DeepRHP: A Hybrid Variational Autoencoder for Designing Random Heteropolymers as Protein Mimics

Shuni Li, Zhiyuan Ruan, Andy Shen, Ivan Jayapurna, Ting Xu, Haiyan Huang

University of California Berkeley; Berkeley, California, 94720, USA.
{shuni_li, ruanzy, aashen, ivanfj, tingxu, hyh0110}@berkeley.edu

Abstract

Synthetic random heteropolymers (RHPs), consisting of a predefined set of monomers, offer an approach toward the design of protein-like materials. These RHPs, if designed appropriately, can mimic protein behavior and function. As such, there is a need for computational tools to efficiently guide RHP design. We bridge this gap by developing DeepRHP, a modified variational autoencoder (VAE) model under a semi-supervised framework. By equipping a classical VAE with an additional feature-based VAE, DeepRHP forces the latent space to capture structures of critical chemical features as well as individual RHP sequence patterns. In this sense, our method is versatile by allowing any relevant features to be incorporated in a hybrid manner. We demonstrate the effectiveness of DeepRHP by suggesting potential monomer compositions that stabilize membrane proteins (e.g. Aquaporin Z) in non-native environments and cross-validating our prediction with published results. The concordance between our model and true RHP function suggests strong potential in utilizing hybrid autoencoder architectures to guide RHP design for proteins and other biological compounds.

1 Introduction

There is a significant interest in engineering synthetic materials capable of replicating protein functions while satisfying stability and compatibility with device fabrication and integration. However, it remains an insurmountable challenge to synthesize sequence-specific polymers. This has led to a recent surge of research in designing protein-like random heteropolymers. Random heteropolymers (RHPs) are an ensemble of many polymer chains with each being composed of monomers arranged in random order (Hilburg et al. 2020). Recent developments have demonstrated that RHPs can act as chaperone proteins for protein stabilization in non-biological environments (Panganiban et al. 2018), a critical bottleneck to fabricate protein-embedded plastics for end-of-life plastic degradation (DelRe et al. 2021). In addition, RHPs can be designed to act as channel proteins for rapid and selective proton transportation (Jiang et al. 2020), important for fuel cells and energy storage.

Despite the fact that RHPs can serve as great biofunctional materials, designing RHPs with desired function is challenging because both the exact monomeric sequences and conformations of synthetic RHP chains are not deterministic. Traditional protein design methods rely heavily on high-throughput sequencing data and 3D structures. For example, directed evolution methods evolve protein function by iteratively mutating a selected protein sequence (Arnold 2018), while *de novo* methods build novel proteins that fold into a certain structure (Huang, Boyken, and Baker 2016). Without exact sequences and structures, there are no rational design principles for creating suitably functional RHP chains. Current RHP designs are largely empirical and depend on time-intensive lab screenings over various monomer compositions and chain lengths. For each RHP made in the lab, ensembles of thousands of sequences are simulated under the same monomer composition in order to understand why certain compositions perform better than others. In this process, scientists face two practical design questions that can potentially accelerate progress if answered:

- How many monomers should be included in a RHP system? Recent results show that RHPs can mimic protein function with only four monomers (Panganiban et al. 2018; Jiang et al. 2020), but it remains unclear how many monomers are enough to include in the alphabet.
- How can one find monomer compositions corresponding to specific protein functions?

Answering these questions requires new methods to model and analyze RHP sequences as an ensemble instead of as individual chains. To our knowledge, there is very limited literature on computational methods of modeling RHPs. As the only two examples, Zhou et al. (2022) used Hidden Markov Models to characterize the functionality of proton-transporting RHPs and Tamasi et al. (2022) utilized Gaussian process regression coupled with Bayesian optimization for optimal copolymer identification.

Here we propose DeepRHP, a modified variational autoencoder trained in a semi-supervised manner, for modeling general RHP sequence data and discovering RHP compositions for protein function. This tool serves as a first step that can guide RHP design by examining their protein-mimicking behavior. The key contributions of this study are:

- We are the first to answer RHP design questions with deep learning. DeepRHP learns interpretable latent representations for RHP sequences and provides a platform to perform similarity analysis between target proteins and RHP sequences in an ensemble.

