# A high-level programming language for generative protein design

# 用于生成蛋白质设计的高级编程语言

Brian Hie Salvatore Candido Zeming Lin Ori Kabeli Roshan Rao Nikita Smetanin Tom Sercu Alexander Rives

布莱恩·希 萨尔瓦托雷·坎迪多 林泽明 奥里·卡贝利 罗山·拉奥 尼基塔·斯梅塔宁 汤姆·塞尔库 亚历山大·里夫斯

# Abstract

# 摘要

Combining a basic set of building blocks into more complex forms is a universal design principle. Most protein designs have proceeded from a manual bottom-up approach using parts created by nature, but top-down design of proteins is fundamentally hard due to biological complexity. We demonstrate how the modularity and programmability long sought for protein design can be realized through generative artificial intelligence. Advanced protein language models demonstrate emergent learning of atomic resolution structure and protein design principles. We leverage these developments to enable the programmable design of de novo protein sequences and structures of high complexity. First, we describe a high-level programming language based on modular building blocks that allows a designer to easily compose a set of desired properties. We then develop an energy-based generative model, built on atomic resolution structure prediction with a language model, that realizes all-atom structure designs that have the programmed properties. Designing a diverse set of specifications, including constraints on atomic coordinates, secondary structure, symmetry, and multimerization, demonstrates the generality and controllability of the approach. Enumerating constraints at increasing levels of hierarchical complexity shows that the approach can access a combinatorially large design space.

将一组基本构建模块组合成更复杂的形式是一种普遍的设计原则。大多数蛋白质设计都是从使用自然界创造的部分进行手动自下而上的方法开始的，但由于生物复杂性，自上而下的蛋白质设计从根本上来说是困难的。我们展示了如何通过生成式人工智能实现长期以来在蛋白质设计中追求的模块化和可编程性。先进的蛋白质语言模型展示了原子分辨率结构和蛋白质设计原则的涌现学习。我们利用这些进展来实现高复杂度的从头蛋白质序列和结构的可编程设计。首先，我们描述了一种基于模块化构建块的高级编程语言，允许设计者轻松组合一组所需属性。然后，我们开发了一种基于能量的生成模型，该模型建立在语言模型的原子分辨率结构预测基础上，实现了具有编程属性的全原子结构设计。设计包括原子坐标、二级结构、对称性和多聚化约束在内的多样化规格，展示了该方法的通用性和可控性。在不断增加层次复杂性的约束条件下进行枚举表明，该方法可以访问组合庞大的设计空间。

# Introduction

# 引言

Protein design would benefit from the regularity, simplicity, and programmability provided by a basic set of abstractions (1-4) like those used in the engineering of buildings, ma-

蛋白质设计将受益于一组基本抽象(1-4)所提供的规律性、简单性和可编程性，这些抽象类似于建筑工程中使用的那些。

Preprint. Copyright 2022 by the authors. chines, circuits, and computer software. But unlike these artificial creations, proteins cannot be decomposed into easily recombinable parts because the local structure of the sequence is entangled in its global context(5,6). Classical de novo protein design has attempted to determine a fundamental set of structural building blocks, which could then be assembled into higher-order structures (7-11). Likewise, traditional protein engineering often recombines segments or domains of natural protein sequences into hybrid chimeras (12-14). However, existing approaches have not been able to achieve the high combinatorial complexity that is necessary for true programmability.

预印本。版权归作者所有，2022年。中文、电路和计算机软件。但与这些人工创造物不同，蛋白质不能被分解成易于重组的部分，因为序列的局部结构与其全局背景相互纠缠(5,6)。经典的从头蛋白质设计试图确定一组基本的结构构建模块，然后将其组装成更高层次的结构(7-11)。同样，传统的蛋白质工程通常将天然蛋白质序列的片段或结构域重组为混合嵌合体(12-14)。然而，现有方法尚未能实现真正可编程所需的高组合复杂性。

We show modern generative models realize these classical goals of modularity and programmability at a new level of combinatorial complexity. Our idea is to place the modularity and programmability at a higher level of abstraction, where a generative model bridges the gap between human intuition and the production of specific sequences and structures. In this setting, the protein designer needs only to recombine high-level directives, while the task of obtaining a protein that fulfills those directives is placed on the generative model.

我们展示了现代生成模型在新的组合复杂性水平上实现了这些模块化和可编程性的经典目标。我们的想法是将模块化和可编程性置于更高的抽象层次，生成模型在此层次上弥合了人类直觉与特定序列和结构生成之间的差距。在这种设置下，蛋白质设计者只需重新组合高级指令，而生成模型则负责获取满足这些指令的蛋白质。

We propose a programming language for generative protein design, which allows a designer to specify intuitive, modular, and hierarchical programs. We show that high-level programs can be translated into low-level sequences and structures by a generative model. Our approach leverages advances in protein language models, which learn structural information(15,16)and the design principles of proteins (see accompanying paper by Verkuil et al.).

我们提出了一种用于生成蛋白质设计的编程语言，该语言允许设计者指定直观、模块化和层次化的程序。我们展示了高级程序可以通过生成模型转化为低级序列和结构。我们的方法利用了蛋白质语言模型的进展，这些模型学习了结构信息(15,16)以及蛋白质的设计原则(参见Verkuil等人的伴随论文)。

In this study, our specific implementation is based on an energy-based generative model. First, a protein designer specifies a high-level program consisting of a set of hierarchically organized constraints (Figure 1A). Then, this program compiles to an energy function that evaluates compatibility with the constraints, which can be arbitrary and non-differentiable (Figure 1B). We apply constraints on structure by incorporating atomic-level structure predictions, enabled by a language model, into the energy function. This approach enables the generation of a wide set of complex designs (Figure 1C).

在本研究中，我们的具体实现基于一种基于能量的生成模型。首先，蛋白质设计者指定一个由一组层次化组织的约束组成的高级程序(图1A)。然后，该程序编译为一个评估与约束兼容性的能量函数，这些约束可以是任意的且不可微的(图1B)。我们通过将原子级结构预测(由语言模型实现)纳入能量函数，对结构施加约束。这种方法能够生成一系列复杂的设计(图1C)。

The use of a high-level language allows the protein designer

使用高级语言使蛋白质设计者

[[1]](#footnote-26)

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.21.521526; this version posted December 22, 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-NC-ND 4.0 International license.

bioRxiv预印本doi: https://doi.org/10.1101/2022.12.21.521526; 此版本发布于2022年12月22日。本预印本的版权持有者(未经同行评审认证)为作者/资助者，其已授予bioRxiv永久展示该预印本的许可。本作品依据CC-BY-NC-ND 4.0国际许可协议提供。

A high-level programming language for generative protein design to systematically reason about the design space and specify very general, modular, and composable programs. To demonstrate this, we generate proteins that realize a variety of constraints that include secondary structure, symmetry, multimerization, and atomic-level coordination in the predicted structures.We apply these constraints in complex, hierarchical settings, where we can enumerate a space of highly idealized forms that have low similarity to natural structures. As de novo design progresses to more complex proteins and protein assemblies, high-level abstractions such as the programming language described in this study should facilitate the systematic exploration and design of complex artificial proteins.

一种用于生成蛋白质设计的高级编程语言，能够系统地推理设计空间并指定非常通用、模块化和可组合的程序。为了证明这一点，我们生成了实现各种约束的蛋白质，这些约束包括二级结构、对称性、多聚化以及预测结构中的原子级配位。我们在复杂的分层设置中应用这些约束，从而枚举出一系列与天然结构相似度极低的高度理想化形式。随着从头设计逐渐发展到更复杂的蛋白质和蛋白质组装体，如本研究中描述的编程语言这样的高级抽象应有助于系统地探索和设计复杂的人工蛋白质。

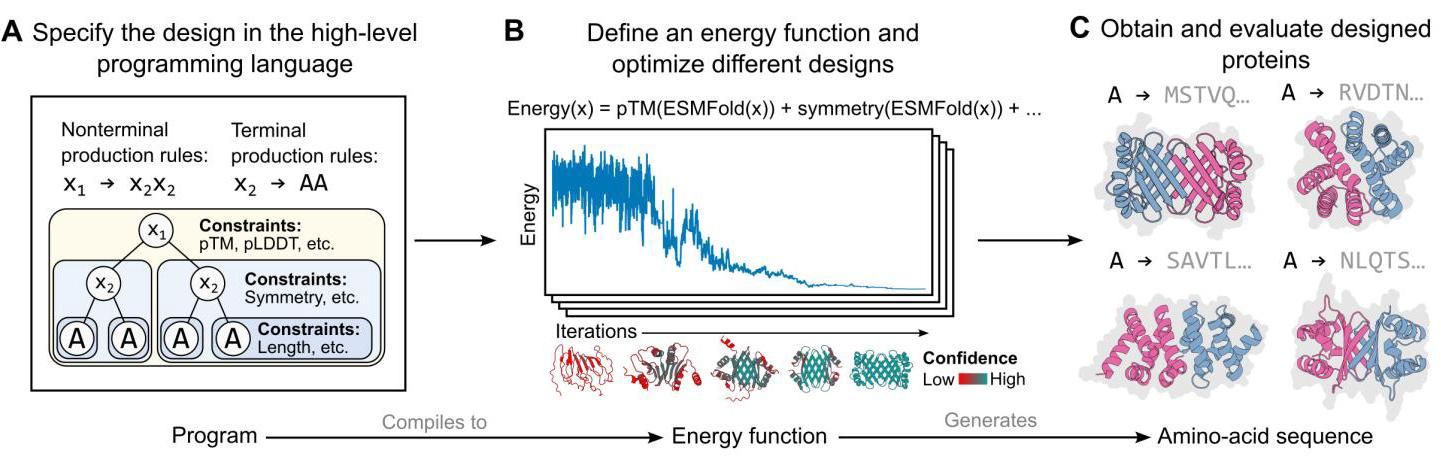


Figure 1. Overview of the high-level programming language and the optimization algorithm. (A) We propose a high-level programming language in which each program consists of (1) a syntax tree (corresponding to a set of nonterminal and terminal production rules) that enables modular and hierarchical organization of protein subunits and (2) a set of constraint functions that can be defined at each node of the syntax tree, where a given constraint is applied to the entire subtree rooted at the corresponding node. (B) This program is then compiled to a single energy function, which in our study is a simple linear combination of the specified constraint functions. The energy function is used to guide an optimization procedure based on simulated annealing, of which a key component is the use of an accurate and efficient structure predictor to evaluate the energy function at each step of the optimization. The same energy function can guide multiple optimization trajectories. (C) Each of these trajectories produces a protein sequence design and an associated predicted structure. These sequences and predicted structures can then be evaluated downstream using in silico and experimental metrics.

图1. 高级编程语言和优化算法的概述。(A) 我们提出了一种高级编程语言，其中每个程序由(1)一个语法树(对应于一组非终结符和终结符产生规则)组成，该语法树支持蛋白质亚基的模块化和分层组织，以及(2)一组可以在语法树的每个节点定义的约束函数，其中给定的约束应用于以相应节点为根的整个子树。(B) 然后，该程序被编译为单个能量函数，在我们的研究中，该能量函数是指定约束函数的简单线性组合。能量函数用于指导基于模拟退火的优化过程，其中一个关键组成部分是使用准确高效的结构预测器来评估优化过程中每一步的能量函数。相同的能量函数可以指导多个优化轨迹。(C) 每个轨迹都会生成一个蛋白质序列设计和相关的预测结构。这些序列和预测结构随后可以通过计算机模拟和实验指标进行下游评估。

# A generative programming language for protein design

# 用于蛋白质设计的生成编程语言

We introduce a high-level programming language for generative protein design. This language first requires a syntax tree (Figure 1A) consisting of terminal symbols (i.e., the leaves of the tree) that each corresponds to a unique protein sequence (which is potentially repeated within the protein) and nonterminal symbols (i.e., the internal nodes of the tree) that enable hierarchical organization. Second, the language requires a set of constraints: at each node in the tree, a protein designer can specify any number of constraints, which are applied to the entire subtree. The syntax tree and its constraints fully specify a program in our high-level language. We provide a more extended description of this language in the Methods section.

我们介绍了一种用于生成蛋白质设计的高级编程语言。该语言首先需要一个语法树(图1A)，该语法树由终结符(即树的叶子)组成，每个终结符对应于一个唯一的蛋白质序列(可能在蛋白质中重复)，以及非终结符(即树的内部节点)，这些非终结符支持分层组织。其次，该语言需要一组约束:在树的每个节点，蛋白质设计者可以指定任意数量的约束，这些约束应用于整个子树。语法树及其约束完全指定了我们高级语言中的一个程序。我们在方法部分提供了对该语言的更详细描述。

Each program is compiled into an energy function that specifies a generative model for that program in the form of a distribution

每个程序都被编译成一个能量函数，该能量函数以分布的形式指定了该程序的生成模型

over protein sequences conforming to the program. The constraints are encoded as weighted terms that are additively combined into the total energy. Since the partition function , which is a function of the parameters , is intractable, low temperature samples can be taken with MCMC and simulated annealing (Figure 1B). The generative capacity of this approach is built on recent developments in deep learning for protein biology. Specifically, each step in the optimization loop has access to a fast and accurate atomic-level structure prediction enabled by the ESM-2 protein language model.

符合该程序的蛋白质序列 。约束被编码为加权项，这些项被加性组合到总能量中。由于配分函数 (它是参数 的函数)难以处理，因此可以使用MCMC和模拟退火(图1B)获取低温样本。该方法的生成能力建立在蛋白质生物学深度学习的最新进展之上。具体而言，优化循环中的每一步都可以通过ESM-2蛋白质语言模型实现快速准确的原子级结构预测。

Given a single program, the generative model can create potentially diverse designs that fulfill the user-specified constraints (Figure 1C). These constraints can be arbitary and nondifferentiable, and can span multiple scales of biological complexity, from atomic-level coordinates to abstract plans of the protein including the overall topology and symmetry. This approach allows the model to propose diverse solutions where many potential designs may satisfy the program. By leveraging an expressive model of structure in a generative capacity, the resulting designs respect the various constraints individually and are also globally coherent.

给定一个程序，生成模型可以创建可能满足用户指定约束的多样化设计(图1C)。这些约束可以是任意且不可微的，并且可以跨越生物复杂性的多个尺度，从原子级坐标到包括整体拓扑和对称性在内的蛋白质抽象计划。这种方法允许模型提出多样化的解决方案，其中许多潜在设计可能满足该程序。通过在生成能力中利用表达性结构模型，最终的设计不仅单独尊重各种约束，而且在全局上也是连贯的。

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.21.521526; this version posted December 22, 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-NC-ND 4.0 International license.

bioRxiv预印本 doi: https://doi.org/10.1101/2022.12.21.521526; 此版本发布于2022年12月22日。该预印本的版权持有者(未经同行评审认证)是作者/资助者，他们已授予bioRxiv永久展示该预印本的许可。该预印本在CC-BY-NC-ND 4.0国际许可下提供。

