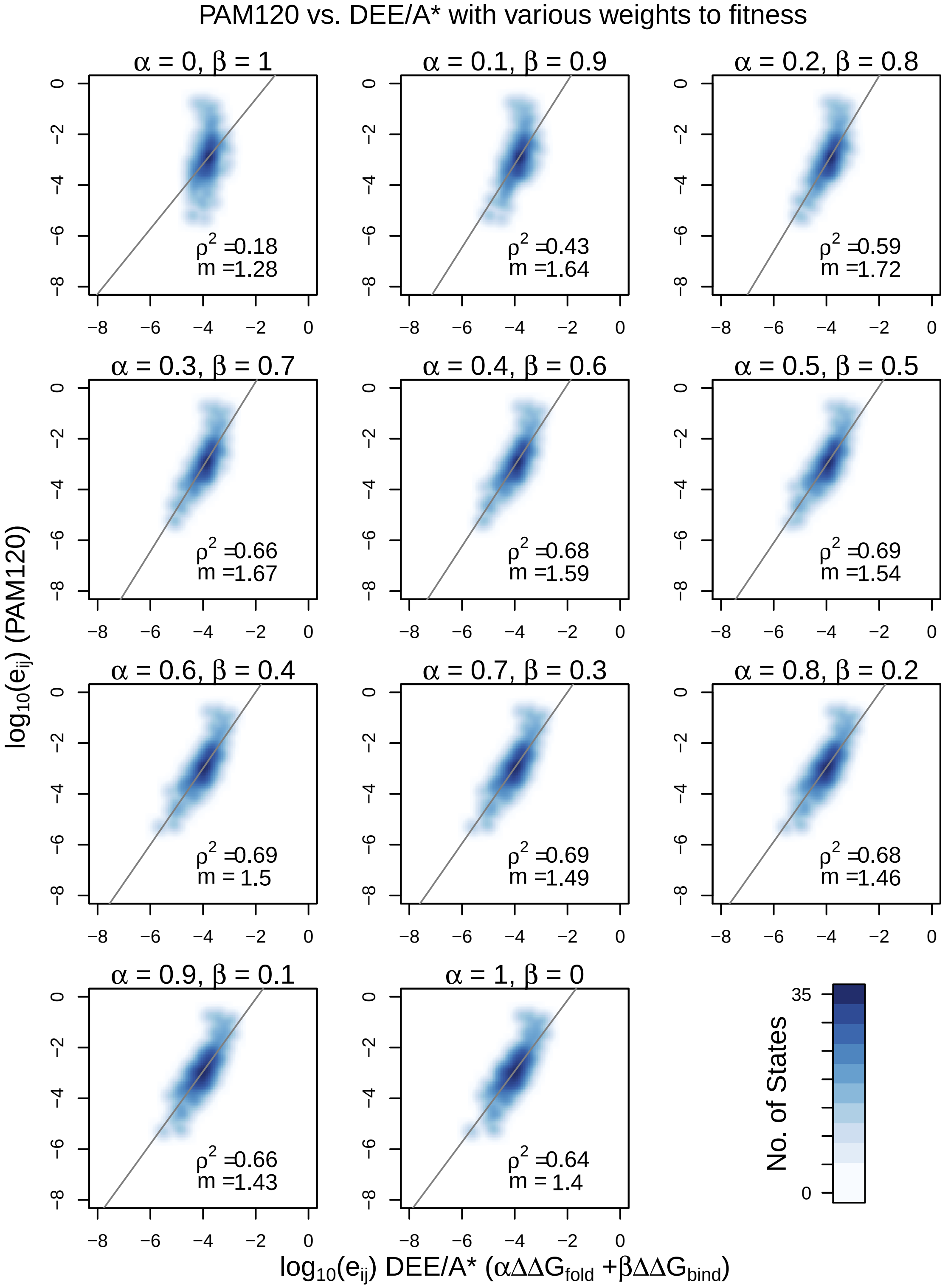

Figure S19: Different proportions of stability and binding were used to define the energetic criteria for survival. Correlation between the expected rate of substitution of amino acid i with j, e_{ij}, is compared between PAM120 and DEE/A* data. Pearson's correlation coefficient and the slope of the least-squares fit are included. The best-fit line is shown in black.