- DeepRHP provides insights into the two important design parameters: monomer alphabet size and monomer composition. We show that the best monomer composition suggested by DeepRHP matches published experimental results.
- DeepRHP is flexible enough to incorporate any function-related chemical features for a wide variety of protein functions.

VAE-based architectures are some of the first model classes used to identify latent representations for biological sequences, and are useful in downstream tasks like identifying mutation effects (Sinai et al. 2017; Riesselman, Ingraham, and Marks 2018) and designing novel functional proteins (Greener, Moffat, and Jones 2018; Costello and Martin 2019). Therefore, we should expect to leverage the same machine learning theory in macromolecular cheminformatics, specifically in this instance of using RHPs to mimic natural biopolymers.

2 Data

Our work utilizes the RHP system developed in both Panganiban et al. (2018) and Jiang et al. (2020). This system consists of four methacrylate-based monomers: methyl methacrylate (MMA), 2-ethylhexyl methacrylate (EHMA), oligo (ethylene glycol) methacrylate (OEGMA), and 3-sulfopropyl methacrylate potassium salt (SPMA). MMA and EHMA are the hydrophobic monomers used to tailor overall hydrophobicity, while OEGMA and SPMA are the hydrophilic monomers used to reduce the aggregation propensity of RHPs.

We used Compositional Drift, a software developed by Smith et al. (2019) to simulate 10,000 sequences per monomer composition listed in Table 1. This software uses established mathematical copolymer models in tandem with Monte-Carlo simulation to calculate RHP sequences based on experimental conditions. The authors showed that, while each chain simulated is random at the sequence level, it contains characteristic segments that have a well-defined statistical distribution (Smith et al. 2019). The reasoning behind the monomeric compositions for each specific RHP is further discussed in Section 4.

We also collected 30,000 membrane protein sequences and 30,000 globular protein sequences with 50% identity threshold from the UniProt database (UniProt Consortium 2020). Some common pre-processing procedures were performed, including discarding sequences with uncommon amino acids and lengths. Each protein was then reduced into its monomer-equivalent form according to the assignment in Table 2. Note that the reduction of protein alphabet is not uncommon in protein sequence analysis, see Liang et al. (2022) for a comprehensive review. Here our reduction rule is based on monomer hydrophobicity and charge.

3 DeepRHP Methodology

In order to address the domain questions raised in Section 1, we developed DeepRHP, a modified variational autoencoder under semi-supervised framework for learning low-dimensional RHP sequence representations. The model architecture is illustrated in Figure 1. We assume the sequence family X follows a probability distribution $p(x)$ and there exists an underlying latent variable $z \sim N(\mu_z, \Sigma_z)$ that captures intrinsic unobserved sequence properties. For each sequence x, there also exists a function-related feature y, which can be considered as a deterministic transformation of x. In the application case presented in Section 4, y is the average hydrophilic–lipophilic balance (HLB) value of sliding windows along each sequence (Kyte and Doolittle 1982). HLB measures local hydrophobicity and solubility distributions and is closely related to RHP functions (Panganiban et al. 2018; Jiang et al. 2020). The motivation for introducing other function-related chemical features (e.g. HLB) is for them to guide the formation of the latent space.

