Gradient-Based Image and Protein Generation

Jun W. Kim
Stanford University, Department of Biomedical Data Science
443 Via Ortega MC 4245, 94305, CA, USA
junwkim@stanford.edu

Abstract

Designing proteins with targeted properties is a key challenge in drug discovery. Traditional approaches often require tedious wet-lab experiments or specialized, property-specific models. Here, we investigate a flexible gradient-based strategy to guide protein generation without repeatedly retraining new models. By leveraging pretrained neural networks, we compute gradients of property-specific classifiers with respect to the protein (or image) representation, then incorporate these gradients into a diffusion-based generative framework. As a proof-of-concept, we first apply our approach to image generation using pretrained ImageNet models, recovering Inceptionism-like visualizations. We then shift focus to protein design, integrating an enzyme and nonenzyme classifier into a generative pipeline. Our results suggest that gradient-based conditioning can guide the sampling process toward desired protein properties, demonstrating an efficient alternative to building new models for every property of interest.

1. Introduction

Proteins play various roles in biological systems, from mediating signaling pathways to providing structural support. Dysfunctional or misfolded proteins have been implicated in neurodegenerative disorders and cancer, and leveraging protein function has become a major focus in both research and therapeutic development [7]. Proteins represent one of the fastest-growing categories of approved therapies, driven by their high specificity and capacity to be engineered for new applications. Consequently, developing computational methods to generate novel, foldable protein structures holds the potential to revolutionize drug development. Historically, efforts to design proteins have often employed wet-lab-based libration generation and screening approaches that assemble experimentally derived motifs [3]. More recent strategies based on deep generative models have shown promise but introduce complexities in downstream post-processing and risk generating unphysical symmetries. Advances like diffusion-based networks (e.g., RFDiffusion) offer robust capabilities for unconditional and conditional protein design [6]. Despite these advancements, a common limitation remains: to tune a generative model for new protein properties, one typically must retrain or fine-tune the model from scratch. In this study, we evaluate gradient-based methods, applying principles from backpropagation, to incorporate existing classifiers or regressors for protein properties directly into a generative pipeline. Rather than building a new model each time a different property is desired, our framework conditions the protein sequence or structure, x, on multiple target characteristics, y and z, through a joint distribution p(x—y,z). As a preliminary demonstration, we use this gradient-based idea in image generation tasks with pretrained ImageNet models to validate proof-of-concept. We then shift focus to protein design, seeking to optimize for enzymatic function. Lastly, we evaluate the resulting designs with distinct property predictors to ensure that the newly generated proteins indeed exhibit the targeted attributes. By circumventing the need for repeated retraining, this paradigm offers a flexible tool for guiding protein generation and exemplifies a step toward more efficient protein engineering.

2. Materials and Methods

2.1. Convolutional neural network (CNN) architectures

We evaluated four primary CNN architectures via PyTorch: • VGG16Experimental – A variant of the VGG16 model by Simonyan and Zisserman [4]. This particular "experimental" version exposes intermediate feature layers, allowing deeper inspection of learned representations. • ResNet50 – A residual network architecture by He et al. [1], which can also optionally load Places365 weights for scene-centric images (Zhou et al.) [8]. For ImageNet-based tasks, PyTorch's pretrained ResNet50 is used. • GoogLeNet – A pretrained variant available in torchvision.models, trained on ImageNet (1.2 million images). This model provides general-purpose feature extraction and was used to demon-

strate "inceptionism" style gradient-based visualizations. • EfficientNetV2-S – An efficient, mobile-friendly backbone that extends the EfficientNet paradigm. It was loaded from TorchVision with ImageNet1K V1 weights [5]. All models were set to eval mode, not requiring additional parameter updates so that only the input (image or protein representation) could be optimized via gradient-based methods.

2.2. Preprocessing and Image Handling

Images were read using OpenCV in BGR format. Intensities were scaled from [0–255] to [0–1] and normalized by subtracting channel-wise ImageNet means and dividing by standard deviations. Then images were resized for computational efficiency. For synthetic, random starting images (noise), pixels were randomly sampled from a Gaussian distribution. For consistent reproducibility random seeds for NumPy and PyTorch were fixed.

