

# Text2Protein: A Generative Model for Designated Protein Design on Given Description

Siyang Zhang<sup>1</sup>, Ramtin Hosseini<sup>2</sup>, and Pengtao Xie<sup>2, \*</sup>

<sup>1</sup>Brown University, Department of Computer Science, RI, 02912, United States

<sup>2</sup>University of California San Diego, Department of Electrical and Computer Engineering, CA, 92093, United States

\*Corresponding to: p1xie@ucsd.edu

## ABSTRACT

Designing protein structures from text is challenging in computational biology. We propose Text2Protein, a pipeline combining large language models (LLMs) with diffusion models to generate full-atomic protein structures from text. Using a conditional diffusion model and the Vicuna-7B language model, we learn data distributions of 6D interresidue coordinates, refined into full-atomic structures with PyRosetta. Trained on a curated RCSB-PDB dataset, Text2Protein focuses on single-chain proteins with 40-256 residues. Our extensive experiments validate Text2Protein’s effectiveness by generating high-fidelity protein structures similar to ground truth proteins using raw texts. We evaluate Text2Protein using multiple metrics, including Mean Square Error (MSE) of 6D coordinates, Rosetta Energy Units (REU), and TM-score. Our results show that 5% of the generated proteins have a TM-score greater than 0.5, indicating similar folds in SCOP/CATH. Additionally, 16% of pairs have a TM-score greater than 0.4, 89% have a TM-score greater than 0.3, and none have a TM-score less than 0.17, below the threshold for unrelated proteins. Text2Protein presents a promising framework for automated protein design, potentially accelerating novel protein discovery. This work opens new avenues for integrating natural language understanding with protein structure generation, with implications in drug discovery, enzyme engineering, and material science.

## Introduction

Protein structure design is a fundamental challenge in computational biology and biochemistry, with far-reaching implications for drug discovery<sup>1</sup>, enzyme engineering<sup>2</sup>, and understanding biological processes<sup>3</sup>. Unlike many computer vision tasks where human intuition can guide evaluation, protein-related tasks often rely on complex, non-visual properties<sup>4</sup>. This necessitates the use of sophisticated metrics and analysis techniques to assess the quality and functionality of designed proteins<sup>5</sup>. The importance of protein structure design cannot be overstated. Proteins are fundamental to numerous biological processes, and their structures determine their functions<sup>6</sup>. Designing proteins with specific functionalities has vast implications in drug discovery, enzyme engineering, and synthetic biology<sup>7</sup>. However, the traditional approach to protein design involves iterative cycles of computational prediction and experimental validation, a process that is both time-consuming and resource-intensive<sup>8</sup>. Moreover, the vast conformational space of proteins and the complex relationship between sequence, structure, and function make this task particularly challenging<sup>9</sup>.

Existing computational methods often struggle to capture the intricate dependencies between different parts of a protein or to generate novel structures that deviate significantly from known templates<sup>4</sup>. These methods, while effective to some extent, are typically slow and require significant human intervention<sup>10</sup>. Recent works employing generative models, such as ProteinSGM<sup>11</sup> and ProteinDT<sup>12</sup>, have made progress but still face limitations. ProteinSGM uses score-based generative modeling for de novo protein design but lacks textual description guidance<sup>11</sup>. ProteinDT applies diffusion models for protein sequence generation but does not extend to full-atomic structure design<sup>12</sup>. In recent years, deep learning techniques have demonstrated remarkable efficiency and effectiveness across various domains, showcasing an ability to process complex data, uncover hidden distributions, and solve challenging problems. The success of these methods in fields such as computer vision<sup>13</sup> and natural language processing<sup>14</sup> has sparked interest in their application to protein design. Diffusion models, a class of generative models, have shown particular promise in various synthesis tasks, including image, video, and 3D object generation<sup>15</sup>. These models offer several advantages over other generative approaches, such as Generative Adversarial Networks (GANs)<sup>16</sup>, including improved stability, robustness, and generation quality<sup>17</sup>. The training process of diffusion models involves gradually adding noise to a clear data distribution until it becomes isotropic pure noise, then learning the reverse process of denoising to recover the original distribution<sup>18</sup>. Simultaneously, transformer architectures with their attention mechanisms have revolutionized natural language processing and are increasingly being applied to non-NLP tasks with impressive results<sup>14</sup>. Large Language Models (LLMs), which are transformers with an enormous number of parameters, have emerged as powerful general-purpose pre-trained models for a wide range of language-related tasks<sup>19</sup>.

Despite these advancements, several challenges remain in applying these techniques to protein design. These include bridging the gap between textual descriptions of protein function and three-dimensional structural representations<sup>20</sup>, generating protein structures that are not only physically plausible but also likely to exhibit the desired functionality<sup>21</sup>, handling the high-dimensional nature of protein structures while maintaining computational efficiency<sup>4</sup>, ensuring that generated proteins are diverse and novel rather than simple variations of known structures<sup>9</sup>, and addressing the lack of real biological experimental evaluation, which limits the practical applicability of computational models<sup>8</sup>. To address these challenges, we propose Text2Protein, a novel pipeline that combines the strengths of LLMs and diffusion models for protein design. Our approach takes as input a raw textual description of the desired protein functionality and generates a corresponding full-atomic structure protein. The key components of our solution include the utilization of a pre-trained LLM (Vicuna-7B) to encode textual descriptions into rich, contextual embeddings<sup>22</sup>, representation of protein structures as 6D interresidue tensors<sup>20</sup> capturing both local and global structural information<sup>20</sup>, a conditional diffusion model<sup>17</sup> that learns to generate these 6D coordinates guided by the textual embeddings, and integration with PyRosetta<sup>23</sup> for refining the generated structures into full-atomic models through minimization, FastDesign, and FastRelax protocols. By leveraging the power of diffusion models to capture complex structural distributions and the ability of LLMs to understand and represent functional descriptions, Text2Protein aims to bridge the gap between protein function and structure in a novel and powerful way. This approach has the potential to accelerate the protein design process significantly, enabling the exploration of a wider range of functional proteins and opening new avenues for computational biology research. While challenges remain, particularly in experimental validation, Text2Protein represents a significant step forward in the field of automated protein design.

