Multi-State Design of Flexible Proteins Predicts Sequences Optimal for Conformational Change

Marion Sauer[1,2], Alexander M. Sevy[1,2], James E. Crowe, Jr.[2,3,4], and Jens Meiler[1,5]

[1] Center for Structural Biology, Vanderbilt University, Nashville, Tennessee, United States of America; [2] Vanderbilt Vaccine Center, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3] Department of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [4] Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [5] Department of Chemistry, Vanderbilt University, Nashville, Tennessee, United States of America

Abstract

Computational protein design of an ensemble of conformations for one protein – *i.e.*, multi-state design – determines the side chain identity by optimizing the energetic contributions of that side chain in each of the backbone conformations. Sampling the resulting large sequence-structure search space limits the number of conformations and the size of proteins in multi-state design algorithms. Here, we demonstrated that the REstrained CONvergence (RECON) algorithm can simultaneously evaluate the sequence of large proteins that undergo substantial conformational changes, such as viral surface glycoproteins. Simultaneous optimization of side chain conformations across all

conformations resulted in an increase of 30% to 40% in sequence conservation when compared to single-state designs. More importantly, the sampled sequence space of RECON designs resembled the evolutionary sequence space of functional proteins. This finding was especially true for sequence positions that require substantial changes in their local environment across an ensemble of conformations. To quantify this rewiring of contacts at a certain position in sequence and structure, we introduced a new metric designated 'contact proximity deviation' that enumerates contact map changes. This measure allows mapping of global conformational changes into local side chain proximity adjustments, a property not captured by traditional global similarity metrics such as RMSD or local similarity metrics such as changes in φ and ψ angles.

Author Summary

Multi-state design can be used to engineer proteins that need to exist in multiple conformations or that bind to multiple partner molecules. In essence, multi-state design selects a compromise of protein sequences that allow for an ensemble of protein conformations, or states, associated with a particular biological function. In this paper, we used the REstrained CONvergence (RECON) algorithm with Rosetta to show that multi-state design of flexible proteins predicts sequences optimal for conformational change, mimicking mutation preferences sampled in evolution. Modeling optimal local side chain physicochemical environments within an ensemble selected significantly more native-like sequences than selections performed when all conformations states are designed independently. This outcome was particularly true for amino acids whose local side chain environment change between conformations. To quantify such contact map changes, we

44 introduced a novel metric to show that sequence conservation is dependent on protein
45 flexibility, *i.e.*, changes in local side chain environments between stated limit the space of
46 tolerated mutations. Additionally, such positions in sequence and structure are more likely
47 to be energetically frustrated, at least in some states. Importantly, we showed that multi-
48 state design over an ensemble of conformations (space) can explore evolutionary tolerated
49 sequence space (time), thus enabling RECON to not only design proteins that require
50 multiple states for function but also predict mutations that might be tolerated in native
51 proteins but have not yet been explored by evolution. The latter aspect can be important to
52 anticipate escape mutations, for example in pathogens or oncoproteins.

53 **Introduction**

54 The central dogma of structural biology states that amino acid sequence defines the
55 three-dimensional protein structure, which in turn, defines protein function, highlighting
56 the central importance of the so-called 'folding problem': which three-dimensional
57 structure adopts a given amino acid sequence? Computational protein design solves the so-
58 called 'inverse folding problem' by identifying an amino acid sequence that is compatible
59 with a given protein structure, *i.e.*, backbone conformation and possibly interactions with
60 partner biomolecules. This approach allows for the molecule to conduct its function in this
61 single state. Protein function, however, often relies on the transition between multiple
62 conformations – a protein must be thermodynamically stable in multiple conformations
63 before it is capable of achieving a defined function. Thus, for a protein to conserve its
64 function, we hypothesized that the conservation of protein flexibility limits the protein's
65 sequence space to be consistent with the conformational changes needed for function.

66 Determining functionally relevant sequence tolerance, or rather, the set of amino acid
67 sequences that are allowable given a protein's function, therefore depends on identifying
68 the set of amino acid sequences that is stable in each of the conformations needed. Testing
69 this hypothesis is complicated, as typically not all functionally relevant conformations have
70 been determined experimentally. The picture gets even more complicated if we look not
71 only at functionally relevant conformations that are by definition local free energy minima
72 (*i.e.*, thermodynamics) but also include an analysis of the height of barriers connecting
73 these states that determine the kinetics of interconversion.

74 Humphris-Narayanan and colleagues demonstrated that prediction of mutation
75 preferences of HIV-1 protease and HIV-1 reverse transcriptase was improved up to 25%
76 when structural ensembles were included during protein design, as opposed to design of a
77 single conformation (1, 2). Friedland and colleagues compared ubiquitin design on a single
78 backbone conformation with design ensembles created through ROSETTA and showed that
79 the sampled sequence space is distinct. When the ensemble conformations were selected
80 to match experimental restraints for dynamics (nuclear magnetic resonance, or NMR,
81 residual dipolar couplings), the sequences consistent with the ensemble match the
82 sequences occurring in the family of ubiquitin proteins (3). These studies demonstrated
83 that a requirement for protein flexibility substantially alters and reduces the sequence space
84 available for evolution. One limitation of this approach, however, is that it assumes that
85 the tolerated sequence space for a conformationally flexible protein can be determined by
86 integrating over single-state design (SSD) of all of its conformations, *i.e.*, enumerating the
87 most energetically favorable amino acid for each position and each conformation.
88 However, the most energetically favorable amino acid, as determined by the

89 aforementioned methods, may be the most energetically favorable for a single
90 conformation, but may not be energetically tolerable at the same position in another
91 conformation. For instance, in some conformations, the energetically most favorable amino
92 acid might be the only allowed amino acid, with all others prohibited (imagine a tiny space
93 where only glycine fits). At the same position in other conformations, there may be
94 acceptable alternatives with more energetically favorable scores, but those residues could
95 not be tolerated as an acceptable mutation in the aforementioned more constrained position.
96 Thus, we hypothesized that multi-state design (MSD) over all conformations relevant for
97 function will yield a more accurate representation of the biologically relevant sequence
98 space compatible with function.

99 MSD approaches evaluate the stability of sequences across all conformations to
100 select the lowest-energy sequence. The Best Max-Marginal First (BMMF) algorithm was
101 used to demonstrate that the MSD of 16 unique calmodulin-substrate complexes increased
102 the similarity of the designed calmodulin binding site to evolutionary sequence profiles by
103 two-fold, and increased in native sequence recovery from 52.5% for single-state designs to
104 80% for the 16-state design scenario (4). Challenges for applying MSD methods like the
105 BMMF algorithm, however, are the efficiency of the search algorithm, large memory
106 requirements, and extended computational time needed. MSD methods up to now have
107 been limited to designing a small number of amino acid positions across all states, with the
108 largest number of simultaneously evaluated design positions being 27 designed positions
109 across 60 states using the MSD FASTER algorithm (4-8).

110 We sought to study large proteins that undergo conformational rearrangements that
111 include domain or hinge displacements of greater than a few Å in root mean square distance

112 (RMSD). We expected that the tolerated sequence space must be restricted in some regions
113 to allow for substantial 'rewiring' of contact networks when transitioning from one state to
114 another. The tolerated sequence space of these types of conformational changes is not
115 limited to local regions, such as protein-protein interfaces, but instead distributed over the
116 entire amino acid sequence. Thus, an MSD approach that seeks to explore such sequence
117 spaces needs to include the entire protein. The REstrained CONvergence (RECON)
118 algorithm was used previously to estimate the sequence tolerance within protein-protein
119 interfaces. However, already at that time this approach proved to be more computationally
120 efficient than the generic ROSETTA MSD algorithm (8). With the addition of a message-
121 passage interface (MPI), RECON MSD can combine the single-state design (SSD)
122 efficiency of evaluating the sequence tolerance of a full-length protein with the MSD
123 capability of evaluating the fitness function of a sequence across multiple conformations
124 (9).

125 Highly flexible viral glycoproteins, such as the influenza A hemagglutinin protein
126 and its stem domain (HA2), undergo conformational rearrangements of greater than 30 Å
127 and have been shown to be conserved in sequence greater than 90% percent across subtypes
128 (10). For other highly flexible proteins, such as calmodulin, kinases, and voltage-gated
129 sensory channels, regions known to mediate conformational change can be conserved up
130 to 100% across phylogenies, suggesting that a limited set of sequences is suitable for select
131 conformation transitions (11-13). Here we use the ROSETTA RECON MSD algorithm to
132 demonstrate that the sequence space consistent with all experimentally determined
133 conformations of a protein approximates sequence profiles observed in evolution.

134 ## Results

135 It is our hypothesis that simultaneous evaluation of sequence space across an
136 ensemble of conformations improves the correspondence of the designed sequences to an
137 evolutionary sequence profile, but limits the set of possible sequences by selecting
138 mutations that are tolerated across all states within an ensemble, as opposed to individually
139 selecting sequences that are optimal for individual states. To address this hypothesis, we
140 perform RECON MSD and compare the designed sequences to SSD results and PSI-
141 BLAST profiles. We quantify the similarity between designs and evolutionary sequence
142 profiles (Fig 1A and 1B).

143 **Fig 1. Graphical representation of hypothesis and experimental design.** (A) Schematic
144 of sequence space and the impact of flexibility on sequence tolerance. S_1 and S_2 represent
145 two unique conformations of the same residue length separated by some RMSD that
146 populate two local energy minima. Black lines with end caps represent unique sequences
147 that are energetically most favorable for a single conformation. The dark shaded area
148 encircles sequences that are energetically favorable for both conformations. Here we
149 illustrate that by using multiple conformations during protein design, we identify sequences
150 that are energetically suitable for conformational flexibility, yet are not necessarily the
151 most stable sequence for any given conformation. Additionally, the requirement to adopt
152 multiple conformations constrains the number of suitable sequences (B) Flow chart of
153 benchmark design.

154

155

156 **Compilation of a benchmark set of eight proteins**

157 We selected proteins with multiple known conformations of identical sequence
158 from the PDBFlex database (14). The benchmark included eight proteins, requiring that
159 each benchmark case have at least two published conformations with an RMSD greater
160 than 5 Å, and an identical sequence greater than 100 amino acids in length (Table 1).We
161 omitted duplicate conformations, which we define as conformations with and RMSD of
162 less than 0.5 Å, to avoid bias towards similar conformations. In addition, we used a
163 resolution cutoff of 5 Å with the requirement that greater than 75% of the included models
164 within each design ensemble were determined at a resolution of better than 3 Å. We also
165 omitted any models with longer sequence gaps or missing density. For structural models
166 with chain breaks that had missing density for only one or two consecutive residues (PDB
167 IDs 1OK8, 3C5X, and 3C6E of the dengue virus E protein monomer) we added the missing
168 densities with the Rosetta loop modeling application (15). All structural models were
169 gently relaxed with a restraint to start coordinates to remove any energetic frustrations
170 frequent in models derived from low-resolution experimental structures.

