# Robust deep learning-based protein sequence design using ProteinMPNN

J. Dauparas[1,2], I. Anishchenko[1,2], N. Bennett[1,2,3], H. Bai[1,2,4], R. J. Ragotte[1,2],

L. F. Milles[1,2], B. I. M. Wicky[1,2], A. Courbet[1,3,4], R. J. de Haas[5], N. Bethel[1,3,4], P. J. Y. Leung[1,2,3], T. F. Huddy[1,2], S. Pellock[1,2], D. Tischer[1,3], F. Chan[1,2], B. Koepnick[1,2], H. Nguyen[1,2], A. Kang[1,2], B. Sankaran[6], A. K. Bera[1,2], N. P. King[1,2], D. Baker[1,3,4*]

*Corresponding author. Email: dabaker@uw.edu

Abstract

While deep learning has revolutionized protein structure prediction, almost all experimentally characterized de novo protein designs have been generated using physically based approaches such as Rosetta. Here we describe a deep learning-based protein sequence design method, ProteinMPNN, with outstanding performance in both in silico and experimental tests. On native protein backbones, ProteinMPNN has a sequence recovery of 52.4%, compared to 32.9% for Rosetta. The amino acid sequence at different positions can be coupled between single or multiple chains, enabling application to a wide range of current protein design challenges. We demonstrate the broad utility and high accuracy of ProteinMPNN using X-ray crystallography, cryoEM and functional studies by rescuing previously failed designs, made using Rosetta or AlphaFold, of protein monomers, cyclic homo-oligomers, tetrahedral nanoparticles, and target binding proteins.

尽管深度学习彻底革新了蛋白质结构预测领域，但几乎所有已被实验验证的全新蛋白质设计仍是通过基于物理的方法（如 Rosetta）生成的。在此我们提出一种基于深度学习的蛋白质序列设计方法——ProteinMPNN，在计算模拟和实验测试中均表现出色。对于天然蛋白质骨架，ProteinMPNN 的序列恢复率达到 52.4%，而 Rosetta 为 32.9%。不同位置的氨基酸序列可在单链或多链之间耦合，从而适用于多种当前的蛋白质设计挑战。我们通过 X 射线晶体学、冷冻电镜和功能研究验证了 ProteinMPNN 的广泛适用性与高精度，成功挽救了使用 Rosetta 或 AlphaFold 设计失败的蛋白质单体、环状同源寡聚体、四面体纳米颗粒和靶向结合蛋白的设计。

## Introduction 引言

The protein sequence design problem is to find, given a protein backbone structure of interest, an amino acid sequence that will fold to this structure. Physically based approaches like Rosetta treat sequence design as an energy optimization problem, searching for the combination of amino acid identities and conformations that have the lowest energy for a given input structure. Recently deep learning approaches have shown promise in rapidly generating candidate amino acid sequences given monomeric protein backbones without need for compute intensive explicit consideration of sidechain rotameric states (1-7). However, the methods described thus far do not apply to the full range of current protein design challenges, and have not been extensively validated experimentally.

蛋白质序列设计问题旨在：在给定一个感兴趣的蛋白质骨架结构的情况下，找到一个可折叠为该结构的氨基酸序列。基于物理的方法，如 Rosetta，将序列设计视为一个能量优化问题，寻找在给定结构下氨基酸种类与构象组合的最低能量状态。近年来，深度学习方法在无需对侧链旋转异构状态进行密集计算的前提下，能快速生成候选氨基酸序列，显示出良好前景（1-7）。然而，目前所描述的方法尚无法适用于当前全部的蛋白质设计挑战，并且尚未经过广泛的实验验证。

We sought to develop a deep learning-based protein sequence design method broadly applicable to design of monomers, cyclic oligomers, protein nanoparticles, and protein-protein interfaces. We began from a previously described message passing neural network (MPNN) with 3 encoder and 3 decoder layers and 128 hidden dimensions which predicts protein sequences in an autoregressive manner from N to C terminus using protein backbone features - distances between $C\alpha-C\alpha$ atoms, relative $C\alpha-C\alpha-C\alpha$ frame orientations and rotations, and backbone dihedral angles-as input (1). We first sought to improve performance of the model on recovering the amino acid sequences of native single-chain proteins given their backbone structures. A set of 19,700 high resolution single-chain structures from the PDB were split into train, validation and test sets (80/10/10) based on the CATH (8) protein classification (see methods). We found that including distances between $N, C\alpha, C, O$ and a virtual $C\beta$ placed based on the other backbone atoms as additional input features resulted in a sequence recovery increase from 41.2% (baseline model) to 49.0% (experiment 1), see Table 1; interatomic distances evidently provide a better inductive bias to capture interactions between residues than dihedral angles or $N-C\alpha-C$ frame orientations.

我们旨在开发一种基于深度学习的蛋白质序列设计方法，可广泛用于单体、环状寡聚体、蛋白质纳米颗粒以及蛋白质-蛋白质界面的设计。我们以先前描述的消息传递神经网络（MPNN）为基础，网络包含 3 个编码器层和 3 个

解码器层，具有 128 个隐藏维度，采用自回归方式从 N 端到 C 端预测蛋白质序列，输入为蛋白质骨架特征——$C\alpha - C\alpha$ 原子之间的距离、相对的 $C\alpha - C\alpha - C\alpha$ 构象朝向与旋转角度、以及主链的二面角（1）。我们首先致力于提升模型在恢复天然单链蛋白氨基酸序列方面的性能。我们从 PDB 中选取了 19,700 个高分辨率单链结构，依据 CATH 蛋白分类标准（见方法）划分为训练、验证和测试集（比例为 80/10/10）。结果显示，将 $N, C\alpha, C, O$ 及虚拟 $C\beta$ 原子间的距离作为附加输入特征后，序列恢复率从 41.2%（基线模型）提升至 49.0%（实验 1），见表 1；原子间距离显然比二面角或 $N - C\alpha - C$ 构架方向更有利于捕捉残基之间的相互作用。

We next introduced edge updates in addition to the node updates in the backbone encoder neural network (experiment 2). Combining additional input features and edge updates leads to a sequence recovery of 50.5% (experiment 3). To determine the range over which backbone geometry influences amino acid identity, we tested $16, 24, 32, 48$, and 64 nearest $C\alpha$ neighbor neural networks (fig. S1A), and found that performance was saturated at 32-48 neighbors. Unlike the protein structure prediction problem, locally connected graph neural networks can accurately model the structure to sequence mapping problem because the optimality of an amino acid at a particular position is largely determined by the immediate protein environment.
随后我们在骨架编码神经网络中引入了边更新机制，除了原有的节点更新外（实验 2）。结合额外输入特征与边更新后，序列恢复率提升至 50.5%（实验 3）。为了确定骨架几何对氨基酸身份影响的范围，我们测试了 $16, 24, 32, 48$ 和 64 个最近 $C\alpha$ 邻居的神经网络（见图 S1A），结果表明模型性能在 32 至 48 个邻居时趋于饱和。与蛋白质结构预测问题不同，局部连接的图神经网络可准确建模结构到序列的映射问题，因为某个位置上氨基酸的最优性主要由其局部蛋白质环境决定。

To enable application to a broad range of single and multi-chain design problems, we replaced the fixed N to C terminal decoding order with an order agnostic autoregressive model in which the decoding order is randomly sampled from the set of all possible permutations (9). This also resulted in a modest improvement in sequence recovery (Table 1, experiment 4). Order agnostic decoding enables design in cases where, for example, the middle of the protein sequence is fixed and the rest needs to be designed, as in protein binder design where the target sequence is known; decoding skips the fixed regions but includes them in the sequence context for the remaining positions (Fig. 1B). For multi-chain design problems (see below), to make the model equivariant to the order of the protein chains, we kept the per chain relative positional encoding capped at $\pm 32$ residues (10) and added a binary feature indicating if the interacting pair of residues are from the same or different chains.
为使该方法可广泛应用于单链和多链的蛋白质设计问题，我们将固定的从 N 端到 C 端的解码顺序替换为一种顺序无关的自回归模型，其解码顺序是从所有可能的排列中随机采样而得（9）。这种改进也带来了序列恢复率的适度提升（见表 1，实验 4）。顺序无关的解码使得在某些情形下的设计成为可能，例如当蛋白质序列中间部分被固定、其余部分需要设计时，如在靶向蛋白设计中目标序列已知；在这种情况下，解码过程会跳过固定区域，但将其纳入上下文中，以指导剩余位置的预测（见图 1B）。针对多链设计问题（见下文），为使模型对蛋白链的排列顺序保持等变性，我们将每条链的相对位置编码限制在 $\pm 32$ 个残基内（10），并新增一个二元特征，用于指示相互作用残基是否来自同一条链。

We used the flexible decoding order to fix residue identities in sets of corresponding positions (the residues at these positions are decoded at the same time). For example, for a homodimer backbone with two chains A and B with sequence $A_1$, $A_2, \ldots$, and $B_1$, $B_2, \ldots$, the amino acids for chains A and B have to be the same for corresponding indices; we implement this by predicting unnormalized probabilities for $A_1$ and $B_1$ first and then combine these two predictions to construct a normalized probability distribution from which a joint amino acid is sampled (Fig. 1C). For pseudosymmetric sequence design, residues within, or between chains can be similarly constrained; for example for repeat protein design, the sequence in each repeat unit can be kept fixed. Multi-state design of single sequences that encodes two or more desired states can be achieved by predicting unnormalized probabilities for each state and then averaging; more generally a linear combination of predicted unnormalized probabilities with some positive and negative coefficients can be used to upweight, or downweight specific backbone states to achieve explicit positive or negative sequence design. The architecture of this multichain and symmetry aware (positionally coupled) model, which we call ProteinMPNN, is outlined schematically in Fig. 1A. We trained ProteinMPNN on protein assemblies in the PDB (as of Aug 02, 2021) determined by X-ray crystallography or cryoEM to better than 3.5 resolution and with less than 10,000 residues (see methods).
我们利用灵活的解码顺序来固定对应位置上残基的身份（这些位置上的残基将同时进行解码）。例如，在具有两条链 A 和 B 的同源二聚体骨架中，其序列为 $A_1$, $A_2, \ldots$ 和 $B_1$, $B_2, \ldots$，则链 A 和链 B 在相应索引位置上的氨基酸应相同；我们通过先分别预测 $A_1$ 和 $B_1$ 的非归一化概率分布，再将这两个预测合并构建一个归一化的概率分布，并从中采样一个联合氨基酸来实现这一点（见图 1C）。对于拟对称序列设计，链内或链间的残基也可以类似地进行约束；例如，在重复蛋白设计中，每个重复单元中的序列可以保持不变。对于需要编码两个或多个目标构象的单序列的多状态设计，可通过分别预测每种状态的非归一化概率再进行平均来实现；更一般地，可使用一组带有正负权重的线性组合对预测的非归一化概率进行加权，以强化或削弱特定骨架构象，实现显式的正向或负向序列设计。我们称这一多链与对称性感知（位置耦合）模型为 ProteinMPNN，其架构如图 1A 所示。我们在 PDB 中通过 X 射线晶体学或冷冻电镜解析、分辨率优于 3.5 且残基数少于 10,000 的蛋白质装配体数据（截至 2021 年 8 月 2 日）上对 ProteinMPNN 进行了训练（见方法）。