A high-level programming language for generative protein design

一种用于生成蛋白质设计的高级编程语言

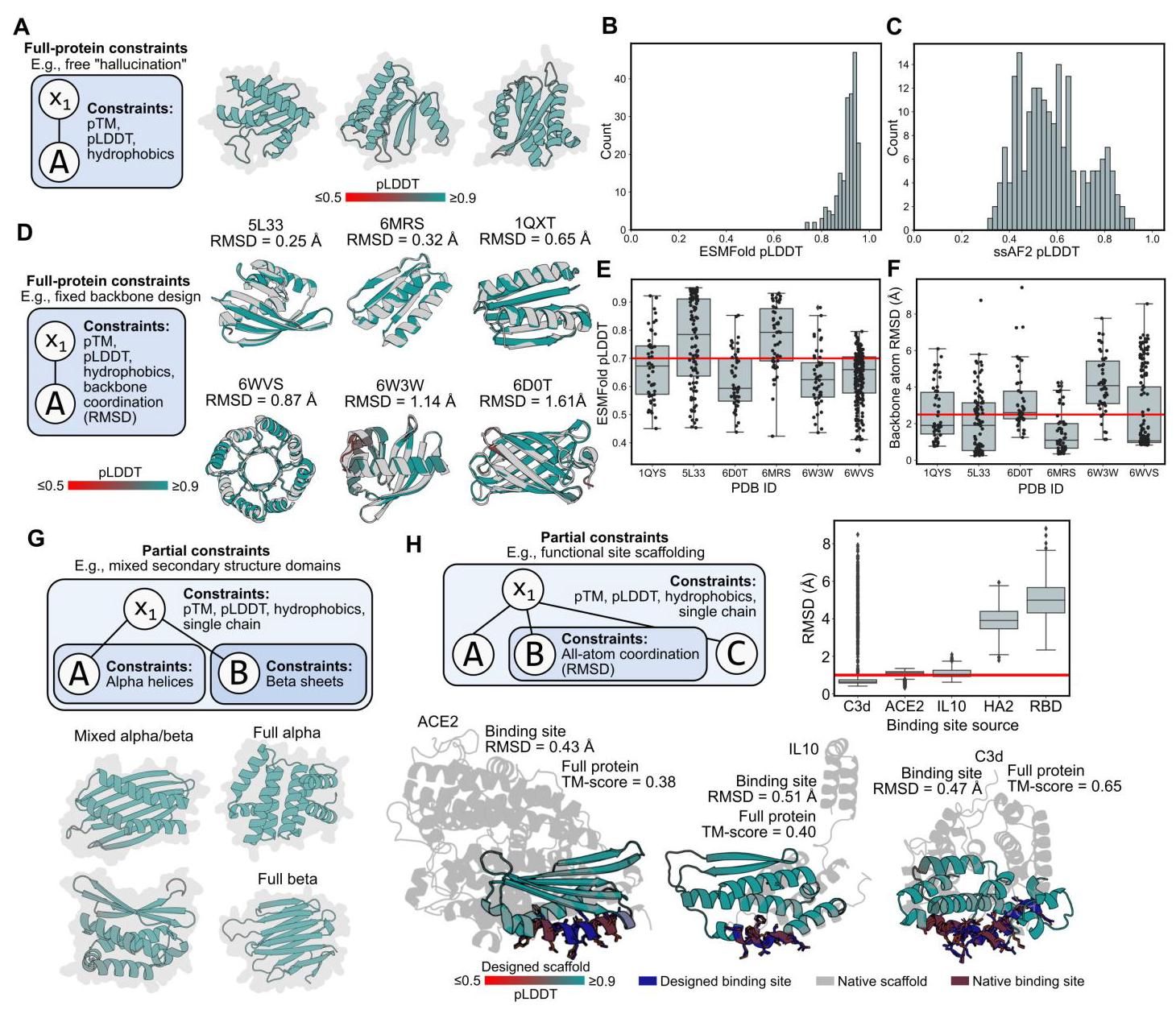


Figure 2. Programming full-protein or partial constraints. (A) A graphical representation of a program for protein "free hallucination" (left) along with three example designed structures (right). (B) The distribution of ESMFold pLDDT values over 200 free-hallucinated structures. Of these, 100% have good confidence (ESMFold pLDDT ). (C) The distribution of single-sequence AlphaFold2 (ssAF2) over the same 200 structures; note that ssAF2 was not used in the design procedure. Of these, 22% have good confidence (ssAF2 pLDDT . (D) A graphical representation of a program for fixed backbone design (left) along with example designs for six de novo target backbones. The experimental backbone is colored gray; the designed backbone is colored by ESMFold pLDDT. (E) For each target backbone, the distribution of the ESMFold pLDDT values of the final designs from 50 or more fixed backbone design seeds is plotted as a boxplot (for all boxplots in this figure, the box extends from first to third quartile, black line indicates the median, and whiskers indicate 2.5 times the interquartile range) with each seed also plotted as a black circle. A horizontal red line indicates pLDDT = 0.7. (F) For each target backbone, the distribution of RMSD values between the target and design backbone atoms from 50 or more fixed backbone design seeds is plotted as a boxplot with each seed also plotted as a black circle. A horizontal red line indicates RMSD = 2.5 Å. (G) A graphical representation of a program for designing a protein with mixed secondary structure (top) along with example designs in which secondary structure was explicitly specified (bottom). (H) Top-left: A graphical representation of a program for functional site scaffolding. Top-right: For each scaffolded binding site, the distribution of RMSD between the native and designed binding site atoms (including side chains) from 2,000 seeds is plotted as a boxplot. A horizontal red line indicates RMSD = 2 Å. Bottom: Example designs that achieve sub-angstrom atomic coordination in the scaffolded binding site atoms, high model confidence in the associated scaffold, and low similarity (quantified by TM-score) to the natural protein.

图2. 编程全蛋白或部分约束。(A) 蛋白质“自由幻觉”程序的图形表示(左)以及三个示例设计结构(右)。(B) 200个自由幻觉结构的ESMFold pLDDT值分布。其中，100%具有高置信度(ESMFold pLDDT )。(C) 单序列AlphaFold2 (ssAF2) 在相同200个结构上的分布；注意ssAF2未用于设计过程。其中，22%具有高置信度(ssAF2 pLDDT )。(D) 固定骨架设计程序的图形表示(左)以及六个从头目标骨架的示例设计。实验骨架为灰色；设计骨架按ESMFold pLDDT着色。(E) 对于每个目标骨架，从50个或更多固定骨架设计种子中最终设计的ESMFold pLDDT值分布以箱线图形式绘制(本图中所有箱线图，箱体从第一四分位数延伸到第三四分位数，黑线表示中位数，须线表示2.5倍四分位距)，每个种子也以黑圈表示。红色水平线表示pLDDT = 0.7。(F) 对于每个目标骨架，从50个或更多固定骨架设计种子中目标与设计骨架原子之间的RMSD值分布以箱线图形式绘制，每个种子也以黑圈表示。红色水平线表示RMSD = 2.5 Å。(G) 设计具有混合二级结构蛋白质的程序的图形表示(上)以及明确指定二级结构的示例设计(下)。(H) 左上:功能位点支架程序的图形表示。右上:对于每个支架结合位点，从2000个种子中天然与设计结合位点原子(包括侧链)之间的RMSD分布以箱线图形式绘制。红色水平线表示RMSD = 2 Å。下:在支架结合位点原子中实现亚埃级原子配位的示例设计，相关支架具有高模型置信度，且与天然蛋白质的相似性低(通过TM-score量化)。

A high-level programming language for generative protein design

用于生成蛋白质设计的高级编程语言

# Results

# 结果

# Full-protein constraints

# 全蛋白约束

We first demonstrate that our approach can design proteins where the constraints are simply applied to the entire sequence and structure without any hierarchical organization. An especially valuable constraint, which we apply generally across all our design efforts, steers the optimization toward predicted structures with higher model confidence, i.e., high pTM and mean pLDDT. For proteins that we desire to have soluble, monomeric expression, we also steer the optimization to minimize hydrophobic residues that are solvent-exposed (Methods). Using only these constraints on structural confidence and hydrophobic residue placement, our model is able to generate or "freely hallucinate" (17) high-confidence structures (Figures 2A and 2B); across 200 seeds, all optimization loops produced predicted structures with an ESMFold mean pLDDT greater than 0.7 (Figure 2B). Of these, a large portion (44, or 22%) also had high predicted confidence (pLDDT ) by single-sequence AlphaFold2 (18) (Figure 2C), a separate structure prediction model that was not used in our optimization procedure.

我们首先证明我们的方法可以设计蛋白质，其中约束简单地应用于整个序列和结构，而无需任何层次组织。一个特别有价值的约束，我们普遍应用于所有设计工作中，引导优化朝向具有更高模型置信度的预测结构，即高pTM和平均pLDDT。对于我们希望具有可溶性、单体表达的蛋白质，我们还引导优化以最小化溶剂暴露的疏水残基(方法)。仅使用这些结构置信度和疏水残基放置的约束，我们的模型能够生成或“自由幻觉”高置信度结构(图2A和2B)；在200个种子中，所有优化循环产生的预测结构的ESMFold平均pLDDT均大于0.7(图2B)。其中，很大一部分(44个，或22%)也通过单序列AlphaFold2 (18) 具有高预测置信度(pLDDT )(图2C)，这是一个未用于我们优化过程的独立结构预测模型。

Our objective function also enables other full-protein constraints, such as specifying the positions of the backbone atoms while allowing the algorithm to design the corresponding sequence, a design task referred to as fixed backbone design (19). To achieve this, we can add a term to the energy function that minimizes the root-mean-square deviation (RMSD) between the corresponding designed and target backbone atoms (Methods). Our simulated annealing procedure successfully produces high-confidence designs with low RMSD ( ) across diverse de novo backbones (Figures 2D and 2E), and can do so reproducibly over different optimization runs (Figure 2F).

我们的目标函数还支持其他全蛋白约束，例如指定骨架原子的位置，同时允许算法设计相应的序列，这种设计任务称为固定骨架设计 (19)。为了实现这一点，我们可以在能量函数中添加一个项，以最小化相应设计与目标骨架原子之间的均方根偏差 (RMSD)(方法)。我们的模拟退火程序成功地在各种从头骨架上生成了具有低RMSD ( ) 的高置信度设计(图2D和2E)，并且可以在不同的优化运行中重复实现(图2F)。

# Partial constraints

# 部分约束

We next sought to increase the complexity of our designable space by varying the constraints enforced on different parts of a protein. For example, a simple mixed-constraint setting is to specify a two-domain protein with different combinations of secondary structure composition (Figures 2G and S1A-S1C). In our programs, we can represent this setting by a syntax tree containing two or more subtrees, where different constraints are only applied within the discrete subtrees.

我们接下来通过改变对蛋白质不同部分施加的约束来增加可设计空间的复杂性。例如，一个简单的混合约束设置是指定具有不同二级结构组合的双域蛋白质(图2G和S1A-S1C)。在我们的程序中，我们可以通过包含两个或更多子树的语法树来表示此设置，其中不同的约束仅应用于离散子树内。

A more complex mixed-constraint setting is to design functional proteins by constraining one region of the protein design to have the same all-atom positions (including protein side chains) as a functional site from nature, while allowing the design procedure to freely generate the remainder of the protein; this design setting is sometimes referred to as functional site "scaffolding" (20). Importantly, in contrast to fixed backbone design, in which constraints are only placed on backbone atomic coordinates, functional site scaffolding requires constraints on side-chain atoms as well, since these are critical to achieving function. Because our optimization procedure produces an all-atom structure prediction at each step of the optimization, we can readily incorporate this constraint as part of the energy function by minimizing the all-atom RMSD between the natural and designed atomic coordinates of the functional site (Methods).

一个更复杂的混合约束设置是通过限制蛋白质设计的一个区域具有与自然界功能位点相同的全原子位置(包括蛋白质侧链)来设计功能蛋白质，同时允许设计过程自由生成蛋白质的其余部分；这种设计设置有时被称为功能位点“支架”(20)。重要的是，与固定骨架设计(其中约束仅施加在骨架原子坐标上)相比，功能位点支架还需要对侧链原子施加约束，因为这些对实现功能至关重要。由于我们的优化过程在每一步都生成全原子结构预测，我们可以通过最小化功能位点的天然和设计原子坐标之间的全原子RMSD(方法)来轻松地将此约束纳入能量函数中。

Across functional sites involving sequence-contiguous or -discontiguous residues from a variety of natural proteins, our algorithm is able to produce designs that scaffold the site with sub-angstrom RMSD between the experimental and predicted structure in three out of five functional sites attempted (Figures and ). Moreover, the algorithm produces designed scaffolds that depart from the native protein (Figure ). The ability to move natural functional sites onto designed backbones has many practical applications, including the design of functional proteins that are smaller or stabler than their natural counterparts.

在涉及来自各种天然蛋白质的序列连续或不连续残基的功能位点中，我们的算法能够在五个尝试的功能位点中的三个中生成支架，实验和预测结构之间的RMSD低于埃(图 和 )。此外，该算法生成的支架设计偏离了天然蛋白质(图 )。将天然功能位点移动到设计骨架上的能力具有许多实际应用，包括设计比天然对应物更小或更稳定的功能蛋白质。

# Symmetric and multimeric group constraints

# 对称和多聚体群约束

Beyond proteins containing partial constraints, we next increase the complexity of our protein designs by generating structures that contain constraints over multiple subunits. A foundational design task for the generation of idealized, de novo proteins is to constrain structural symmetry (7, 22). To generate symmetric proteins, we first enforce the notion of a repeated unit that is repeated times when designing a -fold symmetry (where we can control the value of ). To guide the optimization toward symmetric structures, we add various constraints on the distances among the centroids of each repeated unit as part of the energy function (Methods). In our high-level language, a symmetric protein would be encoded by repeating the same non-terminal symbol times (corresponding to the repeated unit); the symmetry constraint is then placed at the level of the syntax tree containing these repeated non-terminals (Figure 3A).

除了包含部分约束的蛋白质外，我们接下来通过生成包含多个亚基约束的结构来增加蛋白质设计的复杂性。生成理想化的从头蛋白质的一个基本设计任务是约束结构对称性(7, 22)。为了生成对称蛋白质，我们首先在设计 重对称时强制执行重复单元的概念，该单元重复 次(我们可以控制 的值)。为了引导优化朝向对称结构，我们在每个重复单元的质心之间的距离上添加各种约束作为能量函数的一部分(方法)。在我们的高级语言中，对称蛋白质将通过重复相同的非终结符号 次(对应于重复单元)进行编码；然后对称约束被放置在包含这些重复非终结符的语法树级别(图3A)。

Using these symmetric constraints, we show that we can program the level of symmetry within a protein design. When directed to design 3- to 8-fold symmetry, the generative model produces a diverse set of high-confidence structures (Figures 3B, S2A, and S2B), including folds that have common analogs in nature (including coiled-coils, beta propellers, beta sandwiches, beta barrels, and TIM barrels) as well as highly-idealized designs that are different from natural structures, including a pentagonal star-shaped protein (with a TM-score of 0.48 to the nearest PDB structure 3S38; row 1 and column 3 in Figure 3B) and a cube-shape protein (nearest-PDB TM-score of 0.51 to PDB 7DEG; row 2 and column 2 in Figure 3B). The highlighted symmetric proteins in Figure 3B have nearest-PDB TM-scores ranging from bioRxiv preprint doi: https://doi.org/10.1101/2022.12.21.521526; this version posted December 22, 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-NC-ND 4.0 International license.

使用这些对称约束，我们展示了我们可以在蛋白质设计中编程对称级别。当指导设计3到8重对称时，生成模型生成了一组多样化的高置信度结构(图3B、S2A和S2B)，包括自然界中常见的类似折叠(包括卷曲螺旋、β螺旋桨、β三明治、β桶和TIM桶)以及高度理想化的设计，这些设计与天然结构不同，包括五角星形蛋白质(与最近的PDB结构3S38的TM-score为0.48；图3B第1行第3列)和立方体形状的蛋白质(与PDB 7DEG的最近PDB TM-score为0.51；图3B第2行第2列)。图3B中突出显示的对称蛋白质的最近PDB TM-score范围从bioRxiv预印本doi: https://doi.org/10.1101/2022.12.21.521526；此版本发布于2022年12月22日。本预印本的版权持有者(未经同行评审认证)是作者/资助者，他们已授予bioRxiv永久展示预印本的许可。本预印本在aCC-BY-NC-ND 4.0国际许可下提供。

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.21.521526; this version posted December 22, 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-NC-ND 4.0 International license.

bioRxiv预印本doi: https://doi.org/10.1101/2022.12.21.521526；此版本发布于2022年12月22日。本预印本的版权持有者(未经同行评审认证)是作者/资助者，他们已授予bioRxiv永久展示预印本的许可。本预印本在aCC-BY-NC-ND 4.0国际许可下提供。

A high-level programming language for generative protein design

用于生成蛋白质设计的高级编程语言

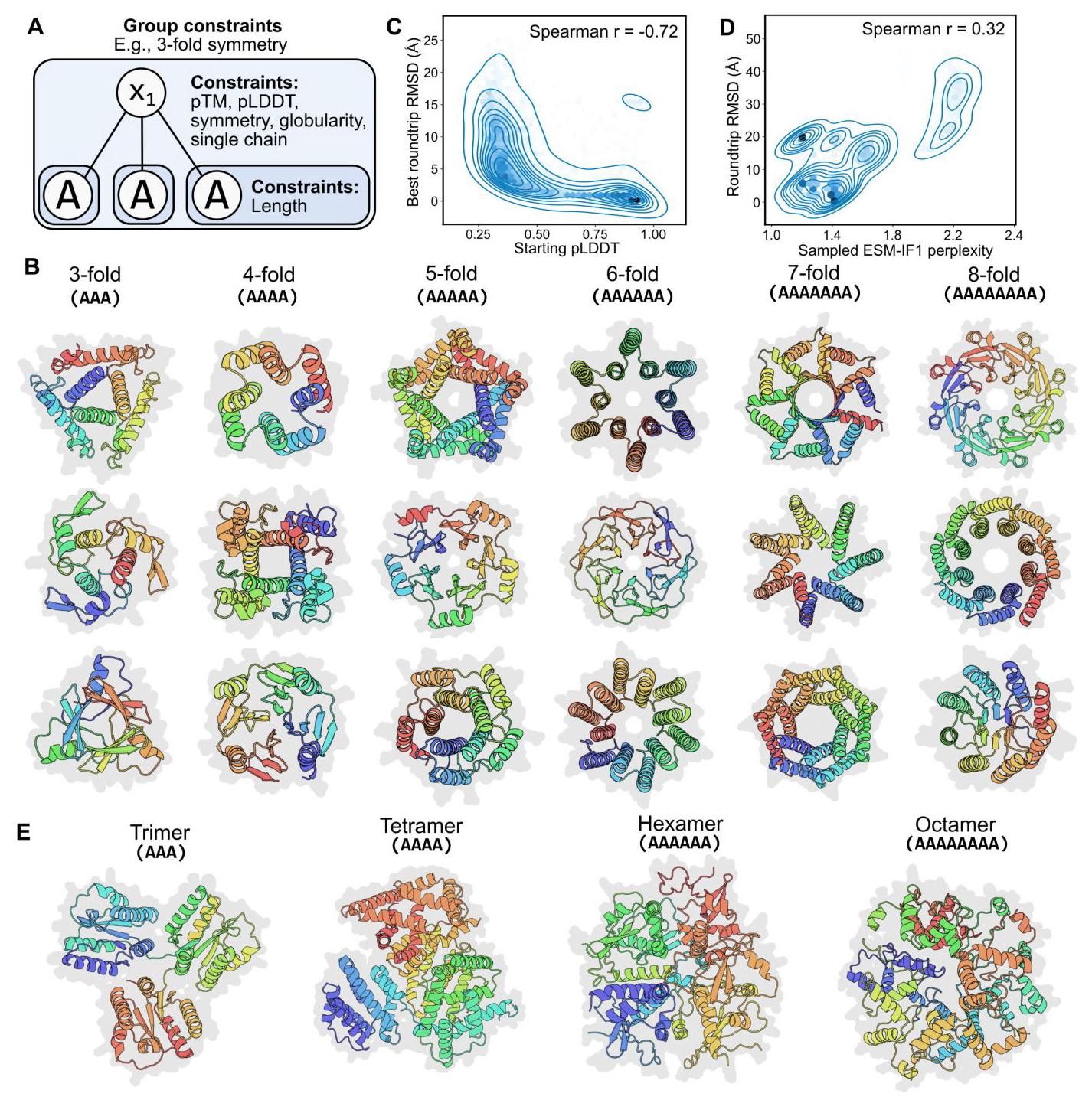


Figure 3. Programming symmetry and homo-oligomerization. (A) A graphical representation of a program for designing a single protein chain with 3-fold symmetry based on a repeated subsequence. (B) Example designs varying fold symmetry from 3- to 8-fold. (C) 1000 randomly sampled symmetric protein designs were "roundtripped" by sampling ten sequences via ESM-IF1 inverse folding (21) of their backbones followed by ESMFold structure prediction. The ESMFold pLDDT of the starting backbone is indicated on the horizontal axis. The lowest of the 10 RMSDs comparing the starting and roundtripped backbone atoms is indicated on the vertical axis. Blue lines indicate density contours and hexagonal bins are darker with greater density. We observed that a more confident design is associated with roundtrip success. (D) 1,000 randomly sampled inverse folding samples are plotted according to their ESM-IF1 perplexity on the horizontal axis and their roundtrip RMSD on the vertical axis. We observed that a lower perplexity sequence is associated with roundtrip success. (E) Example homo-oligomers with increasing numbers of individual protomers. The tetrameric, hexameric, and octameric oligomers depicted here form globular polyhedral shapes rather than the rotational symmetry of designs in (B).

