A Strategy for Proline and Glycine Mutations to Proteins with Alchemical Free Energy Calculations

Ryan L. Hayes[†] and Charles L. Brooks III[*,†,‡]

[†]*Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109, United States*

[‡]*Biophysics Program, University of Michigan, Ann Arbor, Michigan 48109, United States*

E-mail: brookscl@umich.edu

Phone: (734) 647-6682

Abstract

Computation of the thermodynamic consequences of protein mutations holds great promise in protein biophysics and design. Alchemical free energy methods can give improved estimates of mutational free energies, and are already widely used in calculations of relative and absolute binding free energies in small molecule design problems. In principle, alchemical methods can address any amino acid mutation with an appropriate alchemical pathway, but identifying a strategy that produces such a path for proline and glycine mutations is an ongoing challenge. Most current strategies perturb only side chain atoms, while proline and glycine mutations also alter the backbone parameters and backbone ring topology. Some strategies also perturb backbone parameters and enable glycine mutations. This work presents a strategy that enables both proline and glycine mutations and comprises two key elements: a dual backbone with restraints and scaling of bonded terms, facilitating backbone parameter changes, and a soft bond in the proline ring, enabling ring topology changes in proline mutations. These elements also have utility for core hopping and macrocycle studies in computer-aided drug design. This new strategy shows slight improvements over an alternative side chain perturbation strategy for a set T4 lysozyme mutations lacking proline and glycine, and yields good agreement with experiment for a set of T4 lysozyme proline and glycine mutations not previously studied. To our knowledge this is the first report comparing alchemical predictions of proline mutations with experiment. With this strategy in hand, alchemical methods now have access to the full palette of amino acid mutations.

1 Introduction

The effects of amino acid mutations in proteins are of great importance in medicine, where they determine the mechanism of genetic diseases[1] and control evolutionary pathways of drug resistance,[2,3] and in biotechnology, where protein design relies on iterative mutations to optimize target properties.[4–6] The ability to predict the effect of these mutations using computational methods is highly desirable both to streamline experimental efforts and aid in their interpretation. Consequently, many methods have been developed to compute mutational free energies with physics or knowledge-based potentials, machine learning, or genomic sequencing data.[3,7–12] These methods enable rapid estimation of mutational free energy changes, but can suffer in accuracy due to approximations in the

equilibrium ensemble and force field, or from poor generalizability to new ligands, nonnatural amino acids, and problems beyond the training data. Alchemical free energy methods can offer better accuracy and generalizability at an increased computational cost, and have already found widespread use in computer-aided drug design.[13,14] This has motivated a growing interest in applying alchemical methods to protein mutations.[15–22]

While alchemical free energy calculations of protein mutations have shown great promise, none of these previous studies[15–22] have been able to treat mutations to or from proline, and only a few included mutations to or from glycine.[17,19,20] This may seem like a minor limitation in testing and validation studies and some design studies when one can choose to avoid inconvenient mutations, but in many cases, such as comparing evolutionarily related sequences[23] or evaluating redesigned proteins against their natural homologues,[24,25] the sequences are already defined and often include a few proline mutations. In principle there is no reason alchemical calculations cannot address proline mutations given an appropriate alchemical pathway; the limitation lies in the perturbation strategies employed in previous studies, which do not generalize to proline. Indeed, two previous studies have examined a single proline perturbation, but they neither compared to experimental measurement of the free energy change, nor described the perturbation strategy in sufficient detail.[26,27] Consequently, description and experimental testing of a proline perturbation strategy is needed.

In this work, we present a perturbation strategy that enables treatment of proline mutations. This strategy also enables glycine mutations, which can be problematic for some free energy approaches. We begin with a discussion of alchemical free energy methods and the perturbation strategy. Next, the new strategy is validated on a previous T4 lysozyme data set lacking proline and glycine mutations to ensure it does not degrade accuracy for mutations that can be treated with other strategies. Finally, the strategy is tested on a new set of ten proline and glycine mutations in T4 lysozyme. We anticipate this strategy will inspire treatment of proline mutations for several alchemical methods, and the underlying principles will facilitate core hopping and macrocycle calculations in computer-aided drug design.

2 Alchemical Methods and a Proline Perturbation Strategy

Alchemical methods all use a similar approach to calculate free energy differences (Figure 1). Because free energy is a state function, the relative free energy difference upon mutation for a physical process like folding can be expressed as either the difference of the horizontal physical processes or the vertical alchemical processes in Figure 1. Alchemical methods utilize the alchemical processes because they converge much more rapidly. Most alchemical free energy methods introduce an alchemical coupling parameter λ into the potential energy function for the system that mutates from one sequence to the other. In the conventional equilibrium methods of thermodynamic integration,[28] free energy perturbation,[29] and the multistate Bennett acceptance ratio,[30] several simulations are run at closely spaced, fixed values of λ. In nonequilibrium methods like fast growth thermodynamic integration,[31] λ is a continuous driving variable. Finally, in the multisite λ dynamics (MSλD) technique pioneered in our lab,[32,33] λ is a continuous degree of freedom that fluctuates on equal footing with spatial degrees of freedom. While examples of each of these methods have shown impressive accuracy in predicting the effect of protein mutations,[18,19,22] MSλD is unique in that λ can be generalized to a multidimensional alchemical space, allowing scalable and efficient treatment of multiple mutations. This makes MSλD uniquely well suited to the combinatorial sequence spaces encountered in protein design.

Alchemical methods allow representation of multiple sequences by partitioning the system into environment atoms, which are present in all sequences, and mutating or alchemical atoms,

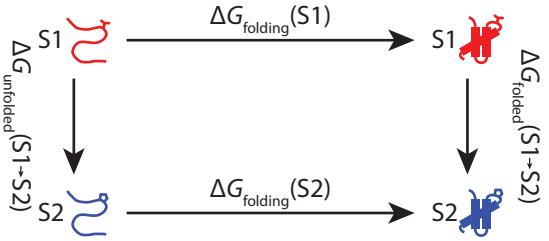


Figure 1: To determine the relative free energy difference of a process like folding upon mutation, alchemical methods take the difference of the two vertical alchemical processes rather than the difference of the two horizontal physical processes, because simulations of the alchemical processes converge more rapidly.

which are unique to a particular mutation. For convenience, an atom may be removed from the environment by creating a copy of it in each alchemical set; for example, while both mutating residues may contain a Cγ atom, it may be more convenient to include a Cγ atom in each alchemical set so that the same alchemical set can be used when mutating to alanine, which lacks a Cγ atom. Within MSλD, the potential energy representing the hybrid system is

$$U = U_{0,0} + \sum_{s}^{M}\sum_{i}^{N_s} \lambda_{si}(U_{0,si} + U_{si,si}) + \sum_{s}^{M}\sum_{t>s}^{M}\sum_{i}^{N_s}\sum_{j}^{N_t} \lambda_{si}\lambda_{tj}U_{si,tj} + U_{\text{bias}} \quad (1)$$

where λ_{si} is the alchemical scaling parameter of mutation i at site s, the sum of λ_{si} over i at each site is constrained with implicit constraints,[34] $U_{0,0}$ are the interaction terms involving only environment atoms, $U_{0,si}$ and $U_{si,si}$ are the interactions of mutating atoms with the environment and among themselves, $U_{si,tj}$ are interactions between mutating atoms at different sites, and U_{bias} is a biasing potential typically obtained with adaptive landscape flattening (ALF) to optimize sampling.[22,35] This reduces to the potential energy function for a particular sequence (plus some noninteracting dummy atoms) at alchemical endpoints where all λ values are either 0 or 1, but allows transformation between sequences through nonphysical alchemical intermediates where λ values are between 0 and 1.