171

172 **Table 1. Proteins used in conformation-dependent sequence tolerance benchmark.**

PROTEIN	PDB ID OF STATES USED WITHIN EACH ENSEMBLE	DETERMINATION METHOD	DESIGNED POSITIONS	AVERAGE PAIRWISE RMSD OF DESIGNED STATES (Å)	LARGEST PAIRWISE RMSD (Å)
5´-NUCLEOTIDASE	1HPU	X-ray	523	5.18 ± 1.12	9.20
	1OI8	X-ray			
	1IOD, chain A	X-ray			
	1OID, chain B	X-ray			
	4WWL	X-ray			
ADENYLATE KINASE	1AKE	X-ray	214	7.19	7.19
	4AKE	X-ray			
CAGL	3ZCJ	X-ray	169	27.0 ± 22.2	39.9
	4CII	X-ray			
	4YVM	X-ray			
CALMODULIN	1A29	X-ray	139	10.4 ± 2.44	17.9
	1CFC	NMR			
	1CFD	NMR			
	1CFF	NMR			
	1CKK	NMR			
	1CLL	X-ray			
	1CM1	X-ray			
	1CM4	X-ray			

	1G4Y	X-ray			
	1LIN	X-ray			
	1MUX	NMR			
	1NIW	X-ray			
	1NWD	NMR			
	2F2P	X-ray			
	2N8J	NMR			
	2WEL	X-ray			
	3EWT	X-ray			
	3EWV	X-ray			
	4DJC	X-ray			
	4HEX	X-ray			
DENGUE VIRUS ENVELOPE PROTEIN (MONOMER)	1OAN	X-ray	394	6.63 ± 2.89	13.3
	1OK8	X-ray			
	3C5X	X-ray			
	3C6E	X-ray			
	3J27	Cryo-EM			
	3J2P	Cryo-EM			
GROEL SUBUNIT	1AON, chain A	X-ray	523	9.06 ± 1.13	13.1
	1AON, chain N	X-ray			
	2C7E	Cryo-EM			
	3WVL	X-ray			
	4AB3	Cryo-EM			
	4KI8	X-ray			
INFLUENZA HEMAGGLUTININ	1QU1	X-ray	344	23.7 ± 17.4	34.9
	1HTM	X-ray			

STEM (TRIMER)	2HMG	X-ray			
	3EYM	X-ray			
RESPIRATORY SYNCYTIAL VIRUS FUSION PROTEIN (TRIMER)	3RKI	X-ray	1252	29.9 ± 22.0	46.2
	3RRR	X-ray			
	4MMS	X-ray			
	4ZYP	X-ray			

173

174 For a complete description of Protein Data Bank (PDB) identification and sequence information included in the benchmark, see S1

175 Table.

176

177 **Metrics to measure amplitudes of local and global conformational change**

178 Quantification of protein flexibility commonly relies on the structural comparison
179 of two structural models, whether that be through the similarity of equivalent atoms in
180 three-dimensional space, calculated as root mean square distance (RMSD), or by the
181 similarity of equivalent φ and ψ backbone dihedral angles, calculated as root mean square
182 deviation ($RMSD_{da}$) (16). RMSD is used frequently as a global metric used to describe the
183 overall similarity of two conformations of the same protein and has been a powerful metric
184 to quantify overall structural similarity. $RMSD_{da}$, on the other hand, is used to describe
185 local backbone displacements and is well-established, for example, to compare loop
186 conformations. The disadvantage of both metrics is that they do not capture whether or not
187 a particular residue is reconfigured in its interactions with neighboring amino acids.
188 However, we hypothesize that such a metric of local rewiring driven by a global
189 conformational space will best correlate with restrictions in sequence space introduced
190 through conformational flexibility. Thus, we settled on three metrics that capture the
191 structural *dis*similarity of a protein ensemble in terms of its maximum global structural
192 dissimilarity, local backbone dissimilarity, or contact map dissimilarity: 1) The maximum
193 pairwise RMSD of all atom coordinates of two superimposed structures within a set of n
194 superimposed structures was used as a metric to describe the maximal global conformation
195 change an ensemble undergoes (Fig 2A). To allow for comparison of RMSD values
196 between benchmark cases that involve proteins of different size, we used RMSD100, a
197 RMSD value normalized to protein of length 100 amino acids. (17) 2) Residue ϕ and φ
198 $RMSD_{da}$ was used as a local metric of similarity (Fig 2B). This metric will directly identify
199 hinge regions between moving domains. 3) Lastly, we designed a metric that captures

200 changes in the contact map computed as $C_\beta - C_\beta$ distance variation. This metric captures
201 local changes in the environment of a residue by including non-local tertiary contacts in
202 the analysis. Thus, it is designed to capture the local and global changes of the
203 physicochemical environment of a residue and thus defines which amino acids are tolerated
204 in a certain position (Fig 2C and 2D). For a complete description of each metric, see
205 Methods.

206 **Fig 2. Metrics used to quantify conformational flexibility.** (A) Illustration of maximum
207 RMSD100, the metric used to quantify large-scale, or global, conformational flexibility.
208 For simplicity, we only represent RMSD on a two-dimensional plane, where the x and y
209 axes represent the difference in distance of cartesian space if two conformations were
210 superimposed onto the same coordinate system. Each protein conformation of identical
211 sequence is represented as a circle, and is separated by some distance vector evaluated as
212 the RMSD100 of two conformations. The maximum RMSD100 describes the greatest
213 pairwise RMSD100 within an ensemble. (B) Illustration of dihedral angle ϕ and φ
214 variation used to calculate dihedral angle RMSD ($RMSD_{da}$). Orientation of atoms is color-
215 coded and corresponds to the diagram drawn at the bottom of the panel. $RMSD_{da}$ is
216 illustrated as the range of dotted lines, corresponding to the deviation in relative orientation
217 of the third and fourth atoms. (C) Explanation of contact proximity deviation. Two
218 conformations of the same protein are depicted in the left, with two residues, outlined in
219 cyan or orange, shown in their respective positions. These two residues are magnified (top
220 right) in their local side chain environment in Conformation A on the top and Conformation
221 B on the bottom. Contact residues in Conformation A are colored yellow. If the same
222 contacts are maintained in Conformation B, contact residues remain colored yellow in the

223 bottom two boxes. If new contacts are made, contact residues are colored in purple. Even
224 though the cyan residue changes slightly in its relative orientation between conformations,
225 the same contacts are maintained so that the degree of conformational flexibility is
226 relatively low in comparison to the heptad trimer refolding, and would have a low contact
227 proximity deviation score. In contrast, the orange residue completely rearranges its local
228 side chain contacts between conformations as a result of the large conformational
229 rearrangement, and would have a high contact proximity deviation score. (D) Explanation
230 of contact proximity deviation. We assigned a score to each C_β–C_β distance by applying a
231 soft-bounded, continuously differentiable function that accounts for the proximity of two
232 side chains and approximates the likelihood of two side chains forming a contact,
233 illustrated in the top left of Panel D. We then calculated the deviation of each C_β–C_β
234 distance across an ensemble as shown in the matrix, with low deviation scores in white and
235 high scores in black. The contact proximity deviation score represents the sum of all C_β–C_β
236 proximity deviations a single residue undergoes within an ensemble, as shown in the
237 bottom row separated from the matrix.
238

239 **RECON MSD samples sequence profiles that are more similar to evolutionary**
240 **observed sequence profiles when compared to SSD**

241 We first examined the correspondence of native sequence recovery determined by
242 MSD versus SSD designed sequences with evolution conservation rates. Native sequence
243 recovery was calculated as the mean percentage of conservation of the starting, or native,
244 sequence for all designed positions. PSI-BLAST profiles were generated using the native
245 sequence, and for consistency we term the percentage of the starting sequence in a PSI-

246 BLAST profile as the percent native sequence recovery. Simultaneously sampling across
247 multiple conformations significantly restricted sequence sampling, or in other words, was
248 more likely to conserve the native sequence, where RECON MSD had total native
249 sequence recovery of 87.8 ± 4.5% versus SSD with 48.9 ± 11.1% native sequence recovery
250 (Fig 3A). In contrast, PSI-BLAST profiles had a native sequence recovery of 82.11 ±
251 11.2%. Qualitatively, the PSI-BLAST profiles were much more similar to the predicted
252 sequence tolerance of RECON MSD compared to SSD, yet a Mann-Whitney U test (18)
253 indicated a significant difference of mean native sequence recovery of either design
254 protocol compared to PSI-BLAST sequence tolerances, with a significance of $p = 0.0029$
255 for RECON MSD and $p < 0.00001$ for SSD. Total sequence recovery is a coarse
256 approximation of sequence similarity, and fails to determine if the designed sequence
257 profiles are sampling similar mutation preferences as observed in evolution. Therefore, we
258 calculated a total variance score of each observed position-specific mutation profile to the
259 corresponding PSI-BLAST profile of each protein (Fig 3B and S1 Fig). We found that in
260 seven out of eight cases, a RECON MSD mutation profile resembled its corresponding
261 PSI-BLAST profile more closely than the SSD mutation profile.

262 **Fig 3. Design native sequence recovery and mutation profile variability comparisons**
263 **to PSI-BLAST profiles.** (A) Comparison of total native sequence recovery of relaxed and
264 unminimized RECON MSD and SSD designs to PSI-BLAST sequence profiles generated
265 using the native sequence. For this figure and all subsequent boxplots, shaded regions of
266 each box plot denote values within the first and third quartiles (interquartile range, or IQR),
267 with the median indicated as a solid line and whiskers representing values $\pm\ 1.5 \times IQR$.
268 Outliers are represented as dots. Asterisks indicate the significance of difference of means

269 of each design in comparison to the PSI-BLAST profile, with a *z*-test *p*-value < 0.01
270 represented by one asterisk, and a *p*-value < 0.00001 by three asterisks. The p-value
271 provided in this figure and all subsequent figures represents a two-sided, 95% confidence
272 interval. (B) Mutation frequency variances of designs in comparison to a PSI-BLAST
273 profile, normalized by protein length. The y-axis values represent the average variability
274 of mutation profiles for each designed residue in relation to a PSI-BLAST profile,
275 represented as:

276 $$y = \frac{\sum_{i=1}^{n}\sum_{j=1}^{20}(aa_{PSI-BLAST} - aa_{Design})^2}{n}$$

277 where aa_j represents the frequency of an amino acid observed at position i for each of all
278 twenty amino acids (j), and y is the sum of all i differences for all amino acids within a
279 protein of length n residues. A y-value of 0 would indicate that the design profile is
280 identical to the PSI-BLAST profile, and an increase in y-value indicates the average
281 frequency variance of the sequence profile for each residue is more dissimilar to a PSI-
282 BLAST profile.
283

284 **RECON MSD underestimates amino acid exchangeability, but samples a more**
285 **evolutionarily relevant sequence space than SSD**

286 Although RECON MSD more closely resembled PSI-BLAST sequence profiles on
287 a per-case basis, we wanted to identify trends in sequence sampling in relation to the PSI-
288 BLAST profiles to highlight design-sampling biases. This task was achieved by calculating
289 the frequency an amino acid is conserved or mutated to another residue, or, the mean amino

290 acid substitution frequency. In general, RECON MSD is more likely to conserve a native
291 amino acid compared to a PSI-BLAST profile, whereas SSD is much more likely to replace
292 the native amino acid (Fig 4A and S2 Fig). We examined amino acid exchangeability as
293 the frequency of exchanging a native for a non-native amino acid. On average, PSI-BLAST
294 profiles exchanged a native for non-native amino acid 1.32 ± 0.03% of the time, versus
295 0.77 ± 0.02% for RECON MSD and 2.45 ± 0.07% for SSD (Fig 4B). Additionally, we
296 compared the average difference of exchangeability for each residue as observed in the
297 PSI-BLAST profiles versus either RECON MSD or SSD and found that RECON MSD
298 average exchangeability rates of each residue are more similar to PSI-BLAST values than
299 SSD (Fig 4C). With the exception of phenylalanine or tyrosine, the difference between
300 exchangeability rates for residues with larger side chains diminishes for RECON MSD, but
301 becomes more exaggerated for SSD, as compared to observed mutation rates in evolution.
302 This finding suggests that the inclusion of multiple conformations during design
303 encourages better placement of bulky side chains, albeit with conservative placement.
304 However, when comparing the linear regression model of individual exchangeability rates
305 of either RECON MSD or SSD to PSI-BLAST rates, both designs were roughly equally
306 dissimilar to PSI-BLAST exchangeability rates, with RECON MSD having a correlation
307 coefficient of $r = 0.35$ and SSD with $r = 0.64$ (S3 Fig). Given that exchangeability rates
308 were not normally distributed, a Kendall τ_β rank correlation coefficient was computed to
309 measure the ordinal association of design and PSI-BLAST amino acid exchangeability
310 rates, where a coefficient of $\tau_\beta = 0$ would indicate that the amino acid exchangeability
311 rates are identical. We found RECON MSD to have a $\tau_\beta = 0.283,\ p \leq 2.22 \times 10^{-16}$
312 versus $\tau_\beta = 0.372,\ p \leq 2.22 \times 10^{-16}$ for SSD when measured for its association to PSI-

313 BLAST amino acid exchangeability rates. In addition, we compared the difference in
314 exchangeability rates between design and PSI-BLAST by calculating the ratio of
315 transformed exchangeability rates (e) as $\frac{e_{design} + 0.00001}{e_{PSI-BLAST} + 0.0001}$ to avoid division by zero, where
316 an individual exchangeability rate would be equivalent between design and PSI-BLAST if
317 $e_{ratio} = 1$. We found that for RECON MSD the mean $e_{ratio} = 2.49$ and for SSD the mean
318 $e_{ratio} = 22.7$. A Mann-Whitney U test of matched individual ratios found a significant
319 difference between RECON MSD and SSD exchangeability rate ratios to PSI-BLAST
320 exchangeability rates, with $p < 0.0001$. Taken together, RECON MSD is sampling
321 individual mutation preferences significantly more closely to that observed in evolution
322 than SSD.