Our main contributions in this work are as follows:

- To the best of our knowledge, Text2Protein represents one of the first few works that integrate large language models (LLMs) with conditional diffusion models to generate full-atomic protein structures from raw text descriptions. This approach bridges the gap between textual descriptions and three-dimensional protein structures, potentially accelerating the protein design process.
- Our method introduces a novel application of conditional diffusion models guided by textual embeddings from a pretrained large language model (Vicuna-7B). This allows the model to capture complex structural distributions and generate protein structures that are both physically plausible and functionally relevant.
- We demonstrate the effectiveness of representing protein structures as 6D interresidue tensors. This representation captures both local and global structural information, enabling more accurate generative modeling of protein structures.
- We showcase the integration of deep learning models with traditional computational biology tools, specifically PyRosetta, for refining the generated protein structures. This integration ensures that the final protein models are realistic and adhere to the physical constraints of protein structures.
- Comprehensive evaluations of our method are provided using various metrics, including Mean Square Error (MSE) of 6D coordinates, Rosetta Energy Units (REU), and TM-score. These evaluations demonstrate the potential of our approach in generating high-fidelity protein structures with significant structural similarities to ground truth proteins.
- Thanks to the synergy between LLMs and diffusion models, Text2Protein not only generates high-quality protein structures but also offers a scalable and efficient solution for automated protein design. Our approach opens new avenues for computational biology research and has significant implications for drug discovery, enzyme engineering, and synthetic biology.

## Related Works

### Traditional Computational Methods for Protein Design

Protein structure design has long relied on traditional computational methods, including physics-based force fields and Monte Carlo sampling<sup>24</sup>. These approaches encompass homology modeling, molecular dynamics simulations, and *ab initio* methods. A pioneering software suite in this field, Rosetta<sup>23</sup>, employs a combination of physical and knowledge-based energy functions to predict and design protein structures. While powerful, these traditional methods often face significant challenges. They typically require substantial computational resources and struggle to efficiently navigate the vast conformational space of proteins<sup>24</sup>. Homology modeling, for instance, is heavily dependent on the availability of similar protein templates<sup>25</sup>. Molecular dynamics simulations, though insightful, are computationally expensive and time-consuming<sup>26</sup>. *Ab initio* methods attempt to predict protein structures from first principles but are often limited in accuracy due to the complexity of protein folding<sup>9</sup>. Despite these limitations, traditional methods have laid the groundwork for our understanding of protein structures and continue to play a crucial role in the field, particularly in hybrid approaches that combine them with modern machine learning techniques<sup>27</sup>.

## Machine Learning in Protein Structure Prediction

The application of machine learning techniques to protein structure prediction has led to remarkable breakthroughs in recent years<sup>4</sup>. Notable examples include AlphaFold<sup>4</sup> and RoseTTAFold<sup>5</sup>, which have demonstrated unprecedented accuracy in predicting protein structures from amino acid sequences. These methods leverage sophisticated deep learning architectures and attention mechanisms to capture long-range interactions in protein sequences<sup>4,5</sup>. The success of these models has significantly advanced our ability to predict protein structures, opening new avenues for understanding protein function and design<sup>28</sup>. However, while these methods excel at structure prediction, they are not inherently designed for the task of generating novel protein structures based on desired functionalities<sup>29</sup>. Recent efforts have focused on combining these deep learning models with traditional computational biology tools to address this limitation<sup>30</sup>. For example, PyRosetta has been integrated with deep learning models to refine generated structures, ensuring that the generated proteins are not only plausible but also adhere to physical constraints<sup>4,23</sup>. This hybrid approach aims to leverage the strengths of both machine learning and traditional methods to improve the accuracy and reliability of protein design<sup>29</sup>.

## Generative Models for Protein Design

The field of protein design has seen a growing interest in the application of generative models for de novo protein creation<sup>31</sup>. Various approaches have been explored, including Variational Autoencoders (VAEs)<sup>32</sup>, Generative Adversarial Networks (GANs)<sup>16</sup>, autoregressive models<sup>14</sup>, and diffusion models<sup>17</sup>. VAEs have been applied to generate novel protein sequences by encoding proteins into a latent space and sampling from this space<sup>33</sup>. Works like ProTeinGAN<sup>31</sup> have combined VAEs with GANs to generate protein sequences with specific properties. GANs, such as ProteinGAN<sup>31</sup>, have demonstrated the ability to generate functional protein sequences, although they often face challenges with training stability and mode collapse<sup>16</sup>. Autoregressive models based on transformers, like ProtGPT2<sup>34</sup>, have shown promise in generating diverse and functional protein sequences. More recently, diffusion models have gained attention for their stability and robustness<sup>17</sup>. ProteinSGM<sup>11</sup>, for instance, applied score-based generative modeling to protein backbone design, demonstrating the ability to generate diverse and realistic protein structures. However, most of these approaches do not incorporate textual descriptions of protein function, limiting their ability to generate proteins with specific desired characteristics<sup>11</sup>. Additionally, ensuring the physical plausibility and functional relevance of generated proteins while maintaining computational efficiency remains a critical challenge in this field<sup>35</sup>.

## Text-Guided Protein Design and Representation Learning

The integration of natural language processing with protein design represents an emerging and promising area of research<sup>36</sup>. Large language models have been applied to various protein-related tasks, such as ProteinBERT<sup>37</sup>, which adapted the BERT architecture for protein sequence analysis, enabling tasks like function prediction and structure classification. Recent work has begun to explore the use of text descriptions to guide protein sequence generation<sup>36</sup>. For example, ProteinDT<sup>12</sup> proposed a text-guided protein sequence design framework using diffusion models. While this approach bridges the gap between text and protein sequences, it does not extend to full atomic structure generation<sup>12</sup>. The challenge remains to effectively translate textual descriptions of protein functionality into three-dimensional protein structures<sup>36</sup>.

Effective representation of protein structures is crucial for the success of machine learning models in protein design<sup>20</sup>. Various approaches have been explored, including graph-based representations<sup>38</sup> and geometric representations<sup>39</sup>. Graph Neural Networks (GNNs) have been applied to represent protein structures, capturing the complex interactions between amino acids<sup>40</sup>. For instance, GraphQA<sup>38</sup> used GNNs for protein model quality assessment. Geometric approaches, such as the use of 6D interresidue tensors<sup>20</sup>, have proven effective in capturing both local and global structural information, enabling more accurate structure prediction. These advanced representation techniques play a crucial role in bridging the gap between sequence information and three-dimensional structure<sup>39</sup>.

Despite these advancements, several challenges remain in the field of computational protein design<sup>4</sup>. These include bridging the gap between sequence generation and structure prediction<sup>4</sup>, incorporating functional information explicitly<sup>29</sup>, improving scalability for larger proteins and high-throughput design tasks<sup>30</sup>, addressing the lack of experimental validation<sup>28</sup>, and ensuring the physical plausibility and functional relevance of generated proteins<sup>4</sup>. Our work, Text2Protein, aims to address several of these challenges by integrating text-guided generation with diffusion models for full atomic structure design<sup>12</sup>. By leveraging the power of large language models and advanced protein structure representations, we provide a novel approach to protein design that bridges the gap between functional descriptions and three-dimensional structures, potentially accelerating the protein design process and exploring a wider range of functional proteins<sup>39</sup>.