In practice, typically only nonbonded electrostatic and Lennard-Jones interactions are scaled by λ, while bonded interactions, referring to bonds, angles, dihedrals, impropers, and CMAP interactions,[36,37] are not scaled by λ. If bonded terms are scaled by λ, mutating atoms can occupy unreasonable geometries when λ is 0, which introduces serious convergence artifacts; for example in MSλD, λ cannot fluctuate back away from 0 because of the large force $-\partial U/\partial\lambda$ on λ once atoms are out of position. Conversely, artifacts can arise if the unscaled bonded terms for an alchemical set of atoms affect the ensemble beyond this set of atoms when λ is 0. When a special allowed set of alchemical bonded terms remains unscaled and other alchemical bonded terms are scaled to zero, their contribution to the partition function may be factored out by a change of variables, resulting in an additive constant to the free energy, which cancels out in the difference between the two alchemical processes.[38] This allowed set of unscaled bonded interactions includes all bonded interactions between a set of alchemical atoms, their bonds to one environment atom, their angles involving that atom and one additional environment atom, and their dihedrals involving those two atoms and one additional environment atom. Alchemical calculations sometimes include extra unscaled bonded interactions or restraints between an alchemical region and the environment or other alchemical regions, but these cannot be guaranteed to cancel out. The dual concerns of ensuring mutating atoms maintain reasonable geometries yet do not perturb the ensemble when λ is 0 strongly influence the development of our perturbation strategy below.

To motivate the new perturbation strategy for proline and glycine, we outline the previous MSλD side chain perturbation strategy[22] and the two fundamental problems that must be addressed for any amino acid perturbation strategy to treat proline and glycine, followed by the new perturbation strategy that addresses these problems. Simulations were carried out using the CHARMM36 forcefield[39,40] in

the CHARMM software package[41,42] using the block module. Backbone atoms (N, HN, Cα, Hα, C, and O) were considered part of the environment, and atoms for each mutating side chain were included with unscaled bond, angle, and improper interactions. Each mutating side chain has its own Cβ atom with three unscaled angle interactions Cβ-Cα-N, Cβ-Cα-C, and Cβ-Cα-Hα, which is two more unscaled angles than allowed as outlined above.[38] This effectively double counts and rigidifies these angles for each Cβ present, but the high accuracy of the approach suggests that the decreased amplitude of angle vibrations has similar effects on both ensembles.[22] The validity of this assumption is verified below by scaling some or all of these angles. In contrast, the double counting of the ϕ dihedral Cβ-Cα-N-C and ψ dihedral Cβ-Cα-C-O would affect not just vibrations but also Ramachandran distributions, so all perturbed dihedral interactions were scaled by λ.

There are two fundamental problems with this approach. First, for proline and glycine mutations, backbone parameters change, which cannot be implemented in the block module of CHARMM without increasing the set of mutating atoms to include the backbone, and can also lead to problems for some other alchemical methods implemented in NAMD[43] and GROMACS.[44] Second, a problem for all alchemical methods is that in proline the side chain is bonded to the backbone at both C_α and N, which is one more bond than allowed above[38] and perturbs the ensemble by preventing free rotation around the backbone ϕ angle, even when λ for proline is zero.

The first problem is that proline and glycine mutations change parameters of backbone atoms generally included in the environment. Changing parameters of environment atoms has been mostly implemented in GROMACS,[45] with the notable exception of CMAP interactions. Implementing changing parameters of environment atoms within the block module of CHARMM would have required extensive code restructuring, so the mutating region was expanded to include the entire residue, leading to multiple copies of the backbone atoms. The whole residue is connected to the environment by two bonds, so care must be taken to avoid artifacts. A simple test system mutating glycine to glutamine in a pentapeptide environment revealed that dihedral and CMAP scaling were required to obtain the correct Ramachandran distribution. With CMAP and dihedral terms scaled, the remaining bonded terms perturbed the glycine N-Cα-C angle from 115.0° to 113.7°, indicating the two unscaled glutamine bonds to neighboring residues distort the glycine even when λ for glutamine is zero.

Therefore, we apply a strategy that allows one to factor out the contribution of the side chain from the partition function, followed by factoring out the contributions of the backbone atoms when λ for that residue is 0, and rigorously guarantees the endpoint ensembles are not perturbed. Bond and angle terms are scaled if they contain only environment and analogous backbone atoms (N, Cα, Hα, C, and O), but are left unscaled if they contain any side chain atoms (or HN, which is missing from proline), while all other bonded terms are scaled regardless. The only obstacle to factoring out the side chain (and HN) is that three unscaled Cβ-Cα-X angles and two unscaled HN-N-X angles remain when only one of each is allowed. Three different treatments of these unscaled angles are tested below. To prevent the analogous backbone atoms from adopting distorted configurations, they are tightly harmonically restrained together (see Supporting Information for details), similar to a recent ligand perturbation approach in NAMD using holonomic constraints.[46] This approach is rigorous, because after the side chain and HN are factored out of the partition function, each analogous backbone atom is an isolated harmonic oscillator that may also be factored out. For generalization to multiple mutation sites, if N_s and N_t mutations to adjacent residues are made, all $N_s \times N_t$ inter-residue C-N bonds are included and scaled by the product of their λ values. While most backbone parameters can be changed as a function of λ in GROMACS, CMAP scaling is not yet implemented,[45] and the distortion of the Ramachandran distribution in our pentapeptide system

highlights that CMAP scaling must be implemented before glycine mutations can be performed in GROMACS with the CHARMM36 force field. Furthermore, by replacing tight harmonic restraints with holonomic constraints, this strategy may be adapted to enable glycine mutations NAMD.

This approach is still not sufficient for proline, where unscaled bonds in the ring prevent free rotation around the backbone ϕ dihedral even when λ is 0, and perturb the Ramachandran distribution of the alternative residues unphysically. Fundamentally, one of the bonds in the proline ring must be scaled to zero with λ, but the two previous studies of a proline perturbation failed to mention this or or describe their solution.[26,27] In this work, we use recently developed soft bonds[47,48] to break the ring:

$$U = \frac{\frac{1}{2}{\lambda_{si}}^{n_\alpha} k(r - r_0)^2}{1 + (1 - {\lambda_{si}}^{n_\alpha})(r - r_0)^2/{r_\alpha}^2} \qquad (2)$$

where k and r_0 are the bond spring constant and equilibrium distance. Previous work chose $r_\alpha = 0.7$ Å for core hopping[47] and $r_\alpha = 1.4$ Å for macrocycle applications,[48] and we choose $r_\alpha = 1$ Å in this work. Previous work only included the special case $n_\alpha = 1$, but we used $n_\alpha = 2$ because it gives smoother free energy profiles (see Supporting Information and Table S1 for details). We apply the soft bond to the Cβ-Cγ bond to avoid any dihedrals through the soft bond that include atoms from the previous residue, which could also be mutating. Any Urey-Bradley interactions through this bond are also treated with soft bonds, and other bonded terms through this bond (e.g. angles) are scaled linearly by ${\lambda_{si}}^{n_\theta}$ with $n_\theta = 1$, rather than by ${\lambda_{si}}^{n_\alpha}$. The two sets of side chain atoms bonded to Cα and N can then be factored out of the partition function separately because they no longer interact when λ is 0. In testing soft bonds on perturbations between 5, 6, and 7 membered ring inhibitors of BACE1 previously studied in our lab,[49] we discovered that soft-core interactions,[35] which were previously not applied to 1-4 nonbonded interactions, had to be applied to 1-4 interactions as well to prevent serious artifacts. Thus, we apply soft-core interactions to 1-4 nonbonded interactions to make the approach easily generalizable, as well as to prevent possible artifacts for 1-4 interactions between Hβ and Hγ atoms that could possibly overlap.