For a test set of 402 monomer backbones we redesigned sequences using Rosetta fixed backbone combinatorial sequence design [one round of the PackRotamersMover (11, 12) with default options and the beta_nov16 score function] and ProteinMPNN. Although requiring only a small fraction of the compute time ( 1.2 s versus 258.8 s on a single CPU for 100 residues), ProteinMPNN had a much higher overall native sequence recovery ( 52.4%

vs 32.9% ), with improvements across the full range of residue burial from protein core to surface (Fig. 2A). Differences between designed and native amino acid biases for the core, boundary and surface regions for the two methods are shown in fig. S2.
我们对一个包含 402 个蛋白质单体骨架的测试集进行了序列再设计，分别使用 Rosetta 的固定骨架组合序列设计方法（使用默认参数和 beta_nov16 评分函数的 PackRotamersMover 执行一次循环）(11, 12) 与 ProteinMPNN。尽管计算时间大大缩短（对于 100 个残基，单核 CPU 上为 1.2 秒对比 258.8 秒），ProteinMPNN 的整体天然序列恢复率显著更高（52.4% 对比 32.9%），在从蛋白质核心到表面整个残基埋藏范围内均表现出提升（见图 2A）。两种方法在核心、边界和表面区域设计氨基酸与天然氨基酸偏差的差异见图 S2。

We further evaluated ProteinMPNN on a test set of 690 monomers, 732 homomers (with less than 2000 residues), and 98 heteromers. The median sequence recoveries over all residues were 52% for monomers, 55% for homomers, and 51% for heteromers and over interface residues, 53% for homomers and 51% for heteromers (Fig. 2B). In all three cases, sequence recovery correlated closely with residue burial ranging from $90-95\%$ in the deep core to 35% on the surface (fig. S1B): the amount of local geometric context determines how well residues can be recovered at specific positions. For homomers, we found best results with averaging unnormalized probabilities (rather than normalized probabilities) between symmetry related positions (fig. S1C); because of the non-local context sequence recovery is no longer a monotonic function of the average C$\beta$ neighbor distance (fig. S1B).
我们进一步在一个测试集中评估了 ProteinMPNN，该测试集包含 690 个单体、732 个同源复合体（残基数少于 2000）和 98 个异源复合体。所有残基的中位数序列恢复率分别为：单体 52%，同源复合体 55%，异源复合体 51%；对于界面残基，同源体为 53%，异源体为 51%（见图 2B）。在这三种情况下，序列恢复率与残基埋藏程度密切相关：从深埋核心的 90 – 95% 到表面的 35%（见图 S1B）；局部几何上下文的丰富程度决定了在特定位置上残基的可恢复性。对于同源复合体，我们发现，在对对称相关位置的非归一化概率进行平均（而非归一化后）时结果最佳（见图 S1C）；由于非局部上下文的存在，序列恢复率不再是平均 C$\beta$ 邻居距离的单调函数（见图 S1B）。

## Training with Backbone Noise **带有骨架噪声的训练**

While protein sequence design approaches have often focused on maximizing sequence recovery for protein backbones from high resolution crystal structures, this is not necessarily optimal for actual protein design applications. We found that training models on backbones to which Gaussian noise (std=0.02 ) had been added improved sequence recovery on confident protein structure models generated by AlphaFold (average pLDDT>80.0) from UniRef50, while the sequence recovery on unperturbed PDB structures significantly decreased (Table 1); crystallographic refinement may impart some memory of amino acid identity in the backbone coordinates which is captured by models trained on crystal structure backbones and reduced by the addition of noise. Robustness to small displacements in atomic coordinates is a desirable feature in real world applications where the protein backbone geometry is not known at atomic resolution.
尽管蛋白质序列设计方法常常致力于最大化高分辨率晶体结构蛋白骨架的序列恢复率，但这并不一定最适用于实际的蛋白质设计任务。我们发现，将高斯噪声（标准差为 0.02 ）添加至蛋白骨架后进行模型训练，可提升在 AlphaFold（平均 pLDDT > 80.0）使用 UniRef50 数据库生成的高置信度蛋白质结构模型上的序列恢复率；而在未扰动的 PDB 结构上则显著下降（见表 1）。晶体结构的精修过程可能会在骨架坐标中引入氨基酸身份的 “记忆”，这种信息可被在晶体骨架结构上训练的模型捕捉到，而通过添加噪声可减弱这种记忆。在实际应用中，骨架几何通常并非原子级分辨率，因此对于原子坐标的小幅位移具备鲁棒性是一项非常理想的特性。

AlphaFold (10) and RoseTTAfold (13) produce remarkably good structure predictions for native proteins given multiple sequence alignments which can contain substantial co-evolutionary and other information reflecting aspects of the 3D structure, but generally produce much poorer structures when provided only with a single sequence. We reasoned that ProteinMPNN might generate single sequences for native backbones more strongly encoding the structures than the original native sequence, as evolution in most cases does not optimize for stability. Indeed, we found that ProteinMPNN sequences generated for native backbones were predicted to fold to these structures much more confidently and accurately by AlphaFold than the original native sequences (Fig. 2E). ProteinMPNN also strengthened the sequence to structure mapping for designed backbones: over a set of de novo designed ligand binding pocket containing scaffolds generated using Rosetta, only 2.7% of the original designed sequences were predicted to fold to the design target structures, but following ProteinMPNN redesign 54.1% were confidently predicted to fold to close to the target structures (Fig. 2F). This should substantially increase the utility of these scaffolds for design of small molecule binding and enzymatic functions.
AlphaFold (10) 和 RoseTTAfold (13) 在提供多个序列比对时（其中包含大量共进化和其他能反映三维结构特征的信息）能对天然蛋白质生成极为出色的结构预测；但在仅提供单一序列时通常表现较差。我们推测，ProteinMPNN 能为天然骨架生成比原始天然序列更强结构编码性的单一序列，因为在大多数情况下，进化并未优化序列的稳定性。事实上，我们发现，ProteinMPNN 为天然骨架生成的序列比原始天然序列更能被 AlphaFold 自信且准确地预测为目标结构（见图 2E）。此外，ProteinMPNN 也增强了设计骨架上序列与结构的映射关系：在一组通过 Rosetta 新设计的、含有配体结合口袋的支架结构中，原始设计的序列中仅有 2.7% 被预测能折叠为目标结构；而经过 ProteinMPNN 重新设计后，有 54.1% 的序列被高置信度预测能接近目标结构折叠（见图 2F）。这将极大提升这些支架在小分子结合与酶功能设计中的实用性。

We found further that the strength of the single sequence to structure mapping, as assessed by AlphaFold,

was higher for models trained with additional backbone noise. As noted above, the average sequence recovery for crystallographically refined backbones decreases with increasing amounts of noise added during training (Fig. 2C) as these models blur out local details of the backbone geometry. However, sequences generated by noised ProteinMPNN models are more robustly decoded into 3D coordinates by AlphaFold, likely because noised models focus more on overall topological features, as encoded by for example the overall polar-nonpolar sequence pattern, than local structural details. For example, a model trained with 0.3 noise generated 2-3 times more sequences with AlphaFold predictions within lDDT-C $\alpha$ (14) of 95.0 and 90.0 of the true structures than unnoised or slightly noised models (Fig. 2C; training with higher levels of noise increases success rates for less stringent lDDT cutoffs). In protein design calculations, the models trained with larger amounts of noise have the advantage of generating sequences which more strongly map to the target structures by prediction methods (this increases the frequency of designs passing prediction based filters, and may correspondingly also increase the frequency of folding to the desired target structure).
我们进一步发现，在 AlphaFold 的评估下，添加骨架噪声训练的模型在“单序列-结构”映射的强度上表现更优。如上所述，随着训练中添加噪声的增多，晶体精修骨架的平均序列恢复率会下降（见图 2C），因为这些模型模糊了骨架几何的局部细节。然而，噪声模型生成的序列被 AlphaFold 更加鲁棒地解码为三维结构，这可能是因为添加噪声的模型更关注整体拓扑特征（例如极性-非极性的整体序列模式），而非局部结构细节。例如，使用 0.3 噪声训练的模型生成的序列中，有 2 到 3 倍的数量在 AlphaFold 预测中达到与真实结构 lDDT-C $\alpha$ (14) 接近 95.0 和 90.0 的标准，相较于无噪声或低噪声训练模型（见图 2C；更高噪声水平的训练可提升对较宽松 lDDT 截断的成功率）。在蛋白质设计计算中，这些高噪声模型有生成更强结构映射序列的优势（这可提升通过预测过滤器的设计频率，也可能提高最终折叠至目标结构的成功率）。

Because the sequence determinants of protein expression, solubility and function are not perfectly understood, in most protein design applications it is desirable to test multiple designed sequences experimentally. We found that the diversity of sequences generated by MPNN could be considerably increased, with only a very small decrease in average sequence recovery, by carrying out inference at higher temperatures (Fig. 2D). We also found that a measure of sequence quality derived from the ProteinMPNN, the averaged log probability of the sequence given the structure, correlated strongly with native sequence recovery over a range of temperatures (fig. S3A), enabling rapid ranking of sequences for selection for experimental characterization.
由于影响蛋白表达、溶解度和功能的序列因素尚未被完全理解，因此在大多数蛋白质设计应用中，实验测试多个设计序列是较为理想的策略。我们发现，通过在较高的温度下进行推理，可以显著增加 MPNN 所生成序列的多样性，同时平均序列恢复率仅有轻微下降（见图 2D）。我们还发现，从 ProteinMPNN 中获得的一种序列质量衡量指标——给定结构条件下序列的平均对数概率——在多个温度下与天然序列恢复率高度相关（见图 S3A），这使得可以快速对候选序列进行排序，以便选择用于实验验证。