	RHP	MMA	OEGMA	EHMA	SPMA
2 Mon.	A	0	10	90	0
	B	0	30	70	0
	C	0	50	50	0
	D	0	70	30	0
	E	0	90	10	0
4 Mon.	1	70	25	0	5
	2	65	25	5	5
	3	60	25	10	5
	4	50	25	20	5
	5	40	25	30	5
	6	20	25	50	5
	7	0	25	70	5

Table 1: Two and four-monomer composition of RHPs used for training

Amino acid	Monomer equiv.	Property
C, Y, A, T, G	MMA	Hydrophobic
S, Q, H, N, P	OEGMA	Hydrophilic
L, I, F, W, V, M	EHMA	Very Hydrophobic
E, D, R, K	SPMA	Charged

Table 2: Amino acid (protein) to monomer (RHP) conversion

To incorporate a chemical feature y into our VAE model, we add a feature-driven VAE in parallel with the classical VAE. y and x share the common latent variable z. This is equivalent to simultaneously training two VAEs with shared latent embeddings, and the encoder relies only on x since y is a direct transformation of x, as indicated by the dashed lines in Figure 1.

The objective is still to maximize the log-likelihood $\log p(x)$ given sequence data X as shown in equation:

$$\log p(x) = \log \int p(x \mid z)\, p(z)\, dz. \tag{1}$$

Under the regular VAE setting, Equation 1 can be bound by the well-known evidence lower bound (ELBO) (Kingma and Welling 2013; Rezende, Mohamed, and Wierstra 2014):

$$\log p(x) \geq \mathbb{E}_q \left[\log p(x \mid z)\right] - D_{KL}\left(q(z \mid x) \parallel p(z)\right), \tag{2}$$

where q is the learned posterior of the normal distribution family. In practice, p and q are learned by the encoder and

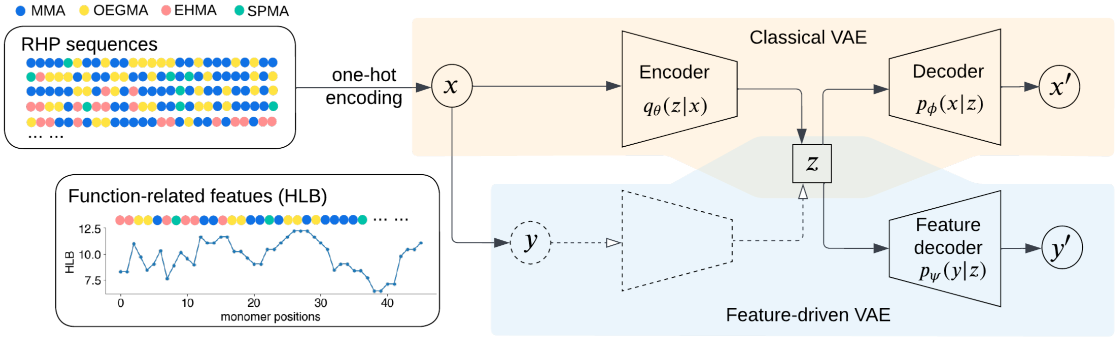

Figure 1: DeepRHP model architecture consisting of a classical VAE equipped with an additional feature-based VAE.

decoder and their weights are optimized through gradient descent.

Traditionally, the reconstruction loss term is approximated by mean-squared error for continuous input, or cross-entropy loss for discrete input. By imposing this hybrid architecture, we can approximate the reconstruction loss through both the classical VAE on x, the feature-driven VAE on y, or a weighted sum of both. Our modified ELBO that considers both sequence structures and chemical features is then formulated as

$$\begin{aligned}\log p(x) \geq \alpha \mathbb{E}_q \left[\log p(x \mid z)\right] + (1-\alpha)\mathbb{E}_q \left[\log p(y \mid z)\right] - \\ D_{KL}\left(q(z \mid x) \parallel p(z)\right), \qquad (3)\end{aligned}$$

where α is a hyperparameter that dictates how much weight is placed on each approximation term. In our case, the first two terms of Equation 3 are approximated as follows:

$$\mathbb{E}_q \left[\log p(x \mid z)\right] \approx \sum_x \sum_l p(x_l) * \log(p(x_l \mid z) \qquad (4)$$

$$\mathbb{E}_q \left[\log p(y \mid z)\right]) \approx -\sum_y ||y - y'||_2^2, \qquad (5)$$

where y' is the output of feature-based decoder denoted by the blue shading in Figure 1.