图3. 编程对称性和同源寡聚化。(A) 基于重复子序列设计具有三重对称性的单一蛋白质链的程序的图形表示。(B) 从三重到八重对称性的示例设计。(C) 通过ESM-IF1逆折叠(21)对其骨架进行十次序列采样，然后进行ESMFold结构预测，对1000个随机采样的对称蛋白质设计进行了“往返”测试。起始骨架的ESMFold pLDDT在横轴上表示。比较起始和往返骨架原子的10个RMSD中的最小值在纵轴上表示。蓝线表示密度等高线，六边形箱体颜色越深表示密度越大。我们观察到，更自信的设计与往返成功相关。(D) 1000个随机采样的逆折叠样本根据其ESM-IF1困惑度在横轴上和其往返RMSD在纵轴上绘制。我们观察到，较低的困惑度序列与往返成功相关。(E) 具有越来越多单个原聚体的同源寡聚体示例。这里描绘的四聚体、六聚体和八聚体寡聚体形成球状多面体形状，而不是(B)中设计的旋转对称性。

A high-level programming language for generative protein design A high-level programming language for generative protein design 0.47 to 0.86, with a median TM-score of 0.64; TM-scores for all seeds are plotted in Figure S2C.

用于生成蛋白质设计的高级编程语言 用于生成蛋白质设计的高级编程语言 0.47到0.86，中位数TM得分为0.64；所有种子的TM得分在图S2C中绘制。

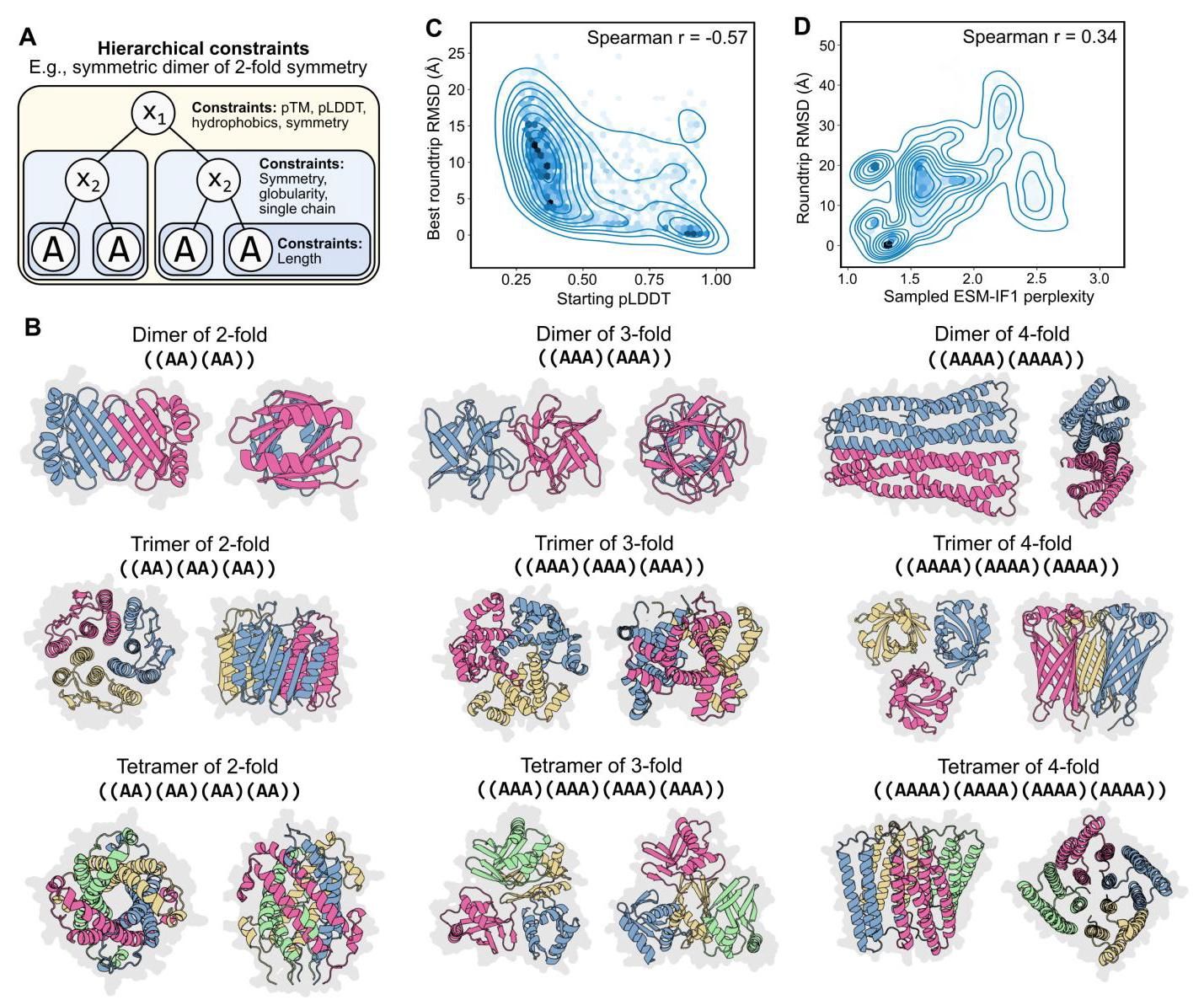


Figure 4. Programming two levels of symmetry. (A) A graphical representation of a program for designing two levels of symmetry in which a homo-oligomeric symmetric dimer represents the top level of symmetry and each unit within the dimer also has two-fold symmetry. (B) Example oligomers with two levels of symmetry, in which we procedurally enumerate across a grid in which we vary the top-level symmetry across the rows and the bottom-level symmetry across the columns. Discrete chains are indicated by different colors. (C) 1,000 randomly sampled two-level symmetric protein oligomer designs were "roundtripped" by sampling ten sequences via ESM-IF1 inverse folding (21) of their backbones followed by ESMFold structure prediction (Methods). The ESMFold pLDDT of the starting backbone is indicated on the horizontal axis. The lowest of the 10 RMSDs comparing the starting and roundtripped backbone atoms is indicated on the vertical axis. Blue lines indicate density contours and hexagonal bins are darker with greater density. We observed that a more confident design is associated with roundtrip success. (D) 1,000 randomly sampled inverse folding samples are plotted according to their ESM-IF1 perplexity on the horizontal axis and their roundtrip RMSD on the vertical axis. We observed that a lower perplexity sequence is associated with roundtrip success.

图4. 编程两级对称性。(A) 设计两级对称性的程序的图形表示，其中同源寡聚对称二聚体代表顶级对称性，二聚体中的每个单元也具有二重对称性。(B) 具有两级对称性的示例寡聚体，其中我们在网格上程序化枚举，在行中变化顶级对称性，在列中变化底层对称性。离散链用不同颜色表示。(C) 通过ESM-IF1逆折叠(21)对其骨架进行十次序列采样，然后进行ESMFold结构预测(方法)，对1000个随机采样的两级对称蛋白质寡聚体设计进行了“往返”测试。起始骨架的ESMFold pLDDT在横轴上表示。比较起始和往返骨架原子的10个RMSD中的最小值在纵轴上表示。蓝线表示密度等高线，六边形箱体颜色越深表示密度越大。我们观察到，更自信的设计与往返成功相关。(D) 1000个随机采样的逆折叠样本根据其ESM-IF1困惑度在横轴上和其往返RMSD在纵轴上绘制。我们观察到，较低的困惑度序列与往返成功相关。

To increase our confidence that these idealized structures correspond to valid and designable backbones, we observe that sampling sequences via inverse folding with the ESM-IF1 and ProteinMPNN models (21, 23) followed by structure prediction of the sequence samples can reproducibly recover the original backbone geometry (Methods), in many instances with sub-angstrom backbone-atom RMSD. To increase our confidence that a successful "roundtrip" through inverse folding indicates designable backbones, we observe that high confidence designs indicates better roundtrip success (Figures 3C and S2D) and that the ability to sample a low perplexity sequence also indicates roundtrip success (Figure 3D).

为了增加我们对这些理想化结构对应于有效且可设计的骨架的信心，我们观察到通过ESM-IF1和ProteinMPNN模型(21, 23)进行逆折叠采样序列，然后对序列样本进行结构预测，可以可重复地恢复原始骨架几何结构(方法)，在许多情况下具有亚埃级骨架原子RMSD。为了增加我们对通过逆折叠成功“往返”表明可设计骨架的信心，我们观察到高置信度设计表明更好的往返成功(图3C和S2D)，并且能够采样低困惑度序列也表明往返成功(图3D)。

Beyond single-chain symmetries, we can also design multi-meric proteins similarly. We enforce the notion of single or multiple chains in our programming language with a "single chain" constraint that dictates that all terminal elements in a subtree belong to the same chain (Methods); to design multimeric proteins, we need only remove this constraint. Example multimeric symmetric proteins involving 4- to 8-mers are provided in Figure 3E.

除了单链对称性，我们还可以类似地设计多聚体蛋白质。我们在编程语言中通过“单链”约束强制执行单链或多链的概念，该约束规定子树中的所有终端元素属于同一链(方法)；要设计多聚体蛋白质，我们只需移除此约束。图3E中提供了涉及4到8聚体的多聚体对称蛋白质示例。

# Hierarchical constraints

# 层次约束

Formalizing our constraints into a syntax tree naturally enables the specification of hierarchical constraints, which enables more complex protein designs. As an initial demonstration, guided by our high-level language’s formalization, we design different levels of symmetry at two levels of hierarchy, where the lower level of symmetry is specified within a chain and the upper level of symmetry is specified among protomers in a homo-oligomer (Figure 4A). We procedurally enumerate over examples that range from a dimer of units with 2 -fold symmetry to a tetramer of units with 4-fold symmetry (Figures 4B, S3A, and S3B). As in the single-chain symmetric design setting, we observe successful structure prediction roundtrips through inverse folding, and that both high-confidence predicted structures and low inverse folding perplexity indicate roundtrip success (Figures and ). Many of the designs with two levels of symmetry have low overall similarity to structures in the PDB; for example, a dimer of 2-fold symmetry in which opposing beta sheets form a regular checkerboard pattern (nearest-PDB TM-score of 0.49 to PDB 3W38; row 1 and column 1 in Figure 4B). The highlighted homo-oligomers of two-level symmetry in Figure 4B have nearest-PDB TM-scores ranging from 0.25 to 0.52, with a median TM-score of 0.48 ; TM-scores for all seeds are plotted in Figure S3C.

将我们的约束形式化为语法树自然能够实现层次化约束的规范，从而支持更复杂的蛋白质设计。作为初步演示，在我们的高级语言形式化的指导下，我们在两个层次上设计了不同级别的对称性，其中较低级别的对称性在链内指定，而较高级别的对称性在同源寡聚体中的原聚体之间指定(图4A)。我们通过程序化枚举从具有2重对称性的单元二聚体到具有4重对称性的单元四聚体的示例(图4B、S3A和S3B)。与单链对称设计设置一样，我们观察到通过逆折叠成功进行结构预测的往返过程，并且高置信度的预测结构和低逆折叠困惑度都表明往返成功(图 和 )。许多具有两级对称性的设计与PDB中的结构整体相似性较低；例如，一个具有2重对称性的二聚体，其中相对的β片形成规则的棋盘图案(与PDB 3W38的最近PDB TM得分为0.49；图4B中的第1行和第1列)。图4B中突出显示的两级对称性同源寡聚体的最近PDB TM得分范围从0.25到0.52，中位TM得分为0.48；所有种子的TM得分在图S3C中绘制。

Another hierarchical design setting is to combine the function-scaffolding and the symmetric design tasks described above, as some functions are enhanced by repetition of a functional site; for example, when improving the strength of a binding interaction, multiple binding sites on a protein could synergize such that the overall binding avidity is greater than the sum of the individual affinities (24). This task requires two levels of hierarchy: the top level specifies symmetry while the bottom level specifies the side-chain atomic coordination constraint (Figure 5A). With this corresponding program, we can generate designs in which an atomic-level constraint is enforced on multiple functional sites over the protein, the overall protein organization is constrained to be symmetric, and we can control the level of designed symmetry (Figures 5B and S4A-S4C).

另一个层次化设计设置是将上述功能支架和对称设计任务结合起来，因为某些功能通过功能位点的重复而增强；例如，当提高结合相互作用的强度时，蛋白质上的多个结合位点可以协同作用，使得整体结合亲和力大于单个亲和力的总和(24)。此任务需要两个层次:顶层指定对称性，而底层指定侧链原子配位约束(图5A)。通过相应的程序，我们可以生成在蛋白质的多个功能位点上强制执行原子级约束的设计，整体蛋白质组织被约束为对称，并且我们可以控制设计的对称性级别(图5B和S4A-S4C)。

We lastly show that we can specify protein designs that have even deeper levels of hierarchy in their constraints (Figure 5C) by designing protein assemblies that combine both symmetry and asymmetry. For example, we designed a protein complex composed of four units in which a pair of the chains are symmetric to each other (and each unit internally has two-fold symmetry) and where another pair of chains are symmetric to each other (and each unit also internally has two-fold symmetry), but the two pairs are asymmetric to each other (Figures 5C, 5D, S4D, and S4E). Our high-level programming language readily enables us to control the complexity of the generated complexes such that, for example, one of the pairs consists of chains with threefold symmetry (Figure 5E) or that the complex consists of five chains (a pair of symmetric chains of two-fold symmetry asymmetrically complexed with a triple of symmetric chains of two-fold symmetry) (Figure 5F). We find that our optimization procedure can produce designed structures consistent with all of these hierarchical specifications.

最后，我们展示了我们可以通过设计结合对称性和非对称性的蛋白质组装体来指定具有更深层次约束的蛋白质设计(图5C)。例如，我们设计了一个由四个单元组成的蛋白质复合物，其中一对链彼此对称(并且每个单元内部具有二重对称性)，另一对链彼此对称(并且每个单元内部也具有二重对称性)，但这两对彼此不对称(图5C、5D、S4D和S4E)。我们的高级编程语言使我们能够轻松控制生成复合物的复杂性，例如，其中一对链具有三重对称性(图5E)或复合物由五条链组成(一对具有二重对称性的对称链与三条具有二重对称性的对称链不对称地复合)(图5F)。我们发现我们的优化程序可以生成与所有这些层次化规范一致的设计结构。

# Related work

# 相关工作

This paper is related to classical work that attempts to (i) classify a set of common sequence or structure motifs(3,4) and (ii) manually combine these motifs to generate new proteins (7-14). More recently, deep-learning-based methods have increased the complexity of designable structures (17,20,22)and machine-learning-based generative models have shown increasingly sophisticated design capabilities. These include sequence-based Potts models and autoregressive language models for designing sequences (25-27), Markov Chain Monte Carlo algorithms combined with structure prediction for jointly designing sequences and structures(17,20,22), inverse folding models that use structural backbone coordinates to design sequences(21,23), and concurrent work using diffusion models for designing protein backbones(28,29). A key contribution of this study is to combine the modularity aspired to by classical methods with the power of modern generative models, in particular improvements in the accuracy and efficiency of language-model-based protein structure prediction (16). bioRxiv preprint doi: https://doi.org/10.1101/2022.12.21.521526; this version posted December 22, 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-NC-ND 4.0 International license.