## Experimental Evaluation of ProteinMPNN 实验评估

While in silico native protein sequence recovery is a useful benchmark, the ultimate test of a protein design method is its ability to generate sequences which fold to the desired structure and have the desired function when tested experimentally. We evaluated ProteinMPNN on a representative set of design challenges encompassing protein monomer design, protein nanocage design, and protein function design. In each case, we attempted to rescue previous failed designs with sequences generated using Rosetta or AlphaFold—we kept the backbones of the original designs fixed but discarded the original sequences and generated new ones using ProteinMPNN. Synthetic genes encoding the designs were obtained, and the proteins expressed in E. coli and characterized biochemically and structurally.
尽管通过计算模拟恢复天然蛋白质序列是一个有用的评估基准，但评判蛋白质设计方法的终极标准，是其实验中能否生成可折叠为目标结构且具备预期功能的序列。我们在一组具有代表性的设计挑战上评估了 ProteinMPNN，涵盖了蛋白质单体设计、纳米笼设计以及功能性蛋白质设计。在每种情形中，我们都尝试通过使用 ProteinMPNN 生成新序列，挽救先前使用 Rosetta 或 AlphaFold 失败的设计——保持原设计的骨架不变，舍弃原序列并重新设计。我们合成了编码这些设计的基因，在大肠杆菌中表达蛋白质，并进行生化及结构表征。

We first tested the ability of ProteinMPNN to design amino acid sequences for protein backbones generated by deep network hallucination using AlphaFold (AF). Starting from a random sequence, a Monte Carlo trajectory is carried out optimizing the extent to which AF predicts the sequence to fold to a well-defined structure (15). These calculations generated a wide range of protein sequences and backbones for both monomers and oligomers that differ considerably from those of native structures. In initial tests, the sequences generated by AF were encoded in synthetic genes, and we attempted to express 150 proteins in E. coli. However, the AF generated sequences were mostly insoluble (median soluble yield: 9 mg per liter of culture equivalent Fig. 3A). To determine whether ProteinMPNN could overcome this problem, we generated sequences for a subset of these backbones with ProteinMPNN; residue identities at symmetry-equivalent positions were tied by averaging unnormalized probabilities as described above. The designed sequences were again encoded in synthetic genes and the proteins produced in E. coli. The success rate was far higher: of 96 designs we attempted to express in E. coli, 73 were expressed solubly (median soluble yield: 247 mg per liter of culture equivalent; Fig. 3A) and 50 had the target monomeric or oligomeric state as assessed by SEC (Fig. 3, A and C). Many of the proteins were highly thermo-

stable, with secondary structure being maintained up to 95°C (Fig. 3B).
我们首先测试了 ProteinMPNN 是否能为通过 AlphaFold（AF）深度网络幻觉方法生成的蛋白质骨架设计氨基酸序列。该方法从一个随机序列出发，使用蒙特卡洛轨迹优化 AlphaFold 对其折叠成明确定义结构的预测程度（15）。这类计算生成了与天然结构差异显著的单体与寡聚体蛋白质骨架及序列。在初步实验中，我们将 AlphaFold 生成的序列编码进合成基因，尝试在大肠杆菌中表达 150 种蛋白质。然而，这些序列大多数不溶（中位可溶产率：9 mg/L 培养液，见图 3A）。为验证 ProteinMPNN 是否能克服此问题，我们用 ProteinMPNN 为其中部分骨架生成新序列；对称位置的残基身份通过平均非归一化概率进行绑定（如前所述）。这些新设计序列再次被编码进合成基因，在大肠杆菌中表达。成功率显著提高：在尝试表达的 96 个设计中，有 73 个表达为可溶形式（中位产率为 247 mg/L 培养液；图 3A），其中 50 个通过 SEC 被验证具有目标单体或寡聚体状态（图 3A 和 C）。许多蛋白质表现出极高的热稳定性，其二级结构在高达 95°C 下仍可保持（图 3B）。

We solved the X-ray crystal structure of one of the ProteinMPNN monomer designs with a fold more complex (TMscore = 0.56 against PDB) than most de novo designed proteins (Fig. 3D). The alpha-beta protein structure contains 5 beta strands and 4 alpha helices, and is close to the design target backbone ( 2.35 over 130 residues), demonstrating that ProteinMPNN can accurately encode monomer backbone geometry in amino acid sequences. The accuracy was particularly high in the central core of the structure, with sidechains predicted using AlphaFold from the ProteinMPNN sequence fitting nearly perfectly into the electron density (Fig. 3D). Crystal structures and cryo-EM structures of ten cyclic homo-oligomers with 130-1800 amino acids were also very close to the design target backbones (15). Thus, ProteinMPNN can robustly and accurately design sequences for both monomers and cyclic oligomers.
我们解析了一个由 ProteinMPNN 设计的蛋白质单体的 X 射线晶体结构，该结构的折叠复杂度（与 PDB 比对的 TMscore 为 0.56）超过了大多数全新设计的蛋白质（图 3D）。该 alpha-beta 结构包含 5 条 链和 4 条 螺旋，其骨架与设计目标非常接近（130 个残基上的 RMSD 为 2.35），表明 ProteinMPNN 能够精确地将蛋白质单体骨架几何编码进氨基酸序列中。尤其在结构核心区域，使用 AlphaFold 对 ProteinMPNN 序列预测的侧链几乎完美匹配电子密度图（图 3D）。十个包含 130 至 1800 个氨基酸的环状同源寡聚体的晶体结构和冷冻电镜结构同样与设计骨架非常接近（15）。因此，ProteinMPNN 可稳健且准确地设计单体和环状寡聚体的蛋白质序列。

We next took advantage of the flexible decoding order of ProteinMPNN to design sequences for proteins containing internal repeats, tying the identities of proteins in equivalent positions. We focused on previously suboptimal Rosetta designs of repeat protein structures and found that many could be rescued by ProteinMPNN redesign; an example is shown in Fig. 3, E and F.
随后，我们利用 ProteinMPNN 的灵活解码顺序功能，为含有内部重复单元的蛋白质设计序列，并将等效位置的残基身份进行绑定。我们集中在先前 Rosetta 表现不佳的重复蛋白结构设计上，发现其中许多可以通过 ProteinMPNN 重新设计挽救；其中一个示例见图 3E 和 F。

We next experimented with enforcing both cyclic and internal repeat symmetry by tying positions both within and between subunits, as illustrated in Fig. 3G. We experimentally characterized a set of $C_5/C_6$ cyclic oligomers built with Rosetta based on sequences designed with Rosetta, and a second set with sequences designed using ProteinMPNN. For the Rosetta designed set, 40% (out of total 10) were soluble and none had the correct oligomeric state confirmed by SECMALS. For the ProteinMPNN designed set, 88% (out of total 18) were soluble and 27.7% had the correct oligomeric state. We characterized the structure of one of the designs that was large enough for resolution of structural features by negative stain EM (Fig. 3I), and image averages were closely consistent with the design model (Fig. 3J).
我们进一步在设计中强制实现环状与内部重复对称性，即在亚基内外的相应位置进行绑定，如图 3G 所示。我们对一组 Rosetta 设计的 $C_5/C_6$ 环状寡聚体进行了实验表征（基于 Rosetta 设计的序列），并对另一组基于 ProteinMPNN 设计的序列进行了相同表征。Rosetta 设计组中，有 40%（共 10 个）可溶，但无一个通过 SECMALS 确认具有正确的寡聚体状态；而 ProteinMPNN 设计组中，有 88%（共 18 个）可溶，其中 27.7% 具有正确的寡聚体状态。我们还对其中一个尺寸足够大、可分辨结构特征的设计进行了负染电镜结构表征（图 3I），其图像平均结果与设计模型高度一致（图 3J）。

We next evaluated the ability of ProteinMPNN to design sequences that assemble into target protein nanoparticle assemblies. We started with a set of previously described protein backbones for two-component tetrahedral designs generated using a compute- and effort-intensive procedure that involved Rosetta sequence design followed by more than a week of manual intervention to decrease surface hydrophobicity and improve interface packing (16). We used ProteinMPNN to design 76 sequences spanning 27 of these tetrahedral nanoparticle backbones, tying identities at equivalent positions in the 12 copies of each subunit in the assemblies, and tested these sequences without further intervention. Upon expression in E. coli and purification by SEC, 13 designs formed assemblies with the expected MW ( 1 MDa ) (fig. S4), including several new tetrahedral assemblies that had failed using Rosetta. We solved the crystal structure of one of these, and found that it was very close to the design model ( 1.2 Cu RMSD over two subunits; Fig. 3K). Thus ProteinMPNN can robustly design sequences that assemble into designed nanoparticle structures, which have proven useful for structure-based vaccine design (17-19). Sequence generation with ProteinMPNN is fully automated and requires only about 1 s per backbone, vastly streamlining the design process compared to the earlier Rosetta-based procedure.
我们还评估了 ProteinMPNN 是否能设计出可自组装为目标蛋白质纳米颗粒的序列。我们从一组先前提出的双组分四面体设计的蛋白骨架出发，这些骨架通过 Rosetta 序列设计加上超过一周的人为干预（用于减少表面疏水性和优化界面堆积）生成，计算与人工成本极高（16）。我们使用 ProteinMPNN 为其中 27 个骨架设计了 76 种序列，

对应组装体中每个亚基的 12 个拷贝在等价位置上绑定相同的氨基酸身份，并在无其他人为干预的前提下进行测试。将这些设计在大肠杆菌中表达并经 SEC 纯化后，有 13 种形成了与预期分子量（1 MDa）一致的组装体（见图 S4），其中包括多个使用 Rosetta 失败的四面体组装体。我们解析了其中一个设计的晶体结构，发现其与设计模型非常接近（两个亚基上的 RMSD 为 1.2；图 3K）。因此，ProteinMPNN 可稳健地设计组装成纳米颗粒结构的序列，这类结构已被证明在基于结构的疫苗设计中具有重要应用价值（17-19）。ProteinMPNN 的序列生成过程完全自动化，每个骨架仅需约 1 秒，极大简化了相较于以往 Rosetta 流程的设计流程。

As a final test, we evaluated the ability of ProteinMPNN to rescue previously failed designs of new protein functions using Rosetta. We chose as a challenging example the design of proteins scaffolding polyproline II helix motifs recognized by SH3 domains, where portions of the protein scaffold outside of the core SH3-binding motif make additional interactions with the target (the goal is to generate protein reagents with high affinity and specificity for individual SH3 family members). Backbones scaffolding a proline rich SH3-binding motif (PPPRPPK) recognized by the Grb2 SH3 domain were generated using Rosetta remodel (see legend of Fig. 4; the SH3-binding motif is colored in green in Fig. 4A), but sequences designed for these backbones and expressed in E. coli did not fold to structures that bind Grb2 (Fig. 4B; the design problem is challenging as very few native proteins have proline rich secondary structure elements that closely interact with the core of the protein). To test whether ProteinMPNN could overcome this problem, we generated sequences for the same backbones while keeping the core SH3binding motif sequence (PPPRPPK) fixed, and expressed the proteins in E. coli. Biolayer interferometry experiments showed strong binding to the Grb2 SH3 domain (Fig. 4B), with considerably higher signal than the free proline rich peptide; point mutations predicted to disrupt the design completely eliminated the binding signal. Thus ProteinMPNN can generate sequences for challenging protein design problems even when traditional RosettaDesign fails.