2.3. Gradient-Ascent Visualization

To intuitively examine how a classifier "sees" an image or what features define a certain class, we implemented a gradient-ascent method inspired by DeepDream. Focusing on a particular "target neuron" or logit, we computed the gradient d(activation)/d(image) at each iteration, scaling and adding it back to the image to maximize that neuron's response. Where necessary, we applied Gaussian blur to balance fine and coarse details. Multiple scales could be explored by successively resizing the image, performing gradient steps, and restoring dimensions.

2.4. Protein Sequence Encoding

Protein sequences were encoded using a Python script that converts amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, X) to integer indices. Sequences were padded or truncated to a specified maximum length (300 residues), generating a (300 × 21) one-hot matrix cast to a float tensor. This representation is compatible with single-channel CNN inputs after reshaping to (batch size, 1, 300, 21).

2.5. Modified ResNet18 for Enzyme Classification

For protein-related classification, we adopted a custom ResNet18 with minimal downsampling to preserve spatial resolution along the sequence dimension. The first convolution layer was adapted to accept a single-channel input, the initial maxpool layer was removed, and the stride in layers 3 and 4 was reduced to 1. An attention-based mechanism was added, providing summarized attention maps at each layer. The entire architecture outputs both a classification logit (e.g., enzyme vs. nonenzyme) and multiple attention maps for interpretability as an option.

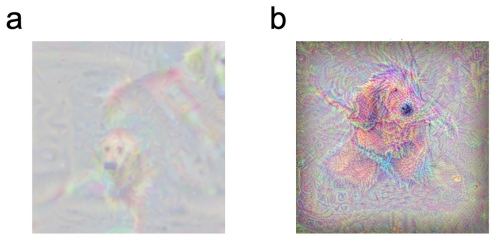

Figure 1. Gradient-based image generation using GoogLeNet (a) and EfficientnetV2 (b). Activations were maximized for the class representing golden retriever

2.6. Coupling with Chroma for Protein Generation

We then integrated this modified ResNetAttentionModel into Chroma, a diffusion-based protein design framework [2]. A custom conditioner class computes the logit or attention map for a given protein sequence and adjusts Chroma's energy function U to guide the model toward sequences meeting the desired classification outcome. Specifically, the energy term is updated via a penalty or reward derived from the classification or attention-based signals. Through iterative sampling, Chroma shifts the protein distribution towards sequences predicted to be enzymes.

3. Results

3.1. Class-Specific Inceptionism in Image Space

As an initial proof-of-concept, we tested whether gradient ascent could synthesize recognizable images from pretrained classifiers. Using GoogLeNet trained on ImageNet, we maximized the logit for class index 207 (associated with golden retrievers). Starting from random noise, the image underwent iterative gradient adjustments, ultimately visualizing coherent "golden retriever-like" patterns. Similar results were obtained with VGG16, ResNet50, and EfficientNet, confirming that gradient-based input modification is a general phenomenon spanning multiple CNN architectures (Fig. 1).

3.2. Dual-Mode Classifier Combinations

Next, we explored combining two pretrained classifiers (e.g., VGG and ResNet) by assuming conditional independence of their target classes probability. By summing each network's class logit (or other activation measure), the model performed gradient ascent on both targets simultaneously. The resulting images displayed semantically mixed or fused characteristics (Figure 2). This outcome suggests that with well-chosen hyperparameters (e.g. controlling gradient magnitude), dual-mode or even multi-mode inceptionism can approximate joint distributions p(y, z—x) to yield composite image features.

