COLOR: A COMPOSITIONAL LINEAR OPERATION-BASED REPRESENTATION OF PROTEIN SEQUENCES FOR IDENTIFICATION OF MONOMER CONTRIBUTIONS TO PROPERTIES

Akash Pandey* **& Wei Chen**
Department of Mechanical Engineering
Northwestern University
Evanston, IL 60208, USA
{akash.pandey, weichen}@northwestern.edu

Sinan Keten
Department of Mechanical Engineering &
Department of Civil and Environmental Engineering
Northwestern University
Evanston, IL 60208, USA
{s-keten}@northwestern.edu

ABSTRACT

The properties of biological materials like proteins and nucleic acids are largely determined by their primary sequence. Certain segments in the sequence strongly influence specific functions, identifying these segments, or so-called motifs, is challenging due to the complexity of sequential data. While deep learning (DL) models can accurately capture sequence-property relationships, the degree of non-linearity in these models limits the assessment of monomer contributions to a property - a critical step in identifying key motifs. Recent advances in explainable AI (XAI) offer attention and gradient-based methods for estimating monomeric contributions. However, these methods are primarily applied to classification tasks, such as binding site identification, where they achieve limited accuracy (40–45%) and rely on qualitative evaluations. To address these limitations, we introduce a DL model with interpretable steps, enabling direct tracing of monomeric contributions. Inspired by the masking technique commonly used in vision and language processing domains, we propose a new metric ($\mathcal{J}$) for quantitative evaluation on datasets mainly containing distinct properties of anti-cancer peptides (ACP), antimicrobial peptides (AMP), and collagen. Our model exhibits 22% higher explainability than the gradient and attention-based state-of-the-art models, recognizes critical motifs (RRR, RRI, and RSS) that significantly destabilize ACPs, and identifies motifs in AMPs that are 50% more effective in converting non-AMPs to AMPs. These findings highlight the potential of our model in guiding mutation strategies for designing protein-based biomaterials.

1 INTRODUCTION

Machine learning (ML) models have emerged as a powerful tool for establishing primary sequence-to-property relationships in proteins (Brandes et al., 2022; Xu et al., 2020; Elnaggar et al., 2021). This remains an active research area, as sequence data is more accessible than structural data. Deep learning (DL) models like AlphaFold2/3 (Jumper et al., 2021; Abramson et al., 2024) predict structures from sequences but often show low confidence for amorphous and fold-switching proteins (Chakravarty & Porter, 2022; Chakravarty et al., 2024). This limits their reliability for materials with more disordered regions, notably structural proteins such as silks (Lefèvre et al., 2007). Models like Transformers (Vaswani, 2017), Long Short-Term Memory (LSTM) networks (Hochreiter, 1997), and 1D convolution Neural Networks (1D CNN) (Kiranyaz et al., 2019) excel at capturing sequential dependencies. These models enable accurate prediction of various properties of proteins (Yu et al., 2022; Gupta & Zou, 2019; Pandey et al., 2023; Sun et al., 2019; Liu et al., 2022), often achieving R^2 and accuracies over 0.8. Transformers have also enabled pre-trained models like ProtBERT (Brandes et al., 2022; Elnaggar et al., 2021), ESM (Lin et al., 2023), and ProtTXL (Elnaggar

*code: https://github.com/pandeyakash23/COLOR.git

et al., 2021), which differ in self-supervised training strategies. Their success has driven transfer learning frameworks, leveraging pre-trained model outputs as inputs to neural networks for predicting protein properties (Lin et al., 2023; Brandes et al., 2022; Khare et al., 2022). Despite advances in sequence-property prediction, DL models lack interpretability due to several non-linear transformations. This limits their ability to dissect monomers' contribution to the property. Understanding these contributions is essential for identifying critical motifs (Vig et al., 2021) to design mutations for enhanced protein properties (Arakawa et al., 2022; Pandey et al., 2024). Thus, there is a great need for a model that elucidates monomer-level contributions while establishing sequence-property relationships.

With the growing emphasis on Explainable AI (XAI) and interpretability (Ali et al., 2023; Wu et al., 2023), some progress has been made in understanding monomeric contributions in proteins (Danilevicz et al., 2023; Avsec et al., 2021; Vig et al., 2021; Chen et al., 2021; Monteiro et al., 2022; Yu et al., 2024; Jiménez-Luna et al., 2020) while handling primary sequence as an input. One widely used XAI approach is based on the self-attention mechanism in transformers, where a monomer's contribution is determined by the attention it receives from other monomers in the sequence (Wu et al., 2020b; Karimi et al., 2020; Liu et al., 2024). Another popular method is Grad-CAM (Class Activation Mapping) (Selvaraju et al., 2017), which attributes monomeric contribution to the gradient of the output with respect to the monomer's latent space representation. While attention and gradient-based methods have made some strides in providing interpretability for protein sequences, they come with limitations (Danilevicz et al., 2023; Avsec et al., 2021; Vig et al., 2021; Chen et al., 2021; Monteiro et al., 2022; Vangala et al., 2023; Jiménez-Luna et al., 2021). The self-attention mechanism has been shown to be unreliable as an XAI tool to identify critical segments in a sequence (Serrano & Smith, 2019; Bai et al., 2021; Wiegreffe & Pinter, 2019). Similarly, Grad-CAM typically relies on the embeddings from certain Transformer or LSTM-based Large Language Models (LLM) which already have layers of non-linear transformations (Chen et al., 2023; Gligorijević et al., 2021). Furthermore, Grad-CAM has been mostly employed for images and graph-based input data (Selvaraju et al., 2017; Gligorijević et al., 2021; Walter et al., 2024). Therefore, there exists a gap for a more explainable model that can effectively elucidate the contributions of individual monomers within protein sequences.

Current XAI methods for proteins focus primarily on classification tasks like binding site identification, with limited accuracy (40–45%) in detecting all the sites (Chen et al., 2023). These approaches have not been extended to continuous properties like melting temperature and are mainly used for qualitative analyses, such as identifying critical regions in the vicinity of high-contributing monomers (Chen et al., 2021; Monteiro et al., 2022; Vig et al., 2021; Chen et al., 2023). Additionally, no comprehensive evaluation strategy exists to validate the monomeric contribution scores in proteins. In contrast, image analysis (Covert et al., 2023; Nazir et al., 2023; Yoshikawa & Iwata, 2024; Zhou et al., 2022) and NLP (Jahromi et al., 2024) have advanced in quantifying contribution scores using insertion and deletion techniques, which evaluate the impact by systematically adding or removing input segments based on score rankings. To our knowledge, this approach has not been applied to validate monomeric contributions in protein property prediction. We propose leveraging this method to quantify and validate monomeric contribution scores in proteins.

Our Contributions. Building on the above discussion, our work aims to develop an interpretable model to explain monomeric contributions within primary sequences and systematically evaluate the contribution scores generated by the model. As the first step, we develop a novel DL model that establishes a primary sequence-property relationship while enabling the tracing of monomeric contributions from predicted outputs. Our analysis involves: (1) benchmarking the predictive performance of our architecture against state-of-the-art (SOTA) models like Transformers, LSTMs, and 1D CNNs; (2) introducing an insertion/deletion-based parameter inspired by image and NLP techniques to evaluate monomeric contribution scores; and (3) using this metric to compare our model's performance with attention- and gradient-based XAI methods. We mainly evaluate our model's performance on diverse datasets, including Anti-Cancer Peptide (ACP) properties (Sun et al., 2024), protein solubility (Hon et al., 2021), binding affinity (Olson et al., 2014), collagen thermal stability (Khare et al., 2022), and Antimicrobial Peptide (AMP) classification (Gupta & Zou, 2019). Our results demonstrate the model's ability to capture monomeric importance across sequences with varying motif sizes and long-range dependencies. Furthermore, we demonstrate our model's ability to identify critical motifs in ACPs and AMPs and validate these findings through mutation analysis.

2 APPROACH

2.1 DEEP-LEARNING FRAMEWORK

We develop a novel deep learning (DL) architecture with interpretable steps to estimate the contribution of each monomer in the primary sequence to the property. Our DL model uses a **Co**mpositional **L**inear **O**peration-based **R**epresentation (COLOR) unit, a key contribution of the work. A COLOR unit consists of 3 modules namely: sequence-to-motif conversion module, motif composition module, and linear weighted summation module. COLOR unit architecture is shown in Fig.1. Before describing these modules, we introduce the term *number of qualitative variables* (q), representing the total distinct qualitative variables that can appear in the sequence. For proteins, q is typically considered to be 21, accounting for the 20 most common amino acids and one additional for uncommon amino acids such as hydroxyproline and hydroxylysine in collagen protein (Ricard-Blum, 2011). Additionally, the term "motif" will be used frequently throughout the paper and, in this context, refers to any sub-segment of the primary sequence.

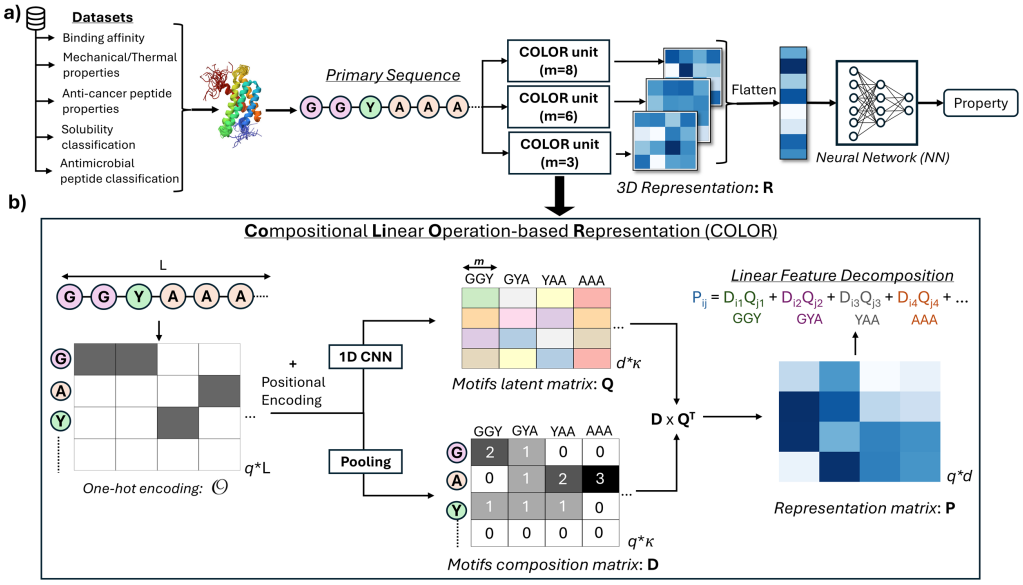

Figure 1: (a) The complete data flow for predicting properties from the primary sequence using COLOR units, (b) Components of the COLOR unit, illustrating the linear decomposition of elements in **P** into motifs, thereby showcasing its interpretability. The m values displayed in the COLOR units are just for reference.