Two control tests were performed to test the thermodynamic rigor of the perturbation strategy. First, the Ramachandran distributions were compared for plain molecular dynamics and the present perturbation strategy with λ fixed at a $\lambda = 1$ endpoint to ensure the perturbation strategy does not perturb the endpoint ensemble. For 40 ns simulations of the pentapeptide model of the unfolded state, deviations were observed due to slow transitions between basins, but for longer 400 ns simulations, both methods converged to the same distribution (Figure S2 & S3). Second, the free energy around a closed thermodynamic cycle proline to glycine to alanine and back to proline was computed. Unlike other alchemical methods, MSλD need not use closed thermodynamic cycles to connect a network of pairwise free energy comparisons since all perturbations can be evaluated in the same simulation, but cycles can still highlight potential artifacts. We find the proline to glycine leg exhibits substantial variability on the 40 ns time scale, while the other legs appear converged. For longer 400 ns simulations, all legs converge to give a cycle closure error of less than 0.2 kcal/mol (Table S4). Only one of the ten mutations examined subsequently involves a proline to glycine mutation, and while we observe slightly improved agreement with experiment for 400 ns simulations (see Supporting Information), the results highlighted in the main text use 40 ns simulations.

This perturbation strategy incorporating both scaled bonded interactions of restrained analogous atoms and soft bonds allows mutation between any amino acid including proline at several sites by MSλD within the CHARMM molecular dynamics package, and should give insight into how to treat proline and glycine mutations with other alchemical methods in other software packages.

3 T4 Lysozyme Control Mutations

To test this perturbation strategy we first sought to ensure it gave consistent results for non-proline and glycine mutations with the previous side chain perturbation strategy. Therefore, the set of previously calculated T4 lysozyme point mutations were recalculated as described previously,[22] changing only the perturbation strategy. Simulations used particle mesh Ewald electrostatics,[50] modeled the folded alchemical transformation starting from PDB 1L63,[51] and approximated the unfolded alchemical transformation with a capped pentapeptide centered on the mutating residue.

It is often more informative to compare computational results obtained with different methods with each other than to compare with experiment, because the goal in methods development (in contrast to force field development or design applications) is to converge to the force field correct answer, which may or may not agree with experiment, depending on the quality of the force field. However, it is also useful to compare with experiment, because artifacts in the method can lead to systematic errors that tend to increase the deviation from experimental values. Experimental values are taken from reference 52, and Pearson correlation (R), mean unsigned error $\langle|\Delta x|\rangle$ (MUE), and root mean squared error $(\langle\Delta x^2\rangle - \langle\Delta x\rangle^2)^{1/2}$ (RMSE) are evaluated. We evaluate the centered RMSE, which includes the native sequence in the averages, rather than the larger uncentered RMSE ($\langle\Delta x^2\rangle^{1/2}$) because it is more appropriate for relative free energies and for consistency with our previous study of T4 lysozyme.[22] We only include neutral mutations in the statistics, otherwise statistical variation in the single M102K mutation dominates the statistics.

To determine whether the three unscaled angles through the Cβ-Cα bond to the backbone caused artifacts, since only one is rigorously allowed, additional simulations were run with some or all of these angles scaled. In one case, all but one of these angles were scaled (see Supporting Information for details), but this allowed free rotation of the Cα-Cβ bond into nonphysical orientations when λ was small. Though no chirality flips were observed, the increased rigor translated to poorer results (Table 1). Therefore another set of simulations was run with a harmonic angular restraint between Cα-Cβ vectors; the restraint counted as the one allowed angle term, thus all three of the Cα angles were scaled by λ. This gave comparable results to the simulations without the angles scaled, suggesting that there are not substantial artifacts when the angles are unscaled (Table 1).

Figure 2 shows the whole residue strategy with scaling of all three Cα angles achieves excellent agreement with experiment, and nearly identical results with the side chain strategy without Cα angle scaling. Furthermore, the new whole residue strategy seems to give slightly improved results relative to the original side chain strategy (Table 1), though this is likely just statistical variation. These findings suggest the high accuracy previously reported with the side chain strategy can also be expected from the whole residue perturbation strategy.

4 T4 Lysozyme Proline and Glycine Mutations

Having shown the whole residue strategy gives comparable or improved results in T4 lysozyme mutations previously evaluated with the side chain strategy, we turn our attention to proline and glycine mutations that could not be addressed with the side chain strategy. As a test set, we chose all mutations between neutral amino acids and either proline or glycine made to T4 lysozyme in the C54T/C97A background listed in reference 52. This comprises ten mutations: Y25P, L33G, P37G, S44G, S44P, G56M, T59G, Q69P, L99G, and V149G. The folded protein and unfolded pentapeptides were set up as described previously, including protonation states at a pH of 3.0 and 5.4.[22] Two mutations in this set were measured experimentally at a pH of 6.5, but PROPKA calculations[53] indicated protonation states of all residues were the

Table 1: Comparison with Experiment of Neutral, Non-Proline/Glycine Mutations in T4 Lysozyme

	RMSE	MUE	R
Side chain - scale 0 Cα angles	1.08	0.91	0.896
Whole res. - scale 0 Cα angles	1.02	0.80	0.894
Side chain - scale 2 Cα angles	1.25	1.00	0.862
Whole res. - scale 2 Cα angles	1.13	0.97	0.897
Side chain - scale 3 Cα angles	1.07	0.91	0.901
Whole res. - scale 3 Cα angles	1.05	0.94	0.886

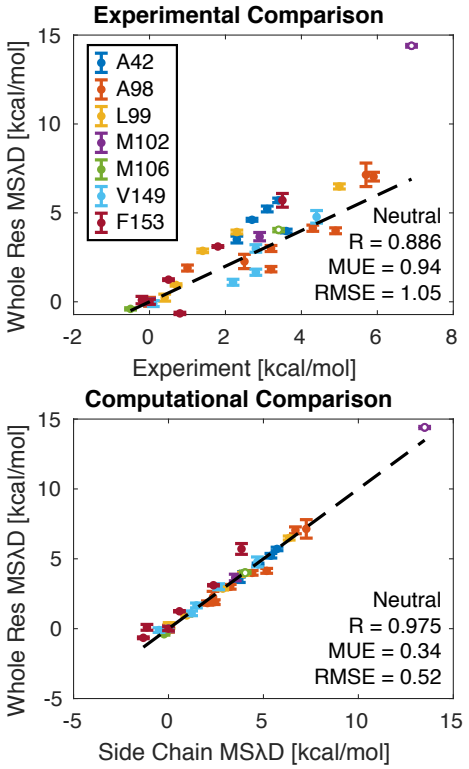


Figure 2: For a previous mutation set lacking proline or glycine, the whole residue perturbation strategy agrees well with experiment (top) and gives virtually identical results to the side chain perturbation strategy (bottom). Statistics exclude the two charge changing mutations, shown as open circles. The dashed line is $y = x$; experimental and side chain data are from references 52 and 22, respectively.

same as at a pH of 5.4. Production simulations were run with 5 independent trials of 40 ns each. To our knowledge, this is the first alchemical study of proline mutations that compares with experimentally measured free energies.