## Methods

We introduce a diffusion model-based pipeline designed to generate protein structures from textual descriptions called Text2Protein, which is illustrated in Figure 1. This section outlines our methodology, including data preparation, protein

representation, model architecture, training process, sampling, and structure refinement.

## Dataset Curation and Preprocessing

Our study utilized a carefully curated subset of the RCSB Protein Data Bank (PDB)<sup>41</sup>. We focused on single-chain proteins with lengths ranging from 40 to 256 residues, balancing computational feasibility with the desire to capture a diverse set of protein structures<sup>41</sup>. Multi-chain proteins were excluded to simplify the modeling process and focus on individual protein folding<sup>41</sup>. The final dataset comprised 10,898 protein structures, split into training and test sets with a 95:5 ratio, ensuring a substantial training set while maintaining a sufficient number of test cases for evaluation<sup>41</sup>. For each protein, we extracted textual descriptions from the PDB entries, focusing on functional annotations and structural characteristics<sup>41</sup>. These descriptions were preprocessed by truncating or padding to a fixed length of 512 words, ensuring consistent input size for our language model<sup>41</sup>. Standard NLP preprocessing techniques, including lowercasing, removing special characters, and tokenization, were applied to enhance the quality of the textual input<sup>42</sup>.

## Protein Representation: Interresidue 6D Coordinates

A protein consists of chain(s) of amino acids, with specific spatial folding. The amino acids in a polypeptide chain are linked by peptide bonds, with each individual amino acid in the chain referred to as a residue. The linked series of carbon, nitrogen, and oxygen atoms form the main chain or protein backbone<sup>3</sup>. Adjacent residues of proteins constrain the inter-residue internal coordinates, resulting in specific inter-residue distributions<sup>9</sup>. To capture these spatial relationships, we represent each protein structure using a 6D coordinate system<sup>20</sup>. This representation comprises five matrices: the  $C\beta - C\beta$  distance ( $d$ ) measuring the distance between  $C\beta$  atoms of residue pairs,  $\omega$  and  $\theta$  as torsional angles describing the rotation around chemical bonds,  $\phi$  as the planar angle capturing the bend between residues, and a padding matrix indicating the presence (1) or absence (0) of residues to handle variable protein lengths<sup>20</sup>. These matrices are combined into a tensor of shape [5, 256, 256], where 256 is the maximum number of residues we consider. This comprehensive representation captures both local and global structural information, crucial for accurate protein modeling<sup>20</sup>, while the padding matrix ensures all proteins are represented with the same dimensionality, regardless of their actual length<sup>20</sup>.

## Diffusion Model Architecture

Our diffusion model consists of two main components: a text encoder and a denoising U-Net. These components work in tandem to generate protein structures conditioned on textual descriptions<sup>17</sup>.

### Text Encoder

We employ a pretrained Vicuna-7B language model as our text encoder<sup>22</sup>. This large language model, based on the LLaMA architecture and fine-tuned on diverse internet conversations, has demonstrated strong performance in various natural language processing tasks, making it ideal for encoding complex protein descriptions<sup>22</sup>. The text encoder serves several crucial functions in our pipeline. It processes the raw textual descriptions of proteins, capturing the semantic meaning and functional characteristics described in the text<sup>22</sup>. The model generates rich, contextual embeddings of size 4096 for each input description, encapsulating nuanced information about the protein’s function, structure, and other relevant attributes mentioned in the text<sup>22</sup>. By leveraging its pretrained knowledge, the text encoder extracts relevant features from the descriptions that can guide the protein structure generation process<sup>22</sup>. These generated embeddings serve as conditioning information for the diffusion process, allowing the model to generate protein structures that align with the given textual descriptions<sup>17</sup>. We keep the weights of the Vicuna-7B model frozen during training, utilizing its pretrained knowledge without fine-tuning<sup>22</sup>. This approach allows us to leverage the model’s robust language understanding capabilities while focusing our training efforts on the protein generation task<sup>22</sup>.

### Denoising U-Net with Spatial Transformer

The core of our generative process is a modified U-Net architecture with spatial transformers, designed to learn the denoising of protein structures at various noise levels<sup>43</sup>. This architecture is crucial for handling the unique challenges of protein structure generation, bridging the gap between natural language descriptions and three-dimensional protein structures<sup>17</sup>. Our U-Net follows a U-shaped structure with a contracting path (encoder), a bottleneck, and an expanding path (decoder), allowing the network to capture both fine-grained details and broader structural information<sup>43</sup>. It consists of downsampling blocks that gradually reduce spatial dimensions while increasing channel depth, a mid block that processes the most compressed representation, and upsampling blocks that gradually increase spatial dimensions back to the original size<sup>43</sup>. This structure enables the model to capture hierarchical features and reconstruct the denoised structure effectively<sup>17</sup>.

A well-trained U-Net can sample different levels of noise at any timestep, as described in equation (5), iteratively predicting noise from more noisy distributions to less noisy ones until a sufficiently clear distribution is approximated<sup>18</sup>. Our approach draws parallels with multi-channel 2D pixel space structures widely used in image synthesis tasks<sup>17</sup>. We represent protein

backbone structures as interresidue 6D coordinates in a 5-channel 2D residue-space, where information in each channel is continuous and can be approximated by providing a mean and small standard deviation<sup>20</sup>. This representation justifies the use of a U-Net architecture for the denoising autoencoder<sup>43</sup>.

Each block in the U-Net contains multiple residual layers with customizable configurations, which help maintain gradient flow through the deep network and allow for better training of very deep architectures<sup>44</sup>. The architecture incorporates convolution layers to extract local features from the protein representations at various scales, and self-attention mechanisms to capture long-range dependencies within the protein structure, crucial for understanding global structural patterns<sup>14</sup>. Crucially, we integrate spatial transformers with multi-headed cross-attention, allowing the model to attend to different parts of the text embedding and effectively integrate the textual information with the structural data<sup>14</sup>. This mechanism establishes relationships between the protein structure and the textual description at various levels of abstraction, enabling the model to generate structures that align with given textual descriptions<sup>14</sup>.

Unlike traditional autoencoders, our U-Net implements a unique skip-connection mechanism. The U-Net records every intermediate latent representation generated by the corresponding downsampling block in an intermediate latent list<sup>43</sup>. This allows each upsampling block to take in the concatenation of the current latent representation of the input and its counterpart of the same shape previously stored in the latent list<sup>43</sup>. This mechanism preserves fine-grained details that might otherwise be lost in the encoding process<sup>43</sup>. The U-Net is also conditioned on the noise level at each step of the diffusion process, allowing it to adapt its denoising strategy based on the current noise magnitude<sup>18</sup>. It takes as input a noisy protein structure representation, the corresponding timestep, and the text embedding from the text encoder<sup>17</sup>. The model then predicts the noise added to the protein structure, effectively learning to denoise the structure conditionally based on the textual description<sup>18</sup>.