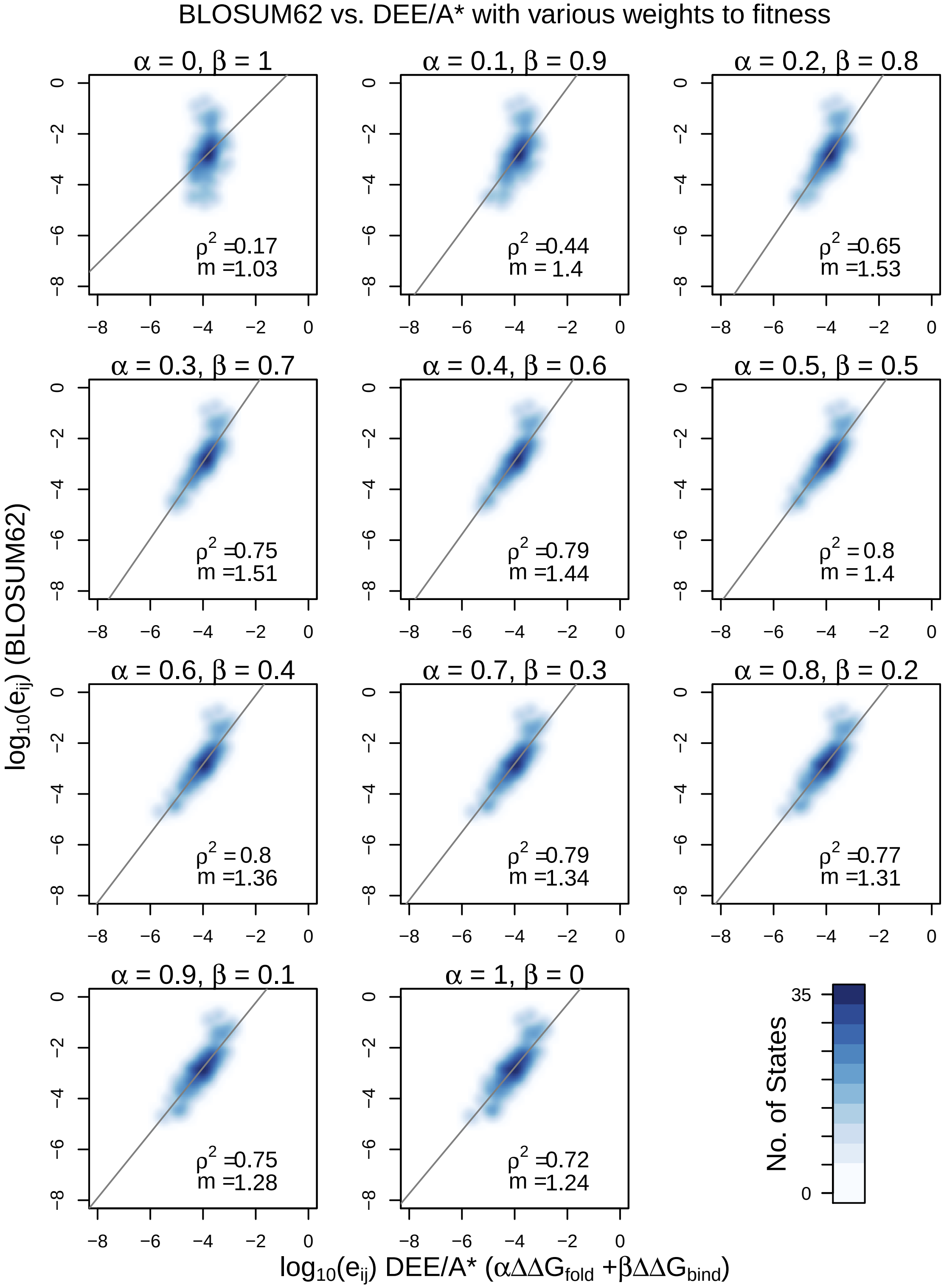

Figure S20: Starting with BLOSUM62 scores, the expected frequency of finding amino acid i replacing j, e_{ij}, was calculated with DEE/A* data, for different combinations of energy contribution from stability and binding interactions. Analogous values were computed from BLOSUM62, so that the two sets of substitution rates can be compared. Pearson's correlation coefficient as well as the slope of the least-squares fit is shown for each; the best-fit line is also drawn.