MSλD simulations of proline and glycine mutations agree well with experiment, but not as well as simulations of mutations excluding proline and glycine (Figure 3). The statistics of R = 0.876, MUE = 2.05 kcal/mol, and RMSE = 1.65 kcal/mol in Figure 3 were obtained scaling all Cα angles; statistics without Cα angles scaled were comparable with R = 0.870, MUE = 1.93 kcal/mol, and RMSE = 1.59 kcal/mol (see Supporting Information Figure S1 and Table S2). Two of the largest studies including glycine mutations both observed poorer results for glycine mutations than other kinds of mutations,[17,19] and our glycine MUE of 1.97 kcal/mol (or 1.84 kcal/mol with unscaled angles) is comparable to the glycine MUE of 2.1 kcal/mol in reference 19. To our knowledge this study is the first comparison of alchemical simulations with experiment for proline, and suggests that like glycine, they will also have larger errors than other mutations.

The most likely source of increased error for proline and glycine mutations is the strong effect on the flexibility of the backbone, though the two worst outliers are both buried mutations from leucine to glycine whose stability changes are driven instead by creation of a buried cavity. The destabilizing effect of proline and glycine mutations is generally overpredicted, suggesting the 40 ns simulations may be too short for the protein to relax to accommodate the mutation. Overprediction can occur if a relaxation process that mitigates the destabilizing effect of a mutation is too slow to ob-

serve computationally. It is also possible the experimental results are partially responsible for the discrepancy; free energies were reported at high temperatures, and extrapolating back to the simulation temperature of 300 K gives marginally improved RMSE and substantially improved mean signed error (see Supporting Information). Overall, it is unsurprising that the computational results are poorer for these difficult mutations, yet it is encouraging that the results are still reasonably accurate.

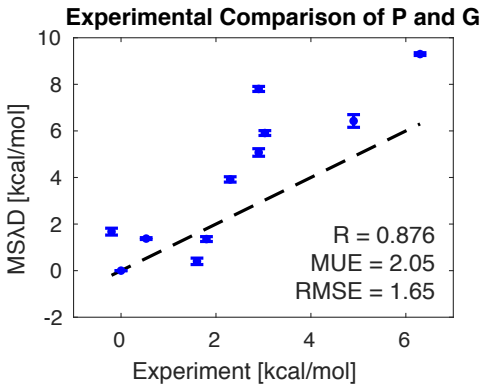


Figure 3: Comparison of MSλD with experiment for proline and glycine mutations. The whole residue strategy was required to evaluate proline and glycine mutations. The dashed line is $y = x$; experimental data is from reference 52.

5 Discussion

In this paper we described the implementation of a protocol for mutating amino acids in proteins where one endpoint contained either glycine or proline. Our results demonstrate that the whole residue perturbation strategy enables accurate computation of mutational free energies for mutations to and from all twenty amino acids. As mentioned in the introduction, the ability to compute the effect of proline mutations is useful in studies of evolution,[23] where the effects of mutations including proline constrain viable evolutionary paths. It is especially important in protein design,[24,25] where we envision MSλD refining designs from less rigorous, but much faster, methods like Rosetta, because proline mutations often have larger effects than other mutations. This new perturbation strategy opens these and other applications of MSλD.

We expect this perturbation strategy will also be relevant to studies of protein mutations with other alchemical free energy methods. Scaling bonded interactions of restrained analogous atoms may be helpful in some alchemical software implementations like NAMD but unnecessary in others where the parameters of bonded interactions can vary as a function of λ. The demonstration that soft bonds enable accurate calculation of the effects of proline mutations is useful for all alchemical methods and should encourage future studies to include proline mutations. Other details, such as noting dihedrals and CMAP terms should always be scaled, that 1-4 interactions should be treated with soft cores, and that angular restraints allow extra angles to be scaled by λ without sacrificing sampling should aid in crafting perturbation strategies for other alchemical methods.

We also anticipate the two key techniques introduced in the whole residue perturbation strategy, namely scaling bonded interactions of restrained analogous atoms and judicious use of soft bonds, will be useful in many other MSλD studies of ligand perturbations in drug design. Scaling and restraining can be used when perturbations involve core atoms that cannot easily be treated as substituents, or for atoms whose parameters change only slightly in response to a perturbation. Soft bonds represent a more aggressive approach that is warranted when perturbations open, close, or resize a ring, or when a perturbation to a core changes connectivity. The use of soft bonds has already enabled studies of core hopping and macrocyles with free energy perturbation,[47,48] and should now enable them within the MSλD framework as well. During the D3R grand challenge 2, the core hopping transformation between ligands 91 and 93 could have been easily achieved by scaling and restraining, rather than the less rigorous approach that we improvised at that time.[54] Soft bonds would have been necessary to efficiently study the macrocycle perturbations with MSλD in the D4R grand challenge 4.[55] Finally, scal-

ing and restraining enables a broader scope of MSλD multisite systems, because alchemical regions may be directly bonded to each other rather than requiring two intervening environment atoms.

6 Conclusions

We have presented a perturbation strategy that allows proline mutations, and demonstrated that it gives accurate predictions of the effects of mutations for all amino acids including proline and glycine. The underlying principles will also enable a wider array of small molecule perturbations in computer-aided drug design. With this strategy, MSλD is now poised to study and design proteins with the full palette of amino acid mutations.

Acknowledgement We gratefully acknowledge funding from the NIH (GM130587 and GM37554) and the NSF (CHE 1506273). We also thank Susan Marqusee and Charlotte Nixon for presenting us with a ribonuclease H problem including proline and glycine mutations that required the development of these methods.

Supporting Information Available

Supporting information contains technical details of the implementation in CHARMM, free energies for individual mutations, a discussion of sources of error, and controls examining thermodynamic cycles and Ramachandran distributions.

7 Data Availability Statement|

Example CHARMM input scripts are available for download at `https://brooks.chem.lsa.umich.edu/index.php?page=proline_and_glycine_perturbations&subdir=articles/resources/data`

8 Note Added in Proof

After the manuscript was accepted, we became aware of a recently published systematic alchemical FEP study involving 20 proline mutations. See reference 56.

References

(1) Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S. Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods. *Phys. Rev. Lett.* **2003**, *91*, 140601.

(2) Rizzo, R. C.; Wang, D.-P.; Tirado-Rives, J.; Jorgensen, W. L. Validation of a Model for the Complex of HIV-1 Reverse Transcriptase with Sustiva through Computation of Resistance Profiles. *J. Am. Chem. Soc.* **2000**, *122*, 12898–12900.

(3) Figliuzzi, M.; Jacquier, H.; Schug, A.; Tenaillon, O.; Weigt, M. Coevolutionary Landscape Inference and the Context- Dependence of Mutations in Beta-Lactamase TEM-1. *Mol. Biol. Evol.* **2015**, *33*, 268–280.

(4) Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* **2003**, *21*, 1364–1368.

(5) Savile, C. K.; Janey, J. M.; Mundorff, E. C.; Moore, J. C.; Tam, S.; Jarvis, W. R.; Colbeck, J. C.; Krebber, A.; Fleitz, F. J.; Brands, J. et al. Biocatalytic Asymmetric Synthesis of Chiral Amines from Ketones Applied to Sitagliptin Manufacture. *Science* **2010**, *329*, 305–309.

(6) Silva, D.-A.; Yu, S.; Ulge, U. Y.; Spangler, J. B.; Jude, K. M.; ao Almeida, C. L.; Ali, L. R.; Quijano-Rubio, A.; Ruterbusch, M.; Leung, I. et al. De Novo Design of Potent and Selective Mimics of IL-2 and IL-15. *Nature* **2019**, *565*, 186âĂŞ191.

(7) Potapov, V.; Cohen, M.; Schreiber, G. Assessing Computational Methods for Predicting Protein Stability Upon Mutation: Good on Average but Not in the Details. *Protein Eng., Des. Sel.* **2009**, *22*, 553–560.