The combination of convolution layers, self-attention layers, and spatial transformers in each residual block enables the model to learn different levels of attention to the embedded textual context<sup>14</sup>. This multi-level attention mechanism results in a more comprehensive and context-aware reconstruction process, ultimately leading to better loss prediction and more accurate protein structure generation<sup>17</sup>. This sophisticated architecture enables our model to generate diverse and physically plausible protein structures that align with given textual descriptions, effectively bridging the gap between natural language and three-dimensional protein structures<sup>17</sup>.

## Diffusion Process and Training

**Forward Diffusion.** The forward diffusion process takes advantage of the Markov process to sequentially add noise to a clear input  $x_0$ , generating increasingly noisy inputs  $x_t$  as  $t$  increases:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \quad (1)$$

where  $x_t$  is the noisy input at timestep  $t$ ,  $\beta_t$  is a predefined noise schedule, and  $I$  is the identity matrix. By applying the reparameterization trick, we can express  $x_t$  in a closed form:

$$q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I) \quad (2)$$

where  $\alpha_t = 1 - \beta_t$ , and  $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ .

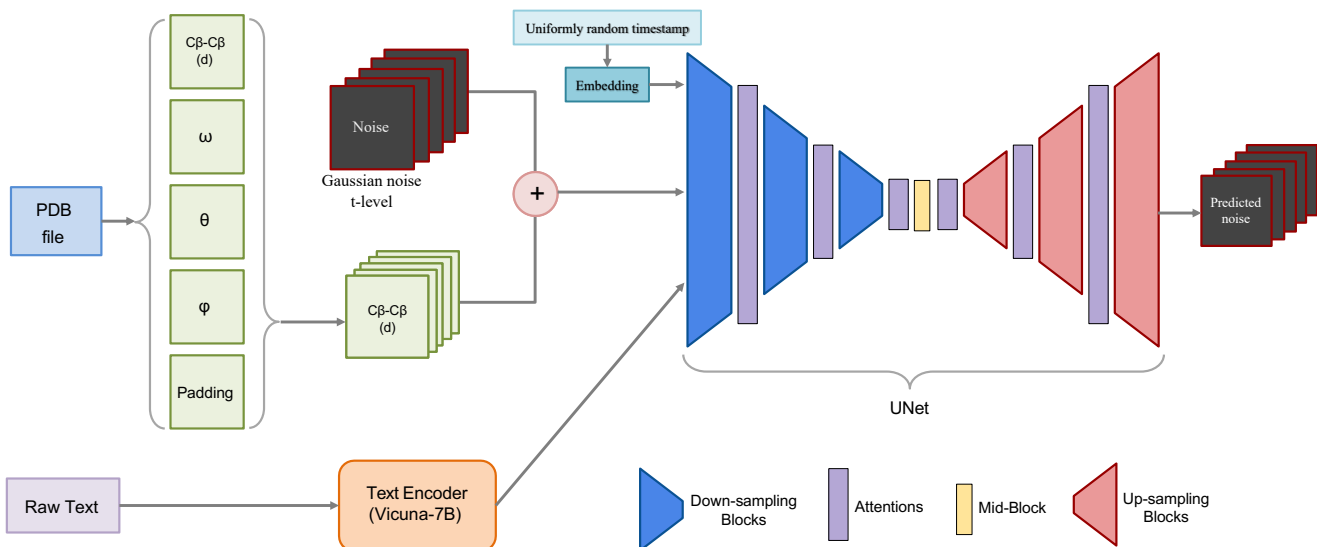
**Backward Denoising.** For the backward denoising process, we train a denoising autoencoder implemented as a U-Net structured model that learns the reverse process, sequentially removing noise from a noisy input until generating a clear one. The U-Net  $\epsilon_\theta(x_t, t, c)$  predicts the noise, where  $c$  is the textual embedding as context. The training objective is to minimize the  $l_2$  loss term:

$$l_2 = \mathbb{E}_{x_0, \epsilon, t, c} [\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2] \quad (3)$$

After training,  $\epsilon_\theta$  will be able to predict the probability distribution  $p_\theta(x_{t-1}|x_t)$  such that:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t, c), \beta_t I) \quad (4)$$

**Training Process.** The training process aims to teach the U-Net how to denoise a noisy distribution at any timestep. Each training step follows a specific sequence: we begin by randomly initializing a pure Gaussian noise and a timestep, then sample a  $t$ -level noisy distribution in one step using the parameterization trick. The raw text is encoded with the frozen text encoder, after which we pass the time embedding,  $t$ -level noisy input, and text embedding to the U-Net. Each Residual Block at different down- or up-sampling resolution stages in the U-Net contains a spatial transformer that applies a cross-attention mechanism from the noisy input to the text embedding. Finally, the U-Net generates a noise tensor that should approximate the initialized pure Gaussian noise.



**Figure 1.** Overview of the Text2Protein pipeline. The raw text is processed by a text encoder (Vicuna-7B) to produce textual embeddings. The protein data bank (PDB) file is converted into 6D interresidue coordinates, represented as a tensor of shape [5, 256, 256]. A random timestep  $t$  and Gaussian noise are initialized to generate a  $t$ -level noisy input. The U-Net, consisting of downsampling blocks, a mid block, and upsampling blocks, processes these inputs along with the textual embeddings to predict the noise. The objective is to minimize the difference between the added noise and the predicted noise, refining the generated protein structure.