本文与经典工作相关，这些工作试图(i)分类一组常见的序列或结构基序(3,4)和(ii)手动组合这些基序以生成新蛋白质(7-14)。最近，基于深度学习的方法增加了可设计结构的复杂性(17,20,22)，基于机器学习的生成模型展示了越来越复杂的设计能力。这些包括基于序列的Potts模型和自回归语言模型用于设计序列(25-27)，结合结构预测的马尔可夫链蒙特卡罗算法用于联合设计序列和结构(17,20,22)，使用结构骨架坐标设计序列的逆折叠模型(21,23)，以及使用扩散模型设计蛋白质骨架的并行工作(28,29)。本研究的一个关键贡献是将经典方法所追求的模块化与现代生成模型的力量结合起来，特别是基于语言模型的蛋白质结构预测的准确性和效率的改进(16)。bioRxiv预印本doi: https://doi.org/10.1101/2022.12.21.521526；此版本发布于2022年12月22日。本预印本的版权持有者(未经同行评审认证)是作者/资助者，他们已授予bioRxiv永久展示预印本的许可。本预印本根据CC-BY-NC-ND 4.0国际许可提供。

A high-level programming language for generative protein design

用于生成蛋白质设计的高级编程语言

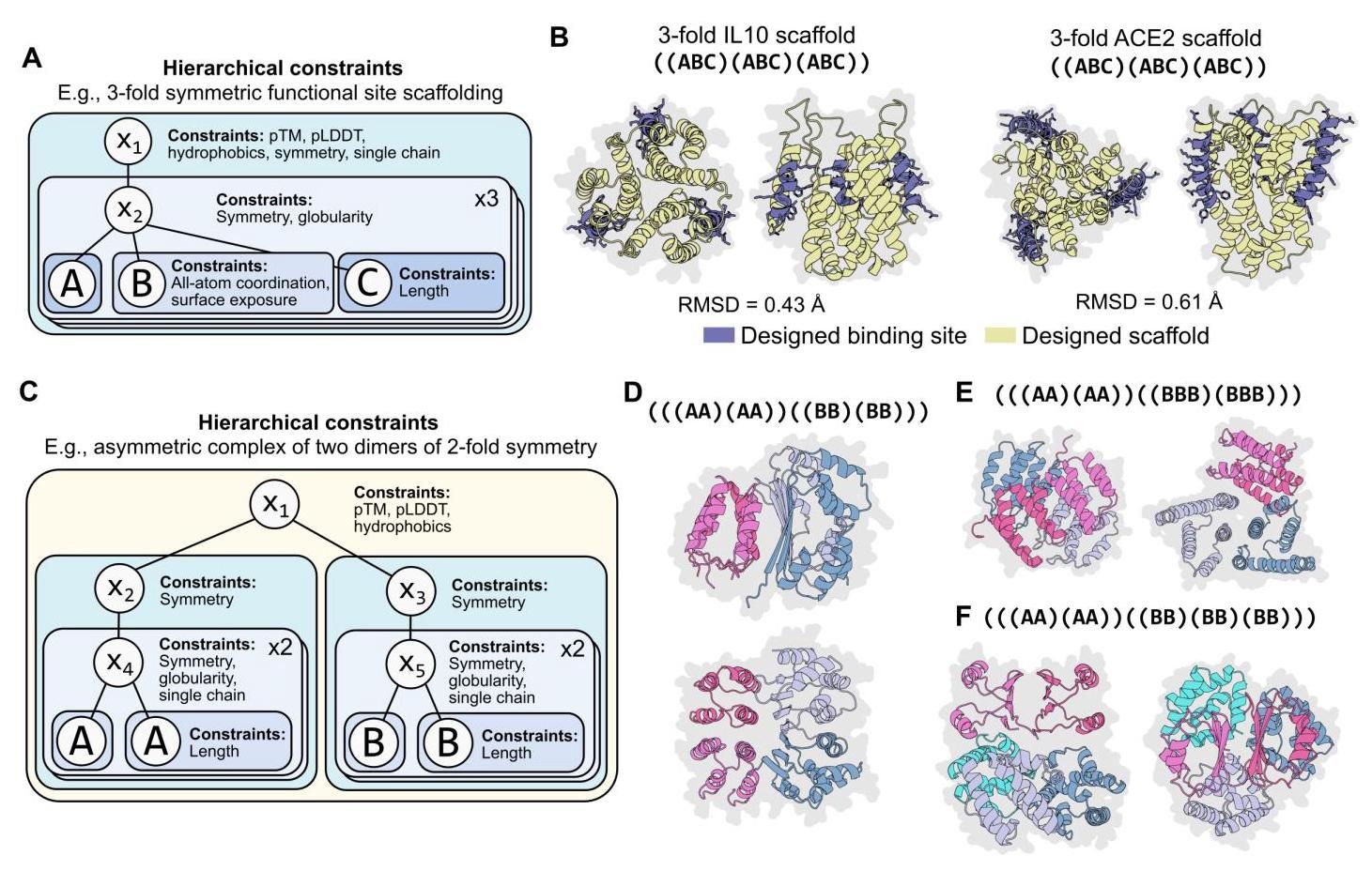


Figure 5. Programming complex hierarchical constraints. (A) A graphical representation of a program for scaffolding three functional sites in which those sites have a 3-fold symmetry. (B) Example 3-fold symmetric scaffolds for the IL10 and ACE2 binding sites that achieve sub-angstrom RMSD averaged across the three sites. (C) A graphical representation of a program that specifies an asymmetric protein complex consisting of two pairs of chains. Each pair is constrained to have 2-fold symmetry between the constituent chains. Furthermore, each constituent chain itself has 2-fold symmetry. (D) A generated protein structure as specified by the program depicted in(C). Discrete chains are indicated by different colors.(E)A generated protein structure as specified by the program depicted in(C) except where one of the pairs has constituent chains that have three-fold symmetry. Discrete chains are indicated by different colors. (F) A generated protein structure as specified by the program depicted in (C) except where one of the pairs is replaced with a symmetric trimer (where each constituent chain in the trimer has two-fold symmetry). Discrete chains are indicated by different colors.

图5. 编程复杂的层次约束。(A) 一个用于搭建三个功能位点的程序的图形表示，这些位点具有三重对称性。(B) 为IL10和ACE2结合位点设计的三重对称支架示例，三个位点的平均RMSD达到亚埃级。(C) 一个程序的图形表示，该程序指定了一个由两对链组成的不对称蛋白质复合物。每对链之间被约束为具有二重对称性。此外，每个组成链本身也具有二重对称性。(D) 由(C)中描述的程序生成的蛋白质结构。不同的链用不同的颜色表示。(E) 由(C)中描述的程序生成的蛋白质结构，但其中一对链的组成链具有三重对称性。不同的链用不同的颜色表示。(F) 由(C)中描述的程序生成的蛋白质结构，但其中一对链被替换为一个对称三聚体(三聚体中的每个组成链具有二重对称性)。不同的链用不同的颜色表示。

A high-level programming language for generative protein design

用于生成蛋白质设计的高级编程语言

# Discussion

# 讨论

In this study, we show that generative artificial intelligence enables high-level programmability at a new level of combinatorial complexity. We propose a programming language that can express high-level programs for the design of proteins at diverse biological scales, including atomic-level coordinates, secondary structure, and high-level symmetries within single chains and the units of self-assembling multi-chain complexes. We show that programs written in the abstract language can be compiled into an energy function and that the corresponding generative model is capable of fulfilling complex constraints within an overall coherent structure.

在本研究中，我们展示了生成式人工智能在组合复杂性新水平上实现了高级可编程性。我们提出了一种编程语言，能够表达用于设计蛋白质的高级程序，涵盖原子级坐标、二级结构以及单链和自组装多链复合体单元中的高级对称性。我们展示了用这种抽象语言编写的程序可以编译成能量函数，并且相应的生成模型能够在整体一致的结构中满足复杂的约束条件。

We demonstrate programs of increasing levels of complexity, including the design of homo-oligomers with two levels of symmetry, symmetric functional scaffolds, and asymmetric complexes of subunits that themselves have two levels of symmetry. The approach reveals a large space of idealized protein designs created from top-down design principles. Especially as the complexity of the constraints increases, many of the corresponding designs are highly idealized, analogous to the regularity of artificially created machines and systems.

我们展示了复杂度逐渐增加的程序，包括具有双重对称性的同源寡聚体设计、对称功能支架设计以及本身具有双重对称性的不对称亚基复合体设计。该方法揭示了从自上而下设计原则创建的大量理想化蛋白质设计空间。特别是随着约束复杂度的增加，许多相应的设计高度理想化，类似于人工创建的机器和系统的规律性。

Our computational results using two independent inverse folding methods suggest that the generated structures are designable, since inverse folding models have demonstrated high experimental success rates (23). We are also obtaining data to experimentally validate the designs.

我们使用两种独立的逆折叠方法的计算结果表明，生成的结构是可设计的，因为逆折叠模型已显示出较高的实验成功率(23)。我们还在获取数据以实验验证这些设计。

More broadly, the formalization offered by a high-level programming language enables logical design principles to be applied to protein design as in other fields of engineering. This has been especially challenging in biology due to the way that the amino acid sequence opaquely encodes the structure and function. As protein design moves toward the engineering of more complex functions, we anticipate that such a system will become increasingly useful.

更广泛地说，高级编程语言提供的形式化使得逻辑设计原则能够像其他工程领域一样应用于蛋白质设计。这在生物学中尤其具有挑战性，因为氨基酸序列以不透明的方式编码结构和功能。随着蛋白质设计向更复杂功能的工程化迈进，我们预计这样的系统将变得越来越有用。

# Acknowledgements

# 致谢

We would like to thank Halil Akin, Adam Lerer, Wenting Lu, Yaniv Shmueli, Robert Verkuil, and Zhongkai Zhu for technical help, feedback, and discussions that helped shape this project. We thank Laurens van der Maaten, Ammar Rizvi, Jon Shepard, and Joe Spisak for program support.

我们要感谢Halil Akin、Adam Lerer、Wenting Lu、Yaniv Shmueli、Robert Verkuil和Zhongkai Zhu的技术帮助、反馈和讨论，这些帮助塑造了这个项目。我们感谢Laurens van der Maaten、Ammar Rizvi、Jon Shepard和Joe Spisak的程序支持。

# References

# 参考文献

[1] Cyrus Levinthal. Are there pathways for protein folding? Extrait du Journal de Chimie Physique, 65(1):44, 1968.

Cyrus Levinthal. 蛋白质折叠有路径吗？《物理化学杂志》摘录，65(1):44, 1968.

[2] A M Lesk and G D Rose. Folding units in globular proteins. Proc. Natl. Acad. Sci. U. S. A., 78(7):4304- 4308, July 1981.

A M Lesk 和 G D Rose. 球状蛋白质中的折叠单元。《美国国家科学院院刊》，78(7):4304-4308, 1981年7月.

[3] William R Taylor. A ’periodic table’ for protein structures. Nature, 416(6881):657-660, April 2002.

William R Taylor. 蛋白质结构的“周期表”。《自然》，416(6881):657-660, 2002年4月.

[4] Craig O Mackenzie, Jianfu Zhou, and Gevorg Grig-oryan. Tertiary alphabet for the observable protein structural universe. Proc. Natl. Acad. Sci. U. S. A., 113 (47):E7438-E7447, November 2016.

Craig O Mackenzie、Jianfu Zhou 和 Gevorg Grigoryan. 可观察蛋白质结构宇宙的三级字母表。《美国国家科学院院刊》，113(47):E7438-E7447, 2016年11月.

[5] W Kabsch and C Sander. On the use of sequence homologies to predict protein structure: Identical pen-tapeptides can have completely different conformations. Proc. Natl. Acad. Sci. U. S. A., 81(4):1075-1078, February 1984.

W Kabsch 和 C Sander. 关于使用序列同源性预测蛋白质结构:相同的五肽可以具有完全不同的构象。《美国国家科学院院刊》，81(4):1075-1078, 1984年2月.

[6] D L Minor, Jr and P S Kim. Context-dependent secondary structure formation of a designed protein sequence. Nature, 380(6576):730-734, April 1996.

D L Minor, Jr 和 P S Kim. 设计蛋白质序列的上下文依赖性二级结构形成。《自然》，380(6576):730-734, 1996年4月.

[7] P B Harbury, J J Plecs, B Tidor, T Alber, and P S Kim. High-resolution protein design with backbone freedom. Science, 282(5393):1462-1467, November 1998.

P B Harbury, J J Plecs, B Tidor, T Alber, 和 P S Kim. 具有骨架自由的高分辨率蛋白质设计. 科学, 282(5393):1462-1467, 1998年11月.

[8] Brian Kuhlman, Gautam Dantas, Gregory C Ireton, Gabriele Varani, Barry L Stoddard, and David Baker. Design of a novel globular protein fold with atomic-level accuracy. Science, 302(5649):1364-1368, November 2003.

Brian Kuhlman, Gautam Dantas, Gregory C Ireton, Gabriele Varani, Barry L Stoddard, 和 David Baker. 以原子级精度设计新型球状蛋白质折叠. 科学, 302(5649):1364-1368, 2003年11月.

[9] Hui-Hsu Gavin Tsai, Chung-Jung Tsai, Buyong Ma, and Ruth Nussinov. In silico protein design by combinatorial assembly of protein building blocks. Protein Sci., 13(10):2753-2765, October 2004.

Hui-Hsu Gavin Tsai, Chung-Jung Tsai, Buyong Ma, 和 Ruth Nussinov. 通过蛋白质构建块的组合组装进行计算机蛋白质设计. 蛋白质科学, 13(10):2753-2765, 2004年10月.

[10] Nobuyasu Koga, Rie Tatsumi-Koga, Gaohua Liu, Rong Xiao, Thomas B Acton, Gaetano T Montelione, and David Baker. Principles for designing ideal protein structures. Nature, 491(7423):222-227, November 2012.

Nobuyasu Koga, Rie Tatsumi-Koga, Gaohua Liu, Rong Xiao, Thomas B Acton, Gaetano T Montelione, 和 David Baker. 设计理想蛋白质结构的原则. 自然, 491(7423):222-227, 2012年11月.

[11] Che Yang, Fabian Sesterhenn, Jaume Bonet, Eva A van Aalen, Leo Scheller, Luciano A Abriata, Johannes T Cramer, Xiaolin Wen, Stéphane Rosset, Sandrine Georgeon, Theodore Jardetzky, Thomas Krey, Martin Fussenegger, Maarten Merkx, and Bruno E Correia. Bottom-up de novo design of functional proteins with complex structural features. Nat. Chem. Biol., 17(4): 492-500, April 2021.

Che Yang, Fabian Sesterhenn, Jaume Bonet, Eva A van Aalen, Leo Scheller, Luciano A Abriata, Johannes T Cramer, Xiaolin Wen, Stéphane Rosset, Sandrine Georgeon, Theodore Jardetzky, Thomas Krey, Martin Fussenegger, Maarten Merkx, 和 Bruno E Correia. 自下而上的功能性蛋白质从头设计，具有复杂的结构特征. 自然化学生物学, 17(4): 492-500, 2021年4月.

[12] H Zhao, L Giver, Z Shao, J A Affholter, and F H Arnold. Molecular evolution by staggered extension process (StEP) in vitro recombination. Nat. Biotech-nol., 16(3):258-261, March 1998. A high-level programming language for generative protein design

H Zhao, L Giver, Z Shao, J A Affholter, 和 F H Arnold. 通过交错延伸过程(StEP)进行体外重组的分子进化. 自然生物技术, 16(3):258-261, 1998年3月. 一种用于生成蛋白质设计的高级编程语言

[13] Christopher A Voigt, Carlos Martinez, Zhen-Gang Wang, Stephen L Mayo, and Frances H Arnold. Protein building blocks preserved by recombination. Nat. Struct. Biol., 9(7):553-558, July 2002.

Christopher A Voigt, Carlos Martinez, Zhen-Gang Wang, Stephen L Mayo, 和 Frances H Arnold. 通过重组保存的蛋白质构建块. 自然结构生物学, 9(7):553-558, 2002年7月.

[14] Marc Ostermeier. Engineering allosteric protein switches by domain insertion. Protein Eng. Des. Sel., 18(8):359-364, August 2005.

Marc Ostermeier. 通过域插入工程化变构蛋白质开关. 蛋白质工程设计与选择, 18(8):359-364, 2005年8月.

[15] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. U. S. A., 118(15):e2016239118, April 2021.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, 和 Rob Fergus. 通过将无监督学习扩展到2.5亿个蛋白质序列，生物结构和功能得以涌现. 美国国家科学院院刊, 118(15):e2016239118, 2021年4月.

[16] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model. bioRxiv, 2022. doi: 10.1101/2022. 07.20.500902.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, 和 Alexander Rives. 使用语言模型进行进化尺度的原子级蛋白质结构预测. bioRxiv, 2022. doi: 10.1101/2022. 07.20.500902.

[17] Ivan Anishchenko, Samuel J Pellock, Tamuka M Chidyausiku, Theresa A Ramelot, Sergey Ovchin-nikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K Bera, Frank DiMaio, Lauren Carter, Cameron M Chow, Gaetano T Montelione, and David Baker. De novo protein design by deep network hallucination. Nature, 600(7889):547-552, December 2021.

Ivan Anishchenko, Samuel J Pellock, Tamuka M Chidyausiku, Theresa A Ramelot, Sergey Ovchin-nikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K Bera, Frank DiMaio, Lauren Carter, Cameron M Chow, Gaetano T Montelione, 和 David Baker. 通过深度网络幻觉进行蛋白质从头设计. 自然, 600(7889):547-552, 2021年12月.

[18] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielin-ski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873):583-589, August 2021.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielin-ski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, 和 Demis Hassabis。使用AlphaFold进行高精度蛋白质结构预测。《自然》，596(7873):583-589，2021年8月。

[19] Po-Ssu Huang, Scott E Boyken, and David Baker. The coming of age of de novo protein design. Nature, 537 (7620):320-327, September 2016.

Po-Ssu Huang, Scott E Boyken, 和 David Baker。从头蛋白质设计的成熟。《自然》，537(7620):320-327，2016年9月。

[20] Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L Watson, Karla M Castro, Robert Ragotte, Amijai Saragovi, Lukas F Milles, Minkyung Baek, Ivan Anishchenko, Wei Yang, Derrick R Hicks, Marc Expösit, Thomas Schlichthaerle, Jung-Ho Chun, Justas Dauparas, Nathaniel Bennett, Basile I M Wicky, Andrew Muenks, Frank DiMaio, Bruno Correia, Sergey Ovchinnikov, and David Baker. Scaffolding protein functional sites using deep learning. Science, 377(6604):387-394, July 2022.

Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L Watson, Karla M Castro, Robert Ragotte, Amijai Saragovi, Lukas F Milles, Minkyung Baek, Ivan Anishchenko, Wei Yang, Derrick R Hicks, Marc Expösit, Thomas Schlichthaerle, Jung-Ho Chun, Justas Dauparas, Nathaniel Bennett, Basile I M Wicky, Andrew Muenks, Frank DiMaio, Bruno Correia, Sergey Ovchinnikov, 和 David Baker。使用深度学习搭建蛋白质功能位点。《科学》，377(6604):387-394，2022年7月。

[21] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In Proceedings of the 39th International Conference on Machine Learning, pages 8946- 8970, 2022.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, 和 Alexander Rives。从数百万预测结构中学习逆向折叠。第39届国际机器学习会议论文集，8946-8970页，2022年。

[22] B I M Wicky, L F Milles, A Courbet, R J Ragotte, J Dauparas, E Kinfu, S Tipps, R D Kibler, M Baek, F DiMaio, X Li, L Carter, A Kang, H Nguyen, A K Bera, and D Baker. Hallucinating symmetric protein assemblies. Science, 378(6615):56-61, October 2022.