(8) Guerois, R.; Nielsen, J. E.; Serrano, L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *J. Mol. Biol.* **2002**, *320*, 369–387.

(9) Getov, I.; Petukh, M.; Alexov, E. SAAFEC: Predicting the Effect of Single Point Mutations on Protein Folding Free Energy Using a Knowledge-Modified MM/PBSA Approach. *Int. J. Mol. Sci.* **2016**, *17*, 512.

(10) Park, H.; Bradley, P.; Greisen, P., Jr.; Liu, Y.; Mulligan, V. K.; Kim, D. E.; Baker, D.; DiMaio, F. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **2016**, *12*, 6201–6212.

(11) Barlow, K. A.; Conchúir, S. O.; Thompson, S.; Suresh, P.; Lucas, J. E.; Heinonen, M.; Kortemme, T. Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. *J. Phys. Chem. B* **2018**, *122*, 5389–5399.

(12) Ding, X.; Zou, Z.; Brooks, C. L., III Deciphering Protein Evolution and Fitness Landscapes with Latent Space Models. *Nat. Commun.* **2019**, *10*, 5644.

(13) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J. et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

(14) Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911–2937.

(15) Rao, S. N.; Singh, U. C.; Bash, P. A.; Kollman, P. A. Free Energy Perturbation Calculations on Binding and Catalysis after Mutating Asn 155 in Subtilisin. *Nature* **1987**, *328*, 551–554.

(16) Pitera, J. W.; Kollman, P. A. Exhaustive Mutagenesis in Silico: Multicoordinate Free Energy Calculations on Proteins and Peptides. *Proteins: Struct., Funct., Bioinf.* **2000**, *41*, 385–397.

(17) Seeliger, D.; de Groot, B. L. Protein Thermostability Calculations Using Alchemical Free Energy Simulations. *Biophys. J.* **2010**, *98*, 2309–2316.

(18) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan. *Angew. Chem.* **2016**, *55*, 7364–7368.

(19) Steinbrecher, T.; Zhu, C.; Wang, L.; Abel, R.; Negron, C.; Pearlman, D.; Feyfant, E.; Duan, J.; Sherman, W. Predicting the Effect of Amino Acid Single-Point Mutations on Protein Stability: Large-Scale Validation of MD-Based Relative Free Energy Calculations. *J. Mol. Biol.* **2017**, *429*, 948–963.

(20) Clark, A. J.; Gindin, T.; Zhang, B.; Wang, L.; Abel, R.; Murret, C. S.; Xu, F.; Bao, A.; Lu, N. J.; Zhou, T. et al. Free Energy Perturbation Calculation of Relative Binding Free Energy between Broadly Neutralizing Antibodies and the gp120 Glycoprotein of HIV-1. *J. Mol. Biol.* **2017**, *429*, 930–947.

(21) Jespers, W.; Isaksen, G. V.; Andberg, T. A.; Vasile, S.; van Veen, A.; Åqvist, J.; Brandsdal, B. O.; Gutiérrez-de-Terán, H. QresFEP: An Automated Protocol for Free Energy Calculations of Protein Mutations in Q. *J. Chem. Theory Comput.* **2019**, *15*, 5461–5473.

(22) Hayes, R. L.; Vilseck, J. Z.; Brooks, C. L., III Approaching Protein Design with Multisite λ Dynamics: Accurate and Scalable Mutational Folding Free Energies in T4 Lysozyme. *Protein Sci.* **2018**, *27*, 1910–1922.

(23) Hart, K. M.; Harms, M. J.; Schmidt, B. H.; Elya, C.; Thornton, J. W.; Marqusee, S. Thermodynamic System Drift in Protein Evolution. *PLoS Biol.* **2014**, *12*, e1001994.

(24) Dantas, G.; Kuhlman, B.; Callender, D.; Wong, M.; Baker, D. A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. *J. Mol. Biol.* **2003**, *332*, 449–460.

(25) Dantas, G.; Corrent, C.; Reichow, S. L.; Havranek, J. J.; Eletr, Z. M.; Isern, N. G.; Kuhlman, B.; Varani, G.; Merritt, E. A.; Baker, D. High-resolution Structural and Thermodynamic Analysis of Extreme Stabilization of Human Procarboxypeptidase by Computational Protein Design. *J. Mol. Biol.* **2007**, *366*, 1209–1221.

(26) Petrov, D.; Daura, X.; Zagrovic, B. Effect of Oxidative Damage on the Stability and Dimerization of Superoxide Dismutase 1. *Biophys. J.* **2016**, *110*, 1499–1509.

(27) Yee, A. W.; Aldeghi, M.; Blakeley, M. P.; Ostermann, A.; Mas, P. J.; Moulin, M.; de Sanctis, D.; Bowler, M. W.; Mueller-Dieckmann, C.; Mitchell, E. P. et al. A Molecular Mechanism for Transthyretin Amyloidogenesis. *Nat. Commun.* **2019**, *10*, 925.

(28) Straatsma, T. P.; Berendsen, H. J. C. Free Energy of Ionic Hydration: Analysis of a Thermodynamic Integration Technique to Evaluate Free Energy Differences by Molecular Dynamics Simulations. *J. Chem. Phys.* **1988**, *89*, 5876–5886.

(29) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420–1426.

(30) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.

(31) Goette, M.; Grubmüller, H. Accuracy and Convergence of Free Energy Differences Calculated from Nonequilibrium Switching Processes. *J. Comput. Chem.* **2009**, *30*, 447–456.

(32) Kong, X.; Brooks, C. L., III λ-Dynamics: A New Approach to Free Energy Calculations. *J. Chem. Phys.* **1996**, *105*, 2414–2423.

(33) Knight, J. L.; Brooks, C. L., III Multisite λ Dynamics for Simulated Structure-Activity Relationship Studies. *J. Chem. Theory Comput.* **2011**, *7*, 2728–2739.

(34) Knight, J. L.; Brooks, C. L., III Applying Efficient Implicit Nongeometric Constraints in Alchemical Free Energy Simulations. *J. Comput. Chem.* **2011**, *32*, 3423–3432.

(35) Hayes, R. L.; Armacost, K. A.; Vilseck, J. Z.; Brooks, C. L., III Adaptive Landscape Flattening Accelerates Sampling of Alchemical Space in Multisite λ Dynamics. *J. Phys. Chem. B* **2017**, *121*, 3626–3635.

(36) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., III Improved Treatment of the Protein Backbone in Empirical Force Fields. *J. Am. Chem. Soc.* **2004**, *126*, 698–699.

(37) Mackerell, A. D., Jr.; Feig, M.; Brooks, C. L., III Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of GasâĂŘPhase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(38) Liu, S.; Wang, L.; Mobley, D. L. Is Ring Breaking Feasible in Relative Binding Free Energy Calculations? *J. Chem. Inf. Model.* **2015**, *55*, 727–735.

(39) Best, R. B.; Mittal, J.; Feig, M.; MacKerell, A. D., Jr. Inclusion of Many-Body Effects in the Additive CHARMM Protein CMAP Potential Results in Enhanced Cooperativity of α-Helix and β-Hairpin Formation. *Biophys. J.* **2012**, *103*, 1045–1051.

(40) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Jr. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

(41) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

(42) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S. et al. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

(43) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; ao V. Ribeiro, J.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W. et al. Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *J. Chem. Phys.* **2020**, *153*, 044130.

(44) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1-2*, 19–25.

(45) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. pmx: Automated Protein Structure and Topology Generation for Alchemical Perturbations. *J. Comput. Chem.* **2015**, *36*, 348–354.