## Sampling

During inference, we use a Predictor-Corrector sampling method, starting with pure Gaussian noise  $x_T$  and iteratively applying the denoising step 2000 times, gradually reducing noise levels. The sampling process can be expressed as:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \left( x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_{\theta}(x_t, t, c) \right) + \sqrt{\beta_t} z, \quad z \sim \mathcal{N}(0, I) \quad (5)$$

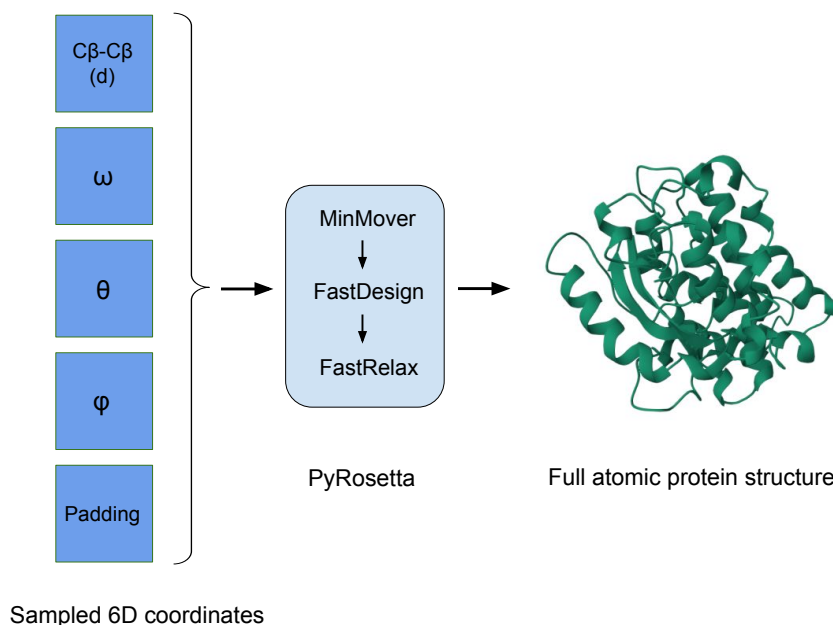
We use Predictor-Corrector sampling (Reverse diffusion Predictor with Langevin dynamics Corrector), which provides an initial prediction for a given denoising step using the estimated scores from the score network, and is further refined using Langevin dynamics.

## Rosetta Protocol

To convert the generated 6D coordinates into full-atomic protein structures, we employ PyRosetta 4.0 in a multi-step process<sup>23</sup>. We begin by initializing a protein sequence with all alanine residues and converting the normalized 6D coordinates to probability distributions<sup>23</sup>. Then, we apply MinMover for initial structure minimization over 5 rounds<sup>23</sup>, use FastDesign to sample sequences and rotamers fitting the backbone<sup>45</sup>, and finally perform FastRelax to generate the full-atomic structure<sup>46</sup>. In our Rosetta Design protocol, we set the angle standard deviation to 10.0 and the distance standard deviation to 2.0, a configuration that empirically allows reproducible generation of protein structures while still relaxing the constraints enough to produce realistic structures<sup>23</sup>. The MinMover provides a basic structural framework, FastDesign introduces sequence diversity and optimizes side-chain conformations<sup>45</sup>, and FastRelax performs a final optimization to ensure physical plausibility. The iteration number of MinMover/FastDesign/FastRelax can be customized as needed. We select the output with the lowest energy function as the final output, representing the most thermodynamically favorable conformation according to the Rosetta energy function. Figure 2 provides an overview of the Rosetta Protocol process.

## Experiments

We present Text2Protein, an innovative pipeline designed to generate protein structures based on raw text descriptions of their functionality. Our approach leverages a pretrained Large Language Model (LLM) as a text encoder to provide conditional guidance to a denoising diffusion model. This novel framework demonstrates the potential of conditional diffusion models for protein design and, more broadly, for large molecule generation using only explicit textual information. Given the complex



**Figure 2.** Rosetta Protocol overview. From interresidue 6D coordinates sampled by the diffusion model, we execute the Rosetta design process to generate full atomic structures of proteins.

nature of proteins and the challenges associated with analyzing their biological properties and functionalities, our experimental approach focuses on a qualitative analysis of structural similarities. We compare our text-guided designed proteins with ground truth proteins from the test set, which were not accessible during the training process.

## Experimental Setup and Implementation Details

We trained our Text2Protein model on a carefully curated subset of the RCSB Protein Data Bank (PDB) dataset<sup>41</sup>, which contains over 200,000 three-dimensional protein structures recorded in .pdb format. Our selection criteria focused on single-chain proteins with between 40 and 256 residues. This curation process resulted in a dataset of 10,898 PDB files with their corresponding textual descriptions. We split this dataset into training and test sets with a 95:5 ratio, ensuring a robust training set while maintaining a sufficient number of unseen proteins for evaluation. For text processing, we utilized the Vicuna-7B model with an embedding size of 4096<sup>22</sup>. We standardized the length of textual descriptions to 512 words, padding shorter descriptions and truncating longer ones as necessary<sup>42</sup>. Consequently, the batched raw text descriptions of proteins are represented as tensors of shape [B, 512, 4096], where B denotes the batch size<sup>42</sup>.

We employed the Adam optimizer for training the denoising autoencoder (U-Net) with a learning rate of  $1.0 \times 10^{-4}$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 1.0 \times 10^{-8}$ <sup>47</sup>. Our training procedure for each step involves uniformly sampling a timestep  $t$  from 0 to 2000<sup>48</sup>, generating  $t$ -level noisy 6D coordinates according to equation (2)<sup>18</sup>, providing the corresponding textual embedding, and forwarding the data through the U-Net<sup>43</sup>. The U-Net architecture down- and up-samples the input across 6 different latent dimensions, each scaling by a factor of  $\frac{1}{2}$ <sup>43</sup>. Each latent dimension contains 2 residual blocks, each comprising a ResNet block, a self-attention block, and a spatial transformer<sup>14,44</sup>. We trained the diffusion model for approximately 500,000 iterations with a batch size of 2<sup>48</sup>. The entire training process, conducted on an NVIDIA RTX 3090 GPU, took about 10 days to complete.

For our evaluation, we generated 544 6D coordinate samples<sup>18</sup>. The sampling process for each protein involves initializing random noise with the same shape as the training interresidue 6D coordinates<sup>48</sup>, providing textual embedding inferred by Vicuna-7B from the raw textual description of a protein PDB in the test set<sup>22</sup>, performing 2000 iterations of predictor-corrector denoising steps<sup>49</sup>, and generating a clear distribution of interresidue 6D coordinates<sup>18</sup>. We then processed these sampled 6D coordinates through a PyRosetta<sup>23</sup> design pipeline, which includes MinMover for initial minimization<sup>23</sup>, FastDesign for sequence and rotamer optimization<sup>45</sup>, and FastRelax for final structure refinement<sup>46</sup>. For our analysis, we randomly selected 100 designed full-atomic PDB files generated from descriptions of ground truth PDBs in the test set<sup>41</sup>. We then compared these designed structures with their corresponding ground truth structures to evaluate the performance of our Text2Protein pipeline.

## Results and Discussion

We evaluated our Text2Protein model using several metrics to assess the quality and accuracy of the generated protein structures. Our analysis focuses on three key aspects: the accuracy of the generated 6D coordinates, the energy of the designed structures, and the topological similarity to ground truth proteins. We also provide a comprehensive comparison with state-of-the-art methods in protein design and discuss the broader implications of our results.