B I M Wicky, L F Milles, A Courbet, R J Ragotte, J Dauparas, E Kinfu, S Tipps, R D Kibler, M Baek, F DiMaio, X Li, L Carter, A Kang, H Nguyen, A K Bera, 和 D Baker。幻觉对称蛋白质组装。《科学》，378(6615):56-61，2022年10月。

[23] J Dauparas, I Anishchenko, N Bennett, H Bai, R J Ragotte, L F Milles, B I M Wicky, A Courbet, R J de Haas, N Bethel, P J Y Leung, T F Huddy, S Pellock, D Tischer, F Chan, B Koepnick, H Nguyen, A Kang, B Sankaran, A K Bera, N P King, and D Baker. Robust deep learning-based protein sequence design using ProteinMPNN. Science, 378(6615):49-56, October 2022.

J Dauparas, I Anishchenko, N Bennett, H Bai, R J Ragotte, L F Milles, B I M Wicky, A Courbet, R J de Haas, N Bethel, P J Y Leung, T F Huddy, S Pellock, D Tischer, F Chan, B Koepnick, H Nguyen, A Kang, B Sankaran, A K Bera, N P King, 和 D Baker。使用ProteinMPNN进行稳健的深度学习蛋白质序列设计。《科学》，378(6615):49-56，2022年10月。

[24] Georges Vauquelin and Steven J Charlton. Exploring avidity: understanding the potential gains in functional affinity and target residence time of bivalent and heterobivalent ligands. Br. J. Pharmacol., 168(8): 1771-1785, April 2013.

Georges Vauquelin 和 Steven J Charlton。探索亲和力:理解双价和异双价配体在功能亲和力和目标停留时间上的潜在增益。《英国药理学杂志》，168(8):1771-1785，2013年4月。

[25] William P Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. An evolution-based model for designing chorismate mutase enzymes. Science, 369(6502):440-445, July 2020.

William P Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, 和 Rama Ranganathan。基于进化的模型设计分支酸变位酶。《科学》，369(6502):440-445，2020年7月。

[26] Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Deep neural language modeling enables functional protein generation across families. bioRxiv, 2021. doi: 10.1101/2021.07.18.452833.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, 和 Nikhil Naik。深度神经语言模型实现跨家族功能蛋白质生成。bioRxiv，2021年。doi: 10.1101/2021.07.18.452833。

[27] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. Nat. Commun., 13(1):4348, July 2022.

Noelia Ferruz, Steffen Schmidt, 和 Birte Höcker。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(1):4348，2022年7月。

[28] John Ingraham, Max Baranov, Zak Costello, Vincent Frappier, Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer, Andrew Beam, and Gevorg A high-level programming language for generative protein design

John Ingraham, Max Baranov, Zak Costello, Vincent Frappier, Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer, Andrew Beam, 和 Gevorg 一种用于生成蛋白质设计的高级编程语言

Grigoryan. Illuminating protein space with a programmable generative model. bioRxiv, 2022. doi: 10.1101/2022.12.01.518682.

Grigoryan. 用可编程生成模型照亮蛋白质空间. bioRxiv, 2022. doi: 10.1101/2022.12.01.518682.

[29] 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 Shef-fler, Jue Wang, Preetham Venkatesh, Isaac Sapping-ton, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. bioRxiv, 2022. doi: 10.1101/2022.12.09.519842.

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, 和 David Baker. 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. bioRxiv, 2022. doi: 10.1101/2022.12.09.519842.

[30] A Shrake and J A Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. . Mol. Biol., 79(2):351-371, September 1973.

A Shrake 和 J A Rupley. 蛋白质原子的环境和溶剂暴露. 溶菌酶和胰岛素. . Mol. Biol., 79(2):351-371, 1973年9月.

[31] Patrick Kunzmann and Kay Hamacher. Biotite: a unifying open source computational biology framework in python. BMC Bioinformatics, 19(1):346, 2018.

Patrick Kunzmann 和 Kay Hamacher. Biotite: Python中统一的开源计算生物学框架. BMC Bioinformatics, 19(1):346, 2018.

[32] G Labesse, N Colloc’h, J Pothier, and J P Mornon. P-SEA: a new efficient assignment of secondary structure from alpha trace of proteins. Comput. Appl. Biosci., 13(3):291-295, June 1997.

G Labesse, N Colloc’h, J Pothier, 和 J P Mornon. P-SEA: 从 蛋白质α轨迹中高效分配二级结构的新方法. Comput. Appl. Biosci., 13(3):291-295, 1997年6月.

[33] W. Kabsch. A solution for the best rotation to relate two sets of vectors. Acta Crys-tallographica Section A, 32(5):922-923, 1976. doi: https://doi.org/10.1107/S0567739476001873.URL https://onlinelibrary.wiley.com/ doi/abs/10.1107/S0567739476001873.

W. Kabsch. 两组向量之间最佳旋转的解决方案. Acta Crystallographica Section A, 32(5):922-923, 1976. doi: https://doi.org/10.1107/S0567739476001873.URL https://onlinelibrary.wiley.com/ doi/abs/10.1107/S0567739476001873.

[34] Sung Il Yoon, Brandi C Jones, Naomi J Logsdon, and Mark R Walter. Same structure, different function crystal structure of the Epstein-Barr virus IL-10 bound to the soluble IL-10R1 chain. Structure, 13(4):551- 564, April 2005.

Sung Il Yoon, Brandi C Jones, Naomi J Logsdon, 和 Mark R Walter. 相同结构，不同功能:Epstein-Barr病毒IL-10与可溶性IL-10R1链结合的晶体结构. Structure, 13(4):551-564, 2005年4月.

[35] Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, Xuanling Shi, Qisheng Wang, Linqi Zhang, and Xinquan Wang. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature, 581(7807):215-220, May 2020.

Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, Xuanling Shi, Qisheng Wang, Linqi Zhang, 和 Xinquan Wang. SARS-CoV-2刺突蛋白受体结合域与ACE2受体结合的结构. Nature, 581(7807):215-220, 2020年5月.

[36] G Szakonyi, J M Guthridge, D Li, K Young, V M Holers, and X S Chen. Structure of complement receptor 2 in complex with its C3d ligand. Science, 292 (5522):1725-1728, June 2001.

G Szakonyi, J M Guthridge, D Li, K Young, V M Holers, 和 X S Chen. 补体受体2与其C3d配体复合物的结构. Science, 292(5522):1725-1728, 2001年6月.

[37] Nicole L Kallewaard, Davide Corti, Patrick J Collins, Ursula Neu, Josephine M McAuliffe, Ebony Benjamin, Leslie Wachter-Rosati, Frances J Palmer-Hill, Andy Q Yuan, Philip A Walker, Matthias K Vorlaender, Siro Bianchi, Barbara Guarino, Anna De Marco, Fabrizia Vanzetta, Gloria Agatic, Mathilde Foglierini, Debora Pinna, Blanca Fernandez-Rodriguez, Alexander Fruehwirth, Chiara Silacci, Roksana W Ogrodowicz, Stephen R Martin, Federica Sallusto, Joann A Suzich, Antonio Lanzavecchia, Qing Zhu, Steven J Gamblin, and John J Skehel. Structure and function analysis of an antibody recognizing all influenza a subtypes. Cell, 166(3):596-608, July 2016.

Nicole L Kallewaard, Davide Corti, Patrick J Collins, Ursula Neu, Josephine M McAuliffe, Ebony Benjamin, Leslie Wachter-Rosati, Frances J Palmer-Hill, Andy Q Yuan, Philip A Walker, Matthias K Vorlaender, Siro Bianchi, Barbara Guarino, Anna De Marco, Fabrizia Vanzetta, Gloria Agatic, Mathilde Foglierini, Debora Pinna, Blanca Fernandez-Rodriguez, Alexander Fruehwirth, Chiara Silacci, Roksana W Ogrodowicz, Stephen R Martin, Federica Sallusto, Joann A Suzich, Antonio Lanzavecchia, Qing Zhu, Steven J Gamblin, 和 John J Skehel. 识别所有甲型流感亚型的抗体的结构和功能分析. 细胞, 166(3):596-608, 2016年7月.

[38] Kathryn Westendorf, Stefanie Zentelis, Lingshu Wang, Denisa Foster, Peter Vaillancourt, Matthew Wiggin, Erica Lovett, Robin van der Lee, Jörg Hendle, Anna Pustilnik, J Michael Sauder, Lucas Kraft, Yuri Hwang, Robert W Siegel, Jinbiao Chen, Beverly A Heinz, Richard E Higgs, Nicole L Kallewaard, Kevin Jepson, Rodrigo Goya, Maia A Smith, David W Collins, Davide Pellacani, Ping Xiang, Valentine de Puyrai-mond, Marketa Ricicova, Lindsay Devorkin, Caitlin Pritchard, Aoise O’Neill, Kush Dalal, Pankaj Panwar, Harveer Dhupar, Fabian A Garces, Courtney A Cohen, John M Dye, Kathleen E Huie, Catherine V Badger, Darwyn Kobasa, Jonathan Audet, Joshua J Fre-itas, Saleema Hassanali, Ina Hughes, Luis Munoz, Holly C Palma, Bharathi Ramamurthy, Robert W Cross, Thomas W Geisbert, Vineet Menachery, Kumari Lokugamage, Viktoriya Borisevich, Iliana Lanz, Lisa Anderson, Payal Sipahimalani, Kizzmekia S Corbett, Eun Sung Yang, Yi Zhang, Wei Shi, Tongqing Zhou, Misook Choe, John Misasi, Peter D Kwong, Nancy J Sullivan, Barney S Graham, Tara L Fernandez, Carl L Hansen, Ester Falconer, John R Mascola, Bryan E Jones, and Bryan C Barnhart. LY-CoV1404 (bebtelovimab) potently neutralizes SARS-CoV-2 variants. Cell Rep., 39(7):110812, May 2022.

Kathryn Westendorf, Stefanie Zentelis, Lingshu Wang, Denisa Foster, Peter Vaillancourt, Matthew Wiggin, Erica Lovett, Robin van der Lee, Jörg Hendle, Anna Pustilnik, J Michael Sauder, Lucas Kraft, Yuri Hwang, Robert W Siegel, Jinbiao Chen, Beverly A Heinz, Richard E Higgs, Nicole L Kallewaard, Kevin Jepson, Rodrigo Goya, Maia A Smith, David W Collins, Davide Pellacani, Ping Xiang, Valentine de Puyrai-mond, Marketa Ricicova, Lindsay Devorkin, Caitlin Pritchard, Aoise O’Neill, Kush Dalal, Pankaj Panwar, Harveer Dhupar, Fabian A Garces, Courtney A Cohen, John M Dye, Kathleen E Huie, Catherine V Badger, Darwyn Kobasa, Jonathan Audet, Joshua J Fre-itas, Saleema Hassanali, Ina Hughes, Luis Munoz, Holly C Palma, Bharathi Ramamurthy, Robert W Cross, Thomas W Geisbert, Vineet Menachery, Kumari Lokugamage, Viktoriya Borisevich, Iliana Lanz, Lisa Anderson, Payal Sipahimalani, Kizzmekia S Corbett, Eun Sung Yang, Yi Zhang, Wei Shi, Tongqing Zhou, Misook Choe, John Misasi, Peter D Kwong, Nancy J Sullivan, Barney S Graham, Tara L Fernandez, Carl L Hansen, Ester Falconer, John R Mascola, Bryan E Jones, 和 Bryan C Barnhart. LY-CoV1404 (bebtelovimab) 强效中和SARS-CoV-2变体. 细胞报告, 39(7):110812, 2022年5月.

[39] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. Nucleic Acids Res., 28(1):235-242, January 2000.

H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, 和 P E Bourne. 蛋白质数据库. 核酸研究, 28(1):235-242, 2000年1月.

[40] Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res., 33(7):2302-2309, April 2005. A high-level programming language for generative protein design

Yang Zhang 和 Jeffrey Skolnick. TM-align: 基于TM-score的蛋白质结构比对算法. 核酸研究, 33(7):2302-2309, 2005年4月. 用于生成蛋白质设计的高级编程语言

# A. Methods

# A. 方法

# A.1. High-level programming language and energy-based optimization

# A.1. 高级编程语言和基于能量的优化

A program in our language is fully specified by (1) a syntax tree and (2) a set of constraints. This program compiles to an energy function, which is used to guide black-box optimization of a protein sequence while also leveraging its predicted structure.

我们的语言程序完全由(1)语法树和(2)一组约束条件指定。该程序编译为一个能量函数，用于指导蛋白质序列的黑箱优化，同时利用其预测结构。

# A.1.1. Syntax tree

# A.1.1. 语法树

The syntax tree consists of nonterminal symbols, which we denote as , as well as terminal symbols, which in our examples we denote as uppercase alphabetic characters such as , etc. Each terminal symbol defines a unique protein sequence. The nonterminal symbol is designated as the special start symbol; all programs must have . Additional nonterminal symbols are used to define hierarchical complexity. For example, specifying two levels of hierarchy requires a nonterminal production rule in addition to a terminal production rule, for example,

语法树由非终结符(我们表示为 )和终结符(在我们的示例中表示为大写字母字符，如 等)组成。每个终结符定义一个唯一的蛋白质序列。非终结符 被指定为特殊的起始符号；所有程序都必须包含 。额外的非终结符用于定义层次复杂性。例如，指定两个层次需要非终结符产生规则和终结符产生规则，例如，

In the example above, the nonterminal enables an intermediate level of hierarchy. A nonterminal can produce any finite-length permutation of higher-numbered nonterminals (for example, is permitted but not or ). A nonterminal can also produce any finite-length permutation of terminals (for example, ) or any finite-length permutation of mixed terminals and higher-numbered nonterminals (for example, or .

在上面的例子中， 非终结符(nonterminal)启用了一个中间层次的层级结构。一个非终结符可以生成任何有限长度的高编号非终结符的排列(例如， 是允许的，但 或 则不允许)。一个非终结符也可以生成任何有限长度的终结符(terminal)的排列(例如， )，或者任何有限长度的混合终结符和高编号非终结符的排列(例如， 或 )。

A complete syntax tree is built by fully expanding the nonterminal into a set of terminals. The production rules define the connectivity structure of the tree, where the parent node corresponds to the left side of the production rule and the child node(s) corresponds to the right side of the production rule. Across the entire syntax tree, each internal node corresponds to a nonterminal symbol and each leaf corresponds to a terminal symbol. Example syntax trees are provided in the main text figures.

通过将非终结符 完全扩展为一组终结符，构建了一个完整的语法树。产生式规则定义了树的连接结构，其中父节点对应于产生式规则的左侧，子节点对应于产生式规则的右侧。在整个语法树中，每个内部节点对应于一个非终结符号，每个叶节点对应于一个终结符号。示例语法树在正文图中提供。

# A.1.2. CONSTRAINTS

# A.1.2. 约束

A program in our language also requires a set of constraints, where a single constraint is defined with respect to a single node and all of its descendants in the syntax tree. Note that this includes constraints on the leaves of the tree (corresponding to the terminal symbols). More specifically, a constraint is a function that takes as input the (sub)tree, as well as its corresponding (sub)sequence and (sub)structure, and outputs a real number. For example, a constraint defined with respect to the node corresponding to simply receives as input the entire syntax tree, the full-length sequence, and the full protein structure. The same constraint (i.e., the same function) can be applied to multiple nodes in the tree. We will use to denote a constraint defined with respect to the node.

我们语言中的程序还需要一组约束，其中单个约束是相对于语法树中的单个节点及其所有后代定义的。请注意，这包括对树的叶节点(对应于终结符号)的约束。更具体地说，约束是一个函数，它接受(子)树及其对应的(子)序列和(子)结构作为输入，并输出一个实数。例如，相对于对应于 的节点定义的约束，只需接收整个语法树、完整长度的序列和完整的蛋白质结构作为输入。相同的约束(即相同的函数)可以应用于树中的多个节点。我们将使用 来表示相对于 节点定义的约束 。

# A.1.3. COMPILATION OF CONSTRAINTS INTO AN ENERGY FUNCTION

# A.1.3. 将约束编译为能量函数

We compile a program into an energy function. In our study, we simply compute a linear combination of all the constraints in the user-specified set, i.e., , where is defined as zero when a constraint is not applied to a given node. In practice, we explicitly keep track of a scalar multiplicative weight on each constraint, i.e., . This energy is used in the simulated annealing optimization procedure described below. Specific examples of constraint functions used in our study are also provided below.

我们将程序编译为能量函数。在我们的研究中，我们简单地计算用户指定集合中所有约束的线性组合，即 ，其中 在约束未应用于给定节点时定义为零。在实践中，我们明确跟踪每个约束的标量乘法权重，即 。该能量用于下面描述的模拟退火优化过程。我们研究中使用的约束函数的具体示例也在下面提供。

Linear combinations work well for our choice of generative model, but in principle any combination of the energy terms could be used here. For example, if we compiled our program into an energy function for a generative model that used a reward function (like a reinforcement learning agent), we might prefer a multiplicative combination of the inverse of our current energy terms.