作为最终测试，我们评估了 ProteinMPNN 是否能挽救此前使用 Rosetta 失败的新蛋白功能设计。我们选择了一个具有挑战性的例子：设计能支撑 SH3 结构域识别的多脯氨酸 II 型螺旋结构的蛋白质，这类设计中，除了核心 SH3 结合基序以外的支架部分还需与靶标形成额外相互作用（目标是生成具有高亲和力和高特异性的蛋白试剂，用于识别特定 SH3 家族成员）。我们使用 Rosetta remodel 生成了支撑 Grb2 SH3 结构域识别的富含脯氨酸的结合基序（PPPRPPK）的骨架（详见图 4注释；该基序在图 4A 中标为绿色），但基于这些骨架设计的序列在大肠杆菌中表达后，并未折叠为可与 Grb2 结合的结构（图 4B；该设计难点在于，很少有天然蛋白具有可与蛋白核心紧密相互作用的富脯氨酸二级结构元素）。为验证 ProteinMPNN 能否克服此问题，我们为这些相同骨架在保留核心 SH3 结合基序（PPPRPPK）的前提下生成新序列，并在大肠杆菌中表达。生物层干涉实验表明，这些设计的蛋白质与 Grb2 SH3 结构域结合紧密（图 4B），信号远强于游离脯氨酸肽；对预测中关键位点的点突变完全消除了结合信号。因此，ProteinMPNN 可为具有挑战性的蛋白设计问题生成有效序列，哪怕传统 RosettaDesign 方法失败。

## Conclusion **结论**

ProteinMPNN solves sequence design problems in a fraction of the time required for physically based approaches such as Rosetta, which carry out large scale sidechain packing calculations, achieves much higher protein sequence recovery on native backbones ( 52.4% vs 32.9% ), and rescues previously failed designs made using Rosetta or AlphaFold for protein monomers, assemblies, and protein-protein interfaces. Machine learning sequence design approaches have been developed previously (1-7), including the message passing method on which ProteinMPNN is based, but have focused on the monomer design problem, achieved lower native sequence recoveries, and with the exception of a TIM barrel design study (6) have not been extensively validated using crystallography and cryoEM. Whereas structure prediction methods can be evaluated purely in silico, this is not the case for protein design methods: In silico metrics such as native sequence recovery are very sensitive to crystallographic resolution (fig. S3, B and C) and may not correlate with proper folding (even a single residue substitution, while causing little change in overall sequence recovery, can block folding); in the same way that language translation accuracy must ultimately be evaluated by human users, the ultimate test of sequence design methods is experimental characterization.

ProteinMPNN 在解决蛋白质序列设计问题方面所需时间远少于传统物理模型方法（如 Rosetta，该方法需进行大规模侧链构象搜索），在天然骨架上的序列恢复率显著更高（52.4% vs 32.9%），并能成功挽救此前使用 Rosetta 或 AlphaFold 在蛋白质单体、组装体及蛋白-蛋白界面设计中失败的案例。虽然此前已有多种基于机器学习的序列设计方法被提出（1-7），包括构成 ProteinMPNN 基础的消息传递网络，但它们多聚焦于单体设计问题，序列恢复率相对较低，除一项 TIM 桶蛋白设计研究（6）外，很少通过晶体学或冷冻电镜进行充分验证。与结构预测方法可通过计算模拟完全评估不同，蛋白设计方法的评估不能仅依赖计算：例如，天然序列恢复率对晶体分辨率非常敏感（见图 S3B 和 C），且与蛋白是否能正确折叠未必相关（即使只改变一个残基，在整体序列恢复率上变化不大，也可能完全阻止蛋白折叠）；正如语言翻译的准确性最终需由人类评估，序列设计方法的终极检验是实验验证。

Unlike Rosetta and other physically based methods, ProteinMPNN requires no expert customization for specific design challenges, and it should thus make protein design more broadly accessible. This robustness reflects fundamental differences in how the sequence design problem is framed. In traditional physically based approaches, sequence design maps to the problem of identifying an amino acid sequence whose lowest energy state is the desired structure. This is, however, computationally intractable as it requires computing energies over all possible structures, including unwanted oligomeric and aggregated states; instead as a proxy Rosetta and other approaches

carry out a search for the lowest energy sequence for a given backbone structure, and structure prediction calculations are required in a second step to confirm that there are no other structures in which the sequence has still lower energy. Because of the lack of concordance between the design objective and what is being explicitly optimized, considerable customization can be required to generate sequences which fold; for example in Rosetta design calculations hydrophobic amino acids are often restricted on the protein surface as they can stabilize undesired multimeric states, and at the boundary region between the protein surface and core there can be considerable ambiguity about the extent to which such restrictions should be applied. While deep learning methods lack the physical transparency of methods like Rosetta, they are trained directly to find the most probable amino acid for a protein backbone given all the examples in the PDB, and hence such ambiguities do not arise, making sequence design more robust and less dependent on the judgement of a human expert.
与 Rosetta 等基于物理的方法不同，ProteinMPNN 无需专家对特定设计任务进行定制调整，因此可大幅提升蛋白设计的普及性。这种鲁棒性反映出序列设计问题建模方式上的根本差异。在传统物理方法中，序列设计被等同于在所有可能结构中识别出最低能量态对应于目标结构的氨基酸序列。然而，这在计算上几乎是不可行的，因为这需要评估包括不期望的寡聚体或聚集态在内的所有结构的能量。因此，Rosetta 等方法采用的近似方式是：寻找在给定骨架结构下能量最低的序列，随后还需通过结构预测步骤确认该序列是否不会在其他构象中能量更低。由于设计目标与优化目标不一致，为获得可折叠序列往往需要进行大量定制，例如在 Rosetta 中常限制蛋白表面的疏水氨基酸使用，以防其稳定不需要的多聚体状态；在蛋白表面与核心的边界区域，对于是否应实施这些限制则常常存在不确定性。相比之下，深度学习方法虽缺乏物理可解释性，但其直接从 PDB 数据中学习，在给定骨架条件下找到最可能的氨基酸，因此避免了这些歧义，使得序列设计更加鲁棒，且不再依赖专家判断。

The high rate of experimental design success of ProteinMPNN, together with the compute efficiency, applicability to almost any protein sequence design problem, and lack of requirement for customization should make it very broadly useful for protein design. ProteinMPNN generated sequences also have a much higher propensity to crystallize, greatly facilitating structure determination of designed proteins (15). The observation that ProteinMPNN generated sequences are predicted to fold to native protein backbones more confidently and accurately than the original native sequences (using single sequence information in both cases) suggests that ProteinMPNN may also be widely useful in improving expression and stability of recombinantly expressed native proteins (with residues required for function kept fixed).
ProteinMPNN 在实验设计中的高成功率，加之其高效的计算性能、几乎可适用于所有蛋白质序列设计任务、且无需定制设置的特性，预示其在蛋白设计中具有广泛应用前景。此外，ProteinMPNN 所生成的序列更容易结晶，大大促进了所设计蛋白的结构解析（15）。更重要的是，我们观察到 ProteinMPNN 生成的序列在仅使用单一序列信息的情况下，比原始天然序列更有信心、更准确地被预测为折叠成目标结构，表明 ProteinMPNN 也可能广泛用于提高重组表达的天然蛋白的表达量和稳定性（在保留功能性残基的前提下）。

# References and Notes


1. L. J. Ingraham, V. K. Garg, R. Barzilay, T. Jaakkola, "Generative models for graph-based protein design" in Advances in Neural Information Processing Systems 32 (NeurIPS 2019), H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett, Eds. (Neural Information Processing Systems Foundation, 2019), pp. 15741-15752.

2. Y. Zhang, Y. Chen, C. Wang, C. C. Lo, X. Liu, W. Wu, J. Zhang, ProDCoNN: Protein design using a convolutional neural network. Proteins 88, 819-829 (2020). doi:10.1002/prol. 25868 Medline

3. Y. Qi, J. Z. H. Zhang, DenseCPD: Improving the accuracy of neural-network-based computational protein sequence design with DenseNet. J. Chem. Inf. Model. 60, 1245-1252 (2020). doi:10.1021/acs.jcom.0c00043 Medline

4. B. Jing, S. Eismann, P. Suriana, R. J. L. Townshend, R. Dror, "Learning from protein structure with geometric vector perceptrons," paper presented at the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.

5. A. Strokach, D. Becerra, C. Corbi-Verge, A. Perez-Riba, P. M. Kim, Fast and flexible protein design using deep graph neural networks. Cell Syst. 11, 402-411.e4 (2020). doi:10.1016//csis.2020.08.016 Medline

6. N. Anand, R. Eguchi, I. I. Mathews, C. P. Perez, A. Derry, R. B. Altman, P. S. Huang, Protein sequence design with a learned potential. Nat. Commun. 13, 746 (2022). doi:10.1038/s41467-022-28313-9 Medline

7. C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, A. Rives, Learning inverse folding from millions of predicted structures. bioRxiv 2022.04.10.487779 [Preprint] (2022): https://doi.org/10.1101/2022.04.10.487779.

8. C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J. M. Thornton, CATH-A hierarchic classification of protein domain structures. Structure 5, 1093-1108 (1997). doi:10.1016/S0969-2126(97)00260-8 Medline

9. B. Uria, I. Murray, H. Larochelle, "A deep and tractable density estimator" in Proceedings of the 31st International Conference on Machine Learning, E. P. Xing, T. Jebara, Eds. (JMLR, 2014), pp. 467-475.

10. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. Nature 596, 583-589 (2021). doi:10.1038/s41586-021-03819-2 Medline

11. A. Leaver-Fay, M. J. O'Meara, M. Tyka, R. Jacak, Y. Song, E. H. Kellogg, J. Thompson, I. W. Davis, R. A. Pache, S. Lyskov, J. J. Gray, T. Kortemme, J. S. Richardson, J. J. Havranek, J. Snoeyink, D. Baker, B. Kuhlman, Scientific benchmarks for guiding macromolecular energy function improvement. Methods Enzymol. 523, 109-143 (2013). doi:10.1016/B978-0-12-394292-0.00006-0 Medline

12. J. K. Leman, B. D. Weitzner, S. M. Lewis, J. Adolf-Bryfogle, N. Alam, R. F. Alford, M. Aprahamian, D. Baker, K. A. Barlow, P. Barth, B. Basanta, B. J. Bender, K. Blacklock, J. Bonet, S. E. Boyken, P. Bradley, C. Bystroff, P. Conway, S. Cooper, B. E. Correia, B. Coventry, R. Das, R. M. De Jong, F. DiMaio, L. Dsilva, R. Dunbrack, A. S. Ford, B. Frenz, D. Y. Fu, C. Geniesse, L. Goldschmidt, R. Gowthaman, J. J. Gray, D. Gront, S. Guffy, S. Horowitz, P.-S. Huang, T. Huber, T. M. Jacobs, J. R. Jeliazkov, D. K. Johnson, K. Kappel, J. Karanicolas, H. Khakzad, K. R. Khar, S. D. Khare, F. Khatib, A. Khramushin, I. C. King, R. Kleffner, B. Koeprick, T. Kortemme, G. Kuenze, B. Kuhlman, D. Kuroda, J. W. Labonte, J. K. Lai, G. Lapidoth, A. LeaverFay, S. Lindert, T. Linsky, N. London, J. H. Lubin, S. Lyskov, J. Maguire, L. Malmström, E. Marcos, O. Marcu, N. A. Marze, J. Meiler, R. Moretti, V. K. Mulligan, S. Nerli, C. Norn, S. Ó Conchúr, N. Ollikainen, S. Ovchinnikov, M. S. Pacella, X. Pan, H. Park, R. E. Pavlovicz, M. Pethe, B. G. Pierce, K. B. Pilla, B. Raveh, P. D. Renfrew, S. S. R. Burman, A. Rubenstein, M. F. Sauer, A. Scheck, W. Schief, O. SchuelerFurman, Y. Sedan, A. M. Sevy, N. G. Sgourakis, L. Shi, J. B. Siegel, D.-A. Silva, S. Smith, Y. Song, A. Stein, M. Szegedy, F. D. Teets, S. B. Thyme, R. Y.-R. Wang, A. Watkins, L. Zimmerman, R. Bonneau, Macromolecular modeling and design in Rosetta: Recent methods and frameworks. Nat. Methods 17, 665-680 (2020). doi:10.1038/s41592-020-0848-2 Medline

13. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhtheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R.

J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871-876 (2021). doi:10.1126/science.ab8754 Medline

14. V. Mariani, M. Biasini, A. Barbato, T. Schwede, lDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 29, 2722-2728 (2013). doi:10.1093/bioinformatics/bt Medline

15. B. I. M. Wicky, L. F. Milles, A. Courbet, R. J. Ragotte, J. Dauparas, E. Kinfu, S. Tipps, R. D. Kibler, M. Baek, F. DiMaio, X. Li, L. Carter, A. Kang, H. Nguyen, A. K. Bera, D. Baker, Hallucinating symmetric protein assembly. Science 10.1126/science.add1964 (2022).

16. N. P. King, J. B. Bale, W. Sheffler, D. E. McNamara, S. Gonen, T. Gonen, T. O. Yeates, D. Baker, Accurate design of co-assembling multi-component protein nanomaterials. Nature 510, 103-108 (2014). doi:10.1038/nature13404 Medline

17. S. Boyoglu-Barnum, D. Ellis, R. A. Gillespie, G. B. Hutchinson, Y.-J. Park, S. M. Moin, O. J. Acton, R. Ravichandran, M. Murphy, D. Pettie, N. Matheson, L. Carter, A. Creanga, M. J. Watson, S. Kephart, S. Ataca, J. R. Valle, G. Ueda, M. C. Crank, L. Stewart, K. K. Lee, M. Guttman, D. Baker, J. R. Mascola, D. Veesler, B. S. Graham, N. P. King, M. Kanekiyo, Quadrivalent influenza nanoparticle vaccines induce broad protection. Nature 592, 623-628 (2021). doi:10.1038/s41586-021-03365-x Medline

18. A. C. Walls, B. Fiala, A. Schäfer, S. Wrenn, M. N. Pham, M. Murphy, L. V. Tse, L. Shehata, M. A. O'Connor, C. Chen, M. J. Navarro, M. C. Miranda, D. Pettie, R. Ravichandran, J. C. Kraft, C. Ogishara, A. Palser, S. Chalk, E.-C. Lee, K. Guerriero, E. Kepl, C. M. Chow, C. Sydeman, E. A. Hodge, B. Brown, J. T. Fuller, K. H. Dinnon III, L. E. Gralinski, S. R. Leist, K. L. Gully, T. B. Lewis, M. Guttman, H. Y. Chu, K. K. Lee, D. H. Fuller, R. S. Baric, P. Kellam, L. Carter, M. Pepper, T. P. Sheahan, D. Veesler, N. P. King, Elicitation of potent neutralizing antibody responses by designed protein nanoparticle vaccines for SARS-CoV-2. Cell 183, 1367-1382.e17 (2020). doi:10.1016/j.cell.2020.10.043 Medline

19. J. Marcandalli, B. Fiala, S. Ols, M. Perotti, W. de van der Schueren, J. Snijder, E. Hodge, M. Benhaim, R. Ravichandran, L. Carter, W. Sheffler, L. Brunner, M. Lawrenz, P. Dubois, A. Lanzavecchia, F. Sallusto, K. K. Lee, D. Veesler, C. E. Correnti, L. J. Stewart, D. Baker, K. Loré, L. Perez, N. P. King, Induction of potent neutralizing antibody responses by a designed protein nanoparticle vaccine for respiratory syncytial virus. Cell 176, 1420-1431.e17 (2019). doi:10.1016/j.cell.2019.01.040 Medline

20. L. Cao, B. Coventry, I. Goreshnik, B. Huang, W. Sheffler, J. S. Park, K. M. Jude, I. Marković, R. U. Kadam, K. H. G. Verschueren, K. Verstraete, S. T. R. Walsh, N. Bennett, A. Phal, A. Yang, L. Kozodoy, M. DeWitt, L. Picton, L. Miller, E.-M. Strauch, N. D. DeBouver, A. Pires, A. K. Bera, S. Halabiya, B. Hammerson, W. Yang, S. Bernard, L. Stewart, I. A. Wilson, H. Ruzhola-Baker, J. Schlessinger, S. Lee, S. N. Savvides, K. C. Garcia, D. Baker, Design of protein-binding proteins from the target structure alone. Nature 605, 551-560 (2022). doi:10.1038/s41586-022-04654-9 Medline

21. J. Dauparas, S. O. S. Duerr, dauparas/ProteinMPNN: ProteinMPNN (v1.0.0). Zenodo (2022): https://doi.org/10.5281/z

22. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need" in Advances in Neural Information Processing Systems 30 (NeurIPS 2017), I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Eds. (Neural Information Processing Systems Foundation, 2017), pp. 5999-6009.

23. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929-1958 (2014).

24. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for computer vision" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE, 2016), pp. 2818-2826.

25. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. 35, 1026-1028 (2017). doi:10.1038/nbt. 3988 Medline

26. Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302-2309 (2005). doi:10.1093/nar/ykj524 Medline

27. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. Devito, M. Raison, A. Tejani, S. Chilamkurth, "Pytorch: An imperative style, high-performance deep learning library" in Advances in Neural Information Processing Systems 32 (NeurIPS 2019), H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett, Eds. (Neural Information Processing Systems Foundation, 2019), pp. 7994-8005.

28. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, "Neural message passing for quantum chemistry" in Proceedings of the 34th International Conference on Machine Learning, D. Precup, Y. W. The, Eds. (JMLR, 2017), pp. 1263 – 1272.

29. B. Dang, M. Mravic, H. Hu, N. Schmidt, B. Mensa, W. F. DeGrado, SNAC-tag for sequence-specific chemical protein cleavage. Nat. Methods 16, 319-322 (2019). doi:10.1038/s41992-019-0357-3 Medline

30. W. Kabsch, XDS, Acta Crystallogr. D Biol. Crystallogr. 66, 125-132 (2010). doi:10.1007/S0907444900047337 Medline

31. M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. 67, 235-242 (2011). doi:10.1107/S0907444910045749 Medline

32. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. J. Appl. Crystallogr. 40, 658-674 (2007). doi:10.1107/S0021889807021206 Medline

33. P. Emsley, K. Cowtan, Coot: Model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. 60, 2126-2132 (2004). doi:10.1107/S0907444904019158 Medline

34. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: A comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol. Crystallogr. 66, 213221 (2010). doi:10.1107/S0907444909052925 Medline

35. C. J. Williams, J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma, D. A. Keedy, B. J. Hintze, V. B. Chen, S. Jain, S. M. Lewis, W. B. Arendall III, J. Snoeyink, P. D. Adams, S. C. Lovell, J. S. Richardson, D. C. Richardson, MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. 27, 293-315 (2018). doi:10.1002/prp. 3330 Medline