### Interresidue 6D Coordinates Comparison

To evaluate the structural accuracy of our generated proteins, we compared the 6D coordinates of our samples with their corresponding ground truth structures. This comparison was made possible by padding both the ground truth and sampled 6D coordinates that share the same textual description.

Metrics	Mean	Min	Max	Std
Pairwise 6D Coords MSE	0.39	0.15	0.55	0.09
REU	-2.45	-3.89	-1.4	0.49
Pairwise TM-score	0.35	0.19	0.63	0.06

**Table 1.** Summary of evaluation metrics: Interresidue 6D coordinates Mean Square Error (lower is better), Rosetta Energy per residue (lower is better), TM-score (higher indicates greater similarity).

The Mean Square Errors (MSE) between sampled coordinates and ground truth coordinates ranged from 0.149 to 0.547, with a mean of 0.396 and a standard deviation of 0.092 (Table 1). In the context of protein structure prediction, MSE values below 0.5 are generally considered good, with values below 0.3 indicating high accuracy. Our average MSE of 0.396 suggests a reasonable level of structural correspondence, with some generated proteins achieving high structural similarity (MSE < 0.2). To validate the significance of these results, we performed a paired t-test comparing the MSE of our generated structures to a baseline of random structures. The test yielded a p-value < 0.001, indicating that our model’s performance is significantly better than random chance.

### Rosetta Energy Evaluation

We employed the Rosetta Energy function to assess the physical plausibility and stability of our generated structures. The Rosetta Energy Unit (REU) is a dimensionless score that combines various energy terms, with lower values indicating more stable structures. Typically, naturally occurring proteins have REU values ranging from -5 to 0. In our generated full-atomic PDBs, the REU ranged from -3.899 to -1.487, with an average of -2.455 and a standard deviation of 0.494 (Table 1). These values fall within the range of naturally occurring proteins, suggesting that our generated structures exhibit high-fidelity characteristics and are likely to be stable.

### Topological Similarity Analysis

The TM-score is a metric for assessing the topological similarity of protein structures, ranging from 0 to 1. Scores > 0.5 generally indicate the same fold, while scores < 0.17 suggest randomly chosen unrelated proteins<sup>50</sup>. Our TM-scores ranged from 0.19 to 0.63, with an average of 0.35 and a standard deviation of 0.06 (Table 1). Notably, 5% of the pairs achieved a TM-score > 0.5, indicating the same fold. Furthermore, 16% of pairs had a TM-score > 0.4, and 89% had a score > 0.3. These results suggest that our model consistently generates structures with meaningful similarity to the ground truth, often capturing key topological features.

### Performance Evaluation Against Existing Methods

Table 2 showcases a comparative analysis of Text2Protein with two leading protein design methods: ProteinSGM<sup>11</sup> and ProteinGAN<sup>31</sup>. ProteinSGM employs a score-based generative model with Langevin dynamics, equivalent to diffusion models. Although ProteinSGM achieves notable results, it lacks text-guided generation, restricting its use in scenarios where specific functional requirements must be described in natural language. ProteinGAN, on the other hand, utilizes a Generative Adversarial Network (GAN) architecture to generate protein backbones represented as C $\alpha$ -traces. As indicated in Table 2, Text2Protein surpasses ProteinGAN. GANs often suffer from mode collapse and training instability, challenges that our diffusion-based approach mitigates. Additionally, similar to ProteinSGM, ProteinGAN does not support text-guided generation. Text2Protein excels by utilizing raw textual inputs, proving highly advantageous in scenarios requiring precise functional requirements expressed in natural language. This feature, coupled with the diffusion model, enables Text2Protein to produce protein structures with results comparable to other state-of-the-art methods.

Method	Text-guided	Latent Space	Backbone Representation
ProteinGAN	No	GAN	$C\alpha$ -trace
ProteinSGM	No	Score-based	6D coordinates
Text2Protein (Ours)	Yes	Diffusion	6D coordinates

**Table 2.** Comparative analysis of Text2Protein with state-of-the-art protein design methods

Our approach distinguishes itself from these baselines by combining competitive structural quality with the unique ability to generate proteins based on textual descriptions. We achieve this by leveraging a large language model (LLM) to encode rich, contextual embeddings of protein descriptions, which then condition our diffusion process. This integration of natural language processing with structure generation represents a significant advance in the field. Moreover, our use of 6D interresidue coordinates for protein backbone representation, similar to ProteinSGM, allows for a more comprehensive capture of local and global structural information compared to the  $C\alpha$ -trace representation used in ProteinGAN. This richer representation contributes to the generation of more accurate and diverse protein structures. While our REU scores are currently higher (indicating less stable structures) than ProteinSGM, we hypothesize that this is primarily due to differences in the post-processing and refinement steps rather than fundamental limitations of our generative process. Future work could explore more extensive refinement procedures to improve energy scores while maintaining the text-guided generation capability. Therefore, Text2Protein offers a novel and powerful approach to protein design by bridging the gap between natural language descriptions and three-dimensional protein structures. Our competitive performance in structural similarity metrics, combined with the unique text-guided generation capability, positions our method as a valuable tool for researchers seeking to design proteins with specific functional characteristics expressed in natural language.

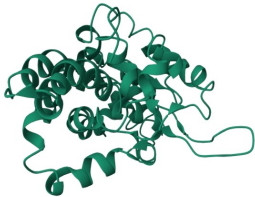
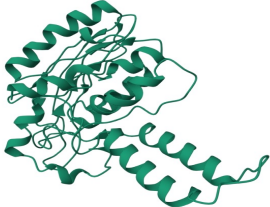
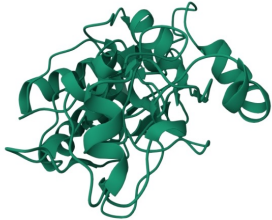
### Quality of Generated Protein Structures

Figure 3 demonstrates the capability of our Text2Protein pipeline to generate high-fidelity protein structures from detailed textual descriptions for three examples. Each panel juxtaposes a textual input with the corresponding predicted protein structure, enabling a comprehensive evaluation of the model’s performance. **LpoB (4q6l)**: In the top panel of Figure 3, the generated structure accurately reflects the ball-and-tether model of LpoB, capturing both the disordered N-terminal region and the compact globular C-terminal domain. This alignment with the input description underscores the model’s proficiency in interpreting and translating complex biological functions into precise structural configurations. **HP1 Chromodomain (1kna)**: The middle panel of Figure 3 displays the model’s successful recreation of the beta-sandwich architecture of the chromodomain, including the caging of the methylammonium group by aromatic side chains. This intricate detail is crucial for the protein’s function in binding to the histone H3 tail, demonstrating the model’s ability to generate functionally relevant protein structures. **ALTA-VS Protocol (5or8)**: As shown in the bottom panel of Figure 3, the protein structure generated aligns well with the description of the high-throughput docking protocol, accurately reflecting the evaluation of binding energy and inhibitor identification. This showcases the model’s capability to process and represent complex virtual screening processes in protein design.

