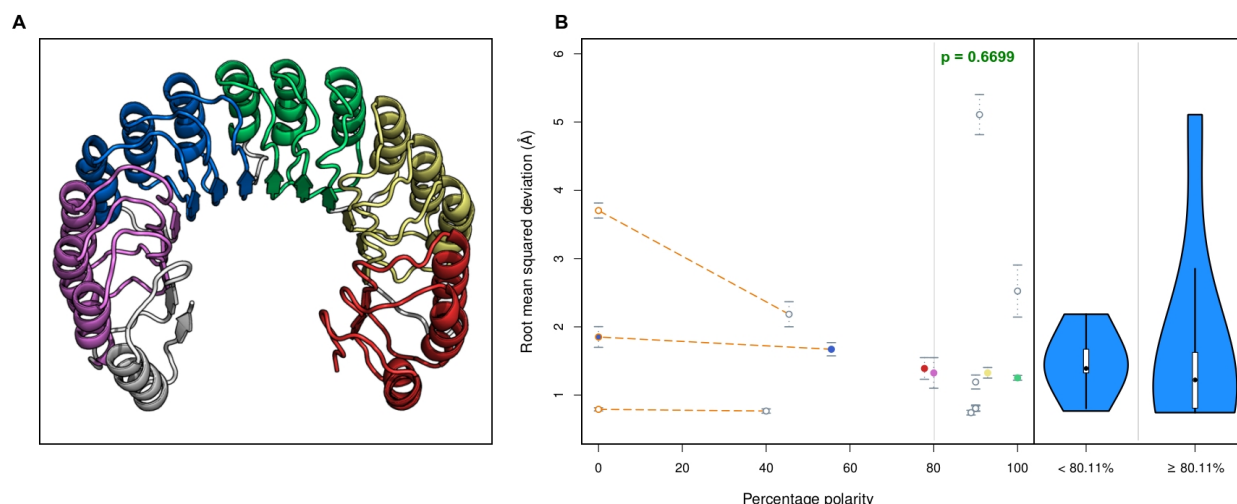


# 1 Text S3. Modeling of exposed hydrophobic patches using Rosetta.



**Figure(Text S3). Ab initio modeling of proteins adopting the horseshoe fold.** (A) The structure of porcine ribonuclease inhibitor (PDB ID: 2bnh, chain A). The ( $\alpha\beta$ )<sub>3</sub> modules used for ab initio modeling are coloured. (B) Ab initio modeling of select proteins adopting the horseshoe fold. Right: Scatterplots describe the ab initio models mean RMSD distribution (with error bars). Coloured whole dots represent ab initio modeling simulations performed on porcine ribonuclease inhibitor (PDB ID: 2bnh, chain A) ( $\alpha\beta$ )<sub>3</sub> modules. Hollow, grey dots represent simulations performed on other proteins adopting the horseshoe fold. Hollow, orange dots represent simulations performed of isosteric mutants. Dashed, orange lines connect isosteric mutants to their parent modules. A grey line divides the data into relatively apolar pores (<80.11% polarity) and relatively polar pores (≥80.11% polarity). The P value (Welch two sample t-test) is depicted in green. Left: Violin plots depict the data distribution across apolar and polar pores.

2 The Rosetta energy function strongly emphasizes hydrophobic collapse. It is possible that the  
 3 energy function is biased against exposed hydrophobic patches that naturally form during the  
 4 process of protein folding. In order to eliminate this possibility, we performed Rosetta ab initio  
 5 modeling on ( $\alpha\beta$ )<sub>n</sub> proteins with naturally exposed hydrophobic patches. We chose the horseshoe  
 6 fold, which contains solvent-exposed pore beta sheets arranged as a crescent. The buried  $\alpha\beta$   
 7 interface is characteristically leucine-rich. Exposed residues on pore beta sheets vary in  
 8 hydrophilicity. It is not uncommon to observe hydrophobic patches interspersed among the usually  
 9 hydrophilic surface. The sequences of 5 ( $\alpha\beta$ )<sub>3</sub> modules of varying pore polarity (55.56% to 100%)  
 10 were extracted from porcine ribonuclease inhibitor (PDB ID: 2bnh, chain A), a rather large protein  
 11 adopting the horseshoe fold (Figure (Text S3)A). These sequences were used for Rosetta ab initio  
 12 modeling and mean C $\alpha$  RMSD calculation as previously described for KLGP decarboxylase  
 13 structural homologues. We observed that all 5 sequences were well-modeled, with a mean RMSD  
 14 not exceeding 1.67Å. Additional ab initio modeling on ( $\alpha\beta$ )<sub>3</sub> modules extracted from other proteins  
 15 adopting the horseshoe fold (PDB ID: 1ogq, chain A; 2ast, chain B; 2omz, chain A) revealed no  
 16 decrease in foldability with increasing pore polarity. Considering all ( $\alpha\beta$ )<sub>3</sub> modules, most modules  
 17 appeared well folded (>90% possessing a mean RMSD of <3Å), and no statistically significant

18 difference was observed in the foldability between relatively hydrophilic and hydrophobic modules  
19 ( $p=0.6699$ ) (Figure (Text S3)B).

20

21 We further tested the ability of Rosetta to correctly fold exposed hydrophobic patches by  
22 performing tests on 3 naturally hydrophobic  $(\alpha\beta)_3$  modules (PDB ID: 2bnh, chain A, module 4,  
23 1ogq, chain A, module1, 2omz, chain A, module 1). Isosteric mutations were performed on the  
24 polar pore residues of these 3 modules to decrease their percentage pore polarities to 0%. Ab initio  
25 modeling was performed on the resulting sequences. In 2 of 3 cases, Rosetta was able to  
26 successfully model  $(\alpha\beta)_3$  modules (Figure (Text S3)B), despite the complete absence of polar pore  
27 residues. These results indicate that Rosetta ab initio simulations are not biased against exposed  
28 hydrophobic patches. Therefore, if  $(\alpha\beta)_3$  modules derived from KLGP decarboxylase structural  
29 homologues with hydrophobic pores are stable in solution, then there is reason to believe that  
30 Rosetta is capable of accurately modeling their structures. Structures for all horseshoe modules  
31 described are provided in the supporting information (Dataset S2).