线性组合适用于我们选择的生成模型，但原则上这里可以使用任何能量项的组合。例如，如果我们将程序编译为使用奖励函数(如强化学习代理)的生成模型的能量函数，我们可能更喜欢当前能量项的逆的乘法组合。

# A.1.4. Simulated annealing

# A.1.4. 模拟退火

The energy function is used as part of an iterative black-box optimization loop, where over multiple iterations, a change to a given state (in this case, a protein sequence design) is accepted with some probability. We use a simulated annealing algorithm in which the acceptance probability is controlled by a temperature value such that the optimization can tolerate higher energy changes at the beginning of the optimization before favoring changes that decrease the energy toward the end of the optimization. In our study, we begin by initializing the sequence state (one unique sequence per terminal symbol) with uniform amino-acid probability to a given user-specified length; we also compute an initial structure prediction from this sequence.

能量函数用作迭代黑盒优化循环的一部分，在多次迭代中，以某种概率接受对给定状态(在本例中为蛋白质序列设计)的更改。我们使用模拟退火算法，其中接受概率由温度值控制，以便优化在开始时可以容忍更高的能量变化，然后在优化结束时倾向于减少能量的变化。在我们的研究中，我们首先以均匀的氨基酸概率初始化序列状态(每个终结符号一个唯一序列)到用户指定的长度；我们还从该序列计算初始结构预测。

Each iteration proposes a mutation to the protein sequence. To make this proposal, first, one of the terminal symbols is chosen with uniform probability, and second, one of a substitution, insertion, or deletion is chosen with some probability (we default to , and , respectively). For substitutions and insertions, the new amino acid is chosen with some probability (unless otherwise specified, we apply uniform probability over a reduced amino acid alphabet that excludes cysteine). We default to uniform probability over A high-level programming language for generative protein design all possible sequence positions.

每次迭代都会对蛋白质序列提出一个突变。为了提出这个突变，首先，以均匀概率选择一个终端符号，其次，以某种概率选择替换、插入或删除(我们默认分别为 和 )。对于替换和插入，新氨基酸以某种概率选择(除非另有说明，我们应用均匀概率在排除半胱氨酸的简化氨基酸字母表上)。我们默认在所有可能的序列位置上应用均匀概率。

The next step in the iteration is to obtain a structure prediction corresponding to the sequence with the proposed change. This prediction provides the structural information that is used to compute the values of the individual constraint functions. These values are then combined to produce the value of the energy function, as described above. This energy function is evaluated on the overall design with and without the proposed mutation, which we denote and , respectively. The mutation is accepted with probability

迭代的下一步是获取与提出变化的序列对应的结构预测。该预测提供了用于计算各个约束函数值的结构信息。然后，这些值被组合以产生能量函数的值，如上所述。该能量函数在整体设计上评估，包括和不包括提出的突变，我们分别表示为 和 。突变以概率被接受

where is the temperature parameter at iteration that decays geometrically over the optimization. By default, our optimization leverages user-specified values and , with a decay schedule given as

其中 是迭代 时的温度参数，在优化过程中几何衰减。默认情况下，我们的优化利用用户指定的值 和 ，衰减计划如下

where is the user-specified number of annealing steps. We report specific values for and in the experiment descriptions below and default to .

其中 是用户指定的退火步骤数。我们在下面的实验描述中报告 和 的具体值，并默认 。

# A.1.5. Single chain constraint

# A.1.5. 单链约束

Our language accommodates both single- and multi-chain design through the use of a special "single chain" constraint. By default, without this constraint applied, all terminal symbols are assumed to correspond to separate chains. When this constraint is applied to a given node, it constrains all of the terminal symbols to be part of a single chain according to the left-to-right order defined in the syntax tree. For example, consider a syntax subtree with and productions. A single chain constraint applied to node would create a chain consisting of a contiguous sequence . Unlike other constraints, this constraint is enforced as part of structure prediction, prior to the energy function compilation.

我们的语言通过使用特殊的“单链”约束来适应单链和多链设计。默认情况下，如果没有应用此约束，所有终端符号都被假定为对应不同的链。当此约束应用于给定节点时，它约束所有终端符号成为根据语法树中定义的从左到右顺序的单链的一部分。例如，考虑一个具有 和 生成的语法子树。应用于节点 的单链约束将创建一个由连续序列 组成的链。与其他约束不同，此约束在能量函数编译之前作为结构预测的一部分强制执行。

# A.2. Constraint implementation

# A.2. 约束实现

# A.2.1. ESMFOLD STRUCTURE PREDICTION

# A.2.1. ESMFOLD结构预测

We obtain all-atom structure predictions using ESMFold (16), where the prediction is made over the entire protein sequence and is represented as a set of atomic coordinates and their corresponding residue identities and indices. This predicted structure is the basis for the structural information passed to each of the specific constraint functions. When a constraint is defined on a subtree, that constraint only has access to the structural information (atomic coordinates, etc.) of the sequence encoded by that subtree.

我们使用ESMFold(16)获取全原子结构预测，其中预测是在整个蛋白质序列上进行的，并表示为原子坐标及其对应的残基身份和索引的集合。该预测结构是传递给每个特定约束函数的结构信息的基础。当约束定义在子树上时，该约束只能访问由该子树编码的序列的结构信息(原子坐标等)。

# A.2.2. STRUCTURE PREDICTION CONFIDENCE (PTM AND PLDDT)

# A.2.2. 结构预测置信度(PTM和PLDDT)

ESMFold produces a pTM score, which indicates the model’s confidence in the overall structure prediction, and a per-atom pLDDT score, which indicates the model’s confidence in the specific atomic coordinate prediction. The pTM value and the mean of the backbone pLDDT values are constraints that are meant to steer the optimization toward structures with higher structure prediction confidence, which is associated with naturally plausible and designable structures. We use a linear combination of the quantities 1 - pTM and 1 - pLDDT (since a higher confidence/lower energy is desirable), with user-specified weights, as the returned value of the confidence constraint.

ESMFold生成一个pTM分数，表示模型对整体结构预测的置信度，以及每个原子的pLDDT分数，表示模型对特定原子坐标预测的置信度。pTM值和主链pLDDT值的均值是旨在引导优化朝向具有更高结构预测置信度的结构的约束，这与自然合理和可设计的结构相关。我们使用1 - pTM和1 - pLDDT的线性组合(因为更高的置信度/更低的能量是可取的)，并带有用户指定的权重，作为置信度约束的返回值。

# A.2.3. SURFACE-EXPOSED HYDROPHOBICS

# A.2.3. 表面暴露的疏水性

The surface exposed hydrophobics constraint aims to reduce the hydrophobicity of the protein surface, where high hydrophobicity leads to protein aggregation and insolubility. We implement this constraint using the Shrake-Rupley "rolling probe" algorithm to determine the surface exposed atoms (30) as implemented in the biotite Python package version 0.35.0 (31). We then calculate the fraction of atoms involved in hydrophobic residues that are also surface exposed, and we use this fraction as the output of the constraint function.

表面暴露的疏水性约束旨在减少蛋白质表面的疏水性，高疏水性会导致蛋白质聚集和不溶性。我们使用Shrake-Rupley“滚动探针”算法来确定表面暴露的原子(30)，如biotite Python包版本0.35.0(31)中实现的那样。然后，我们计算涉及疏水性残基且表面暴露的原子的比例，并将此比例用作约束函数的输出。

# A.2.4. GLOBULARITY

# A.2.4. 球状性

It is sometimes desirable to encourage a protein chain to pack into a globular structure. Our globularity constraint is implemented by computing the centroid of a set of atomic coordinates, where the globularity constraint function returns the variance of the distances from all coordinates to this centroid. Intuitively, low variance indicates that all coordinates that are largely equidistant to the centroid, which is more consistent with globular packing.

有时需要促使蛋白质链折叠成球状结构。我们的球状性约束通过计算一组原子坐标的质心来实现，其中球状性约束函数返回所有坐标到该质心距离的方差。直观上，低方差表明所有坐标到质心的距离大致相等，这与球状折叠更为一致。

# A.2.5. SECONDARY STRUCTURE

# A.2.5. 二级结构

The secondary structure constraint steers the energy toward user-defined secondary structure. To annotate residue secondary structure, we use the P-SEA algorithm (32) as implemented by the biotite Python package (31). This constraint function returns one minus the fraction of residues that belong to the desired secondary structure element (since a higher fraction/lower energy is desirable).

二级结构约束将能量引导向用户定义的二级结构。为了注释残基的二级结构，我们使用由biotite Python包(31)实现的P-SEA算法(32)。该约束函数返回1减去属于所需二级结构元素的残基比例(因为较高的比例/较低的能量是理想的)。

# A.2.6. ROTATIONAL SYMMETRY

# A.2.6. 旋转对称性

To design symmetry, we first find it helpful to tie the sequence identities across the subsequences corresponding to the asymmetric units. The first symmetry we consider is rotational symmetry, which only consider the centroids A high-level programming language for generative protein design of the immediate children of the constraint’s node; for example, rotational symmetry defined on a node where would only consider the centroids of the individual substructures defined by , or .

为了设计对称性，我们首先发现将序列身份与对应于不对称单元的子序列绑定是有帮助的。我们考虑的第一个对称性是旋转对称性，它只考虑约束节点直接子节点的质心；例如，定义在节点 上的旋转对称性，其中 将只考虑由 或 定义的各个子结构的质心。

Using the left-to-right order of these children in the production rule, the rotational symmetry function first computes the distances among adjacent centroids, circularly wrapping to include the distance between the first and last symbols; for example, rotational symmetry defined on would compute the set of distances among pairs , , and . The final value returned by this constraint function is the variance among all adjacent distances; intuitively, a rotational or ring-like symmetry would have equal distances among centroids. This rotational symmetry function is adopted from that used by Wicky et al. (22).

使用生产规则中这些子节点的从左到右顺序，旋转对称性函数首先计算相邻质心之间的距离，循环包括第一个和最后一个符号之间的距离；例如，定义在 上的旋转对称性将计算 、 和 对之间的距离集。该约束函数返回的最终值是所有相邻距离的方差；直观上，旋转或环状对称性将使质心之间的距离相等。该旋转对称性函数借鉴了Wicky等人(22)使用的方法。

# A.2.7. GLOBULAR SYMMETRY

# A.2.7. 球状对称性

The globular symmetry constraint is defined on the centroids of the immediate children of the constraint’s node; for example, globular symmetry defined on a node where would only consider the centroids of the individual substructures defined by , or (this is similar to rotational symmetry described above). The globularity symmetry function computes all pairwise distances among centroids and returns the variance of these distances. In practice, this constraint function is useful for defining symmetry that is not rotational, for example, the symmetry observed in polyhedral assemblies.

球状对称性约束定义在约束节点直接子节点的质心上；例如，定义在节点 上的球状对称性，其中 将只考虑由 或 定义的各个子结构的质心(这与上述旋转对称性类似)。球状对称性函数计算所有质心之间的成对距离，并返回这些距离的方差。在实践中，该约束函数对于定义非旋转对称性(例如在多面体组装中观察到的对称性)非常有用。

# A.2.8. ALL-ATOM COORDINATION

# A.2.8. 全原子配位

One approach to designing functional proteins is to constrain (a portion of) the protein to match the structure of a known functional site in nature. We accomplish this with an all-atom coordination constraint. This constraint is first defined with respect to a list of atoms from a native protein structure (outside of our designed protein), which we denote . We then constrain all of the atoms in the corresponding (sub)tree to match, which we denote , as closely as possible, the coordination of the atoms in . We achieve this with two functions. The first is the constrained root mean square deviation (cRMSD),

设计功能蛋白质的一种方法是约束(部分)蛋白质以匹配自然界中已知功能位点的结构。我们通过全原子配位约束来实现这一点。该约束首先相对于天然蛋白质结构(我们设计的蛋白质之外)的原子列表定义，我们将其表示为 。然后我们约束相应(子)树中的所有原子尽可能匹配 中原子的配位，我们将其表示为 。我们通过两个函数实现这一点。第一个是约束均方根偏差(cRMSD)，

where is a structural transformation, denotes the atomic coordinates of the th atom out of total atoms considered, and denotes a vector norm. We implement the structural alignment using the Kabsch algorithm (33) as implemented by biotite (31). The second function for constraining atomic coordination is the distance-matrix RMSD (dRMSD),

其中 是结构变换， 表示所考虑的 个原子中第 个原子的坐标， 表示向量范数。我们使用biotite(31)实现的Kabsch算法(33)进行结构对齐。用于约束原子配位的第二个函数是距离矩阵均方根偏差(dRMSD)，

where is the Euclidean distance between the ith and jth atoms. The returned final value is a linear combination of the cRMSD and dRMSD values with user-specified weights. In practice, cRMSD is sometimes excluded (i.e., its weight is set to zero) in conjunction with dRMSD, as cRMSD alone does not appear sufficiently stable to create a sufficiently smooth energy landscape.

其中 是第i个和第j个原子之间的欧几里得距离。返回的最终值是cRMSD和dRMSD值的线性组合，权重由用户指定。在实践中，cRMSD有时会被排除(即其权重设置为零)，与dRMSD结合使用，因为仅cRMSD似乎不足以稳定地创建足够平滑的能量景观。

# A.2.9. BACKBONE ATOM COORDINATION

# A.2.9. 主链原子配位

For a class of design tasks called fixed backbone design, we desire to only constrain the backbone atoms of the protein structure and have the optimization produce sequences that match a known backbone. This constraint is largely equivalent to the all-atom constraint described above, but rather than constraining all atoms (including side chains), this constraint is only applied to the carbon, -carbon, and nitrogen atoms in the protein backbone.

对于一类称为固定主链设计的设计任务，我们希望仅约束蛋白质结构的主链原子，并通过优化生成与已知主链匹配的序列。这种约束在很大程度上等同于上述的全原子约束，但不是约束所有原子(包括侧链)，而是仅应用于蛋白质主链中的碳、 -碳和氮原子。

# A.2.10. SURFACE EXPOSURE

# A.2.10. 表面暴露

In some cases, we desire that a given set of residues be exposed on the surface of the protein (for example, when scaffolding a protein binding site). As with the hydropho-bics constraint, we leverage the Shrake-Rupley algorithm (30) as implemented by biotite (31). We then calculate the fraction of surface exposed atoms within the structure corresponding to the constraint’s subtree, and we use one minus this fraction as the output of the function.

在某些情况下，我们希望给定的一组残基暴露在蛋白质表面(例如，在支架蛋白质结合位点时)。与疏水性约束一样，我们利用biotite(31)实现的Shrake-Rupley算法(30)。然后，我们计算与约束子树对应的结构中表面暴露原子的比例，并使用1减去该比例作为函数的输出。

# A.2.11. LENGTH

# A.2.11. 长度

The length constraint requires a user-specified number of residues. In practice, we can enforce a hard length constraint by disallowing insertions and deletions during the optimization procedure, or through a function that returns increasingly high values when a sequence length goes beyond a user-specified range. In this study, whenever we apply a length constraint we take the former approach.

长度约束要求用户指定残基的数量。在实践中，我们可以通过在优化过程中禁止插入和删除来强制执行硬长度约束，或者通过一个函数在序列长度超出用户指定范围时返回越来越高的值。在本研究中，每当我们应用长度约束时，我们采用前一种方法。

# A.3. Design tasks and experiments

# A.3. 设计任务和实验

# A.3.1. Free HALLUCINATION

# A.3.1. 自由幻觉

Free hallucination simply requires applying confidence and surface-exposed hydrophobic constraints to the whole protein, where we place equal weight on each term (pTM, pLDDT, and hydrophobics). In the experiments described in this study, we ran simulated annealing over 30,000 iter-A high-level programming language for generative protein design ations with across 200 seeds. We also evaluated single-sequence AlphaFold2 (18) on the final sequences produced by these 200 optimization runs.

自由幻觉只需将置信度和表面暴露的疏水性约束应用于整个蛋白质，其中我们对每个项(pTM、pLDDT和疏水性)赋予相同的权重。在本研究描述的实验中，我们在200个种子中运行了30,000次模拟退火迭代，使用 。我们还对这些200次优化运行生成的最终序列进行了单序列AlphaFold2(18)评估。

# A.3.2. FIXED BACKBONE DESIGN

# A.3.2. 固定主链设计

For fixed backbone design, we apply a weight of 2 on the dRMSD constraint, a weight of 1 on the cRMSD, pTM, and pLDDT constraints, and a weight of 0.5 on the hydropho-bics constraint. As the target backbones, we used the de novo structures with PDB IDs 1QYS, 5L33, 6D0T, 6MRS, , and . In the experiments described in this study, we ran simulated annealing over 30,000 iterations with across at least 50 seeds for each de novo backbone.

对于固定主链设计，我们在dRMSD约束上应用权重2，在cRMSD、pTM和pLDDT约束上应用权重1，在疏水性约束上应用权重0.5。作为目标主链，我们使用了PDB ID为1QYS、5L33、6D0T、6MRS、 和 的从头结构。在本研究描述的实验中，我们在每个从头主链上运行了30,000次模拟退火迭代，使用 ，至少50个种子。

# A.3.3. SECONDARY STRUCTURE DESIGN

# A.3.3. 二级结构设计

We performed protein design with partial constraints on a protein by constraining the secondary structure corresponding to different segments of the protein sequence. We place a weight of 10 on the secondary structure constraint and weights of 1 on pTM, pLDDT and hydrophobics constraints. Our programs specify the secondary structure corresponding to two discrete subsequences, where we program (1) all alpha, (2) all beta, and (3) mixed alpha and beta secondary structure. We ran simulated annealing over 30,000 iterations for 10 seeds for each of these three programs (30 optimization trajectories in total).

