# Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models

# 通过整合结构预测网络和扩散生成模型实现广泛适用且准确的蛋白质设计

Joseph L. Watson , David Juergens , Nathaniel R. Bennett , Brian L. Trippe , Jason Yim , Helen E. Eisenach , Woody Ahern , Andrew J. Borst , Robert J. Ragotte , Lukas F. Milles , Basile I. M. Wicky , Nikita Hanikel , Samuel J. Pellock , Alexis Courbet , William Sheffler , Jue Wang , Preetham Venkatesh , Isaac Sappington , Susana Vázquez Torres , Anna Lauko , Valentin De Bortoli , Emile Mathieu , Regina Barzilay , Tommi S. Jaakkola , Frank DiMaio , Minkyung Baek , David Baker\*1,2,11

约瑟夫·L·沃森，戴维·尤根斯，内森尼尔·R·贝内特，布莱恩·L·特里普，杰森·严，海伦·E·艾森纳赫，伍迪·阿赫恩，安德鲁·J·博尔斯特，罗伯特·J·拉戈特，卢卡斯·F·米勒，巴西尔·I·M·维基，尼基塔·哈尼克尔，塞缪尔·J·佩洛克，亚历克西斯·库尔贝，威廉·谢夫勒，王珏，普里萨姆·文卡特什，艾萨克·萨平顿，苏萨娜·巴斯克斯·托雷斯，安娜·劳科，瓦伦丁·德·博尔托利，埃米尔·马修，雷吉娜·巴齐莱，汤米·S·雅卡拉，弗兰克·迪迈奥，闵京白，大卫·贝克

# \*Equal contribution

# \*平等贡献

\*To whom correspondence should be addressed

\*通讯应寄至

1. Department of Biochemistry, University of Washington, Seattle, WA 98105, USA

1. 华盛顿大学生物化学系，美国华盛顿州西雅图，邮政编码98105

2. Institute for Protein Design, University of Washington, Seattle, WA 98105, USA

2. 华盛顿大学蛋白质设计研究所，美国华盛顿州西雅图，邮政编码98105

3. Graduate Program in Molecular Engineering, University of Washington, Seattle, WA 98105, USA

3. 华盛顿大学分子工程研究生项目，美国华盛顿州西雅图，邮政编码98105

4. Columbia University, Department of Statistics, New York, NY 10027, USA

4. 哥伦比亚大学统计系，美国纽约，邮政编码10027

5. Irving Institute for Cancer Design, Columbia University, New York, NY 10027, USA

5. 哥伦比亚大学癌症设计欧文研究所，纽约，NY 10027，美国

6. Massachusetts Institute of Technology, Cambridge, MA 02139, USA

6. 麻省理工学院，剑桥，MA 02139，美国

7. Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA 98105, USA

7. 保罗·G·艾伦计算机科学与工程学院，华盛顿大学，西雅图，WA 98105，美国

8. Graduate Program in Biological Physics, Structure and Design, University of Washington, Seattle, WA 98105, USA

8. 华盛顿大学生物物理、结构与设计研究生项目，西雅图，WA 98105，美国

9. Centre National de la recherche scientifique, École Normale Supérieure rue d’Ulm, Paris 75005, France

9. 法国国家科学研究中心，巴黎高等师范学校，乌尔姆街，巴黎75005，法国

10. Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom

10. 剑桥大学工程系，剑桥CB2 1PZ，英国

11. Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105, USA

11. 霍华德·休斯医学研究所，华盛顿大学，西雅图，华盛顿州 98105，美国

12. School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea

12. 首尔国立大学生物科学学院，首尔 08826，韩国

# Abstract

# 摘要

There has been considerable recent progress in designing new proteins using deep learning methods . Despite this progress, a general deep learning framework for protein design that enables solution of a wide range of design challenges, including de novo binder design and design of higher order symmetric architectures, has yet to be described. Diffusion models have had considerable success in image and language generative modeling but limited success when applied to protein modeling, likely due to the complexity of protein backbone geometry and sequence-structure relationships. Here we show that by fine tuning the RoseTTAFold structure prediction network on protein structure denoising tasks, we obtain a generative model of protein backbones that achieves outstanding performance on unconditional and topology-constrained protein monomer design, protein binder design, symmetric oligomer design, enzyme active site scaffolding, and symmetric motif scaffolding for therapeutic and metal-binding protein design. We demonstrate the power and generality of the method, called RoseTTAFold Diffusion (RFdiffusion), by experimentally characterizing the structures and functions of hundreds of new designs. In a manner analogous to networks which produce images from user-specified inputs, RFdiffusion enables the design of diverse, complex, functional proteins from simple molecular specifications.

最近在使用深度学习方法设计新蛋白质方面取得了相当大的进展 。尽管取得了这些进展，但尚未描述一种通用的深度学习框架，用于蛋白质设计，能够解决广泛的设计挑战，包括新型结合物设计和更高阶对称结构的设计。扩散模型 在图像和语言生成建模方面取得了显著成功，但在蛋白质建模中的应用成功有限，这可能是由于蛋白质主链几何形状和序列-结构关系的复杂性。在这里，我们展示了通过对RoseTTAFold结构预测网络进行微调，以处理蛋白质结构去噪任务，我们获得了一种蛋白质主链的生成模型，在无条件和拓扑约束的蛋白质单体设计、蛋白质结合物设计、对称寡聚体设计、酶活性位点支架设计以及用于治疗和金属结合蛋白设计的对称基序支架设计中表现出色。我们通过实验表征数百种新设计的结构和功能，展示了这种方法的强大和通用性，称为RoseTTAFold扩散(RFdiffusion)。类似于从用户指定输入生成图像的网络，RFdiffusion能够从简单的分子规格设计出多样、复杂、功能性强的蛋白质。

# Main

# 主要

De novo protein design seeks to generate proteins with specified structural and/or functional properties, for example making a binding interaction with a given target , folding into a particular topology , or stabilizing a desired functional "motif" (geometries and amino acid identities that produce a desired activity) . Denoising diffusion probabilistic models (DDPMs), a powerful class of machine learning models recently demonstrated to generate novel photorealistic images in response to text prompts , have several properties well-suited to protein design. First, DDPMs generate highly diverse outputs - DDPMs are trained to denoise data (for instance images or text) that have been corrupted with Gaussian noise; by learning to stochastically reverse this corruption, diverse outputs closely resembling the training data are generated. Second, DDPMs can be guided at each step of the iterative generation process towards specific design objectives through provision of conditioning information. Third, for almost all protein design applications it is necessary to explicitly model 3D structure; SE(3)-equivariant DDPMs are able to do this in a representation-frame independent manner. Recent work has adapted DDPMs for protein monomer design by conditioning on small protein “motifs”5,9 or on secondary structure and block-adjacency (“fold”) information”. While promising, these attempts have shown limited success in generating sequences that fold to the intended structures in silico , likely due to the limited ability of the denoising networks to generate realistic protein backbones, and have not been tested experimentally.

从头设计蛋白质旨在生成具有特定结构和/或功能特性的蛋白质，例如与给定靶标的结合相互作用 ，折叠成特定拓扑结构 ，或稳定所需的功能“基序”(产生所需活性的几何形状和氨基酸特征) 。去噪扩散概率模型(DDPMs)是一类强大的机器学习模型，最近被证明能够根据文本提示生成新颖的照片级真实图像 ，具有几种非常适合蛋白质设计的特性。首先，DDPMs生成高度多样化的输出——DDPMs被训练去噪那些被高斯噪声污染的数据(例如图像或文本)；通过学习随机逆转这种污染，生成与训练数据高度相似的多样化输出。其次，DDPMs可以在每一步迭代生成过程中通过提供条件信息来引导特定的设计目标。第三，对于几乎所有的蛋白质设计应用，明确建模三维结构是必要的；SE(3)-等变DDPMs能够以与表示框架无关的方式做到这一点。最近的研究通过对小蛋白“基序”5,9或二级结构和块邻接(“折叠”)信息进行条件化，调整了DDPMs用于蛋白质单体设计。尽管前景看好，但这些尝试在生成能够在计算机中折叠成预期结构的序列方面显示出有限的成功 ，这可能是由于去噪网络生成现实蛋白质骨架的能力有限，并且尚未进行实验测试。

We reasoned that improved diffusion models for protein design could be developed by taking advantage of the deep understanding of protein structure implicit in powerful structure prediction methods like AlphaFold2 (AF2) and RoseTTAFold (RF). RF has properties particularly well suited for use in a protein design DDPM (Fig. 1A). First, RF can generate protein structures with very high precision, and in our previous work we demonstrated considerable success in accurately scaffolding motifs following fine tuning of RF for protein design ("RF Inpainting") . Second, RF operates on a rigid-frame representation of residues with rotational and translational equivariance. Third, the RF architecture enables conditioning on design specifications at three different levels: individual residue properties, pairwise distances and orientations between residues, and coordinates. In Inpainting, we fine-tuned to design protein scaffolds in a single step. Experimental characterization showed that the method can scaffold a wide range of protein functional motifs with atomic accuracy , but the approach fails on minimalist site descriptions that do not sufficiently constrain the overall fold, and because it is deterministic, can produce only a limited diversity of designs for a given problem. We reasoned that by instead fine-tuning RoseTTAFold as the denoising network in a generative diffusion model, we could overcome both problems: because the starting point is random noise, each denoising trajectory yields a different solution, and because structure is built up progressively through many denoising iterations, little to no starting structural information should be required.

我们推测，通过利用像AlphaFold2(AF2)和RoseTTAFold(RF)这样的强大结构预测方法中隐含的对蛋白质结构的深刻理解，可以开发出改进的蛋白质设计扩散模型。RF具有特别适合用于蛋白质设计DDPM(图1A)的特性。首先，RF能够以非常高的精度生成蛋白质结构，在我们之前的工作中，我们展示了在对RF进行微调以进行蛋白质设计后，准确构建基序的显著成功(“RF 修复”) 。其次，RF在具有旋转和位移等变性的残基刚性框架表示上运行。第三，RF架构能够在三个不同层次上对设计规范进行条件化:单个残基属性、残基之间的成对距离和方向，以及 坐标。在 修复中，我们对 进行了微调，以在单一步骤中设计蛋白质支架。实验表征表明，该方法能够以原子级精度构建广泛的蛋白质功能基序 ，但该方法在对整体折叠约束不足的极简位点描述上失败，并且由于其确定性，对于给定问题只能产生有限的设计多样性。我们推测，通过将RoseTTAFold微调为生成扩散模型中的去噪网络，我们可以克服这两个问题:因为起始点是随机噪声，每个去噪轨迹产生不同的解决方案，并且由于结构是通过许多去噪迭代逐步构建的，因此几乎不需要起始结构信息。

We construct a RoseTTAFold-based diffusion model, RFdiffusion, using the RF frame representation which comprises a coordinate and rigid orientation for each residue. We generate training inputs by simulating the noising process for a random number of steps (up to 200) on structures sampled from the Protein Data Bank (PDB) . For translations, we perturb coordinates with 3D Gaussian noise. For residue orientations, we use Brownian motion on the manifold of rotation matrices (building on refs ). To enable RFdiffusion to learn to reverse each step of the noising process, we train the model by minimizing a mean squared error (MSE) loss between frame predictions and the true protein structure (without alignment), averaged across all residues (Methods 2.5). This loss drives denoising trajectories to match the data distribution at each timestep and hence to converge on structures of designable protein backbones (Fig. S1A). MSE contrasts to the loss used in RF structure prediction training ("frame aligned point error", FAPE) in that unlike FAPE, MSE loss is not invariant to the global reference frame and therefore promotes continuity of the global coordinate frame between timesteps (Methods 2.5). While in this study we use RoseTTAFold as the basis for the denoising network architecture, other SE(3)-equivariant structure prediction networks (AF2 , OmegaFold , ) could in principle be substituted into an analogous DDPM.

我们构建了一个基于RoseTTAFold的扩散模型RFdiffusion，使用RF框架表示法，该表示法为每个残基包含一个 坐标和一个 刚性方向。我们通过对从蛋白质数据银行(PDB) 中采样的结构进行随机步数(最多200步)的噪声过程模拟来生成训练输入。对于坐标，我们用3D高斯噪声进行扰动。对于残基方向，我们在旋转矩阵的流形上使用布朗运动(基于参考文献 )。为了使RFdiffusion能够学习逆转每一步的噪声过程，我们通过最小化帧预测与真实蛋白质结构(不进行对齐)之间的均方误差(MSE)损失来训练模型，该损失在所有残基上进行平均(方法2.5)。该损失驱动去噪轨迹在每个时间步匹配数据分布，从而收敛到可设计的蛋白质骨架结构(图S1A)。MSE与RF结构预测训练中使用的损失(“帧对齐点误差”，FAPE)形成对比，因为与FAPE不同，MSE损失对全局参考框架不是不变的，因此促进了时间步之间全局坐标框架的连续性(方法2.5)。虽然在本研究中我们使用RoseTTAFold作为去噪网络架构的基础，但其他SE(3)-等变结构预测网络(AF2 ，OmegaFold ， )原则上可以替换为类似的DDPM。

To generate a new protein backbone, we first initialize random residue frames and RFdiffusion makes a denoised prediction. Each residue frame is updated by taking a step in the direction of this prediction with some noise added to generate the input to the next step. The nature of the noise added and the size of this reverse step is chosen such that the denoising process matches the distribution of the noising process (Methods 2.2-2.3, Figure S1A). RFdiffusion initially seeks to match the full breadth of possible protein structures compatible with the purely random frames with which it is initialized, and hence the denoised structures do not initially appear protein-like (Fig. 2A left). However, through many such steps, the breadth of possible protein structures from which the input could have arisen narrows, and RFdiffusion predictions come to closely resemble protein structures (Fig. 2A right). We use the ProteinMPNN network to subsequently design sequences encoding these structures. We also considered simultaneously designing structure and sequence within RFdiffusion, but given the excellent performance of combining ProteinMPNN with the diffusion of structure alone, we did not extensively explore this possibility.

为了生成新的蛋白质骨架，我们首先初始化随机残基框架，RFdiffusion进行去噪预测。每个残基框架通过朝着这个预测的方向迈出一步并添加一些噪声来更新，以生成下一步的输入。添加的噪声性质和这个反向步骤的大小选择得使得去噪过程与噪声过程的分布相匹配(方法2.2-2.3，图S1A)。RFdiffusion最初试图匹配与其初始化时的完全随机框架兼容的所有可能的蛋白质结构，因此去噪后的结构最初并不显得像蛋白质(图2A左)。然而，通过许多这样的步骤，输入可能来源的蛋白质结构的范围逐渐缩小，RFdiffusion的预测开始与蛋白质结构密切相似(图2A右)。我们使用ProteinMPNN网络 随后设计编码这些结构的序列。我们也考虑在RFdiffusion中同时设计结构和序列，但鉴于将ProteinMPNN与单独的结构扩散结合的优秀表现，我们没有深入探索这个可能性。

Fig. 1A highlights the similarities between RoseTTAFold structure prediction and an RF diffusion denoising step: in both cases, the networks transform coordinates into a predicted structure, conditioned on inputs to the model. In RoseTTAFold, sequence is the primary input, with additional structural information provided as templates and initial coordinates to the model. In RFdiffusion, the primary input is the noised coordinates from the previous step. For design tasks, we optionally provide a range of auxiliary conditioning information, including partial sequence, fold information, or fixed functional motif coordinates (see Methods 3). bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

图1A突出了RoseTTAFold结构预测与RF扩散去噪步骤之间的相似性:在这两种情况下，网络将坐标转换为预测结构，基于模型的输入。在RoseTTAFold中，序列是主要输入，额外的结构信息作为模板和初始坐标提供给模型。在RFdiffusion中，主要输入是来自前一步的噪声坐标。对于设计任务，我们可以选择提供一系列辅助条件信息，包括部分序列、折叠信息或固定功能基序坐标(见方法3)。bioRxiv预印本doi:https://doi.org/10.1101/2022.12.09.519842；该版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审的)是作者/资助者，他们已授予bioRxiv永久展示该预印本的许可。该文档根据CC-BY-ND 4.0国际许可证提供。

We explored two different strategies for training RFdiffusion: 1) in a manner akin to "canonical" diffusion models, with predictions at each timestep independent of predictions at previous timesteps (as in previous work ), and 2) with self-conditioning , where the model can condition on previous predictions between timesteps (Fig. 1A bottom row, Methods 2.4). The latter strategy was inspired by the success of "recycling" in AF2, which is also central to the more recent RF model used here (Methods 1). Self-conditioning within RFdiffusion dramatically improved performance on in silico benchmarks encompassing both conditional and unconditional protein design tasks (Fig. S2E). Increased coherence of predictions within self-conditioned trajectories may, at least in part, explain these performance increases (Fig. S2H). Fine-tuning RFdiffusion from pre-trained RF weights was far more successful than training for an equivalent length of time from untrained weights (Fig. S2F) and the MSE loss was also crucial (Fig. S2D). For all in silico benchmarks in this paper, we use the AF2 structure prediction network for validation and define an in silico "success" as an RFdiffusion output for which the AF2 structure predicted from a single sequence is (1) of high confidence (mean predicted aligned error, pAE,<5),(2) globally within backbone-RMSD of the designed structure, and (3) within 1 Å backbone-RMSD on any scaffolded functional-site. This definition of success is significantly more stringent than those described elsewhere (refs [5,8,16,25], Fig. S3A-B) but is a good predictor of experimental success .

我们探索了两种不同的RFdiffusion训练策略:1)类似于“标准”扩散模型的方式，在每个时间步的预测独立于之前时间步的预测(如之前的工作 )，2)自条件化 ，模型可以在时间步之间对之前的预测进行条件化(图1A底部，方法2.4)。后一种策略受到AF2中“回收”成功的启发，这在这里使用的更近期的RF模型中也占据中心地位(方法1)。在RFdiffusion中进行自条件化显著提高了在计算基准测试中的表现，包括条件和无条件蛋白质设计任务(图S2E)。自条件化轨迹中预测的一致性增加，至少在某种程度上，可以解释这些性能的提升(图S2H)。从预训练的RF权重微调RFdiffusion比从未训练的权重训练相同时间长度要成功得多(图S2F)，均方误差损失也至关重要(图S2D)。在本文的所有计算基准测试中，我们使用AF2结构预测网络 进行验证，并将计算“成功”定义为RFdiffusion输出，其中从单一序列预测的AF2结构是(1)高置信度(平均预测对齐误差，pAE，<5)，(2)在设计结构的 主链-RMSD范围内，且(3)在任何支架功能位点的主链-RMSD范围内为1 Å。这个成功的定义比其他地方描述的要严格得多(参考文献[5,8,16,25]，图S3A-B)，但它是实验成功的良好预测指标 。

# Unconditional protein monomer generation

# 无条件蛋白质单体生成

Physically-based protein design methodologies have struggled in unconstrained generation of diverse protein monomers due to the difficulty of sampling on the very large and rugged conformational landscape , and overcoming this limitation has been a primary test of deep learning based protein design approaches . As illustrated in Fig. 2B-D, Fig. S4B-C, starting from random noise, RFdiffusion can readily generate elaborate protein structures with little overall structural similarity to any known protein structures, indicating considerable generalization beyond the PDB training set. The designs are diverse (Fig. S4A), spanning a wide range of alpha-, beta- and mixed alpha-beta- topologies, with AF2 and ESMFold (Fig. S2B-C, Fig. S3A) predictions very close to the design structure models for de novo designs with as many as 600 residues. RFdiffusion generates plausible structures for even very large proteins, but these are difficult to validate in silico as they are likely beyond the single sequence prediction capabilities of AF2 and ESMFold. The quality and diversity of designs that are sampled is inherent to the model, and does not require any auxiliary conditioning input (for example secondary structure information ). Characterization of two of these 300 amino acid proteins is shown in Figure 2G, demonstrating circular dichroism (CD) spectra consistent with the mixed alpha-beta topologies of the two designs, and CD melts showing that designs are extremely thermostable. RFdiffusion strongly outperforms Hallucination (Fig. 2E), the only previously described deep learning method for unconditional protein structure generation that has been experimentally validated . Hallucination uses Monte Carlo search or gradient descent to identify sequences predicted to fold into stable structures; in contrast to RFdiffusion, Hallucination success rates deteriorate beyond 100 amino acids. RFdiffusion is also more compute efficient than unconstrained Hallucination, requiring 2̃.5 minutes on an NVIDIA RTX bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license. A4000 GPU to generate a 100 residue structure compared to 8̃.5 minutes for Hallucination. The computational efficiency of RFdiffusion can be further improved by taking larger steps at inference time, and by truncating trajectories early - an advantage of predicting the final structure at each timestep (Fig. S3C-D). For design problems where a particular fold or architecture is desired (such as TIM barrels or cavity-containing NTF2s for small molecule binder and enzyme design ), we further fine-tuned RFdiffusion to condition on secondary structure and/or fold information, enabling rapid and accurate generation of diverse designs with the desired topologies (Fig. 2H, Fig. S5). In silico success rates were 42.5% and 54.1% for TIM barrels and NTF2 folds respectively (Fig. S5G), and experimental characterization of 11 TIM barrel designs indicated that at least 9 designs were soluble, thermostable, and had circular dichroism (CD) spectra consistent with the design model (Fig. 2H, Fig. S5D-F).

基于物理的蛋白质设计方法在无约束生成多样化蛋白质单体方面面临挑战，因为在非常大且复杂的构象空间中进行采样非常困难，克服这一限制一直是基于深度学习的蛋白质设计方法的主要考验。如图2B-D和图S4B-C所示，从随机噪声开始，RFdiffusion能够轻松生成与任何已知蛋白质结构几乎没有整体结构相似性的复杂蛋白质结构，表明其在PDB训练集之外具有相当大的泛化能力。这些设计多样(图S4A)，涵盖了广泛的α、β和混合α-β拓扑，AF2和ESMFold(图S2B-C，图S3A)对新设计的结构模型的预测与设计结构模型非常接近，设计中包含多达600个残基。RFdiffusion甚至能够为非常大的蛋白质生成合理的结构，但由于这些结构可能超出了AF2和ESMFold的单序列预测能力，因此在计算机上验证这些结构是困难的。所采样设计的质量和多样性是模型固有的，不需要任何辅助条件输入(例如二级结构信息)。这300个氨基酸蛋白质中的两个的特征在图2G中展示，表明其圆二色性(CD)光谱与这两个设计的混合α-β拓扑一致，CD熔融显示这些设计具有极高的热稳定性。RFdiffusion的表现远超Hallucination，这是唯一一种经过实验验证的无条件蛋白质结构生成的深度学习方法。Hallucination使用蒙特卡洛搜索或梯度下降来识别预测折叠为稳定结构的序列；与RFdiffusion相比，Hallucination的成功率在超过100个氨基酸后下降。RFdiffusion的计算效率也优于无约束的Hallucination，生成一个100个残基的结构大约需要2.5分钟，而Hallucination则需要约8.5分钟。通过在推理时采取更大的步长和提前截断轨迹，可以进一步提高RFdiffusion的计算效率，这是在每个时间步预测最终结构的优势。对于希望获得特定折叠或结构的设计问题(例如TIM桶或含腔的NTF2用于小分子结合剂和酶设计)，我们进一步微调RFdiffusion以条件化二级结构和/或折叠信息，从而快速准确地生成具有所需拓扑的多样化设计。在计算机上的成功率分别为TIM桶和NTF2折叠的42.5%和54.1%，对11个TIM桶设计的实验特征表明至少有9个设计是可溶的、热稳定的，并且具有与设计模型一致的圆二色性(CD)光谱。

# Design of higher order oligomers

# 高阶寡聚物的设计

There is considerable interest in designing symmetric oligomers, which can serve as vaccine platforms , delivery vehicles , and catalysts . Cyclic oligomers have been designed using structure prediction networks with an adaptation of Hallucination that searches for sequences predicted to fold to the desired cyclic symmetry, but this approach fails for higher order dihedral, tetrahedral, octahedral, and icosahedral symmetries, likely in part because of the much lower representation of such structures in the .

设计对称寡聚物引起了相当大的兴趣，这些寡聚物可以作为疫苗平台 、递送载体 和催化剂 。使用结构预测网络和幻觉(Hallucination)方法设计了环状寡聚物，该方法搜索预测折叠为所需环状对称性的序列，但这种方法在高阶二面体、四面体、八面体和二十面体对称性方面失败，部分原因可能是这些结构在 中的表示非常少。

We set out to generalize RFdiffusion to create symmetric oligomeric structures with any specified point group symmetry. Given a specification of a point group symmetry for an oligomer with chains, and the monomer chain length, we generate random starting residue frames for a single monomer subunit as in the unconditional generation case, and then generate N-1 copies of this starting point arranged with the specified point group symmetry. Because RFdiffusion exhibits equivariance (inherited from RF) with respect to rotation and relabelings of chains, symmetry is largely maintained in the denoising predictions; although we explicitly re-symmetrize at each step, this changes the structures only slightly (Compare gray and colored chains in Fig. S6A, Methods Proposition 2). For octahedral and icosahedral architectures, we explicitly model only the smallest subset of monomers required to generate the full assembly (e.g. for icosahedra, the subunits at the five-fold, three-fold, and two-fold symmetry axes) to reduce the computational cost and memory footprint.

我们着手将RFdiffusion推广，以创建具有任何指定点群对称性的对称寡聚体结构。给定一个具有 链的寡聚体的点群对称性规范，以及单体链的长度，我们为单个单体亚单位生成随机起始残基框架，如同无条件生成的情况，然后生成N-1个按照指定点群对称性排列的起始点副本。由于RFdiffusion在旋转和链的重新标记方面表现出等变性(继承自RF)，因此在去噪预测中对称性基本保持；尽管我们在每一步都明确重新对称化，但这仅会稍微改变结构(比较图S6A中的灰色和有色链，方法命题2)。对于八面体和二十面体结构，我们仅明确建模生成完整组装所需的最小单体子集(例如，对于二十面体，在五重、三重和二重对称轴上的子单位)，以降低计算成本和内存占用。

Despite not being trained on symmetric inputs, RFdiffusion is able to generate symmetric oligomers with high in silico success rates (Fig. S6B), particularly when guided by an auxiliary inter- and intra-chain contact potential (Fig. S6C). As illustrated in Fig. 3 and Fig. S6E, RFdiffusion-generated designs are nearly indistinguishable from AF2 predictions of the structures adopted by the designed sequences (predictions of the full assemblies for the cyclic and dihedral designs, and trimeric substructures of the octahedral and icosahedral designs), and show little resemblance to proteins in the PDB (Fig. S6D). These include a number of oligomeric topologies not seen in nature, including two-layer beta strand barrels (Fig. 3A, C10 symmetry) and complex mixed alpha/beta topologies (Fig. 3A, C8 symmetry) (closest TM align in PDB: 6BRP, 0.47; 6BRO, 0.43 respectively). bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

尽管没有针对对称输入进行训练，RFdiffusion仍能够生成具有高计算成功率的对称寡聚体(图S6B)，特别是在辅助的链间和链内接触势的指导下(图S6C)。如图3和图S6E所示，RFdiffusion生成的设计与AF2对设计序列所采用结构的预测几乎无法区分(对循环和二面体设计的完整组装的预测，以及八面体和二十面体设计的三聚体亚结构)，并且与PDB中的蛋白质几乎没有相似之处(图S6D)。这些包括一些在自然界中未见的寡聚体拓扑结构，包括双层β链桶(图3A，C10对称性)和复杂的混合α/β拓扑结构(图3A，C8对称性)(在PDB中最接近的TM比对:6BRP，0.47；6BRO，0.43)。bioRxiv预印本doi:https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审的)是作者/资助者，已授予bioRxiv永久展示该预印本的许可。该文档根据CC-BY-ND 4.0国际许可协议提供。

We selected 608 designs for experimental characterization, and found using size exclusion chromatography (SEC) that at least 70 had oligomerization states closely consistent with the design models (within the 95% CI, 109 designs within the 99% CI, as determined by SEC calibration curves) (Fig. S11, S12). We collected negative stain electron microscopy (nsEM) data on a subset of these designs across different symmetry groups and, for most, distinct particles were evident with shapes resembling the design models in both the raw micrographs and subsequent 2D classifications (Fig. 3, and Fig. S6F). We describe these designs in the following paragraphs; most have structures that are, to our knowledge, unprecedented in nature.