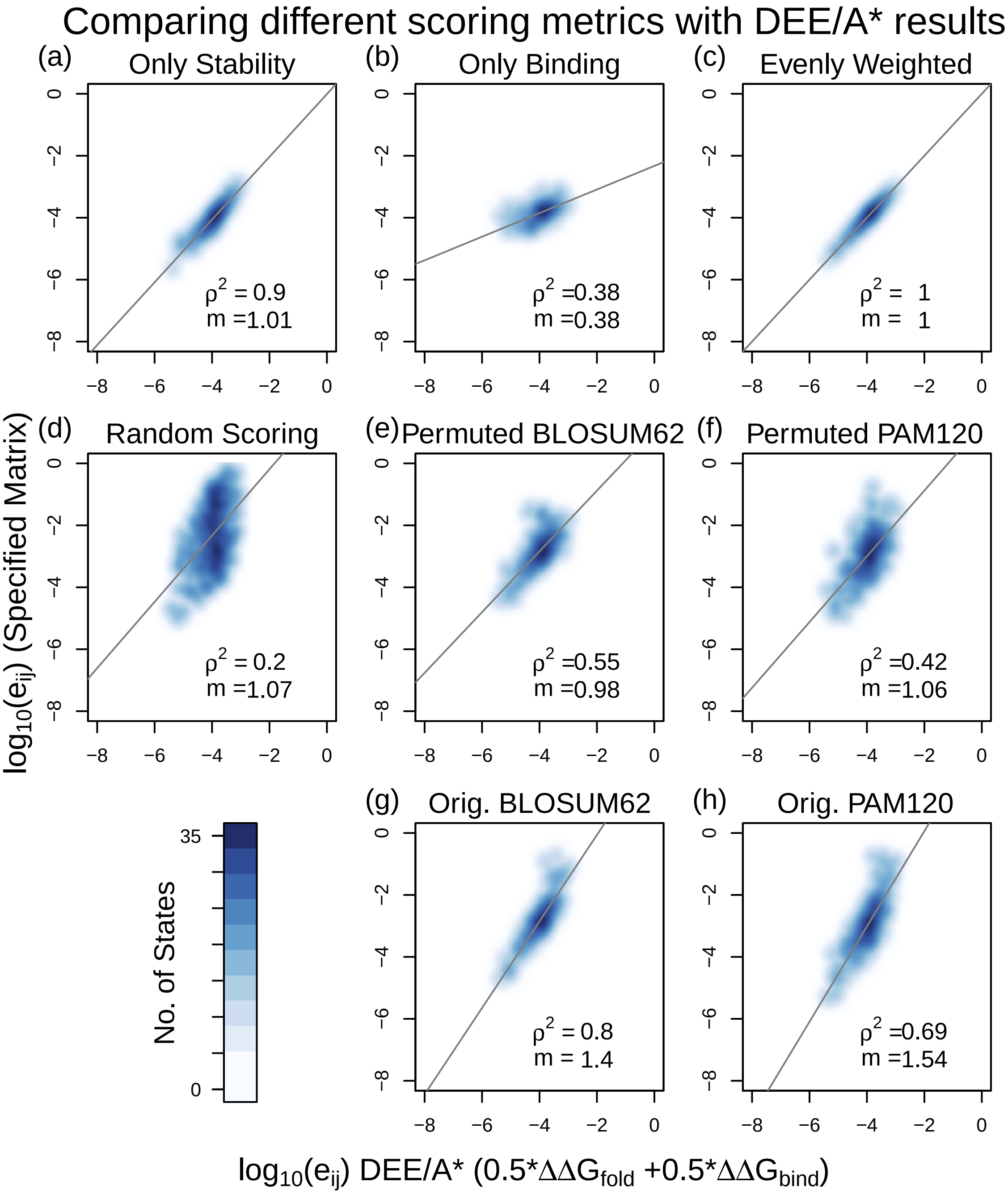

Figure S21: Assuming a uniform contribution from stability and binding, expected frequencies of substitution, e_{ij}, were compared to (from left to right) scores: (a) accounting only for stability; (b) accounting only for binding; (c) in which stability and binding are evenly weighted (50–50); (d) generated from a uniform distribution, bounded by $\max(BLOSUM62, PAM120)$ (e) permuted BLOSUM62 matrix; (f) permuted PAM120 matrix; (g) original BLOSUM62 matrix; and (h) original PAM120 matrix.