323 **Fig 4. Comparison of exchangeability rates.** (A) Average amino acid exchangeability of
324 PSI-BLAST, RECON MSD, and SSD sequence profiles. Single-letter amino acid codes
325 were used for both x and y axes, with the x axis representing the original amino acid and
326 the y axis representing the average mutation frequency the original amino acid to the
327 indicated mutation. (B) Comparison of exchangeability rates between profiles, excluding
328 rates of native sequence conservation rates. The y axis represents the mean frequency a
329 native amino acid is replaced with a specific, non-native amino acid, which we term as
330 amino acid exchangeability. (C) Difference of mean amino acid-specific exchangeability
331 observed in a PSI-BLAST profile compared to a design profile. The x axis represents each
332 type of amino acid present in the native sequence. The y axis represents the difference in
333 average exchangeability frequency of each amino acid type, or rather, the average
334 frequency a native amino acid type is replaced with any other non-native amino acid. A
335 positive value indicates the native amino acid is less likely to be exchanged for a non-native

336 amino acid during design, whereas a negative value indicates the native amino acid is more
337 likely to be exchanged, as compared to a PSI-BLAST profile.
338

339 **Sequence conservation is dependent on its contact map as computed by Cβ–Cβ**
340 **distance deviations**

341 To consider the effect of conformational flexibility on sequence conservation, we
342 examined the dependency of native sequence recovery on different aspects of
343 conformational flexibility using the aforementioned metrics, maximum RMSD100,
344 $RMSD_{da}$, and contact proximity deviation. We performed a Kendall τ_β rank correlation test
345 on each profile to test for the strength of dependency of native sequence recovery on each
346 metric (Fig 5). Of the three metrics, the native sequence recovery, or rather percent
347 conservation, observed in PSI-BLAST profiles was only dependent on contact proximity
348 deviation z-score, with $p = 1.79 \times 10^{-6}$, versus $p \geq 0.144$ for all other tests. RECON
349 MSD and SSD native sequence recovery depended on both $RMSD_{da}$ and contact proximity
350 deviation z-score ($p < 0.01$). Native sequence recoveries of both designed profiles
351 depended strongly on dihedral angle deviation, with $p \leq 2.22 \times 10^{-16}$, and had similar τ_β
352 coefficients, with $\tau_\beta = 0.101$ for RECON MSD and $\tau_\beta = 0.0806$ for SSD. This finding
353 suggests that the Rosetta scoring function employed by both protein design algorithms is
354 too restrictive in sampling for residues at hinge points, given that the same dependency on
355 $RMSD_{da}$ is not observed for PSI-BLAST sequence conservation. However, both PSI-
356 BLAST and RECON MSD had similar τ_β coefficients predicted with the same confidence,
357 with $\tau_\beta = 0.0639, p = 1.79 \times 10^{-6}$ and $\tau_\beta = 0.0787, p = 1.19 \times 10^{-7}$ respectively, for
358 the dependence of native sequence recovery on contact proximity deviation z-score. This

359 observation suggests that there is an evolutionary constraint on residues that are required
360 to maintain a re-assortment of their local physicochemical environments necessary for a
361 conformational change, and that RECON MSD closely models this evolutionary constraint
362 by considering the multiple local side-chain environments within a protein ensemble.

363 **Fig 5. Relationship of conformational flexibility and native sequence recovery by**
364 **sequence profiles.** The x axis is binned into three groups of equal number of data points
365 to show the distribution of native sequence recovery between groups of low, middle, and
366 high values for each metric. A Kendall τ_β rank correlation test was performed on each
367 profile to measure the strength of dependence of native sequence recovery on each metric,
368 indicated in each plot along with its associated p-value. (A) Comparison of native sequence
369 recovery dependence on maximum RMSD100 between sequence profiles. (B) Comparison
370 of native sequence recovery dependence on $RMSD_{da}$ between sequence profiles. $RMSD_{da}$
371 values of each protein were not equally distributed, nor were of similar range. Therefore, a
372 z-score of was used to normalize $RMSD_{da}$ values of each protein to compare dihedral angle
373 deviation scores, shown along the x axis. A similar approach was implemented to
374 normalize contact map deviation scores. (C) Comparison of native sequence recovery
375 dependence on contact deviation scores.

376

377 **Regions critical for conformational plasticity are energetically frustrated**

378 The encouraged sequence convergence employed by the RECON MSD algorithm
379 identifies amino acid sequences that have the lowest total energy across all states (8). To
380 examine the energetic impact of requiring a single amino acid sequence to adopt multiple
381 states, we use a similar energy score term described previously as the sum total energy

382 score normalized by the number of designed positions (see Methods). For RECON MSD
383 designs, this approach would include lowest mean energy score of the designed ensemble,
384 whereas the SSD energy score would include the lowest energy scores for each state. In all
385 eight cases, RECON MSD selects sequences with a significantly higher energy score than
386 SSD with a paired student's t-test (19), with $p < 1 \times 10^{-4}$. (Fig 6). We also compared the
387 design energy scores to the ten lowest-energy relaxed structures, which only included the
388 native sequences, and found that RECON MSD samples a lower energy sequences than the
389 relaxed native structures. Even though SSD sequences are the most stable, RECON MSD
390 identifies more favorable mutations within the native ensemble of states.

391 **Fig 6. Average per-residue total energy score of the lowest ten percent scoring**
392 **models for RECON MSD, SSD, and starting relaxed (Native) models.** One hundred
393 simulations were performed for each group and the lowest ten total energy scoring
394 models were used for the comparison. The total scores were normalized so that the
395 calculated total score was divided by the number of residues within each model to obtain
396 a mean residue score. For RECON MSD models, the total calculated score also had to be
397 normalized by the number of states within each model. The violin plot width indicates the
398 normalized energy score density of each group.

399

400 **Stability decreases for residues with larger Cβ–Cβ contact map deviations**

401 We used a Kendall τ_β rank correlation test to analyze the dependency of the
402 modeled sequence energy score on global and local conformational changes. For the
403 comparison with global conformational changes, we compared the mean total score of the
404 ten lowest-energy scoring design models, normalized by the number of residues within

each protein, to the maximum RMSD100 of an ensemble. We found that there is a negative dependence of mean total score on the maximum RMSD100 for SSD models ($\tau_\beta = -0.143, p = 1.16 \times 10^{-5}$), but not so for RECON MSD models ($\tau_\beta = 0.0177, p = 0.586$; **Error! Reference source not found.**). Conversely, there was a small, but significant positive dependence of individual residue scores on contact proximity deviation z-scores for RECON MSD models ($\tau_\beta = 0.0356, p = 0.00584$), but not so SSD models ($\tau_\beta = -0.00538, p = 0.677$). There was no dependence of individual residue scores on dihedral RMSD for either design approach (Fig 7). This finding suggests that RECON MSD is restricted in optimizing the stability of residues that must rearrange their local side-chain environments, but not for optimizing local backbone flexibility. SSD, on the other hand, is not restricted in optimizing side-chain placement within an ensemble, and therefore can select amino acid sequences that are more stabilizing for individual conformations.

Fig 7. Comparison of conformational diversity and per-residue total scores. All panels are binned into low, medium, and high x values, with equal number of data points for each bin. A Kendall τ_β rank correlation test was performed on each profile to measure the strength of dependence of native sequence recovery on the x axis value, indicated in each plot along with its associated p-value. (A) Comparison of maximum RMSD100 and mean total energy score, normalized by the number of residues. (B) Comparison of normalized $\mathrm{RMSD_{da}}$ z-score and mean total energy score of each residue. (C) Comparison of normalized contact proximity deviation z-score and mean total energy score of each residue.

427 # Discussion

428 **Contact proximity deviation captures local and global conformational**
429 **rearrangements as a single metric**

430 Methods like Local-Gobal Alignment and contact area differences are useful in
431 circumventing the over-estimation of global structural dissimilarity by either emphasizing
432 local backbone segment structure similarity or by side chain placement similarity,
433 respectively (20, 21). In particular, contact area differences between two homologues have
434 been shown to be as accurate as RMSD, if not more, in comparing the structural similarity
435 of proteins with very high sequence similarity (22). However, contact area differences are
436 sensitive to errors in side chain atom placement within a structural model, so that
437 accounting for side chain mutations while measuring the relative position of equivalent
438 residues is not feasible.

439 For this paper, we introduced the contact proximity deviation metric that quantifies
440 the relative placement of a side chain within a structure and its variability within an
441 ensemble without having to predict side-chain contacts. The soft-bounded function we use
442 to score the distance of a side chain C_β atom to all other surrounding side chains is less
443 sensitive to the misplaced assignment of atom coordinates in an individual structure.
444 Additionally, because we use the deviation of contact proximity within an ensemble rather
445 than the contact proximity score of an individual $C_\beta - C_\beta$ distance, small errors in the C_β
446 atom assignment of one structural model are less likely to substantially influence the score
447 deviation of a residue within an ensemble. Therefore, scoring $C_\beta - C_\beta$ distance deviations
448 of each residue allows for the comparison of C_β placement and local protein fold variability
449 within an ensemble, independent of side chain identity.