# Acknowledgments

The authors wish to thank Sergey Ovchinnikov, Chris Norn, David Juergens, Jue Wang, Frank DiMaio, Ryan Kibler, Minkyung Baek, Sanaa Mansoor, Luki Goldschmidt, and Lance Stewart for helpful discussions. The authors would also like to thank the Meta AI protein team for sharing AlphaFold models generated for UniRef50 sequences. The Berkeley Center for Structural Biology is supported in part by the National Institutes of Health (NIH), National Institute of General Medical Sciences. Crystallographic data collected at The Advanced Light Source (ALS) and is supported by the Director, Office of Science, Office of Basic Energy Sciences and US Department of Energy under contract number DEACO2- OSCH11231. Funding: This work was supported with funds provided by a gift from Microsoft (J.D., D.T., D.B.), the Audacious Project at the Institute for Protein Design (A.B., A.K., B.K., F.C., T.F.H., R.J.dH., N.P.K., D.B.), a grant from the NSF (DBI 1937533 to D.B. and I.A.), an EMBO long-term fellowship ALTF 1392018 (B.I.M.W.), the Open Philanthropy Project Improving Protein Design Fund (R.J.R., D.B.), Howard Hughes Medical Institute Hanna Gray fellowship grant GT11817 (N.Beth.), The Donald and Jo Anne Petersen Endowment for Accelerating Advancements in Alzheimer's Disease Research (N.Ben.), a Washington Research Foundation Fellowship (S.P.), an Alfred P. Sloan Foundation Matter-to-Life Program Grant (G-2021-16899, A.C., D.B.), a Human Frontier Science Program Cross Disciplinary Fellowship (LTD00395/2020-C, L.F.M.), an EMBO Non-Stipendiary Fellowship (ALTF 1047-2019, L.F.M.), the National Science Foundation Graduate Research Fellowship (DGE-2140004, P.J.Y.L), the Howard Hughes Medical Institute (A.C., H.B., D.B.), and the National Institutes of Health, National Institute of General Medical Sciences, P30 GM124169-01(B.S.). We thank Microsoft and AWS for generous gifts of cloud computing credits. Author contributions: Conceptualization: JD, LFM, BIMW, AC, RJdH, HB, NBen; Methodology: JD, IA, PJYL; Software: JD, TFH, DT, BK, FC; Validation: JD, NBen, HB, AKB, BS, AK, HN, SP, PJYL, NBeth, RJdH, LFM, BIMW, AC, RJR; Formal analysis: JD, LFM, BIMW, RJR, NBen; Resources: JD, DB; Data curation: IA, JD, HB, ; Writing - original draft: JD, DB; Writing - review and editing: JD, DB; Visualization: JD, RJR, RJdH, HB, LFM, BIMW, PJYL, HB; Supervision: DB, NPK; Project administration: JD; Funding acquisition: JD, DB. Competing interests: Authors declare that they have no competing interests. Data and materials availability: All data are available in the main text or as supplementary materials. ProteinMPNN code (21) is available at https://github.com/dauparas/ProteinMPNN. 

# Supplementary Materials

science.org/doi/10.1126/science.add2187 Materials and Methods
Figs. S1 to S12
Table S1
References (22-35)

Submitted 27 May 2022; accepted 7 September 2022
Published online 15 September 2022
10.1126/science.add2187

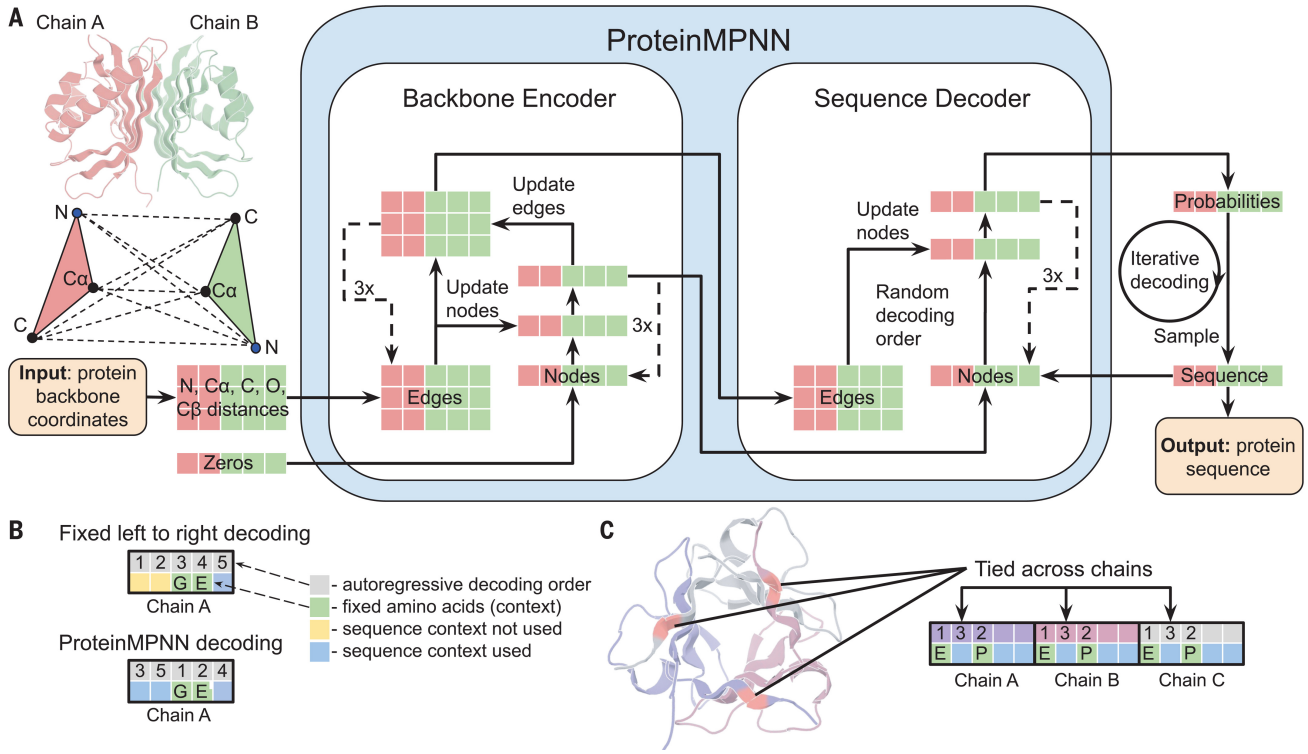


Figure 1: ProteinMPNN architecture. (A) Distances between N, C$\alpha$, C, O, and virtual C$\beta$ are encoded and processed using a message passing neural network (Encoder) to obtain graph node and edge features. The encoded features together with a partial sequence are used to generate amino acids iteratively in a random decoding order. (B) A fixed left to right decoding cannot use sequence context (green) for preceding positions (yellow) whereas a model trained with random decoding orders can be used with arbitrary decoding order during the inference. The decoding order can be chosen such that the fixed context is decoded first. (C) Residue positions within and between chains can be tied together, enabling symmetric, repeat protein, and multistate design. In this example, a homo-trimer is designed with coupling of positions in different chains. Predicted unnormalized probabilities for tied positions are averaged to get a single probability distribution from which amino acids are sampled.
ProteinMPNN 架构。（A）N、C$\alpha$、C、O 和虚拟 C$\beta$ 之间的距离被编码，并通过消息传递神经网络（编码器）处理以获得图节点和边特征。将编码特征与部分已知序列结合后，模型按随机解码顺序迭代生成氨基酸。（B）固定从左到右的解码顺序无法使用前一个位置的序列上下文信息（绿色）来预测当前位置（黄色），而使用随机解码顺序训练的模型在推理阶段可采用任意解码顺序。可以选择优先解码固定区域，从而更好利用上下文信息。（C）链内和链间的残基位置可以绑定在一起，支持对称性、重复蛋白及多状态设计。在此示例中，设计了一个同源三聚体，并对不同链之间的相应残基位置进行了耦合。对于绑定的残基位置，预测出的非归一化概率被取平均，以生成一个联合的概率分布，用于采样生成氨基酸。