我们选择了608个设计进行实验表征，并通过尺寸排除色谱法(SEC)发现至少有70个的聚合状态与设计模型密切一致(在95%置信区间内，109个设计在99%置信区间内，依据SEC校准曲线确定)(图S11，S12)。我们对这些设计中的一部分在不同对称组中收集了负染色电子显微镜(nsEM)数据，对于大多数设计，明显可见的颗粒形状与设计模型相似，既在原始显微照片中，也在后续的二维分类中(图3和图S6F)。我们在以下段落中描述这些设计；据我们所知，大多数具有前所未有的自然结构。

As described above, RFdiffusion was able to unconditionally generate a wide range of monomer structures, and while AF2 predictions and circular dichroism measurements were consistent with the design models, the structures were too small for electron microscopy-based structure validation. We took advantage of the increase in size upon oligomerization to evaluate, using electron microscopy, unconstrained structure generation for oligomers with subunits over 175 amino acids in length (Fig. 3B, top row). Electron microscopy characterization of a C3 design (HE0822) with 350 residue subunits (1050 residues in total) suggests that the actual structure is very close to the design, both over the 350 residue subunits and the overall C3 architecture. 2D class averages are clearly consistent with both top- and side-views of the design model, and a 3D reconstruction of the density had key features consistent with the design, including the distinctive pinwheel shape. Electron microscopy 2D class averages of C5 and C6 designs with greater than 750 residues (HE0795, HE0789, HE0841) were also consistent with the respective design models (Fig. S6F).

如上所述，RFdiffusion能够无条件地生成多种单体结构，尽管AF2预测和圆二色性测量与设计模型一致，但这些结构对于基于电子显微镜的结构验证来说过于微小。我们利用寡聚化时尺寸的增加，通过电子显微镜评估了长度超过175个氨基酸的亚单位的无约束结构生成(图3B，顶部行)。对具有350个残基亚单位(总共1050个残基)的C3设计(HE0822)的电子显微镜表征表明，实际结构与设计非常接近，无论是在350个残基的亚单位上还是整体C3架构上。2D类平均值与设计模型的顶部和侧面视图明显一致，密度的3D重建具有与设计一致的关键特征，包括独特的风车形状。C5和C6设计(HE0795、HE0789、HE0841)中超过750个残基的电子显微镜2D类平均值也与各自的设计模型一致(图S6F)。

RFdiffusion also generated cyclic oligomers with alpha/beta barrel structures that resemble expanded TIM barrels and provide an interesting comparison between innovation during natural evolution and innovation through deep learning. The TIM barrel fold, with 8 strands and 8 helices, is one of the most abundant folds in nature . Electron microscopy characterization validated two RFdiffusion cyclic oligomers which considerably extend beyond this fold (Fig. 3B, bottom rows). HE0626 is a C6 alpha/beta barrel composed of 18 strands and 18 helices, and HE0675 is a C8 octamer composed of an inner ring of 16 strands and an outer ring of 16 helices arranged locally in a very similar repeating pattern to the TIM barrel (1:1 helix:strand). By nsEM, we observed 2D class averages for HE0626 that resemble this two ring organization, and for both HE0626 and HE0675 we were able to obtain 3D reconstructions that are in agreement with the computational design models. The HE0600 design is also an alpha-beta barrel (Fig. S6F), but has two strands for every helix (24 strands and 12 helices in total) and is hence locally quite different from a TIM barrel. Whereas natural evolution has extensively explored structural variations of the classic 8-strand/8-helix TIM barrel fold, RFdiffusion can more readily explore global changes in barrel curvature, enabling discovery of TIM barrel-like structures with many more helices and strands.

RFdiffusion还生成了具有α/β桶结构的环状寡聚物，这些结构类似于扩展的TIM桶，并提供了自然进化中的创新与深度学习中的创新之间有趣的比较。TIM桶折叠结构由8条链和8个螺旋组成，是自然界中最丰富的折叠结构之一 。电子显微镜表征验证了两个RFdiffusion环状寡聚物，它们的结构明显超出了这一折叠(图3B，底部行)。HE0626是一个C6 α/β桶，由18条链和18个螺旋组成，而HE0675是一个C8八聚体，由内环的16条链和外环的16个螺旋组成，这些螺旋在局部以与TIM桶非常相似的重复模式排列(1:1螺旋:链)。通过nsEM，我们观察到HE0626的二维类平均图像，类似于这种双环结构，对于HE0626和HE0675，我们能够获得与计算设计模型一致的三维重建。HE0600设计也是一个α-β桶(图S6F)，但每个螺旋有两条链(总共24条链和12个螺旋)，因此在局部上与TIM桶有很大不同。虽然自然进化广泛探索了经典8条链/8个螺旋TIM桶折叠的结构变异，RFdiffusion则可以更容易地探索桶曲率的全球变化，从而发现具有更多螺旋和链的TIM桶样结构。

RFdiffusion readily generated structures with dihedral and tetrahedral symmetries (Fig. 3C, Fig. S6E, F). SEC characterization indicated that 38 D2, 7 D3, and 3 D4 designs had the expected bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

RFdiffusion 轻松生成了具有二面角和四面体对称性的结构(图 3C，图 S6E, F)。SEC 表征表明，38 个 D2、7 个 D3 和 3 个 D4 设计具有预期的 bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842；该版本于 2022 年 12 月 10 日发布。该预印本的版权持有者(未经过同行评审认证)是作者/资助者，已授予 bioRxiv 永久展示该预印本的许可。该文档根据 CC-BY-ND 4.0 国际许可协议提供。

molecular weights (these have 4, 6, and 8 chains, respectively) (Fig. S12). While the D2 dihedrals are too small for nsEM, 2D class averages—and for some, 3D reconstructions— of D3 and D4 designs were congruent with the overall topologies of the design models (Fig. 3C, Fig. S6F). The reconstruction for the D3 HE0490 shows the characteristic triangular shape of the design. Similarly, the 3D reconstruction of the D4 HE0537 closely matches the design model, recapitulating the approximate offset between tetramic subunits. We were also able to obtain cryogenic electron microscopy (cryo-EM) data for HE0537; however, a preferred orientation precluded generation of a reliable 3D reconstruction for in-depth structural analysis. Nonetheless, the resulting 2D class averages bear a striking level of secondary-structure similarity to generated 2D projections of the corresponding design model (Fig. 3D). 2D class averages for a 12 chain tetrahedron (HE0964) were consistent with the design, but we were unable to generate a 3D reconstruction of high confidence due to a lack of clear discernable design features visible at the resolution range provided by nsEM (Fig. S6F).

分子量(分别为4、6和8链)(图S12)。虽然D2二面角对于nsEM来说太小，但D3和D4设计的2D类平均值——对于某些设计，还有3D重建——与设计模型的整体拓扑结构一致(图3C，图S6F)。D3 HE0490的重建显示了设计的特征性三角形状。同样，D4 HE0537的3D重建与设计模型紧密匹配，重现了四聚体亚单位之间的近似 偏移。我们还能够获得HE0537的低温电子显微镜(cryo-EM)数据；然而，优先取向阻碍了生成可靠的3D重建以进行深入的结构分析。尽管如此，生成的2D类平均值与相应设计模型的2D投影在二级结构上表现出惊人的相似性(图3D)。12链四面体(HE0964)的2D类平均值与设计一致，但由于在nsEM提供的分辨率范围内缺乏明显可辨识的设计特征，我们无法生成高可信度的3D重建(图S6F)。

Icosahedra have 60 subunits arrayed around 2-fold, 3-fold and 5-fold symmetry axes. Of the 48 icosahedra selected for experimental validation, one was confirmed by nsEM to form the intended assembly. As shown in Fig. 3E on the left, HE0902 is a 15nm (diameter) highly-porous icosahedron composed of alpha helical subunits. The nsEM micrographs reveal highly homogeneous particles, and the corresponding 2D class averages and 3D reconstruction nearly perfectly match the design model (Fig. 3E), with triangular hubs arrayed around the empty C5 axes. Designs such as HE0902 (and future similar large assemblies) should be useful as new nanomaterials and vaccine scaffolds, with robust assembly and (in the case of HE0902) the outward facing N- and C-termini offering multiple possibilities for antigen display.

二十面体有60个子单元围绕2重、3重和5重对称轴排列。在为实验验证选择的48个二十面体中，有一个通过nsEM确认形成了预期的组装。如左侧图3E所示，HE0902是一个直径为15纳米的高孔隙率二十面体，由α螺旋子单元组成。nsEM显微照片显示出高度均匀的颗粒，相应的二维类别平均和三维重建几乎完美匹配设计模型(图3E)，三角形中心围绕空的C5轴排列。像HE0902这样的设计(以及未来类似的大型组装)应作为新的纳米材料和疫苗支架，具有稳健的组装能力，并且(在HE0902的情况下)外向的N端和C端提供了多种抗原展示的可能性。

# Functional motif scaffolding

# 功能性基序支架

We next investigated the use of RFdiffusion for scaffolding protein structural motifs that carry out binding and catalytic functions, where the role of the scaffold is to hold the motif in precisely the 3D geometry needed for optimal function. In RFdiffusion, we input motifs as 3D coordinates (including sequence and sidechains) both during conditional training and inference, and RFdiffusion builds scaffolds that hold the motif atomic coordinates in place. A number of deep learning methods have been developed recently to address this problem, including Inpainting , constrained Hallucination , and other DDPMs . To rigorously evaluate the performance of these methods in comparison to RFdiffusion across a broad set of design challenges, we established an in silico benchmark test comprising 25 motif-scaffolding design problems addressed in six recent publications encompassing several design methodologies . The benchmark includes 25 challenges that span a broad range of motifs, including simple "inpainting" problems, viral epitopes, receptor traps, small molecule binding sites, binding interfaces and enzyme active sites. Full details of this benchmark set are described in Supplementary Table 9.

我们接下来研究了RFdiffusion在支架蛋白结构基序中的应用，这些基序执行结合和催化功能，支架的作用是将基序保持在所需的最佳功能的精确三维几何形状中。在RFdiffusion中，我们在条件训练和推理过程中将基序作为三维坐标(包括序列和侧链)输入，RFdiffusion构建支架以固定基序的原子坐标。最近开发了多种深度学习方法来解决这个问题，包括 修复 、受限幻觉 和其他DDPMs 。为了严格评估这些方法与RFdiffusion在广泛设计挑战中的性能比较，我们建立了一个包含25个基序支架设计问题的计算基准测试，这些问题在六篇最近的出版物中得到了解决，涵盖了几种设计方法 。该基准包括25个挑战，涵盖了广泛的基序，包括简单的“修复”问题、病毒表位、受体陷阱、小分子结合位点、结合界面和酶活性位点。该基准集的完整细节在补充表9中描述。

RFdiffusion solves all but two of the 25 benchmark problems, with greater success (23/25) than both Hallucination (15/25) and Inpainting (12/25) (Fig. 4A-B). For 22/23 of the problems solved by RFdiffusion, it also has a higher fraction of successful designs than either bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made

RFdiffusion解决了25个基准问题中的所有问题，除了两个，其成功率(23/25)高于Hallucination(15/25)和 Inpainting(12/25)(图4A-B)。在RFdiffusion解决的22/23个问题中，它的成功设计比例也高于bioRxiv预印本doi: https://doi.org/10.1101/2022.12.09.519842；该版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审认证)是作者/资助者，他们已授予bioRxiv永久展示该预印本的许可。它被制作

Hallucination or Inpainting. The excellent performance of RFdiffusion required no hyperparameter tuning or external potentials; this contrasts with Hallucination, for which problem-specific optimization can be required. In 17/23 of the problems, RFdiffusion generated successful solutions with higher success rates when noise was not added during the reverse diffusion trajectories (see Fig. S2I for further discussion of the effect of noise on design quality).

幻觉或 修复。RFdiffusion的卓越表现无需超参数调整或外部势能；这与幻觉形成对比，后者可能需要特定问题的优化。在23个问题中，有17个问题在反向扩散轨迹中未添加噪声时，RFdiffusion生成了成功的解决方案，成功率更高(有关噪声对设计质量影响的进一步讨论，请参见图S2I)。

One of the benchmark problems is the scaffolding of the p53 helix that binds MDM2. Inhibiting this interaction through high-affinity competitive inhibition by scaffolding the p53 helix and making additional interactions with MDM2 is a promising avenue for therapeutics . In silico success has been described elsewhere , but experimental success has not been reported. We tested 96 designs scaffolding this helix, which were predicted to make additional interactions with MDM2, and identified 0.5nM and 0.7nM binders (Fig. 4C-D), three orders of magnitude higher affinity than the reported affinity of the p53 peptide alone . The success rate for this problem was particularly striking with 55/95 designs showing some detectable binding at 10μM (Fig. 4E) and multiple designs with affinities in the low- to sub- nanomolar range (Fig. 4D).

其中一个基准问题是支架化结合MDM2的p53螺旋。通过支架化p53螺旋并与MDM2进行额外相互作用来抑制这种相互作用的高亲和力竞争抑制是一条有前景的治疗途径 。在计算机模拟中取得的成功在其他地方已有描述 ，但实验成功尚未报道。我们测试了96种设计，这些设计支架化了该螺旋，并预测能够与MDM2进行额外相互作用，识别出0.5nM和0.7nM的结合剂(图4C-D)，其亲和力比报道的p53肽单独的亲和力 高出三个数量级 。这个问题的成功率尤其引人注目，55/95种设计在10μM时显示出可检测的结合(图4E)，并且有多种设计的亲和力在低至亚纳摩尔范围内(图4D)。

# Scaffolding enzyme active sites

# 支架酶活性位点

A grand challenge in protein design is to scaffold minimal descriptions of enzyme active sites comprising a few single amino acids. While some in silico success has been reported previously , a general solution that can readily produce high-quality, orthogonally-validated outputs remains elusive. Following fine-tuning on a task mimicking this problem (Methods 4.2), RFdiffusion was able to scaffold enzyme active sites comprising multiple sidechain and backbone functional groups with high accuracy and in silico success rates across a range of enzyme classes (Fig. 4F-H). While RFdiffusion is currently unable to explicitly model bound small molecules (see conclusion), the substrate can be implicitly modeled using an external potential to guide the generation of "pockets" around the active site. As a demonstration, we scaffold a retroaldolase active site triad while implicitly modeling its substrate (Fig. S7).

蛋白质设计中的一个重大挑战是构建包含少量单个氨基酸的酶活性位点的最小描述。虽然之前已有一些计算成功的报道 ，但能够轻松产生高质量、正交验证输出的一般解决方案仍然难以实现。在模拟这一问题的任务上进行微调后(方法4.2)，RFdiffusion能够以高精度构建包含多个侧链和主链功能基团的酶活性位点，并在多种酶类别中实现计算成功率(图4F-H)。虽然RFdiffusion目前无法明确建模结合的小分子(见结论)，但可以使用外部势能隐式建模底物，以引导在活性位点周围生成“口袋”。作为演示，我们在隐式建模其底物的同时构建了一个逆醛缩酶活性位点三元组(图S7)。

# Symmetric functional-motif scaffolding for metal coordinating assemblies and antiviral therapeutics and vaccines

# 对金属配位组装体及抗病毒治疗和疫苗的对称功能基元支架

A number of important design challenges involve the scaffolding of multiple copies of a functional motif in symmetric arrangements. For example, many viral glycoproteins are trimeric, and symmetry matched arrangements of inhibitory domains can be extremely potent . Conversely, symmetric presentation of viral epitopes in an arrangement that mimics the virus could induce new classes of neutralizing antibodies . To explore this general direction, we sought to design trimeric multivalent binders to the SARS-CoV-2 spike protein. In previous work, flexible linkage of a binder to the ACE2 binding site (on the spike protein receptor binding domain) to a trimerization domain yielded a high-affinity inhibitor that had potent and broadly neutralizing antiviral activity in animal models . Ideally, however, symmetric fusions to binders would be rigid, so as to reduce the entropic cost of binding while maintaining the avidity benefits from multivalency. We used RFdiffusion to design C3 symmetric trimers which rigidly hold three binding domains (the functional motif in this case) such that they exactly match the ACE2 bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license. binding sites on the SARS-CoV-2 spike protein trimer. Design models were confidently predicted by AF2 to both assemble as C3-symmetric oligomers, and to scaffold the AHB2 SARS-CoV-2 binder interface with high accuracy (Fig. 5A).

许多重要的设计挑战涉及在对称排列中支架多个功能基元的复制。例如，许多病毒糖蛋白是三聚体的，对称匹配的抑制域排列可以极其有效。相反，病毒表位的对称呈现以模仿病毒的方式排列可能会诱导新类型的中和抗体。为了探索这一总体方向，我们试图设计针对SARS-CoV-2刺突蛋白的三聚体多价结合物。在之前的工作中，将结合物灵活连接到ACE2结合位点(在刺突蛋白受体结合域上)与三聚化域结合，产生了一种具有强效和广泛中和抗病毒活性的高亲和力抑制剂。在理想情况下，结合物的对称融合应是刚性的，以减少结合的熵成本，同时保持多价性带来的亲和力优势。我们使用RFdiffusion设计了C3对称三聚体，刚性地固定三个结合域(在这种情况下是功能基元)，使其完全匹配SARS-CoV-2刺突蛋白三聚体上的ACE2结合位点。设计模型被AF2自信地预测为组装成C3对称寡聚体，并以高精度支架AHB2 SARS-CoV-2结合物界面。

The ability to scaffold functional sites with any desired symmetry opens up new approaches to designing metal-coordinating protein assemblies. Divalent transition metal ions exhibit distinct preferences for specific coordination geometries (e.g., square planar, tetrahedral, and octahedral) with ion-specific optimal sidechain-metal bond lengths. RFdiffusion provides a general route to building up symmetric protein assemblies around such sites, with the symmetry of the assembly matching the symmetry of the coordination geometry. As a first test, we sought to design square planar binding sites. We designed protein assemblies with four central histidine imidazoles arranged in an ideal -binding site with square planar coordination geometry. Diverse designs starting from various different C4-symmetric histidine square planar sites (Fig. 5B, Fig. S8A, B, C) had good in silico success (Fig. S8D), with the histidine residues in near ideal geometries for coordinating metal in the AF2 predicted structures (Fig. 5E Fig. S8B, C, E, F).

能够以任何所需的对称性构建功能性位点，为设计金属配位蛋白组装开辟了新的方法。二价过渡金属离子对特定配位几何形状(例如，平面正方形、四面体和八面体)表现出明显的偏好，并具有离子特定的最佳侧链-金属键长。RFdiffusion提供了一种通用的方法，可以围绕这些位点构建对称的蛋白组装，组装的对称性与配位几何形状的对称性相匹配。作为第一次测试，我们试图设计平面正方形 结合位点。我们设计了 蛋白组装，其中四个中心组氨酸咪唑以理想的 结合位点排列，具有平面正方形的配位几何形状。从各种不同的C4对称组氨酸平面正方形位点(图5B，图S8A, B, C)出发的多样化设计在计算机模拟中取得了良好的成功(图S8D)，在AF2预测结构中，组氨酸残基的几何形状接近理想状态，以协调金属(图5E 图S8B, C, E, F)。

We expressed and purified 44 designs in E. coli., and found that 37 had SEC chromatograms consistent with the intended oligomeric state (Fig. S9B). 36 of these designs were tested for coordination by isothermal titration calorimetry. 18 designs bound with dissociation constants ranging from low nanomolar to low micromolar (Fig. 5C, E and Fig. S9A). The inflection points in the wild-type isotherms indicate binding with the designed stoichiometry, a 1:4 ratio of ion:monomer. While most of the designed proteins displayed exothermic metal coordination, in a few cases binding was endothermic (Fig. 5E, left, Fig. S9A), suggesting that coordination is entropically driven in these assemblies. To confirm that binding was indeed mediated by the scaffolded histidine 52, we mutated this residue to alanine, which abolished or dramatically reduced binding in all cases (Fig. S9A, C and Fig. 5C, E) . We structurally characterized by nsEM a subset of the designs - E1, C10, G3, and A5 - that displayed histidine-dependent binding. All four designs exhibited clear 4-fold symmetry both in the raw micrographs and in 2D class averages (Fig. 5D, E), with design E1 also clearly displaying 2-fold axis "side-views" with a measured diameter approximating the design model. A 3D reconstruction of E1 was in close agreement to the design model (Fig. 5D).

我们在大肠杆菌中表达并纯化了44个设计，发现其中37个的SEC色谱图与预期的聚合状态一致(图S9B)。对这36个设计进行了等温滴定热量法测试以评估 的配位情况。18个设计与 结合，解离常数范围从低纳摩尔到低微摩尔(图5C、E和图S9A)。野生型等温线中的拐点表明以设计的化学计量比结合，离子与单体的比例为1:4。虽然大多数设计的蛋白质表现出放热的金属配位，但在少数情况下，结合是吸热的(图5E，左，图S9A)，这表明在这些组装中 的配位是由熵驱动的。为了确认 的结合确实是由支架组氨酸52介导的，我们将该残基突变为丙氨酸，这在所有情况下都消除了或显著降低了结合(图S9A、C和图5C、E)。我们通过纳秒电子显微镜(nsEM)对一部分显示组氨酸依赖性结合的设计进行了结构表征 - E1、C10、G3和A5。所有四个设计在原始显微照片和二维类平均中均表现出明显的4倍对称性(图5D、E)，设计E1还清晰地显示出2倍轴的“侧视图”，测得的直径接近设计模型。E1的三维重建与设计模型非常吻合(图5D)。

# Design of de novo protein-binding proteins

# 新型蛋白质结合蛋白的设计

The design of high-affinity binders to target proteins is a grand challenge in protein design, with numerous therapeutic applications . A general method to de novo design binders to protein binders from target structure information alone using the physically-based Rosetta method was recently described . Subsequently, utilizing ProteinMPNN for sequence design and AF2 for design filtering was found to improve design success rates . However, experimental success rates were low, requiring many thousands of designs to be screened for each design campaign , and the approach relied on pre-specifying a particular set of protein scaffolds as the basis for the designs, inherently limiting the diversity and shape complementarity of possible solutions . To our knowledge, no deep-learning method has yet demonstrated experimental general success in designing completely de novo binders. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made

针对蛋白质的高亲和力结合物的设计是蛋白质设计中的一项重大挑战，具有众多治疗应用 。最近描述了一种基于物理的Rosetta方法，仅利用目标结构信息从头设计结合物的通用方法 。随后，利用ProteinMPNN进行序列设计和AF2进行设计筛选被发现可以提高设计成功率 。然而，实验成功率较低，每次设计活动需要筛选数千个设计 ，而且该方法依赖于预先指定一组特定的蛋白质支架作为设计的基础，固有地限制了可能解决方案的多样性和形状互补性 。据我们所知，目前尚无深度学习方法在设计完全从头的结合物方面展示出实验上的普遍成功。bioRxiv预印本doi:https://doi.org/10.1101/2022.12.09.519842；该版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审认证)是作者/资助者，已授予bioRxiv永久展示该预印本的许可。它被制作

We reasoned that RFdiffusion might be able to address this challenge by directly generating binding proteins in the context of the target. For many therapeutic applications, for example blocking a protein-protein interaction, it is desirable to bind to a particular site on a target protein. To enable this, we fine-tuned RFdiffusion on protein complex structures, providing as input a subset of the residues on the target chain (called "interface hotspots") to which the diffused chain binds (Fig. 6A, Fig. S10A, B). To enable control over binder scaffold topology, we fine-tuned an additional model to condition binder diffusion on secondary structure and block-adjacency information, in addition to conditioning on interface hotspots (Fig. S10C-D, Methods 4.3).

我们推测RFdiffusion可能通过在目标的背景下直接生成结合蛋白来解决这一挑战。例如，对于许多治疗应用，如阻断蛋白-蛋白相互作用，理想情况下希望能够结合到目标蛋白的特定位点。为此，我们在蛋白质复合物结构上对RFdiffusion进行了微调，输入了目标链上一部分残基(称为“界面热点”)，以便扩散链能够结合(图6A，图S10A, B)。为了控制结合体支架的拓扑结构，我们还微调了一个额外的模型，以在界面热点的基础上，结合二级结构和块邻接信息来调节结合体的扩散(图S10C-D，方法4.3)。

To compare RFdiffusion to previous binder design methods, we performed binder design campaigns against 4 targets: Influenza A H1 Hemagglutinin (HA) (HA), Interleukin-7 Receptor-a (IL-7Ra) , Programmed Death-Ligand 1 (PD-L1) , and Tropomyosin Receptor Kinase A (TrkA) (we also designed against the insulin receptor, and although these binder designs expressed solubly and monomerically (Fig. S15), we were not able to obtain suitable target receptor for experimental testing). We designed putative binders to each target, both with and without conditioning on compatible fold information, with high success rates (Fig. S10E, F). Designs were filtered by AF2 confidence in the interface and monomer structure , and 95 were selected for each target for experimental characterization.

为了将RFdiffusion与之前的结合剂设计方法进行比较，我们针对4个靶标进行了结合剂设计活动:流感A H1血凝素(HA)、白介素-7受体α(IL-7Ra)、程序性死亡配体1(PD-L1)和肌动蛋白受体激酶A(TrkA)(我们还设计了针对胰岛素受体的结合剂，尽管这些结合剂设计以可溶性和单体形式表达(图S15)，但我们未能获得适合实验测试的靶标受体)。我们为每个靶标设计了假定结合剂，既有基于兼容折叠信息的设计，也有不基于此信息的设计，成功率很高(图S10E, F)。设计通过AF2在界面和单体结构上的信心进行筛选，最终为每个靶标选择了95个进行实验表征。

The designed binders were expressed in . coli and purified, and binding was assessed through single point biolayer interferometry (BLI) screening at binder (Fig. 6B). In each case a positive control was included that binds to the site targeted by the designs on the target protein . The overall success rate, as defined by binding at at or above of the maximal response for the positive control, was 18% (this is a conservative estimate as some designs which showed binding had insufficient material to permit screening at 10μM (Fig. 6B)). This is a success rate increase of approximately 2 orders-of-magnitude over our previous Rosetta-based method on the same targets (Fig. 6C). Binders were identified for all 4 targets, with fewer than 100 designs tested per target compared to thousands in previous studies. Full BLI titrations for a subset of the designs showed moderate to high affinities with no further experimental optimization, including HA and IL-7Ra binders with affinities of approximately (Fig. 6D). To assess binder specificity, 6 of the highest affinity IL-7Ra binders were assessed via competition BLI, and all 6 competed for binding with the structurally validated positive control (Fig. S14).

设计的结合剂在 大肠杆菌中表达并纯化，通过在 结合剂处进行单点生物层干涉仪(BLI)筛选来评估结合(图6B)。在每种情况下，均包含一个阳性对照，该对照与目标蛋白 上设计所针对的位点结合。总体成功率定义为在 处的结合达到或超过阳性对照最大反应的 ，为18%(这是一个保守估计，因为一些显示结合的设计由于材料不足，无法在10μM下进行筛选(图6B))。与我们之前基于Rosetta的方法相比，这一成功率提高了大约两个数量级，针对相同的目标(图6C)。所有4个目标均识别出结合剂，每个目标测试的设计少于100个，而之前的研究中则有数千个。对部分设计的完整BLI滴定显示出中等到高的亲和力，无需进一步的实验优化，包括HA和IL-7Ra结合剂，其亲和力约为 (图6D)。为了评估结合剂的特异性，对6个最高亲和力的IL-7Ra结合剂进行了竞争BLI评估，所有6个结合剂均与结构验证的阳性对照竞争结合(图S14)。

# Conclusion

# 结论

RFdiffusion is a major improvement over current physically-based and deep learning protein design methods over a wide range of design challenges. Substantial progress was recently made using Rosetta in designing binding proteins from target structural information alone , but this required testing tens of thousands of designs. RFdiffusion achieves experimental success rates that are two orders of magnitude higher. Consequently, high affinity binders (at least to the targets experimentally characterized here) can be identified through testing only dozens of bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license. designs. In the accompanying paper (Vázquez Torres et al.), we demonstrate the ability of RFdiffusion to design picomolar affinity binders to flexible helical peptides, further highlighting the utility of RFdiffusion for de novo binder design. Vázquez Torres et al. also show that RFdiffusion can be used to improve upon starting designs by partial noising and denoising, which enables tunable sampling around a given input structure. For peptide binder design, this enabled increases in affinity of nearly three orders of magnitude, without high-throughput screening of designs.