By optimizing the reconstruction loss in this hybrid manner, we obtain a meaningful low-dimensional latent space that captures the sequence structure relevant to the desired protein function. Additionally, our method comes with interpretability benefits that classical VAEs often lack. Existing works usually concatenate all features together into a single vector for the encoder. The resulting latent space is then obscured, as no physical meanings can be derived for the principal directions. In contrast, our hybrid training leads to meaningful visualizations of the data because the latent variables are directly linked to the chemical features.

Both the encoder and the decoder were implemented with multilayer perceptrons using PyTorch. Each has three fully connected layers with 256, 128, and 64 hidden units, respectively. The feature decoder has two fully connected layers with 32 hidden units. ReLU activation functions were used as non-linearities throughout the network, except in the output layer of the decoder where Sigmoid activation was used instead. The model was trained using the ADAM optimizer with a learning rate of 0.0001. A learning rate scheduler was used when validation loss stopped improving.

4 Results and Discussion

Aquaporins (Aqp) are membrane channel proteins that facilitate water transport between cells. Membrane proteins are unstable and prone to aggregation even under mild experimental conditions. Panganiban et al. successfully stabilized Aquaporin Z (AqpZ) and preserved its function in non-native environments with the presence of RHPs. We demonstrate how DeepRHP can be used to accelerate RHP design by identifying promising monomer compositions.

Panganiban et al. (2018) chose to use 70% hydrophobic monomers and 30% hydrophilic monomers in their RHP system based on a crude protein surface analysis on four protein sequences. We first validate this distribution of monomer hydrophobicities using our model. The latent factors of the two-monomer RHPs and natural proteins are projected onto a two-dimensional space using Principal Component Analysis (PCA), as shown in Figure 2(a). All two-monomer RHPs are composed of one hydrophobic monomer (EHMA) and one hydrophilic monomer (OEGMA). The compositions of *RHP A* through *RHP E* listed in Table 1 are selected to sufficiently reflect this hydrophobicity range. We observe that PC1 correlates with hydrophobicity as RHPs span left to right, with left being least hydrophobic to right being most hydrophobic. The majority of membrane and globular proteins overlap with *RHP B* and *RHP C*, suggesting these two RHP compositions are most similar to natural proteins. On the other hand, most hydrophobic membrane proteins overlap with *RHP B* (30% hydrophilic, 70% hydrophobic), confirming that 30:70 is a good balance for the two-monomer system.