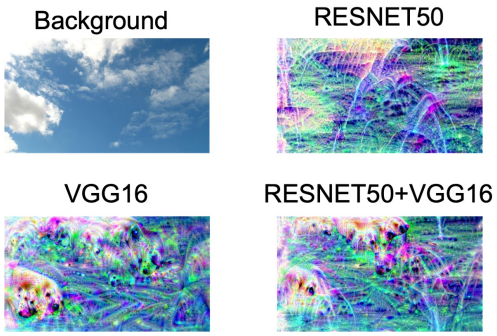

Figure 2. Gradient-based image generation by combining two different classifier (ResNet50: fountain; VGG16: Golden Retriever

3.3. Protein Design with Diffusion and Classifiers

Extending these principles to protein sequences presented unique challenges, as verifying correctness of newly generated sequences or structure is not as straightforward as visual inspection of a golden retriever. We therefore coupled our gradient-based conditioning with Chroma—a diffusion-based protein generative framework. A ResNetAttentionModel served as the classifier to differentiate enzyme from nonenzyme sequences, providing a suitable penalty term to guide diffusion sampling. Generated proteins consistently scored as enzymes, according to the same classifier's logits. Although these sequences' functional validity requires wet-lab confirmation, our results demonstrate that gradient-based property conditioning can be integrated into a diffusion approach, testing the generative process for desirable biochemical traits without retraining an entire generative model from scratch.

4. Discussion

Our study illustrates how pretrained classifiers can be used "on demand" to shape generative outputs. By taking gradients with respect to the input, we effectively implement an iterative feedback loop that prioritizes outputs satisfying a chosen property, whether it is a visual class (e.g., dog breed) or a protein function (e.g., enzyme classification). Although popular deep generative models like GANs or diffusion networks can embed properties via conditional training, they often necessitate specialized datasets or model modifications. Our gradient-based approach circumvents this by reusing existing high-accuracy classifiers, drastically reducing the need for property-specific retraining. However, several issues remain. First, balancing multiple properties simultaneously can be non-trivial, especially if properties are mutually exclusive or highly correlated. Second, the gradient-based modifications may generate unrealistic or artifact-ridden structures, underscoring the importance of carefully chosen fixed parameters such as learning rates, normalization steps, and feasibility checks (e.g., for protein designs). Finally, thorough domain-specific validation is crucial for tasks such as protein engineering, where real-world efficacy cannot be inferred directly from computational outputs alone. In conclusion, gradient-based property conditioning offers a promising route toward more flexible, modular design processes. For proteins, this is a promising advancement toward an integrated frameworks that combine generative sampling with multiple orthogonal property predictors, potentially accelerating progress in biological therapeutics, industrial enzymes, and more.

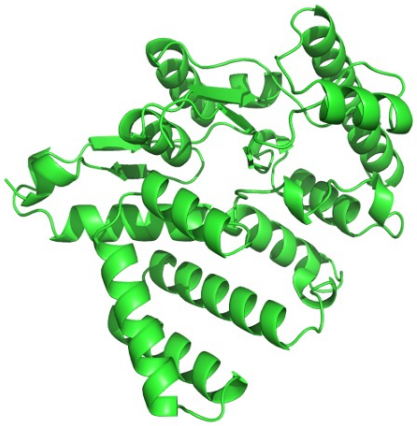

Figure 3. Enzyme generation using conditional protein generation. Chroma was used for diffusion-based sampling; enzyme classifier was used for conditioning. The generated sequence was validated with the same enzyme classifier

5. Acknowledgements

The authors thank the instructors and the teaching staff of CS231N for great lectures and support. We would also like to thank for the cloud credits.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] J. B. Ingraham, M. Baranov, Z. Costello, K. W. Barber, W. Wang, A. Ismail, V. Frappier, D. M. Lord, C. Ng-Thow-Hing, E. R. Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

[3] H. Lu, Q. Zhou, J. He, Z. Jiang, C. Peng, R. Tong, and J. Shi. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal transduction and targeted therapy*, 5(1):213, 2020.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] M. Tan and Q. Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[6] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[7] X. Xie, T. Yu, X. Li, N. Zhang, L. J. Foster, C. Peng, W. Huang, and G. He. Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials. *Signal transduction and targeted therapy*, 8(1):335, 2023.

[8] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.