RFdiffusion在当前基于物理和深度学习的蛋白质设计方法中是一项重大改进，适用于广泛的设计挑战。最近，使用Rosetta从目标结构信息中设计结合蛋白质取得了显著进展，但这需要测试数万个设计。RFdiffusion的实验成功率高出两个数量级。因此，可以通过仅测试几十个bioRxiv预印本doi: https://doi.org/10.1101/2022.12.09.519842来识别高亲和力结合物(至少对于这里实验表征的目标)。该预印本的版权持有者(未经过同行评审的)是作者/资助者，他们已授予bioRxiv永久展示该预印本的许可。该文档根据CC-BY-ND 4.0国际许可证提供。在随附的论文中(Vázquez Torres等)，我们展示了RFdiffusion设计皮摩尔亲和力结合物对灵活螺旋肽的能力，进一步突显了RFdiffusion在新型结合物设计中的实用性。Vázquez Torres等还表明，RFdiffusion可以通过部分噪声和去噪声来改进初始设计，从而实现围绕给定输入结构的可调采样。在肽结合物设计中，这使得亲和力提高了近三个数量级，而无需对设计进行高通量筛选。

There has been recent progress in scaffolding protein functional motifs using deep learning methods (Hallucination, Inpainting, and diffusion), but Hallucination becomes very slow for large systems, inpainting fails when insufficient starting information is provided, and previous diffusion methods had quite low accuracy. Our benchmark tests show that RFdiffusion considerably outperforms all previous methods in the complexity of the motifs that can be scaffolded, the ability to precisely position sidechains (for catalysis and other functions), and the accuracy of motif recapitulation by AF2. The robust design of MDM2 binding proteins with three orders of magnitude higher binding affinities than the scaffolded P53 motif experimentally demonstrates the power of RFdiffusion for motif scaffolding.

最近在使用深度学习方法(幻觉、 修复和扩散)构建蛋白质功能基序方面取得了进展，但对于大型系统，幻觉变得非常缓慢，当提供的起始信息不足时，修复失败，而以前的扩散方法的准确性相当低。我们的基准测试表明，RFdiffusion在可以构建的基序复杂性、精确定位侧链(用于催化和其他功能)的能力以及AF2对基序的再现准确性方面显著优于所有以前的方法。MDM2结合蛋白的稳健设计，其结合亲和力比构建的P53基序高出三个数量级，实验证明了RFdiffusion在基序构建中的强大能力。

For the classic unconstrained protein structure generation problem, RFdiffusion readily generates novel protein structures with as many as 600 residues that are accurately predicted by AF2 (and ESMFold), far exceeding the complexity and accuracy achieved by previously described diffusion and other methods. Experimental data demonstrate that designs express solubly, with CD spectra consistent with the design models. That the designs are also extremely thermostable also shows that RFdiffusion designs retain the desirable ideality and stability of previous de novo design methods, while achieving considerably increased complexity. The versatility and control provided by diffusion models enabled extension of RFdiffusion unconditional generation to higher order architectures with any desired symmetry (Hallucination methods are primarily limited to cyclic symmetries); experimental characterization of a subset of these designs using electron microscopy revealed structures very similar to the design models and largely without precedent in nature. Combining the accurate motif scaffolding with the ability to design symmetric assemblies, we were able to scaffold functional motifs spanning multiple symmetrically arranged chains.

对于经典的无约束蛋白质结构生成问题，RFdiffusion能够轻松生成多达600个残基的新型蛋白质结构，这些结构的预测准确性得到了AF2(和ESMFold)的验证，远远超过了之前描述的扩散和其他方法所达到的复杂性和准确性。实验数据表明，设计的蛋白质能够以可溶形式表达，其圆二色谱(CD谱)与设计模型一致。这些设计的极高热稳定性也表明，RFdiffusion设计保留了之前从头设计方法所期望的理想性和稳定性，同时实现了显著增加的复杂性。扩散模型提供的多样性和控制能力使得RFdiffusion的无条件生成能够扩展到具有任何所需对称性的高阶结构(幻觉方法主要限于循环对称性)；使用电子显微镜对这些设计的一个子集进行实验表征，揭示了与设计模型非常相似的结构，并且在自然界中几乎没有先例。结合准确的基序支架与设计对称组装的能力，我们能够支架跨越多个对称排列链的功能基序。

Overall, the complexity of the problems solvable with RFdiffusion and the robustness and accuracy of the solutions (extensively validated both in silico and experimentally) far exceeds what has been achieved previously. In a manner reminiscent of the generation of images from text prompts, RFdiffusion makes possible, with minimal specialist knowledge, the generation of proteins from very simple molecular specifications (for example, from a specification of a target protein, high affinity binders to that protein, and from specification of a desired symmetry, diverse protein assemblies with that symmetry).

总体而言，RFdiffusion能够解决的问题的复杂性以及其解决方案的稳健性和准确性(在计算机模拟和实验中均经过广泛验证)远远超过了之前的成就。RFdiffusion以一种类似于从文本提示生成图像的方式，使得在最少的专业知识下，从非常简单的分子规格生成蛋白质成为可能(例如，从目标蛋白质的规格、高亲和力结合物的规格，以及从期望对称性的规格，生成具有该对称性的多样化蛋白质组装)。

The power and scope of RFdiffusion can be extended in several directions. RF has recently been extended to nucleic acids and protein-nucleic acid complexes , which should enable RFdiffusion to design nucleic acid binding proteins, and perhaps folded RNA structures. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

RFdiffusion的能力和范围可以在多个方向上扩展。RF最近已扩展到核酸和蛋白质-核酸复合物，这应该使RFdiffusion能够设计核酸结合蛋白，甚至可能设计折叠的RNA结构。bioRxiv预印本doi: https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审的)是作者/资助者，他们已授予bioRxiv永久展示该预印本的许可。该文档根据CC-BY-ND 4.0国际许可证提供。

Extension of RF to incorporate ligands should similarly enable extension of RFdiffusion to explicitly model ligand atoms, allowing the design of protein-ligand interactions. The ability to customize RFdiffusion to specific design challenges by addition of external potentials and by fine-tuning (as illustrated here for catalytic site scaffolding, binder-targeting and fold-specification), along with continued improvements to the underlying methodology, should enable protein design to achieve still higher levels of complexity, to approach and - in some cases - surpass what natural evolution has achieved.

扩展射频(RF)以纳入配体应同样能够扩展射频扩散(RFdiffusion)，以明确建模配体原子，从而允许设计蛋白质-配体相互作用。通过添加外部势能和微调(如这里所示的催化位点支架、结合物靶向和折叠规范)，定制射频扩散以应对特定设计挑战的能力，以及对基础方法的持续改进，应该能够使蛋白质设计达到更高的复杂性，接近并在某些情况下超越自然进化所达到的水平。

# Acknowledgements

# 致谢

We thank Namrata Anand and Doug Tischer for helpful discussions, and Indrek Kalvet and Yakov Kipnis for providing helpful Rosetta scripts. We thank Ian Haydon for help with graphics. We also thank Luki Goldschmidt and Kandise VanWormer, respectively, for maintaining the computational and wet lab resources at the Institute for Protein Design.

我们感谢Namrata Anand和Doug Tischer的有益讨论，以及Indrek Kalvet和Yakov Kipnis提供的有用的Rosetta脚本。我们感谢Ian Haydon在图形方面的帮助。我们还特别感谢Luki Goldschmidt和Kandise VanWormer，感谢他们维护蛋白质设计研究所的计算和湿实验室资源。

This work was supported by gifts from Microsoft (D.J., M.B., D.B.), Amgen (J.L.W.), the Audacious Project at the Institute for Protein Design (B.L.T., I.S., J.Y., H.E., D.B.), the Washington State General Operating Fund supporting the Institute for Protein Design (P.V., I.S.), grant INV-010680 from the Bill and Melinda Gates Foundation Grant (W.B.A., D.J., J.W., D.B.), grant DE-SC0018940 MOD03 from the U.S. Department of Energy Office of Science (A.J.B., D.B.), grant 5U19AG065156-02 from the National Institute for Aging (S.V.T., D.B.), an EMBO long-term fellowship ALTF 139-2018 (B.I.M.W.), the Open Philanthropy Project Improving Protein Design Fund (R.J.R., D.B.), The Donald and Jo Anne Petersen Endowment for Accelerating Advancements in Alzheimer’s Disease Research (N.R.B.), a Washington Research Foundation Fellowship (S.J.P.), a Human Frontier Science Program Cross Disciplinary Fellowship (LT000395/2020-C, L.F.M.), an EMBO Non-Stipendiary Fellowship (ALTF 1047-2019, L.F.M.), the Defense Threat Reduction Agency grants HDTRA1-19-1-0003 (N.H., D.B.) and HDTRA12210012 (F.D.), the Institute for Protein Design Breakthrough Fund (A.C., D.B.), an EMBO Postdoctoral Fellowship (ALTF 292-2022, J.L.W.) and the Howard Hughes Medical Institute (A.C., W.S., R.R., D.B.), an NFS-GRFP (J.Y), an NSF Expeditions grant (1918839, J.Y, R.B., T.S.J.), the Machine Learning for Pharmaceutical Discovery and Synthesis consortium (J.Y, R.B., T.S.J.), the Abdul Latif Jameel Clinic for Machine Learning in Health (J.Y, R.B., T.S.J), the DTRA Discovery of Medical Countermeasures Against New and Emerging threats program (J.Y, R.B., T.S.J), and EPSRC Prosperity Partnership EP/T005386/1 (E.M.), the DARPA Accelerated Molecular Discovery program and the Sanofi Computational Antibody Design grant (J.Y, R.B., T.S.J.). We thank Microsoft and AWS for generous gifts of cloud computing resources.

本研究得到了来自微软(D.J.，M.B.，D.B.)、安进(J.L.W.)、蛋白质设计研究所的大胆计划(B.L.T.，I.S.，J.Y.，H.E.，D.B.)、支持蛋白质设计研究所的华盛顿州一般运营基金(P.V.，I.S.)、比尔和梅琳达·盖茨基金会的赠款INV-010680(W.B.A.，D.J.，J.W.，D.B.)、美国能源部科学办公室的赠款DE-SC0018940 MOD03(A.J.B.，D.B.)、国家衰老研究所的赠款5U19AG065156-02(S.V.T.，D.B.)、EMBO长期奖学金ALTF 139-2018(B.I.M.W.)、开放慈善项目改善蛋白质设计基金(R.J.R.，D.B.)、唐纳德和乔安·彼得森加速阿尔茨海默病研究进展的基金(N.R.B.)、华盛顿研究基金会奖学金(S.J.P.)、人类前沿科学计划跨学科奖学金(LT000395/2020-C，L.F.M.)、EMBO非津贴奖学金(ALTF 1047-2019，L.F.M.)、国防威胁减少局的赠款HDTRA1-19-1-0003(N.H.，D.B.)和HDTRA12210012(F.D.)、蛋白质设计研究所突破基金(A.C.，D.B.)、EMBO博士后奖学金(ALTF 292-2022，J.L.W.)以及霍华德·休斯医学研究所(A.C.，W.S.，R.R.，D.B.)、NFS-GRFP(J.Y.)、NSF探险赠款(1918839，J.Y.，R.B.，T.S.J.)、制药发现与合成的机器学习联盟(J.Y.，R.B.，T.S.J.)、阿卜杜勒·拉蒂夫·贾米尔健康机器学习诊所(J.Y.，R.B.，T.S.J.)、DTRA针对新兴威胁的医疗对策发现计划(J.Y.，R.B.，T.S.J.)以及EPSRC繁荣伙伴关系EP/T005386/1(E.M.)、DARPA加速分子发现计划和赛诺菲计算抗体设计赠款(J.Y.，R.B.，T.S.J.)的支持。我们感谢微软和AWS慷慨提供的云计算资源。

# Author Contributions

# 作者贡献

Conceived the study: J.L.W., D.J., N.R.B, B.L.T., J.Y., D.B.; Trained RF diffusion: J.L.W., D.J., N.R.B, W.A., B.L.T., J.Y.; Extended diffusion to residue orientations: B.L.T., J.Y. with assistance from V.D.B., E.M.; Generated experimentally characterized designs: H.E.E., D.J., J.L.W., N.R.B., N.H., W.S., P.V., I.S.; Generated computational designs: W.A., B.L.T., J.Y., D.J., J.L.W., N.R.B.; Experimentally characterized designs: H.E.E., A.J.B., R.J.R., L.F.M., B.I.M.W., S.J.P., N.H., A.C., S.V.T., J.L.W., B.L.T.; Contributed additional code: J.W., A.L., W.S.; Trained RF: M.B., F.D.; Offered supervision throughout the project: D.B., T.S.J. and R.B.; Wrote the manuscript: J.L.W., D.J., B.L.T., J.Y., N.R.B., D.B. All authors read and contributed to the manuscript. J.L.W. and D.J. agree that for personal pursuits, the order of their respective names may be changed to best their own interests. Figures

构思研究:J.L.W.、D.J.、N.R.B、B.L.T.、J.Y.、D.B.；训练RF扩散:J.L.W.、D.J.、N.R.B、W.A.、B.L.T.、J.Y.；将扩散扩展到残基取向:B.L.T.、J.Y.，在V.D.B.、E.M.的协助下；生成实验表征的设计:H.E.E.、D.J.、J.L.W.、N.R.B.、N.H.、W.S.、P.V.、I.S.；生成计算设计:W.A.、B.L.T.、J.Y.、D.J.、J.L.W.、N.R.B.；实验表征的设计:H.E.E.、A.J.B.、R.J.R.、L.F.M.、B.I.M.W.、S.J.P.、N.H.、A.C.、S.V.T.、J.L.W.、B.L.T.；贡献额外代码:J.W.、A.L.、W.S.；训练RF:M.B.、F.D.；在整个项目中提供监督:D.B.、T.S.J.和R.B.；撰写手稿:J.L.W.、D.J.、B.L.T.、J.Y.、N.R.B.、D.B. 所有作者阅读并贡献了手稿。J.L.W.和D.J.同意为了个人事务，可以根据自身利益调整各自名字的顺序。图表

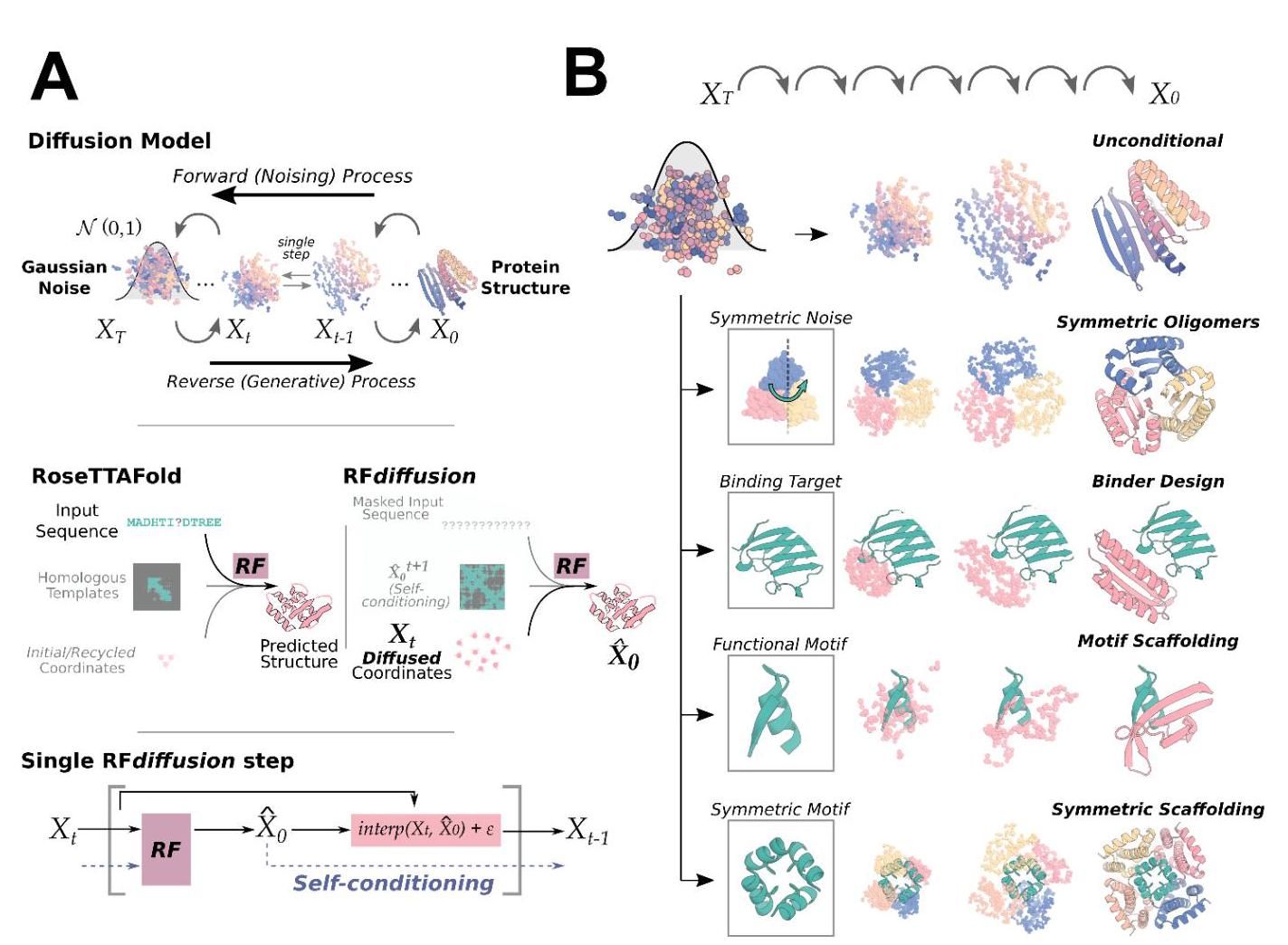
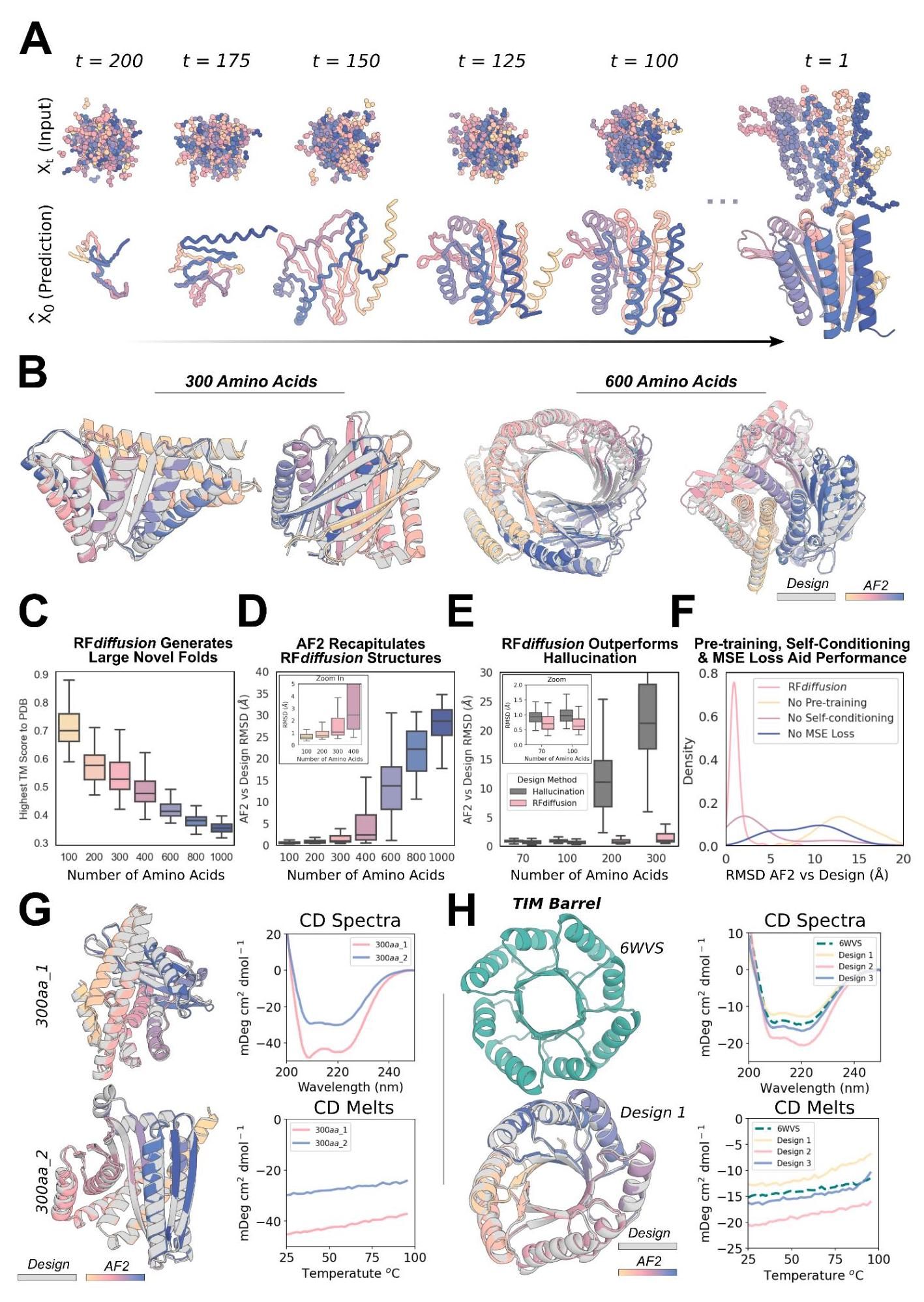


Figure 1: RFdiffusion is a denoising diffusion probabilistic model with RoseTTAFold fined-tuned as the denoising network. A) Top panel: Diffusion models for proteins are trained to recover structures of proteins corrupted with noise, and generate new structures by reversing the corruption process through iterative denoising of initially random noise into a realistic structure . Middle panel: RoseTTAFold (RF, left) can be fine-tuned as the denoising network in a DDPM. RFdiffusion (right) is trained from a pre-trained RF network with minimal architectural changes. While in RF, the primary input to the model is sequence, in RFdiffusion, the primary input is diffused residue frames. In both cases, the model predicts final 3D coordinates directly (denoted in RFdiffusion). In RFdiffusion, the model receives its previous prediction as a template input ("self-conditioning", see Methods 2.4). Bottom panel: At each timestep "t" of a design trajectory (typically 200 steps), RFdiffusion takes and from the previous step and then predicts an updated structure . The coordinate input to the model at the next time step is generated by a noisy interpolation toward . B) RFdiffusion is of broad applicability to protein design. RFdiffusion generates protein structures either without additional input (top row), or by conditioning on: symmetric inputs to design symmetric oligomers (second row); a binding target (third row); protein functional motifs (fourth row); symmetric functional motifs to design symmetric oligomers scaffolds (bottom row). In each case random noise, along with conditioning information, is input to RFdiffusion, which iteratively refines that noise until a final protein structure is designed. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

图1:RFdiffusion是一种去噪扩散概率模型，使用RoseTTAFold作为去噪网络进行微调。A) 上面板:蛋白质的扩散模型经过训练以恢复被噪声破坏的蛋白质结构，并通过对最初随机噪声 的迭代去噪过程生成新的结构 。中间面板:RoseTTAFold(RF，左)可以作为DDPM中的去噪网络进行微调。RFdiffusion(右)是在预训练的RF网络基础上进行的，架构变化最小。在RF中，模型的主要输入是序列，而在RFdiffusion中，主要输入是扩散的残基帧。在这两种情况下，模型直接预测最终的3D坐标(在RFdiffusion中表示为 )。在RFdiffusion中，模型将其之前的预测作为模板输入(“自我条件化”，见方法2.4)。下面板:在设计轨迹的每个时间步“t”(通常为200步)中，RFdiffusion从上一步获取 和 ，然后预测更新后的 结构 。模型在下一个时间步的坐标输入 是通过向 的噪声插值生成的。B) RFdiffusion在蛋白质设计中具有广泛的适用性。RFdiffusion生成蛋白质结构，既可以不需要额外输入(第一行)，也可以通过条件化:对称输入以设计对称寡聚体(第二行)；结合目标(第三行)；蛋白质功能基序(第四行)；对称功能基序以设计对称寡聚体支架(最后一行)。在每种情况下，随机噪声与条件信息一起输入RFdiffusion，RFdiffusion迭代地精炼该噪声，直到设计出最终的蛋白质结构。bioRxiv预印本doi:https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审的)是作者/资助者，已授予bioRxiv永久展示该预印本的许可。根据CC-BY-ND 4.0国际许可证提供。

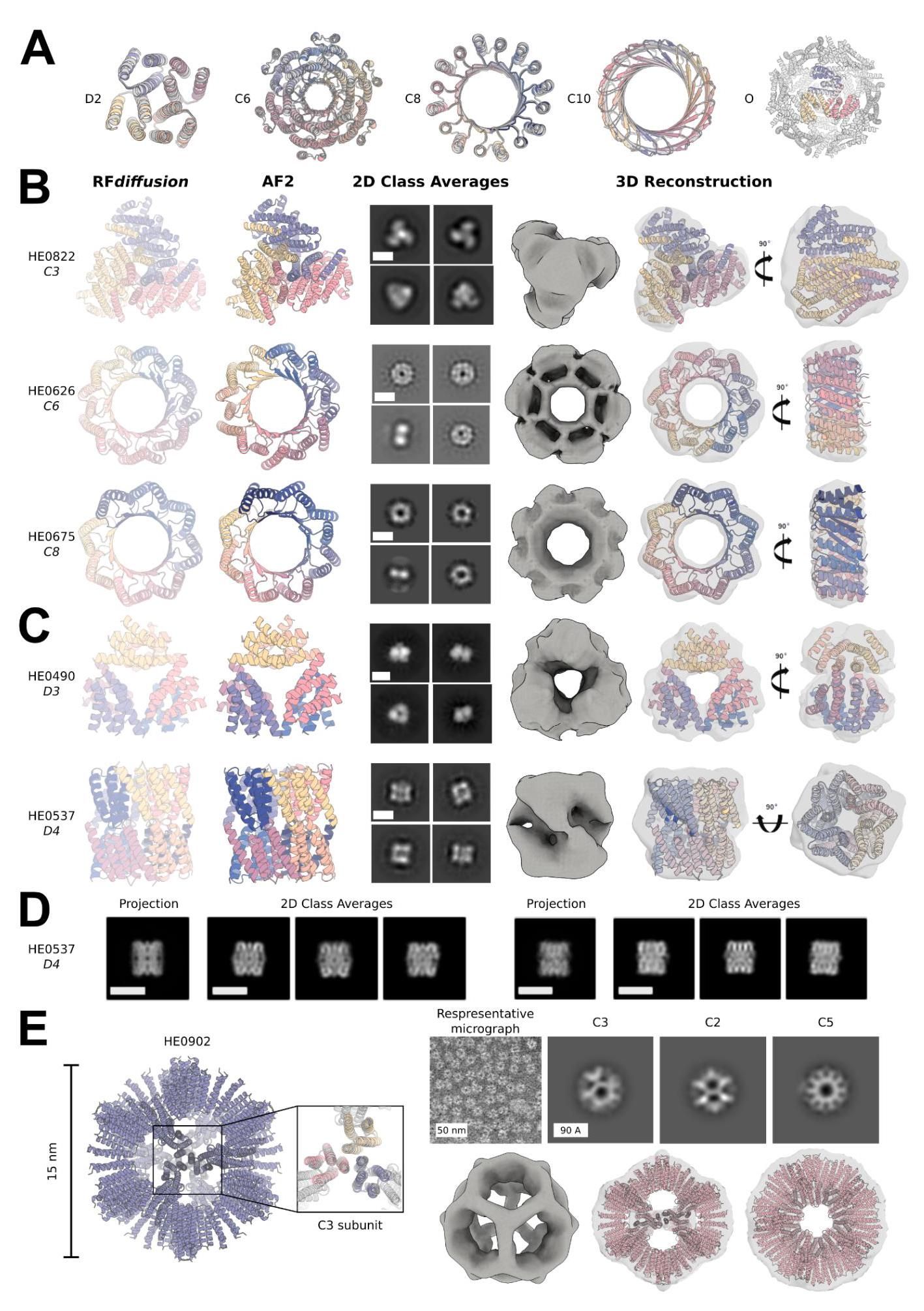


bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审认证)是作者/资助方，他们已授予bioRxiv永久展示该预印本的许可。该文档根据CC-BY-ND 4.0国际许可协议提供。

Figure 2: Outstanding performance of RFdiffusion for monomer generation. A) An example trajectory of an unconditional 300 amino acid design, depicting the input to the model and the corresponding prediction. At early timesteps (high ), predictions bear little resemblance to a protein, but are gradually refined into a protein structure. B) RFdiffusion can generate new monomeric proteins of different lengths (left: 300, right: 600) with no conditioning information. Gray=design model; colors=AlphaFold2 (AF2) prediction. RMSD AF2 vs design ( Å ), left to right: 0.90, 0.98, 1.15, 1.67. C) Unconditional designs from RFdiffusion are novel and not present in the training set as quantified by highest TM score to the protein data bank (PDB). Designs are increasingly novel with increasing length. D) Unconditional samples are closely re-predicted by AF2. Beyond 400 amino acids, the recapitulation by AF2 deteriorates. E) RFdiffusion significantly outperforms Hallucination (with RoseTTAFold) at unconditional monomer generation (two-way ANOVA & Tukey’s test, ). While Hallucination successfully generates designs up to 100 amino acids in length, success rates rapidly deteriorate beyond this length. F) Ablating pre-training (by starting from untrained RF), self-conditioning, or MSE losses (by training with FAPE) each dramatically decrease the performance of RFdiffusion. RMSD between design and AF2 is shown, for the motif-scaffolding problem "5TPN" (see Methods 5.2). G) Two example 300 amino acid proteins that expressed as soluble monomers. Designs (gray) overlaid with AF2 predictions (colors) are shown on the left, alongside CD spectra (top) and melt curves (bottom) on the right. The designs are highly thermostable. H) RFdiffusion can condition on fold information. An example TIM barrel is shown (bottom left), conditioned on the secondary structure and block-adjacency of a previously designed TIM barrel, PDB: 6WVS (top left). Designs have very similar CD spectra to 6WVS (top right), and are highly thermostable (bottom right).