We then fine-tune the performance of the 30:70 distri-

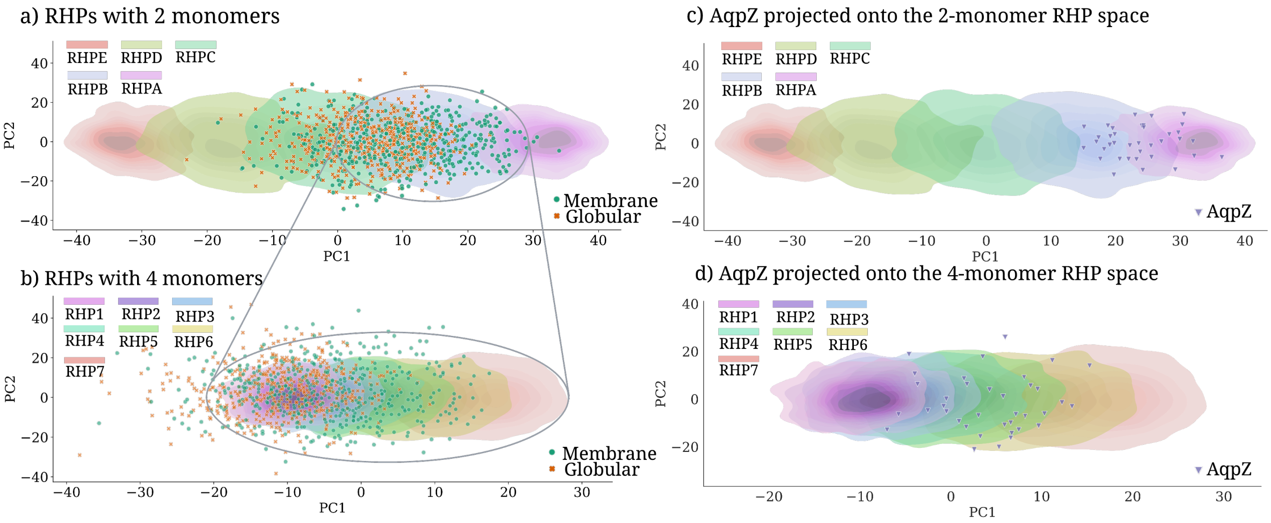

Figure 2: PCA projections of RHP and protein latent factors. Panels (a) and (b) project membrane and globular proteins onto two and four-monomer RHP space, respectively. Panels (c) and (d) project AqpZ onto the same two RHP spaces.

bution of hydrophilic and hydrophobic monomers by increasing the number of monomers from two to four as shown in Figure 2(b). A library of four-monomer-based RHPs was designed by varying the MMA:EHMA ratio. The specific monomer composition is shown in Table 1. Each of *RHP 1* through *RHP 7* is still composed of 30% hydrophilic monomers (OEGMA + SPMA) and 70% hydrophobic monomers (MMA + EHMA).

Panganiban et al. (2018) did not rationalize the choice of four monomers for their design of protein-like RHPs. Our approach explains why the two-monomer alphabet size is insufficient. In Figure 2(b), each of the RHP ensembles can be considered as a subset of *RHP B* and occupies a much more localized natural protein sequence space with smaller variance. In Figures 2(c) and (d), we project AqpZ onto the two-monomer and four-monomer PCA spaces, respectively. In the two-monomer setting, the *RHP B* space is much larger than the span of AqpZ. In the four-monomer setting, however, the AqpZ projections cover the *RHP 4* and *RHP 5* spaces almost entirely. Therefore, we believe the two-monomer sequence space is too broad with respect to proteins while the four-monomer sequence space is more localized, offering stability in synthesizing RHPs.

In addition to providing heuristics regarding the number of monomers, DeepRHP sheds light on the choice of monomer compositions. In Figure 2(d), there is a large overlap between the projected proteins and the *RHP 4* and *RHP 5* contours. Wet-lab experiments in Panganiban et al. (2018) demonstrated that the optimal RHP has the same monomer ratio as that of *RHP 4* and is capable of stabilizing AqpZ. Thus, the overlap between RHPs and AqpZ in the PCA space can modulate their sequence correlation and molecular interactions in the aqueous solution. This indicates that the latent embeddings discovered by DeepRHP are chemically meaningful and play a key role in discovering RHPs that provide strong performance.

5 Conclusion

In this study, we developed DeepRHP, a hybrid variational autoencoder model to guide RHP design. Our model suggests the feasibility of four-monomer compositions to stabilize ApqZ, matching the respective wet-lab experiment. In ablation studies, our model outperforms a singular classical VAE without the additional decoding regressor.

Overall, DeepRHP holds much promise for the future of integrating deep learning techniques, specifically VAEs, into RHP design. Hybrid VAE architectures like DeepRHP possess many advantages. First, they are flexible and can be trained on any sequence family with variable sequence lengths and no multiple sequence alignment is needed. DeepRHP is also flexible due to its flexibility in supervision. It can be totally unsupervised when no prior knowledge on RHP subpopulations is available, or it can also be semi-supervised by combining function-related chemical features with vast amounts of sequence data to improve interpretability of latent variables.

Future work in this regime includes strengthening the quantitative assessment of DeepRHP. Our model is currently assessed in a qualitative manner and validated using laboratory results. We hope to improve DeepRHP by developing a quantitative measure to evaluate the quality of the latent representations. For instance, we hope to complete further downstream tasks such as classifying specific membrane proteins and evaluating similarities between each RHP and their target proteins.