450 The measurement of local structural variability by NMR residual dipolar coupling
451 has shown that regions of high flexibility within ubiquitin ensembles align with multiple
452 protein-protein interfaces (23). Large protein dynamics are not suitable for NMR and are
453 difficult to experimentally determine as an ensemble. In the few cases where multiple
454 conformations of large proteins have been resolved, contact proximity deviation provides
455 a unique method to explore how side chain interaction flexibility or protein re-folding
456 affects binding interfaces. Additionally, in cases such as protein switch design or structural
457 analysis of mutations where protein flexibility, but not sequence, needs to be conserved, it
458 is necessary to have a metric that allows for the comparison of local structural variability
459 due to protein refolding, such as contact proximity deviation.
460
461 **Sampling functional mutation preferences requires evaluation of sequence stability**
462 **as an ensemble.**
463 MSD approaches use the energetic contributions of multiple conformations to steer
464 sequence selection and cull any sequences that do not improve the energy score of the
465 designed ensemble as a whole (24). For most approaches, sequences that do not improve
466 all or the majority of conformations within an ensemble are culled, which is appropriate
467 when the goal is to stabilize a protein ensemble within an energy minimum. However,
468 protein function may not select for sequences that are limited to a single energy minimum.
469 To better estimate the mutational preferences critical for function, it is necessary to use an
470 approach that models local side-chain environments within the context of an ensemble.
471 Within Fig 4, we illustrated that, although RECON MSD failed to accurately
472 predict all amino acid exchangeability rates, the consideration of multiple local side chain

473 environments during protein design improved the prediction of sequence conservation and
474 overall accuracy of amino acid substitution frequencies as compared to modeling local side
475 chain environments independently, particularly when modeling bulky side chains. In
476 combination with Fig 6 and Fig 7, we demonstrated that the consideration of local side
477 chain stability within the context of an ensemble restricts stability optimization of the
478 ensemble, especially for residues that require side-chain rearrangements during a
479 conformational rearrangement. Taken together, we demonstrated that selection of mutation
480 profiles by RECON MSD is much more similar to mutation rates observed in homologs if
481 each mutation is evaluated across every conformation, or state, within an ensemble, and
482 then culled if the mutation is evaluated to be destabilizing for an individual state within an
483 ensemble. This approach does not necessarily select sequences that improve the stability
484 of every state within an ensemble, but rather places the importance of modeling an ideal
485 conformation-specific, local side-chain environment to prevent local side-chain
486 destabilization within the context of an ensemble.

487 **RECON MSD can be used to predict evolutionary sequence conservation of flexible**
488 **proteins**

489 Sequence similarity searches, such as PSI-BLAST, are fast and easy to use.
490 Predicting mutation preferences from structure, especially if the sequence is known to form
491 multiple conformations, remains to be a challenge. Advances in structure-based evolution
492 design methods rely on iterative approaches that match sequence and structure similarities
493 to predict sequence entropy (25). For proteins that undergo conformational rearrangements,
494 using this type of approach to search for structural similarity limits the sequence search
495 space to similar conformations, possibly preventing the identification of sequences capable

496 of adopting multiple conformations. Although other MSD methods have improved the
497 selection of more evolutionarily similar sequences, they are limited in their capacity to
498 simultaneously sample conformation and sequence space so that the relevance of
499 conformational plasticity in evolutionary dynamics have not been fully interrogated.

500 For this benchmark, we used RECON MSD to demonstrate that sequence
501 conservation and mutation preferences of a single sequence can be approximated using the
502 evaluation of local residue physicochemical changes, provided that this one sequence folds
503 into select, multiple conformations. In Fig 3, we showed that the estimated sequence
504 conservation of RECON MSD designs differs by roughly 5% from the sequence
505 conservation observed in PSI-BLAST profiles, with RECON MSD being more
506 conservative. We demonstrated in Fig 5 that sequence conservation, as observed in
507 evolutionarily-related sequences, depends on the degree of local side-chain environment
508 changes a residue must undergo to assist in the refolding of a protein during a conformation
509 rearrangement, and that RECON MSD approximates the same restriction in amino acid
510 conservation as C_β–C_β proximity variation increases.

511 The caveat to using RECON MSD to predict mutation preferences is accounting
512 for ROSETTA sampling biases. First, RECON MSD does not currently allow for the
513 formation or destruction of disulfide bonds, which is critical for conformation stability, and
514 does not accurately model the frequency of cysteine conservation. Consideration of
515 alternate protonation states due changes in pH are also not explicitly modeled, which we
516 see from our amino acid exchangeability comparisons that RECON MSD underrepresents
517 exchangeability of polar residues and frequently mutates histidine to lysine or arginine,
518 which has a pK_a much higher than histidine or which is not charged. Additionally, in Fig

519 5, we showed that RECON MSD is likely to overestimate sequence conservation of hinge
520 regions that have large dihedral angle RMSDs. Even though we used a gentle minimization
521 prior to design, minimization significantly increases the estimated stability of the native
522 residue, making the replacement of the native amino acid unfavorable, as shown in S1 Fig.
523 Given that residues located at hinge points within flexible loops are intrinsically disordered
524 and typically contain less than ideal Ramachandran dihedral angles, it is likely that
525 minimization specifically overcorrects these bond angles to fit the energy scoring function,
526 preventing accurate sampling of rotamer placement. With the addition of explicit disulfide
527 bond formation, use of a pK_a-dependent rotamer library, and improvement of minimization
528 prior to design, the RECON MSD algorithm could prove to be a valuable tool in predicting
529 accurate mutation profiles.

530 The computational time required for the RECON MSD design simulations within
531 this benchmark ranged from 2 - 36 hours. Compared to experimental approaches that have
532 tested for functionally tolerated mutations in either dengue virus envelope protein or
533 influenza hemagglutinin protein (2, 26, 27), RECON MSD is much faster and less costly
534 in identifying biologically relevant mutations. Additionally, RECON MSD is not limited
535 to sampling mutations singly, pairwise, or as limited networks, but rather can sample
536 mutations as an interaction network of each local side-chain environment. Traditional intra-
537 protein co-evolution methods, such as direct coupling analysis (28), mutual information
538 (29-31), or McLachlan-based substitution correlation methods (32, 33), are not reliable in
539 detecting co-variation or correlation of mutation frequencies of highly conserved
540 sequences (34), and so they fail to detect contact dependencies of sequences with low
541 sequence variation. In the case of this benchmark, we see that flexible sequences tend to

542 be more highly conserved, especially when residues need to maintain distinct contacts
543 between conformations. Therefore, current co-evolution methods cannot be used to detect
544 residue contact dependencies of flexible, highly conserved sequences, whereas this
545 benchmark suggests that RECON MSD is well-suited to identifying the evolutionary
546 potential of a flexible sequence.

547 **Conclusions**

548 We demonstrated that RECON MSD significantly improves the similarity to
549 evolutionary mutation preferences from SSD selected mutation profiles by selecting
550 sequences which are energetically favorable for an ensemble of local side-chain
551 interactions. Specifically, in instances where the goal of protein design is to preserve an
552 ensemble of conformations for functionality, we suggest a greater emphasis on designing
553 local physicochemical environments for each and all conformations within an ensemble,
554 and to place less of an emphasis of finding sequences representing the most
555 thermostabilizing for either each state individually or as an average of all states.
556 Furthermore, the new conformational diversity metric contact proximity deviation we
557 describe in this paper allows for the comparison of protein ensembles, assuming they are
558 of similar length but not sequence, by quantifying position-specific relocation due to one
559 or more conformational changes. Therefore, in conjunction with contact proximity
560 deviation, RECON MSD warrants further use as a bioinformatic tool to estimate mutation
561 preferences of homologous proteins, especially for proteins known to undergo similar
562 domain or fold reorganization between conformations.

563 # Methods

564 **Selection and preparation of benchmark datasets**

565 Our criteria for benchmark datasets included proteins that had at least two published
566 conformations with greater than 5 Å RMSD and at least one peptide chain greater than 100
567 residues in length. To identify these proteins, we performed a BLAST search to identify
568 proteins with 100% sequence identity and with gaps of three or less residues in length.
569 Structures with similar backbone conformations of less than 1 Å RMSD were excluded
570 from design so that the structure with the longest matching consecutive sequence was kept
571 as the template structure.

572 Structures were downloaded from the Protein Data Bank (PDB; www.rcsb.org) and
573 processed manually to remove all atoms other than the residue atoms intended for design.
574 Any residues that did not align or positions that were not present in all template structures
575 were not considered for design and were removed from the template. For a detailed
576 description of which residues were included for design, see S1 Table. Template structures
577 were subject to minimization and repacking in ROSETTA using FastRelax constrained to
578 start coordinates with a standard deviation of ± 0.5 Å with the talaris 2013 score function
579 (35, 36). The lowest total energy score model of 100 relaxed models was selected for
580 design. For comparisons using un-relaxed models, the template structure was the input
581 model structure used for relaxation.

582 **RECON MSD multi-specificity and single-state design**

583 Benchmarking using RECON MSD multi-specificity design was performed using
584 four rounds of fixed backbone design and a convergence step using the greedy selection
585 algorithm, as previously described (8), with the exception that only repacking, and not

586 backbone minimization was allowed following the convergence step to prevent over-
587 optimization of the energy score following design. Similarly, single-state design was
588 performed using four rounds of fixed backbone rotamer optimization followed by
589 repacking using the identical designable residues as specified for RECON MSD designs.
590 The talaris 2013 scoring function was used for both RECON MSD and single-state designs.
591 For each protein dataset, 100 designs were generated for each benchmark structure using
592 either RECON MSD or SSD.

593 **Generation of sequence profiles**

594 The lowest ten out of a hundred scoring models were used for quantification of
595 sequence tolerance. In the case of SSD, the ten lowest total scoring models were used from
596 the design simulation of each PDB structure. For RECON MSD, the total score of each
597 model designed within an ensemble design run were added to create a fitness score, which
598 then was sorted to identify the ten designed ensembles with the lowest fitness score. A
599 Shannon entropy bitscore was calculated for each designed position within an ensemble as
600 $I_i = p_i \times \log_2 (20 \times p_i)$, with i as the amino acid and p_i as the frequency of that amino
601 acid. Here, the calculated amino acid frequency includes the frequency at the same position
602 within the ten lowest-scoring models of all designed states, whether designed
603 independently by SSD or designed simultaneously by RECON MSD, such that an amino
604 acid represented a 100% of the time at a particular position in all states has a bitscore of
605 4.32 (37).

606 PSI-BLAST profiles were obtained by querying a non-redundant protein database
607 using default parameters, increasing the number of iterations to ten iterations, as well as
608 querying the database with *e*-value thresholds ranging from 1×10^{-5} to 1×10^{2}. We

609 reported only PSI-BLAST profiles generated using default parameters, which includes two
610 iterations and an *e*-value of 0.005. Comparisons to PSI-BLAST profiles using non-default
611 parameters were qualitatively identical to the PSI-BLAST profiles generated using the
612 default parameters, and were therefore not reported. We omitted any sequences within the
613 queried sequence profiles which were not included for design and calculated the Shannon
614 entropy bitscore for each position using the position specific-scoring matrix (PSSM)
615 frequencies as described in the previous paragraph.

616 **Comparison of sequence profiles**

617 We compared the position specific-scoring matrix (PSSM) generated from the PSI-
618 BLAST query to the PSSM generated by either RECON MSD or SSD by calculating the
619 percentage of native sequence recovery, which was determined as the sum of the bitscores
620 of the native amino acids at each position divided by the sum of the information bitscore
621 of all amino acids at all positions (38). Additionally, we calculated the difference in
622 designed position mutation frequencies from each position with the PSSM, as described in
623 **Error! Reference source not found.**:

624 $$y = \frac{\sum_{i=1}^{n}\sum_{j=1}^{20}(aa_{PSI-BLAST} - aa_{Design})^2}{n}$$

625 where aa_j represents the frequency of an amino acid observed at position i for each of all
626 twenty amino acids (j), and y is the sum of all i differences for all amino acids within a
627 protein with a length of n residues. Average amino acid substitution rates were determined
628 as the mean cumulative substitution frequency of each amino acid i to amino acid j, where

$$\overline{aa_{i,j}} = \begin{bmatrix} \frac{1}{aa_1}\sum aa_{1,1} & \cdots & \frac{1}{aa_1}\sum aa_{1,j} \\ \vdots & \ddots & \vdots \\ \frac{1}{aa_i}\sum aa_{i,1} & \cdots & \frac{1}{aa_i}\sum aa_{i,j} \end{bmatrix}.$$

We define amino acid exchangeability as the subset of average amino acid substation rates that exclude the substitution rates of amino acid i to amino acid j, where j is identical to i, or in other words, all substitution rates that include the average conservation frequencies of the native amino acid. The mean amino acid exchangeability rates, as shown in **Error! Reference source not found.**C, were calculated as

$$\overline{aa_i} = \frac{1}{19}\sum_{j=1}^{19}\frac{1}{aa_i}\sum aa_{i,j}$$

where the mean exchangeability rate of each amino acid i is the mean of all exchangeability rates of amino acid i to amino acid j, excluding the conservation rate of amino acid i. Kendall $\tau_\beta$rank comparison tests, linear regression, Wilcox comparison of means, and student t tests were performed in R.