Sequence-to-Motif Module. This module divides the primary sequence into several motifs using a 1D convolution network (CNN) (Kiranyaz et al., 2019). For example, GGYAAA can be divided into motifs GGY, GYA, YAA, and AAA of size 3. The motif size (m) can be controlled by regulating the kernel size in the 1D CNN. The kernel size controls the number of neighboring monomers that the 1D CNN considers in front of each monomer for feature extraction. It is important to note that motifs are created by sweeping with a filter of size m across the primary sequence with the stride of 1. Therefore, the number of motifs (κ) of size m obtained from a primary sequence of length L is $(L - m + 1)$. The 1D CNN uses the one-hot encoded representation ($\mathcal{O} \in R^{\mathrm{q}*L}$) of the primary protein sequence as input (Harding-Larsen et al., 2024). As illustrated in Fig.1, the 1D CNN divides the primary sequence into κ motifs and generates a latent space vector of size d for each motif, leading to matrix $\mathbf{Q} \in \mathbb{R}^{d*\kappa}$.

Motif Composition Module. In this module, the model captures the composition of each motif i.e., it captures the number of different qualitative variables present in each motif. This is obtained by performing a pooling operation (shown in Fig.1) on $\mathcal{O}$ as

$$\mathbf{D}_{ij} = \sum_{k=j}^{j+m} \mathcal{O}_{ik} \tag{1}$$

Linear Weighted Summation Module. Till this stage, the model has converted the primary sequence into motifs and has computed latent space (**Q**) and composition (**D**) matrix. However, to predict the property based on the primary sequence it is important to accumulate the impact of all the motifs. Hence in this module, a representation matrix **P** is obtained by linearly combining the properties of the motifs in the latent space as follows:

$$\mathbf{P} = \mathbf{D} \times \mathbf{Q}^T \tag{2}$$

Element $\mathbf{P}_{ij}$ captures the linear combination of j^{th} latent property (where, j = 1,2...d) of motifs weighted by the quantity of i^{th} (where, i = 1,2...q) qualitative variable in each motif. It is important to note that the size of the matrix **P** is independent of the sequence length L, unlike other architectures like Transformers, LSTM, and 1D CNNs, where the output size depends on L. It is important to note that the linear weighted sum of the property of motifs in the latent space to obtain **P** does not consider the order of occurrence of motifs in the primary sequence. This can lead to the loss of any sequential information and poor prediction. Therefore, we add sine and cosine positional encoding, given in (Vaswani, 2017), to $\mathcal{O}$ before inputting it into the CNN layers.

End-to-End Architecture. In the above sections, we discussed the method to obtain sequence-length independent representation matrix **P** for a particular motif size m using the COLOR unit. However, using **P** pertaining to just one motif size,m, can be insufficient to fully capture the behavior of the protein since motifs of varying size may contribute strongly to a given property. For this reason, we can use several COLOR units to obtain different **P** based on several motif sizes as shown in Fig.1a with the detailed structure of the COLOR unit depicted in Fig.1b. Different **P** matrices can be assembled into a 3D representation matrix **R** . Subsequently **R** is flattened and fed into a fully-connected neural network (NN) to predict the property. Although **R** is a 3D matrix and could theoretically be processed using a CNN to distill information and make predictions, we chose not to use a CNN because there is no meaningful spatial relationship between neighboring elements in **R** . This decision is further supported by an average drop of 4–5% in predictive performance when a fully connected network was replaced with a CNN. The cardinality of the set m, representing the number of COLOR units in **R** , is denoted as $|m|$. It is important to note that the performance and size of the COLOR unit-based model depend on the values of m , $|m|$, and d. Optimal values can be determined through parametric studies, as detailed in Appendix I. As a note, we would like to highlight that whenever the term '*COLOR method*' is used in the text, it refers to the DL model based on COLOR units. Based on our observations throughout this study, we also document certain pointers in Appendix D to better train COLOR-based models.

Quantifying Predictability. To quantify the predictive performance of COLOR, we adopt an approach inspired by (Bornschein et al., 2020) to calculate the area under error versus training data size (N_T) curve. For our analysis, we use two terms calculated as

$$\mathscr{A} = \int_0^{\infty} e(n)dn \ \text{, and } \mathscr{A}_{500} = \int_0^{500} e(n)dn \tag{3}$$

, where $e(n)$ represents the mean absolute error (MAE) for regression tasks and accuracy for classification tasks. In the case of the classification datasets, the classes are balanced; therefore, accuracy is an appropriate metric for comparing the models. The term $\mathscr{A}$ represents the overall predictive performance, whereas the term $\mathscr{A}_{500}$ represents the performance in a low-data regime. In regression tasks, lower values of the terms in Eq.3 indicate a superior model, whereas higher values are better for classification tasks.

2.2 Estimating Monomeric Contribution using COLOR method

As discussed in the Introduction, studying the contribution of monomers in the primary sequence can help estimate the motifs responsible for modulating properties within proteins. The layers of non-linearity in deep-learning models render them uninterpretable to estimate the impact of each monomer on the property. In this section, we show that the architectural decisions in the COLOR unit make it interpretable to calculate the impact of each monomer. Estimating the monomeric contribution based on COLOR is a two-step process. To elucidate these steps, we examine the case with $|m|$=1, which renders **R** =**P** . Let y^p represent the predicted property. As the first step, we estimate the importance of P_{ij} in **P** as $\left|\frac{\partial y^p}{\partial \mathrm{P}_{ij}}\right|$. Based on Eq.2, P_{ij} can be expanded as

$$\mathrm{P}_{ij} = \sum_{k=1}^{\kappa} \mathrm{D}_{ik}\mathrm{Q}_{jk} \tag{4}$$

, where $\mathrm{D}_{ik}\mathrm{Q}_{jk}$ corresponds to k^{th} motif in a sequence. The greater the magnitude of $\mathrm{D}_{ik}\mathrm{Q}_{jk}$, the stronger the influence of the motif on P_{ij}. Hence, the contribution score ϕ_m is assigned to each motif in a sequence as

$$\phi_m = \sum_{i=0}^{q-1} \sum_{j=0}^{\mathscr{d}-1} \left| \frac{\partial y^p}{\partial \mathrm{P}_{ij}} \right| \times \frac{|\mathrm{D}_{im}\mathrm{Q}_{jm}| - \min_k(|\mathrm{D}_{ik}\mathrm{Q}_{jk}|)}{\max_k(|\mathrm{D}_{ik}\mathrm{Q}_{jk}|) - \min_k(|\mathrm{D}_{ik}\mathrm{Q}_{jk}|)} \tag{5}$$

The non-linearity introduced by the neural network (NN) in the model architecture can lead to noisy latent properties (Q_{jk}) for motifs, particularly affecting the Q_{jk} with smaller magnitudes. Hence, to mitigate the effect of such noise on ϕ_m, we apply min-max scaling of $\mathrm{D}_{ik}\mathrm{Q}_{jk}$ in Eq.5, effectively reducing the contribution of noisy, smaller $\mathrm{D}_{ik}\mathrm{Q}_{jk}$ values, to nearly zero. Once ϕ_m is assigned to all motifs, a contribution score will be associated with every monomer in the sequence.

Quantifying Explainability. Due to the absence of any quantitative evaluation metrics for monomeric contribution scores, we propose a novel metric to quantitatively compare contribution scores calculated by different XAI methods. we draw inspiration from the masking-based method in the field of computer vision (Hooker et al., 2019) and NLP (Pham et al., 2022). Firstly, based on ϕ values, the monomers in the sequence are ranked. Subsequently, all monomers are masked except for the top $u\%$, after which the model is re-trained to assess performance. The value of u is incrementally increased, and with each step, the model is re-trained and the error (or accuracy) on the test data is recorded. The area under error (or accuracy) versus u curve, $\mathscr{I}$, calculated as

$$\mathscr{I} = \int_0^{100} e(u)du \tag{6}$$

is used as a metric to evaluate the explainability of the model. The COLOR method is rigorously evaluated against state-of-the-art XAI models, including Grad-CAM (Selvaraju et al., 2017), Attention Tracing (Wu et al., 2020a), and Grad-SAM (Barkan et al., 2021), whose details presented in Appendix A.

2.3 DATASET

We present results for 7 unique properties derived from distinct datasets. These datasets include continuous properties analyzed as regression tasks and categorical properties analyzed as classification tasks. The details of all these datasets will follow shortly. We also conduct additional analysis to further demonstrate the robustness of the COLOR method using two toy datasets and a computational silk dataset (Kim et al., 2023) which are discussed in detail in Appendix C.1. The data split for all the datasets mentioned above is given in Tab.3.

Anti-Cancer Peptide (ACP) Properties. The instability index of the protein captures the intracellular stability of the protein (Guruprasad et al., 1990). (Sun et al., 2024) have constructed a comprehensive dataset documenting the instability index of several ACPs. We utilize this documented instability index for different primary sequences as one of the key datasets. Additionally, we incorporated the Aliphatic Index (Ikai, 1980) and GRAVY Index (Kyte & Doolittle, 1982a) of ACPs from the same database as two distinct datasets, as the ground truth for monomeric contribution scores is known in these cases.

Collagen Melting Temperature (T_m). Collagen is one of the most abundant proteins in animals with numerous applications in medicine (Deshmukh et al., 2016). Khare et al. (2022) have experimentally gathered the melting temperature, directly proportional to thermal stability, of 633 different primary sequences of collagen.

GB1 binding affinity. Olson et al. (2014) developed a dataset containing an experimentally calculated binding affinity of double mutated protein G domain B1 (GB1) to immunoglobulin G fragment crystallizable (IgGFC).

Soluprot. Protein solubility is crucial for the production of various therapeutics (Hon et al., 2021), making it an essential property to predict. Hon et al. (2021) used TargetTrack (Berman et al., 2017) to extract the data on the solubility of proteins in E.coli.

Antimicrobial Peptide (AMP) Classification. AMPs are small molecular peptide that possesses anti-microbial functions against a broad range of microorganisms such as bacteria, fungi, parasites, and viruses. Gupta & Zou (2019) curated a dataset of 5200 short peptides, with 2,600 experimentally verified as AMPs, while the remaining sequences are non-AMPs.