6 Acknowledgments

This work is supported by the U.S. Department of Defense (DOD), Army Research Office, under contract W911NF-13-1-0232 and the National Science Foundation under grant number DGE 2146752. We thank Yaodong Yu and Peter Bickel for useful discussions and comments on the model formulation.

References

Arnold, F. H. 2018. Directed Evolution: Bringing New Chemistry to Life. *Angewandte Chemie - International Edition*, 57: 4143–4148.

Costello, Z.; and Martin, H. G. 2019. How to Hallucinate Functional Proteins.

DelRe, C.; Jiang, Y.; Kang, P.; Kwon, J.; Hall, A.; Jayapurna, I.; Ruan, Z.; Ma, L.; Zolkin, K.; Li, T.; Scown, C. D.; Ritchie, R. O.; Russell, T. P.; and Xu, T. 2021. Near-complete depolymerization of polyesters with nano-dispersed enzymes. *Nature*, 592: 558–563.

Greener, J. G.; Moffat, L.; and Jones, D. T. 2018. Design of metalloproteins and novel protein folds using variational autoencoders. *Scientific Reports*, 8: 16189.

Hilburg, S. L.; Ruan, Z.; Xu, T.; and Alexander-Katz, A. 2020. Behavior of Protein-Inspired Synthetic Random Heteropolymers. *Macromolecules*, 53: 9187–9199.

Huang, P. S.; Boyken, S. E.; and Baker, D. 2016. The coming of age of de novo protein design. *Nature*, 537: 320–327.

Jiang, T.; Hall, A.; Eres, M.; Hemmatian, Z.; Qiao, B.; Zhou, Y.; Ruan, Z.; Couse, A. D.; Heller, W. T.; Huang, H.; de la Cruz, M. O.; Rolandi, M.; and Xu, T. 2020. Single-chain heteropolymers transport protons selectively and rapidly. *Nature*, 577: 216–220.

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes.

Kyte, J.; and Doolittle, R. F. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157: 105–132.

Liang, Y.; Yang, S.; Zheng, L.; Wang, H.; Zhou, J.; Huang, S.; Yang, L.; and Zuo, Y. 2022. Research progress of reduced amino acid alphabets in protein analysis and prediction. *Computational and Structural Biotechnology Journal*, 20: 3503–3510.

Panganiban, B.; Qiao, B.; Jiang, T.; DelRe, C.; Obadia, M. M.; Nguyen, T. D.; Smith, A. A.; Hall, A.; Sit, I.; Crosby, M. G.; Dennis, P. B.; Drockenmuller, E.; Cruz, M. O. D. L.; and Xu, T. 2018. Random heteropolymers preserve protein function in foreign environments. *Science*, 359: 1239–1243.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models.

Riesselman, A. J.; Ingraham, J. B.; and Marks, D. S. 2018. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15: 816–822.

Sinai, S.; Kelsic, E.; Church, G. M.; and Nowak, M. A. 2017. Variational auto-encoding of protein sequences.

Smith, A. A.; Hall, A.; Wu, V.; and Xu, T. 2019. Practical Prediction of Heteropolymer Composition and Drift. *ACS Macro Letters*, 8: 36–40.

Tamasi, M. J.; Patel, R. A.; Borca, C. H.; Kosuri, S.; Mugnier, H.; Upadhya, R.; Murthy, N. S.; Webb, M. A.; Gormley, A. J.; Tamasi, M. J.; Kosuri, S.; Mugnier, H.; Upadhya, R.; Murthy, N. S.; Gormley, A. J.; Patel, R. A.; Borca, C. H.; and Webb, M. A. 2022. Machine Learning on a Robotic Platform for the Design of Polymer–Protein Hybrids. *Advanced Materials*, 34: 2201809.

UniProt Consortium, T. 2020. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49: D480–D489.

Zhou, Y.; Gong, B.; Jiang, T.; Xu, T.; and Huang, H. 2022. Stochastic Variational Methods in Generalized Hidden Semi-Markov Models to Characterize Functionality in Random Heteropolymers.