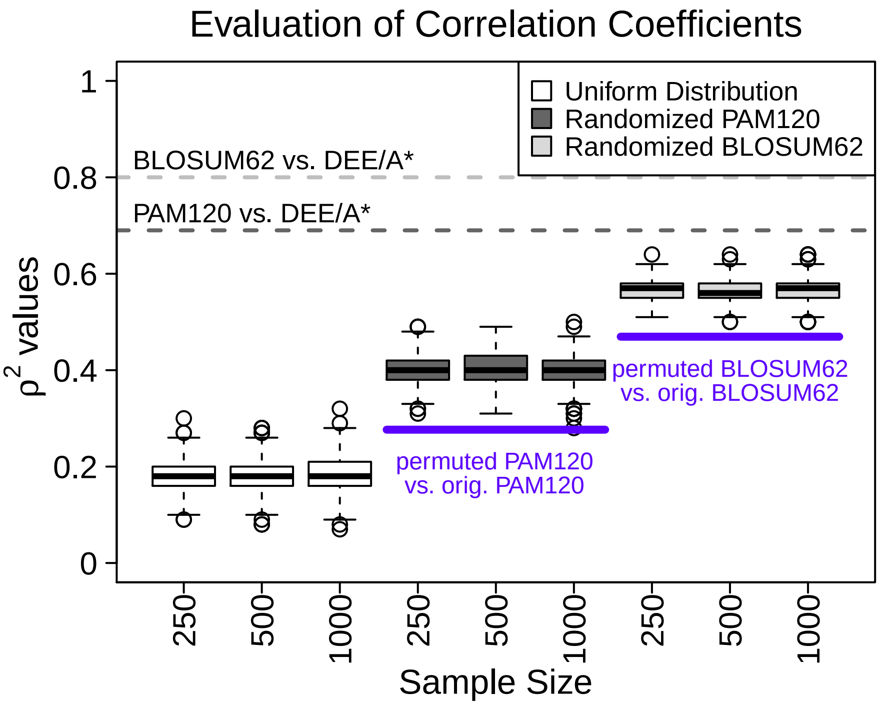

Figure S22: The Pearson's correlation coefficient between DEE/A* data (with 50–50 distribution between stability and binding interactions) and permuted matrix from either a random distribution (white), PAM120 (gray) or BLOSUM62 (light gray) was computed for samples of size $n = 250, 500$ and 1000. Distributions are generally consistent within each family of distributions. Dotted lines indicate the correlation measured between the original PAM120 (gray) or BLOSUM62 (light gray) with DEE/A* based on a 50–50 contribution from each aspect of fitness, as seen in **Fig. S19** and **Fig. S20**. Solid, indigo lines indicate the correlation between either PAM120 or BLOSUM62 with the permuted version of itself.

Table S5: Average Pearson's correlation coefficient between DEE/A* and randomly generated data for various sample sizes. Number of samples given by n.

Randomized matrix	$n = 250$	$n = 500$	$n = 1000$
Uniform distribution	0.18 ± 0.03	0.18 ± 0.04	0.18 ± 0.04
PAM120 values	0.40 ± 0.03	0.40 ± 0.03	0.40 ± 0.03
BLOSUM62 values	0.57 ± 0.02	0.56 ± 0.02	0.57 ± 0.02

Table S6: Average Pearson's correlation coefficient between original similarity matrix and the permuted version. Number of samples given by n.

Similarity matrix	$n = 250$	$n = 500$	$n = 1000$
PAM120	0.27 ± 0.03	0.28 ± 0.03	0.28 ± 0.03
BLOSUM62	0.46 ± 0.03	0.47 ± 0.03	0.47 ± 0.03

7 Comparisons with established experimental data

Alanine scan of Gα. A full-scale alanine scan was performed by Sun, *et al.* for Gα in which thermal stability was measured relative to wild type (ΔT_m) for all single mutants. The corresponding alanine mutations from our DEE/A* were used for comparison to these data. In our analysis, a positive outcome was defined as a mutation that was destabilizing relative to wild type ($\Delta\Delta G_{fold} > +1.5$ kcal/mol), which would suggest that native interactions were important for structural stability. We quantified the proportion of

- true positives ($\Delta\Delta G_{fold} > 1.5$ kcal/mol & $\Delta T_m \leq -2°C$)
- true negatives ($\Delta\Delta G_{fold} \leq 1.5$ kcal/mol & $\Delta T_m > -2°C$)
- false positives ($\Delta\Delta G_{fold} > 1.5$ kcal/mol & $\Delta T_m \leq -2°C$) and
- false negatives ($\Delta\Delta G_{fold} \leq 1.5$ kcal/mol & $\Delta T_m > -2°C$)

The relationship between thermal stability and our DEE/A* calculations is illustrated in **Figure S23**, and we report additional statistics in **Table S7**. To provide a basis for comparing these proportions, the mutation free energy for DEE/A* and thermal stability were randomized for a total of 5000 independent trials, then compared again to measure the proportion of different outcomes (**Table S7** & **Fig. S24**). We found that the observed number of correctly identified outcomes (true positives and true negatives) were consistently higher than expected, and that the proportion of incorrect predictions (false positives and false negatives) were consequently much lower. These results were further quantified in terms of sensitivity and specificity (**Table S8**), and we found that sensitivity was much higher than the randomized energy data. Here, there is a natural trade-off with specificity, which was found to be relatively lower in comparison to the shuffled energy values.