3 Results and Discussion

Predictive Capability. We test the COLOR method on all the datasets discussed in Section 2.3 and compare the results with the current state-of-the-art (SOTA) models such as Transformers, LSTM, and 1D CNN. We also present additional analysis on two toy datasets and a silk dataset in Appendix F. The details of the models used for various datasets are given in Tab.1 and 2. We use metrics $\mathcal{A}$ and $\mathcal{A}_{500}$ derived from the curves shown in Fig.6 for comparing predictive performance. A comparison of models is shown in Fig.2. COLOR outperforms the next best SOTA model by 1-79% across various datasets except for the ACP Instability dataset, where it performs worse by 16%. The lower performance in the case of the instability dataset can be attributed to the lower degree of non-linearity in COLOR as it aggregates the contribution of various motifs through a simple linear operation as shown in Eq.2. However, it will be demonstrated in a later section, that despite the lower predictive performance in the case of this dataset, the model excels at capturing the monomeric contribution within the primary sequence; emphasizing the higher explainability offered by COLOR.

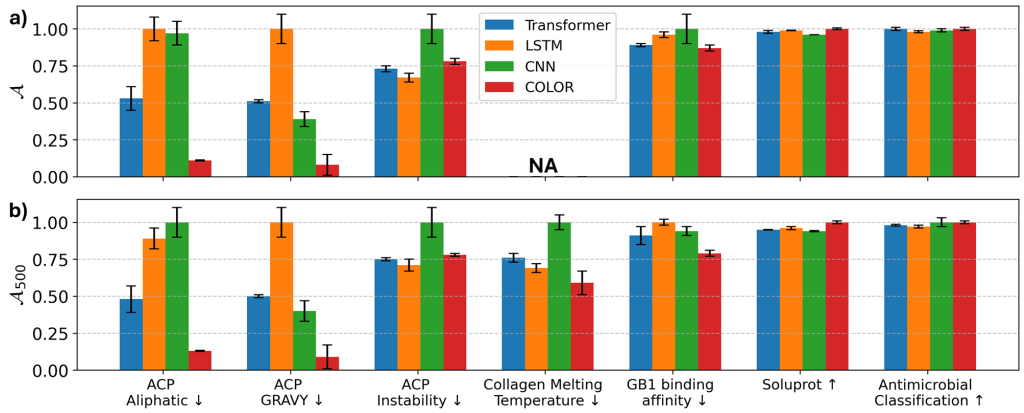

Figure 2: Comparison of the predictive capability of different supervised models. Figure a) and b) shows the comparison of $\mathcal{A}$ and $\mathcal{A}_{500}$ respectively for different datasets. The arrows ↑ and ↓ indicate whether higher or lower values are better, respectively. The results are *normalized* using the highest values of the corresponding dataset.

How Explainable is the model? In this section, we proceed to show COLOR's capability to capture monomeric contribution in the primary sequence as discussed in Section 2.2. For this study, we dropped the GB1 binding affinity and Soluprot dataset. We do not consider the GB1 binding affinity dataset for this study as it contains highly similar sequences with only two mutations in the wild-type protein. Additionally, we also drop the Soluprot dataset as it contains noisy labels for the solubility of proteins (Hon et al., 2021) leading to lower model accuracy as shown in Fig.6. We use random assignment of the monomeric contribution scores within the primary sequence as one of the baseline methods. This random method provides a baseline against which our method should perform better, indicating that it has learned some meaningful information about the sequence. A comparison of COLOR with other SOTA models along with random baseline is given in Fig.3. The $\mathcal{I}$ values reported are obtained from the curves shown in Fig.7. All results are normalized using the results from the random method. The figure shows that the COLOR method achieves the highest performance, outperforming the next-best method by 1–38% across datasets, with an average gain of 22%. Notably, our approach consistently outperforms random baselines, a result not guaranteed by other SOTA models. The above observations suggest that our method offers more explainability, making it more effective in estimating the monomeric contribution within the primary sequence. Additional analyses highlighting the explainability of the COLOR are presented in Appendix G.

Is latent space representation meaningful? To study whether COLOR learns a meaningful latent representation of motifs in matrix $\mathbf{Q}$, we designed two tasks: in the first, we trained the model to predict the sum of monomeric hydropathy (Kyte & Doolittle, 1982b) of the monomers; in the second, we trained it to predict the sum of the isoelectric (pI) point (Ouellette & Rawn, 2014). The pI point of a monomer is reflective of its charge. As the hyperparameters, we fix m and d to be 1, indicating that we generate the latent representation for every single monomer. The choice of d=1 here is sufficient, as it is already known that only a single monomer-level property is necessary to

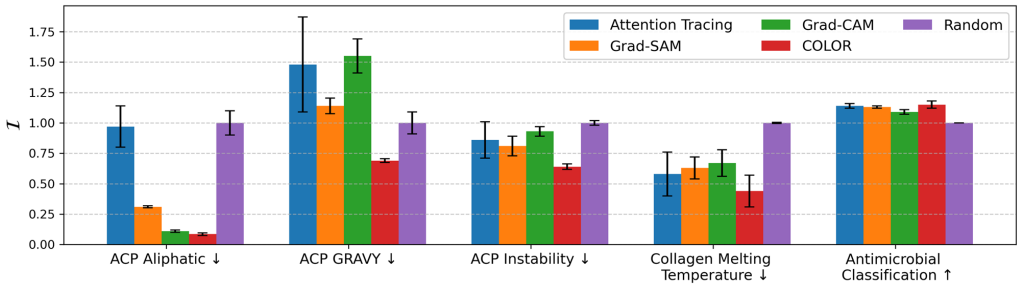

Figure 3: Comparison of explainability offered by different XAI models. The arrows ↑ and ↓ indicate whether higher or lower values are better, respectively.

capture the final property y for the two tasks discussed above. We also replace the neural network with a simple sum of all the elements of $\mathbf{P}$ (i.e., $y^p = \sum_{i,j} \mathbf{P}_{ij}$). After training, upon comparing the actual monomeric hydropathy and pI values with their corresponding latent space representations from $\mathbf{Q}$, we achieve an R^2 of 1.0, indicating that the model learns meaningful and application-specific representations.

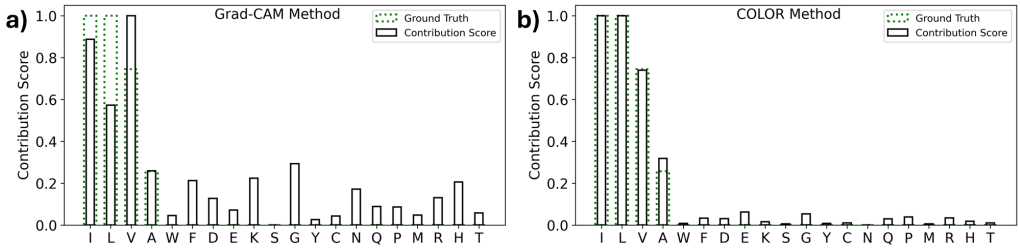

Figure 4: Contribution score of each amino acid in predicting the aliphatic index assigned using a) Grad-CAM method, and b) using COLOR.

Does the contribution score reflect expected patterns? In this section, we study if the contribution score assigned to different amino acids is proportional to their actual contribution to the output y. First, we consider the example of ACP aliphatic index prediction. Given a primary sequence, the aliphatic index is equal to $\chi(A) + 2.9\chi(V) + 3.9(\chi(I) + \chi(L))$, where χ is the amino acid compositional fraction. According to this equation, amino acids isoleucine (I) and leucine (L) have the highest significance, followed by valine (V) and alanine (A). The ϕ values COLOR assigned to amino acids closely follow the expected trend, as shown in Fig.4. In contrast, the Grad-CAM method (second-best) exhibits notable deviations from the expected trend, particularly underestimating the contribution of amino acid L. It is noteworthy that both COLOR and the transformer-based model exhibit excellent predictive capabilities, achieving R^2 >0.99 in both instances. Consequently, the variation in the contribution scores presented in Fig.4 can be attributed solely to the explainability of the respective models. In another study on ACP GRAVY index data, shown in Appendix H, we show that COLOR captures the contribution scores of amino acids 31% accurately.

Application in motif identification. Having quantitatively demonstrated the explainability of COLOR, we now extend its application to motif identification. We first choose the three most unstable peptides from the ACP Instability index dataset and study the monomeric contribution using Grad-SAM (second best) and COLOR. Subsequently, based on the contribution score, we identify the three most important motifs (i_u) in the case of both methods. To validate the impact of identified motifs, all the test sequences, x_t, are mutated to $\tilde{x}_t$ at the three most important positions (p_m) using i_{un}. In short, $\tilde{x}_t = r(x_t, i_u, p_m)$, where r represents the mutation of x_{test} at positions p_m using motif i_u. For a fair comparison between the two methods, we conduct multiple scenarios: in half of the scenarios, the p_m are determined based on contribution scores from the Grad-SAM method, while in the other half, they are selected using the COLOR method. To further avoid any bias, the instability index of $\tilde{x}_t$ is calculated using both methods. The distribution of the instability index of $\tilde{x}_t$ is shown in Fig.5a in comparison with the distribution of the instability index of x_t. The shift is dominant when mutated with i_u identified using the COLOR method (RRR, RSS, and RRI), re-

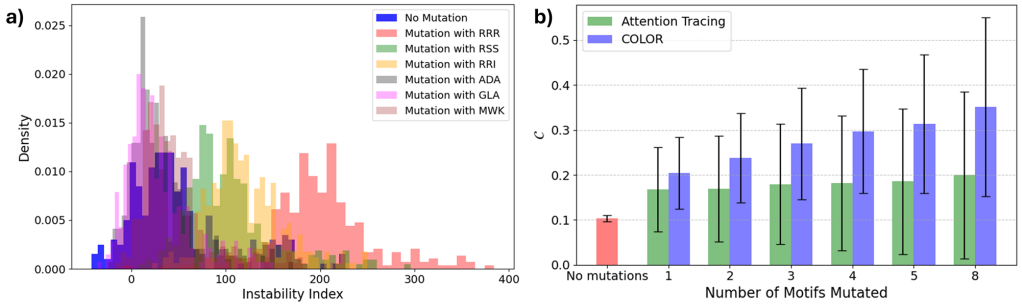

Figure 5: Motif identification and mutation study. a) Illustrates the distribution shift in the instability index of the mutated ACP sequences $\tilde{x}_t$. RRR, RRI, and RSS are the motifs identified by COLOR, b) Depicts the variation of $\mathcal{C}$ with the number of mutations ($|p_m|$) introduced in non-AMPs (x_n).

inforcing its capability to identify impactful motifs. This study also demonstrates that motifs RRR, RSS, and RRI, significantly compromise the stability of ACP.