Description of conformational metrics used in this benchmark

We use the maximum RMSD within an ensemble to represent the largest amplitude of dissimilarity within an ensemble, defined as:

$$RMSD_{max} = max_s\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|v_i - w_i\|^2}$$

where n represents the number of residues, s represents the number of structures within an ensemble, and $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|v_i - w_i\|^2}$ represents each pairwise RMSD within an ensemble (**Error! Reference source not found.**A). For the local backbone dissimilarity metric, we

use dihedral angle RMSD to describe the deviation of each equivalent dihedral angle, or pair ϕ and φ angles, within an ensemble containing *s* structures, as

$$RMSD_{dihedral} = \sqrt{\frac{1}{s}\sum_{i=1}^{s}\|\phi_s - \overline{\phi}\|^2 + \frac{1}{s}\sum_{i=1}^{s}\|\psi_s - \overline{\psi}\|^2}.$$

where $\overline{\phi}$ and $\overline{\psi}$ represent the mean ϕ and ψ angle of each equivalent residue. The contact map dissimilarity metric we introduce here is based off the Durham et al. (39) neighbor count weight metric, which scores the likelihood of a neighboring contact by assigning each C_β–C_β distance a score, which we term contact proximity ($\mathcal{C}$) as

$$\mathcal{C} = \begin{cases} 1, if distance \leq lower\ bound \\ \frac{1}{2}\cos\left(\frac{distance - lower\ bound}{upper\ bound - lowerbound}\right), if\ lower\ bound < distance < upper\ bound \\ 0, if\ distance \geq upper\ bound \end{cases}$$

For glycines, a pseudo-C_β atom is defined using the amide N, C_α, and carboxyl C atom coordinates in the PDBtools package (github.com/harmslab/pdbtools) before calculating C_β–C_β distances. The lower and upper bounds represent thresholds where a C_β–C_β distance certainly does and does not contain any side-chain atoms that are in contact with another residue's side-chain atom. We define the lower bound as 4.0 Å and the upper bound 12.8 Å, where the lower bound was determined to be a reliable threshold to define solvent-inaccessible side-chains due to side-chain contacts, or in other words, a C_β–C_β distance less than the lower bound is very likely to form at least one side chain interaction (39). The upper bound was determined by the maximum C_β–C_β distance where at least one atom from each side chain formed an interaction (40). Finally, the contact proximity deviation for each residue was calculated as the sum of all C_β–C_β contact proximity score deviations

668 for that residue (Fig 2D). With this metric, we can quantify the changes in side chain local
669 environments that are not due to local hinge-points, but instead, show local side chain
670 environment changes that are due to larger conformational rearrangements.

671 **Acknowledgements**

672 The authors would like to thank Amanda Duran, Axel Fischer, and Diego del
673 Alamo for their assistance and helpful discussions during the development of this project.
674 Calculations for the contact map deviations were made possible by members of the Michael
675 Harms Laboratory as contributors to the PDBtools opensource software
676 (github.com/harmslab/pdbtools). We also thank the Vanderbilt University Advanced
677 Computing Center for Research and Education for the computational resources necessary
678 to complete this project. All work was funded by the U19 AI117905 grant for Structure
679 Based Design of Antibodies.
680

681 **References**

682 1. Humphris-Narayanan E, Akiva E, Varela R, S OC, Kortemme T.
683 Prediction of mutational tolerance in HIV-1 protease and reverse
684 transcriptase using flexible backbone protein design. PLoS Comput Biol.
685 2012;8(8):e1002639.
686 2. Christian EA, Kahle KM, Mattia K, Puffer BA, Pfaff JM, Miller A, et
687 al. Atomic-level functional model of dengue virus Envelope protein
688 infectivity. Proc Natl Acad Sci U S A. 2013;110(46):18662-7.
689 3. Friedland GD, Lakomek NA, Griesinger C, Meiler J, Kortemme T. A
690 correspondence between solution-state dynamics of an individual protein
691 and the sequence and conformational diversity of its family. PLoS Comput
692 Biol. 2009;5(5):e1000393.
693 4. Fromer M, Shifman JM. Tradeoff between stability and
694 multispecificity in the design of promiscuous proteins. PLoS Comput Biol.
695 2009;5(12):e1000627.
696 5. Allen BD, Mayo SL. An efficient algorithm for multistate protein
697 design based on FASTER. Journal of computational chemistry.
698 2010;31(5):904-16.
699 6. Leaver-Fay A, Jacak R, Stranges PB, Kuhlman B. A generic program
700 for multistate protein design. PloS one. 2011;6(7):e20937.
701 7. Humphris EL, Kortemme T. Design of multi-specificity in protein
702 interfaces. PLoS Comput Biol. 2007;3(8):e164.
703 8. Sevy AM, Jacobs TM, Crowe JE, Jr., Meiler J. Design of Protein
704 Multi-specificity Using an Independent Sequence Search Reduces the
705 Barrier to Low Energy Sequences. PLoS Comput Biol.
706 2015;11(7):e1004300.
707 9. Sevy AM, Wu NC, Gilchuk IM, Parrish EH, Burger S, Yousif D, et al.
708 Multistate design of influenza antibodies improves affinity and breadth
709 against seasonal viruses. Proc Natl Acad Sci U S A. 2019;116(5):1597-602.
710 10. Heiny AT, Miotto O, Srinivasan KN, Khan AM, Zhang GL, Brusic V,
711 et al. Evolutionarily conserved protein sequences of influenza a viruses,
712 avian and human, as vaccine targets. PloS one. 2007;2(11):e1190.
713 11. Friedberg F, Rhoads AR. Evolutionary aspects of calmodulin.
714 IUBMB Life. 2001;51(4):215-21.
715 12. Kobayashi M, Buck J, Levin LR. Conservation of functional domain
716 structure in bicarbonate-regulated "soluble" adenylyl cyclases in bacteria
717 and eukaryotes. Dev Genes Evol. 2004;214(10):503-9.
718 13. Shealy RT, Murphy AD, Ramarathnam R, Jakobsson E, Subramaniam
719 S. Sequence-function analysis of the K+-selective family of ion channels

720 using a comprehensive alignment and the KcsA channel structure.
721 Biophysical journal. 2003;84(5):2929-42.
722 14. Hrabe T, Li Z, Sedova M, Rotkiewicz P, Jaroszewski L, Godzik A.
723 PDBFlex: exploring flexibility in protein structures. Nucleic Acids Res.
724 2016;44(D1):D423-8.
725 15. Stein A, Kortemme T. Improvements to robotics-inspired
726 conformational sampling in rosetta. PloS one. 2013;8(5):e63090.
727 16. Kufareva I, Abagyan R. Methods of protein structure comparison.
728 Methods in molecular biology. 2012;857:231-57.
729 17. Carugo O, Pongor S. A normalized root-mean-square distance for
730 comparing protein three-dimensional structures. Protein science : a
731 publication of the Protein Society. 2001;10(7):1470-3.
732 18. Mann HBaW, D.R. On a test whether one or two random variables is
733 stochastically larger than the other. The Annals of Mathematical Statistics.
734 1947;18(1):50-60.
735 19. Student. The probable error of a mean. Biometrika. 1908;6(1):1-25.
736 20. Zemla A. LGA: A method for finding 3D similarities in protein
737 structures. Nucleic Acids Res. 2003;31(13):3370-4.
738 21. Abagyan RA, Totrov MM. Contact area difference (CAD): a robust
739 measure to evaluate accuracy of protein models. J Mol Biol.
740 1997;268(3):678-85.
741 22. Marsden B, Abagyan R. SAD--a normalized structural alignment
742 database: improving sequence-structure alignments. Bioinformatics.
743 2004;20(15):2333-44.
744 23. Lange OF, Lakomek NA, Fares C, Schroder GF, Walter KF, Becker
745 S, et al. Recognition dynamics up to microseconds revealed from an RDC-
746 derived ubiquitin ensemble in solution. Science. 2008;320(5882):1471-5.
747 24. Davey JA, Chica RA. Multistate Computational Protein Design with
748 Backbone Ensembles. Methods in molecular biology. 2017;1529:161-79.
749 25. Brender JR, Shultis D, Khattak NA, Zhang Y. An Evolution-Based
750 Approach to De Novo Protein Design. Methods in molecular biology.
751 2017;1529:243-64.
752 26. Tsai F, Homan PJ, Agrawal H, Misharin AV, Abdala-Valencia H,
753 Haines GK, 3rd, et al. Bim suppresses the development of SLE by limiting
754 myeloid inflammatory responses. J Exp Med. 2017;214(12):3753-73.
755 27. Thyagarajan B, Bloom JD. The inherent mutational tolerance and
756 antigenic evolvability of influenza hemagglutinin. eLife. 2014;3.
757 28. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et
758 al. Direct-coupling analysis of residue coevolution captures native contacts

759 across many protein families. Proc Natl Acad Sci U S A.
760 2011;108(49):E1293-301.
761 29. Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual information in
762 protein multiple sequence alignments reveals two classes of coevolving
763 positions. Biochemistry. 2005;44(19):7156-65.
764 30. Tillier ER, Lui TW. Using multiple interdependency to separate
765 functional from phylogenetic correlations in protein alignments.
766 Bioinformatics. 2003;19(6):750-5.
767 31. Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory
768 to search for co-evolving residues in proteins. Bioinformatics.
769 2005;21(22):4116-24.
770 32. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations
771 and residue contacts in proteins. Proteins. 1994;18(4):309-17.
772 33. Olmea O, Rost B, Valencia A. Effective use of sequence correlation
773 and conservation in fold recognition. J Mol Biol. 1999;293(5):1221-39.
774 34. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-
775 evolution. Nat Rev Genet. 2013;14(4):249-61.
776 35. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R,
777 et al. ROSETTA3: an object-oriented software suite for the simulation and
778 design of macromolecules. Methods in enzymology. 2011;487:545-74.
779 36. Combs SA, Deluca SL, Deluca SH, Lemmon GH, Nannemann DP,
780 Nguyen ED, et al. Small-molecule ligand docking into comparative models
781 with Rosetta. Nature protocols. 2013;8(7):1277-98.
782 37. Shannon CE. The mathematical theory of communication. 1963. MD
783 Comput. 1997;14(4):306-17.
784 38. Willis JR, Briney BS, DeLuca SL, Crowe JE, Jr., Meiler J. Human
785 germline antibody gene segments encode polyspecific antibodies. PLoS
786 Comput Biol. 2013;9(4):e1003045.
787 39. Durham E, Dorr B, Woetzel N, Staritzbichler R, Meiler J. Solvent
788 accessible surface area approximations for rapid and accurate protein
789 structure prediction. Journal of molecular modeling. 2009;15(9):1093-108.
790 40. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate
791 prediction of residue-residue interactions across protein interfaces using
792 evolutionary information. eLife. 2014;3:e02030.