图2:RFdiffusion在单体生成方面的卓越表现。A) 一个无条件300个氨基酸设计的示例轨迹，描绘了模型的输入 和相应的 预测。在早期时间步(高 )， 预测与蛋白质几乎没有相似之处，但逐渐被细化为蛋白质结构。B) RFdiffusion可以生成不同长度的新单体蛋白(左:300，右:600)，没有条件信息。灰色=设计模型；颜色=AlphaFold2(AF2)预测。AF2与设计的RMSD(Å)，从左到右:0.90，0.98，1.15，1.67。C) RFdiffusion的无条件设计是新颖的，并且在训练集中不存在，量化标准为与蛋白质数据银行(PDB)的最高TM分数。随着长度的增加，设计的创新性也在增加。D) AF2对无条件样本的重新预测非常接近。超过400个氨基酸后，AF2的重现效果下降。E) RFdiffusion在无条件单体生成方面显著优于Hallucination(与RoseTTAFold一起)(双向ANOVA和Tukey测试， )。虽然Hallucination成功生成长度达到100个氨基酸的设计，但超过此长度后成功率迅速下降。F) 消除预训练(从未训练的RF开始)、自我条件化或MSE损失(通过使用FAPE训练)都会显著降低RFdiffusion的性能。设计与AF2之间的RMSD显示在“5TPN”这个基序支架问题上(见方法5.2)。G) 两个表达为可溶性单体的300个氨基酸蛋白的示例。左侧显示了设计(灰色)与AF2预测(颜色)的叠加，以及右侧的CD光谱(顶部)和熔解曲线(底部)。这些设计具有很高的热稳定性。H) RFdiffusion可以基于折叠信息进行条件化。示例TIM桶显示在左下角，基于先前设计的TIM桶的二级结构和块邻接条件化，PDB: 6WVS(左上角)。设计与6WVS的CD光谱非常相似(右上角)，并且具有很高的热稳定性(右下角)。



bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审认证)是作者/资助者，他们已授予 bioRxiv 永久展示该预印本的许可。该文档在 aCC-BY-ND 4.0 国际许可下提供。

Figure 3: Design and experimental characterization of high-order symmetric oligomers. A) RFdiffusion-generated assemblies overlaid with the AF2 structure predictions based on the designed sequences; in all 5 cases they are nearly indistinguishable. Symmetries are indicated to the left of the design models. The octahedral symmetries were validated by their C3 subunits only, as shown in panel A. B-C) Designed assemblies characterized by negative stain electron microscopy. Model symmetries: B) Cyclic: C3 (HE0822, 350 AA/chain); C6 (HE0626, 100 AA/ chain); C8 (HE0675, 60 AA/ chain) C) Dihedral: D3 (HE0490, 80 AA/ chain); and D4 (HE0537, 100 AA/ chain). From left to right: 1) symmetric design model, 2) AF2 prediction of design following sequence design with ProteinMPNN, 3) 2D class averages showing a combination of (at minimum) top and side views (scale bar = 60 Å for all class averages), 4) 3D reconstructions from class averages with the design model fit into the density map. The overall shapes are closely consistent with the design models, and confirm the intended oligomeric state. As in ), the AF2 predictions of each design are nearly indistinguishable from the original diffusion model (backbone RMSDs ( Å ) for HE0822, HE0626, HE0490, HE0675, and HE0537, are 1.33, 1.03, 0.60, 0.74, and 0.75, respectively). D) Two orthogonal side views of HE0537 by cryo-EM. Representative 2D class averages from the cryo-EM data are shown to the right of the predicted 2D projection images of the computational design model (lowpass filtered to 8 Å), which appear nearly identical to the experimental data. Scale bar shown is 60 Å for all images. E) Characterized icosahedral particle (HE0902, 100 AA/ chain) by negative stain electron microscopy. The design model, including the prediction of the subunit are shown on the left. nsEM data are shown on the right: on top, a representative micrograph is shown alongside 2D class averages representing each axis of symmetry (C3, C2, and C5, from left to right) with their corresponding 3D reconstruction map views shown directly below and demonstrating high agreement to the design model. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

图3:高阶对称寡聚物的设计与实验表征。A) 基于设计序列的RF扩散生成的组装与AF2结构预测重叠；在所有5个案例中，它们几乎无法区分。对称性在设计模型的左侧标示。八面体对称性仅通过其C3亚单位得到验证，如面板A所示。B-C) 通过负染色电子显微镜表征的设计组装。模型对称性:B) 循环:C3(HE0822，350 AA/链)；C6(HE0626，100 AA/链)；C8(HE0675，60 AA/链)C) 二面角:D3(HE0490，80 AA/链)；和D4(HE0537，100 AA/链)。从左到右:1) 对称设计模型，2) 基于ProteinMPNN的序列设计的AF2预测，3) 2D类平均显示(至少)顶部和侧面的组合视图(所有类平均的比例尺=60 Å)，4) 从类平均中重建的3D图像，设计模型与密度图相符。整体形状与设计模型高度一致，确认了预期的寡聚状态。如 所示，每个设计的AF2预测与原始扩散模型几乎无法区分(HE0822、HE0626、HE0490、HE0675和HE0537的骨架RMSD(Å)分别为1.33、1.03、0.60、0.74和0.75)。D) 通过冷冻电子显微镜(cryo-EM)获得的HE0537的两个正交侧视图。右侧显示了来自cryo-EM数据的代表性2D类平均，与计算设计模型的预测2D投影图像(低通滤波至8 Å)几乎相同，实验数据也显示出相似性。所有图像的比例尺为60 Å。E) 通过负染色电子显微镜表征的二十面体颗粒(HE0902，100 AA/链)。设计模型，包括 对 亚单位的预测，显示在左侧。右侧显示nsEM数据:顶部显示一幅代表性显微照片，旁边是表示每个对称轴的2D类平均(从左到右为C3、C2和C5)，其对应的3D重建图视图直接显示在下方，显示出与设计模型的高度一致性。bioRxiv预印本doi:https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审的版本)为作者/资助者，已授予bioRxiv永久展示该预印本的许可。根据CC-BY-ND 4.0国际许可协议提供。

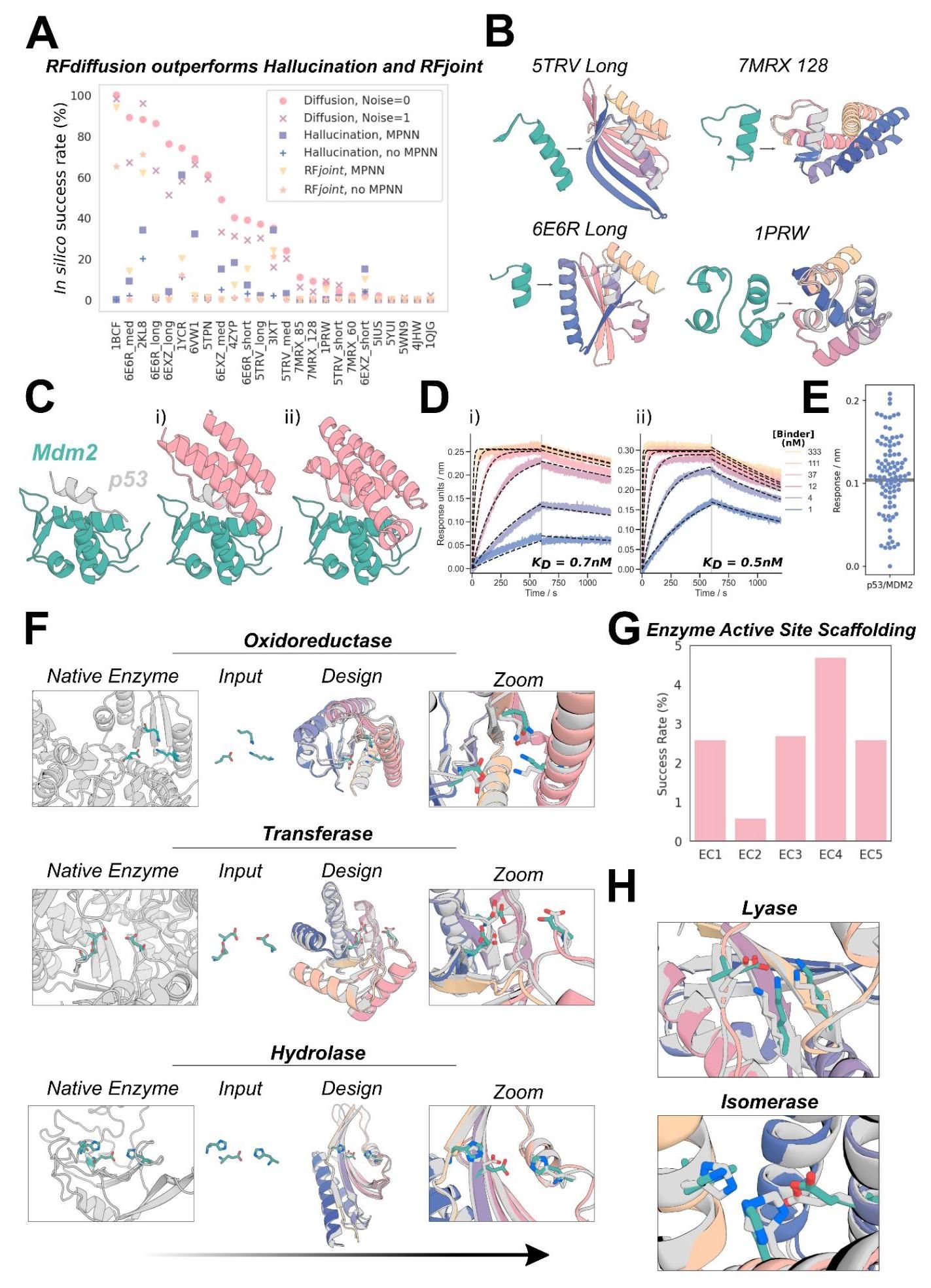
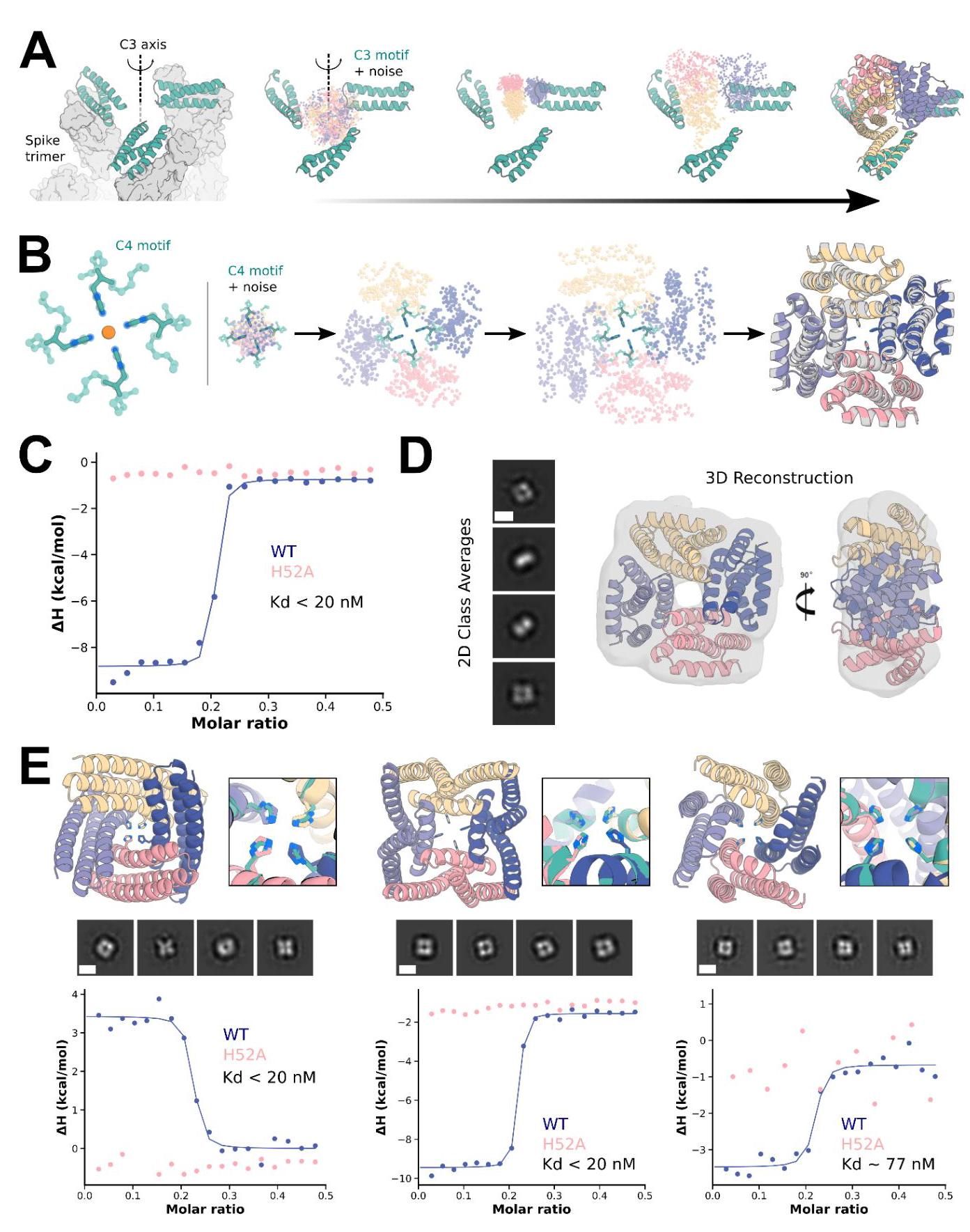


Figure 4: Scaffolding of diverse functional-sites with RFdiffusion. A) RFdiffusion has state of the art performance across 25 benchmark motif scaffolding problems collected from six recent publications, encompassing a broad range of motifs (Supplementary Table 9). Success was defined as AF2 RMSD to design model < 2Å, AF2 RMSD to the native functional site (the "motif") < 1 Å, and AF2 predicted alignment error (pAE) < 5, and the examples are ordered by success rate with RFdiffusion (with noise scale = 0). 100 designs were generated per problem, with no prior optimization on the benchmark set (some optimization was necessary for the Hallucination results). Supplementary Table 10 presents full results. B) Four examples of designs for benchmarking problems where RFdiffusion significantly outperforms existing methods. Teal: native motif; colors: AF2 prediction of an RF diffusion design. Metrics (RMSD AF2 vs design / vs native motif ( Å ), AF2 pAE): 5TRV Long: 1.17/0.57, 4.73; 6E6R Long: 0.89/0.27, 4.56; 7MRX Long: 0.84/0.82 4.32; 1PRW: 0.77/0.89, 4.49. C) RFdiffusion can scaffold the native p53 helix that binds to MDM2 (left) and makes additional contacts with the target (right, average 31% increased surface area). D) Biolayer interferometry (BLI) measurements demonstrate high affinity (0.7nM and 0.5nM) binding to MDM2 for the two designs shown in ; the native p53 helix affinity is . ) Experimental success rates were high, with 55/95 designs showing significant binding to MDM2 (> 50% of maximum response). F) After fine-tuning on a task that mimics active-site scaffolding (Methods 4.2), RFdiffusion can scaffold a broad range of enzyme active sites. Three examples are shown (Enzyme Classes, EC, 1-3; ref [50]). Left to right: native enzyme (PDB: 1A41, 1CWY, 1DE3); catalytic site (teal); RFdiffusion output (gray: model, colors: AF2 prediction); zoom of active site. G) In silico success rates on active sites derived from EC1-5 (AF2 Motif RMSD vs native: backbone < 1 Å, backbone and sidechain atoms < 1.5 Å, RMSD AF2 vs design < 2, AF2 pAE < 5). H) Zoom in views of two further successful designs, for EC4 and EC5 (active sites from PDB: 1P1X, 1SNZ). Metrics for examples in (F) and (H) (AF2 vs design backbone RMSD, AF2 vs design motif backbone RMSD, AF2 vs design motif full-atom RMSD, AF2 pAE): EC2: 0.93Å, 0.50Å, 1.29Å, 3.51; EC3: 0.92Å, 0.60Å, 1.07Å, 4.59; EC4: 0.93Å, 0.80Å, 1.03Å, 4.41; EC5: 0.78Å, . bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license. (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

图4:使用RFdiffusion构建多样化功能位点的支架。A) RFdiffusion在从六篇近期出版物中收集的25个基准图案支架问题上表现出色，涵盖了广泛的图案(补充表9)。成功的定义是AF2 RMSD与设计模型<2Å，AF2 RMSD与天然功能位点(“图案”)<1Å，以及AF2预测的对齐误差(pAE)<5，示例按RFdiffusion的成功率排序(噪声比例=0)。每个问题生成100个设计，基准集上没有先前的优化(某些优化对于幻觉结果是必要的)。补充表10提供了完整结果。B) 四个RFdiffusion显著优于现有方法的基准问题设计示例。青色:天然图案；颜色:RF扩散设计的AF2预测。指标(RMSD AF2与设计/与天然图案(Å)，AF2 pAE):5TRV Long: 1.17/0.57, 4.73；6E6R Long: 0.89/0.27, 4.56；7MRX Long: 0.84/0.82 4.32；1PRW: 0.77/0.89, 4.49。C) RFdiffusion可以支架与MDM2结合的天然p53螺旋(左)并与目标形成额外接触(右，平均增加31%的表面积)。D) 生物层干涉测量(BLI)显示出对MDM2的高亲和力(0.7nM和0.5nM)，如 所示的两个设计；天然p53螺旋的亲和力为 。 )实验成功率很高，55/95个设计显示出对MDM2的显著结合(>最大反应的50%)。F) 在模拟活性位点支架的任务上进行微调后(方法4.2)，RFdiffusion可以支架广泛的酶活性位点。展示了三个示例(酶类别，EC，1-3；参考文献[50])。从左到右:天然酶(PDB: 1A41, 1CWY, 1DE3)；催化位点(青色)；RFdiffusion输出(灰色:模型，颜色:AF2预测)；活性位点的放大图。G) 从EC1-5导出的活性位点的计算成功率(AF2图案RMSD与天然:主链<1Å，主链和侧链原子<1.5Å，RMSD AF2与设计<2，AF2 pAE<5)。H) 对两个进一步成功设计的放大视图，针对EC4和EC5(活性位点来自PDB: 1P1X, 1SNZ)。示例(F)和(H)的指标(AF2与设计主链RMSD，AF2与设计图案主链RMSD，AF2与设计图案全原子RMSD，AF2 pAE):EC2: 0.93Å, 0.50Å, 1.29Å, 3.51；EC3: 0.92Å, 0.60Å, 1.07Å, 4.59；EC4: 0.93Å, 0.80Å, 1.03Å, 4.41；EC5: 0.78Å， 。bioRxiv预印本doi:https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审)是作者/资助者，已授予bioRxiv永久展示该预印本的许可。根据CC-BY-ND 4.0国际许可证提供。



bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint

bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842；此版本发布于2022年12月10日。该预印本的版权持有者

Figure 5: Symmetric motif scaffolding with RFdiffusion. A) Design of C3-symmetric oligomers to scaffold the binding interface of the designed ACE2 mimic, AHB2 (left, teal), against the SARS-CoV-2 spike trimer (left, gray). Starting from AHB2 bound to each of the three ACE2 binding sites on the spike trimer, RFdiffusion was used to generate C3-symmetric oligomers that hold the three AHB2 exactly in place to simultaneously engage the binding sites on all three spike subunits. The first 55 amino-acids of each minibinder copy are used as the symmetric motif input to RFdiffusion (middle). The method produces designs whose AF2 predictions (right) recapitulate the mini-binder motif with high accuracy on the asymmetric unit (0.6 Å RMSD) and good accuracy the symmetric motif (2.9 Å RMSD). B) Design of C4-symmetric oligomers to scaffold a theoretical binding motif (left). Starting from a square-planar set of histidine rotamers within three-residue helical fragments (Methods 5.9) and C4-symmetric noise, an RFdiffusion trajectory iteratively builds a symmetric oligomer scaffolding the theoretical binding domain (middle). AF2 predictions (color) overlaid with the RFdiffusion design model (gray) agree closely, with backbone RMSD for the particular example < 1.0 Å (right). C) Isothermal titration calorimetry (ITC) binding isotherm of design ("E1") and corresponding H52A mutant. The inflection point of the wild-type isotherm (blue) displays an estimated dissociation constant of less than at the designed metal:monomer stoichiometry of 1:4. Importantly, the H52A mutant isotherm (pink) displays complete ablation of binding, indicating the scaffolded histidine at position 52 of each protomer is critical for metal binding. D) 2D class averages (left) and corresponding 3D reconstruction with the model of design E1 docked into the 3D reconstructed density (right). The four-fold symmetry and general shape of the designed oligomer can be readily identified in the 2D class averages, with both top-down views and side views captured (scale bar ). E) Additional experimentally characterized binding oligomers G3 (left), C10 (middle), and A5 (right) from RF diffusion show structural diversity in successful designs. Design models and binding-site zoom (top, AF2 in colors and ideal motif in teal) show close recapitulation of the motif sidechains by AF2. 2D nsEM class averages (middle, scale bar ), and binding isotherms for wild-type and H52A mutant (bottom) indicate tight binding mediated directly by the scaffolded histidines at the designed 1:4 stoichiometry. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

图5:使用RFdiffusion的对称基元支架。A) 设计C3对称寡聚物以支撑设计的ACE2模拟物AHB2(左，青色)与SARS-CoV-2刺突三聚体(左，灰色)的结合界面。从AHB2与刺突三聚体上三个ACE2结合位点结合开始，使用RFdiffusion生成C3对称寡聚物，将三个AHB2精确固定在位，以同时与所有三个刺突亚单位的结合位点结合。每个迷你结合体副本的前55个氨基酸作为对称基元输入RFdiffusion(中)。该方法生成的设计，其AF2预测(右)在不对称单元上以高精度(0.6 Å RMSD)重现迷你结合体基元，并在对称基元上具有良好的精度(2.9 Å RMSD)。B) 设计C4对称寡聚物以支撑理论 结合基元(左)。从三残基螺旋片段中的方形平面组氨酸旋转异构体(方法5.9)和C4对称噪声开始，RFdiffusion轨迹迭代构建支撑理论 结合域的对称寡聚物(中)。AF2预测(颜色)与RFdiffusion设计模型(灰色)重叠，紧密一致，特定示例的主链RMSD < 1.0 Å(右)。C) 设计(“E1”)和相应H52A突变体的等温滴定量热法(ITC)结合等温线。野生型等温线(蓝色)的拐点显示在设计的金属:单体化学计量比1:4下，估计的解离常数小于 。重要的是，H52A突变体的等温线(粉色)显示完全失去结合，表明每个原聚体在52位的支架组氨酸对金属结合至关重要。D) 2D类平均(左)和相应的3D重建，设计E1模型停靠在3D重建密度中(右)。设计的寡聚物的四重对称性和一般形状可以在2D类平均中轻松识别，捕获了俯视图和侧视图(比例尺 )。E) 额外实验表征的 结合寡聚物G3(左)、C10(中)和A5(右)来自RF扩散，显示成功设计中的结构多样性。设计模型和结合位点放大(顶部，AF2以颜色显示，理想基元为青色)显示AF2对基元侧链的紧密重现。2D nsEM类平均(中，比例尺 )，以及野生型和H52A突变体的结合等温线(底部)表明，由设计的1:4化学计量比直接介导的支架组氨酸紧密 结合。bioRxiv预印本doi:https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审)为作者/资助者，已授予bioRxiv永久展示该预印本的许可。根据CC-BY-ND 4.0国际许可协议提供。bioRxiv预印本doi:https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审)为作者/资助者，已授予bioRxiv永久展示该预印本的许可。根据CC-BY-ND 4.0国际许可协议提供。

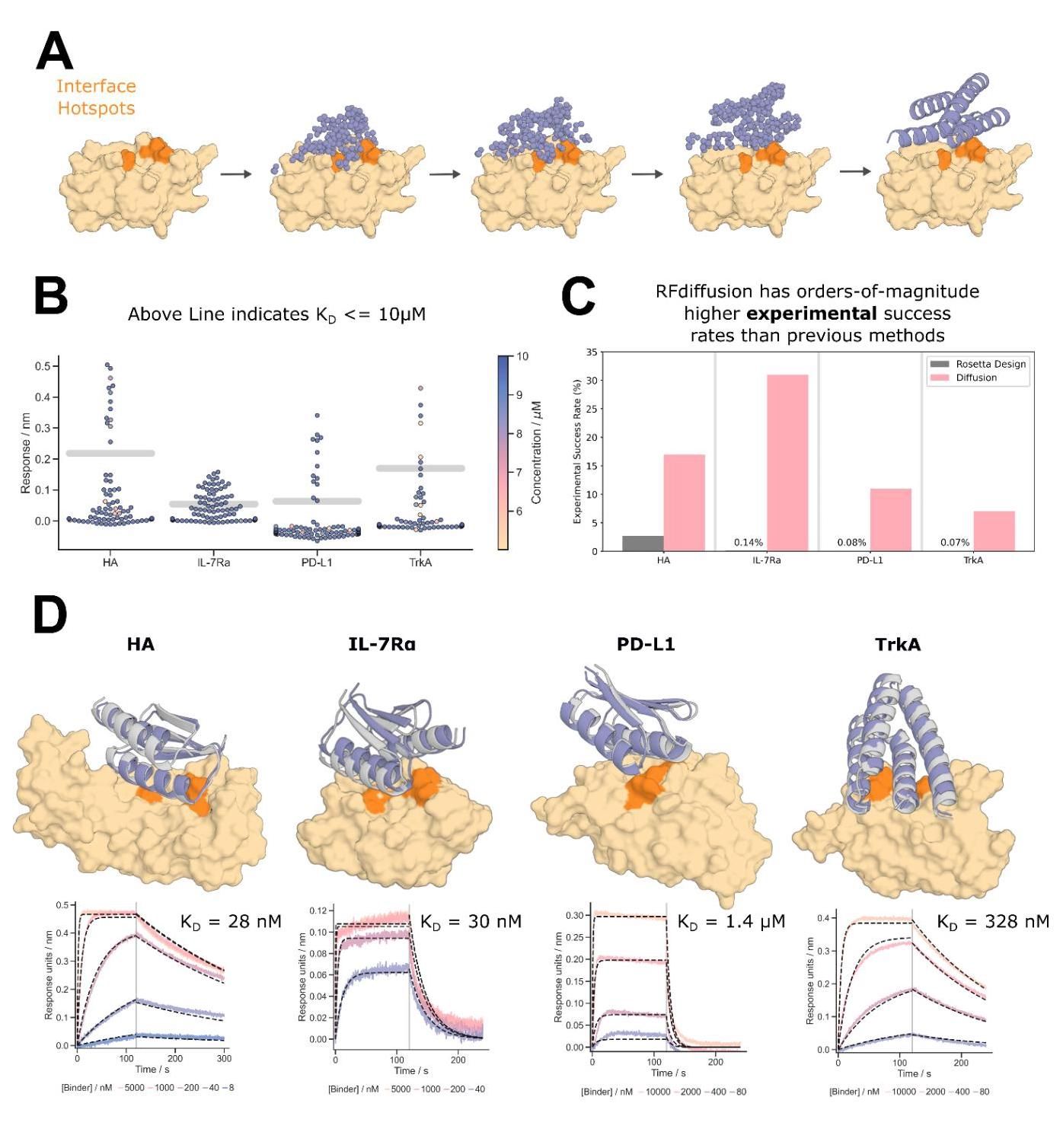


Figure 6: Design of de novo protein-binding proteins. A-B) De novo binders were designed to four protein targets; Influenza Hemagglutinin A, IL-7 Receptor-a, PD-L1, and TrkA receptor. A) RFdiffusion generates protein binders by conditioning on interface hotspot residues. Additionally, the general topology of the binders generated by RFdiffusion can be controlled using fold-conditioning. B) De novo protein binders were identified for all four of the targets for which we could obtain suitable target protein (see Fig. S15 for expression data on Insulin receptor binders). Designs that bound at during single point BLI screening with a response equal to or greater than 50% of the positive control were considered binders. Concentration is denoted by hue for designs that were screened at concentrations less than and thus may be false negatives. C) RFdiffusion designed binders have very high experimental success rates compared to the previous design campaigns against the same targets. For IL-7Ra, PD-L1, and TrkA, RF diffusion has success rates 2̃ orders-of-magnitude higher than the original design campaigns. D) For each target, the highest affinity binder is shown alongside a BLI titration series. Reported are based on global kinetic fitting with fixed global . Yellow/orange: target/hotspot residues; gray: design model; purple: AF2 prediction (RMSD AF2 vs design, left to right: 0.7Å, 0.9Å, 1.2Å, 0.7Å). bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