In a similar study, we identify key motifs in AMP sequences and validate their impact through mutation analyses. We first identify key motifs (i_a) in AMP sequences using Attention tracing (second-best) and COLOR. To evaluate the impact of i_a identified by the two methods, we first select specific positions (p_m) for mutation in non-AMP sequences. Subsequently, all the non-AMP sequences in the test data, x_n, are mutated to $\tilde{x}_n$, where $\tilde{x}_n$=$r(x_n, i_a, p_m)$. It is important to note that for a given i_a, p_m can also be a list for facilitating mutations at multiple positions. Additionally, since $\tilde{x}_n$ is derived from x_n, their cardinality remain the same for a given i_a and p_m, i.e., $|x_n|$=$|\tilde{x}_n|$. Following the mutations, the probability of a $\tilde{x}_n$ belonging to the AMP class, p($\tilde{x}_n \in$AMP class). To avoid bias towards any one model, we follow similar steps as in the case of the ACP dataset above. To quantify the impact of mutation(s), we introduce the variable $\mathcal{C}$ defined as

$$\mathcal{C} = \frac{|\{\tilde{x}_n | p(\tilde{x}_n \in \text{AMP class}) > 0.8\}|}{|\tilde{x}_n|} \tag{7}$$

The term $\mathcal{C}$ represents the fraction of non-AMP sequences that exhibit a high probability (>0.8) of being classified as AMPs following mutation(s). Fig.5b shows the variation of $\mathcal{C}$ with the number of mutations introduced in the non-AMP sequences. The higher $\mathcal{C}$ values observed for motifs identified from AMP sequences using the COLOR method reflect its effectiveness in identifying critical motifs in the sequences. Overall, based on $\mathcal{C}$ values, COLOR demonstrates a 53% mean improvement in the likelihood of converting a non-AMP sequence into an AMP sequence compared to the attention tracing method. Additionally, in Appendix J, we further demonstrate COLOR's enhanced ability to accurately identify motifs in a toy dataset.

4 Limitation

It is evident from Eq.2 that COLOR only captures the linear interaction between all motifs while constructing **P**. Hence, this can affect its predictive capability for sequences that have higher-order interactions. To study COLOR's performance in the dataset with higher-order interactions, we utilize the ACP GRAVY index dataset but change the property y to:

$$y = \sum_{i=1}^{L/2} (\psi_i \times \psi_{L-i})^2 \tag{8}$$

where, ψ_i is the hydropathy of the amino acids present at the i^{th} position. Based on Eq.8, it can be noted that there is higher-order (order=2) interaction between the monomers far away in the sequence; hence making this a good dataset to test COLOR's capability to capture such interactions. Upon training COLOR to predict y in Eq.8, we obtain $< R^2 >$ = 0.88 which is 12% lower than the R^2 value while trained on GRAVY index (order=1). This drop in the performance can be attributed to only capturing linear interactions in Eq.2. Even though we add a fully connected neural network after **R** (Fig.1a) to capture higher-order interactions, the information loss that occurs in Eq.2 about the higher-order interactions cannot be fully retrieved using the neural network. But the Transformer model, the second-best model in this case, also shows a 10% drop in performance

compared to the prediction of the GRAVY index. This highlights that the performance decline when capturing higher-order interactions is common among other state-of-the-art models as well, given the complexity of the task.

Typically, these long-range higher-order interactions arise in proteins due to their complex tertiary structure Kihara (2005). Therefore, to enhance predictive performance, it is essential to explicitly or implicitly incorporate structural information into the model. One approach is to predict the protein's contact map Vendruscolo et al. (1997) and concatenate its elements with rbefore feeding all features into the neural network (NN) for prediction.

In the current method, we have used one-hot encoding ($\mathcal{O}$) to represent the primary sequence. However, this approach can be extended to incorporate other tokenization methods, such as byte-pair encoding (BPE) tokenizer Gage (1994), for sequence optimization. One important consideration when using an alternative tokenization is to control or be aware of the number of monomers that form each token. This will help in accurately assigning the contribution score (ψ) to correct motifs in Eq.5.

5 Conclusion

To address the gap for an explainable model to estimate monomeric contribution in proteins, we developed a novel deep-learning model named COLOR in which every step is interpretable to estimate monomeric contribution scores. Firstly we show that COLOR has superior data efficiency as it outperforms SOTA in a low training data regime (N_T<500) on 7 out of 10 datasets. We also formulate a metric ($\mathcal{J}$) to evaluate the contribution scores calculated by COLOR and compare it against attention and gradient-based explainable models. Our analysis shows that COLOR achieves 22% higher explainability than the SOTA. Therefore, COLOR achieves enhanced explainability without compromising the predictive capability. Additionally, through systematic study, we show that COLOR is more effective in identifying critical motifs in primary sequences. For example, we show that the critical motifs identified by COLOR are 50% more effective in converting non-AMP sequences to AMPs. The motif identification study in our analysis provides the foundation for monomeric contribution score-driven sequence optimization to accelerate the design of $de\ novo$ proteins.

Author Contributions

A.P., S.K., and W.C. conceived the idea. A.P. performed all implementations. A.P., S.K., and W.C. contributed to the manuscript writing.

Acknowledgments

A.P., S.K., and W.C. acknowledge funding from the National Science Foundation's MRSEC program (DMR-2308691) at the Materials Research Center of Northwestern University. A.P. also acknowledges Payal Mohapatra from the Department of Electrical and Computer Engineering at Northwestern University for her valuable input regarding the preparation of this manuscript.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, Sebastian W Bodenstein, David A Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B Fuchs, Hannah Gladman, Rishub Jain, Yousuf A Khan, Caroline M R Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024.

Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera.

Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99:101805, 2023.

Kazuharu Arakawa, Nobuaki Kono, Ali D Malay, Ayaka Tateishi, Nao Ifuku, Hiroyasu Masunaga, Ryota Sato, Kousuke Tsuchiya, Rintaro Ohtoshi, Daniel Pedrazzoli, et al. 1000 spider silkomes: Linking sequences to silk physical properties. *Science advances*, 8(41):eabo6043, 2022.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18 (10):1196–1203, 2021.

Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. Why attentions may not be interpretable? In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 25–34, 2021.

Oren Barkan, Edan Hauon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. Grad-sam: Explaining transformers via gradient self-attention maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 2882–2887, 2021.

Helen M Berman, Margaret J Gabanyi, A Kouranov, DI Micallef, and J Westbrook. Protein structure initiative–targettrack 2000-2017–all data files. *Zenodo. doi*, 10, 2017.

Jorg Bornschein, Francesco Visin, and Simon Osindero. Small data, big decisions: Model selection in the small-data regime. In *International conference on machine learning*, pp. 1035–1044. PMLR, 2020.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Devlina Chakravarty and Lauren L Porter. Alphafold2 fails to predict protein fold switching. *Protein Science*, 31(6):e4353, 2022.

Devlina Chakravarty, Joseph W Schafer, Ethan A Chen, Joseph F Thole, Leslie A Ronish, Myeongsang Lee, and Lauren L Porter. AlphaFold predictions of fold-switched conformations are driven by structure memorization. *Nature Communications*, 15(1):7296, August 2024.

Jiarui Chen, Hong Hin Cheong, and Shirley WI Siu. xdeep-acpep: deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning. *Journal of chemical information and modeling*, 61(8):3789–3803, 2021.

Jiaxiao Chen, Zhonghui Gu, Youjun Xu, Minghua Deng, Luhua Lai, and Jianfeng Pei. Quotetarget: A sequence-based transformer protein language model to identify potentially druggable protein targets. *Protein Science*, 32(2):e4555, 2023.

Ian Connick Covert, Chanwoo Kim, and Su-In Lee. Learning to estimate shapley values with vision transformers. In *ICLR*. OpenReview.net, 2023. URL `http://dblp.uni-trier.de/db/conf/iclr/iclr2023.html#Covert0L23`.

Monica F Danilevicz, Mitchell Gill, Cassandria G Tay Fernandez, Jakob Petereit, Shriprabha R Upadhyaya, Jacqueline Batley, Mohammed Bennamoun, David Edwards, and Philipp E Bayer. Dnabert-based explainable lncrna identification in plant genome assemblies. *Computational and Structural Biotechnology Journal*, 21:5676–5685, 2023.

Shrutal Narendra Deshmukh, Alka M Dive, Rohit Moharil, and Prashant Munde. Enigmatic insight into collagen. *Journal of Oral and Maxillofacial Pathology*, 20(2):276–283, 2016.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

Philip Gage. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, Ramnik J Xavier, Rob Knight, Kyunghyun Cho, and Richard Bonneau. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1):3168, May 2021.

Anvita Gupta and James Zou. Feedback gan for dna optimizes protein functions. *Nature Machine Intelligence*, 1(2):105–111, 2019.

Kunchur Guruprasad, BV Bhasker Reddy, and Madhusudan W Pandit. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering, Design and Selection*, 4(2):155–161, 1990.

David Harding-Larsen, Jonathan Funk, Niklas Gesmar Madsen, Hani Gharabli, Carlos G Acevedo-Rocha, Stanislav Mazurenko, and Ditte Hededam Welner. Protein representations: Encoding biological information for machine learning in biocatalysis. *Biotechnology Advances*, pp. 108459, 2024.

S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

Jiri Hon, Martin Marusiak, Tomas Martinek, Antonin Kunka, Jaroslav Zendulka, David Bednar, and Jiri Damborsky. Soluprot: prediction of soluble protein expression in escherichia coli. *Bioinformatics*, 37(1):23–28, 2021.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Atsushi Ikai. Thermostability and aliphatic index of globular proteins. *The Journal of Biochemistry*, 88(6):1895–1898, 1980.

Mohammad NS Jahromi, Satya M Muddamsetty, Asta Sofie Stage Jarlner, Anna Murphy Høgenhaug, Thomas Gammeltoft-Hansen, and Thomas B Moeslund. Sidu-txt: An xai algorithm for nlp with a holistic assessment approach. *Natural Language Processing Journal*, 7:100078, 2024.

José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, 2020.

José Jiménez-Luna, Miha Skalic, Nils Weskamp, and Gisbert Schneider. Coloring molecules with explainable artificial intelligence for preclinical relevance assessment. *Journal of Chemical Information and Modeling*, 61(3):1083–1094, 2021.