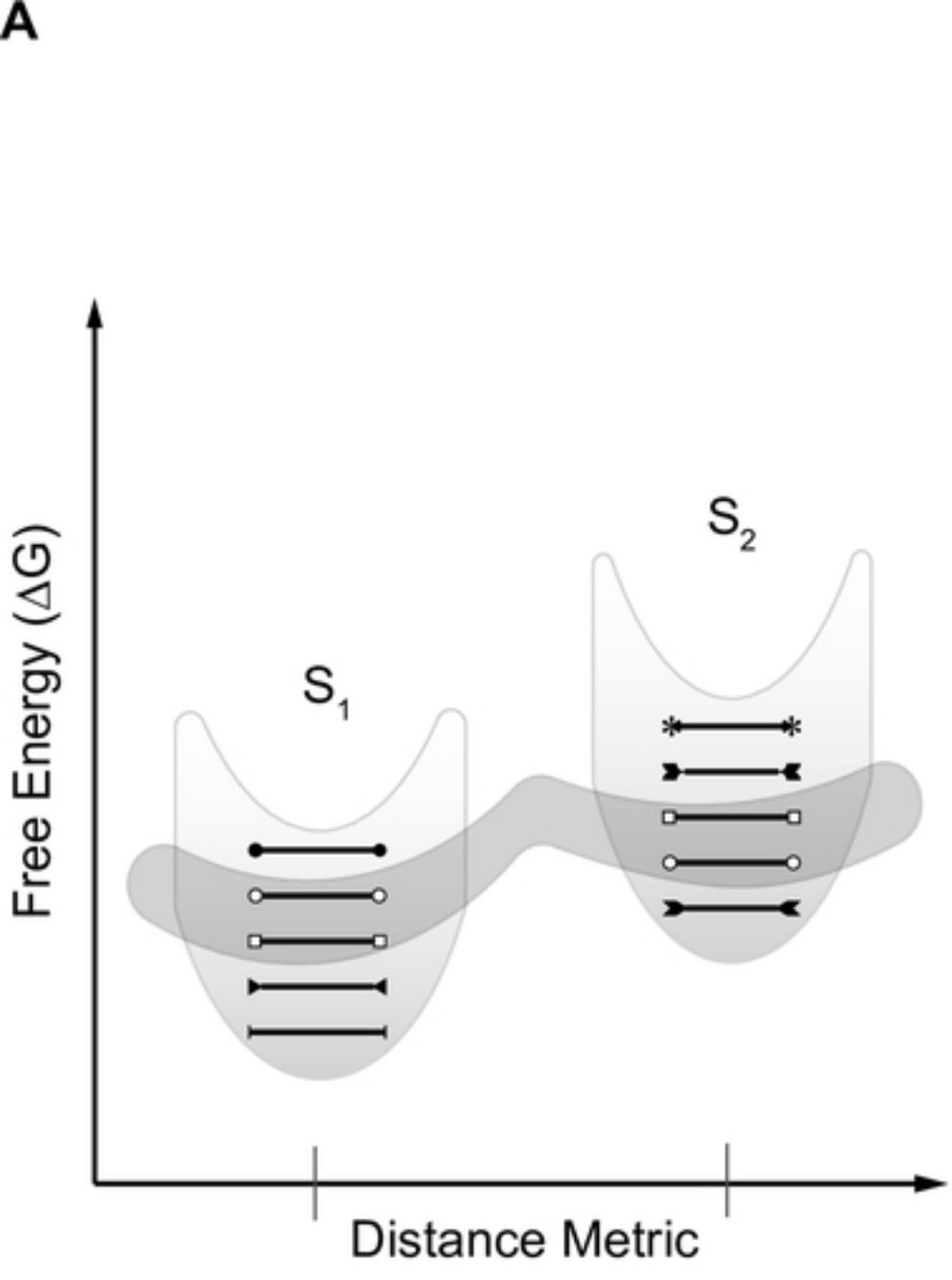
793 Supplementary Information

794 **S1 Fig. Design native sequence recovery and mutation profile variability comparisons**

795 **to PSI-BLAST profiles using relaxed and unminimized starting models.** (A)

796 Comparison of total native sequence recovery of relaxed and unminimized RECON MSD
797 and SSD designs to PSI-BLAST sequence profiles generated using the native sequence.
798 Asterisks indicate the significance of difference of means of each design in comparison to
799 the PSI-BLAST profile, with a *z*-test *p*-value < 0.01 represented by one asterisk, and a *p*-
800 value < 0.00001 by three asterisks. (B) Mutation frequency variances of designs in
801 comparison to a PSI-BLAST profile, normalized by protein length. The y-axis values
802 represent the average variability of mutation profiles for each designed residue in relation
803 to a PSI-BLAST profile, as described in Fig 3.
804
805 **S2 Fig. Difference in average amino acid exchangeability between sequence profiles.**
806 The x axis represents the original amino acid. The y axis represents the difference in
807 average mutation frequencies between two profiles, which are noted above each grid.
808 Along the diagonal axis, indicating native sequence conservation, values less than zero
809 (oranges) signify that the latter profile was more highly conserved, and values greater than
810 zero (blues to black) signify that the native residue was less conserved in the latter profile.
811 Not along the diagonal, values less than zero indicate that exchangeability of the native
812 residue to the indicated residue along the y axis was higher in the latter profile, whereas
813 values greater than zero indicate that exchangeability was lower in the latter profile.
814
815 **S3 Fig. Comparison of individual exchangeability rates.** (A) Scatterplots of each
816 exchangeability rate as observed in a PSI-BLAST profile compared to design profile. Both
817 the x and y axes represent the exchangeability frequency of a native amino acid to a
818 specific, non-native amino acid, with PSI-BLAST exchangeability rates along the x axis,

819 and design exchangeability rates along the y axis. For reference, a grey line drawn is drawn
820 along where the exchangeability rates would be equal between a PSI-BLAST and design
821 profile. Both an adjusted $r^2$and τ_β value is provided, along with the associated two-sided
822 p-value. Lighter points are found along the linear regression model, and darker points
823 represent outliers. (B) Measures of influence for individual exchangeability rate. The index
824 listed along the x axis refers to each exchangeability rate, indexed in order alphabetically.
825 For reference, in the first nineteen indices, the first index refers to the A to C mutation
826 frequency, followed by the next eighteen indices that correspond with A to D through Y
827 mutation frequencies. The measure of influential observation, or DFBETA index, is
828 represented along the y axis. The height and direction of each bar corresponds with the
829 change in regression model correlation coefficient without that particular observation.
830 Influential outliers that have $> \left|\frac{2}{\sqrt{N}}\right|$ index value, or ± 0.106 threshold, are colored in
831 black and are labeled with the associated mutation.
832
833 **S4 Fig. Correlation of dihedral angle RMSD and Cβ–Cβ distance deviation.** (A) The
834 x-axis represents dihedral RMSD, measured in radians, and the y-axis represents contact
835 proximity deviation, measured in Å. The hex bins shaded in grey are the number of residues
836 within the deposited PDB structure have have both a C_β–C_β distance deviation and dihedral
837 angle RMSD within a bin. (B) Axes represent same metrices as in Panel A, normalized by
838 z-score.


A
Free Energy (ΔG)
S1
S2
Distance Metric

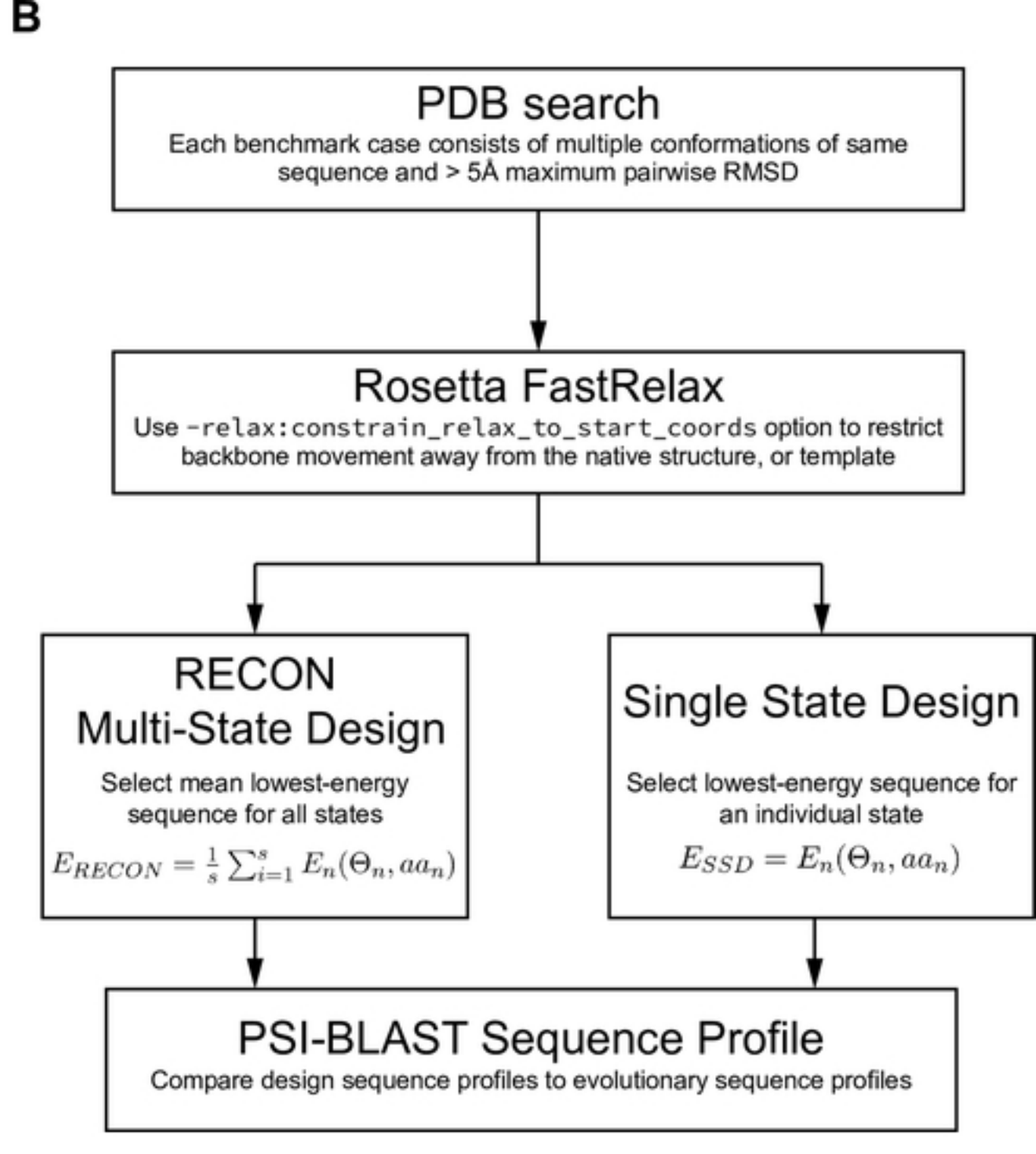

B
PDB search
Each benchmark case consists of multiple conformations of same sequence and > 5Å maximum pairwise RMSD
Rosetta FastRelax
Use -relax:constrain_relax_to_start_coords option to restrict backbone movement away from the native structure, or template
RECON Multi-State Design
Select mean lowest-energy sequence for all states
$E_{RECON} = \frac{1}{s}\sum_{i=1}^{s} E_n(\Theta_n, aa_n)$
Single State Design
Select lowest-energy sequence for an individual state
$E_{SSD} = E_n(\Theta_n, aa_n)$
PSI-BLAST Sequence Profile
Compare design sequence profiles to evolutionary sequence profiles