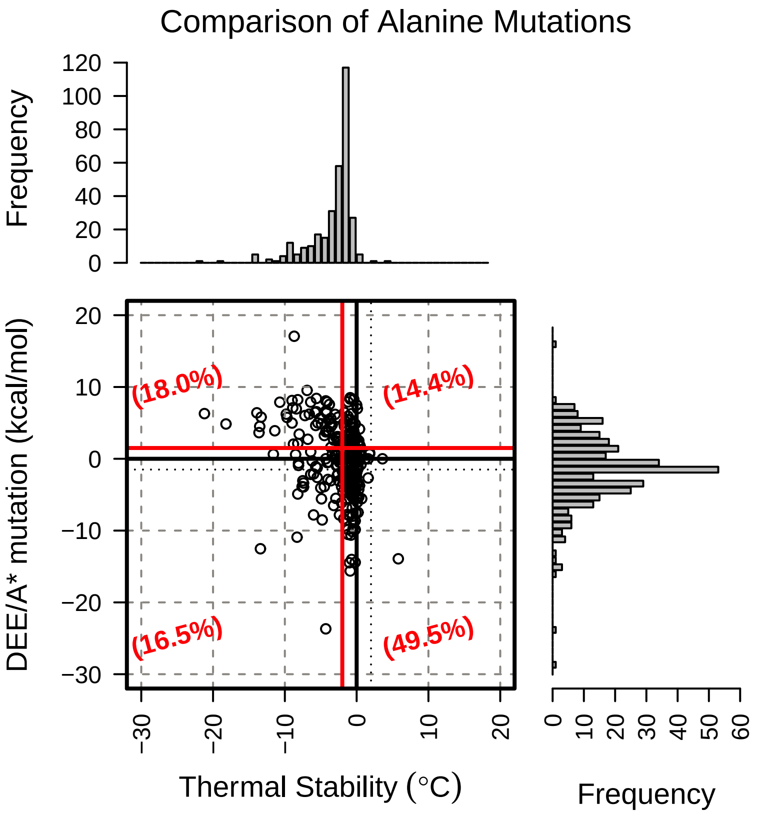

Figure S23: The energetic difference between each wild-type residue and corresponding alanine mutation in the stability of $G\alpha$–GDP are shown here. The thermal stability (ΔT_m) between alanine mutations and wild type measured by Sun, *et al.* were used for comparison with the alanine mutants from our DEE/A* computations for structural stability. The red vertical lines correspond to $-2°C$, the threshold used by Sun, *et al.* to indicate that native interactions were important in stabilizing $G\alpha$–GDP, while the red horizontal line is the cutoff used in our calculations to indicate that a substitution is unfavorable relative to the wild-type residue. These red lines are used to define the four different quadrants, and the percentage in each region is shown in red text within parentheses. (See also **Fig. S7**.)

Table S7: Correlation between alanine mutants for DEE/A* calculations and thermal stability data measured by Sun, *et al.* was quantified by the proportions of true positives, false positives, true negatives and false negatives. The sensitivity and specificity of our computational approach were also calculated. (See also **Fig. S23**.)

		DEE/A* calculations	
		$\Delta\Delta G_{fold} > +1.5$ kcal/mol	$\Delta\Delta G_{fold} \leq 1.5$ kcal/mol
Thermal Stability	$\Delta T_m \leq -2°C$	18.0% ($n = 59$)	14.4% ($n = 47$)
	$\Delta T_m > -2°C$	16.5% ($n = 54$)	49.5% ($n = 162$)
		Sensitivity: 0.52	Specificity: 0.78

Table S8: After randomizing the DEE/A* alanine mutants, the proportion of true positives, false positives, false negatives and true negatives were measured. The reported values are the average and standard deviations for each category after 5000 independent calculations. The sensitivity and specificity have also been measured. (See also **Fig. S24**.)