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

Md Rezaul Karim, Tanhim Islam, Md Shajalal, Oya Beyan, Christoph Lange, Michael Cochez, Dietrich Rebholz-Schuhmann, and Stefan Decker. Explainable ai for bioinformatics: Methods, tools and applications. *Briefings in bioinformatics*, 24(5):bbad236, 2023.

Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Explainable deep relational networks for predicting compound–protein affinities and contacts. *Journal of chemical information and modeling*, 61(1):46–66, 2020.

Eesha Khare, Constancio Gonzalez-Obeso, David L Kaplan, and Markus J Buehler. Collagentransformer: end-to-end transformer model to predict thermal stability of collagen triple helices using an nlp approach. *ACS Biomaterials Science & Engineering*, 8(10):4301–4310, 2022.

Daisuke Kihara. The effect of long-range interactions on the secondary structure formation of proteins. *Protein Science*, 14(8):1955–1963, 2005.

Yoonjung Kim, Taeyoung Yoon, Woo B Park, and Sungsoo Na. Predicting mechanical properties of silk from its amino acid sequences via machine learning. *Journal of the Mechanical Behavior of Biomedical Materials*, 140:105739, 2023.

Serkan Kiranyaz, Turker Ince, Osama Abdeljaber, Onur Avci, and Moncef Gabbouj. 1-d convolutional neural networks for signal processing applications. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8360–8364. IEEE, 2019.

Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982a.

Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982b.

Thierry Lefèvre, Marie-Eve Rousseau, and Michel Pézolet. Protein secondary structure and orientation in silk as revealed by raman spectromicroscopy. *Biophysical journal*, 92(8):2885–2895, 2007.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Frank YC Liu, Bo Ni, and Markus J Buehler. Presto: Rapid protein mechanical strength prediction with an end-to-end deep learning model. *Extreme Mechanics Letters*, 55:101803, 2022.

Mingqing Liu, Xuechun Meng, Yiyang Mao, Hongqi Li, and Ji Liu. Redumixdti: Prediction of drug–target interaction with feature redundancy reduction and interpretable attention mechanism. *Journal of Chemical Information and Modeling*, 64(23):8952–8962, 2024.

Nelson RC Monteiro, Carlos JV Simões, Henrique V Ávila, Maryam Abbasi, José L Oliveira, and Joel P Arrais. Explainable deep drug–target representations for binding affinity prediction. *BMC bioinformatics*, 23(1):237, 2022.

Sajid Nazir, Diane M Dickson, and Muhammad Usman Akram. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in Biology and Medicine*, 156:106668, 2023.

C Anders Olson, Nicholas C Wu, and Ren Sun. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current biology*, 24(22):2643–2651, 2014.

Robert J Ouellette and J David Rawn. *Organic chemistry study guide: Key concepts, problems, and solutions*. Elsevier, 2014.

Akash Pandey, Elaine Liu, Jacob Graham, Wei Chen, and Sinan Keten. B-factor prediction in proteins using a sequence-based deep learning model. *Patterns*, 4(9), 2023.

Akash Pandey, Wei Chen, and Sinan Keten. Sequence-based data-constrained deep learning framework to predict spider dragline mechanical properties. *Communications Materials*, 5(1):83, 2024.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. Double trouble: How to not explain a text classifier's decisions using counterfactuals synthesized by masked language models? In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 12–31, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.aacl-main.2.

Sylvie Ricard-Blum. The collagen family. *Cold Spring Harbor perspectives in biology*, 3(1): a004978, 2011.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282.

Xin Sun, Yanchao Liu, Tianyue Ma, Ning Zhu, Xingzhen Lao, and Heng Zheng. Dctpep, the data of cancer therapy peptides. *Scientific Data*, 11(1):541, 2024.

Zhoutong Sun, Qian Liu, Ge Qu, Yan Feng, and Manfred T Reetz. Utility of b-factors in protein science: interpreting rigidity, flexibility, and internal motion and engineering thermostability. *Chemical reviews*, 119(3):1626–1665, 2019.

Sarveswara Rao Vangala, Sowmya Ramaswamy Krishnan, Navneet Bung, Rajgopal Srinivasan, and Arijit Roy. pbrics: a novel fragmentation method for explainable property prediction of drug-like small molecules. *Journal of Chemical Information and Modeling*, 63(16):5066–5076, 2023.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Michele Vendruscolo, Edo Kussell, and Eytan Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2(5):295–306, 1997.

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Moritz Walter, Samuel J Webb, and Valerie J Gillet. Interpreting neural network models for toxicity prediction by extracting learned chemical features. *Journal of Chemical Information and Modeling*, 64(9):3670–3688, 2024.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002.

Zhengxuan Wu, Thanh-Son Nguyen, and Desmond Ong. Structured self-AttentionWeights encode semantics in sentiment analysis. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 255–264, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.24. URL https://aclanthology.org/2020.blackboxnlp-1.24.

Zhengxuan Wu, Thanh-Son Nguyen, and Desmond C Ong. Structured self-attention weights encode semantics in sentiment analysis. *arXiv preprint arXiv:2010.04922*, 2020b.

Zhenxing Wu, Jihong Chen, Yitong Li, Yafeng Deng, Haitao Zhao, Chang-Yu Hsieh, and Tingjun Hou. From black boxes to actionable insights: a perspective on explainable artificial intelligence for scientific discovery. *Journal of Chemical Information and Modeling*, 63(24):7617–7627, 2023.

Yuting Xu, Deeptak Verma, Robert P Sheridan, Andy Liaw, Junshui Ma, Nicholas M Marshall, John McIntosh, Edward C Sherer, Vladimir Svetnik, and Jennifer M Johnston. Deep dive into machine learning models for protein engineering. *Journal of chemical information and modeling*, 60(6): 2773–2790, 2020.

Yuya Yoshikawa and Tomoharu Iwata. Explanation-based training with differentiable insertion/deletion metric-aware regularizers. In *International Conference on Artificial Intelligence and Statistics*, pp. 370–378. PMLR, 2024.

Chi-Hua Yu, Wei Chen, Yu-Hsuan Chiang, Kai Guo, Zaira Martin Moldes, David L Kaplan, and Markus J Buehler. End-to-end deep learning model to predict and design secondary structure content of structural proteins. *ACS biomaterials science & engineering*, 8(3):1156–1165, 2022.

Haopeng Yu, Heng Yang, Wenqing Sun, Zongyun Yan, Xiaofei Yang, Huakun Zhang, Yiliang Ding, and Ke Li. An interpretable rna foundation model for exploring functional rna motifs in plants. *Nature Machine Intelligence*, pp. 1–10, 2024.

Longxi Zhou, Xianglin Meng, Yuxin Huang, Kai Kang, Juexiao Zhou, Yuetan Chu, Haoyang Li, Dexuan Xie, Jiannan Zhang, Weizhen Yang, et al. An interpretable deep learning workflow for discovering subvisual abnormalities in ct scans of covid-19 inpatients and survivors. *Nature Machine Intelligence*, 4(5):494–503, 2022.

APPENDIX

A RELATED EXPLAINABLE MODELS FOR PROTEIN SEQUENCES

With advancements in AI, models such as CNN, LSTM, and Transformers have been extensively used to predict properties based on the primary sequence. However, these models lack interpretability due to the layers of added non-linearity. But recently there have been some developments in the field of Explainable AI (XAI) methods to improve the interpretability of DL models. Based on the techniques discussed in the comprehensive review of the XAI method for biological application by (Karim et al., 2023), we are going to use three methods as the baseline due to their applicability to the sequence-based models. These methods are Grad-CAM (Chen et al., 2021; Monteiro et al., 2022), Attention Tracing (Vig et al., 2021; Danilevicz et al., 2023; Avsec et al., 2021), and Grad-SAM (Barkan et al., 2021). The grad-SAM method has not been used for proteins but is a simple extension of the Attention Tracing method; hence we have included it as our baseline. The description of the baseline methods is as follows:

Grad-CAM:
The Gradient-weighting Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) was initially proposed for producing a visual explanation for the decision made by a CNN-based model for classification tasks. To understand the formulation of Grad-CAM let us consider z_{ji} (where i=1,...L, and j=1,..d) to be the latent space representation of the primary sequence with L and d being the sequence length and latent vector size for every position in the sequence respectively. Also, let y^p be the output predicted by the DL model. Then the importance of each position in the primary sequence can be obtained using 9.

$$\overline{\phi_i} = \left| \frac{1}{d} \sum_j \frac{\partial y^p}{\partial z_{ji}} \right| \tag{9}$$

Attention Tracing:
Self-attention mechanism in Transformers (Vaswani, 2017) has been used to study the importance of different positions in the field of Natural Language Processing (NLP) (Vaswani, 2017), images (Covert et al., 2023), and protein (Vig et al., 2021). Self-attention captures the impact of one-time point (or pixel in the case of images) on another in the sequential data. Transformer consists of several layers and in each layer, self-attention is calculated. Let us indicate the layer number using n and the self-attention matrix of n^{th} layer as $\alpha^{(n)}$. $\alpha^{(n)}$ is calculated as per 10a using the Query (Q) and Key(K) matrix obtained from the primary sequence as indicated in (Vaswani, 2017) with $\alpha^{(n)}_{i \to j}$ indicating the attention position i places on j in n^{th} layer of the transformer. Using $\alpha^{(n)}$, the importance $\overline{\phi_i^{(n)}}$ (where i= 1,..L) can be obtained for every position in the sequence in every layer

Dataset	q	m	d	**Trainable Parameters**
Toy Dataset 1	5	4	8	10,601
Toy Dataset 2	9	[4,8,6,3]	8	48,657
ACP Aliphatic	20	1	20	7,341
ACP GRAVY	20	1	8	7,337
ACP Instability	20	3	20	11,965
Collagen Melting temperature	21	[4,8,6,3]	8	52,433
GB1 binding affinity	20	4	20	37237
Soluprot	20	[4,8,16,24]	20	166,922
AMP classification	4	8	4	18,958
Silk	20	[4,8,6,3]	8	51,089

Table 1: Details for our model for every dataset used in the study.

of the Transformer with 10b.

$$\alpha^{(n)} = \text{softmax}\left(\frac{Q^{(n)}K^{(n)^T}}{d^{0.5}_{K^{(n)}}}\right), \text{ where } \alpha^{(n)} \in R^{L*L} \tag{10a}$$

$$\overline{\phi_i^{(n)}} = \sum_{j=1}^{L} \alpha^{(n)}_{i\to j}\phi_j^{(n+1)} \tag{10b}$$

Using the same 10b, the importance can be propagated from the topmost layer of the transformer to the input of the transformer. The importance of the input to the transformer ($\overline{\phi_i^{(1)}}$) indicates the importance of positions in the primary sequence.