Fig 1

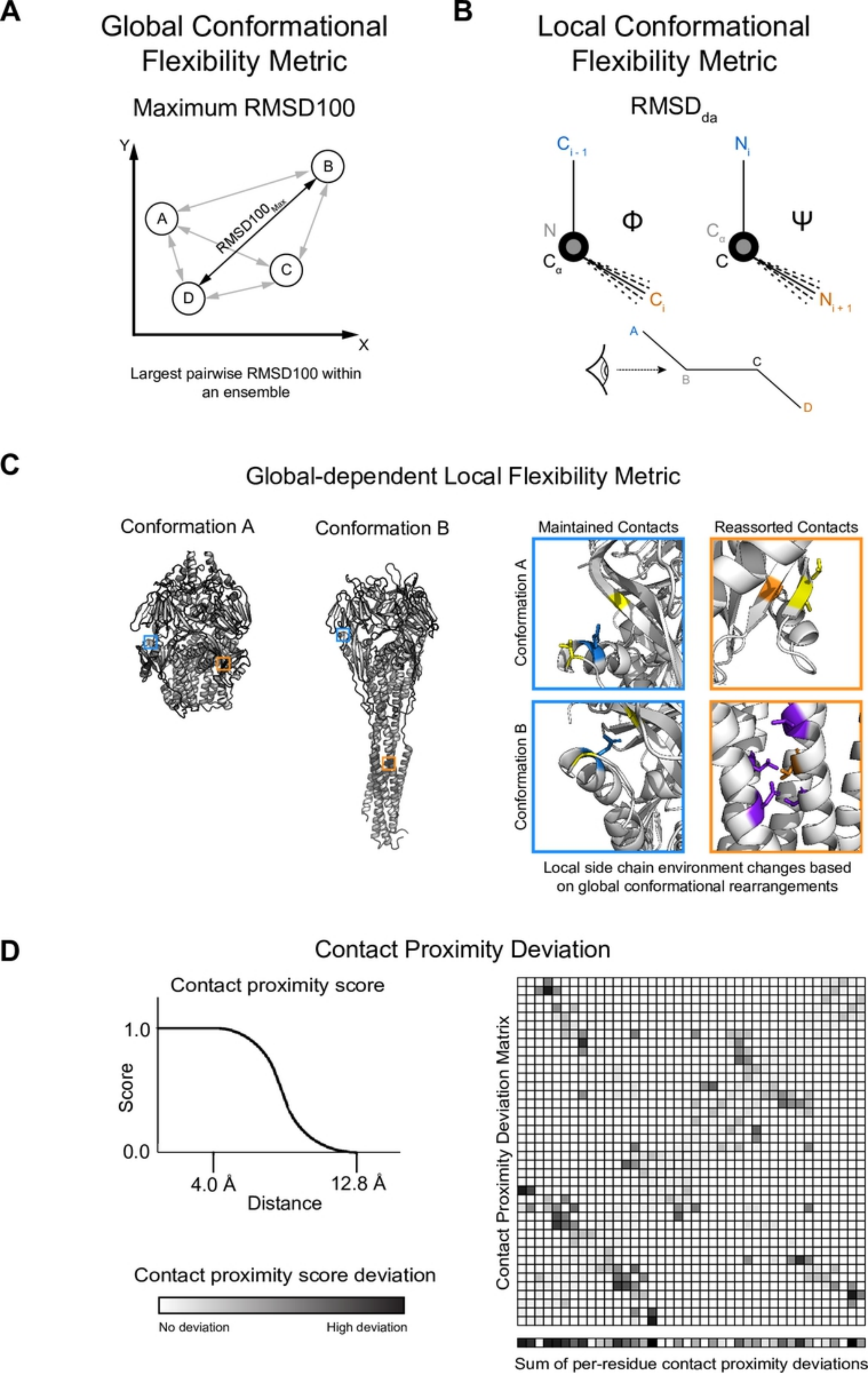

A
Global Conformational Flexibility Metric
Maximum RMSD100
Y
X
A
B
C
D
RMSD100Max
Largest pairwise RMSD100 within an ensemble
B
Local Conformational Flexibility Metric
RMSDda
Ci-1
Ni
N
Cα
Φ
Cα
C
Ψ
Ci
Ni+1
A
B
C
D
C
Global-dependent Local Flexibility Metric
Conformation A
Conformation B
Maintained Contacts
Reassorted Contacts
Conformation A
Conformation B
Local side chain environment changes based on global conformational rearrangements
D
Contact Proximity Deviation
Contact proximity score
1.0
0.0
Score
4.0 Å
12.8 Å
Distance
Contact proximity score deviation
No deviation
High deviation
Contact Proximity Deviation Matrix
Sum of per-residue contact proximity deviations

Fig 2

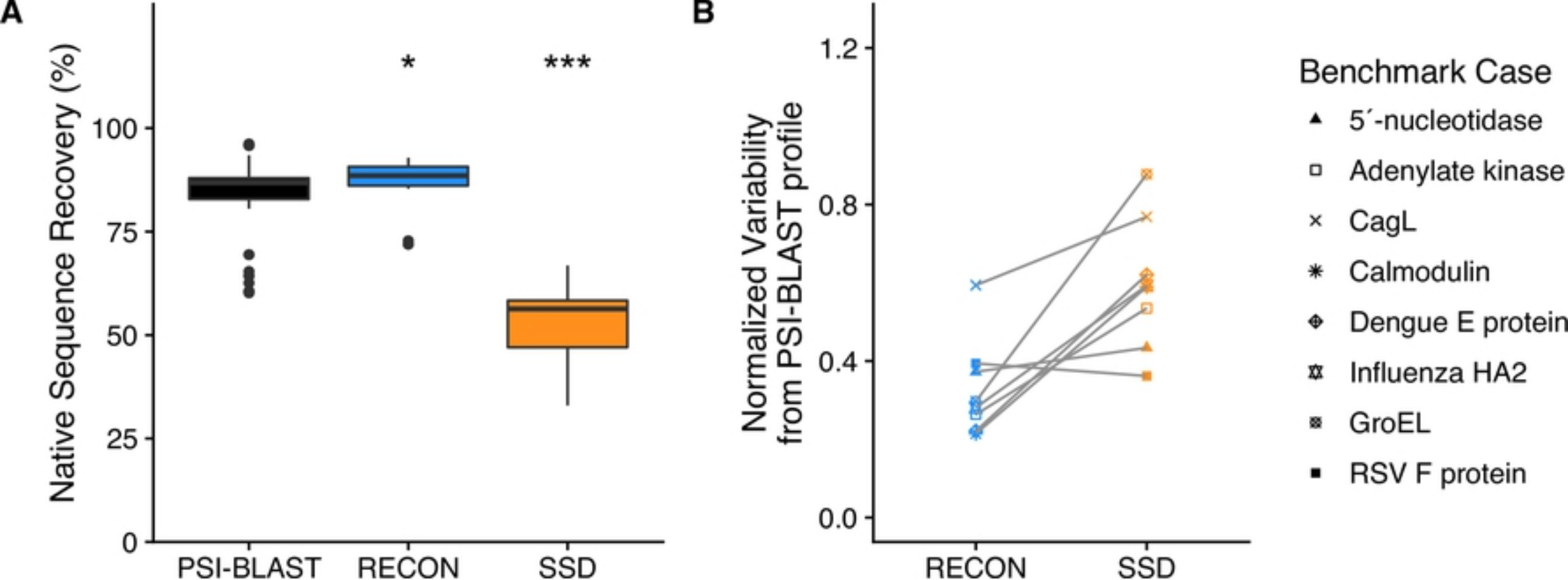

A
Native Sequence Recovery (%)
100
75
50
25
0
*
***
PSI-BLAST
RECON
SSD
B
Normalized Variability from PSI-BLAST profile
1.2
0.8
0.4
0.0
RECON
SSD
Benchmark Case
5´-nucleotidase
Adenylate kinase
CagL
Calmodulin
Dengue E protein
Influenza HA2
GroEL
RSV F protein

Fig 3

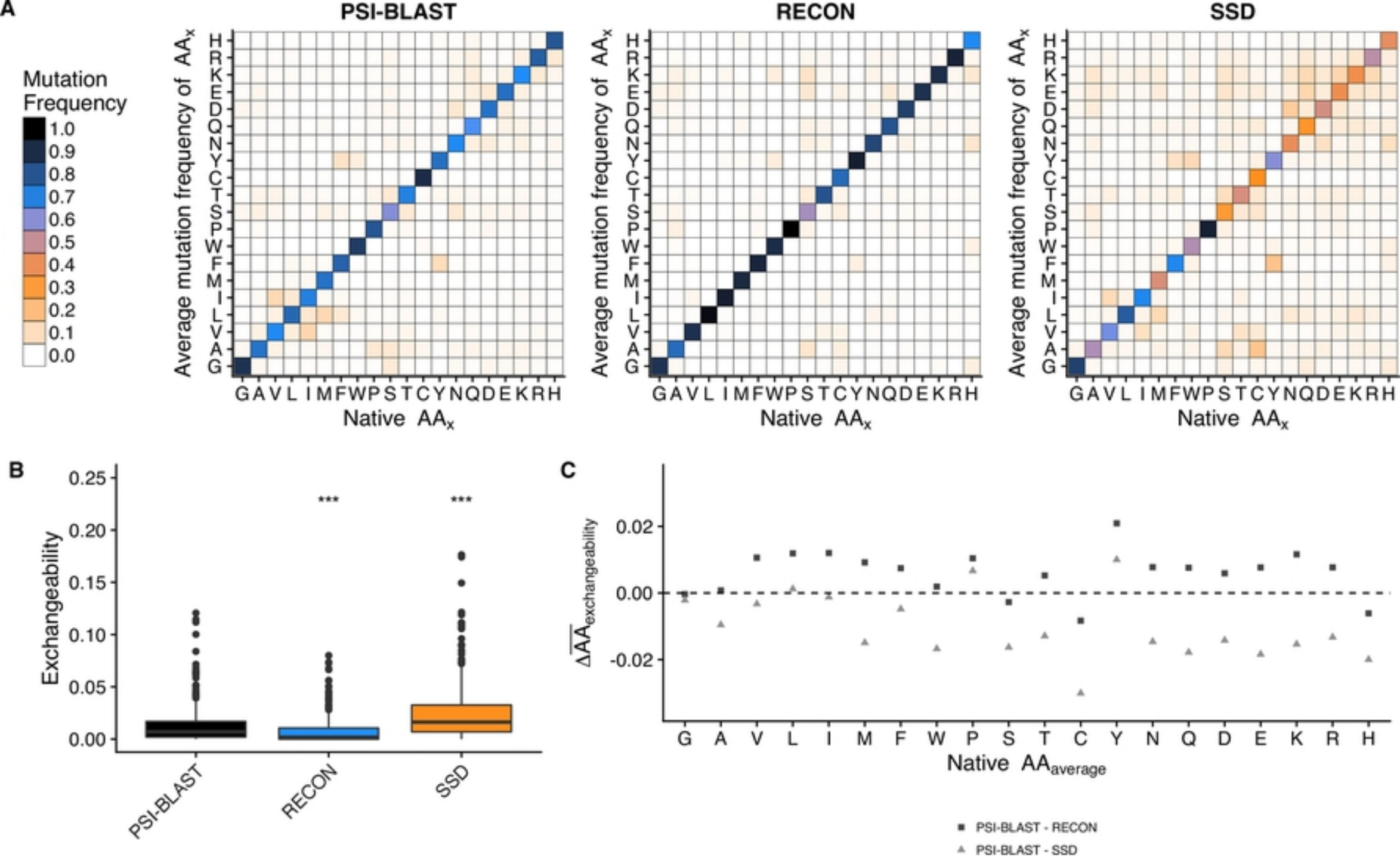

A
PSI-BLAST
RECON
SSD
Mutation Frequency
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Average mutation frequency of AAx
H R K E D Q N Y C T S P W F M I L V A G
G A V L I M F W P S T C Y N Q D E K R H
Native AAx
B
Exchangeability
0.25
0.20
0.15
0.10
0.05
0.00
***
***
PSI-BLAST
RECON
SSD
C
ΔĀA exchangeability
0.02
0.00
-0.02
G A V L I M F W P S T C Y N Q D E K R H
Native AAaverage
PSI-BLAST - RECON
PSI-BLAST - SSD