(46) Jiang, W.; Chipot, C.; Roux, B. Computing Relative Binding Affinity of Ligands to Receptor: An Effective Hybrid Single-Dual-Topology Free-Energy Perturbation Approach in NAMD. *J. Chem. Inf. Model.* **2019**, *59*, 3794–3802.

(47) Wang, L.; Deng, Y.; Wu, Y.; Kim, B.; LeBard, D. N.; Wandschneider, D.; Beachy, M.; Friesner, R. A.; Abel, R. Accurate Modeling of Scaffold Hopping Transformations in Drug Discovery. *J. Chem. Theory Comput.* **2017**, *13*, 42–54.

(48) Yu, H. S.; Deng, Y.; Wu, Y.; Sindhikara, D.; Rask, A. R.; Kimura, T.; Abel, R.; Wang, L. Accurate and Reliable Prediction of the Binding Affinities of Macrocycles to Their Protein Targets. *J. Chem. Theory Comput.* **2017**, *13*, 6290–6300.

(49) Vilseck, J. Z.; Sohail, N.; Hayes, R. L.; Brooks, C. L., III Overcoming Challenging Substituent Perturbations with Multisite λ-Dynamics: A Case Study Targeting β-Secretase 1. *J. Phys. Chem. Lett.* **2019**, *10*, 4875–4880.

(50) Huang, Y.; Chen, W.; Wallace, J. A.; Shen, J. All-Atom Continuous Constant pH Molecular Dynamics with Particle Mesh Ewald and Titratable Water. *J. Chem. Theory Comput.* **2016**, *12*, 5411–5421.

(51) Nicholson, H.; Anderson, D. E.; Pin, S. D.; Matthews, B. W. Analysis of the Interaction Between Charged Side Chains and the α-Helix Dipole Using Designed Thermostable Mutants of Phage T4 Lysozyme. *Biochemistry* **1991**, *30*, 9816–9828.

(52) Baase, W. A.; Liu, L.; Tronrud, D. E.; Matthews, B. W. Lessons from the

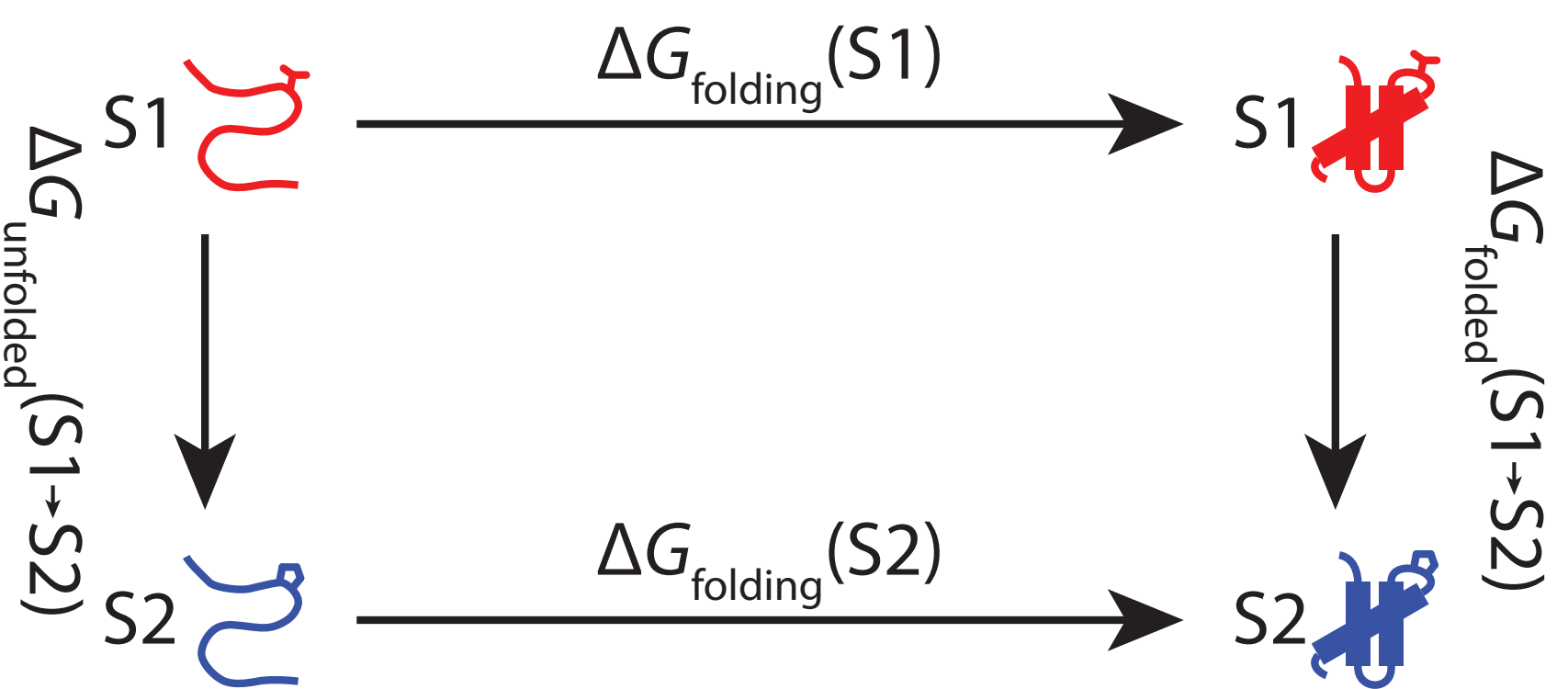
Lysozyme of Phage T4. *Protein Sci.* **2010**, *19*, 631–641.

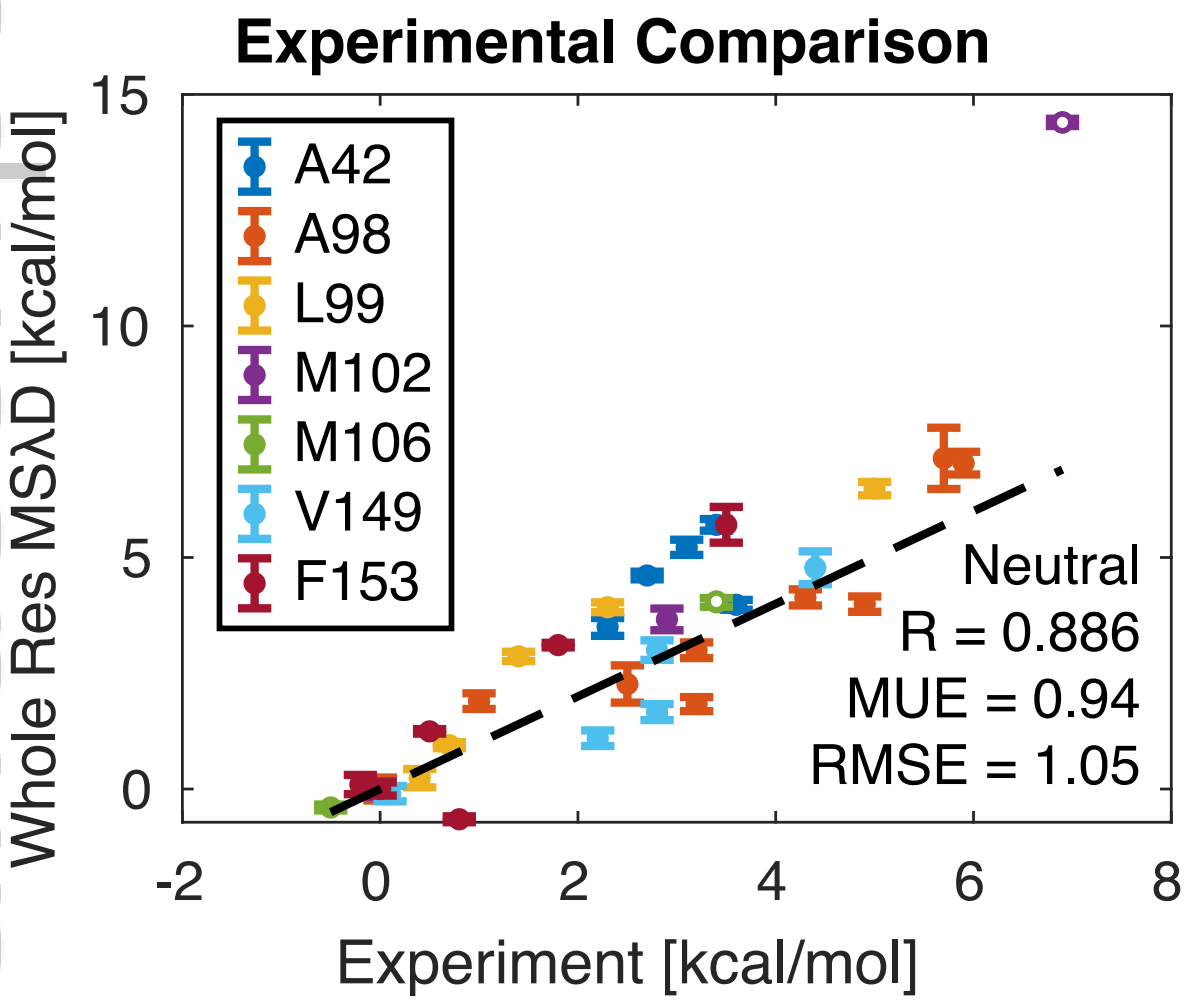
(53) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK_a Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.