		DEE/A* calculations (randomized)	
		$\Delta\Delta G_{fold} > +1.5$ kcal/mol	$\Delta\Delta G_{fold} \leq 1.5$ kcal/mol
Thermal Stability	$\Delta T_m \leq -2°C$	$11.0 \pm 1.2\%$	$21.4 \pm 1.2\%$
	$\Delta T_m > -2°C$	$23.1 \pm 1.2\%$	$42.8 \pm 1.2\%$
		Sensitivity: 0.32 ± 0.04	Specificity: 0.67 ± 0.02

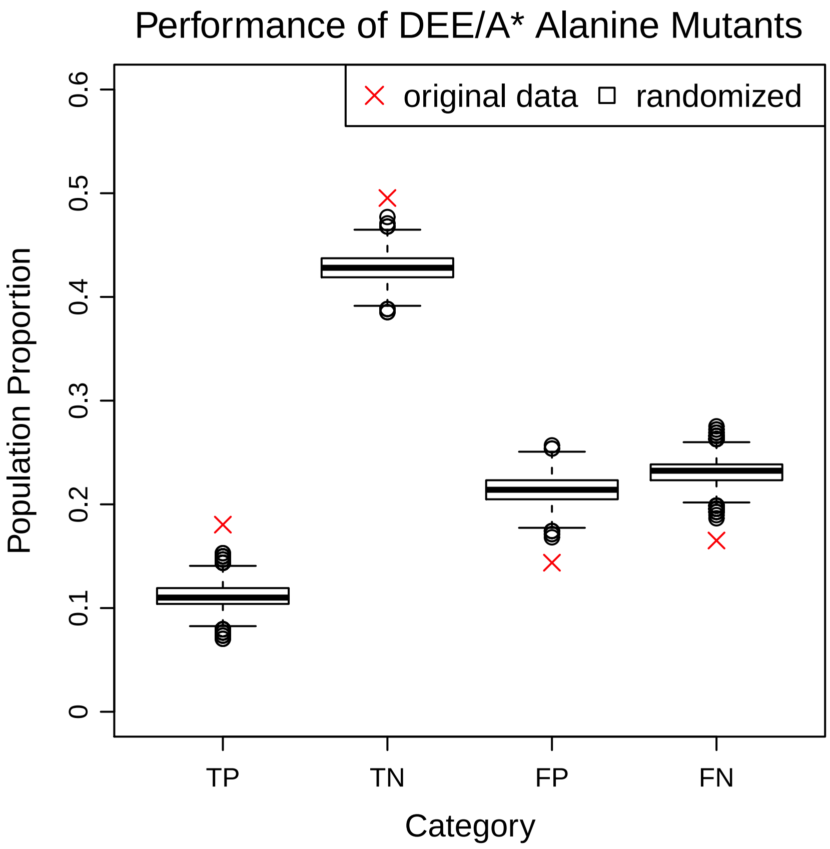

Figure S24: The proportions of different true positive, true negative, false positive and false negative outcomes were calculated, and the distribution of these values from randomized data are shown using box-and-whiskers for each outcome type. Red X's are included to represent the computed values of each category from the initial comparison of DEE/A* and thermal stability.

Oncogenic point mutations in Gβ. The original list of point mutations was provided by Yoda, *et al.* in the supplementary information of their publication. This included a few mutations that were not suitable for comparison, such as mutations to glycine or using splice variants, which are not covered by our computational protocol, and thus excluded from analysis. Mutations to histidine were taken as the average mutation free energy of all three possible histidine states (δ-, ε- and doubly-protonated) as modeled by CHARMM. A total of 36 point mutations were available for analysis, and are listed in **Table S9**. From our DEE/A* calculations for each point mutation listed, the stability and binding interactions relative to wild type were used to categorize mutations as gain-of-function, neutral or loss-of-function. These energetic cutoffs were based on previous definitions using ±1.5 kcal/mol. Mutations were assessed as independent aspects of fitness, and also simultaneously. For the latter, we measured the maximum magnitude of either structural stability or binding interactions. (See main text.)

Table S9: The original list of GNB1 mutations was compiled and amended by Yoda, *et al.* A condensed version of point mutations that could be compared to our DEE/A* data (*e.g.* not a glycine mutation, splice variant or non-specific mutation) is provided here, along with the computed DEE/A* values for Gβ structural stability ($\Delta\Delta G_{fold}$) and binding interactions ($\Delta\Delta G_{bind}$). References may be from Yoda, *et al.*, COSMIC, cBioPortal or a specific publication, in which the PubMed identification number is provided.