Fig 4

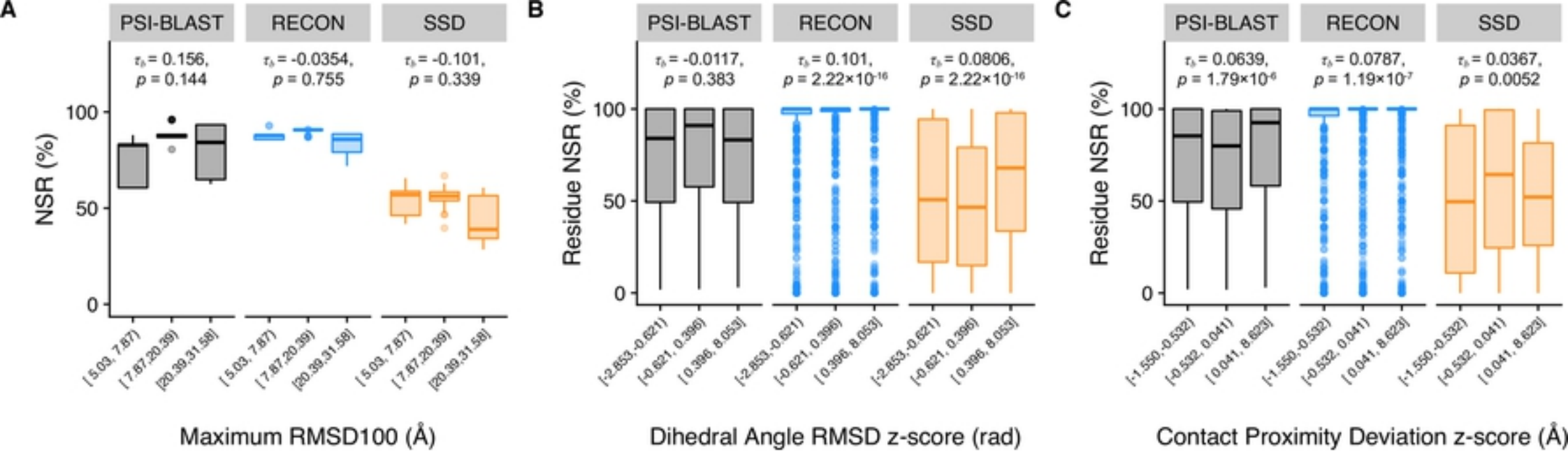

A
PSI-BLAST
RECON
SSD
τ_b = 0.156, p = 0.144
τ_b = -0.0354, p = 0.755
τ_b = -0.101, p = 0.339
NSR (%)
100
50
0
[5.03, 7.87)
[7.87,20.39)
[20.39,31.58]
Maximum RMSD100 (Å)
B
PSI-BLAST
RECON
SSD
τ_b = -0.0117, p = 0.383
τ_b = 0.101, p = 2.22×10^{-16}
τ_b = 0.0806, p = 2.22×10^{-16}
Residue NSR (%)
100
50
0
[-2.853,-0.621)
[-0.621, 0.396)
[0.396, 8.053]
Dihedral Angle RMSD z-score (rad)
C
PSI-BLAST
RECON
SSD
τ_b = 0.0639, p = 1.79×10^{-6}
τ_b = 0.0787, p = 1.19×10^{-7}
τ_b = 0.0367, p = 0.0052
Residue NSR (%)
100
50
0
[-1.550,-0.532)
[-0.532, 0.041)
[0.041, 8.623]
Contact Proximity Deviation z-score (Å)

Fig 5

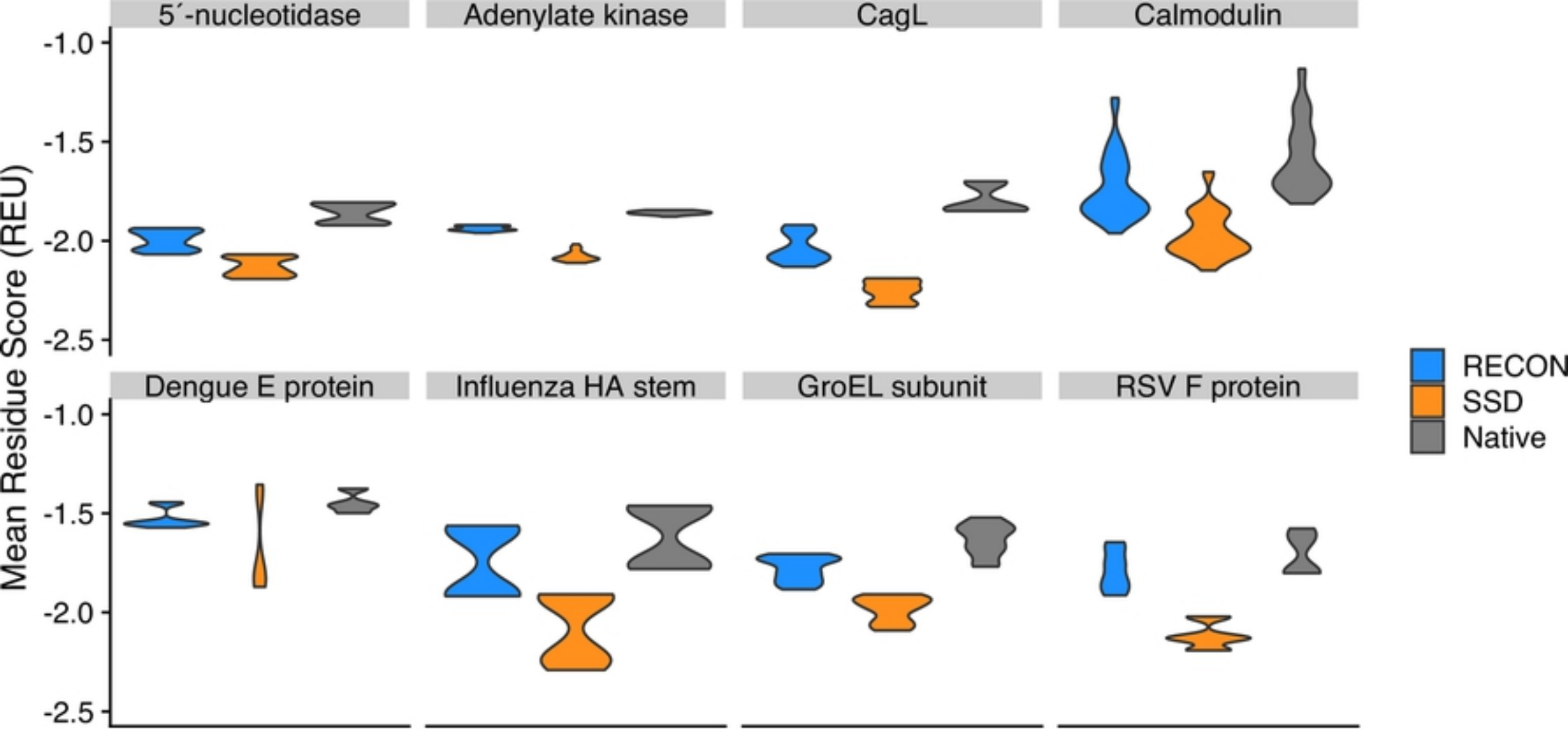

5´-nucleotidase
Adenylate kinase
CagL
Calmodulin
Dengue E protein
Influenza HA stem
GroEL subunit
RSV F protein
Mean Residue Score (REU)
-1.0
-1.5
-2.0
-2.5
RECON
SSD
Native

Fig 6

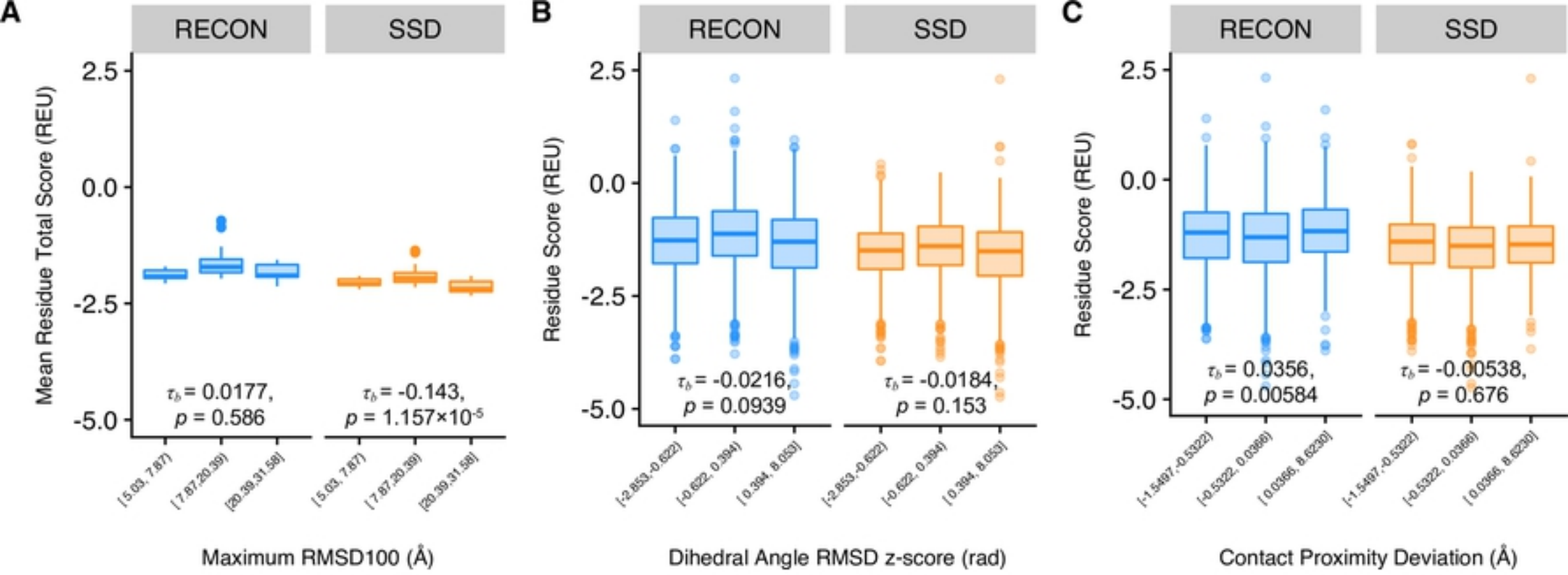

A
RECON
SSD
Mean Residue Total Score (REU)
2.5
0.0
-2.5
-5.0
τ_b = 0.0177,
p = 0.586
τ_b = -0.143,
p = 1.157×10^{-5}
[5.03, 7.87]
(7.87, 20.39]
(20.39, 31.58]
Maximum RMSD100 (Å)
B
RECON
SSD
Residue Score (REU)
τ_b = -0.0216,
p = 0.0939
τ_b = -0.0184,
p = 0.153
[-2.853, -0.622]
(-0.622, 0.394]
(0.394, 8.053]
Dihedral Angle RMSD z-score (rad)
C
RECON
SSD
Residue Score (REU)
τ_b = 0.0356,
p = 0.00584
τ_b = -0.00538,
p = 0.676
[-1.5497, -0.5322]
(-0.5322, 0.0366]
(0.0366, 8.6230]
Contact Proximity Deviation (Å)

Fig 7