图6:新型蛋白质结合蛋白的设计。A-B) 新型结合剂被设计用于四个蛋白质靶标；流感血凝素A、IL-7受体-a、PD-L1和TrkA受体。A) RFdiffusion通过对界面热点残基进行条件化生成蛋白质结合剂。此外，RFdiffusion生成的结合剂的一般拓扑结构可以通过折叠条件进行控制。B) 对于我们能够获得合适靶标蛋白的所有四个靶标，识别出了新型蛋白结合剂(有关胰岛素受体结合剂的表达数据，请参见图S15)。在单点BLI筛选中，结合反应等于或大于阳性对照50%的设计被视为结合剂。浓度通过色调表示，对于在低于 浓度下筛选的设计，可能存在假阴性。C) 与之前针对相同靶标的设计活动相比，RFdiffusion设计的结合剂具有非常高的实验成功率。对于IL-7Ra、PD-L1和TrkA，RF扩散的成功率比原始设计活动高出约两个数量级。D) 对于每个靶标，显示了最高亲和力的结合剂及其BLI滴定系列。报告的 基于固定全局 的全局动力学拟合。黄色/橙色:靶标/热点残基；灰色:设计模型；紫色:AF2预测(AF2与设计的RMSD，从左到右:0.7Å，0.9Å，1.2Å，0.7Å)。bioRxiv预印本doi:https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审的)是作者/资助者，他们已授予bioRxiv永久展示该预印本的许可。该文档根据CC-BY-ND 4.0国际许可协议提供。

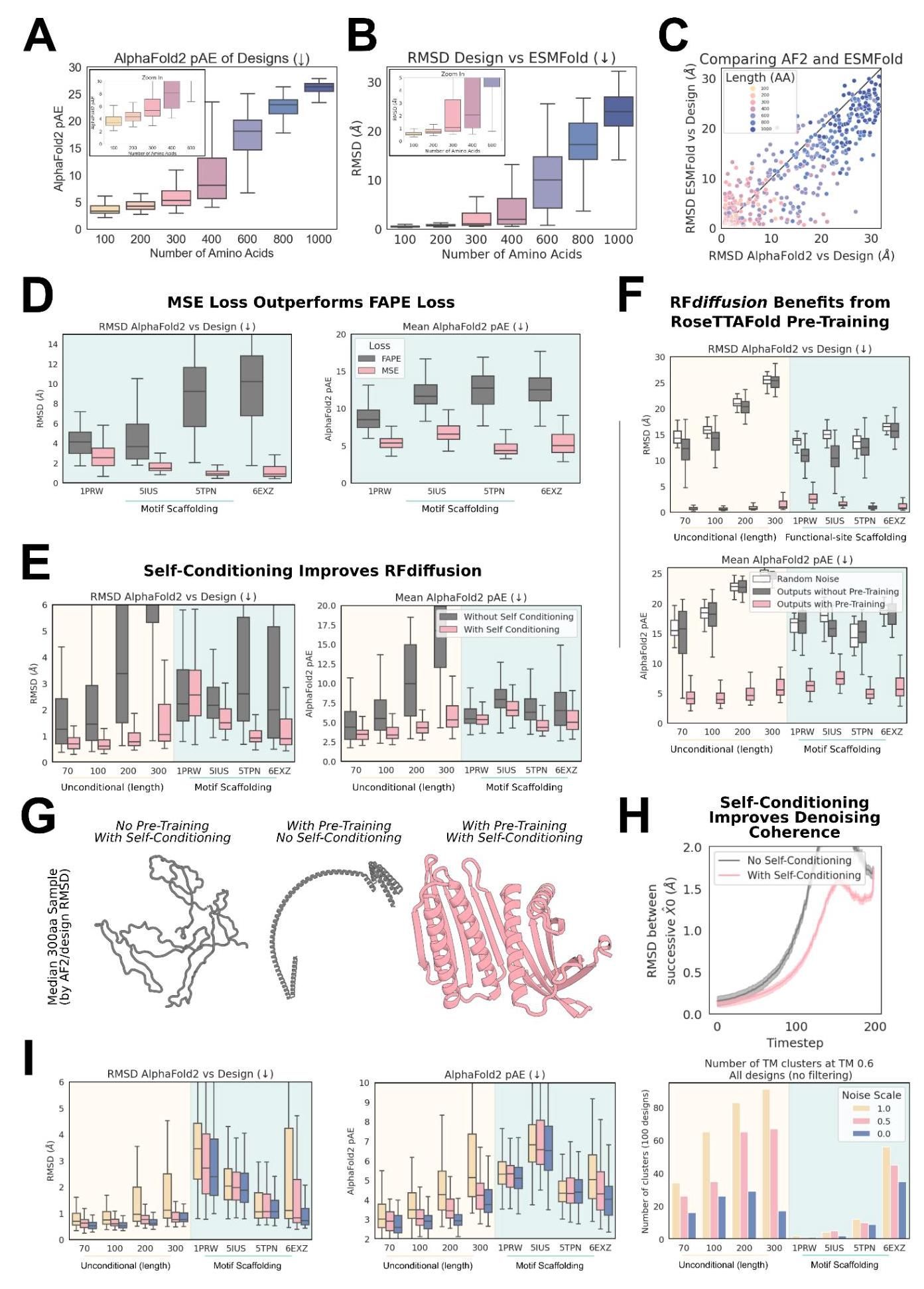
# Supplementary Figures

# 补充图表



Figure S1: RFdiffusion learns the distribution of the denoising process. Analysis of simulated forward (noising) and reverse (denoising) trajectories shows that the distribution of coordinates and residue orientations closely match, demonstrating that RFdiffusion has learned the distribution of the denoising process as desired. Left to right: i) average distance between a coordinate at and its position in ; ii) average distance between a coordinate at and ; iii) average distance between adjacent coordinates at ; iv) average rotation distance between a residue orientation at and ) average rotation distance between a residue orientation at and .

图 S1:RFdiffusion 学习去噪过程的分布。对模拟的正向(加噪)和反向(去噪)轨迹的分析表明， 坐标和残基取向的分布非常接近，证明 RFdiffusion 已如预期学习了去噪过程的分布。从左到右:i) 坐标在 的平均距离及其在 的位置；ii) 坐标在 和 之间的平均距离；iii) 相邻 坐标在 之间的平均距离；iv) 处的残基取向与 之间的平均旋转距离；v) 处的残基取向与 之间的平均旋转距离。



bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842; 此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审认证)是作者/资助方，他们已授予bioRxiv永久展示该预印本的许可。该文档根据CC BY-ND 4.0国际许可证提供。

# Figure S2: Training ablations reveal determinants of RFdiffusion success. A-C)

# 图 S2: 训练消融揭示RF扩散成功的决定因素。A-C)

RFdiffusion can generate high quality large unconditional monomers. Designs are routinely accurately recapitulated by AF2 (see also Fig. 2D), with high confidence (A) for proteins up to approximately 400 amino acids in length. B) Further orthogonal validation of designs by ESMFold. C) Recapitulation of the design structure is often better with ESMFold compared with AF2. For each backbone, the best of 8 ProteinMPNN sequences is plotted, with points therefore paired by backbone rather than sequence. D) Comparing RFdiffusion trained with MSE loss on atoms and backbone frames (Methods 2.5), rather than with FAPE loss . The two models were benchmarked on motif scaffolding problems (see Methods 5.2 for justification of this decision), and across all cases, AF2 recapitulation of the structure (left) and AF2 confidence (right) was improved when RFdiffusion was trained with MSE loss. Two-way ANOVA: Success rate Allowing the model to condition on its prediction at the previous timestep (see Methods 2.4) improves designs. Designs with self-conditioning (pink) have improved recapitulation by AF2 (left) and better AF2 confidence in the prediction (right). Two-way ANOVA, in silico success rate: ) RFdiffusion leverages the protein representations learned during RF pre-training. RFdiffusion fine-tuned from pre-trained RF (pink) comprehensively outperforms a model trained for an equivalent amount of time, from untrained weights (gray). Training RFdiffusion without pre-training (for 5 epochs) showed no significant improvement (in terms of in silico success rates) compared with generating ProteinMPNN sequences from random Gaussian-sampled coordinates (white, two-way ANOVA & Tukey’s test, ; Random noise vs no pre-training, (n.s.); Random noise vs with pre-training, ; Pre-training vs not, . Note that the data in pink in D-F is the same data, reproduced in each plot for clarity. G) The median (by AF2 RMSD vs design) 300 amino acid unconditional sample highlighting the importance of self-conditioning and pre-training. Without pre-training, RFdiffusion outputs bear little resemblance to proteins (gray, left). Without self-conditioning, outputs show characteristic protein secondary structures, but lack core-packing and ideality (gray, middle). With pre-training and self-conditioning, proteins are diverse and well-packed (pink, right). H) Greater coherence during unconditional denoising may partly explain the effect of self-conditioning. Successive predictions are more similar when the model can self-condition (lower RMSD between predictions, pink curve). Data are aggregated from unconditional design trajectories of 100,200 and 300 residues. I) During the reverse (generation) process, the noise added at each step can be scaled (reduced). Reducing the noise scale improves the in silico design success rates (left, middle; two-way ANOVA & Tukey’s test: vs 0.5 : vs 1 : vs 1 : . This comes at the expense of diversity, with the number of unique clusters at a TM score cutoff of 0.6 reduced when noise is reduced (right). Note throughout this figure the 6EXZ\_long benchmarking problem is abbreviated to 6EXZ for brevity. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

RFdiffusion可以生成高质量的大型无条件单体。设计通常通过AF2准确重现(另见图2D)，对于长度约为400个氨基酸的蛋白质具有高置信度(A)。B)通过ESMFold进一步进行设计的正交验证。C)与AF2相比，ESMFold对设计结构的重现通常更好。对于每个主链，绘制了8个ProteinMPNN序列中最好的一个，因此点是按主链而不是序列配对的。D)比较使用MSE损失在 原子和 主链框架上训练的RFdiffusion，而不是使用FAPE损失 。这两个模型在基元支架问题上进行了基准测试(见方法5.2以说明此决定)，在所有情况下，当RFdiffusion使用MSE损失进行训练时，AF2对结构的重现(左)和AF2置信度(右)得到了改善。双向方差分析:成功率 允许模型在前一个时间步的 预测上进行自我条件化(见方法2.4)可以改善设计。具有自我条件化(粉色)的设计通过AF2(左)得到了更好的重现，并且在预测中具有更好的AF2置信度(右)。双向方差分析，计算机模拟成功率: )RFdiffusion利用在RF预训练期间学习的蛋白质表示。经过预训练的RF(粉色)微调的RFdiffusion全面优于从未训练权重(灰色)训练的模型，后者训练时间相当。与从随机高斯采样坐标生成ProteinMPNN序列(白色)相比，未经过预训练的RFdiffusion(训练5个周期)在计算机模拟成功率方面没有显著改善(双向方差分析和Tukey检验， ；随机噪声与未预训练， (不显著)；随机噪声与预训练， ；预训练与未预训练， 。请注意，D-F中粉色的数据在每个图中重复以便于理解。G)中位数(按AF2 RMSD与设计)300个氨基酸无条件样本突显了自我条件化和预训练的重要性。没有预训练，RFdiffusion输出与蛋白质几乎没有相似之处(灰色，左)。没有自我条件化，输出显示出特征性的蛋白质二级结构，但缺乏核心包装和理想性(灰色，中)。经过预训练和自我条件化，蛋白质多样且包装良好(粉色，右)。H)无条件去噪过程中更大的连贯性可能部分解释了自我条件化的效果。当模型能够自我条件化时，连续的 预测更为相似( 预测之间的RMSD较低，粉色曲线)。数据来自100、200和300个残基的无条件设计轨迹。I)在反向(生成)过程中，每一步添加的噪声可以缩放(减少)。减少噪声规模提高了计算机模拟设计的成功率(左，中；双向方差分析和Tukey检验: 与0.5: 与1: 与1: 。这以牺牲多样性为代价，当噪声减少时，在TM评分截止值为0.6时，独特簇的数量减少(右)。请注意，在整个图中，6EXZ\_long基准问题缩写为6EXZ以简化。bioRxiv预印本doi:https://doi.org/10.1101/2022.12.09.519842；此版本发布于2022年12月10日。该预印本的版权持有者(未经过同行评审认证)是作者/资助者，他们已授予bioRxiv永久展示该预印本的许可。根据CC-BY-ND 4.0国际许可证提供。bioRxiv预印本doi:https://doi.org/10.1101/2022.12.09.519842；此版本发布于2022年12月10日。该预印本的版权持有者(未经过同行评审认证)是作者/资助者，他们已授予bioRxiv永久展示该预印本的许可。根据CC-BY-ND 4.0国际许可证提供。

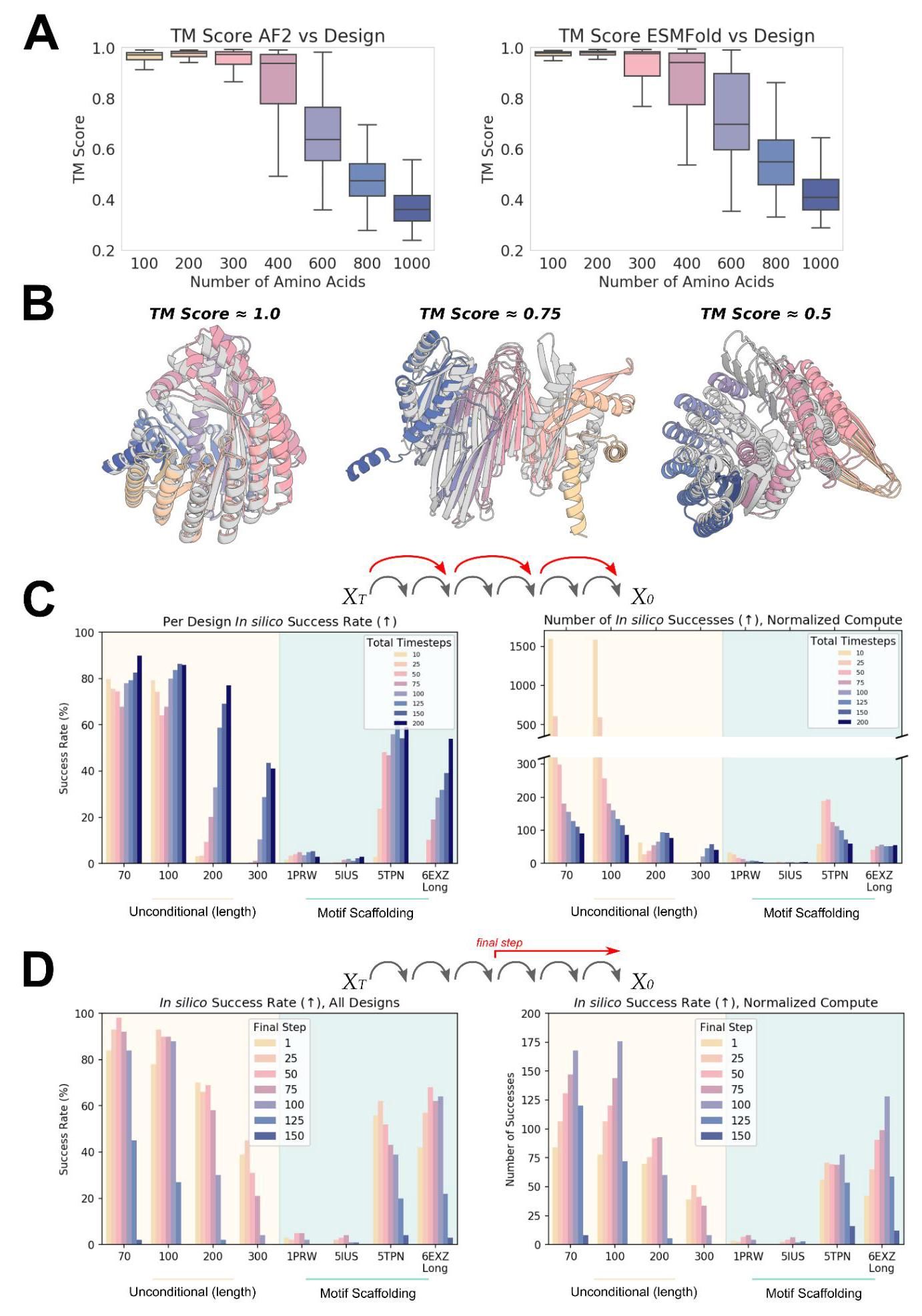


Figure S3: Optimizing inference and improving metrics for in silico success. A-B) TM score between a design and a subsequent orthogonal prediction (e.g. AF2), has been previously used, typically with a threshold of > 0.5, as a metric for design success. A) RFdiffusion designs have high TM score agreement to both the AF2 (left) and ESMFold (right) predictions of the unconditional structures, with for a significant fraction of designs even up to 1000 amino acids in length. B) TM score is, however, much less stringent than RMSD alignment. Depicted here are three unconditional RFdiffusion designs of 600 amino acids in length (gray), overlaid with the AF2 prediction (colors), with TM scores of 0.983, 0.757 and 0.506 respectively. While a TM score of 0.5 clearly shows some resemblance to the designed structure, it differs significantly and should not be classed as "successfully designed". RMSD with a strict threshold (for example, ) is significantly more stringent. RMSDs for the displayed designs are 1.15 Å, 9.78 Å and 21.4 Å respectively. C-D) While RFdiffusion is trained to generate samples over 200 timesteps, in many cases, trajectories can be shortened to improve computational efficiency. C) Bigger steps can be taken between timesteps at inference. While decreasing the number of timesteps typically reduces the per-design success rate (left), when normalized for compute budget (right), it is often more efficient to run more trajectories with fewer timesteps. For example, while generating 100 amino acid unconditional proteins, using a schedule with just 10 timesteps (as opposed to 200) allows the generation of 1584 in silico successful designs in the time taken to generate 86 successful designs with 200 timesteps. As problems get more challenging, however, this no longer remains the case (for example, fourth column, with generation of 300 amino acid designs). D) An alternative to taking larger steps is to stop trajectories early (possible because RFdiffusion predicts at every timestep). In many cases, trajectories can be stopped at timestep 50-75 with little effect on the final success rate of designs (left), and when normalized by compute budget (right), success rates per unit time are typically higher generating more designs with early-stopping. For example, in the 6EXZ\_Long benchmarking motif-scaffolding problem, stopping trajectories at allows the generation of 128 in silico successful designs in the time it takes to generate 42 successful designs running full trajectories. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license. (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license. prediction.

图 S3:优化推理和提高计算成功的指标。A-B) 设计与后续正交预测(例如 AF2)之间的 TM 分数，之前通常使用，通常阈值为 > 0.5，作为设计成功的指标。A) RFdiffusion 设计与 AF2(左)和 ESMFold(右)对无条件结构的预测具有高 TM 分数一致性，对于相当一部分设计，甚至在长度达到 1000 个氨基酸时也能达到 。B) 然而，TM 分数远不如 RMSD 对齐严格。这里描绘的是三个长度为 600 个氨基酸的无条件 RFdiffusion 设计(灰色)，与 AF2 预测(彩色)叠加，TM 分别为 0.983、0.757 和 0.506。虽然 TM 分数为 0.5 显示出与设计结构有一定相似性，但差异显著，不应被归类为“成功设计”。RMSD 具有严格的阈值(例如， )则严格得多。所显示设计的 RMSD 分别为 1.15 Å、9.78 Å 和 21.4 Å。C-D) 虽然 RFdiffusion 被训练生成超过 200 个时间步的样本，但在许多情况下，可以缩短轨迹以提高计算效率。C) 在推理时，可以在时间步之间采取更大的步长。虽然减少时间步的数量通常会降低每个设计的成功率(左)，但当按计算预算归一化(右)时，通常更有效的是用更少的时间步运行更多的轨迹。例如，在生成 100 个氨基酸的无条件蛋白质时，使用仅 10 个时间步的计划(而不是 200)可以在生成 86 个成功设计所需的时间内生成 1584 个计算成功的设计。然而，随着问题变得更加复杂，这种情况不再成立(例如，第四列，生成 300 个氨基酸的设计)。D) 采取更大步长的替代方案是提前停止轨迹(因为 RFdiffusion 在每个时间步都预测 )。在许多情况下，轨迹可以在时间步 50-75 停止，对设计的最终成功率影响不大(左)，当按计算预算归一化(右)时，单位时间内的成功率通常更高，生成更多的设计。举例来说，在 6EXZ\_Long 基准测试的模体支架问题中，在 停止轨迹可以在生成 42 个成功设计所需的时间内生成 128 个计算成功的设计。bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842；此版本于 2022 年 12 月 10 日发布。该预印本的版权持有者(未经过同行评审)是作者/资助者，已授予 bioRxiv 永久展示该预印本的许可。根据 CC-BY-ND 4.0 国际许可证提供。

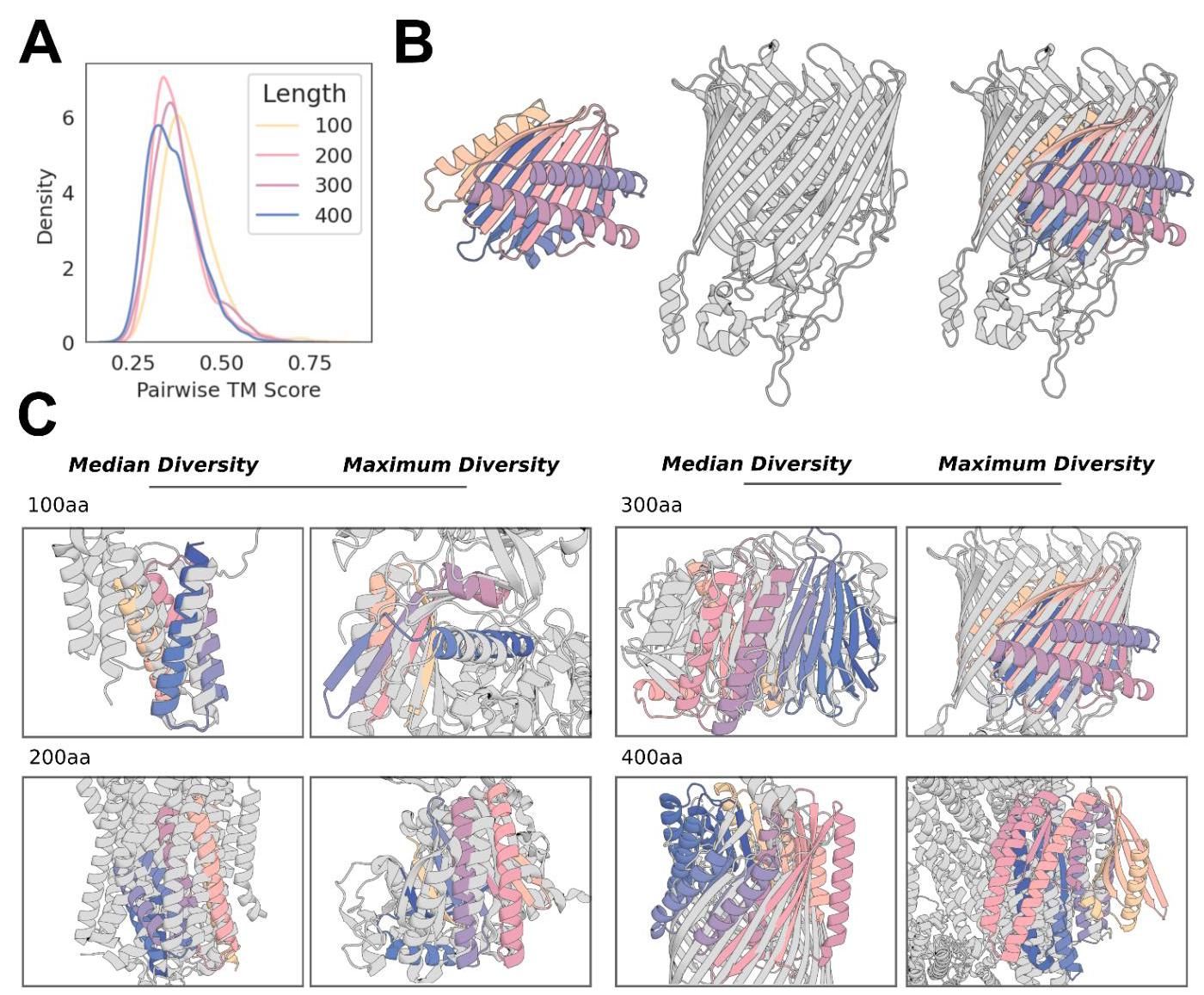


Figure S4: RFdiffusion designs are diverse and dissimilar to proteins in the PDB. A) Comparing unconditional designs to one another (100 designs per length) demonstrates that, by TM score alignment, designs are diverse (medians 100-400aa: 0.39, 0.36, 0.37, 0.35). B-C) Designs also bear little resemblance to the training set (PDB). B) Example of the most diverse (lowest TM score hit) to the PDB for a set of 300 amino acid designs. The folds of the design (left) and native protein (middle) are highly dissimilar, aligning only across a portion of the

图 S4:RFdiffusion 设计多样且与 PDB 中的蛋白质不同。A) 比较无条件设计之间的差异(每个长度 100 个设计)表明，通过 TM 分数对齐，设计是多样的(中位数 100-400aa: 0.39, 0.36, 0.37, 0.35)。B-C) 设计与训练集(PDB)也几乎没有相似之处。B) 300 个氨基酸设计中与 PDB 最多样(最低 TM 分数命中)的示例。设计的折叠(左)与天然蛋白质(中)高度不同，仅在部分区域对齐。

-sheet. C) Example designs demonstrating extrapolation beyond the training set for generating novel folds. Gray: closest protein in the PDB by TM score, colors: RFdiffusion design model, overlaid by TM alignment. For each protein length, the median and most diverse samples are shown (the 300aa design is the same as in ). While for short proteins, designs typically show some similarity to known protein folds, with increasing length, designs become increasingly dissimilar to the PDB. TM score (closest PDB, TM score; median, most diverse): 100aa: 5WVE\_A, 0.71; 4W5T\_A, 0.59; 200aa: 4AV3\_A, 0.58; 4CLY\_A, 0.47; 300aa: 4PEW\_B, 0.53; 4RDR\_A, 0.46; 400aa: 4AIP\_A, 0.49; 6R9T\_A, 0.42.