Mutation	$\Delta\Delta G_{fold}$	$\Delta\Delta G_{bind}$	References	Mutation	$\Delta\Delta G_{fold}$	$\Delta\Delta G_{bind}$	References
A11V	−4.0	0.0	COSMIC	D118Y	−7.3	−7.9	COSMIC
R19L	−11.5	0.0	COSMIC	S147A	−14.7	−3.2	Yoda, *et al.*
A21S	2.1	0.0	COSMIC	B177K	−10	0.0	24220272
Q32K	0.6	0.0	COSMIC	S191C	−13	0.0	COSMIC
T47M	−11.6	0.0	COSMIC	D205N	−14.7	0.5	COSMIC
p54N	−16.8	−17.0	COSMIC	E215D	−5.9	0.0	cBioPortal
K57E	−1.2	4.3	Yoda, *et al.*	d225L	−24.3	−0.1	23292937
			COSMIC	D228N	−17.9	−7.1	cBioPortal
L57E	−1.2	4.3	Yoda, *et al.*	N230S	−4.4	−11.0	COSMIC
			COSMIC	R256H	5.1	5.1	cBioPortal
			24220272	D258N	−25.4	0.0	COSMIC
			23443460	E260K	−9.8	0.0	COSMIC
K57N	−8.4	−2.2	23443460	M262T	2.8	0.0	COSMIC
K57T	−2.7	−1.9	COSMIC	I269T	−22.4	0.0	COSMIC
K78E	−4.0	0.3	COSMIC	K280N	−5.8	0.0	cBioPortal
			cBioPortal	S281N	−1.1	0.0	COSMIC
K78Q	−10.0	−3.7	Yoda, *et al.*	R283C	−15.5	0.0	COSMIC
I80N	−3.8	−1.8	Yoda, *et al.*	R314H	−5.3	−5.3	23699601
			22343534	A326T	4.8	0.0	cBioPortal
I80T	2.5	−11.1	Yoda, *et al.*				
			COSMIC				
			23699601				
			24220272				
N88D	2.4	−18.1	Yoda, *et al.*				
K89E	−17.9	−17.1	Yoda, *et al.*				
K89T	−20.5	−20.6	24220272				
R96H	0.4	−0.4	COSMIC				

8 Statistical analysis for predictions

The Boltzmann-weighted mean of each mutant sequence was computed to determine the average change in all 40 structural states used. For each position, this provided twenty unique values (one for each amino acid) which summarized all mutational effects. From this vector of numbers, values on the $[-1.5, 1.5]$-kcal/mol interval were assigned zero to represent no change. A 20-dimensional zero vector was thus chosen for the null hypothesis, and the Mann–Whitney–Wilcoxon test was performed using R for every position in the heterotrimer. Computed p-values are shown in **Fig. S25** and grouped according to (1) whether or not position is known to have binding interactions (according to Wall, *et al.*) and (2) whether mutation free energy is based on $\Delta\Delta G_{fold}$ or $\Delta\Delta G_{bind}$.

The findings for data based on binding interactions are discussed in the main text. A detailed list for true positives, false negatives and predicted positions can be found in **Tables S10 & S11**. The calculations for stabilizing interactions, however, suggest that nearly all positions have a meaningful contribution to protein tertiary structure (low p-values). Given that this protein family is highly evolved and that the mutational profiles (**Fig. S5–S11**) suggested that most substitutions are unfavorable, alternative metrics would need to be applied to further separate side chains into varying degrees of involvement.