Grad-SAM:
Gradient Self-Attention Map (Grad-SAM) (Barkan et al., 2021) has been employed in NLP to identify the input elements that explain the model prediction. Grad-SAM extends the attention tracing method by incorporating an additional gradient term into 10b, yielding the formulation in 11.

$$\overline{\phi_i^{(n)}} = \sum_{j=1}^{L} \alpha^{(n)}_{i\to j}\phi_j^{(n+1)} \left|\frac{\partial y_p}{\partial \alpha^{(n)}_{i\to j}}\right| \tag{11}$$

Including the gradient term helps capture not only the absolute value of the self-attention ($\alpha_{i\to j}$) but also the sensitivity of the output (y_p) with respect to these self-attention values.

B Model details

The details of the COLOR model used for different datasets are provided in Tab.1, and those of other baselines are given in Tab.2.

C Additional Analysis

We have designed these toy datasets such that the output properties (y) for each primary sequence are determined through an analytical formulation, providing an exact ground truth for model evaluation. This approach is particularly useful for conducting monomeric contribution studies, where the primary sequence is masked based on its contribution score, and the properties are re-evaluated as described in Section 2.2 in the main text. In the case of the toy dataset, re-evaluating the properties is straightforward due to the availability of an analytical formulation.

C.1 Datasets

C.1.1 Toy Dataset 1

In this dataset, the primary sequence consists of 5 qualitative (i.e., categorical) variables: A, B, C, D, and E. At each i^{th} position, the ψ_i is assigned to account for 2 neighboring positions in the sequence

	Models		
Dataset	**Transformer**	**1D CNN**	**LSTM**
Toy Dataset 1	159K	137K	115K
Toy Dataset 2	138K	138K	115K
ACP Aliphatic	139K	87K	139K
ACP GRAVY	139K	87K	139K
ACP Instability	158K	139K	139K
Collagen Melting temperature	158K	139K	139K
GB1 binding affinity	158K	139K	139K
Soluprot	158K	139K	139K
Silk	158K	-	-
Antimicrobial Classification	159K	138K	163K

Table 2: Details of state-of-the-art models used for comparison with our model. For the silk dataset, there is no data available for 1D CNN and LSTM as this dataset was only used for the explainability study.

as:

$$\psi_i = (a_{i-1} + a_i) * a_{i+1} \tag{12}$$

, where, ψ_0 and ψ_L are equal to 0. The term a_i takes on values 5,2,4,1, or 8 depending on whether A, B, C, D, or E is present at that position. The property y is calculated as

$$y = \sum_{i=1}^{L} \psi_i \tag{13}$$

The L of all the primary sequences is 50.

C.1.2 TOY DATASET 2

In this dataset, the primary sequence consists of 9 distinct qualitative variables: A, B, C, D, E, F, G, H, and I. At each i^{th} position, the descriptor a_i is assigned, where a_i takes on values 5, 2, 4, 1, 8, 10, 7, 6, or 3 depending on whether A, B, C, D, E, F, G, H, or I is present at that position. The descriptor at each position is further refined to ψ_i to incorporate the influence of neighboring variables in the sequence as:

$$\psi_i = \sum_{j=i-b_i/2}^{i+b_i/2} a_j \tag{14}$$

, where, b_i takes values 10, 12, 14, 16, 18, 20, 22, 24, or 26, depending on whether A, B, C, D, E, F, G, H, or I is present at the i^{th} position. The property y is calculated as:

$$y = \sum_{i=1}^{L} \psi_i \tag{15}$$

The L of all the primary sequences is 50.

C.1.3 SPIDER SILK PEAK FORCE

Silk has superior mechanical properties, making it a good choice for designing biomaterials. (Kim et al., 2023) collected the peak force data using Molecular Dynamics (MD) simulation for 82 different primary sequences of silk mimicking MaSp1 spidroin of the spider silk (Arakawa et al., 2022). We use this computational data as one of the datasets for our study. For simplicity, we will refer to this dataset as the "Silk dataset".

The data split for the above-mentioned datasets is given in Tab.3.

Dataset	Training	Validation	Test
ACP Aliphatic, GRAVY, Instability	850	150	150
Collagen Melting Temperature	506	63	64
GB1 Binding Affinity	10000	5255	5308
Soluprot	8336	3100	3100
AMP Classification	3200	1000	1000
Toy Dataset 1,2	1000	100	100
Silk	50	15	15

Table 3: Data split for different datasets used in the current study.

D Strategies for Enhancing Model Performance

In Section 2.1, we discuss the overall architecture of our model. Based on the architecture, we outline several strategies to enhance model performance across various applications:

- Adjust the motif size m to identify the optimal choice for different datasets, as they may contain motifs of varying sizes.
- Tune the latent space dimension d to balance representation capacity and model complexity.
- Increase the number of COLOR units, each with unique motif sizes m to capture diverse motif structures.
- Apply layer normalization after each COLOR unit for enhanced optimization in certain applications.
- Introduce layer normalization following the 3D representation $\mathbf{R}$ for further optimization benefits.
- Normalize the motif composition matrix $\mathbf{D}$ by the sequence length to ensure the model processes relative compositions rather than absolute values. This approach is particularly advantageous for properties like the Aliphatic Index, which depends on the relative proportions of amino acids A, V, I, and L.

E Predictability and Explainability Curves

The MAE (or accuracy) versus N_T curves are shown in Fig.6, and these curves are used to calculate A and A_{500}. The metric $\mathcal{I}$ is calculated from MAE (or accuracy) versus the percentage of sequence unmasked ($u\%$) curves, and these curves are shown in Fig.7.

F Comprehensive Predictive Capability

To study the predictive capability of COLOR, we use $\mathcal{A}$ and $\mathcal{A}_{500}$, as introduced in Section 2.2 in the Main text. The $\mathcal{A}$ and $\mathcal{A}_{500}$ values for Toy dataset 1 and 2 are plotted in Fig.8 along with all other datasets discussed in the Main text. On the toy datasets, COLOR outperforms the next-best SOTA model by 34% on average. We do not perform this study on the silk dataset due to the insufficiency of training samples to calculate either $\mathcal{A}$ or $\mathcal{A}_{500}$. An important observation from Fig.8 is the strong performance of COLOR on Toy Dataset 2. In this dataset, the monomeric properties (ψ) depend on a larger neighborhood of monomers, as defined by 12. Interestingly, despite using small values of m (3,4,6,8) in the model, COLOR accurately predicts the property y as evident by the lower values of $\mathcal{A}$ and $\mathcal{A}_{500}$. This highlights the model's ability to capture properties at the motif level through $\mathbf{Q}$, while the matrix multiplication of $\mathbf{D}$ and $\mathbf{Q}^T$, combined with the non-linearities in the neural network, enables the model to effectively capture global interactions.

G Comprehensive Explainability

We study the explainability of the COLOR method in estimating the monomeric contribution scores for two toy datasets and the silk dataset. Even though we do not perform the predictive study on the

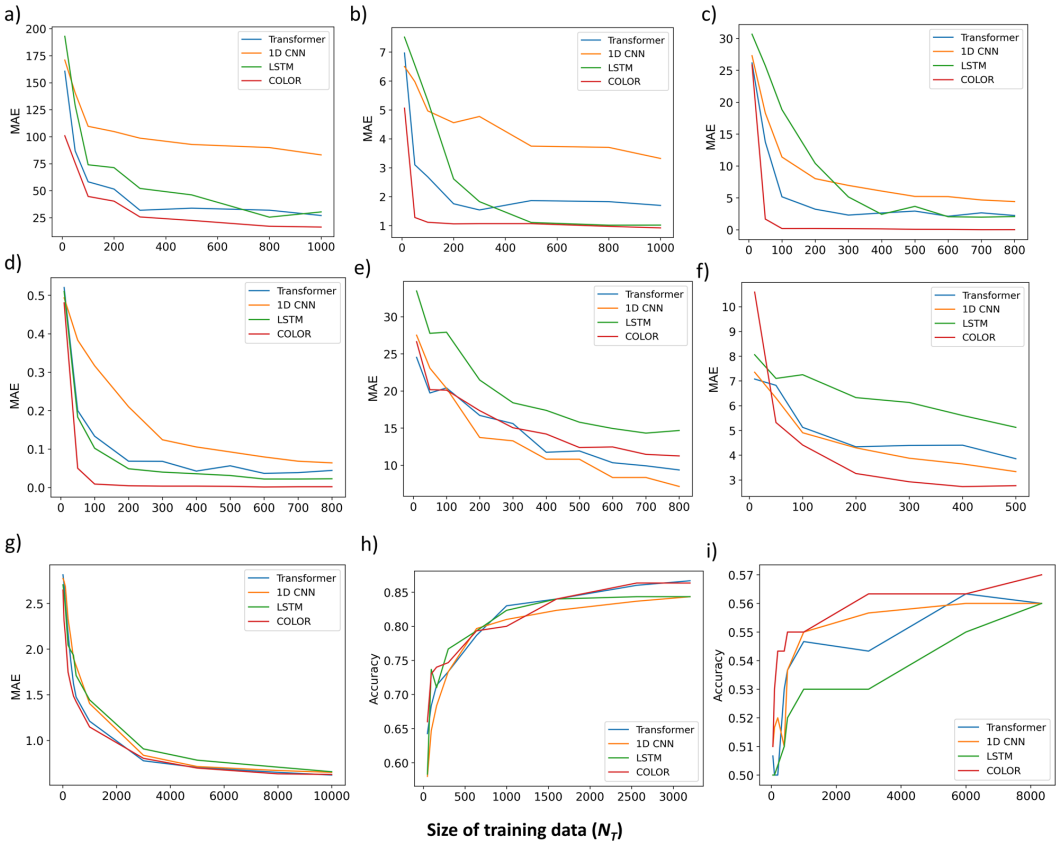

Figure 6: Curves showing the predictive capability of different models. MAE (or accuracy) versus training data size (N_T) curves obtained from different deep-learning models are plotted for a) Toy dataset 1, b) Toy dataset 2, c) ACP Aliphatic index dataset, d) ACP GRAVY index dataset, e) ACP instability dataset, f) Collagen dataset, g) Silk dataset, h) Antimicrobial classification dataset. and i) Soluprot dataset. The curves shown here are the mean of runs using three different seeds.