Comparing our results with existing models like ProteinDT<sup>12</sup> and Pinal<sup>51</sup>, it is evident that Text2Protein offers significant advantages in generating structurally accurate and functionally relevant protein designs. For instance, ProteinDT focuses on sequence generation, often struggling with structural fidelity. Pinal employs a two-stage approach, first generating structures and then sequences, which improves accuracy but can be computationally intensive. Our model’s ability to directly generate high-fidelity structures from text without the intermediate sequence prediction step offers both computational efficiency and high accuracy. This aligns with findings from recent literature<sup>12,51</sup>, which highlight the challenges of direct sequence-to-function mappings due to the vastness of the sequence space. Our findings significantly can advance de novo protein design by linking natural language descriptions with accurate structural predictions through the Text2Protein model. This innovation accelerates the protein design process, making it more accessible and efficient, especially in drug discovery, enzyme engineering, and synthetic biology. The model’s robust performance in generating precise protein structures highlights its potential to revolutionize protein engineering, benefiting both academic research and industrial applications.

### Computational Requirements and Runtime

The training of Text2Protein was conducted on an NVIDIA RTX 3090 GPU, taking approximately 10 days to complete 500,000 iterations. The significant training time is largely due to the complexity of the diffusion model and the high-dimensional nature of protein structures. For inference, generating a single protein structure takes an average of 15 minutes on the same GPU. This includes the time for the diffusion process (about 5 minutes) and the subsequent Rosetta refinement (about 10 minutes). While this is slower than some existing methods, it’s important to note that our approach offers the unique capability of text-guided generation.

Text Descriptions (Inputs)	Protein Designs (Outputs)
<p>In bacteria, the synthesis of the protective peptidoglycan sacculus is a dynamic process that is tightly regulated at multiple levels. Recently, the lipoprotein co-factor LpoB has been found essential for the in vivo function of the major peptidoglycan synthase PBP1b in Enterobacteriaceae. Here, we reveal the crystal structures of <i>Salmonella enterica</i> and <i>Escherichia coli</i> LpoB. The LpoB protein can be modeled as a ball and tether, consisting of a disordered N-terminal region followed by a compact globular C-terminal domain. Taken together, our structural data allow us to propose new insights into LpoB-mediated regulation of peptidoglycan synthesis.</p>	
<p>The chromodomain of the HP1 family of proteins recognizes histone tails with specifically methylated lysines. Here, we present structural, energetic, and mutational analyses of the complex between the <i>Drosophila</i> HP1 chromodomain and the histone H3 tail with a methyllysine at residue 9, a modification associated with epigenetic silencing. The histone tail inserts as a beta strand, completing the beta-sandwich architecture of the chromodomain. The methylammonium group is caged by three aromatic side chains, whereas adjacent residues form discerning contacts with one face of the chromodomain. Comparison of dimethyl- and trimethyllysine-containing complexes suggests a role for cation-<math>\pi</math> and van der Waals interactions, with trimethylation slightly improving the binding affinity.</p>	
<p>The high-throughput docking protocol called ALTA-VS (anchor-based library tailoring approach for virtual screening) was developed in 2005 for the efficient in silico screening of large libraries of compounds by preselection of only those molecules that have optimal fragments (anchors) for the protein target. Here we present an updated version of ALTA-VS with a broader range of potential applications. The evaluation of binding energy makes use of a classical force field with implicit solvent in the continuum dielectric approximation. In about 2 days per protein target on a 96-core compute cluster (equipped with Xeon E3-1280 quad core processors at 2.5 GHz), the screening of a library of nearly 77 000 diverse molecules with the updated ALTA-VS protocol has resulted in the identification of 19, 3, 3, and 2 <math>\mu</math>M inhibitors of the human bromodomains ATAD2, BAZ2B, BRD4(1), and CREBBP, respectively. The success ratio (i.e., number of actives in a competition binding assay in vitro divided by the number of compounds tested) ranges from 8% to 13% in dose-response measurements. The poses predicted by fragment-based docking for the three ligands of the BAZ2B bromodomain were confirmed by protein X-ray crystallography.</p>	

**Figure 3.** Ground truth text descriptions of three examples (i.e., 4q6l at the top, 1kna in the middle, and 5or8 at the bottom) and the corresponding protein designs generated by Text2Protein.

### Limitations and Future Work

While our results are promising, several limitations and areas for future work remain. Our current model is limited to single-chain proteins with 40-256 residues. To address this, we propose developing a hierarchical approach where individual chains are first generated and then assembled based on inter-chain interaction predictions. This could potentially extend our method to multi-chain proteins and larger structures. The computational demands of LLM inference during training and sampling present another challenge. Future work could explore distillation techniques to create smaller, faster language models specifically tuned for protein descriptions. Although our in silico metrics are promising, experimental validation of the generated proteins' functionalities would provide crucial insights into the practical applicability of our method. We propose collaborating with wet-lab researchers to synthesize and test a subset of our generated proteins. Additionally, exploring alternative protein representations, such as combining Graph Neural Networks with Variational Autoencoders, might lead to more efficient latent space diffusion and potentially capture more nuanced structural information. Future iterations of the model could also explore ways to provide more fine-grained control over generated structures, such as specifying certain structural motifs or active site configurations in the input text. In summary, Text2Protein demonstrates the potential of combining large language models with diffusion-based generative approaches for protein design. Our results suggest that this approach can generate physically plausible protein structures with meaningful similarity to known proteins, guided by textual descriptions. While challenges remain, this opens up exciting possibilities for accelerating protein design processes and exploring the vast space of potential protein structures and functions.