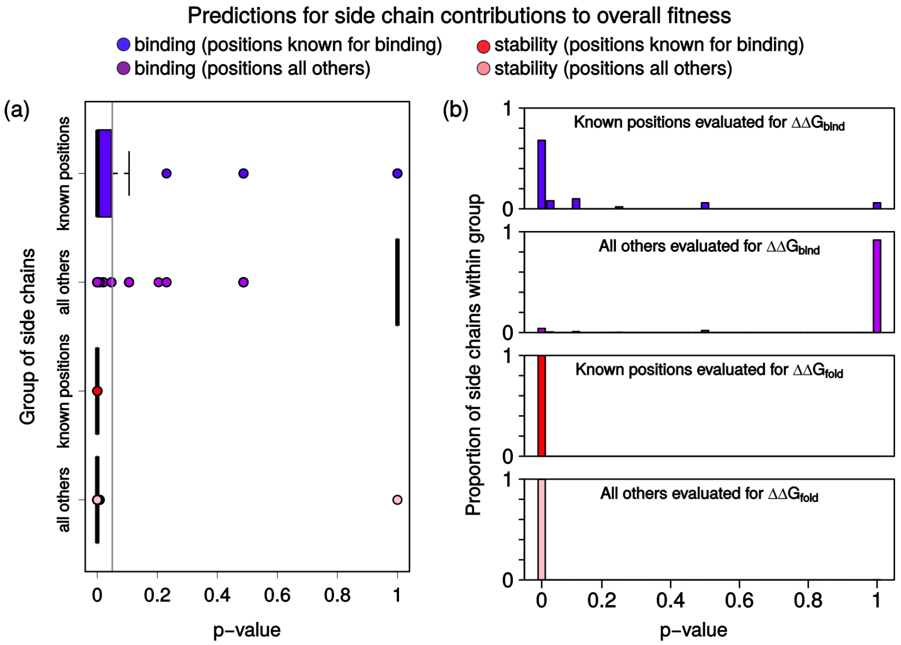

Figure S25: Positions known for binding interactions were separated from all other positions, then mutational differences based on $\Delta\Delta G_{bind}$ and $\Delta\Delta G_{fold}$ were computed (blue and red, respectively.) The analysis was applied to all other positions based on $\Delta\Delta G_{bind}$ and $\Delta\Delta G_{fold}$ (purple and pink, respectively) for comparison. These data are shown as distributions in (a) box-and-whiskers plots and (b) as a histogram to illustrate how the majority of side chains within each subgroup shifts as the premise for analysis changes.

Table S10: The positions known to have binding interactions according to Wall, *et al.* are provided here. The computed *p*-values are the untruncated output from R and based on $\Delta\Delta G_{bind}$ energetic differences from wild type.

Position	*p*-value	Position	*p*-value
A12A	1.453066e-04	B52R	3.046867e-10
A13V	0.4871795	B55L	3.276003e-03
A15R	1.541715e-07	B57K	3.046867e-10
A16S	4.509515e-05	B59Y	4.712405e-02
A19I	1.288433e-05	B75Q	2.019602e-02
A20D	3.351553e-09	B78K	3.046867e-10
A23L	4.359198e-04	B80I	1.288433e-05
A24R	0.4871795	B88N	2.019602e-02
A26D	4.509515e-05	B89K	3.046867e-10
A182T	0.1060291	B90V	0.1060291
A184I	4.359198e-04	B91e	4.712405e-02
A186E	3.351553e-09	B99W	7.708573e-07
A199F	4.712405e-02	B101M	0.1060291
A204Q	4.359198e-04	B117L	2.569524e-08
A206S	4.359198e-04	B119N	1.453066e-04
A207E	2.569524e-08	B132N	1.0000000
A209K	0.2307692	B143T	1.0000000
A210K	1.453066e-04	B145Y	2.019602e-02
A211W	1.228501e-03	B186D	3.276003e-03
A213e	4.509515e-05	B188M	0.4871795
A214C	3.276003e-03	B204C	4.359198e-04
A215F	0.1060291	B228D	0.1060291
A216E	8.316008e-03	B230N	4.712405e-02
A258W	3.046867e-10	B246D	3.276003e-03
		B332W	3.340382e-06

Table S11: Using $\Delta\Delta G_{bind}$ data, these positions were predicted to be at the binding interfaces of the heterotrimer. Values are untruncated R output.

Position	*p*-value	Position	*p*-value
A6S	3.276003e-03	B54p	2.569524e-08
A8E	1.288433e-05	B56A	2.019602e-02
A9D	2.569524e-08	B68R	3.351553e-09
A17K	2.019602e-02	B74S	3.276003e-03
A21R	2.569524e-08	B76D	2.019602e-02
A29K	4.712405e-02	B83D	4.712405e-02
A30A	8.316008e-03	B84S	4.359198e-04
A35K	3.340382e-06	B85Y	4.712405e-02
A197K	3.276003e-03	B86T	4.359198e-04
A218V	4.359198e-04	B92A	1.453066e-04
		B97S	1.453066e-04
		B98S	4.712405e-02
		B118D	7.708573e-07
		B120I	8.316008e-03
		B129R	1.288433e-05
		B147S	4.509515e-05
		B274T	1.453066e-04
		B313N	1.228501e-03
		B314R	7.708573e-07
		B316S	1.453066e-04