-表。C) 示例设计展示了超出训练集的外推以生成新颖的折叠。灰色:在TM评分中与PDB中最接近的蛋白质，颜色:RFdiffusion设计模型，叠加在TM对齐上。对于每种蛋白质长度，显示了中位数和最具多样性的样本(300aa设计与 中的相同)。对于短蛋白质，设计通常显示出与已知蛋白质折叠的某些相似性，随着长度的增加，设计与PDB的相似性逐渐降低。TM评分(最接近的PDB，TM评分；中位数，最具多样性):100aa:5WVE\_A，0.71；4W5T\_A，0.59；200aa:4AV3\_A，0.58；4CLY\_A，0.47；300aa:4PEW\_B，0.53；4RDR\_A，0.46；400aa:4AIP\_A，0.49；6R9T\_A，0.42。

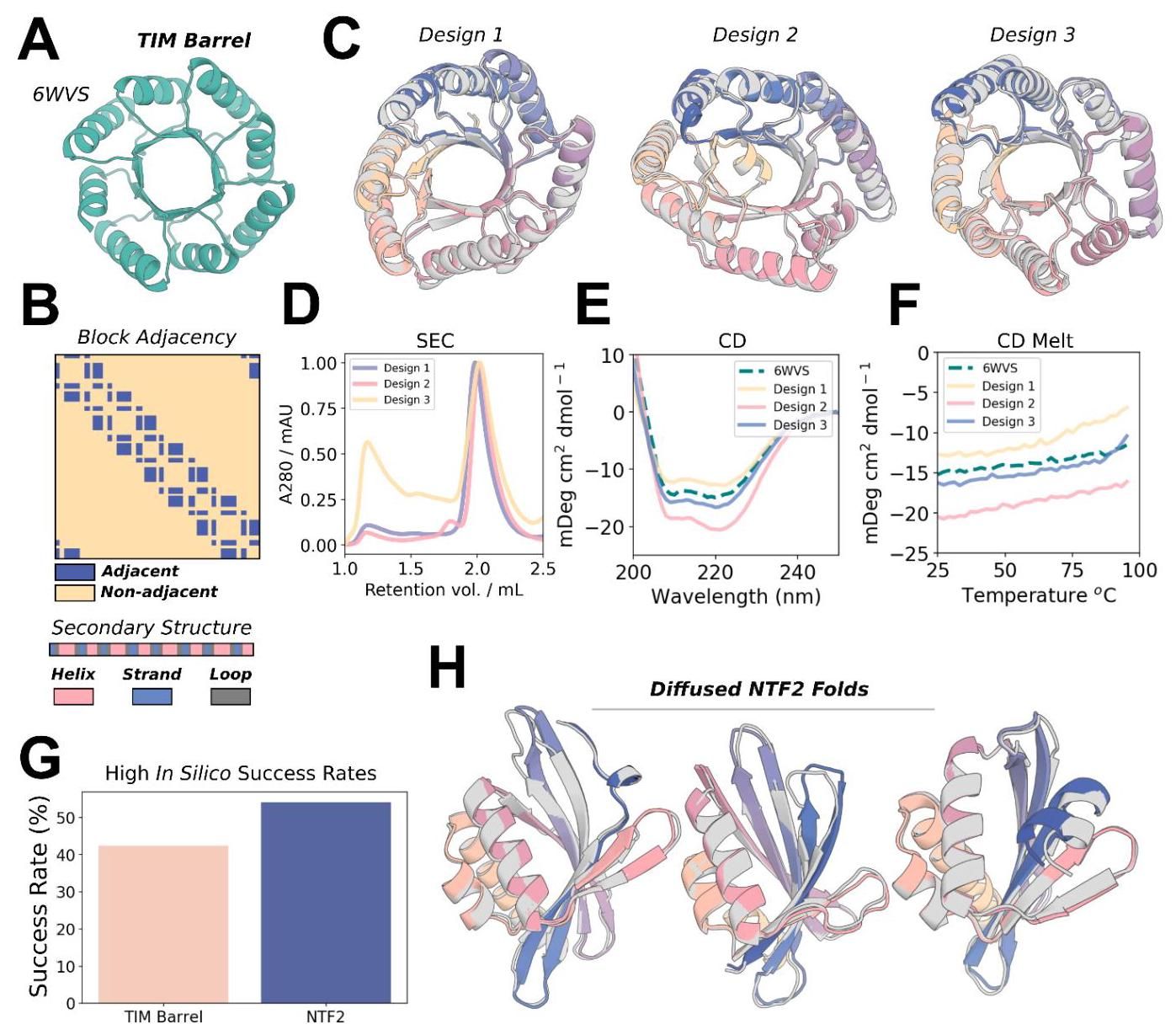
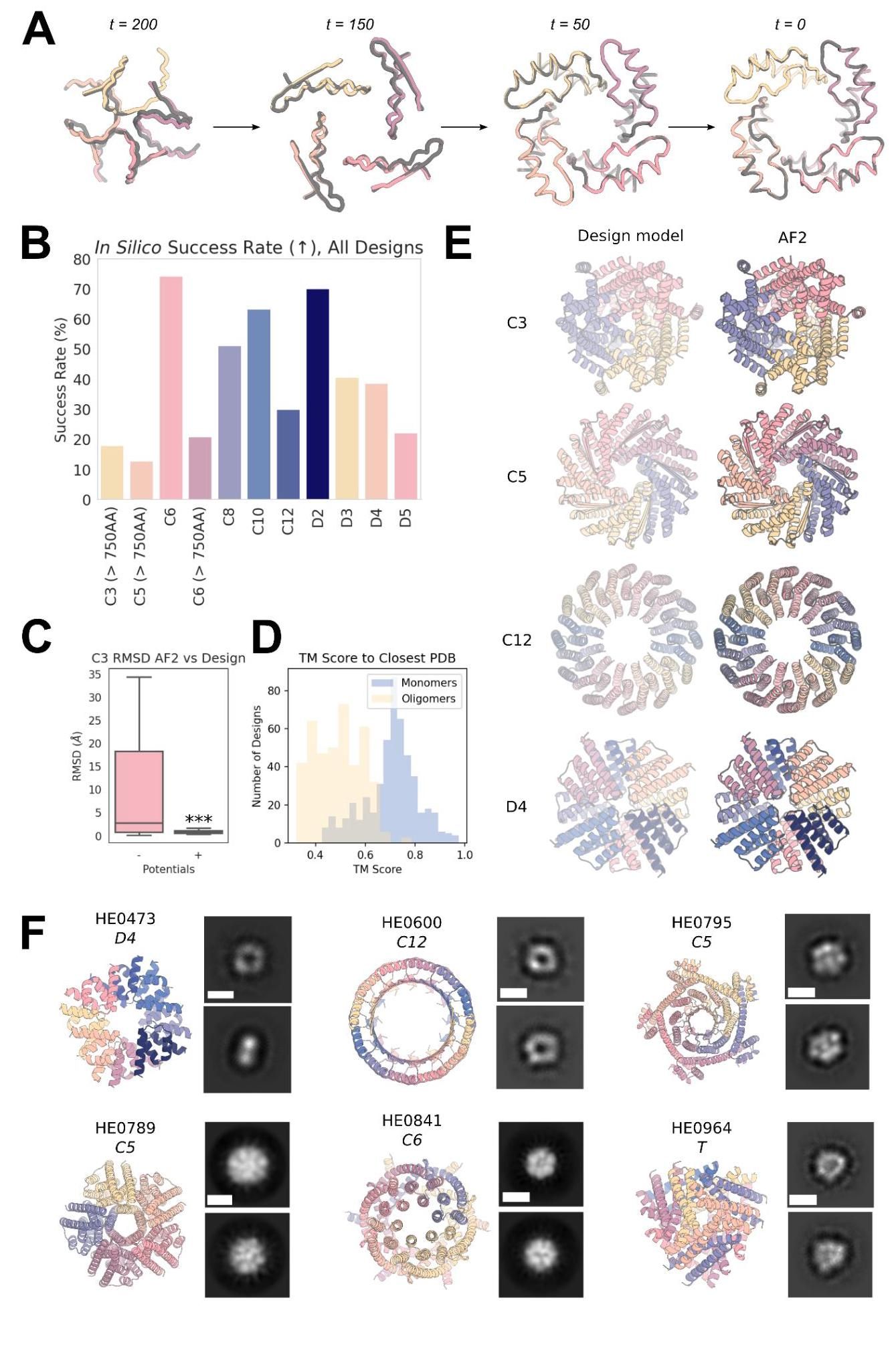


Figure S5: RFdiffusion can condition on fold information to generate specific scaffold types. A-B) 6WVS is a previously-described de novo designed TIM barrel (left). A fine-tuned RFdiffusion model can condition on 1D and 2D inputs representing this protein fold, specifically secondary structure (B, bottom) and block-adjacency information (B, top) (see Methods 4.3.2). C) RFdiffusion readily conditions on fold information and generates a diverse set of TIM barrels. D-F) Purification of the three designs depicted in (C) show elution at the predicted volume (D), circular dichroism (CD) spectra very similar to 6VWS (E), and very high thermal stability (F). Note that ) and ) are reproduced from Fig. 2H, for clarity. G) TIM barrels are generated with an in silico success rate of 42.5% (left bar). Success incorporates AF2 metrics and a TM score vs 6WVS > 0.5. G-H) NTF2 folds are useful scaffolds for de novo enzyme design, and can also be readily generated with fold-conditioning in RFdiffusion. Designs are diverse(H)and designed with an in silico success rate of 54.1% (G, right bar). NTF2 fold design success also included both AF2 metrics and a TM score vs PDB: 1GY6 > 0.5. Gray: RF diffusion design, colors: AF2

图 S5:RFdiffusion 可以根据折叠信息生成特定的支架类型。A-B) 6WVS 是一个先前描述的全新设计的 TIM 桶(左)。经过微调的 RFdiffusion 模型可以根据表示该蛋白质折叠的 1D 和 2D 输入进行条件化，特别是二级结构(B，底部)和块邻接信息(B，顶部)(见方法 4.3.2)。C) RFdiffusion 可以轻松地根据折叠信息生成多样化的 TIM 桶。D-F) 在 (C) 中描绘的三种设计的纯化显示在预测体积下洗脱(D)，圆二色性(CD)光谱与 6VWS 非常相似(E)，以及非常高的热稳定性(F)。请注意 ) 和 ) 为了清晰起见从图 2H 中复制。G) TIM 桶以 42.5% 的计算成功率生成(左侧条)。成功包括 AF2 指标和与 6WVS 的 TM 分数 > 0.5。G-H) NTF2 折叠是全新酶设计的有用支架，并且也可以通过 RFdiffusion 中的折叠条件轻松生成。设计多样(H)并以 54.1% 的计算成功率设计(G，右侧条)。NTF2 折叠设计的成功还包括 AF2 指标和与 PDB: 1GY6 的 TM 分数 > 0.5。灰色:RF diffusion 设计，颜色:AF2

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842; 此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审认证)是作者/资助方，他们已授予bioRxiv永久展示该预印本的许可。该文档根据CC-BY-ND 4.0国际许可协议提供。



bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842; 此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审认证)是作者/资助方，他们已授予bioRxiv永久展示该预印本的许可。该文档根据CC-BY-ND 4.0国际许可协议提供。

Figure S6: Symmetric oligomer design with RF diffusion. A) Due to the (near-perfect - see Methods 3.1) equivariance properties of RFdiffusion, predictions from symmetric inputs are also symmetric, even at very early timepoints (and becoming more symmetric through time; RMSD vs symmetrized: . Gray: symmetrized (top left) subunit; colors: RFdiffusion X0 prediction. B) In silico success rates for symmetric oligomer designs of various cyclic and dihedral symmetries. Success is defined here as the proportion of designs for which AF2 yields a prediction from a single sequence that has mean pLDDT > 80 and backbone RMSD over the oligomer between the design model and AF2 < 2 Å. Note that 16 sequences per RFdiffusion design were sampled. C) Box plots of the distribution of backbone RMSDs between AF2 and the RF diffusion design model with and without the use of external potentials during the trajectory. The external potentials used are the "inter-chain" contact potential (pushing chains together), as well as the "intra-chain" contact potential (making chains more globular). Using these potentials dramatically improves in silico success (Student’s unpaired t-test, ). D) Designs are diverse with respect to the training dataset (the PDB). While the monomers (typically 60-100aa) show reasonable alignment to the PDB (median 0.72), the whole oligomeric assemblies showed little resemblance to the PDB (median 0.50). E) Additional examples of design models (left) against AF2 predictions (right) for C3, C5, C12, and D4 symmetric designs (the symmetries not displayed in Fig. 3) with backbone RMSDs against their AF2 predictions of 0.82, 0.63, 0.79, and 0.78 with total amino acids 750, 900, 960, 640. F) Additional nsEM data for symmetric designs. The model is shown on the left and the 2D class averages on the right for each design. Scale bars shown (white) are . bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

图 S6:具有 RF 扩散的对称寡聚体设计。A) 由于 RF 扩散的(近乎完美 - 见方法 3.1)等变性特性，来自对称输入的 预测也是对称的，即使在非常早的时间点(并且随着时间的推移变得更加对称；RMSD 与对称化: 。灰色:对称化(左上)亚单位；颜色:RF 扩散 X0 预测。B) 各种循环和二面体对称性的对称寡聚体设计的计算成功率。这里定义的成功是指 AF2 从单一序列中产生的预测的设计比例，其平均 pLDDT > 80，且设计模型与 AF2 之间的寡聚体骨架 RMSD < 2 Å。注意每个 RF 扩散设计采样了 16 个序列。C) AF2 与 RF 扩散设计模型之间的骨架 RMSD 分布的箱形图，分别在轨迹中使用和不使用外部势能。使用的外部势能是“链间”接触势能(将链推在一起)，以及“链内”接触势能(使链更球状)。使用这些势能显著提高了计算成功率(学生无配对 t 检验， )。D) 设计在训练数据集(PDB)方面多样化。虽然单体(通常为 60-100aa)与 PDB 显示出合理的对齐(中位数 0.72)，但整个寡聚体组装与 PDB 的相似性较小(中位数 0.50)。E) C3、C5、C12 和 D4 对称设计的设计模型(左)与 AF2 预测(右)的额外示例(这些对称性未在图 3 中显示)，其与 AF2 预测的骨架 RMSD 分别为 0.82、0.63、0.79 和 0.78，总氨基酸数为 750、900、960、640。F) 对称设计的额外 nsEM 数据。模型显示在左侧，每个设计的 2D 类平均值显示在右侧。显示的比例尺(白色)为 。bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842；此版本于 2022 年 12 月 10 日发布。该预印本的版权持有者(未经过同行评审的)是作者/资助者，已授予 bioRxiv 永久展示该预印本的许可。根据 CC-BY-ND 4.0 国际许可证提供。

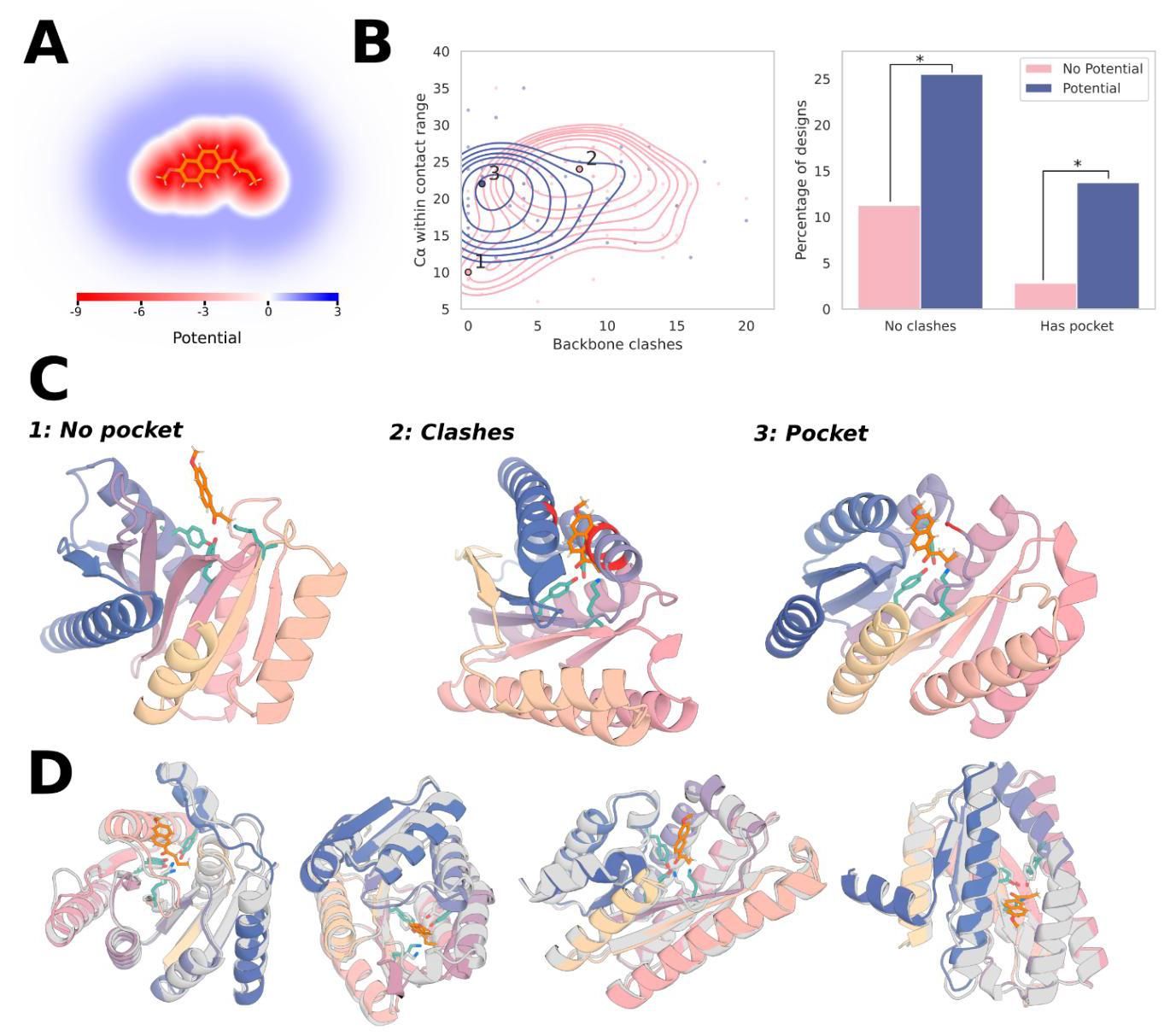


Figure S7: External potentials for generating pockets around substrate molecules. Enzymes generated from a retroaldolase active site triad [TYR1051-LYS1083-TYR1180] of a retro-aldolase: PDB: 5AN7. A) The potential used to implicitly model the substrate, which has both a repulsive and attractive field (see Methods 4.4). B) Left: Kernel densities demonstrate that without using the external potential (pink), designs often fall into two failure modes: (1) no pocket, and (2) clashes with the substrate. Right: clashes (substrate < 3A of the backbone) & pockets (no clash and > 16 Ca within 3-8A of substrate) with and without the potential. Two-proportion z-test: clashes , pocket . Each datapoint represents a design already passing the stringent success metrics (AF2 motif RMSD < 1 Å, AF2 backbone RMSD < 2 Å, AF2 pAE < 5). C) Designs close to the labeled local maxima of the kernel density estimate. Without the potential, the catalytic triad is predominantly (1) exposed on the surface with no residues available to provide substrate stabilization or (2) buried in the protein core, preventing substrate access. With the potential, the catalytic triad is predominantly (3), partially buried in a concave pocket with shape complementary to the substrate. Backbone atoms within of the available under aCC-BY-ND 4.0 International license.

图 S7:生成基质分子周围口袋的外部势能。由逆醛缩酶的活性位点三元组 [TYR1051-LYS1083-TYR1180] 生成的酶:PDB: 5AN7。A) 用于隐式建模基质的势能，具有排斥和吸引场(见方法 4.4)。B) 左侧:核密度显示在不使用外部势能(粉色)的情况下，设计通常会落入两种失败模式:(1)没有口袋，以及(2)与基质发生冲突。右侧:冲突(基质与主链距离 < 3A)和口袋(无冲突且与基质距离在 3-8A 内的 Ca 原子数 > 16)，有无势能的情况。两比例 z 检验:冲突 ，口袋 。每个数据点代表已经通过严格成功指标的设计(AF2 模式 RMSD < 1 Å，AF2 主链 RMSD < 2 Å，AF2 pAE < 5)。C) 接近核密度估计标记局部极大值的设计。在没有势能的情况下，催化三元组主要是(1)暴露在表面，没有残基可提供基质稳定，或(2)埋藏在蛋白质核心，阻碍基质接入。使用势能时，催化三元组主要是(3)，部分埋藏在形状与基质互补的凹口袋中。主链原子在 的可用范围内，遵循 CC-BY-ND 4.0 国际许可协议。

substrate are shown in red. D) A variety of diverse designs with pockets made using the potential, with no clashes between the substrate and the AF2-predicted backbone. The functional form and parameters used for the pocket potential are discussed in Methods 4.4. In each case the substrate is superimposed on the AF2 prediction of the catalytic triad. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license. and the input motif, AF2 PAE < 6, AF2 pLDDT > 90 D) In silico success count for the inverse rotamers from set #1 depicted in panel B. An in silico "success" here is defined as an AF2 prediction for a single sequence which has (1) full-atom RMSD over the four histidine residues between the AF2 prediction and the ideal C4 motif of < 1.0 Å and (2) an AF2 pAE < 10. E) Overlay of various AF2 predictions for designs scaffolding motifs derived from imidazole groups with no shear (panel A, left) shows a diverse array of RF diffusion solutions can all place the histidine imidazole groups at near-ideal distances from a theoretical nickel ion. F) Overlay of various AF2 predictions for motifs derived from imidazole groups with shear (panel A, middle and right) again displays diverse backbone solutions for placing the imidazole groups at near-ideal distances from the theoretical ion. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

基底以红色显示。D) 使用该潜力制作的多种不同设计的口袋，基底与AF2预测的主链之间没有冲突。口袋潜力的功能形式和参数在方法4.4中讨论。在每种情况下，基底都与催化三联体的AF2预测重叠。bioRxiv预印本doi: https://doi.org/10.1101/2022.12.09.519842；该版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审)是作者/资助者，已授予bioRxiv永久展示该预印本的许可。根据CC-BY-ND 4.0国际许可协议提供。输入动机，AF2 PAE < 6，AF2 pLDDT > 90 D) 在面板B中描绘的来自集合#1的逆旋转异构体的计算成功计数。在这里，计算“成功”被定义为单个序列的AF2预测，其具有(1)AF2预测与理想C4动机之间四个组氨酸残基的全原子RMSD < 1.0 Å，以及(2)AF2 pAE < 10。E) 不同AF2预测的叠加，设计支架动机源自咪唑基团且没有剪切(面板A，左侧)，显示出多样的RF扩散解决方案可以将组氨酸咪唑基团放置在理论镍离子附近的理想距离。F) 不同AF2预测的叠加，动机源自咪唑基团且有剪切(面板A，中间和右侧)，再次展示了多样的主链解决方案，将咪唑基团放置在理论 离子附近的理想距离。bioRxiv预印本doi: https://doi.org/10.1101/2022.12.09.519842；该版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审)是作者/资助者，已授予bioRxiv永久展示该预印本的许可。根据CC-BY-ND 4.0国际许可协议提供。

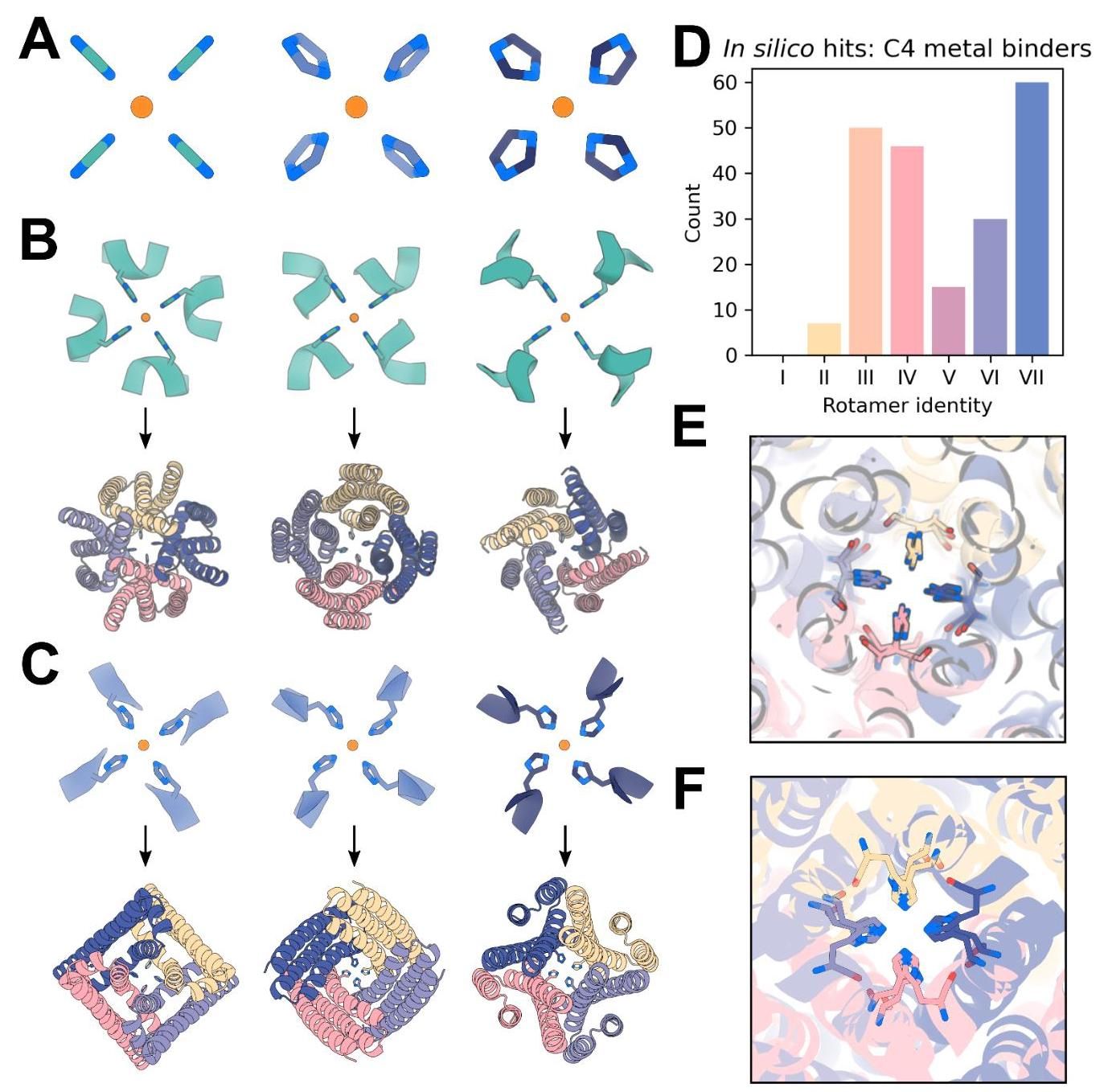
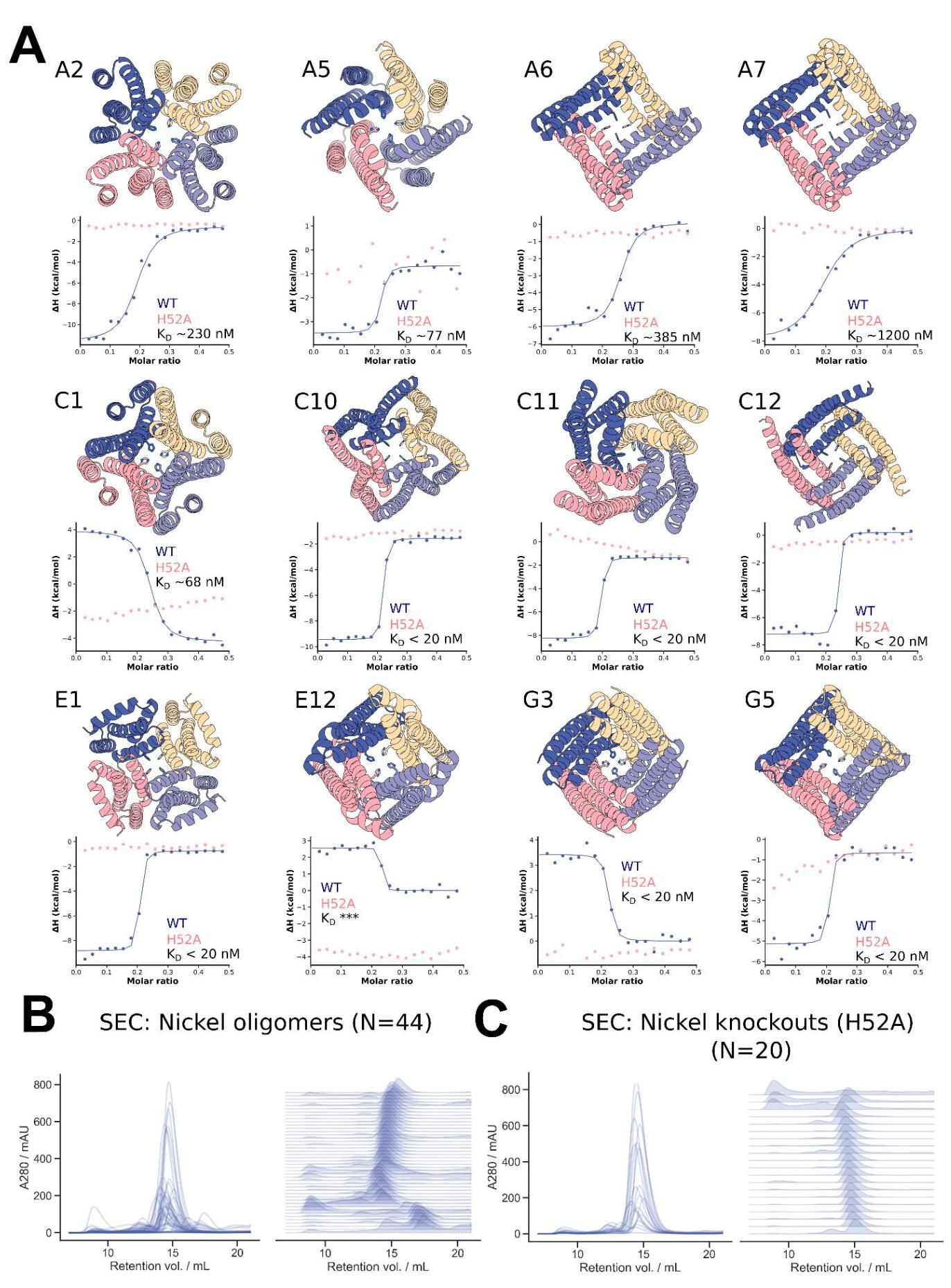