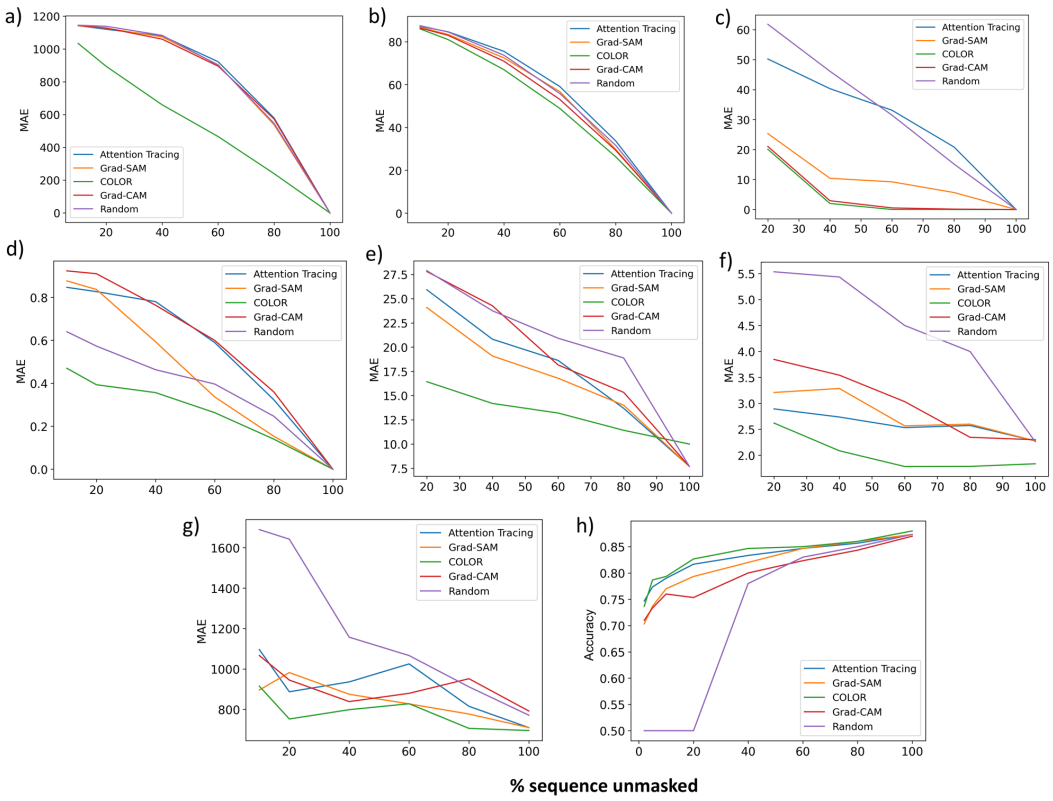

Figure 7: Curves showing the explainability of different models. MAE (or accuracy) versus % of sequence unmasked curve obtained from different XAI models are plotted for a) Toy dataset 1, b) Toy dataset 2, c) ACP Aliphatic index dataset, d) ACP GRAVY index dataset, e) ACP instability dataset, f) Collagen dataset, g) Silk dataset, and h) Antimicrobial classification dataset. The curves shown here are the mean of runs using three different seeds.

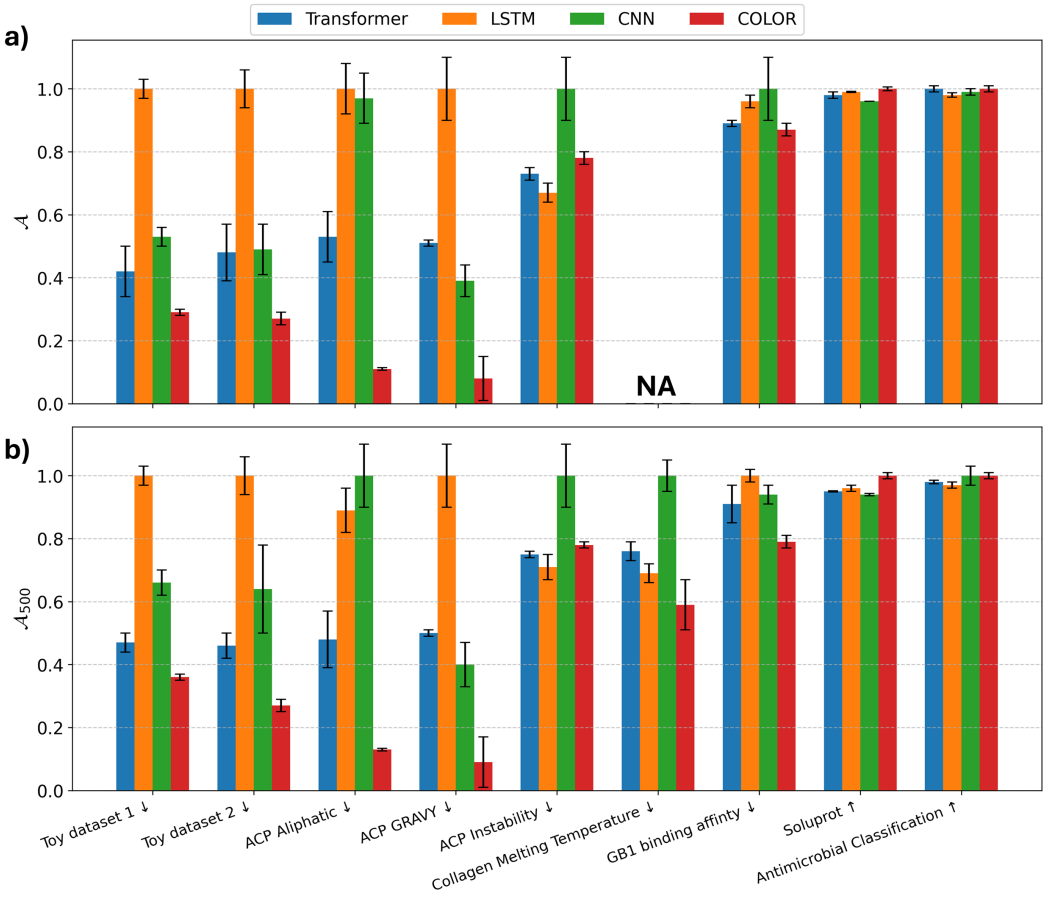

Figure 8: Comparison of the predictive capability of different supervised models. Figure a) shows the comparison of $\mathcal{A}$ and b) illustrates the comparison of $\mathcal{A}_{500}$ obtained for different datasets. The arrows $\uparrow$ and $\downarrow$ in front of dataset names indicate whether higher or lower values are better, respectively. The results are *normalized* using the highest values of the corresponding dataset.

Silk dataset, we include it in the explainability study. Using the data split provided in Tab.3 for the Silk dataset, the mean R^2 value of 0.91 is achieved across all models, indicating strong predictive performance and making the dataset suitable for the explainability study. We use the metric $\mathcal{I}$ discussed in Section 2.2 in the Main text to quantify the explainability. The values of $\mathcal{I}$ are shown in Fig.9 along with all the datasets discussed in the Main text. The COLOR method outperforms the next-best method by 37%, 5%, and 14% on Toy Dataset 1, Toy Dataset 2, and the Silk dataset, respectively.

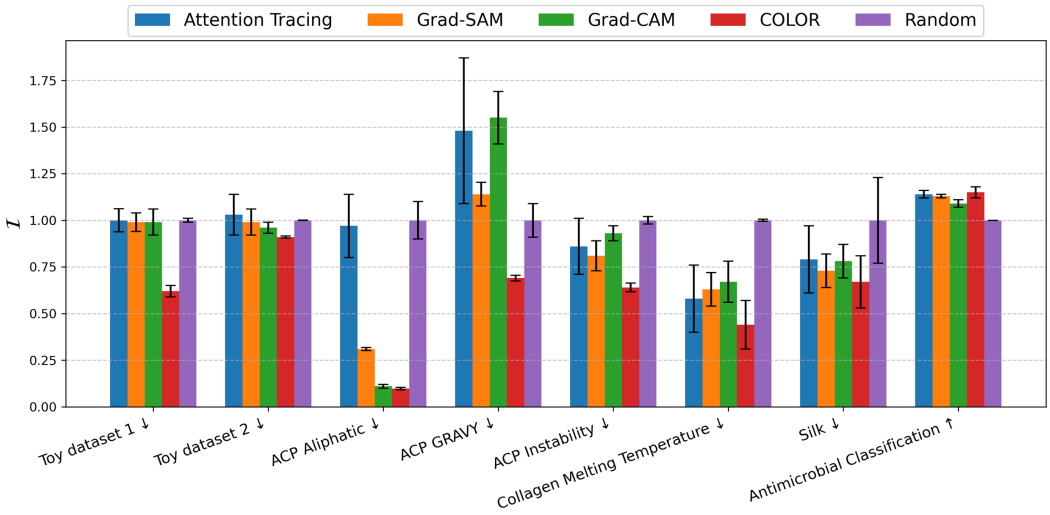

Figure 9: Comparison of explainability offered by different XAI models. The results are normalized using the results from the *Random method* of the corresponding dataset. The arrows ↑ and ↓ in front of dataset names indicate whether higher or lower values are better, respectively.

H Contribution score of amino acids in GRAVY index Dataset

Analytically, the GRAVY index is the sum of the hydropathy value of all the amino acids, divided by the sequence length (L). Fig.10 shows the comparison between the contribution scores and absolute hydropathy of amino acids, wherein we anticipate a strong correlation between the two variables. In Fig.10a & b, the contribution scores are obtained using Grad-SAM (second best interpretable model) and our method, respectively. The contribution scores from COLOR have $\sim$31% higher correlation with the absolute hydropathy value, highlighting the effectiveness of our approach in accurately capturing the significance of various amino acids. It is again important to note that both COLOR and the transformer-based model exhibit high predictive capabilities, with R^2 >0.99. Therefore, the differences illustrated in Fig. 10 arise from the varying explainability of the two models.