我们通过对蛋白质序列的不同片段对应的二级结构进行约束，对蛋白质进行了部分约束的蛋白质设计。我们在二级结构约束上应用权重10，在pTM、pLDDT和疏水性约束上应用权重1。我们的程序指定了对应于两个离散子序列的二级结构，其中我们编程了(1)全α，(2)全β，和(3)混合α和β二级结构。我们在30,000次迭代 中为每个程序运行了10个种子的模拟退火(总共30个优化轨迹)。

# A.3.4. SINGLE FUNCTIONAL SITE SCAFFOLDING

# A.3.4. 单功能位点支架

To program functional site scaffolding on a de novo backbone, we divide a single-chain sequence into three segments: a sequence in the middle segment (with an all-atom coordination constraint and a surface exposure constraint) flanked by two "free" sequences. pTM, pLDDT, and hydrophobics constraints are also applied to the full protein. We apply a weight of 2 to the cRMSD and dRMSD constraints, and a weight of 1 to the pTM, pLDDT, and hydrophobics constraints.

为了在从头主链上编程功能位点支架，我们将单链序列分为三个部分:中间部分的序列(具有全原子配位约束和表面暴露约束)两侧是两个“自由”序列。pTM、pLDDT和疏水性约束也应用于整个蛋白质。我们在cRMSD和dRMSD约束上应用权重2，在pTM、pLDDT和疏水性约束上应用权重1。

We attempted to scaffold five protein binding sites, the first three of which were successfully scaffolded by Wang et al. (20):

我们尝试构建五个蛋白质结合位点，其中前三个已由Wang等人成功构建(20):

1. IL10: We used the residue indices 31-40, inclusive, of chain in the PDB structure , corresponding to the IL10 binding site of IL-10R1 (34).

1. IL10:我们使用了PDB结构 中链 的残基索引31-40，对应于IL-10R1的IL10结合位点(34)。

2. ACE2: We used the residue indices 5-23, inclusive, of chain A in the PDB structure , corresponding to the ACE2 binding site of the SARS-CoV-2 spike receptor binding domain (RBD) (35).

2. ACE2:我们使用了PDB结构 中链A的残基索引5-23，对应于SARS-CoV-2刺突蛋白受体结合域(RBD)的ACE2结合位点(35)。

3. C3d: We used the residue indices 104-126 and 170- 184, inclusive, of chain A in the PDB structure 1GHQ, corresponding to the C3d binding site of complement receptor 2 (36).

3. C3d:我们使用了PDB结构1GHQ中链A的残基索引104-126和170-184，对应于补体受体2的C3d结合位点(36)。

4. HA2: We used the residue indices , and 45-49, inclusive, of chain B in the PDB structure 5JW3, corresponding to the influenza HA2 epitope of the antibody MEDI8825 (37).

4. HA2:我们使用了PDB结构5JW3中链B的残基索引 和45-49，对应于抗体MEDI8825的流感HA2表位(37)。

5. RBD: We used the residue indices 439-450 and 498- 506, inclusive, of chain in the PDB structure 7MMO, corresponding to the SARS-CoV-2 RBD epitope of the antibody bebtelovimab (38).

5. RBD:我们使用了PDB结构7MMO中链 的残基索引439-450和498-506，对应于抗体bebtelovimab的SARS-CoV-2 RBD表位(38)。

We ran simulated annealing over 30,000 iterations with for 1,000 seeds for each of the five binding sites (5,000 optimization trajectories in total).

我们对每个结合位点进行了30,000次模拟退火迭代，使用 进行1,000次种子运行(总共5,000次优化轨迹)。

# A.3.5. Symmetric and Homo-OLIGOMER DESIGN

# A.3.5. 对称和同源寡聚体设计

We first program single-chain symmetry using a rotational symmetry constraint applied to the top-level node. In our program, we also tie the sequences across the subsequences corresponding to the asymmetric units such that we only use a single terminal symbol; an example program for designing 3-fold symmetry is provided in Figure 3A. We place a weight of 1 on the symmetry constraint, as well as weights of 1 on pTM, pLDDT, and hydrophobics constraints. We also place length constraints on the terminal nodes. We specify programs where we increase the fold-symmetry from 3- to 8-fold. We also vary the lengths that constrain the terminal symbol such that the full sequence has approximately 200, 300, or 400 residues (for example, a 200-residue protein with 3 -fold symmetry would have length constraints of 66 on its terminal symbols). We ran simulated annealing over 30,000 iterations with a starting temperature of 1 for 10 seeds for each of the six fold symmetries and each of the three length constraints (for a total of 180 optimization trajectories).

我们首先通过将旋转对称约束应用于顶级节点来编程单链对称性。在我们的程序中，我们还将在对应于不对称单元的子序列中绑定序列，以便仅使用单个终端符号；图3A中提供了一个设计3重对称性的示例程序。我们将对称约束的权重设为1，同时将pTM、pLDDT和疏水性约束的权重也设为1。我们还在终端节点上设置了长度约束。我们指定了将对称性从3重增加到8重的程序。我们还改变了约束终端符号的长度，使得完整序列大约有200、300或400个残基(例如，具有3重对称性的200残基蛋白质将在其终端符号上设置长度为66的约束)。我们对每个六重对称性和每个长度约束进行了30,000次模拟退火迭代，起始温度为1，进行10次种子运行(总共180次优化轨迹)。

We also designed larger homo-oligomers similarly, but removing the single-chain constraint from the top-level node. We designed trimeric, tetrameric, hexameric, and octameric homo-oligomers with a globular symmetry constraint applied to the top level node. We placed a weight of 1 on the symmetry constraint, as well as weights of 1 on pTM, pLDDT, and hydrophobics constraints. We also placed a weight of 0.1 on globularity constraints that are applied to each terminal symbol. We applied length constraints such that the full complex contained 720 residues (for example, the hexamer would consist of length-120 protomers). We ran simulated annealing over 30,000 iterations with for 10 seeds for each of oligomerization levels (for a total of 40 optimization trajectories). A high-level programming language for generative protein design

我们还类似地设计了更大的同源寡聚体，但移除了顶级节点的单链约束。我们设计了具有球状对称约束的三聚体、四聚体、六聚体和八聚体同源寡聚体。我们将对称约束的权重设为1，同时将pTM、pLDDT和疏水性约束的权重也设为1。我们还在每个终端符号上设置了0.1的球状性约束。我们设置了长度约束，使得完整复合物包含720个残基(例如，六聚体将由长度为120的原聚体组成)。我们对每个寡聚化水平进行了30,000次模拟退火迭代，使用 进行10次种子运行(总共40次优化轨迹)。一种用于生成蛋白质设计的高级编程语言

# A.3.6. TWO-LEVEL SYMMETRY DESIGN

# A.3.6. 两重对称性设计

We program two levels of symmetry using the productions

我们使用以下产生式编程了两重对称性

In these programs, we place the single-chain constraint on , so the final designs are protein homo-oligomers. We place a globularity symmetry constraint on ; to control the top-level symmetry, we repeat according to the desired oligomerization. We place a rotational symmetry constraint on ; to control the bottom-level symmetry, we repeat according to desired fold symmetry. We also place pTM, pLDDT, and hydrophobics constraints on the full protein; we place globularity constraints on . We compile constraints into an energy function with weights of 1 on all terms.

在这些程序中，我们对 施加了单链约束，因此最终设计是蛋白质同源寡聚体。我们对 施加了球状对称约束；为了控制顶层对称性，我们根据所需的寡聚化重复 。我们对 施加了旋转对称约束；为了控制底层对称性，我们根据所需的折叠对称性重复 。我们还在整个蛋白质上施加了pTM、pLDDT和疏水性约束；我们对 施加了球状性约束。我们将所有约束编译成一个能量函数，所有项的权重均为1。

We enumerated programs over the grid varying both the top and bottom levels of symmetry from 2 to 4 . We constrained lengths to 200 residues in total for the dimer of 2-fold; length-250 for the dimer of 3 -fold; length-400 for the dimer of 4-fold, the trimer of 2-fold, the trimer of 3-fold, and the tetramer of 2-fold; length-450 for the trimer of 4-fold and the tetramer of 3-fold; and length-500 for the tetramer of 4-fold. We ran simulated annealing over 30,000 iterations with for 10 seeds for each of these programs (for a total of 90 optimization trajectories).

我们在网格上枚举了程序，将顶层和底层的对称性从2到4进行变化。我们将二聚体2倍对称的总长度限制为200个残基；二聚体3倍对称的长度限制为250；二聚体4倍对称、三聚体2倍对称、三聚体3倍对称和四聚体2倍对称的长度限制为400；三聚体4倍对称和四聚体3倍对称的长度限制为450；四聚体4倍对称的长度限制为500。我们对每个程序运行了30,000次模拟退火，使用 对10个种子进行优化(总共90条优化轨迹)。

# A.3.7. Structural NOVELTY

# A.3.7. 结构新颖性

We quantify a given design for structural novelty by running an exhaustive search over the PDB version 2022-08 (http://www.rcsb.org/) (39) to find the experimental structure with the highest TM-score to the designed structure, normalizing by the designed structure length, using TM-align version 20210107 (40).

我们通过运行PDB版本2022-08(http://www.rcsb.org/)(39)的详尽搜索来量化给定设计的结构新颖性，以找到与设计结构具有最高TM分数的实验结构，并使用TM-align版本20210107(40)根据设计结构长度进行归一化。

# A.3.8. Inverse folding roundtrip experiments

# A.3.8. 逆折叠往返实验

We assessed the "designability" of a structure prediction produced by our optimization procedure by "roundtripping" the protein through an inverse folding model. More specifically, given a predicted structure from our optimization loops, we first use ESM-IF1 (21), an independently trained inverse folding model, to sample 10 sequences with temperature 0.1 from the backbone coordinates. We then run these sequences through ESMFold and compute the cRMSD between the starting and the roundtripped backbone atoms of the predicted structure.

我们通过将蛋白质“往返”通过逆折叠模型来评估由我们的优化程序产生的结构预测的“可设计性”。更具体地说，给定来自我们优化循环的预测结构，我们首先使用独立训练的逆折叠模型ESM-IF1(21)，从骨架坐标中以温度0.1采样10个序列。然后，我们通过ESMFold运行这些序列，并计算预测结构的起始骨架原子与往返骨架原子之间的cRMSD。

We performed this roundtrip experiment for 1,000 predicted structures that were obtained by first uniformly sampling one of the 180 symmetric single-chain optimization trajectories and then uniformly sampling one of the intermediate structure predictions within a given design loop (i.e., we do not restrict this analysis to the best pLDDT structure over a design loop, which are highly biased toward high pLDDTs). We report the relationship between ESM-IF1 perplexities of all sample structures and the corresponding cRMSD values. We also report the relationship between the pLDDT of the starting structure and the minimum RMSD over the structure for the 10 inverse-folded sequences. We also repeated the same experiment for 1,000 predicted structures that were obtained by first uniformly sampling over the 90 two-level symmetry optimization trajectories and then uniformly sampling one of the intermediate structure predictions within a given design loop. We also report the same metrics as in the single-chain evaluation.

我们对1,000个预测结构进行了往返实验，这些结构是通过首先均匀采样180个对称单链优化轨迹中的一个，然后在给定设计循环中均匀采样一个中间结构预测获得的(即，我们不将此分析限制在设计循环中具有最佳pLDDT的结构，这些结构高度偏向于高pLDDT)。我们报告了所有样本结构的ESM-IF1困惑度与相应cRMSD值之间的关系。我们还报告了起始结构的pLDDT与10个逆折叠序列的结构最小RMSD之间的关系。我们还对1,000个预测结构重复了相同的实验，这些结构是通过首先均匀采样90个两级对称优化轨迹中的一个，然后在给定设计循环中均匀采样一个中间结构预测获得的。我们还报告了与单链评估相同的指标。

# A.3.9. Symmetric functional site scaffolding

# A.3.9. 对称功能位点支架

We designed proteins that symmetrically scaffold multiple functional sites by using the tree described for the single-site functional scaffold but replicating it according to the desired fold symmetry and adding a rotational symmetry constraint to the top-level node; an example program for a 3- fold functional site scaffold can be found in Figure 5A. We use weights of 10 on the cRMSD and dRMSD constraints and weights of 1 on the pTM, pLDDT, rotational symmetry, binding site surface exposure, and hydrophobics constraints. We ran simulated annealing over 30,000 iterations with a starting temperature of 1 over 20 seeds for the design of 3- fold scaffolds of the IL10 and ACE2 binding sites described above, as well as 20 seeds for the design of 5 -fold scaffolds of the ACE2 binding site (for a total of 60 optimization loops).

我们设计了对称支架多个功能位点的蛋白质，使用单点功能支架描述的树，但根据所需的折叠对称性进行复制，并在顶层节点添加旋转对称约束；3倍功能位点支架的示例程序可以在图5A中找到。我们在cRMSD和dRMSD约束上使用权重10，在pTM、pLDDT、旋转对称性、结合位点表面暴露和疏水性约束上使用权重1。我们对上述IL10和ACE2结合位点的3倍支架设计运行了30,000次模拟退火，起始温度为1，共20个种子，以及ACE2结合位点的5倍支架设计，共20个种子(总共60个优化循环)。

# A.3.10. HIERARCHICAL ASYMMETRIC SYMMETRY DESIGN

# A.3.10. 层次化不对称对称性设计

We increased the level of hierarchical complexity in our programs by designing with three levels of constraints. The top level specifies two asymmetric subunits. Each asymmetric subunit itself has two-level symmetry (similar to the setting described above): we specifically consider the dimer of 2-fold (2x2), the dimer of 3-fold (2x3), and the trimer of 2-fold (3x2). We write programs consisting of (1) two asymmetric complexed together,(2) a and a complexed together, and (3) a and a completed together. An example program for the two asymmetric is provided in Figure 5C. We use weights of 1 on all constraints (pTM, pLDDT, hydrophobics, rotational/globular symmetry, and globularity). We ran simulated annealing over 30,000 iterations with over 10 seeds for each of the three programs described above (for a total of 30 optimization loops).

我们通过设计三个层次的约束来增加程序的层次复杂性。顶层指定了两个不对称的亚单元。每个不对称亚单元本身具有两级对称性(类似于上述设置):我们特别考虑了二聚体的2倍(2x2)、二聚体的3倍(2x3)和三聚体的2倍(3x2)。我们编写的程序包括(1)两个不对称的 复合在一起，(2)一个 和一个 复合在一起，以及(3)一个 和一个 复合在一起。图5C中提供了一个关于两个不对称 的示例程序。我们在所有约束(pTM、pLDDT、疏水性、旋转/全局对称性和全局性)上使用权重1。我们对上述三个程序中的每一个进行了30,000次模拟退火迭代，使用 在10个种子下运行(总共30个优化循环)。

A high-level programming language for generative protein design

用于生成蛋白质设计的高级编程语言

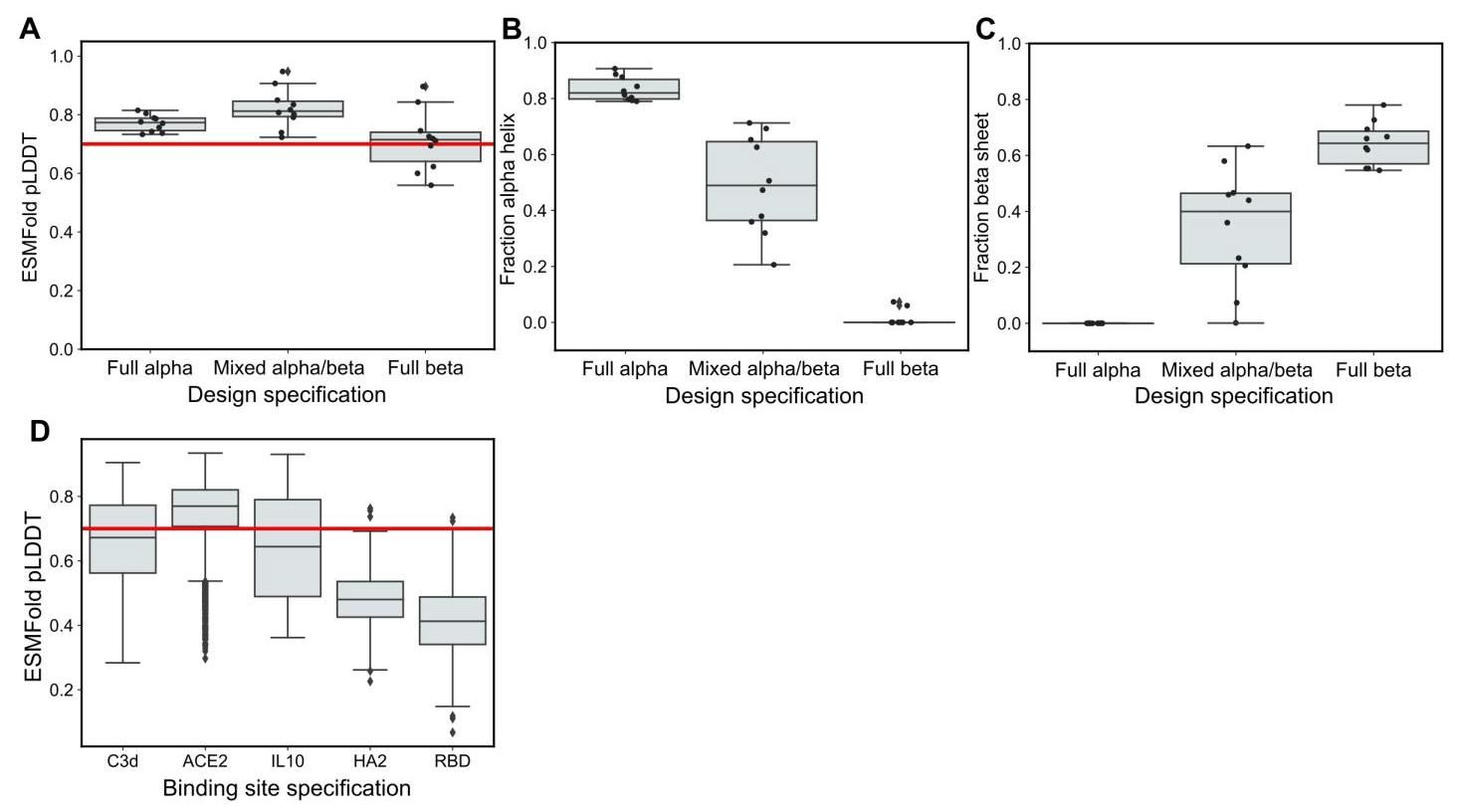


Figure S1. Additional plots for secondary structure design and functional site scaffolding. (A) ESMFold pLDDT values for different secondary structure design specifications (10 seeds per specification). A red line is plotted at pLDDT = 0.7. (B) The fraction of residues that are part of alpha helices for different secondary structure design specifications (10 seeds per specification). (C) The fraction of residues that are part of beta sheets for different secondary structure design specifications (10 seeds per specification). (D) ESMFold pLDDT values for different functional site scaffolding design runs (1,000 seeds per binding site).