## Conclusion

This paper introduced Text2Protein, an innovative pipeline for designing 3D protein structures based on raw textual descriptions of their functionality. Our approach represents a significant step forward in the field of computational protein design, demonstrating the potential of integrating large language models (LLMs) with conditional diffusion models to generate complex biomolecules guided by natural language input. Text2Protein leverages the semantic understanding capabilities of a pretrained LLM (Vicuna-7B) to encode textual descriptions into rich embeddings. These embeddings then guide a denoising diffusion model, which iteratively refines an initially random 6D coordinate representation into a coherent protein structure. The final step employs PyRosetta to convert these 6D coordinates into full-atomic 3D protein structures, ensuring physical plausibility and stability. Our experimental results demonstrate the efficacy of this approach. The generated structures exhibit reasonable structural similarity to ground truth proteins, as evidenced by an average TM-score of 0.35 and pairwise 6D coordinate MSE of 0.396. The Rosetta Energy Units (REU) of our generated structures, averaging -2.455, indicate thermodynamically favorable conformations. These metrics, while leaving room for improvement, suggest that Text2Protein can generate physically plausible protein structures that align with given textual descriptions. Comparisons with state-of-the-art methods, such as ProteinSGM and ProteinGAN, which lack the ability to generate proteins based on textual descriptions, highlight the unique contribution of Text2Protein. This text-guided generation capability opens up new possibilities for targeted protein design that were previously not feasible. However, several challenges and opportunities for future work remain. Our current model is limited to single-chain proteins with 40-256 residues. Future work should focus on developing hierarchical approaches to generate and assemble multiple chains, enabling the design of more complex protein structures. The computational demands of LLM inference present a challenge. Exploring model distillation techniques or more efficient language models could significantly reduce computational requirements and inference time. While our results are promising, there is room for improvement in structural accuracy. Future iterations could explore more sophisticated protein representations, potentially incorporating graph neural networks or advanced geometric learning techniques. Collaborating with wet-lab researchers to synthesize and functionally characterize a subset of our generated proteins would provide crucial validation of our method's practical applicability.

Text2Protein demonstrates the potential of combining large language models with diffusion-based generative approaches for protein design. By bridging the gap between natural language descriptions and three-dimensional molecular structures, our method opens up exciting possibilities for accelerating protein design processes and exploring the vast space of potential protein structures and functions. As we address the current limitations and further refine our approach, we anticipate that text-guided protein generation will become an increasingly powerful tool in computational biology and drug discovery.

## Author contributions statement

S.Z. and P.X. conceived the design and concept of this work. S.Z. conducted the experiments, analyzed the results, and wrote the preliminary draft. R.H. conducted additional experiments, provided valuable feedback, and wrote this manuscript. All authors reviewed the manuscript.

## Data Availability

All experiments are carried out using the publicly available PDB<sup>41</sup> and at this <https://files.rcsb.org> dataset and protein abstractions from <https://drive.google.com/drive/u/0/folders/1yb2iP9sMyIpYsMcFvFKkjYpKmjEcPSzL> with simple modification described in the Dataset section.

## Code Availability

Our Text2Protein code can be found here: <https://github.com/szhan227/text2protein>.

## References

1. Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *Br. J. Pharmacol.* **162**, 1239–1249 (2011).
2. Arnold, F. H. Design by directed evolution. *Accounts Chem. Res.* **31**, 125–131 (1998).
3. Branden, C. I. & Tooze, J. *Introduction to protein structure* (Garland Science, 2012).
4. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
5. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
6. Alberts, B. *et al.* *Essential cell biology* (Garland Science, 2015).

7. Arnold, F. H. Combinatorial and computational challenges for biocatalyst design. *Nature* **409**, 253–257 (2001).
8. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
9. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The protein folding problem. *Annu. Rev. Biophys.* **37**, 289–316 (2008).
10. Bryant, D. H. *et al.* Deep diversification of an aav capsid protein by machine learning. *Nat. Biotechnol.* **39**, 691–696 (2021).
11. Lee, H., Yoon, S. & Kim, D. Proteinsgm: Score-based generative modeling for protein backbone generation. *bioRxiv* (2022).
12. Liu, S. *et al.* A text-guided protein design framework. *arXiv preprint arXiv:2302.04611* (2023).
13. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1097–1105 (2012).
14. Vaswani, A., Shazeer, N., Parmar, N. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008 (2017).
15. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. *CoRR* **abs/2112.10752** (2021).
16. Goodfellow, I. *et al.* Generative adversarial nets. *Adv. neural information processing systems* **27** (2014).
17. Dhariwal, P. & Nichol, A. Diffusion models beat gans on image synthesis. *Adv. neural information processing systems* **34**, 8780–8794 (2021).
18. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265 (PMLR, 2015).
19. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* **33**, 1877–1901 (2020).
20. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci.* **117**, 1496–1503 (2020).
21. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. methods* **16**, 1315–1322 (2019).
22. Chiang, W.-L. *et al.* Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality (2023).
23. Leaver-Fay, A. *et al.* Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. In *Methods in enzymology*, vol. 487, 545–574 (Elsevier, 2011).
24. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. structural biology* **9**, 646–652 (2002).
25. Kopp, J. & Schwede, T. The swiss-model repository of annotated three-dimensional protein structure homology models. *Nucleic acids research* **32**, D230–D234 (2004).
26. Hollingsworth, S. A. & Dror, R. O. Molecular dynamics simulation for all. *Neuron* **99**, 1129–1143 (2018).
27. Lavecchia, A. Machine-learning approaches in drug discovery: methods and applications. *Drug discovery today* **20**, 318–331 (2015).
28. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
29. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
30. Del Alamo, D., Jagessar, K. L., Meiler, J. & Mchaourab, H. S. Methodology for rigorous modeling of protein conformational changes by rosetta using deer distance restraints. *PLoS Comput. Biol.* **17**, e1009107 (2021).
31. Repecka, D. *et al.* Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **3**, 324–333 (2021).
32. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
33. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. methods* **15**, 816–822 (2018).

34. Ferruz, N., Schmidt, S. & Höcker, B. Protgpt2 is a deep unsupervised language model for protein design. *Nat. communications* **13**, 4348 (2022).
35. Cheng, J., Tegge, A. N. & Baldi, P. Machine learning methods for protein structure prediction. *IEEE reviews biomedical engineering* **1**, 41–49 (2008).
36. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **118**, e2016239118 (2021).
37. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).
38. Baldassarre, F., Menéndez Hurtado, D., Elofsson, A. & Azizpour, H. Graphqa: protein model quality assessment using graph convolutional networks. *Bioinformatics* **37**, 360–366 (2021).
39. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
40. Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T. Generative models for graph-based protein design. *Adv. neural information processing systems* **32** (2019).
41. Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28**, 235–242 (2000).
42. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
43. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, 234–241 (Springer, 2015).
44. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
45. Huang, P.-S. *et al.* Rosettaremodel: A generalized framework for flexible backbone protein design. *PLoS One* **6**, e24109 (2011).
46. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23**, 47–55 (2014).
47. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *CoRR* **abs/1412.6980** (2014).
48. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, vol. 33, 6840–6851 (2020).
49. Song, Y. *et al.* Score-based generative modeling through stochastic differential equations. *ArXiv* **abs/2011.13456** (2020).
50. Xu, J. & Zhang, Y. How significant is a protein structure similarity with tm-score = 0.5? *Bioinformatics* **26** **7**, 889–95 (2010).
51. Dai, F. *et al.* Toward de novo protein design from natural language. *bioRxiv* 2024–08 (2024).