I Parametric Study

As discussed in Section 2.1, the motif size m and the dimensionality of the latent space representation d for each motif are key parameters that define a COLOR unit. Hence, in this section, we conduct a parametric study to examine the impact of varying m and d on the model's predictive performance and explainability. To quantify explainability in the parametric study, we introduce a scaled parameter $\overline{M_r}$, defined as

$$\begin{aligned} M_r =& \text{Predictive performance with only 20\% sequence unmasked} \\ \overline{M_r} =& \frac{M_r}{\max_{m \in \mathcal{S}_m}(M_r)}, \text{ where, } \mathcal{S}_m = [1,2,3,6,8,12,18,24] \end{aligned} \tag{16}$$

While computing M_r in Eq.16, the monomers within the top 20% based on their contribution scores are unmasked, while the remainder of the sequence is masked. To evaluate predictive performance,

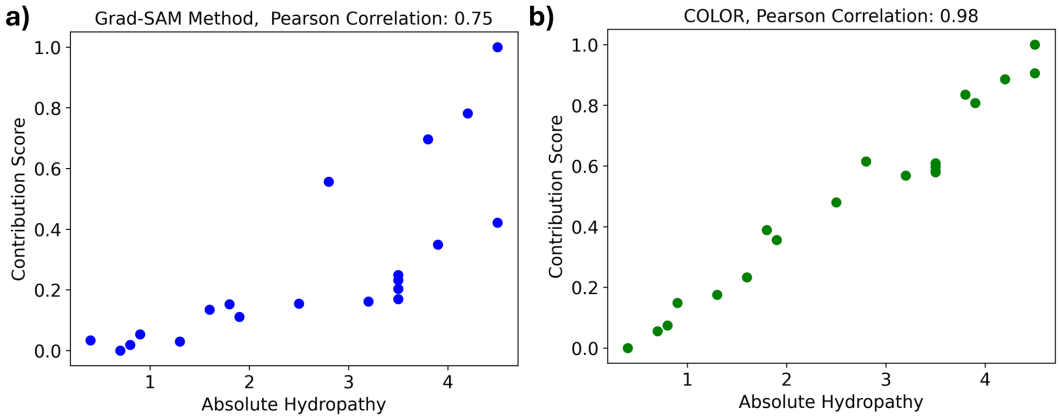

Figure 10: Comparison of contribution scores as a function of monomer hydropathy. a) contribution scores obtained using the Grad-SAM method, and b) contribution scores obtained using the COLOR method.

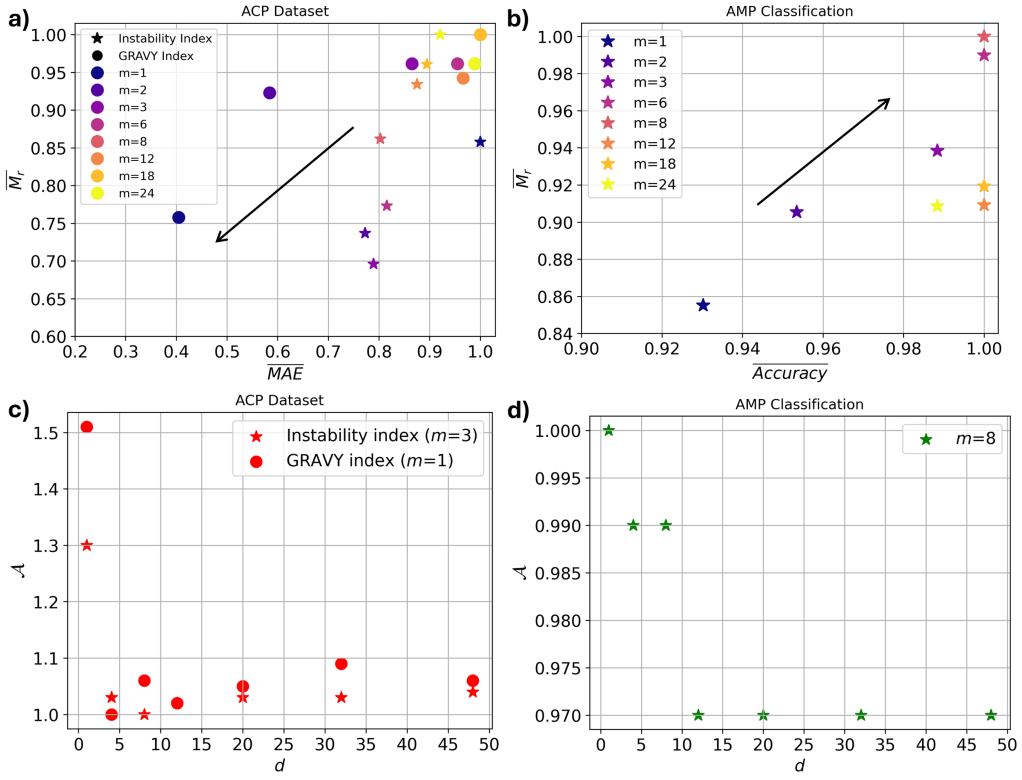

Figure 11: Parametric study depicting the effect of m and d on COLOR's predictive performance and explainability. Figures (a) and (b) show the effect of m on the ACP dataset (Instability and GRAVY index) and AMP classification dataset, respectively, with mean results from three independent runs. Black arrows indicate the direction in which optimal values should trend. Figures (c) and (d) depict the effect of d on predictive performance, where red and green markers indicate lower and higher values corresponding to better model performance, respectively.

we utilize scaled MAE ($\overline{MAE}$) for regression tasks and scaled accuracy ($\overline{Accuracy}$) for classification tasks. The scaling process for MAE and accuracy follows the same approach as outlined in Eq.16. Furthermore, to examine the effect of m, we fix $|m|$ to 1 and vary the value of m. This approach enables us to isolate the influence of motif size on the performance of COLOR.

In Fig.11a&b, we show the effect of m on the model's predictive performance and explainability. For the parametric study in regression tasks, we selected two properties (Instability index and GRAVY index) from the ACP dataset. The GRAVY index was specifically chosen because $m = 1$ is sufficient for accurate prediction, making it an interesting case to explore the impact of increasing m on predictive performance. It is evident from Fig.11a&b that explainability drops while using $m \geq$12 in the case of instability index and AMP classification. This can be attributed to the fact that in the COLOR method, the contribution score $\overline{\phi}$ is assigned at the motif level as per Eq.5. This means that while working with larger m values, the model can end up assigning a higher contribution score to a larger motif of which only a smaller segment is important and the rest of the motif is insignificant. For the same reason, in the case of the GRAVY Index, which is independent of any interactions between neighboring monomers in the sequence, using any m >1 adversely affects explainability as evident from Fig.11a. It can also be noted that in the case of instability index and AMP classification, the model's predictive capability is lower while using m =1. This is because the model does not consider any neighboring monomers while generating matrix $\mathbf{Q}$ and linearly combining the effect of various monomers in Eq.2 might not be sufficient to capture the effect of neighboring monomers in the sequence.

There is also a subtle but important difference in the effect of m on ACP instability and AMP classification dataset. For the ACP instability dataset, a smaller motif size (3< m <6) yields optimal performance, whereas the AMP classification dataset requires a larger motif size (m=6 or 8) for better results. This difference can be attributed to the nature of the datasets: the AMP dataset consists of sequences made up of nucleotide bases (A, G, C, and T), where every three nucleotides correspond to a single amino acid. Since amino acids are crucial for protein function, a larger motif size in nucleotide sequence is necessary to capture 2–3 amino acids in each motif, thereby improving both predictive accuracy and explainability.

For a fixed value of m , varying d can lead to different predictive capabilities. Since d determines the size of the vector representing each motif in matrix $\mathbf{Q}$, adjusting it primarily impacts the number of tunable parameters, thus influencing the model's predictive capability. To study the impact of d, we first fix m to be 1, 3, and 8 for the GRAVY index, Instability index, and AMP classification datasets respectively based on the results shown in Fig.11 a&b. In Fig.11 c&d, we show the variation of $\mathcal{A}$ (see Eq.3) for ACP and AMP datasets. From the figure, it can be noted that in the case of the ACP dataset, choosing d >4 is a robust choice for better predictive performance. On the other hand, for the AMP dataset, $\mathcal{A}$ for all d values are within 3% of each other. The better performance observed with lower d values in the AMP dataset could be attributed to the composition of the sequences, which consist of nucleotides (q = 4), necessitating a lower dimensionality (d) for effective representation. This study suggests that users may consider setting d proportional to q .

J Motif Identification in a Toy Dataset

We show the capability of COLOR to effectively identify important motifs using Toy Dataset 1 as the critical motif is correctly known in this case. Fig.12 showcases the capability of Grad-SAM and COLOR methods in accurately pinpointing the critical motifs within the sequences. The Grad-SAM method is chosen for the comparison as it ranks the second-best interpretable model for toy dataset 1 as shown in Fig.9. Based on the results in Fig.12, COLOR successfully identifies the most important motif in 8 out of 10 instances, compared to 6 out of 10 for the Grad-SAM method.

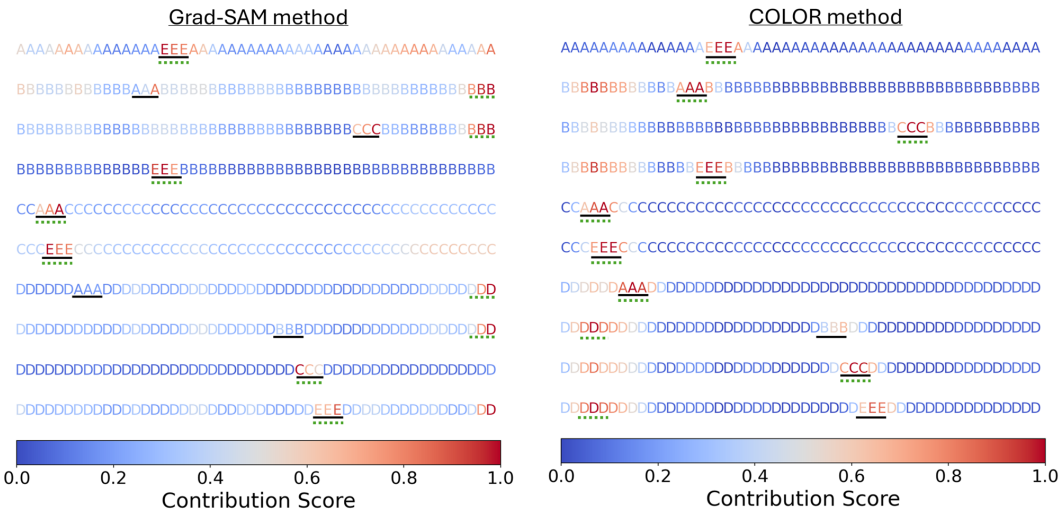

Figure 12: Motif identification in sequences from Toy dataset 1. The figure highlights the most contributing motif in each sequence with a black underline. Every monomer in the sequence is color-coded based on its contribution score, as determined using the Grad-SAM and COLOR methods. Based on the contribution scores, the key motif identified by Grad-SAM and COLOR methods is shown using green underline. With a good interpretable model, the black solid and green dashed underlines are expected to overlap more frequently.