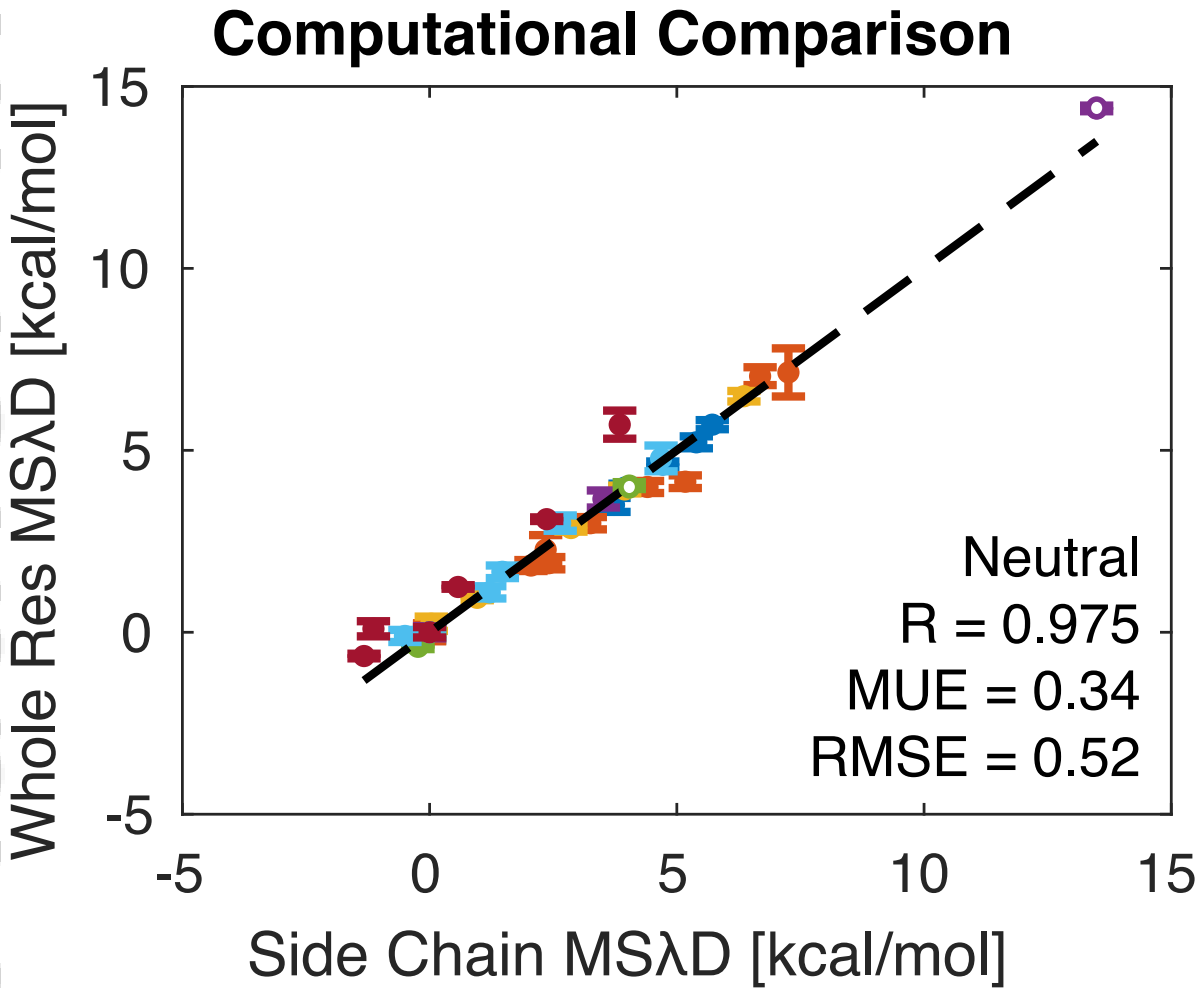
(54) Ding, X.; Hayes, R. L.; Vilseck, J. Z.; Charles, M. K.; Brooks, C. L., III CDOCKER and λ-Dynamics for Prospective Prediction in D3R Grand Challenge 2. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 89–102.