Figure S8: Symmetric motif scaffolding for square-planar binding. A) Symmetrized imidazole groups of varying amounts of shear used for constructing the square-planar motifs to scaffold, with between the theoretically coordinating nitrogen and the symmetry axis. B) Depiction of a subset of the C4-symmetrized backbone-dependent rotamers (“inverse rotamers”, Methods 5.9) used as motifs from set #1 input to RFdiffusion for symmetrically scaffolding the theoretical binding site (teal, top). AF2 predictions of selected in silico successes scaffolding the inverse rotamers show significant structural diversity in RFdiffusion solutions (colors, bottom). All AF2 structures have full-atom RMSD < 1.0 Å between AF2 predictions and the input motif, AF2 PAE < 6, and AF2 pLDDT > 90. C) Depiction of a different subset of the C4-symmetrized backbone-dependent inverse used as motifs from sets #2 and #3 (top), with AF2 predictions of selected in silico successes (bottom). All AF2 structures have full-atom RMSD < 1.0 Å between AF2 predictions

图 S8:用于方平面 结合的对称图案支架。A) 用于构建方平面图案的对称咪唑基团，使用不同量的剪切，理论上 协调氮与对称轴之间的 。B) 描述了一组 C4 对称化的主链依赖性 旋转异构体 (“逆旋转异构体”，方法 5.9)，作为从集合 #1 输入到 RFdiffusion 的图案，用于对称支架理论 结合位点(青色，顶部)。AF2 对选定的计算成功的预测显示，支架 逆旋转异构体在 RFdiffusion 解决方案中具有显著的结构多样性(颜色，底部)。所有 AF2 结构的全原子 RMSD < 1.0 Å，AF2 PAE < 6，AF2 pLDDT > 90。C) 描述了来自集合 #2 和 #3 的不同子集的 C4 对称化主链依赖性 逆 ，作为图案(顶部)，以及选定的计算成功的 AF2 预测(底部)。所有 AF2 结构的全原子 RMSD < 1.0 Å，AF2 预测之间的 RMSD

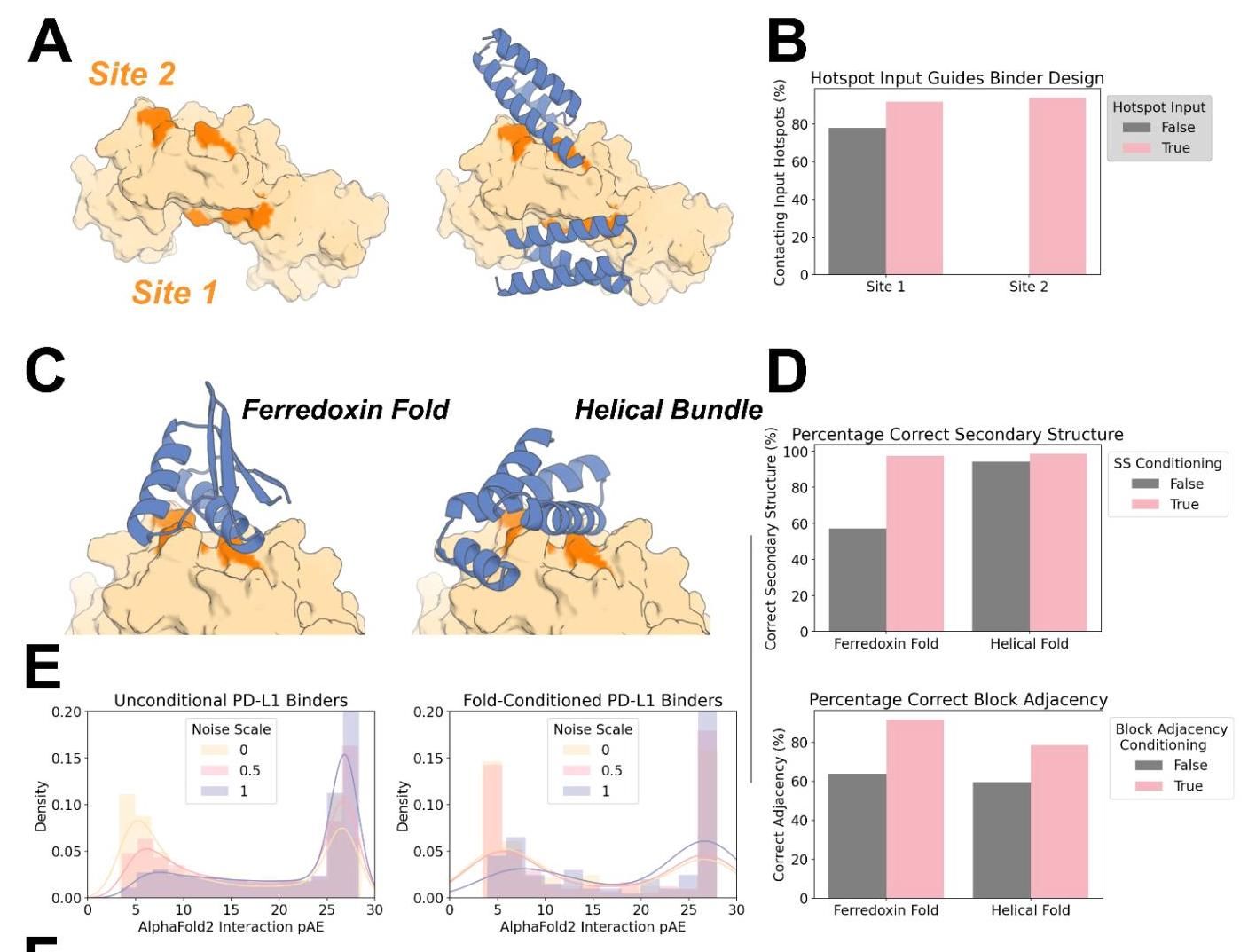


bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842; 此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审认证)是作者/资助方，他们已授予bioRxiv永久展示该预印本的许可。该文档在CC-BY-ND 4.0国际许可下提供。

Figure S9: Additional Ni binding C4 oligomers A) AF2 predictions of a subset of the experimentally verified binding oligomers, with corresponding isothermal titration calorimetry (ITC) binding isotherms for the wild-type (blue) and H52A mutant (pink) below. Wild-type dissociation constants are displayed in each plot. We observe a mixture of endothermic (C1, E12, G3) and exothermic isotherms. For all cases displayed we observe no binding to the ion for H52A mutants, indicating the scaffolded histidine at position 52 is critical for ion binding. Kd values in the isotherms indicate binding of the ion with the designed stoichiometry (1:4 :protein). Note that each backbone depicted is from a unique RFdiffusion sampling trajectory, and that models and data for designs G3, C10, A5 and E1 from Figure 5 are duplicated here for ease of viewing. B) Size exclusion chromatograms for elutions from the 44 purifications indicate the vast majority of designs are soluble and have the correct oligomeric state. C) Size exclusion chromatograms for mutants show that the mutants remain soluble and retain the intended oligomeric state. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

图S9:额外的Ni 结合C4寡聚体 A) AF2对一部分实验验证的 结合寡聚体的预测，下面是野生型(蓝色)和H52A突变体(粉色)的相应等温滴定量热法(ITC)结合等温线。每个图中显示了野生型的解离常数。我们观察到内热(C1，E12，G3)和外热等温线的混合。在所有显示的情况下，我们观察到H52A突变体对离子没有结合，表明位于52位的支架组氨酸对离子结合至关重要。等温线中的Kd值表明离子与设计的化学计量比(1:4 :蛋白质)结合。请注意，每个描绘的主链来自独特的RF扩散采样轨迹，并且图5中的设计G3、C10、A5和E1的模型和数据在此处重复以便于查看。B) 从44个纯化物的洗脱中得到的尺寸排阻色谱图表明绝大多数设计是可溶的，并且具有正确的寡聚状态。C) 突变体的尺寸排阻色谱图显示突变体保持可溶性并保留预期的寡聚状态。bioRxiv预印本doi:https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审的)是作者/资助者，已授予bioRxiv永久展示该预印本的许可。根据CC-BY-ND 4.0国际许可协议提供。



In Silico Success Rate (%) [raw / +FastRelax]

计算机模拟成功率 (%) [原始 / +快速放松]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Noise Scale | PD-L1 | IL7gc Receptor | Insulin Receptor | TrkA Receptor | Hemagglutinin |
| Unconditional | 0.0 | 39.3/47.6 | 9.1/18.8 |  | 14.7/12.1 | 4.2/3.9 |
| 0.5 | 25.4/29.9 | 6.6/9.9 | 7.1/6.7 | 6.5/6.0 | 0.6/1.0 |
| 1.0 | 10.7/14.4 | 2.1/4.4 | 2.8/3.2 |  | 0.1/0.1 |
| Fold-Conditioned | 0.0 | 44.0/46.0 | 11.7/21.9 | 10.6/10.8 |  |  |
| 0.5 |  | 11.9/15.0 | 3.7/3.7 |  | 3.2/4.7 |
| 1.0 | 22.0/26.0 |  |  | 6.6/7.2 | 0.2/2.3 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 噪声尺度 | PD-L1 | IL7gc 受体 | 胰岛素受体 | TrkA 受体 | 血凝素 |
| 无条件 | 0.0 | 39.3/47.6 | 9.1/18.8 |  | 14.7/12.1 | 4.2/3.9 |
| 0.5 | 25.4/29.9 | 6.6/9.9 | 7.1/6.7 | 6.5/6.0 | 0.6/1.0 |
| 1.0 | 10.7/14.4 | 2.1/4.4 | 2.8/3.2 |  | 0.1/0.1 |
| 折叠条件 | 0.0 | 44.0/46.0 | 11.7/21.9 | 10.6/10.8 |  |  |
| 0.5 |  | 11.9/15.0 | 3.7/3.7 |  | 3.2/4.7 |
| 1.0 | 22.0/26.0 |  |  | 6.6/7.2 | 0.2/2.3 |

Figure S10: Targeted unconditional and fold-conditioned protein binder design. A-B) The ability to specify where on a target a designed binder should bind is crucial. Specific "hotspot" residues can be input to a fine-tuned RFdiffusion model, and with these inputs, binders almost universally target the correct site. A) IL-7Ra (PDB: 3DI3) has two patches that are optimal for binding, denoted Site 1 and Site 2 here. For each site, 100 designs were generated (without fold-specification). B) Without guidance, designs typically target Site 1 (left bar, gray), with contact defined as distance between binder and hotspot reside . Specifying Site 1 hotspot residues increases further the efficiency with which Site 1 is targeted (left bar, pink). In contrast, specifying the Site 2 hotspot residues can completely redirect RFdiffusion, allowing it to efficiently target this site (right bar, pink). C-D) As well as conditioning on hotspot residue information, a fine-tuned RFdiffusion model can also condition on input fold information (secondary structure and block-adjacency information - see Methods 4.5). This effectively allows the specification of a (for instance, particularly compatible) fold that the binder should adopt. C) Two examples showing binders can be specified to adopt either a ferredoxin fold (left) or a particular helical bundle fold (right). D) Quantification of the efficiency of fold-conditioning. Secondary structure inputs were accurately respected (top, pink). Note that in this design target and target site, RFdiffusion without fold-specification made generally helical designs (right, gray bar). Block-adjacency inputs were also respected for both input folds (bottom, pink). E) Reducing the noise added at each step of inference improves the quality of binders designed with RFdiffusion, both with and without fold-conditioning. As an example, the distribution of AF2 interaction pAEs (known to indicate binding when pAE ) is shown for binders designed to PD-L1. In both cases, the proportion of designs with interaction pAE < 10 is high (blue curve), and improved when the noise is scaled by a factor 0.5 (pink curve) or 0 (yellow curve). F) Full in silico success rates for the protein binders designed to five targets. In each case, the best fold-conditioned results are shown (i.e. from the most target-compatible input fold), and the success rates at each noise scale are separated. In line with current best practice , we tested using Rosetta FastRelax before designing the sequence with ProteinMPNN, but found that this did not systematically improve designs. Success is defined in line with current best practice : AF2 pLDDT of the monomer > 80, AF2 interaction pAE < 10, AF2 RMSD monomer vs design < . bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

图 S10:针对无条件和折叠条件的蛋白质结合剂设计。A-B) 指定设计的结合剂应在目标的哪个位置结合至关重要。可以将特定的“热点”残基输入到精细调整的 RFdiffusion 模型中，利用这些输入，结合剂几乎普遍能够针对正确的位置。A) IL-7Ra (PDB: 3DI3) 有两个适合结合的区域，这里标记为位置 1 和位置 2。对于每个位置，生成了 100 个设计(没有折叠规格)。B) 在没有指导的情况下，设计通常针对位置 1(左侧条，灰色)，接触定义为结合剂与热点残基之间的 距离 。指定位置 1 热点残基进一步提高了针对位置 1 的效率(左侧条，粉色)。相反，指定位置 2 热点残基可以完全重新引导 RFdiffusion，使其有效地针对该位置(右侧条，粉色)。C-D) 除了基于热点残基信息进行条件设置外，精细调整的 RFdiffusion 模型还可以基于输入的折叠信息进行条件设置(次级结构和块邻接信息 - 见方法 4.5)。这有效地允许指定结合剂应采用的(例如，特别兼容的)折叠。C) 两个示例显示结合剂可以被指定为采用铁氧还蛋白折叠(左)或特定的螺旋束折叠(右)。D) 折叠条件设置效率的量化。次级结构输入得到了准确的尊重(顶部，粉色)。请注意，在这个设计目标和目标位置中，RFdiffusion 在没有折叠规格的情况下通常生成螺旋设计(右侧，灰色条)。块邻接输入也得到了对两个输入折叠的尊重(底部，粉色)。E) 在推理的每一步减少添加的噪声提高了使用 RFdiffusion 设计的结合剂的质量，无论是否进行折叠条件设置。作为示例，显示了针对 PD-L1 设计的结合剂的 AF2 交互 pAEs 的分布(已知在 pAE 时指示结合)。在这两种情况下，交互 pAE < 10 的设计比例较高(蓝色曲线)，并且在噪声缩放因子为 0.5(粉色曲线)或 0(黄色曲线)时有所改善。F) 针对五个目标设计的蛋白质结合剂的完整计算成功率。在每种情况下，显示了最佳的折叠条件结果(即来自最兼容目标的输入折叠)，并且在每个噪声尺度下的成功率被分开。根据当前最佳实践 ，我们在使用 ProteinMPNN 设计序列之前测试了使用 Rosetta FastRelax ，但发现这并没有系统性地改善设计。成功的定义符合当前最佳实践 :单体的 AF2 pLDDT > 80，AF2 交互 pAE < 10，AF2 单体与设计的 RMSD < 。bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842；此版本于 2022 年 12 月 10 日发布。该预印本的版权持有者(未经过同行评审的)是作者/资助者，已授予 bioRxiv 永久展示该预印本的许可。根据 CC-BY-ND 4.0 国际许可证提供。

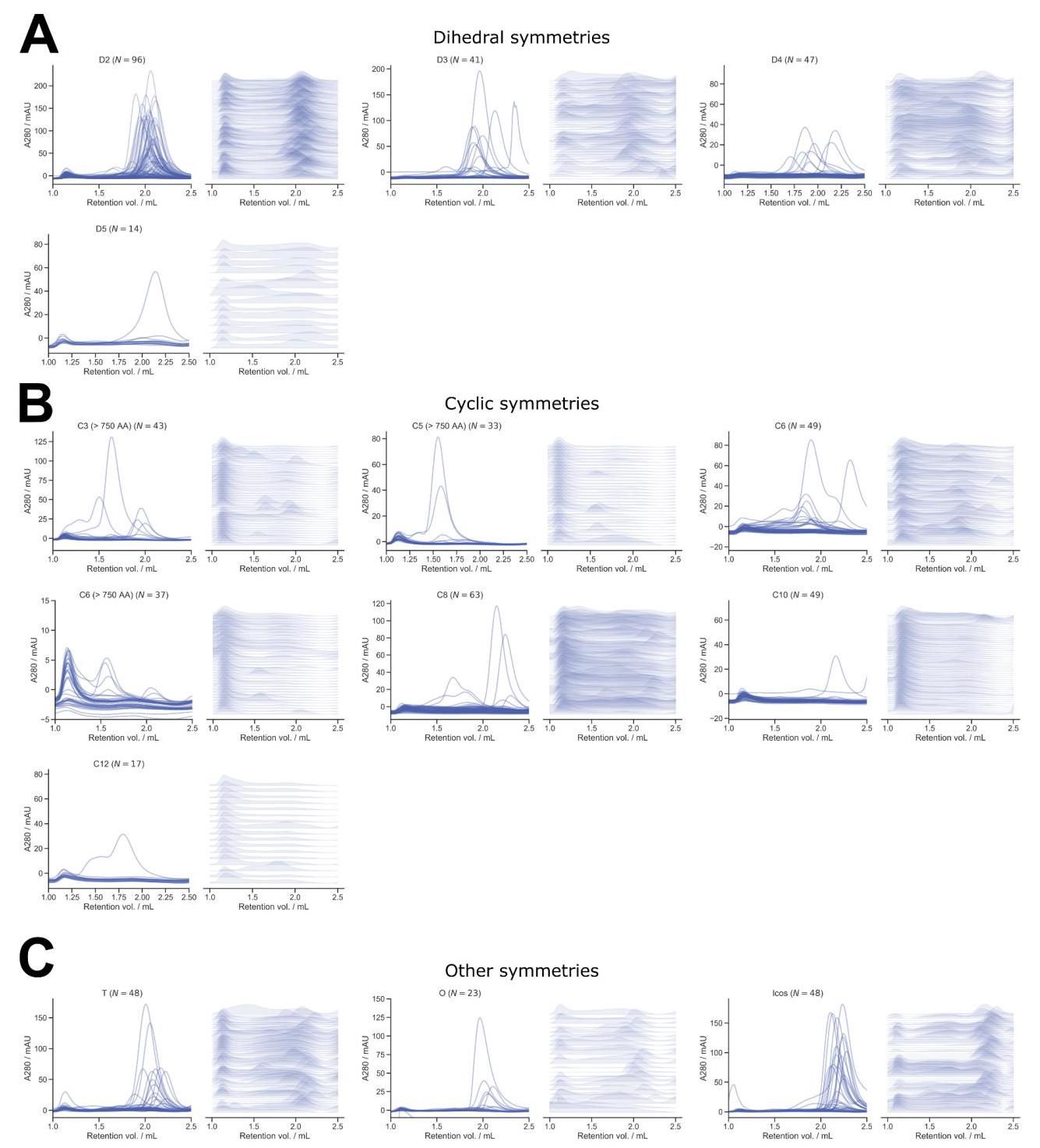
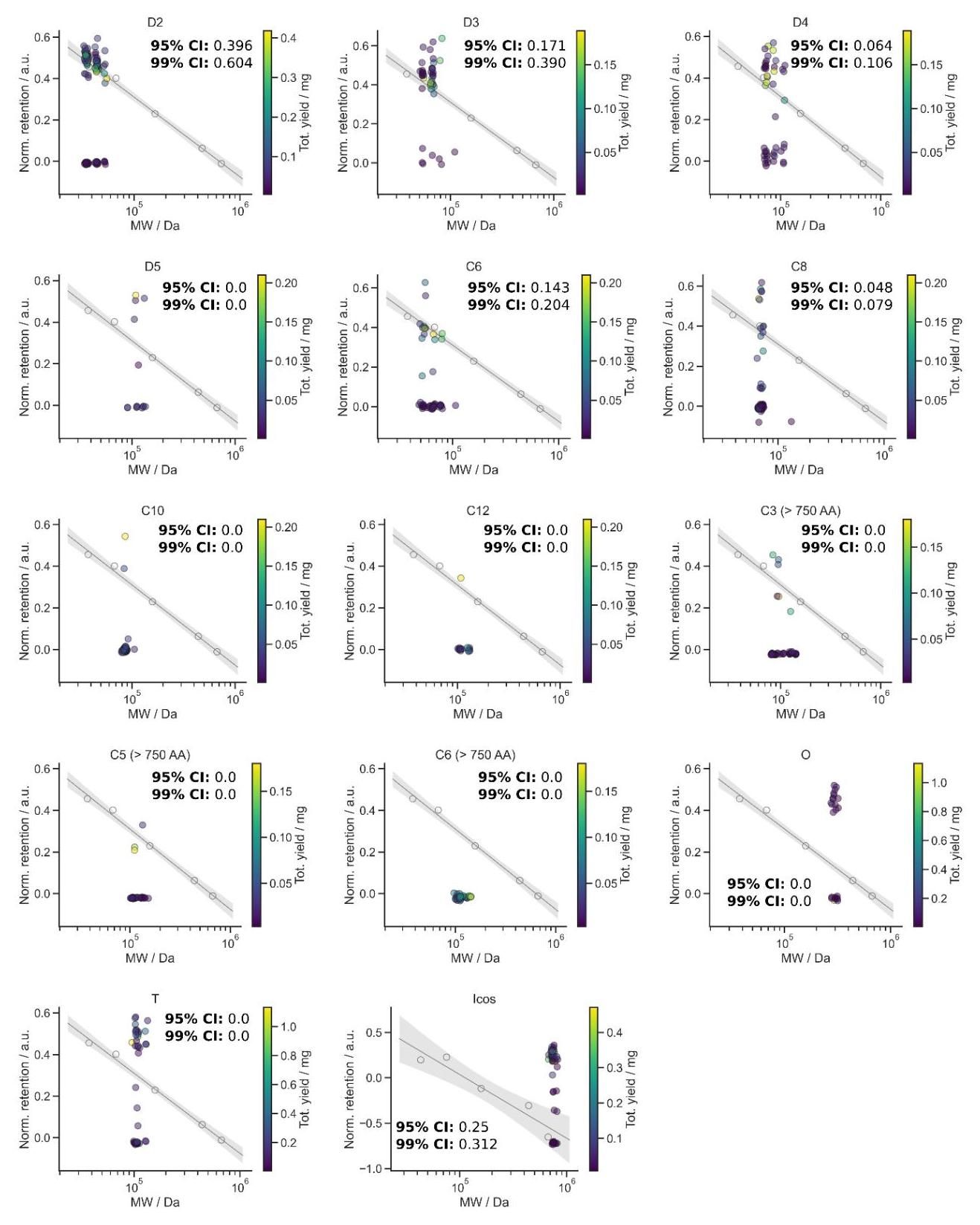


Figure S11: Size exclusion chromatography of symmetric oligomers. A-C) Size exclusion chromatography (SEC) was used as a primary screening method for all RFdiffusion-generated oligomers. Here, SEC traces from 608 oligomers are shown for each of the experimentally tested symmetry groups, excluding the void volume. Panel A) shows dihedral symmetries, B) shows cyclic symmetries, and C) shows all others. For each set of traces, on the left, data are overlaid for all designs, and on the right, traces are normalized and stacked. As designs increase in complexity (higher number of individual subunits), the amount of soluble protein shown by SEC visibly decreases. For tetrahedral, octahedral, and icosahedral designs, many have soluble protein peaks that are possibly dimer and trimer subunits (unassembled cages).

图 S11:对称寡聚物的尺寸排斥色谱。A-C) 尺寸排斥色谱(SEC)被用作所有 RFdiffusion 生成的寡聚物的主要筛选方法。在这里，展示了608个寡聚物在每个实验测试的对称性组中的 SEC 曲线，排除了空体积。面板 A) 显示了二面角对称性，B) 显示了循环对称性，C) 显示了其他所有类型。对于每组曲线，左侧是所有设计的数据叠加，右侧是归一化并堆叠的曲线。随着设计复杂性的增加(单个亚单位数量的增加)，SEC 显示的可溶性蛋白质的量明显减少。对于四面体、八面体和二十面体设计，许多具有可溶性蛋白质峰，可能是二聚体和三聚体亚单位(未组装的笼子)。

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842; 此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审认证)是作者/资助方，他们已授予bioRxiv永久展示该预印本的许可。该文献根据CC-BY-ND 4.0国际许可协议提供。

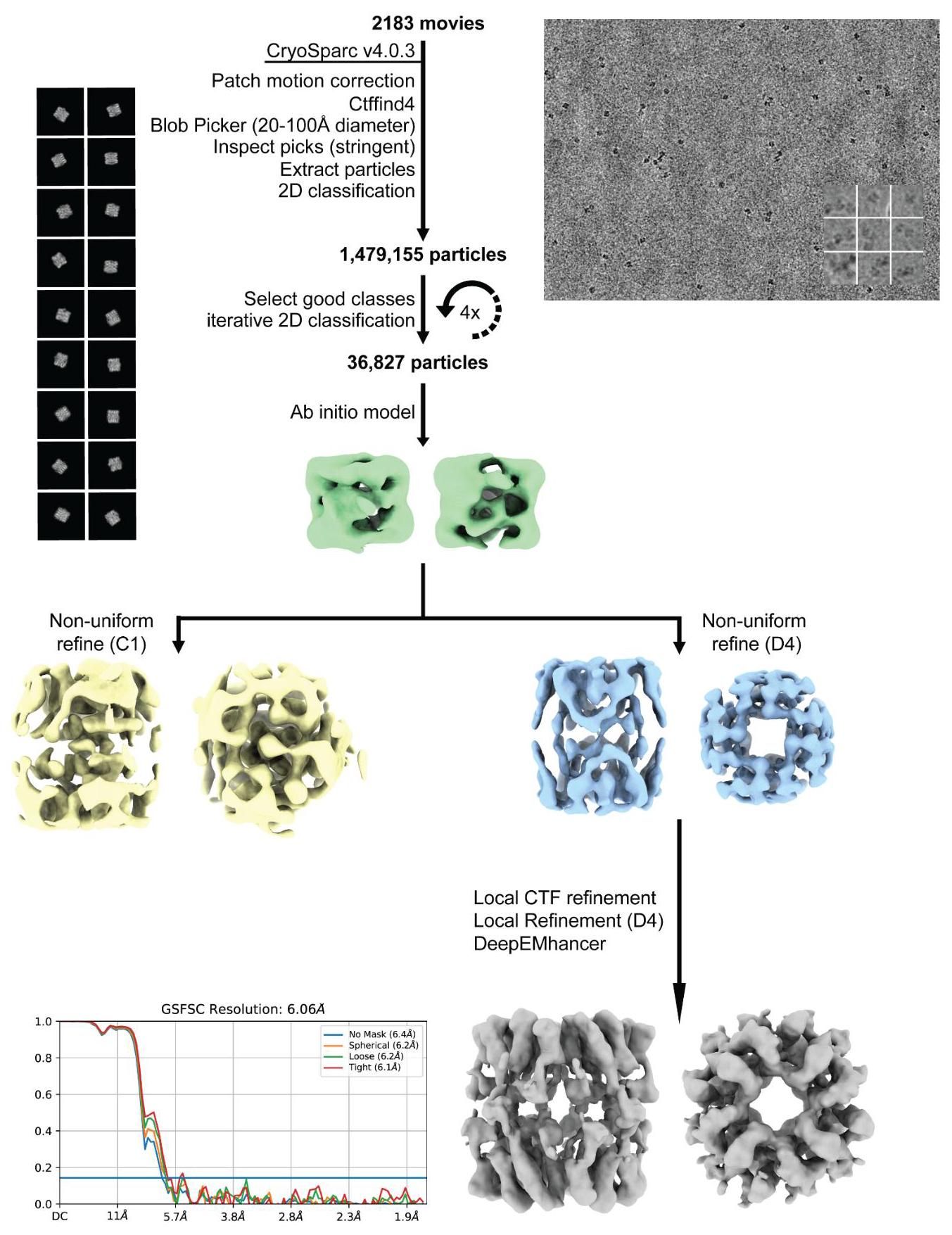


bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842; 此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审认证)是作者/资助方，他们已授予bioRxiv永久展示该预印本的许可。该文献根据CC-BY-ND 4.0国际许可协议提供。