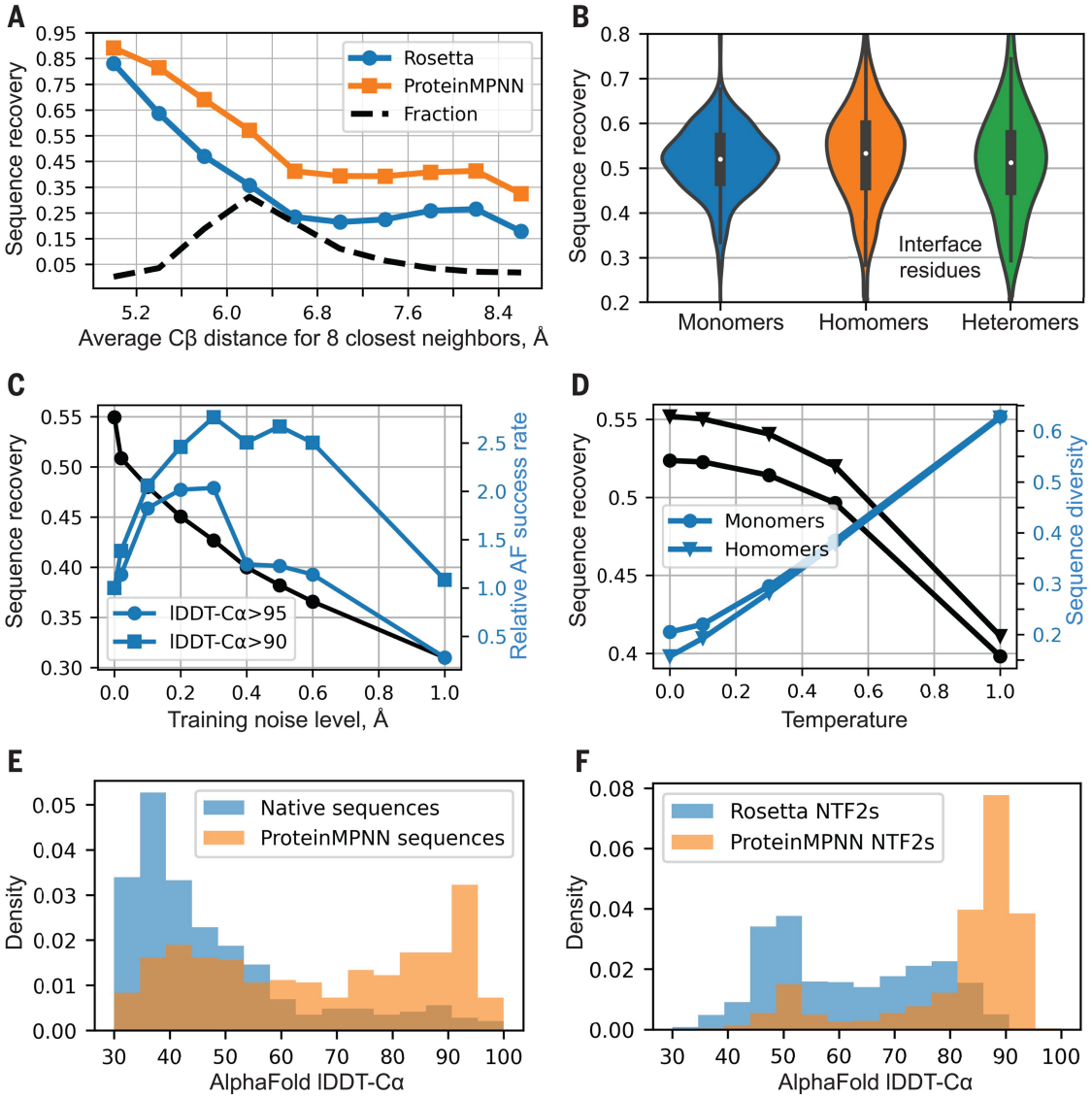


Figure 2: In silico evaluation of ProteinMPNN. (A) ProteinMPNN has higher native sequence recovery than Rosetta. The average C$\beta$ distance of the 8 closest neighbors ( x axis) reports on burial, with most buried positions on the left and more exposed on the right; ProteinMPNN outperforms Rosetta at all levels of burial. Average sequence recovery for ProteinMPNN was 52.4%, compared to 32.9% for Rosetta. (B) ProteinMPNN has high sequence recovery for monomers and for both homo-oligomer and hetero-oligomer interfaces ( $C\beta - C\beta < 8$ ); violin plots are for 690 monomers, 732 homomers, 98 heteromers. (C) Sequence recovery (black) and relative AlphaFold success rates (blue) as a function of training noise level. For higher accuracy predictions (circles) smaller amounts of noise are optimal (1.0 corresponds to 1.8% success rate), while to maximize prediction success at a lower accuracy cutoff (squares), models trained with more noise are better ( 1.0 corresponds to 6.7% success rate). (D) Sequence recovery and diversity as a function of sampling temperature. (E) Redesign of native protein backbones with ProteinMPNN considerably increases AphaFold prediction accuracy compared to the original native sequence using no multiple sequence information. Single sequences (designed or native) were input in both cases. (F) ProteinMPNN redesign of previous Rosetta designed NTF2 fold proteins (3,000 backbones in total) results in considerably improved AlphaFold single sequence prediction accuracy.
ProteinMPNN 的计算模拟评估。（A）ProteinMPNN 的原生序列恢复率高于 Rosetta。图中 x 轴表示 8 个最近邻的平均 C$\beta$ 距离，用以反映残基的埋藏程度——越靠左表示越埋藏，越靠右则越暴露；在所有埋藏水平下，ProteinMPNN 的表现均优于 Rosetta。ProteinMPNN 的平均序列恢复率为 52.4%，而 Rosetta 为 32.9%。（B）ProteinMPNN 在单体、同源寡聚体界面和异源寡聚体界面上均具有较高的序列恢复率（定义为 $C\beta - C\beta < 8$）；小提琴图展示了 690 个单体、732 个同源体、98 个异源体的恢复结果。（C）图中展示了序列恢复率（黑色）与 AlphaFold 相对预测成功率（蓝色）随训练噪声水平的变化趋势。对于更高准确性的预测（圆点），较小的噪声更为理想（噪声水平为 1.0 时对应 1.8% 成功率）；而在较低准确度阈值下最大化预测成功率（方块）时，更大的训练噪声表现更好（1.0 对应 6.7% 成功率）。（D）序列恢复率与多样性随采样温度的变化。（E）使用 ProteinMPNN 对天然蛋白骨架进行重新设计，显著提升了 AlphaFold 在无多序列信息输入下的结构预测准确性。无论是设计序列还是原始天然序列，两者输入的均为单一序列。（F）使用 ProteinMPNN 对先前由 Rosetta 设计的 NTF2 折叠蛋白（共 3,000 个骨架）进行重新设计后，AlphaFold 的单序列预测准确性显著提升。

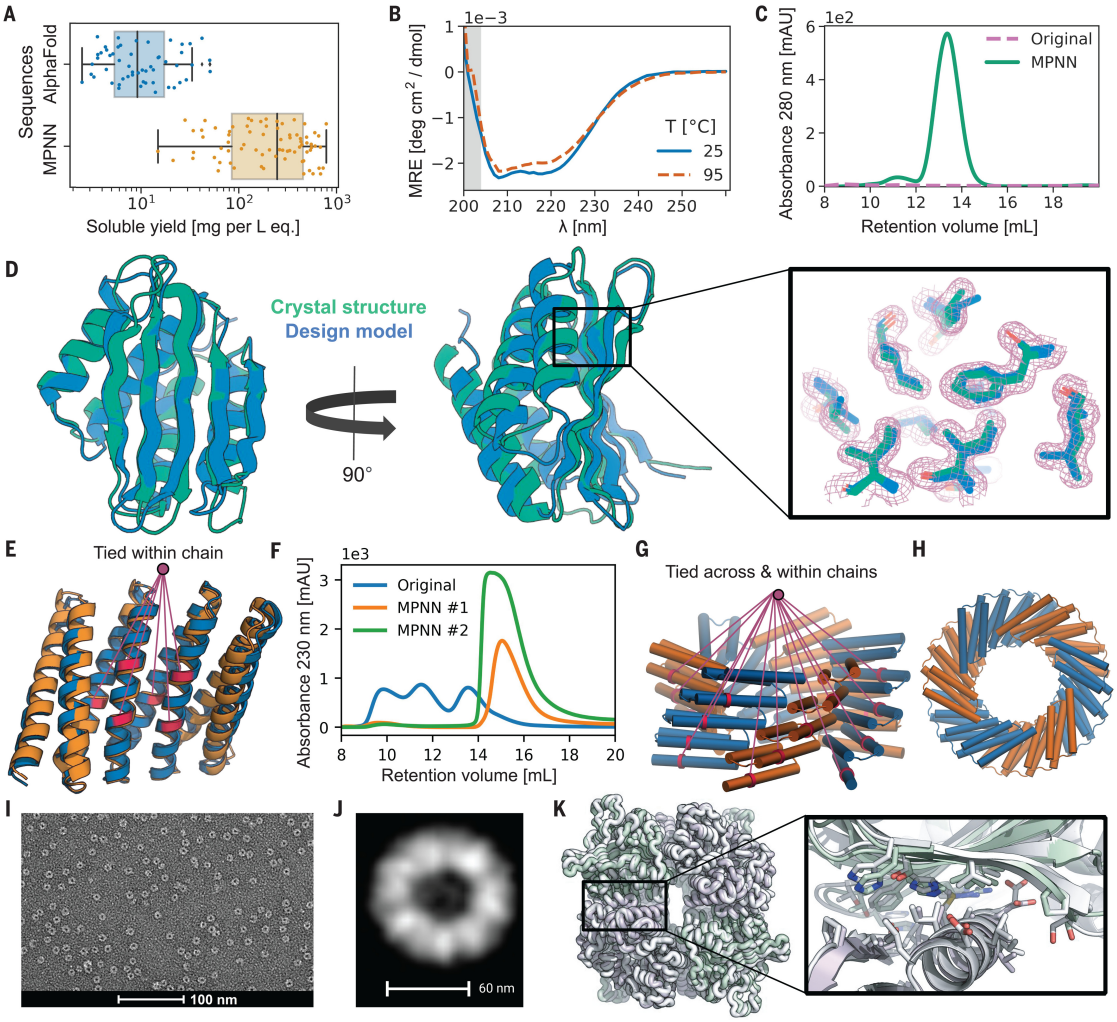


Figure 3: Structural characterization of ProteinMPNN designs. (A) Comparison of soluble protein expression over a set of AlphaFold hallucinated monomers and homo-oligomers (blue) and the same set of backbones with sequences designed using ProteinMPNN (orange), N=129. The total soluble protein yield following expression in E. coli, obtained from the integrated area unders size exclusion traces of nickel-NTA purified proteins, increases considerably from the barely soluble protein of the original sequences following ProteinMPNN rescue (median yields for 1 L of culture equivalent: 9 and 247 mg respectively). (B to D) In depth characterization of a monomer hallucination and corresponding ProteinMPNN rescue from the set in (A).

ProteinMPNN 设计的结构表征。（A）展示了在 AlphaFold 幻觉生成的蛋白单体与同源寡聚体（蓝色）和使用 ProteinMPNN 重新设计序列的相同骨架（橙色）上的可溶蛋白表达比较，样本数为 N=129。将这些蛋白在大肠杆菌中表达、经 Ni-NTA 亲和层析纯化，并通过凝胶过滤色谱的积分面积计算蛋白产量后发现，使用 ProteinMPNN 设计的序列相比原始几乎不可溶的蛋白，显著提高了可溶蛋白的表达（相当于 1 L 培养液的中位产率：原始为 9 mg，ProteinMPNN 设计后为 247 mg）。（B 至 D）为图（A）中某一单体幻觉设计及其对应 ProteinMPNN 拯救设计的深入表征。

Like almost all of the designs in (A), the sequence and structural similarity to the PDB of the design model are very low (E-value=2.8 against UniRef100 using HHblits, TM-score=0.56 against PDB). As shown in (B), the ProteinMPNN rescued design has high thermostability, with a virtually unchanged circular dichroism profile at 95°C compared to 25°C. Shown in (C) is a size exclusion (SEC) profile of the failed original design overlaid with the ProteinMPNN sequence design, which has a clear monodisperse peak at the expected retention volume. As shown in (D), the crystal structure of the ProteinMPNN (8CYK) design is nearly identical to the design model (2.35 Å RMSD over 130 residues); see fig. S5 for additional information. Right panel shows model sidechains in the electron density, in green crystal side chains, in blue AlphaFold side chains.