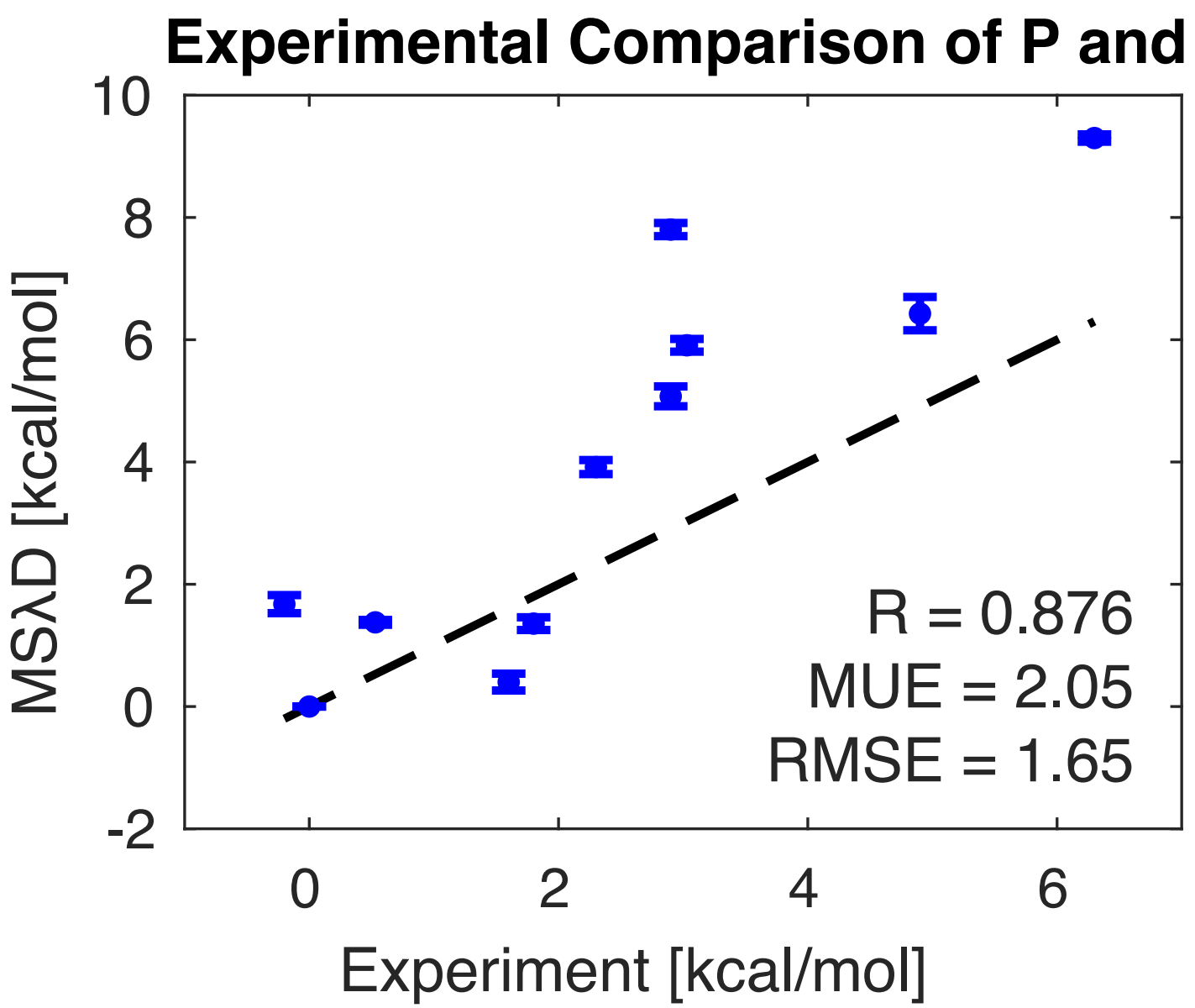
(55) Parks, C. D.; Gaieb, Z.; Chiu, M.; Yang, H.; Shao, C.; Walters, W. P.; Jansen, J. M.; McGaughey, G.; Lewis, R. A.; Bembenek, S. D. et al. D3R Grand Challenge 4: Blind Prediction of Protein-Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 99–119.

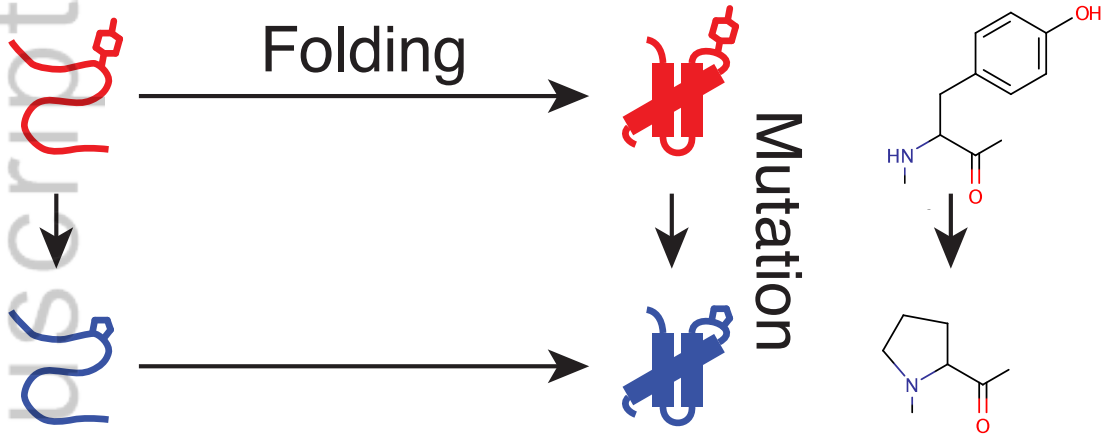
(56) Duan, J.; Lupyan, D.; Wang, L. Improving the Accuracy of Protein Thermostability Predictions for Single Point Mutations. *Biophys. J.* **2020**, *119*, 115–127.



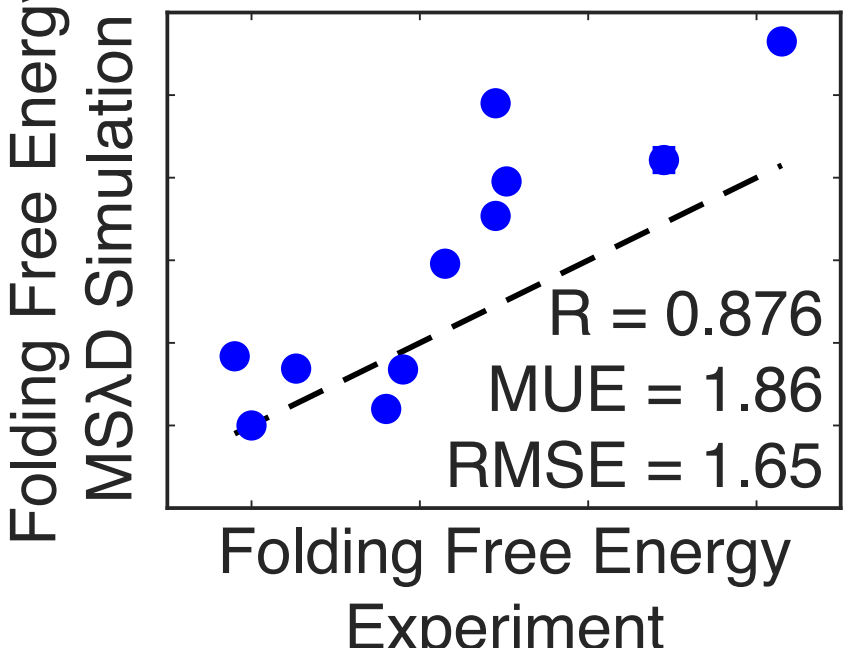


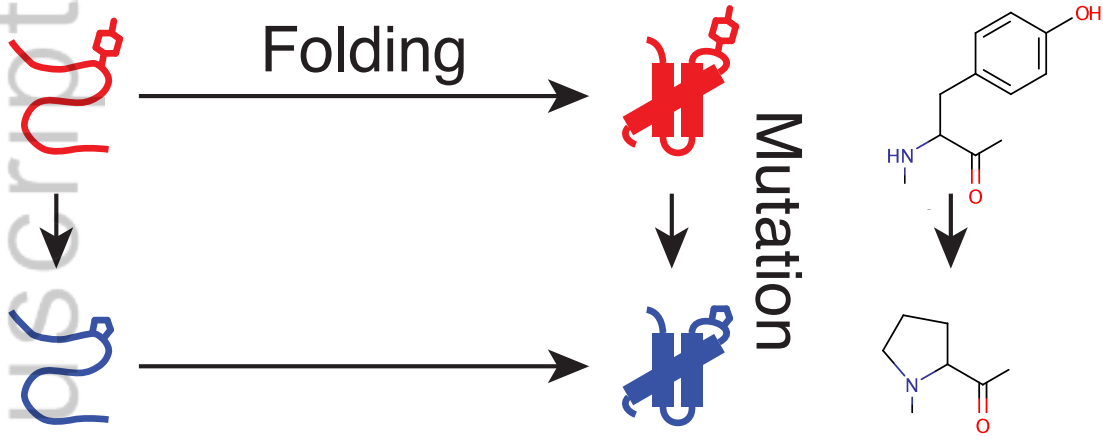






Proline or Glycine Mutations





Proline or Glycine Mutations