Figure S12: SEC elution peaks of symmetric oligomers vs. calibration curves. Retention volume for the major SEC peak versus molecular weight for each design are plotted in comparison to a known calibration curve. The calibration curve is shown in gray, with shading representing the 95% confidence interval. Total yield of each design is indicated by the scale bar on the right of the graphs, and success rates for the 95% CI and 99% CI are denoted on each graph per each symmetry. Given that MW is being used as a proxy for hydrodynamic radius, we expect that some designs (e.g. cycles with large pores) may be true to their design model, but deviate from the standard curve. These calibration curves provide a rough estimate of the success rate of each symmetry group, and help guide the selection process for downstream analysis of any design. In some cases, even though no designs are within the , we still selected designs to screen by nsEM. For example, we are able to confirm HE0822 (C3) by nsEM despite misalignment between the theoretical and actual elution profiles (Fig. 3B). Because of their size, the icosahedra were run on an S6 column with lower resolution; thus, the calibration curve fit results in bigger confidence intervals compared to an S200 column, which was used to screen all other oligomers (See Methods 6.2). We expect that for oligomers run on the S200, reported success rates are fairly conservative, whereas for designs run on the S6, experimental success rates are likely lower. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

图 S12:对称寡聚物的 SEC 洗脱峰与校准曲线。主要 SEC 峰的保留体积与每个设计的分子量进行比较，并与已知的校准曲线进行对比。校准曲线以灰色显示，阴影部分代表 95% 置信区间。每个设计的总产量由图表右侧的比例尺指示，95% CI 和 99% CI 的成功率在每个图表上按对称性标注。考虑到分子量(MW)被用作流体动力半径的代理，我们预计某些设计(例如具有大孔的循环)可能与其设计模型相符，但偏离标准曲线。这些校准曲线提供了每个对称性组成功率的粗略估计，并帮助指导下游分析中任何设计的选择过程。在某些情况下，即使没有设计在 范围内，我们仍然选择设计进行 nsEM 筛选。例如，尽管理论与实际洗脱曲线之间存在不对齐，我们仍然能够通过 nsEM 确认 HE0822 (C3) 的存在(图 3B)。由于其尺寸，二十面体在 S6 柱上运行，分辨率较低；因此，校准曲线的拟合结果导致与用于筛选所有其他寡聚物的 S200 柱相比，置信区间更大。我们预计，对于在 S200 上运行的寡聚物，报告的成功率相对保守，而对于在 S6 上运行的设计，实验成功率可能更低。bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842；此版本于 2022 年 12 月 10 日发布。该预印本的版权持有者(未经过同行评审的)是作者/资助者，他们已授予 bioRxiv 永久展示该预印本的许可。该文档根据 CC-BY-ND 4.0 国际许可证提供。



bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审认证)是作者/资助方，他们已授予 bioRxiv 永久展示该预印本的许可。该文档根据 aCC-BY-ND 4.0 国际许可证提供。

Figure S13: Details of HE0537 cryo-EM data processing pipeline. 2D class averages showing exclusively side-views of HE0537, and an ab initio reconstruction followed by a C1 non-uniform refinement yielding identifiable D4 features corresponding to the size and rough secondary structure of the design model. Further data processing was attempted with D4 symmetry imposed, but the strong preferred orientation precluded generation of a reliable 3D map for detailed structural analysis. At this time, only the predicted 2D projection images of the design model are analyzed/compared alongside the corresponding experimental cryo-EM 2D class average side views in Fig. 3D, which display strikingly high agreement to the design. A representative raw cryo-EM micrograph is shown on the right along with nine example extracted particles and characteristic 2D class averages used in the processing pipeline. An FSC validation curve for the final reconstruction is shown along with the density map. (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made

图 S13:HE0537 冷冻电子显微镜数据处理流程的详细信息。2D 类别平均图仅显示 HE0537 的侧视图，以及一个初步重建，随后进行 C1 非均匀精细化，产生可识别的 D4 特征，对应于设计模型的大小和粗略的二级结构。进一步的数据处理尝试在施加 D4 对称性的情况下进行，但强烈的优选取向阻碍了生成可靠的 3D 图以进行详细的结构分析。目前，仅分析/比较设计模型的预测 2D 投影图像与图 3D 中相应的实验冷冻电子显微镜 2D 类别平均侧视图，这些视图与设计显示出惊人的高度一致性。右侧展示了一幅代表性的原始冷冻电子显微镜显微照片，以及九个提取的粒子示例和用于处理流程的特征 2D 类别平均图。最终重建的 FSC 验证曲线与密度图一起显示。(未经过同行评审)是作者/资助者，已授予 bioRxiv 永久展示该预印本的许可。它被制作

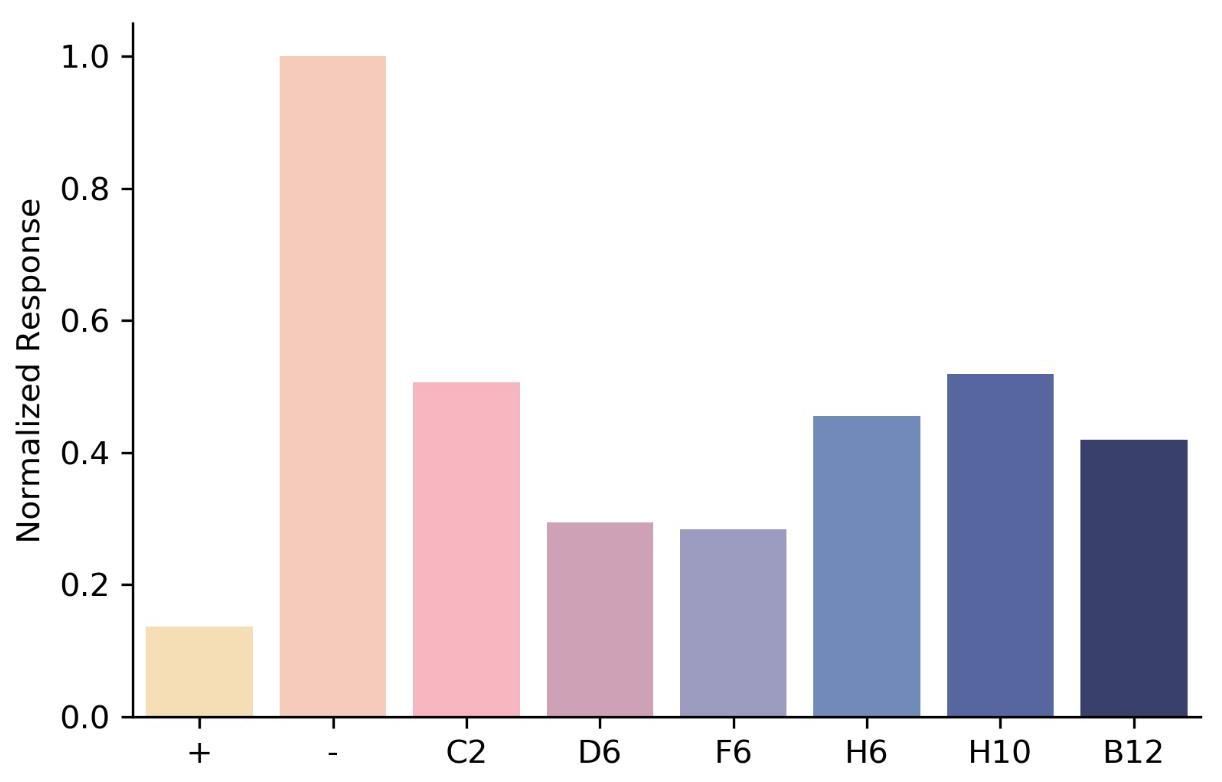


Figure S14: IL-7Ra Competition Assay

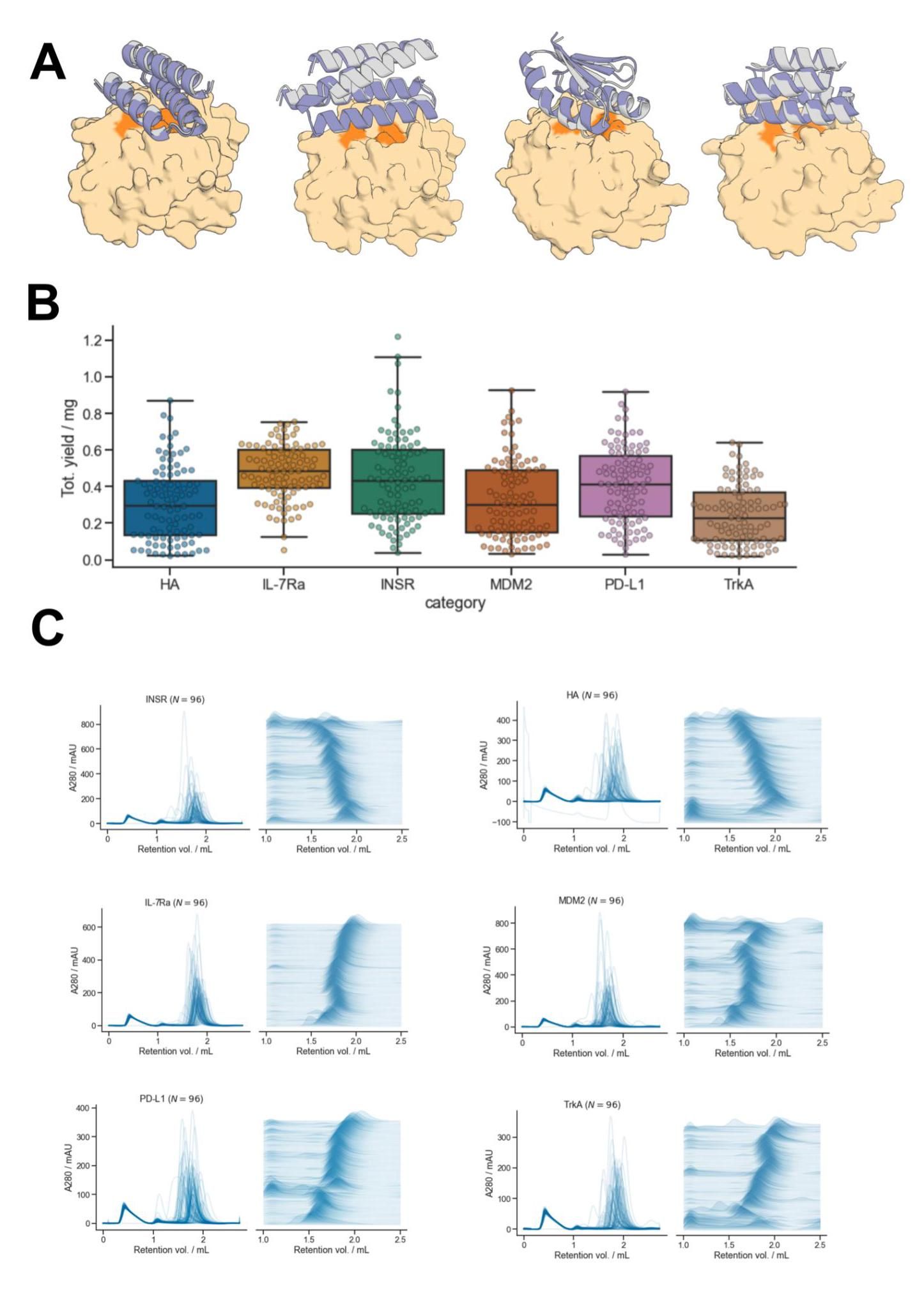
图 S14:IL-7Ra 竞争测定

Positive control (known IL-7Ra binder from ref [ ]) was amine conjugated to ar2g biosensor tips. 100nM IL-7Ra with 1μM of each design then was used as analyte. Positive control was also included as an analyte as there should be no binding. Response is normalized to binding of IL-7Ra on its own. All six diffusion-generated binders compete with the positive control, indicating they bind to the intended site.

阳性对照(来自参考文献[ ]的已知IL-7Ra结合物)与ar2g生物传感器尖端进行胺结合。然后使用100nM IL-7Ra和每种设计的1μM作为分析物。阳性对照也作为分析物被包含，因为应该没有结合。响应归一化为IL-7Ra自身的结合。所有六种扩散生成的结合物与阳性对照竞争，表明它们结合到预期的位点。

(which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(未经过同行评审认证)是作者/资助者，已授予bioRxiv永久展示该预印本的许可。该文档在CC-BY-ND 4.0国际许可下提供。



bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint

bioRxiv预印本doi: https://doi.org/10.1101/2022.12.09.519842；该版本于2022年12月10日发布。该预印本的版权持有者

# Figure S15: Analysis of Insulin Receptor Binding Campaign

# 图S15:胰岛素受体结合实验分析

A) Insulin Receptor binders are well-predicted by AF2. Yellow/orange: target/hotspot residues; gray: design model; purple: AF2 prediction. B) Insulin Receptor binders are expressed with high yield, in line with the rest of the binder campaigns. C) SEC elution profiles indicate most Insulin Receptor binders elute as monomers, in line with the rest of the binder campaigns. bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

A) 胰岛素受体结合物的预测结果良好，符合AF2。黄色/橙色:目标/热点残基；灰色:设计模型；紫色:AF2预测。B) 胰岛素受体结合物的表达产量高，与其他结合物实验一致。C) SEC洗脱曲线表明大多数胰岛素受体结合物以单体形式洗脱，与其他结合物实验一致。bioRxiv预印本doi:https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审)为作者/资助者，已授予bioRxiv永久展示该预印本的许可。该文档根据CC-BY-ND 4.0国际许可证提供。

# Bibliography

# 参考文献

1. Dauparas, J. et al. Robust deep learning-based protein sequence design using

1. Dauparas, J. 等。基于深度学习的蛋白质序列设计的稳健方法

ProteinMPNN. Science eadd2187 (2022) doi:10.1126/science.add2187.

ProteinMPNN。科学 eadd2187 (2022) doi:10.1126/science.add2187。

2. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. Nat. Commun. 13, 4348 (2022).

2. Ferruz, N., Schmidt, S. 和 Höcker, B. ProtGPT2 是一种用于蛋白质设计的深度无监督语言模型。自然通讯 13, 4348 (2022)。

3. Singer, J. M. et al. Large-scale design and refinement of stable proteins using sequence-only models. PLOS ONE 17, e0265020 (2022).

3. Singer, J. M. 等. 使用仅基于序列的模型进行大规模稳定蛋白质的设计和优化。PLOS ONE 17, e0265020 (2022)。

4. Wang, J. et al. Scaffolding protein functional sites using deep learning. Science 377, 387-394 (2022).

4. Wang, J. 等. 利用深度学习构建蛋白质功能位点的支架。科学 377, 387-394 (2022)。

5. Trippe, B. L. et al. Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. (2022) doi:10.48550/ARXIV.2206.04119.

5. Trippe, B. L. 等. 针对基序支架问题的三维蛋白质骨架的扩散概率建模。(2022) doi:10.48550/ARXIV.2206.04119。

6. Anishchenko, I. et al. De novo protein design by deep network hallucination. Nature 600, 547-552 (2021).

6. Anishchenko, I. 等. 通过深度网络幻觉进行新型蛋白质设计。自然 600, 547-552 (2021)。

7. Wicky, B. I. M. et al. Hallucinating symmetric protein assemblies. Science 378, 56-61 (2022).

7. Wicky, B. I. M. 等. 幻想对称蛋白质组装。科学 378, 56-61 (2022)。

8. Anand, N. & Achim, T. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. (2022) doi:10.48550/ARXIV.2205.15019.

8. Anand, N. 和 Achim, T. 采用等变去噪扩散概率模型生成蛋白质结构和序列。(2022) doi:10.48550/ARXIV.2205.15019.

9. Luo, S. et al. Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models. 13.

9. Luo, S. 等. 基于扩散的生成模型进行抗原特异性抗体设计和优化。13.

10. Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N. & Ganguli, S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. Preprint at https://doi.org/10.48550/arXiv.1503.03585 (2015).

10. Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N. 和 Ganguli, S. 使用非平衡热力学的深度无监督学习。预印本见 https://doi.org/10.48550/arXiv.1503.03585 (2015).

11. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. Preprint at https://doi.org/10.48550/arXiv.2006.11239 (2020).

11. Ho, J., Jain, A. 和 Abbeel, P. 去噪扩散概率模型。预印本见 https://doi.org/10.48550/arXiv.2006.11239 (2020).

12. Cao, L. et al. Design of protein-binding proteins from the target structure alone. Nature 605, 551-560 (2022).

12. Cao, L. 等. 仅从目标结构设计蛋白质结合蛋白。自然 605, 551-560 (2022).

13. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. Science 302, 1364-1368 (2003).

13. Kuhlman, B. 等. 设计一种具有原子级精度的新型球状蛋白质折叠. 科学 302, 1364-1368 (2003).

14. Ramesh, A. et al. Zero-Shot Text-to-Image Generation. Preprint at http://arxiv.org/abs/2102.12092 (2021).

14. Ramesh, A. 等. 零样本文本到图像生成. 预印本于 http://arxiv.org/abs/2102.12092 (2021).

15. Saharia, C. et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Preprint at http://arxiv.org/abs/2205.11487 (2022).

15. Saharia, C. 等. 具有深度语言理解的照片级真实文本到图像扩散模型. 预印本于 http://arxiv.org/abs/2205.11487 (2022).

16. Wu, K. E. et al. Protein structure generation via folding diffusion. (2022) doi:10.48550/arXiv.2209.15611.

16. Wu, K. E. 等. 通过折叠扩散生成蛋白质结构. (2022) doi:10.48550/arXiv.2209.15611.

17. Watson, J. L., Bera, A., Juergens, D., Wang, J. & Baker, D. X-ray crystallographic validation of design from this paper | Science | AAAS. (2022).

17. Watson, J. L., Bera, A., Juergens, D., Wang, J. 和 Baker, D. 本文设计的X射线晶体学验证 | 科学 | 美国科学促进会. (2022).

18. Berman, H. M. et al. The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000).

18. Berman, H. M. 等. 蛋白质数据银行. 核酸研究. 28, 235–242 (2000).

19. De Bortoli, V. et al. Riemannian Score-Based Generative Modelling. Preprint at https://doi.org/10.48550/arXiv.2202.02763 (2022).

19. De Bortoli, V. 等. 黎曼评分基础的生成建模. 预印本于 https://doi.org/10.48550/arXiv.2202.02763 (2022).

20. Leach, A., Schmon, S. M., Degiacomi, M. T. & Willcocks, C. G. Denoising Diffusion Probabilistic Models On SO(3) For Rotational Alignment. 8 (2022).

20. Leach, A., Schmon, S. M., Degiacomi, M. T. 和 Willcocks, C. G. 在SO(3)上去噪扩散概率模型用于旋转对齐. 8 (2022).

21. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583-589 (2021).

21. Jumper, J. 等. 使用 AlphaFold 进行高度准确的蛋白质结构预测. 自然 596, 583-589 (2021).

22. Wu, R. et al. High-resolution de novo structure prediction from primary sequence. 2022.07.21.500999 Preprint at https://doi.org/10.1101/2022.07.21.500999 (2022).

22. Wu, R. 等. 从初级序列进行高分辨率的 de novo 结构预测. 2022.07.21.500999 预印本，网址:https://doi.org/10.1101/2022.07.21.500999 (2022).

23. Lin, Z. et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. 31.

23. Lin, Z. 等. 蛋白质序列的语言模型在进化规模上使准确的结构预测成为可能. 31.

24. Chen, T., Zhang, R. & Hinton, G. Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning. Preprint at https://doi.org/10.48550/arXiv.2208.04202 (2022).

24. Chen, T., Zhang, R. & Hinton, G. 模拟位: 使用自我条件的扩散模型生成离散数据. 预印本，网址:https://doi.org/10.48550/arXiv.2208.04202 (2022).

25. Lee, J. S. & Kim, P. M. ProteinSGM: Score-based generative modeling for de novo protein design. bioRxiv (2022) doi:10.1101/2022.07.13.499967.

25. Lee, J. S. & Kim, P. M. ProteinSGM: 基于评分的生成建模用于 de novo 蛋白质设计. bioRxiv (2022) doi:10.1101/2022.07.13.499967.

26. Bennett, N. et al. Improving de novo Protein Binder Design with Deep Learning. 2022.06.15.495993 Preprint at https://doi.org/10.1101/2022.06.15.495993 (2022).

26. Bennett, N. 等. 利用深度学习改进新型蛋白质结合体设计. 2022.06.15.495993 预印本于 https://doi.org/10.1101/2022.06.15.495993 (2022).

27. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of Protein Folding: The Energy Landscape Perspective. Annu. Rev. Phys. Chem. 48, 545-600 (1997).

27. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. 蛋白质折叠理论:能量景观视角. 年度化学物理评论 48, 545-600 (1997).

28. Jendrusch, M., Korbel, J. O. & Sadiq, S. K. AlphaDesign: A de novo protein design framework based on AlphaFold. 2021.10.11.463937 Preprint at https://doi.org/10.1101/2021.10.11.463937 (2021).

28. Jendrusch, M., Korbel, J. O. & Sadiq, S. K. AlphaDesign:基于AlphaFold的新型蛋白质设计框架. 2021.10.11.463937 预印本于 https://doi.org/10.1101/2021.10.11.463937 (2021).

29. Anand, N. & Huang, P. Generative modeling for protein structures. in Advances in Neural Information Processing Systems vol. 31 (Curran Associates, Inc., 2018).

29. Anand, N. & Huang, P. 蛋白质结构的生成建模. 见《神经信息处理系统进展》第31卷 (Curran Associates, Inc., 2018).

30. Basanta, B. et al. An enumerative algorithm for de novo design of proteins with diverse pocket structures. Proc. Natl. Acad. Sci. 117, 22135–22145 (2020).

30. Basanta, B. 等. 一种用于新型蛋白质多样口袋结构设计的枚举算法. 美国国家科学院院刊 117, 22135–22145 (2020).

31. Pan, X. et al. Expanding the space of protein geometries by computational design of de novo fold families. Science 369, 1132-1136 (2020).

31. Pan, X. 等. 通过计算设计新型折叠家族扩展蛋白质几何空间. 科学 369, 1132-1136 (2020).

32. Marcandalli, J. et al. Induction of Potent Neutralizing Antibody Responses by a Designed Protein Nanoparticle Vaccine for Respiratory Syncytial Virus. Cell 176, 1420-1431.e17 (2019).

32. Marcandalli, J. 等. 通过设计的蛋白质纳米颗粒疫苗诱导强效中和抗体反应以对抗呼吸道合胞病毒. 细胞 176, 1420-1431.e17 (2019).

33. Butterfield, G. L. et al. Evolution of a designed protein assembly encapsulating its own RNA genome. Nature 552, 415-420 (2017).

33. Butterfield, G. L. 等. 设计的蛋白质组装体的进化，封装其自身的RNA基因组. 自然 552, 415-420 (2017).

34. Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. Annu. Rev. Biophys. Biomol. Struct. 29, 105-153 (2000).

34. Goodsell, D. S. & Olson, A. J. 结构对称性与蛋白质功能. 生物物理学与生物大分子结构年鉴 29, 105-153 (2000).

35. Sterner, R. & Höcker, B. Catalytic Versatility, Stability, and Evolution of the (βα)8-Barrel Enzyme Fold. Chem. Rev. 105, 4038-4055 (2005).

35. Sterner, R. & Höcker, B. 催化多样性、稳定性及(βα)8-桶状酶折叠的演化。化学评论 105, 4038-4055 (2005)。

36. Sesterhenn, F. et al. De novo protein design enables the precise induction of RSV-neutralizing antibodies. Science 368, (2020).

36. Sesterhenn, F. 等. 新型蛋白质设计使得精确诱导RSV中和抗体成为可能。科学 368, (2020)。

37. Yang, C. et al. Bottom-up de novo design of functional proteins with complex structural features. Nat. Chem. Biol. 17, 492-500 (2021).

37. Yang, C. 等. 自下而上的新型功能蛋白设计，具有复杂的结构特征。自然化学生物学 17, 492-500 (2021)。

38. Glasgow, A. et al. Engineered ACE2 receptor traps potently neutralize SARS-CoV-2. Proc. Natl. Acad. Sci. U. S. A. 117, 28046-28055 (2020).

38. Glasgow, A. 等. 工程化ACE2受体强效中和SARS-CoV-2。美国国家科学院院刊 117, 28046-28055 (2020)。

39. Chène, P. Inhibiting the p53-MDM2 interaction: an important target for cancer therapy. Nat. Rev. Cancer 3, 102-109 (2003).

39. Chène, P. 抑制p53-MDM2相互作用:癌症治疗的重要靶点。自然癌症评论 3, 102-109 (2003)。

40. Kussie, P. H. et al. Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. Science 274, 948-953 (1996).

40. Kussie, P. H. 等. MDM2癌蛋白与p53肿瘤抑制因子转激活域结合的结构. 科学 274, 948-953 (1996).

41. Hunt, A. C. et al. Multivalent designed proteins neutralize SARS-CoV-2 variants of concern and confer protection against infection in mice. Sci. Transl. Med. 14, eabn1252 (2022).

41. Hunt, A. C. 等. 多价设计蛋白中和SARS-CoV-2变异株并在小鼠中提供感染保护. 科学转化医学 14, eabn1252 (2022).

42. Silverman, J. et al. Multivalent avimer proteins evolved by exon shuffling of a family of human receptor domains. Nat. Biotechnol. 23, 1556-1561 (2005).

42. Silverman, J. 等. 通过外显子重组进化的多价avimer蛋白，源自一系列人类受体结构域. 自然生物技术 23, 1556-1561 (2005).

43. Detalle, L. et al. Generation and Characterization of ALX-0171, a Potent Novel Therapeutic Nanobody for the Treatment of Respiratory Syncytial Virus Infection. Antimicrob. Agents Chemother. 60, 6–13 (2016).

43. Detalle, L. 等. ALX-0171的生成与特征分析，一种用于治疗呼吸道合胞病毒感染的强效新型治疗纳米抗体. 抗微生物药物化疗 60, 6–13 (2016).

44. Strauch, E.-M. et al. Computational design of trimeric influenza-neutralizing proteins targeting the hemagglutinin receptor binding site. Nat. Biotechnol. 35, 667–671 (2017).

44. Strauch, E.-M. 等. 计算设计针对血凝素受体结合位点的三聚体流感中和蛋白。自然生物技术 35, 667–671 (2017).

45. Boyoglu-Barnum, S. et al. Quadrivalent influenza nanoparticle vaccines induce broad protection. Nature 592, 623-628 (2021).

45. Boyoglu-Barnum, S. 等. 四价流感纳米颗粒疫苗诱导广泛保护。自然 592, 623-628 (2021).

46. Walls, A. C. et al. Elicitation of Potent Neutralizing Antibody Responses by Designed Protein Nanoparticle Vaccines for SARS-CoV-2. Cell 183, 1367-1382.e17 (2020).

46. Walls, A. C. 等. 通过设计的蛋白质纳米颗粒疫苗诱导针对SARS-CoV-2的强效中和抗体反应。细胞 183, 1367-1382.e17 (2020).

47. Quijano-Rubio, A., Ulge, U. Y., Walkey, C. D. & Silva, D.-A. The advent of de novo proteins for cancer immunotherapy. Curr. Opin. Chem. Biol. 56, 119-128 (2020).

47. Quijano-Rubio, A., Ulge, U. Y., Walkey, C. D. & Silva, D.-A. 新型蛋白质在癌症免疫治疗中的出现。当前化学生物学意见 56, 119-128 (2020).

48. Chevalier, A. et al. Massively parallel de novo protein design for targeted therapeutics. Nature 550, 74-79 (2017).

48. Chevalier, A. 等. 大规模平行的去新蛋白设计用于靶向治疗。自然 550, 74-79 (2017).

49. Baek, M., McHugh, R., Anishchenko, I., Baker, D. & DiMaio, F. Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. 2022.09.09.507333 Preprint at https://doi.org/10.1101/2022.09.09.507333 (2022).

49. Baek, M., McHugh, R., Anishchenko, I., Baker, D. & DiMaio, F. 使用RoseTTAFoldNA准确预测核酸和蛋白质-核酸复合物。2022.09.09.507333 预印本于 https://doi.org/10.1101/2022.09.09.507333 (2022).

50. Ribeiro, A. J. M. et al. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. Nucleic Acids Res. 46, D618-D623 (2018).

50. Ribeiro, A. J. M. 等. 机制与催化位点图谱 (M-CSA):酶反应机制和活性位点的数据库。核酸研究 46, D618-D623 (2018)。

51. Shapovalov, M. V. & Dunbrack, R. L. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. Structure 19, 844-858 (2011).

51. Shapovalov, M. V. & Dunbrack, R. L. 基于平滑骨架依赖的蛋白质旋转体库，源自自适应核密度估计和回归。结构 19, 844-858 (2011)。

52. Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 487, 545–574 (2011). bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

52. Leaver-Fay, A. 等. ROSETTA3:用于大分子模拟和设计的面向对象软件套件。酶学方法 487, 545–574 (2011)。bioRxiv 预印本 doi: https://doi.org/10.1101/2022.12.09.519842；此版本于2022年12月10日发布。该预印本的版权持有者(未经过同行评审)为作者/资助者，已授予bioRxiv永久展示该预印本的许可。根据CC-BY-ND 4.0国际许可协议提供。