图S1。二级结构设计和功能位点支架的附加图。(A)不同二级结构设计规范的ESMFold pLDDT值(每个规范10个种子)。在pLDDT = 0.7处绘制了一条红线。(B)不同二级结构设计规范中α螺旋部分残基的比例(每个规范10个种子)。(C)不同二级结构设计规范中β折叠部分残基的比例(每个规范10个种子)。(D)不同功能位点支架设计运行的ESMFold pLDDT值(每个结合位点1,000个种子)。

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.21.521526; this version posted December 22, 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-NC-ND 4.0 International license.

bioRxiv预印本doi: https://doi.org/10.1101/2022.12.21.521526；此版本发布于2022年12月22日。本预印本的版权持有者(未经同行评审认证)是作者/资助者，他们已授予bioRxiv永久展示预印本的许可。本预印本根据CC-BY-NC-ND 4.0国际许可提供。

A high-level programming language for generative protein design

用于生成蛋白质设计的高级编程语言

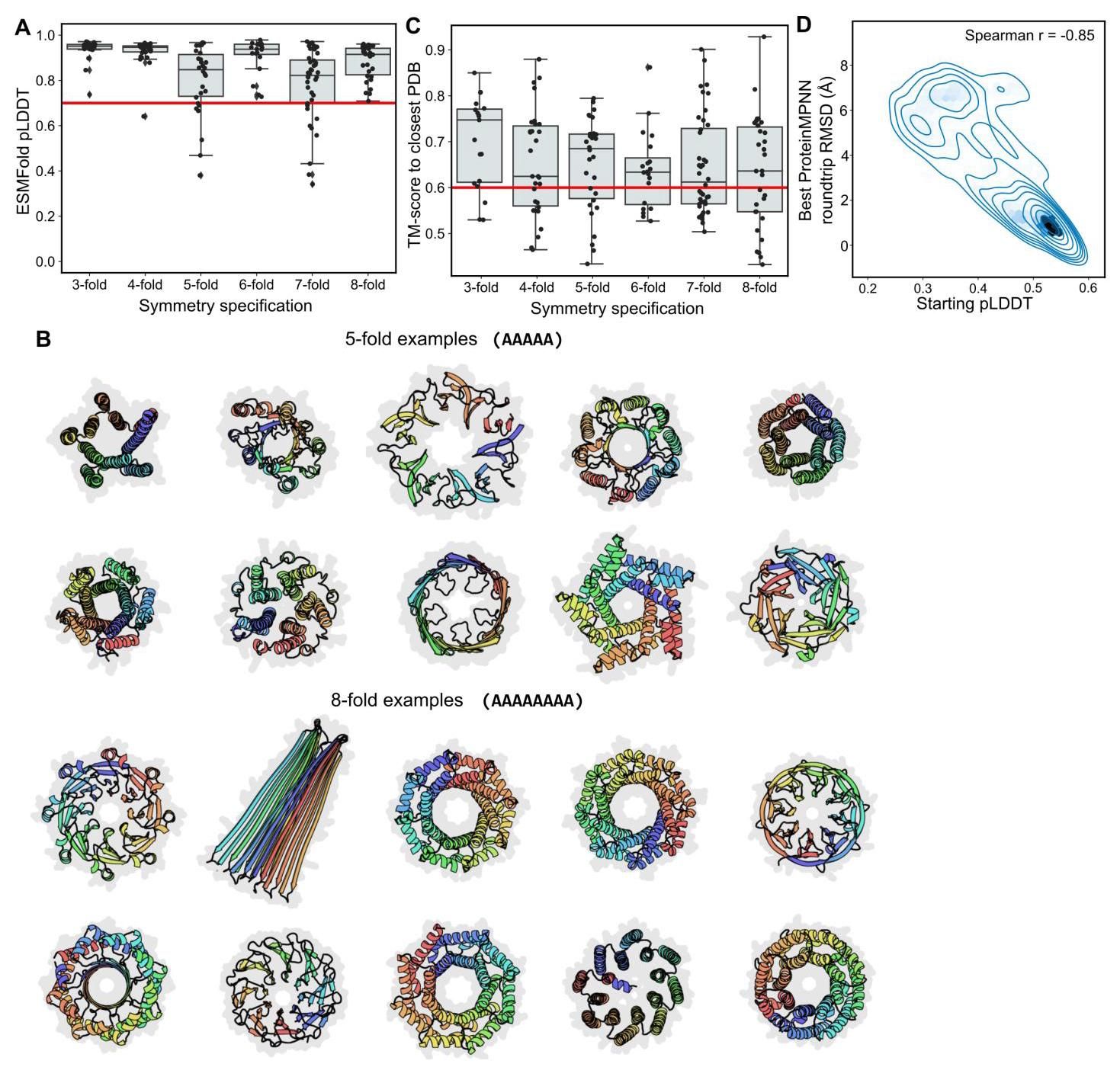


Figure S2. Additional plots for the design of symmetric single chains. (A) ESMFold pLDDT values for different fold symmetry design specifications (30 seeds per specification). A red line is plotted at pLDDT = 0.7 . (B) Ten randomly sampled designs for the design of 5- and 8-fold symmetry. (C) TM-scores for different fold symmetry design specifications (30 seeds per specification); the TM-score is between the best design and the closest structure in the PDB. A red line is plotted at TM-score . (D) Samples obtained by inverse folding with ProteinMPNN. On the -axis is the pLDDT of the designed structure prior to the roundtrip and on the -axis is the roundtrip RMSD.

图S2。对称单链设计的附加图。(A)不同折叠对称设计规范的ESMFold pLDDT值(每个规范30个种子)。在pLDDT = 0.7处绘制了一条红线。(B)5倍和8倍对称设计的十个随机采样设计。(C)不同折叠对称设计规范的TM分数(每个规范30个种子)；TM分数是最佳设计与PDB中最接近结构之间的分数。在TM分数 处绘制了一条红线。(D)通过ProteinMPNN逆向折叠获得的样本。 轴上是往返前设计结构的pLDDT， 轴上是往返RMSD。

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.21.521526; this version posted December 22, 2022. The copyright holder for this preprint

bioRxiv预印本doi: https://doi.org/10.1101/2022.12.21.521526；此版本发布于2022年12月22日。本预印本的版权持有者

A high-level programming language for generative protein design bioRxiv preprint doi: https://doi.org/10.1101/2022.12.21.521526; this version posted December 22, 2022. The copyright holder for this preprint

用于生成蛋白质设计的高级编程语言 bioRxiv预印本doi: https://doi.org/10.1101/2022.12.21.521526；此版本发布于2022年12月22日。本预印本的版权持有者

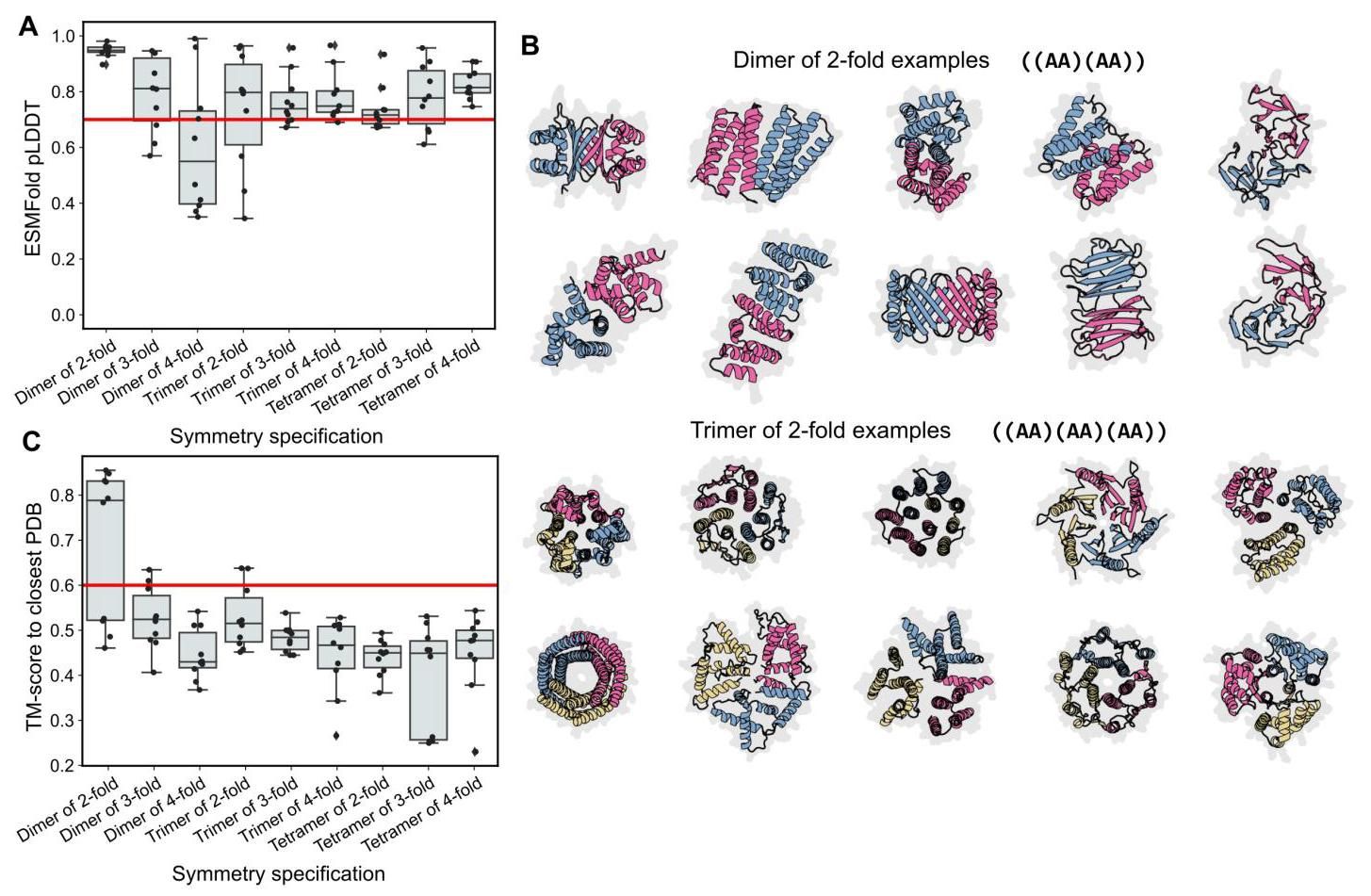


Figure S3. Additional plots for the design of two-level symmetry. (A) ESMFold pLDDT values for different two-level symmetry design specifications (10 seeds per specification). A red line is plotted at pLDDT . (B) All ten of the designs of a dimer of 2-fold and of a trimer of 2-fold symmetry. (C) TM-scores for different two-level symmetry design specifications (10 seeds per specification); the TM-score is between the best design and the closest structure in the PDB. A red line is plotted at TM-score .

图S3。两级对称设计的附加图。(A)不同两级对称设计规范的ESMFold pLDDT值(每个规范10个种子)。在pLDDT 处绘制了一条红线。(B)2倍二聚体和2倍三聚体对称性的所有十个设计。(C)不同两级对称设计规范的TM分数(每个规范10个种子)；TM分数是最佳设计与PDB中最接近结构之间的分数。在TM分数 处绘制了一条红线。

A high-level programming language for generative protein design

用于生成蛋白质设计的高级编程语言

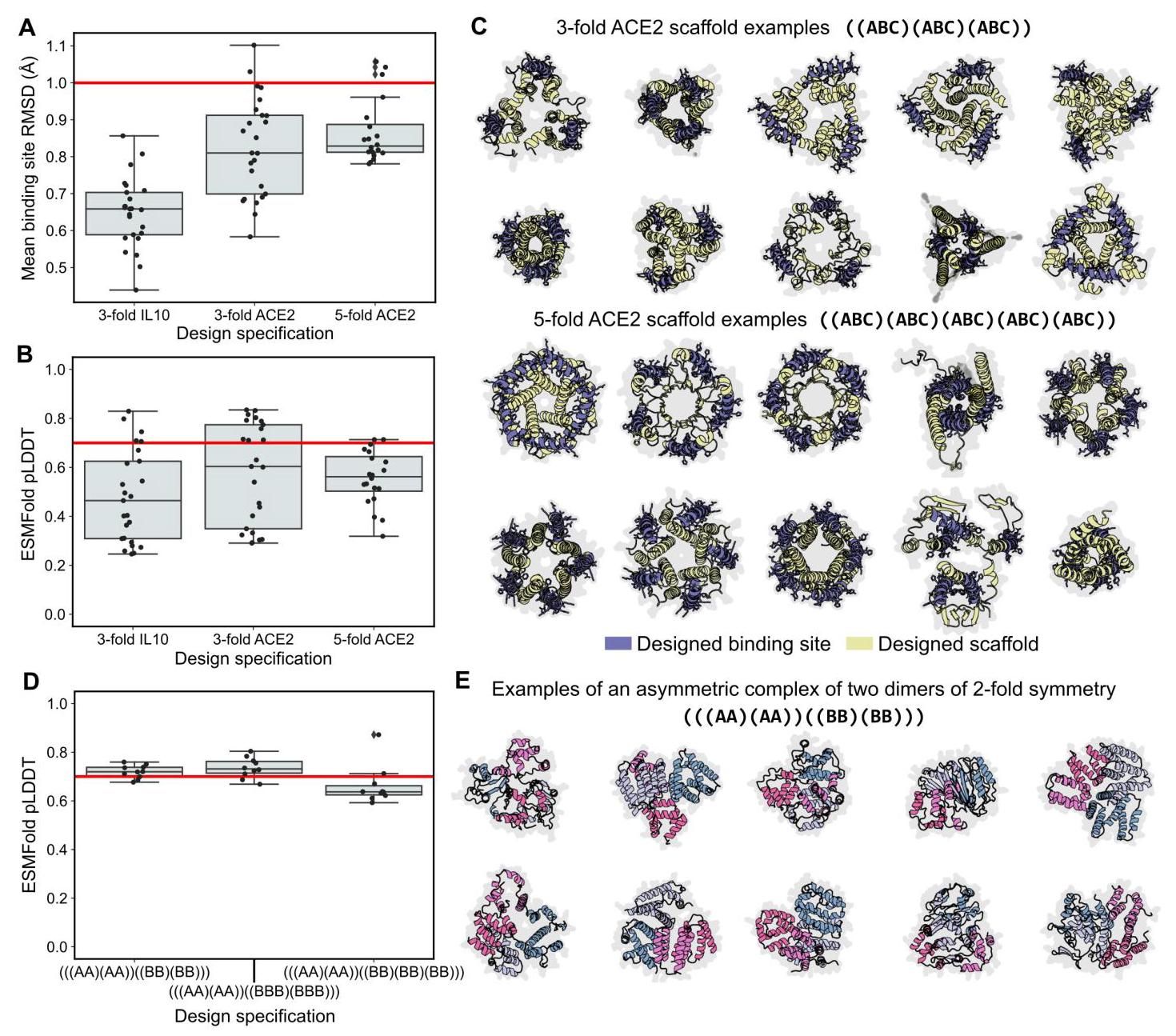


Figure S4. Additional plots for multi-level hierarchical design (A) ESMFold pLDDT values for different binding site scaffolds with fold symmetry (20 seeds per specification). A red line is plotted at pLDDT = 0.7. (B) RMSD values for different binding site scaffolds with fold symmetry (20 seeds per specification). The mean RMSD is reported across either three or five binding sites. A red line is plotted at pLDDT . (C) Ten randomly sampled designs for the design of 3- and 5-fold symmetric ACE2 binding site scaffolds. (D) ESMFold pLDDT values for different asymmetric-symmetric design specifications (10 seeds per specification). A red line is plotted at pLDDT 0.7. (E) All ten of the designs of an asymmetric complex of two dimers of 2-fold symmetry.

图S4. 多层次分层设计的附加图表 (A) 具有折叠对称性的不同结合位点支架的ESMFold pLDDT值(每个规格20个种子)。在pLDDT = 0.7处绘制了一条红线。(B) 具有折叠对称性的不同结合位点支架的RMSD值(每个规格20个种子)。报告了三个或五个结合位点的平均RMSD。在pLDDT 处绘制了一条红线。(C) 随机抽样的十个设计，用于设计3倍和5倍对称的ACE2结合位点支架。(D) 不同非对称-对称设计规格的ESMFold pLDDT值(每个规格10个种子)。在pLDDT 0.7处绘制了一条红线。(E) 所有十个设计，用于设计具有2倍对称性的两个二聚体的非对称复合物。

1. Equal contribution Meta Fundamental AI Research Protein Team (FAIR). Stanford University. Work performed as a visiting researcher at Meta AI. New York University. Work performed as a visiting researcher at Meta AI. New York University. Correspondence to <arives@meta.com>.

   同等贡献 Meta基础人工智能研究蛋白质团队(FAIR)。 斯坦福大学。作为Meta AI的访问研究员完成的工作。 纽约大学。作为Meta AI的访问研究员完成的工作。 纽约大学。 通讯地址:<arives@meta.com>。 [↑](#footnote-ref-26)