与图（A）中几乎所有设计一样，该设计模型在序列和结构上与 PDB 的相似性极低（使用 HHblits 对 UniRef100 的比对 E-value 为 2.8，对 PDB 的 TM-score 为 0.56）。如图（B）所示，ProteinMPNN 拯救设计具有极高的热稳定性，其在 95°C 和 25°C 条件下的圆二色谱（CD）光谱几乎没有变化。图（C）展示了原始失败设计与使用 ProteinMPNN 重新设计序列的 SEC（凝胶过滤）色谱曲线对比，后者在预期保留体积处显示出清晰的单分散峰。如图（D）所示，ProteinMPNN 设计的晶体结构（PDB 编号：8CYK）与设计模型几乎一致（130 个残基上的主链 RMSD 为 2.35 Å）；更多信息见图 S5。右图显示了电子密度图中的模型侧链，绿色为晶体结构侧链，蓝色为 AlphaFold 预测的侧链。

(E and F) ProteinMPNN rescue of Rosetta design made from a perfectly repeating structural and sequence unit (DHR82). Residues at corresponding positions in the repeat unit were tied during ProteinMPNN sequence inference. Shown in (E) are a backbone design model (orange) and MPNN redesigned sequence AlphaFold model (blue) with tied residues indicated by lines ( $\sim$ 1.2 error over 232 residues). Shown in (F) is a SEC profile of the IMAC purified original Rosetta design and two ProteinMPNN redesigns.
(E 和 F）展示了对 Rosetta 所设计的完全重复结构和序列单元（DHR82）进行的 ProteinMPNN 拯救。在 ProteinMPNN 序列推理过程中，重复单元中对应位置的残基被绑定。如图（E）所示，展示了骨架设计模型（橙色）与 ProteinMPNN 重新设计的序列经 AlphaFold 预测的模型（蓝色），通过线条标注了绑定的残基位置（232 个残基范围内误差约为 1.2 Å）。图（F）展示了原始 Rosetta 设计（经 IMAC 纯化）及两个 ProteinMPNN 重新设计的 SEC 曲线对比。
(G and H) Tying residues during ProteinMPNN sequence inference both within and between chains to enforce both repeat protein and cyclic symmetries. Shown in (G) is a side view of the design model. A set of tied residues are shown in red. Shown in (H) is a top-down view of the design model.
（G 和 H）在 ProteinMPNN 序列推理过程中同时在链内与链间绑定残基，以强制实现重复蛋白与环状对称性。如图（G）所示为设计模型的侧视图，红色标示了一组绑定残基。图（H）为同一设计模型的俯视图。
(I) Negative stain electron micrograph of purified design. (J) Class average of images from I closely match top down view in (H).
（I）为所纯化设计的负染电镜图像。（J）为图（I）中图像的类别平均图，与图（H）中的俯视图高度一致。
(K) Rescue of the failed two-component Rosetta tetrahedral nanoparticle design T33-27 (16) by ProteinMPNN interface design. Following ProteinMPNN rescue, the nanoparticle assembled readily with high yield, and the crystal structure (grey) is very nearly identical to the design model (green/purple) (backbone RMSD of 1.2 over two complete asymmetric units forming the ProteinMPNN rescued interface).
（K）展示了对 Rosetta 设计失败的双组分四面体纳米颗粒 T33-27（参考文献 16）通过 ProteinMPNN 界面设计进行的拯救。经 ProteinMPNN 拯救后，纳米颗粒能高效组装，晶体结构（灰色）与设计模型（绿色/紫色）高度一致（在两个完整不对称单元上的主链 RMSD 为 1.2，构成了 ProteinMPNN 拯救成功的界面）。

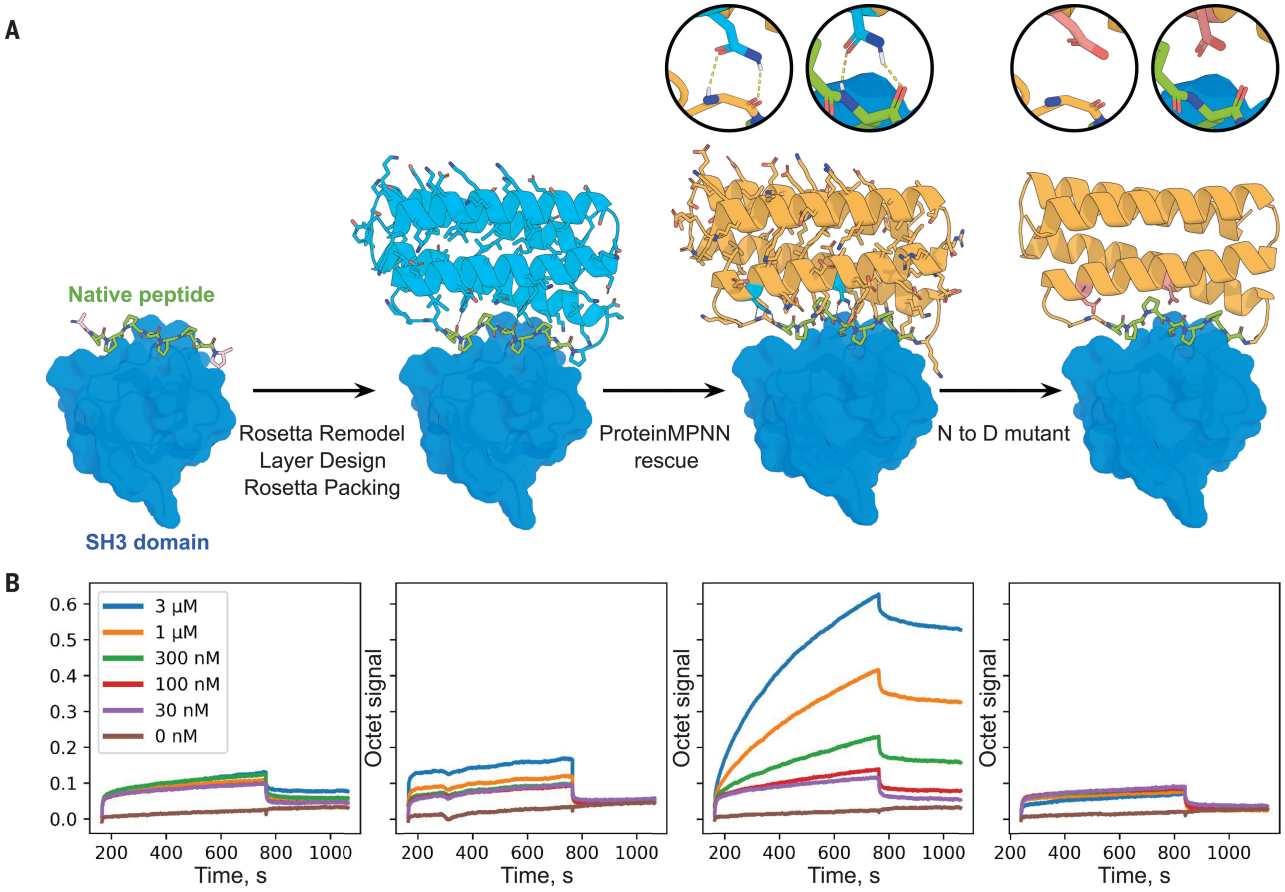


Figure 4: Design of protein function with ProteinMPNN. (A) Design scheme. First panel; structure (PDB 2WOZ) of a fragment of Gab2 peptide bound to the human Grb2 C-term SH3 domain (core SH3-binding motif PPPRPPK is in green, target rendered with surface and colored blue). Second panel: helical bundle scaffolds were docked to the exposed face of the peptide using RIFDOCK (20), and Rosetta remodel was used to build loops connecting the peptide to the scaffolds. Rosetta sequence design with layer design task operations was used to optimize the sequence of the fusion (Cyan) for stability, rigidity of the peptide-helical bundle interface, and binding affinity for the Grb2 SH3 domain. Third panel; ProteinMPNN redesign (orange) of the designed binder sequence; hydrogen bonds involving asparagine sidechains between the peptide and base scaffold are shown in green and in the inset. Fourth panel; Mutation of the two asparagines to aspartates to disrupt the scaffolding of the target peptide.
利用 ProteinMPNN 进行蛋白功能设计。(A) 设计流程。第一图：Gab2 肽段与人类 Grb2 C 端 SH3 结构域复合物的结构（PDB 编号 2WOZ）；核心 SH3 结合基序 PPPRPPK 用绿色表示，Grb2 SH3 靶点以表面模式显示并着蓝色。第二图：利用 RIFDOCK（参考文献 20）将螺旋束支架对接至肽段暴露表面，接着使用 Rosetta remodel 构建连接肽段与支架的环结构。随后，Rosetta 的 layer design 任务操作被用于优化融合蛋白（青色）的序列，以增强稳定性、肽段与螺旋束界面的刚性，以及对 Grb2 SH3 结构域的结合亲和力。第三图：对上述设计的结合体序列进行 ProteinMPNN 重设计（橙色）；肽段与支架之间由天冬酰胺侧链参与形成的氢键以绿色标出，并在插图中进一步放大。第四图：将两个天冬酰胺突变为天冬酸，以破坏目标肽段的支撑结构。
(B) Experimental characterization of binding using biolayer interferometry. Biotinylated C-term SH3 domain from human Grb2 was loaded onto Streptavidin (SA) Biosensors, which were then immersed in solutions containing varying concentrations of SH3binding peptide AIAPPPRPPKPSQ (left), or of the designs (right panels), and then transferred to buffer lacking added protein for dissociation measurements. The ProteinMPNN design (3rd panel from the left) has much greater binding signal than the original Rosetta design (2nd panel from the left); this is greatly reduced by the asparagine to aspartate mutations (last panel). Note that all designs as well as the native peptide are fused with sfGFP at the C terminus.
(B) 利用生物层干涉技术（BLI）对结合能力进行实验表征。将生物素标记的人 Grb2 C 端 SH3 结构域加载至链霉亲和素（SA）生物传感器上，然后分别浸入含不同浓度 SH3 结合肽 AIAPPPRPPKPSQ 的溶液（左图）或不同设计体（右图）中，随后转入无蛋白的缓冲液以测量解离过程。结果显示，ProteinMPNN 设计（左起第三图）的结合信号远高于原始 Rosetta 设计（左起第二图）；而将关键天冬酰胺突变为天冬酸后（最右图），其结合能力显著下降。需要注意的是，所有设计体与天然肽段均在 C 端融合了 sfGFP。

| Noise level when training: | Modification | Number of Parameters in millions | PDB Test Accuracy | PDB Test Perplexity | AlphaFold Model Accuracy 0.00Å/0.02Å |
|---|---|---|---|---|---|
| Baseline model | None | 1.381 | 41.2/40.1 | 6.51/6.77 | 41.4/41.4 Experiment 1 |
| Add N, C$\alpha$, C, C$\beta$, O distances | 1.430 | 49.0/46.1 | 5.03/5.54 | 45.7/47.4 Experiment 2 | Update encoder edges |
| 1.629 | 43.1/42.0 | 6.12/6.37 | 43.3/43.0 Experiment 3 | Combine 1 and 2 | 1.678 |
| 50.5/47.3 | 4.82/5.36 | 46.3/47.9 Experiment 4 | Experiment 3 with random instead of forward decoding | 1.678 | 50.8/47.9 |
| 4.74/5.25 | 46.9/48.5 | | | | |

Table 1: Single chain sequence design performance on CATH held out test split. Test accuracy (percentage of correct amino amino acids recovered) and test perplexity (exponentiated categorical cross entropy loss per residue) for models trained on the native backbone coordinates (left, normal font) and models trained with Gaussian noise (std=0.02 ) added to the backbone coordinates (right, bold font). Noise was only added during training and all test evaluations are with no added noise. The final column shows sequence recovery on 5,000 AlphaFold protein backbone models with average pLDDT $> 80.0$ randomly chosen from UniRef50 sequences.